THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

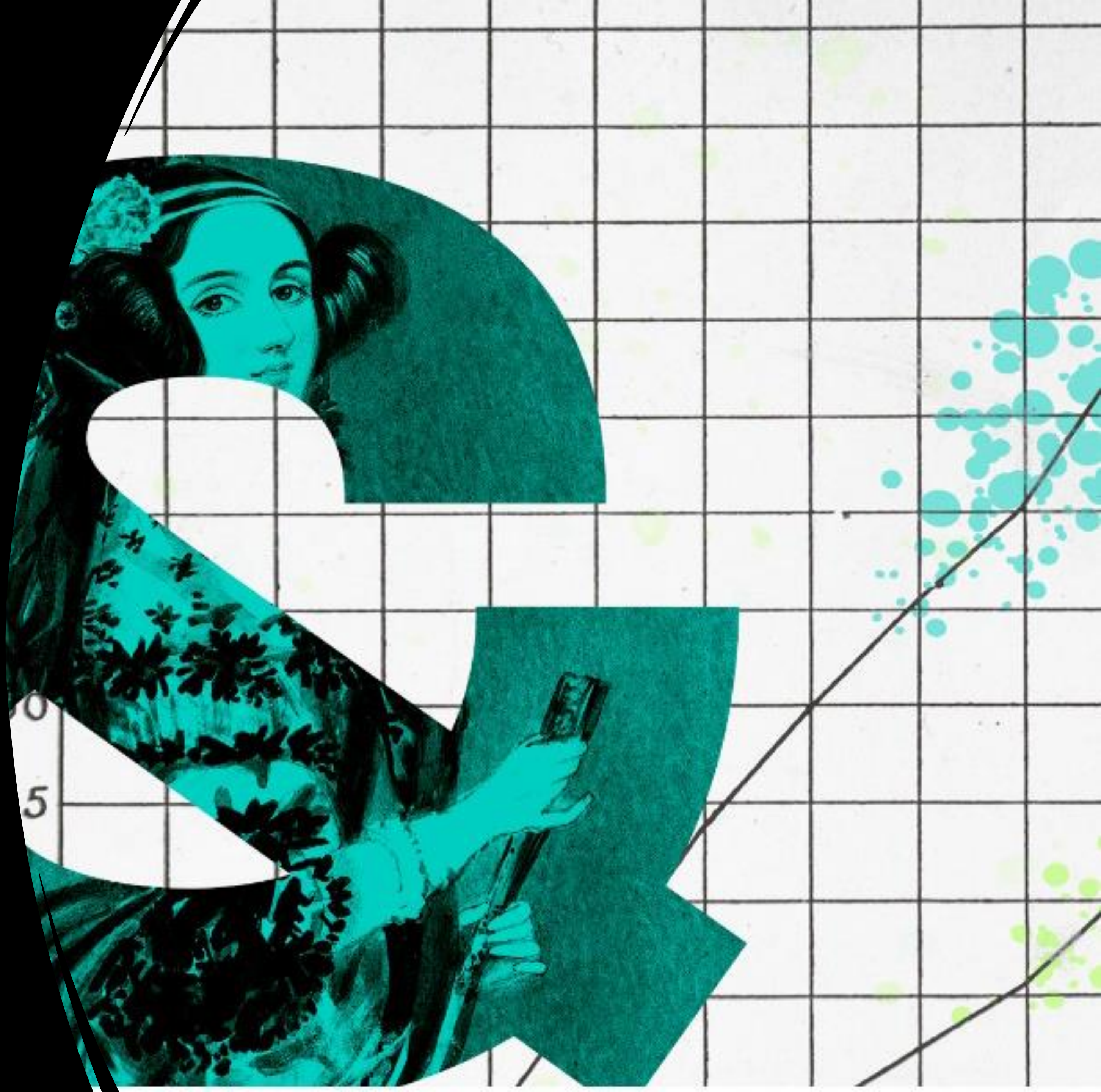
Regression & Mixed-Effects Modelling with R

