Multilevel (or mixed-effect) linear models

Çağrı Çöltekin

University of Groningen, Dept of Information Science

Apr 18, 2013

university of groningen

An example data set

We are interested in speech rate of phrases in three contexts (labeled *A*, *B* and *C*). We recorded participants (or subjects) where they utter a set of phrases (or items) in all three contexts. The data looks like this:

index	subject	time	item	context	speech rate
1	subject_1	10	item_1	A	3.755
2	subject_1	7	item_1	B	4.150
3	subject_1	15	item_1	C	3.060
4	subject_1	11	item_2	A	3.719
...					
135	subject_N	15	item_1	A	5.210
136	subject_N	2	item_1	B	5.670
137	subject_N	3	item_1	C	5.005
138	subject_N	14	item_2	A	5.037
...					

*Note: based on a real research, but data is simplified and randomly generated.

How to analyze it?

How to analyze it?

- The typical analysis would be 'within-subject' *repeated measures* ANOVA.
- This method accounts for the variation due to subject.
- We compare:

$$\left|\begin{array}{l} \text{sbj}_1, \text{ctx}_A, \text{item}_1 \\ \text{sbj}_1, \text{ctx}_A, \text{item}_2 \\ \text{sbj}_1, \text{ctx}_A, \text{item}_3 \end{array}\right| \text{ vs } \left|\begin{array}{l} \text{sbj}_1, \text{ctx}_B, \text{item}_1 \\ \text{sbj}_1, \text{ctx}_B, \text{item}_2 \\ \text{sbj}_1, \text{ctx}_B, \text{item}_3 \end{array}\right| \text{ vs } \left|\begin{array}{l} \text{sbj}_1, \text{ctx}_C, \text{item}_1 \\ \text{sbj}_1, \text{ctx}_C, \text{item}_2 \\ \text{sbj}_1, \text{ctx}_C, \text{item}_3 \end{array}\right|$$

- Effectively, we are comparing means over items.

Language-as-a-fixed-effect fallacy

- With RM ANOVA analysis, we account for the variation between subjects: our results are generalizable to the population at large.
- However, we do not account for variation due to items (words or phrases): our results do not generalize for the 'language', and we are loosing power.
- This problem is called 'language-as-a-fixed-effect fallacy' (Clark 1973; Raaijmakers et al. 1999).
- Workaround within RM ANOVA framework exist. However, multi-level or mixed-effect linear models provide a more general solution.
- Multilevel modeling is a general technique that can be used for a wide range of problems.

Overview

- A short introduction to *general* linear models (not to be confused with *generalized* linear models).
- Multilevel models: introduction.
- Multilevel models: examples.
- Summary.

Simple regression: a reminder

$$y_i = a + bx_i + e_i$$

- y is the *outcome* (or response) index i represent each unit observation/measurement ('index' in our example data).
- x is the *predictor* variable.
- a is the intercept.
- b is the slope of the regression line.
- $a + bx_i$ is the *deterministic* part of the model.
- e is the error, or the variation that is not accounted for by the model. Assumed to be (approximately) normally distributed with 0 mean (e_i are assumed to be i.i.d).

Simple regression: a reminder

$$y_i = a + bx_i + e_i$$

- y is the *outcome* (or response) index i represent each unit observation/measurement ('index' in our example data).
- x is the *predictor* variable.
- a is the intercept.
- b is the slope of the regression line.
- $a + bx_i$ is the *deterministic* part of the model.
- e is the error, or the variation that is not accounted for by the model. Assumed to be (approximately) normally distributed with 0 mean (e_i are assumed to be i.i.d).

Typically, regression is applied when both outcome and predictor variable(s) are continuous values, but it can be extended for categorical variables.

Predicting effect of time on speech rate

We want to see whether time during the experiment has an effect on speech rate.

index	subject	time	item	context	speech rate
1	subject_1	10	item_1	A	3.755
2	subject_1	7	item_1	B	4.150
3	subject_1	15	item_1	C	3.060
...					