Fang Jackson-Yang & Aislinn Keogh

13 May 2025

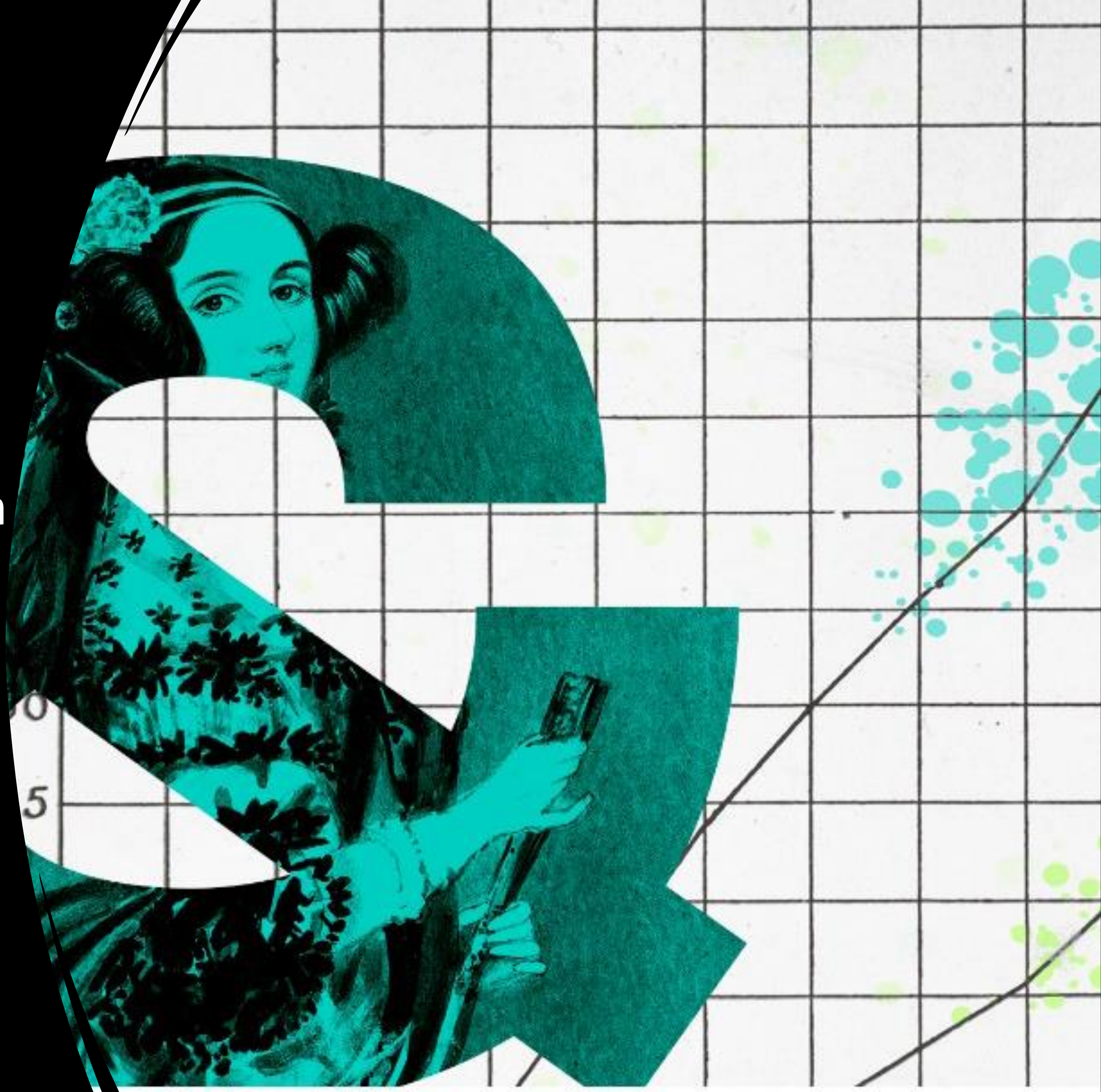
Course outline

- Session 1. Simple Regression; Individual Difference; Intro to Linear mixed-effect models (LMMs)
- Session 2. LMMs (lmer); Generalised LMMs (glmer)
- Session 3. Practical
- Session 4. Ggeneralised LMMs continued; Model assumptions and diagnostics
- Session 5. Practical



Session 3
Roadmap (today)

- Model Comparison & Selection
- Generalised LMMs
- Model Assumptions
- Exercise and Q&A



Model Comparison and Selection

Recall: Structure of LMMS

- *R* code:

  lmer (y ~ $\boxed{x1 \ * x2}$ +
  ($\boxed{1 \ + x1 \ * x2}$ | $\boxed{\text{Grouping}_1}$) +
  ($\boxed{1 \ + x1}$ | $\boxed{\text{Grouping}_2}$) ,
  data = datafilename)

Recall: Model Fit

- No R^2
- Use Maximum likelihood Ratio test to compare models
- R code: anova(model1, modle2)

```
anova(mMixed1_pval, mMixed_reduced)


## Data: vocabdata
## Models:
## mMixed1_pval: vocab_test_score ~ week * proficiency + (1 | participant)
## mMixed_reduced: vocab_test_score ~ week * proficiency + (1 + week | participant)
##                npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## mMixed1_pval      6 2451.4 2474.8 -1219.7   2439.4
## mMixed_reduced    8 2416.2 2447.5 -1200.1   2400.2 39.151  2  3.151e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- NB: when using it to test random structures, make sure to set REML = T

Model Comparison & Selection

- In what order should you build your models?

- Which model should you select as final model?
  - **Barr, Levy, Scheepers, & Tily, (2013) :** keep it maximal
  - **Matuschek, Kliegl, Vasishth, Baayen & Bates (2017) :** make it parsimonious to balance Type1 error and power

Common Issues

- Convergence

```
Warning message:
In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge with max|grad| = 0.0206805 (tol = 0.002, component 1)
```

- Overfitting

```
boundary (singular) fit: see ?isSingular
```

Deal with Convergence Issue

```
Warning message:
In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge with max|grad| = 0.0206805 (tol = 0.002, component 1)
```

Solutions:

- Adjust stopping (convergence) tolerances for the nonlinear optimizer, using the optCtrl() argument to lmerControl.

- Centre and standardise continuous predictor variables

  - the scale() function

Deal with Overfitting Issue