Our model is,

$$\text{speech.rate}_i = a + b \times \text{time}_i + e_i$$

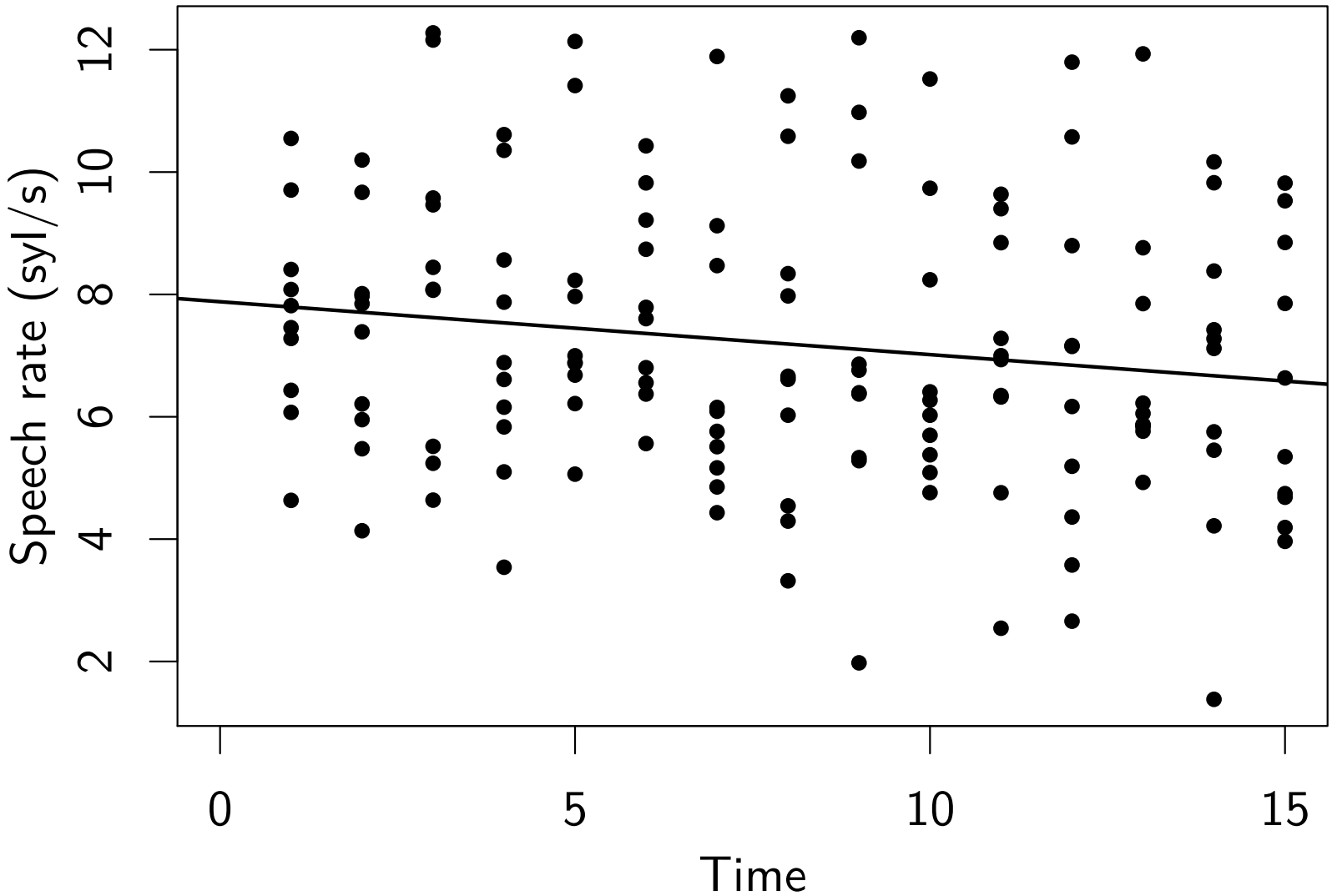
Here is the way to specify in R:

```
> lm(speech.rate ~ time)
```

R output: predicting effect of time on speech rate

```
> summary(lm(rate ~ time))

Call:
lm(formula = rate ~ time)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2888 -1.6484 -0.5204  1.6276  5.1750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.88059    0.39048  20.182   <2e-16 ***
time        -0.08645    0.04295  -2.013   0.0459 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.273 on 148 degrees of freedom
Multiple R-squared: 0.02665,    Adjusted R-squared: 0.02007
F-statistic: 4.052 on 1 and 148 DF,  p-value: 0.04594
```

And the graph...



Multiple regression

Regression analysis can be extended to multiple predictors.

$$y_i = a + b_1 x_{1i} + \ldots + b_k x_{ki} + e_i$$

Example would be predicting children's test scores from mothers' and fathers' IQ scores.
In R:

```
> lm(speech.rate ~ time + subject.age)
```

Note: In R, intercept, a, is implicitly included in a model specification. If you want a model with 0 intercept, you could specify the model as `speech.rate ~ time + subject.age - 1`, and a model without slope (model of the mean) can be specified using `speech.rate ~ 1`.

Regression with categorical predictors

- A categorical variable with N levels converted to $N - 1$ 'indicator' (or dummy) variables.
- Consider 'context' variable with three levels ('A', 'B', 'C'), we can code it as two variables, 'contextB', 'contextC' :

level	contextB	contextC
A	0	0
B	1	0
C	0	1

- Other coding options (contrasts) are possible. With some constraints, the inferences will not change.
- If the levels are ordered, transforming the categorical variable into a numeric variable may be more appropriate.

An example with only two levels

We want to check whether means of two of the contexts differ (labeled as ‘A’ and ‘C’).

An example with only two levels

We want to check whether means of two of the contexts differ (labeled as 'A' and 'C').
Normally we would do a t-test:

```
> t.test(rate2 ~ context2, var.equal=T)

        Two Sample t-test

data:  rate2 by context2
t = -1.4806, df = 98, p-value = 0.1419
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.5596945  0.2267907
sample estimates:
mean in group A mean in group C
       6.428031        7.094483
```

Doing t-test with regression

- We have two levels of the predictor (A and C).
- We code 'A' as 0 and 'C' as 1.