```
boundary (singular) fit: see ?isSingular
```

Solutions:

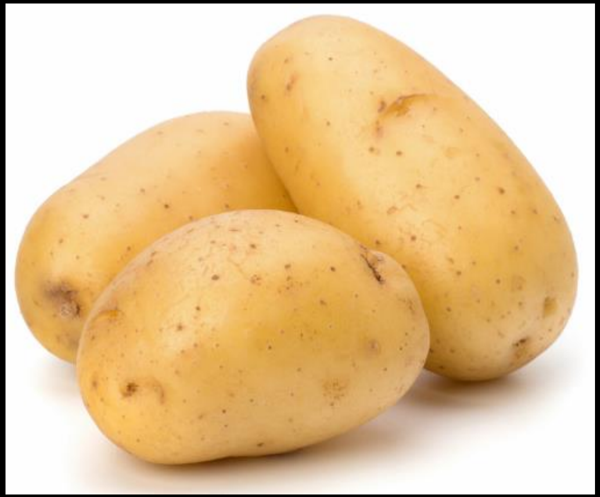
- Remove the most complex part of the random effects structure (i.e. random slopes)
- Maybe acceptable to remove a specific random effect term when its variance estimates are very low

Generalised Linear Mixed-effects Models - glmer()

Binary Outcome

- Pass or Fail
- Invest or Not
- Vote ‘Yes’ or ‘No’
- ‘Correct’ or ‘incorrect’ answer
- Fixation on the target image or not
- Align with partner or not
    - Lexical choice
    - Grammatical choice

‘potato’
or
‘tattie’



(a) *Gromit gave Wallace some cheese.*
(b) *Gromit gave some cheese to Wallace.*

Generalised Linear Mixed-effects Models

- glmer () for binary outcome
- *R* code:

glmer (y ~ $x1 * x2$ +

(1 + $x1 * x2$ | Grouping_1) +

(1 + $x1$ | Grouping_2) ,

data = datafilename,

family = ‘binomial’

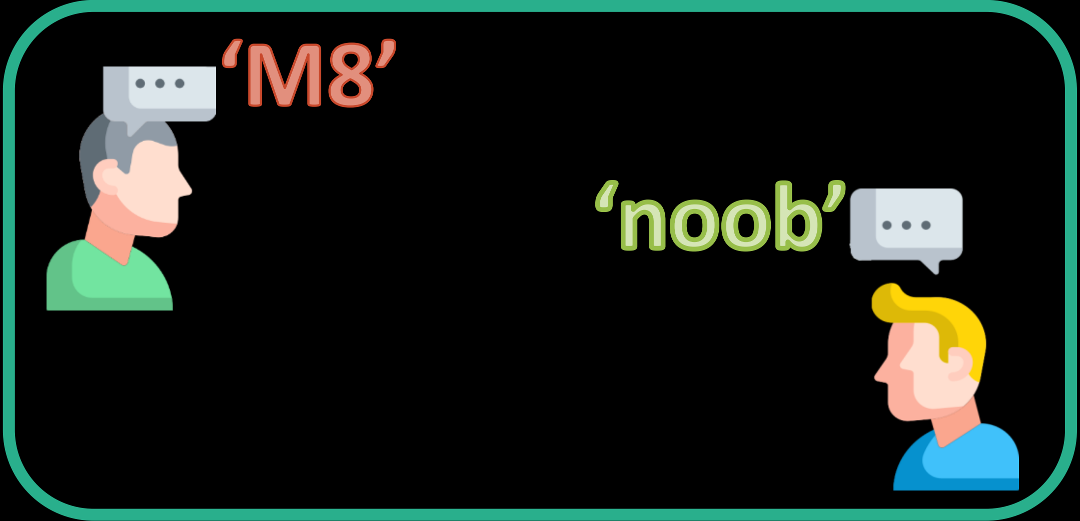
)

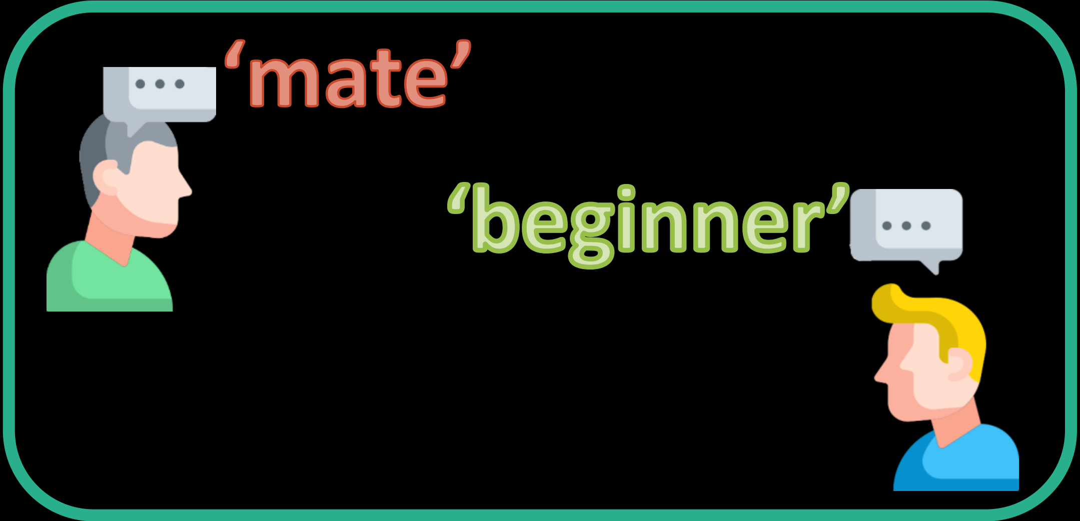
Example: Lexical Choice Study

‘Noob’ or ‘Beginner’ ?

- Internet slang data (simulated data)

- Hypothesis: People were more likely to use internet slang words (as relative to alternative expressions) after seeing their conversational partner uses a slang rather than a stand word.





Example: Lexical Choice Study

- Repeated measure (36 items/participant)
- One predictor (3 conditions):
  'prime' = slang / standard words / other
- Two sources of random variance

Filter

	Participant	Item	List	Trial	Prime	Response
15	1	11	1	81	standard	0
16	1	11	1	109	other	0
17	1	12	1	8	slang	0
18	1	13	1	106	standard	0
19	1	14	1	25	other	0
20	1	15	1	1	slang	1
21	1	16	1	103	standard	0
22	1	17	1	38	slang	1
23	1	18	1	49	standard	0
24	1	18	1	70	other	0
25	1	19	1	42	other	0
26	1	21	1	46	slang	1
27	1	22	1	87	standard	0
28	1	23	1	18	other	0
29	1	24	1	120	slang	0
30	1	25	1	116	standard	0

Showing 15 to 32 of 5,434 entries, 6 total columns