$$y_i = a + b \times \text{contextC}_i + e_i$$

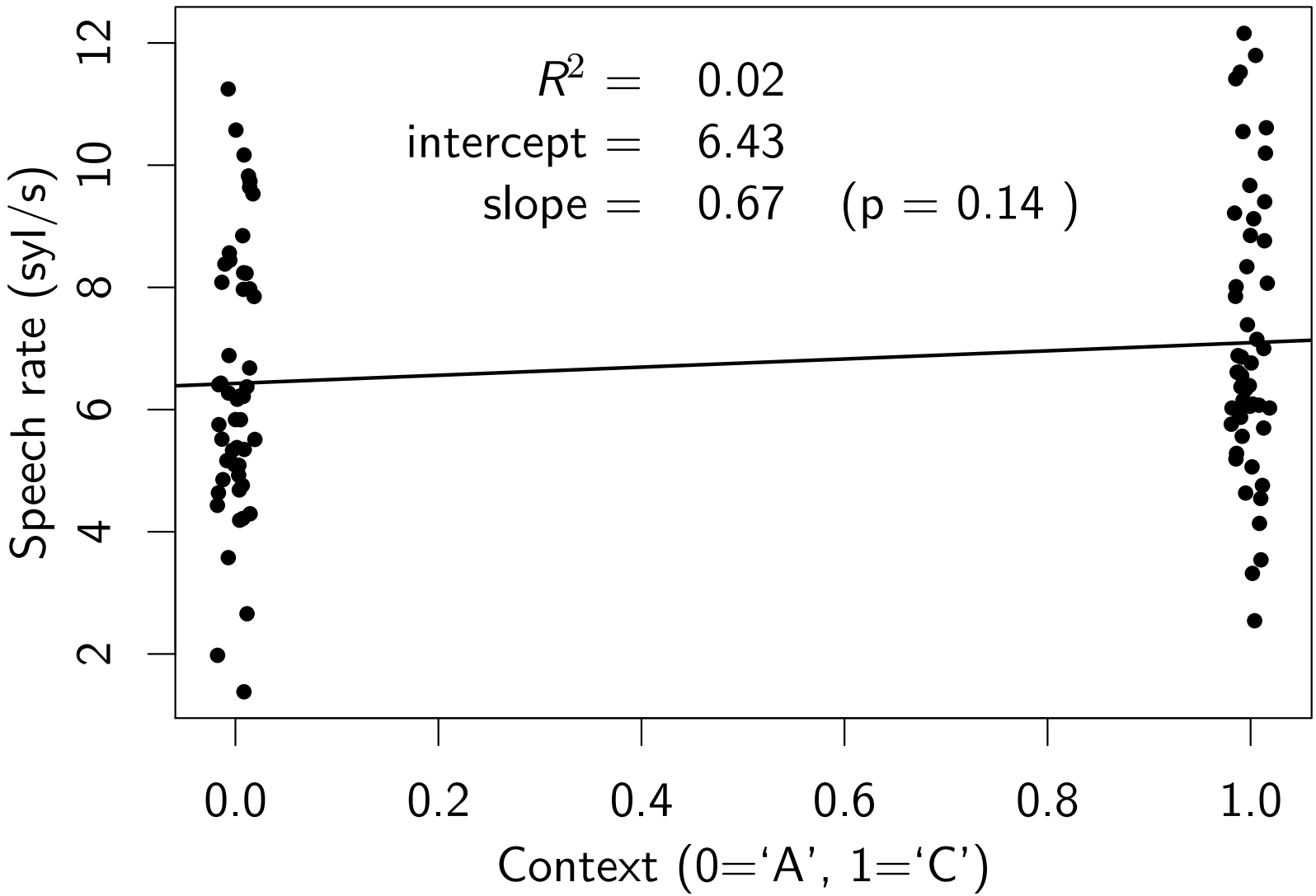
a (intercept) is the mean of level 'A'.

b (slope) is the mean difference between 'A' and 'B'.