```
> summary(data_ISD)
  Participant            Item            List           Trial              Prime        Response
 1       :  36    18       : 248    6       : 457    Min.   :  1.00    other   :1787    0   :2874
 2       :  36    14       : 234    1       : 410    1st Qu.: 31.00    standard:1833    1   :2512
 5       :  36    28       : 215    5       : 401    Median : 63.00    slang   :1814    NA's:  48
 6       :  36    2        : 208    10      : 386    Mean   : 62.45
 8       :  36    19       : 208    14      : 377    3rd Qu.: 95.00
 10      :  36    12       : 201    11      : 374    Max.   :123.00
 (Other):5218    (Other):4120    (Other):3029
```

Build a Glmer Model

Model <- **glmer** ($response \sim prime$ +

(1 | $participant$) +

(1 | $item$) ,

data = $dataISD$,

family = ‘binomial’

)

Interpret Coefficients

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Response ~ Prime + (1 | Participant) + (1 | Item)
   Data: data_ISD
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

     AIC      BIC   logLik deviance df.resid
  5625.4   5658.3  -2807.7   5615.4     5381

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.9925 -0.5758 -0.2008  0.6024  4.8213

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 2.4345   1.5603
 Item        (Intercept) 0.1262   0.3552
Number of obs: 5386, groups:  Participant, 160; Item, 30

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.27907    0.15230  -1.832   0.0669 .
Primestandar  -0.93888    0.08594 -10.925   <2e-16 ***
Primeslang     1.12819    0.08583  13.145   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) Prmstn
Primestndrd -0.266
Primeslang  -0.277  0.441
```

Use $\beta 0$ to calculate the probability of the intercept:

$$\frac{e^{-0.28}}{1+e^{-0.28}} = 0.43$$

Use $\beta 1$ to get the probability of using standard form (slope 1):

$$\frac{e^{(-0.28+(-0.94))}}{1+e^{(-0.28+(-0.94))}} = 0.23$$

Use $\beta 2$ to get the probability of using slang (slope 2):

$$\frac{e^{(-0.28+1.13)}}{1+e^{(-0.28+1.13)}} = 0.70$$

- Coefficients are in logit units (log-odds)
- Can be transformed to probabilities: $prob(x) = \frac{e^x}{1+e^x}$

Model Assumptions

Model Assumptions: Linear regression

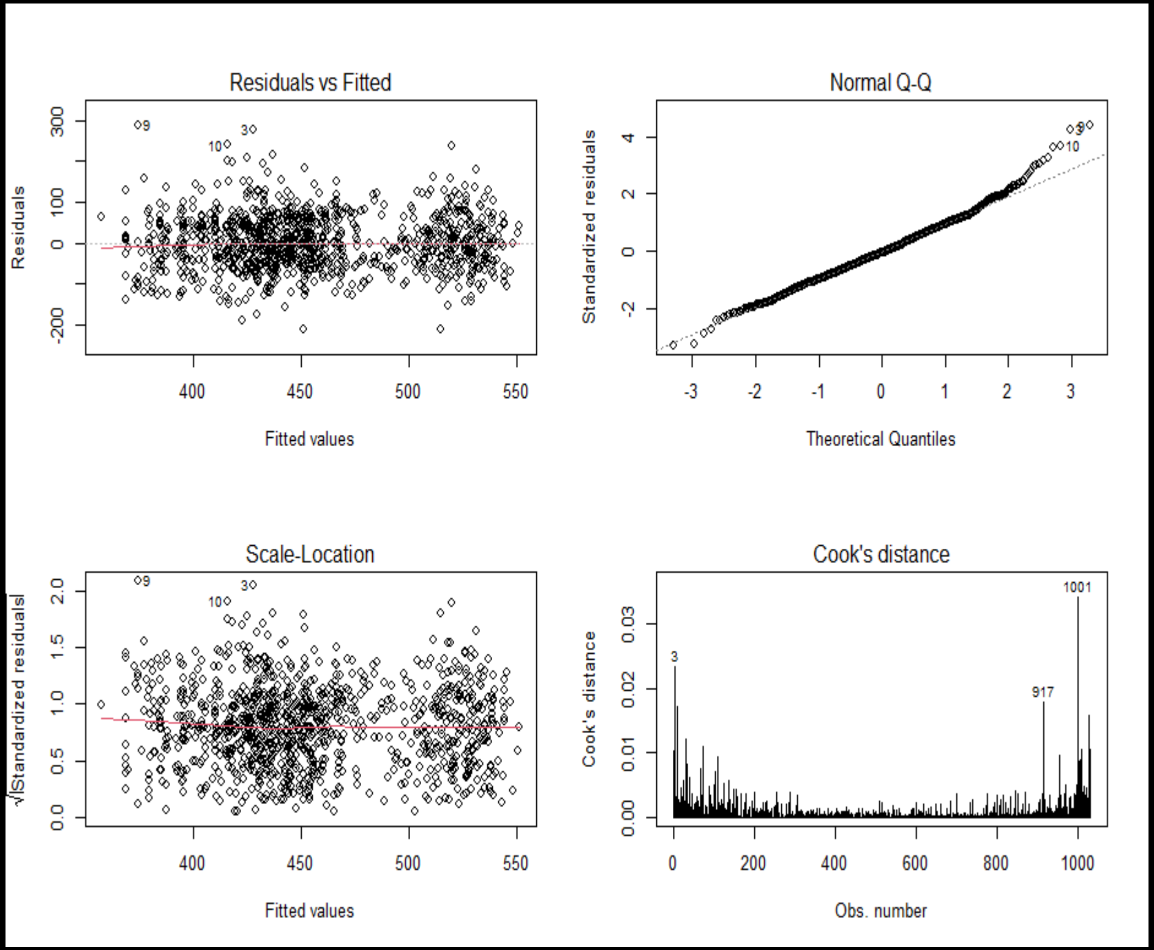
- Nature of the model
  - the relation between predictor and outcome has to be linear

- Nature of the errors (i.e., residuals)
  - normal and independent of each other

Model Assumptions: Simple regression lm()

- Required(LINE):
  - **L**inearity of relationships
  - **Independence of residuals**
  - **N**ormality of residuals
  - **E**qual variances for residuals

- Desirable:
  - uncorrelated predictors (no collinearity)
  - no outliers

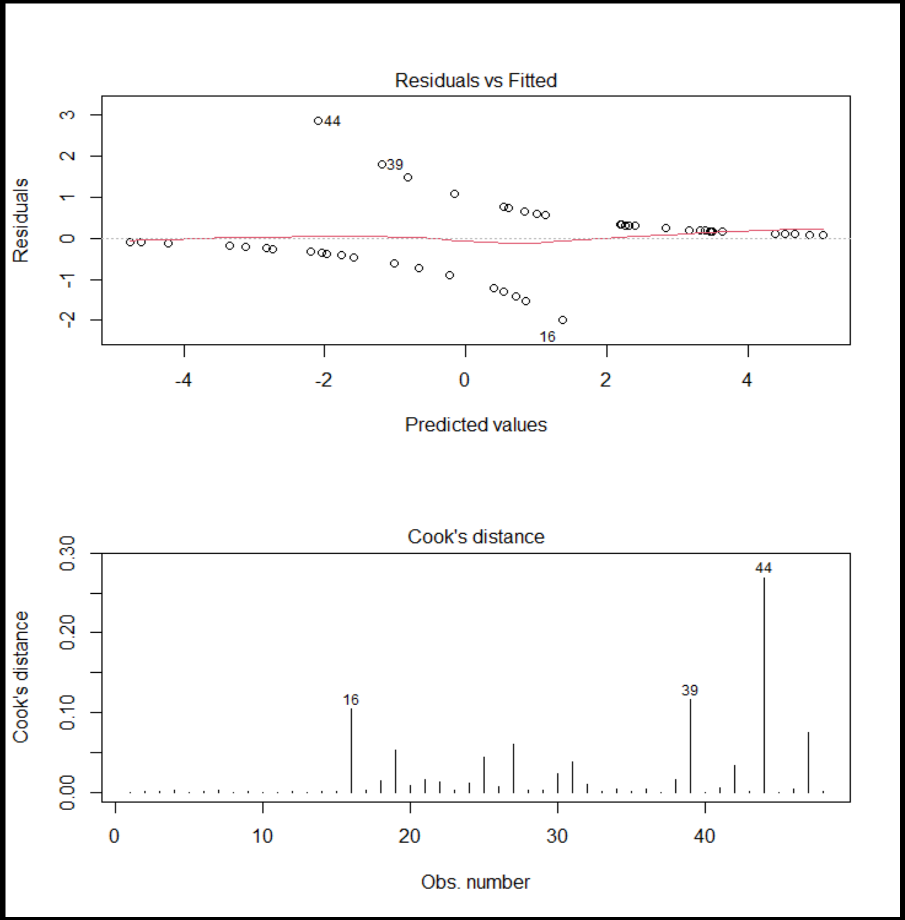


Model Assumptions: Simple regression glm()