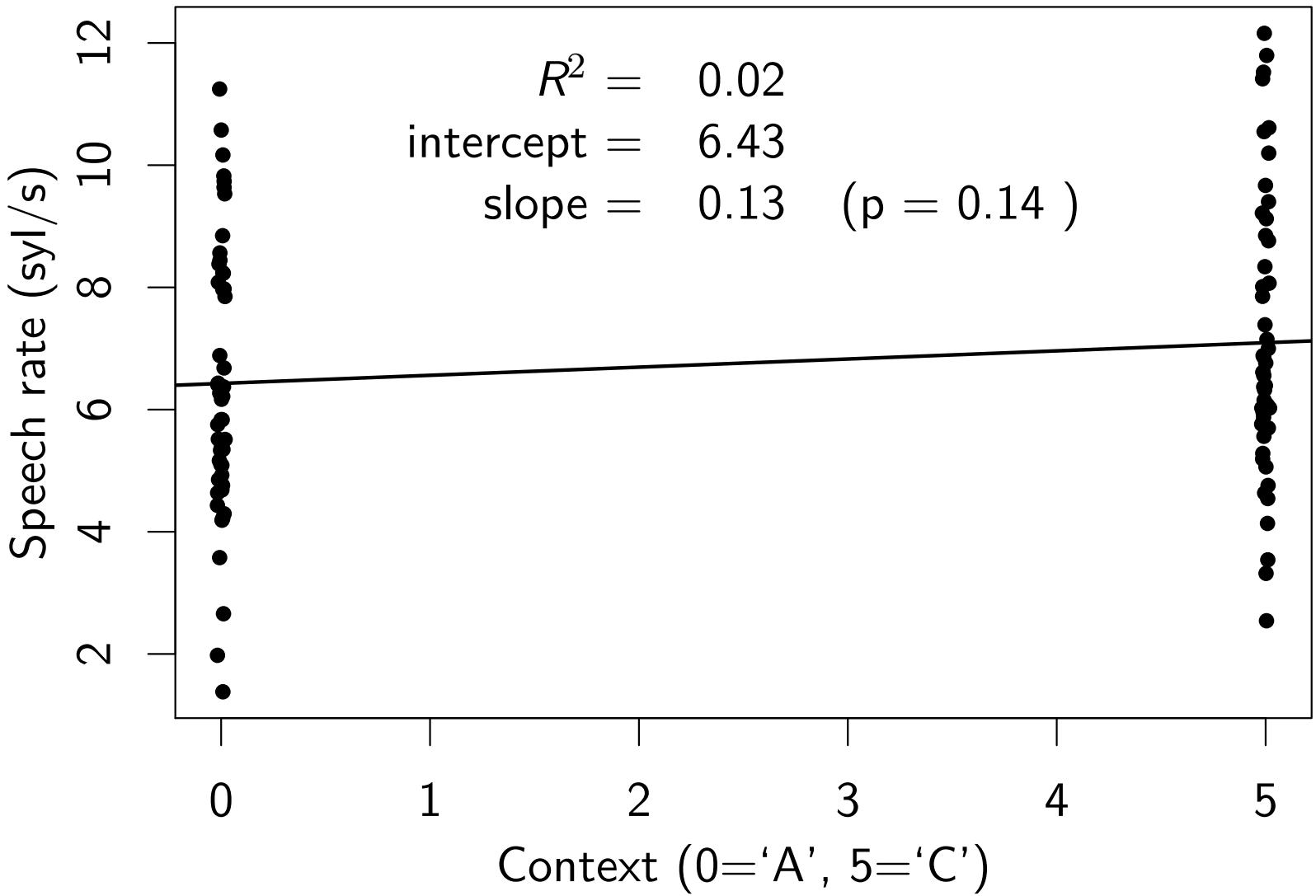
Doing t-test with regression: practice

```
> summary(lm(rate2 ~ context2))

Call:
lm(formula = rate2 ~ context2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0466 -1.3540 -0.4838  1.6895  5.0638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4280     0.3183  20.196   <2e-16 ***
context2C     0.6665     0.4501   1.481    0.142
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.251 on 98 degrees of freedom
Multiple R-squared: 0.02188,    Adjusted R-squared: 0.0119
F-statistic: 2.192 on 1 and 98 DF,  p-value: 0.1419
```

T-test as regression: the picture



T-test as regression: the picture



ANOVA as regression

Remembering that we code three levels as two indicator (dummy) variables:

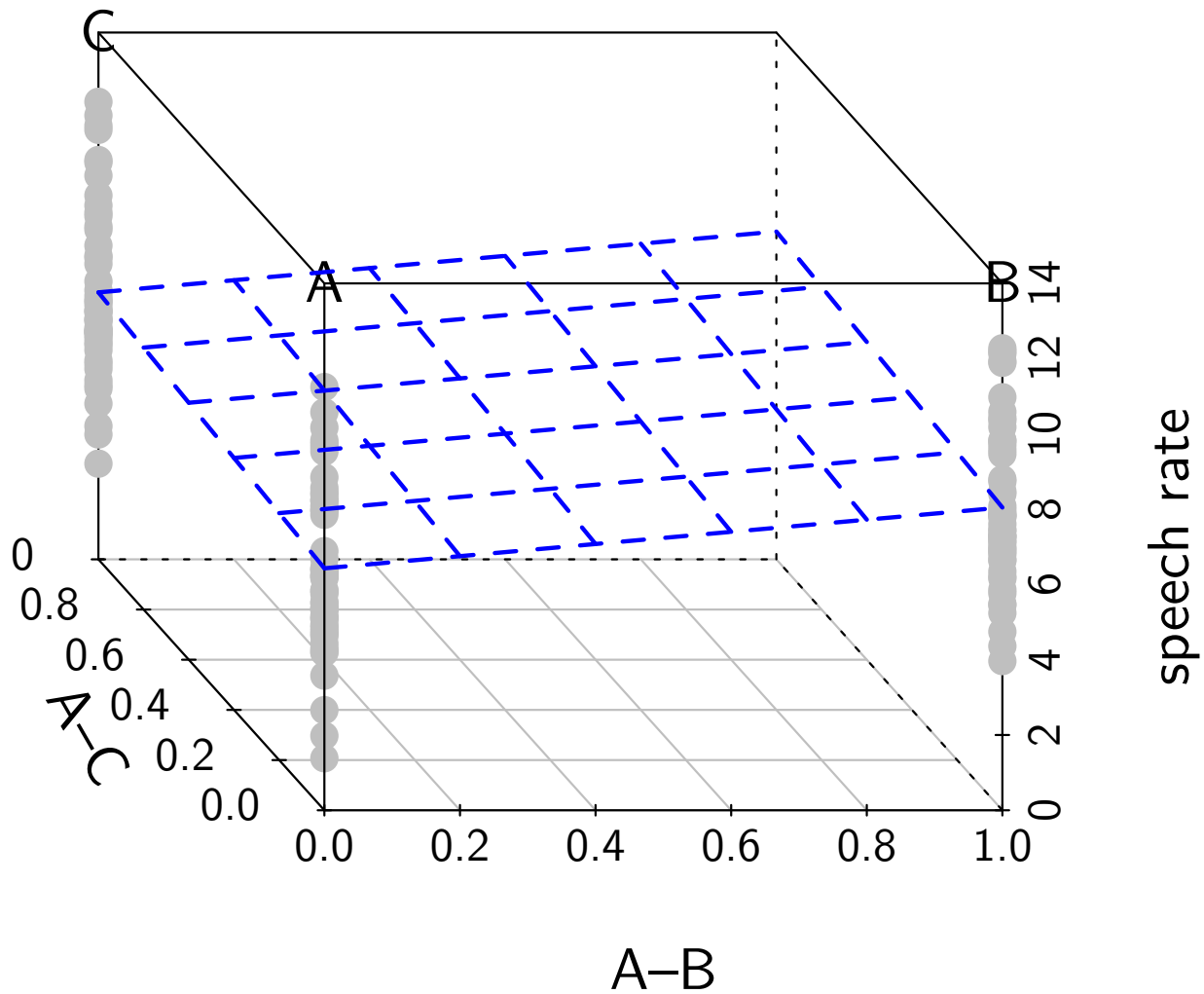
$$y_i = a + b_1 \times \text{contextB}_i + b_2 \times \text{contextC}_i + e_i$$

- a (intercept) is the mean of context 'A'.
- b_1 (slope of contextB) is the mean difference between 'A' and 'B'.
- b_2 (slope of contextC) is the mean difference between 'A' and 'C'.

ANOVA as regression: practice

```
> summary(lm(rate ~ context))

Call:
lm(formula = rate ~ context)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0466 -1.3719 -0.4616  1.6664  5.0638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4280     0.3128  20.548  < 2e-16 ***
contextB      1.6165     0.4424   3.654 0.000359 ***
contextC      0.6665     0.4424   1.506 0.134105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.212 on 147 degrees of freedom
Multiple R-squared: 0.08404,    Adjusted R-squared: 0.07158
F-statistic: 6.744 on 2 and 147 DF,  p-value: 0.001577
```

ANOVA as regression: the picture



ANOVA as regression: ANOVA table

```
> anova(lm(rate ~ context))
Analysis of Variance Table

Response: rate
           Df Sum Sq Mean Sq F value   Pr(>F)
context     2  66.00  32.998  6.7437 0.001577 **
Residuals 147 719.29   4.893
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that the fitted model is the same, we only summarize the results differently.

ANOVA as regression: ANOVA table

```
> anova(lm(rate ~ context))
Analysis of Variance Table

Response: rate
           Df Sum Sq Mean Sq F value   Pr(>F)
context     2  66.00  32.998  6.7437 0.001577 **
Residuals 147 719.29   4.893
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that the fitted model is the same, we only summarize the results differently. Problem with this analysis: the cases (measurements/observations) are not independent. Each case is related to others through ‘subject’ and ‘item’.