For binomial DVs, (logistic regression)

- Required:
  - LINEAR relationships between IVs and log-odds
  - ~~Normality of residuals~~
  - ~~homogeneity of variance~~
  - Independence of residuals

- Desirable:
  - uncorrelated predictors (no collinearity)
  - no “bad” (overly influential) observations
  - large samples (due to maximum likelihood fitting)



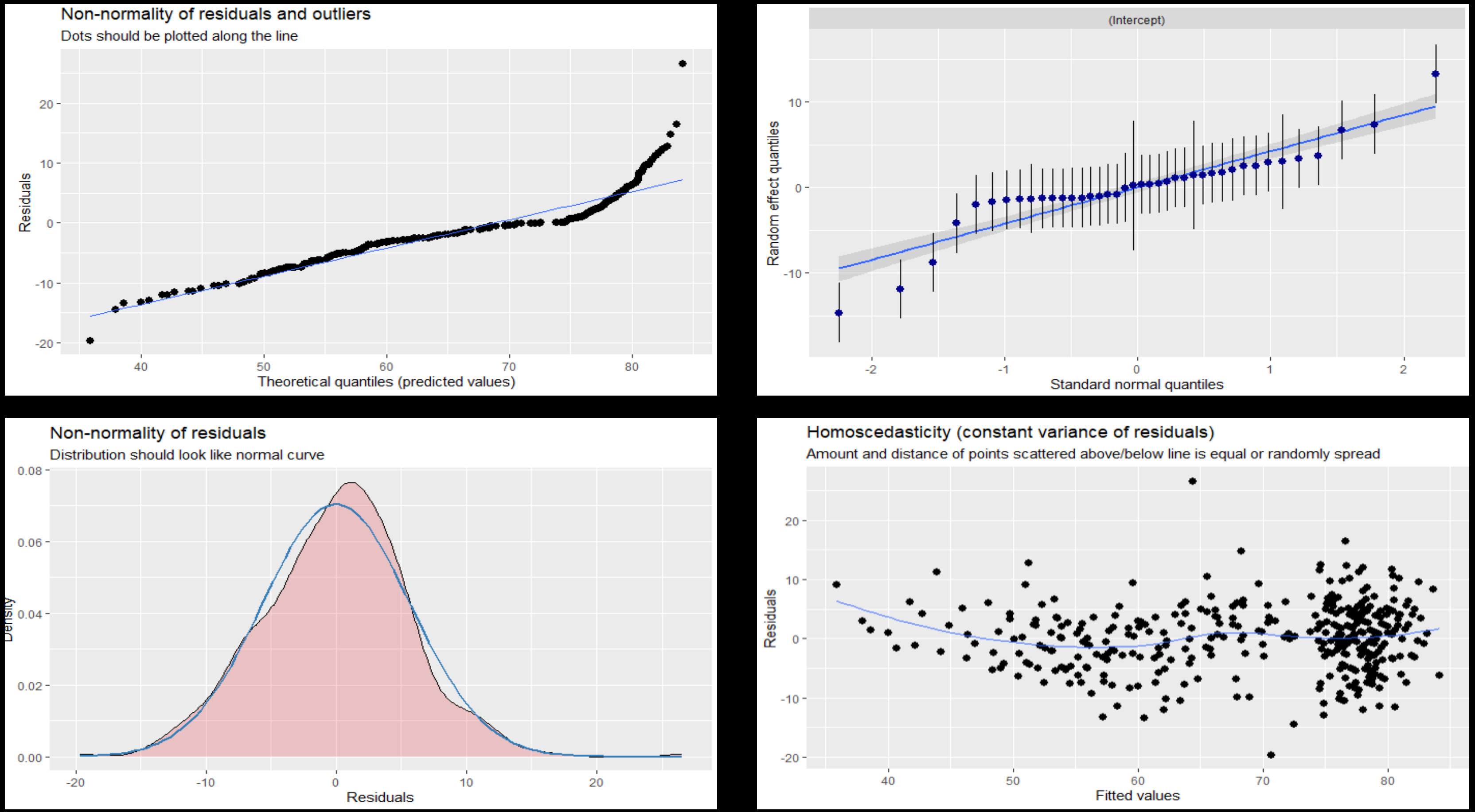
Model Assumptions: Mixed-effects Models

Similar to simple linear regressions model

- Error is random
- Residuals at multiple levels
  - Level1 residuals: mean = 0, variance constant (R code: residual())
  - Level2+ residuals: mean = 0, variance constant (R code: ranef())