What is a multilevel model?

$$y_i = a + b_1 x_{1i} + \ldots + b_k x_{ki} + e_i$$

- In a classical regression model, the parameters (a and b_i) are 'fixed'.
- In multilevel models, we model one or more parameters as 'random', being drawn from a distribution.
- Multilevel modeling is about estimating the main model, and the random parameters simultaneously.

A simple example: accounting for between-subject variation

$$y_i = a_{j[i]} + bx_{1i} + e_i$$

$$a_j = \mu_j + \epsilon_j$$

A simple example: accounting for between-subject variation

$$y_i = a_{j[i]} + bx_{1i} + e_i$$

$$a_j = \mu_j + \epsilon_j$$

For our example,

- y (response variable) is the speech rate, indexed by each measurement unit (speaker $\times$ context $\times$ item).
- a_j is intercept for each subject j, notation $j[i]$ indicates subject associated with i^{th} unit.
- e is the error for each unit.
- μ_j is the mean speech rate for subject j.
- ϵ_j is the error (variation) associated with the speech rate for subject j.

Note: this is equivalent to repeated measures ANOVA.

Analysis using RM ANOVA

Repeated-measures ANOVA is a restricted version of multilevel linear model, which works fine for this example.

```
> summary(aov(rate ~ context + Error(subject)))

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9  75.24   8.359

Error: Within
           Df Sum Sq Mean Sq F value  Pr(>F)
context     2   66.0   33.00    7.07 0.00119 **
Residuals 138  644.1    4.67
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Compare with $\mathrm{F}(2, 147) = 6.7437 (\mathrm{p} = 0.001577)$ from earlier (non-RM) ANOVA.

Analysis with multi-level modeling

```
> library(lme4); summary(lmer(rate ~ context + (1|subject)))
Linear mixed model fit by REML
Formula: rate ~ context + (1 | subject)
   AIC   BIC logLik deviance REMLdev
 670.6 685.7 -330.3    659.6   660.6
Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 0.24616  0.49615
 Residual             4.66703  2.16033
Number of obs: 150, groups: subject, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.4280     0.3434  18.717
contextB      1.6165     0.4321   3.741
contextC      0.6665     0.4321   1.542