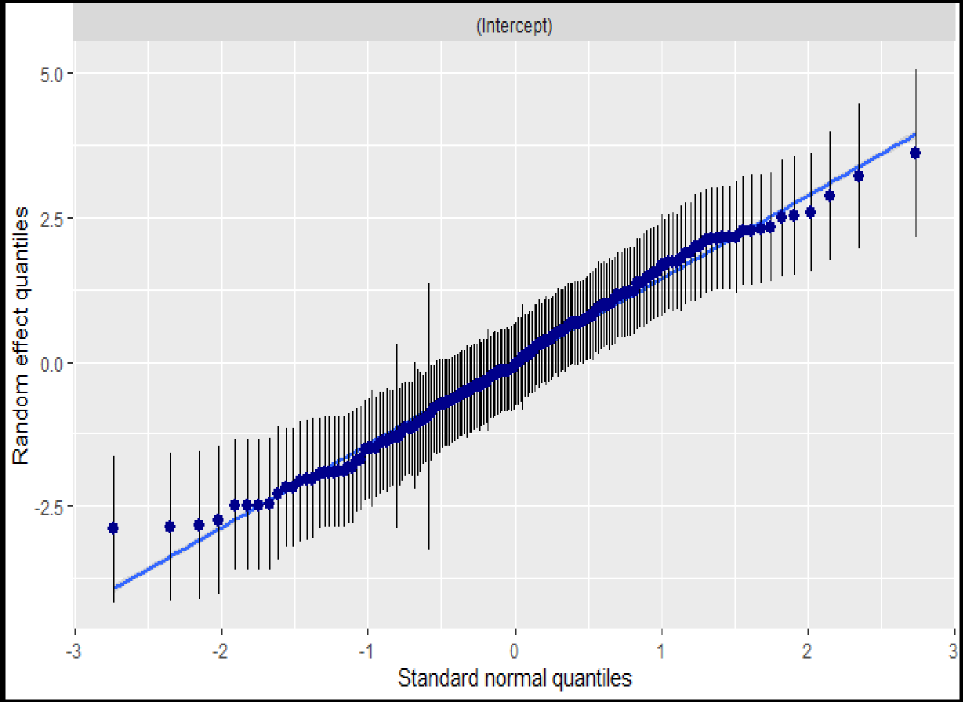
Model Assumptions: Mixed Models – lmer()

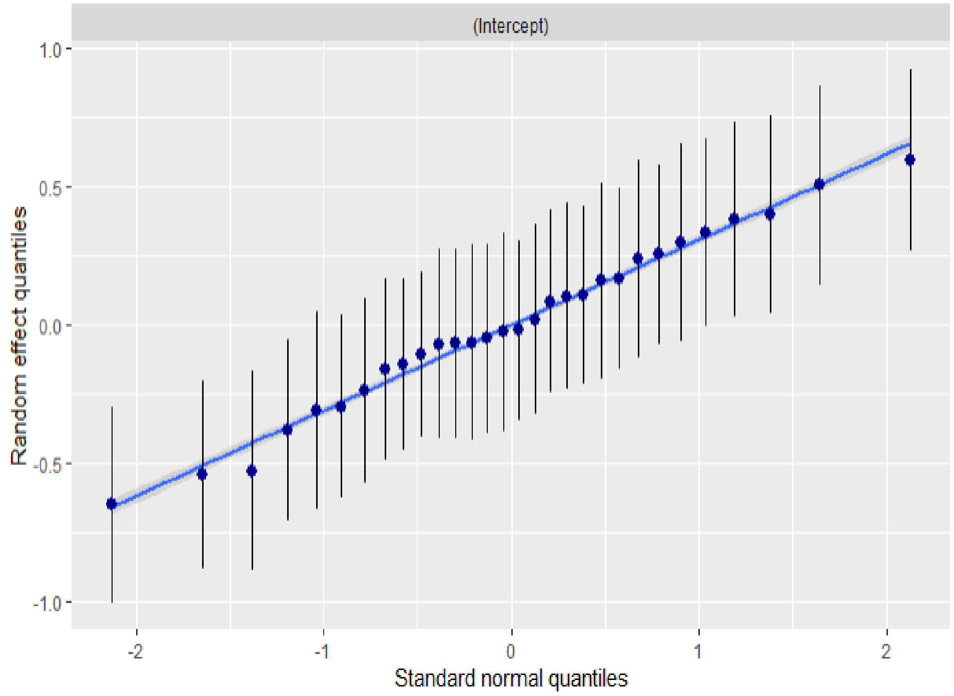
```
plot_model(mMixed_reduced , type = "diag")
```



Model Assumptions: Mixed Models – glmer()

```
plot_model(m3_dataISD, type = "diag")
```





Exercise & Q&A

Exercise for Friday

The ‘cheese data’: Simulated data based on real psycholinguistic findings on structural priming.

- Two Predictors :

  2-level factor “Prime”

  (a): Tom gave Spike some flowers.

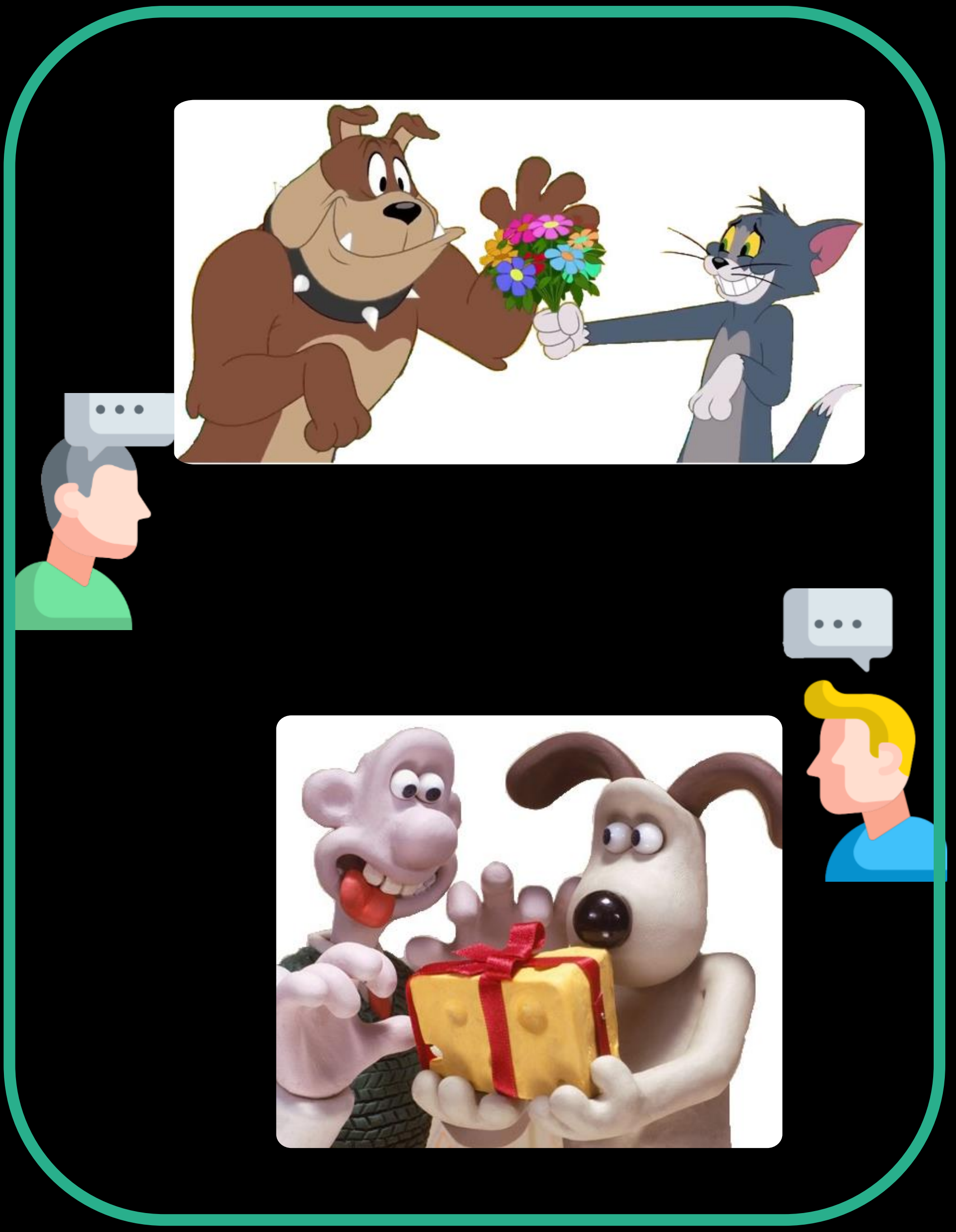
  (b): Tom gave some flowers to Spike.

  2-level factor “communication” (video- vs audio-call)

- Binary outcome: (a) or (b)?

  (a) Gromit gave … (Wallace some cheese)

  (b) Gromit gave … (some cheese to Wallace).



Exercise

Can you investigate the priming effect in the 'cheese data'?

- *R* code:

glmer ($y \sim x1 * x2$ +

($1 + x1 * x2$ | $Grouping_1$) +

($1 + x1$ | $Grouping_2$) ,

data = datafilename,

family = 'binomial'

)

Further Reading

- Paper

Brown, VA. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science.* 4(1). doi:10.1177/2515245920960351

Advances in Methods and Practices in Psychological Science

aps | ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Open access | Research article | First published online March 25, 2021

An Introduction to Linear Mixed-Effects Modeling in R

Violet A. Brown View all authors and affiliations

All Articles | https://doi.org/10.1177/2515245920960351

- E-Book

https://vasishth.github.io/Freq_CogSci/

Shravan Vasishth, Daniel Schad, Audrey Bürki, Reinhold Kliegl

Linear Mixed Models in Linguistics and Psychology: A Comprehensive Introduction

THANK YOU

THE UNIVERSITY *of* EDINBURGH
Centre for Data, Culture & Society