Correlation of Fixed Effects:
         (Intr) cntxtB
contextB -0.629
contextC -0.629  0.500
```

Why use multilevel models?

- RM ANOVA has very strict requirements
  - all assumptions of ANOVA, except independence of samples.
  - balanced design
  - sphericity (homogeneity of variance/covariance)
- We may want to include continuous predictors.
- We may have more than one sources error, e.g., error due to subjects and items, where RM ANOVA is no more applicable.
- We may want to include predictors at higher, e.g., subject, level.
- We may want to arrive at conclusions at multiple levels (more on this later).
- Using 'generalized linear models' (logistic regression, count regression) may be more appropriate.
- The data may be structured in complex ways.

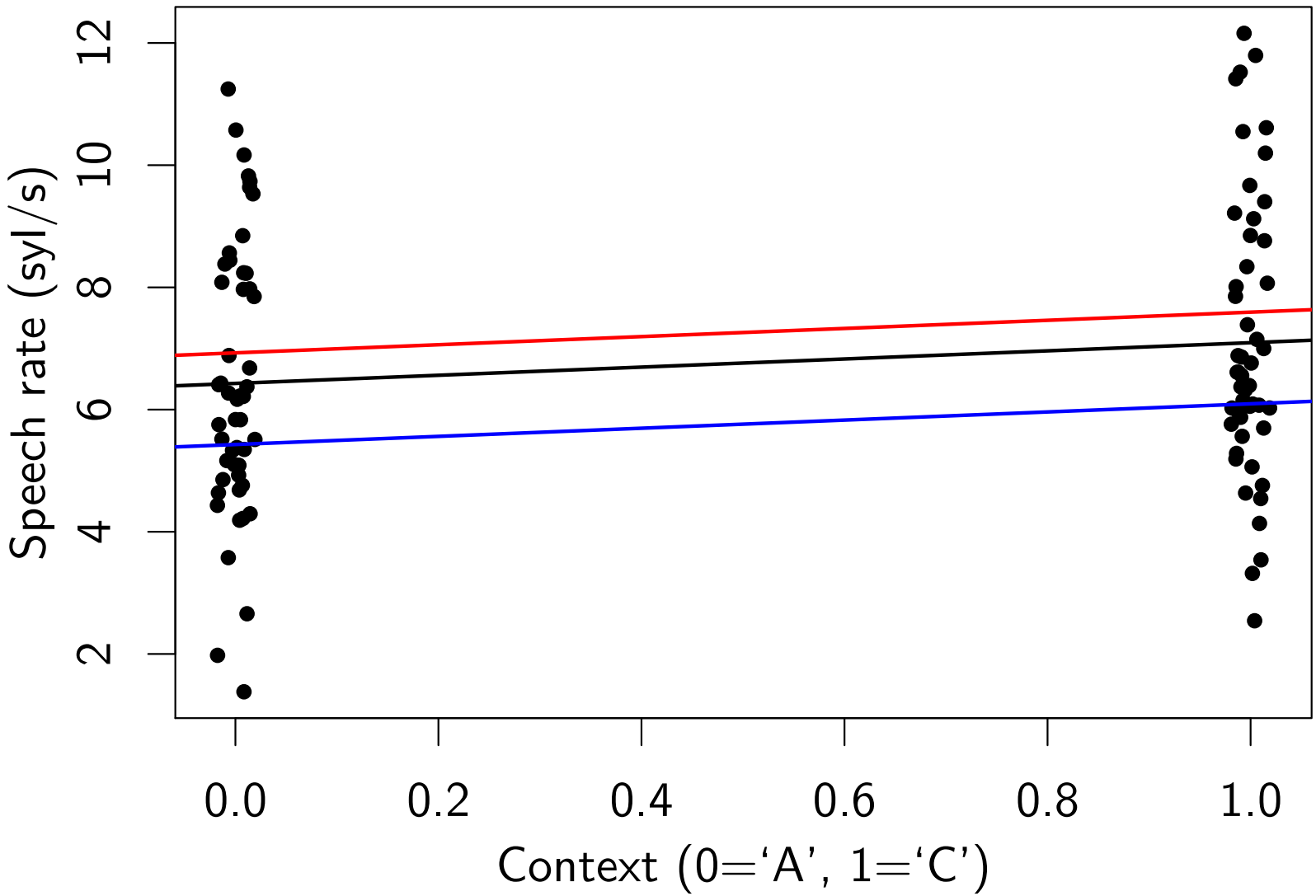
Varying intercepts

$$\begin{aligned} \text{Level 1:} \quad & \text{speech.rate}_i = a_{j[i]} + b \times \text{context}_i + e_i \\ \text{Level 2:} \quad & a_j = \mu_{\text{subject}} + \epsilon_j \end{aligned}$$

The examples we discussed so far are varying-intercept models. In these models,

- Each group has a different intercept: each subject has a characteristic/baseline speech rate.
- All groups share common slope(s): subject and the context does not interact. That is, a subject does not systematically speak slower on one context while speaking faster on the other context(s).
- Level 2 can be more complex, as much as linear modeling allows.

Varying intercepts: visualization



Varying intercepts: multiple sources of variation

Level 1: $\text{speech.rate}_i = a_{j[i]} + b \times \text{context}_i + e_i$

Level 2: $a_j = \mu_{\text{subject}} + \epsilon_{\text{subject}}$

Varying intercepts: multiple sources of variation

$$\text{Level 1:} \quad \text{speech.rate}_i = \alpha + a_{j[i]} + a_{k[i]} + b \times \text{context}_i + e_i$$

$$\text{Level 2:} \quad a_j = \delta_{\text{subject}} + \epsilon_{\text{subject}}$$

$$a_k = \delta_{\text{item}} + \epsilon_{\text{item}}$$

Changes are in Level 2 regression.

- α is the intercept for the Level 1 intercept term (base speech rate without subject and item variation).
- δ_{subject} is the (systematic) change to the base speech rate due to subject.
- δ_{item} is the (systematic) change to the base speech rate due to item. N

Note: this model can be formulated different ways.

Multiple sources of variation: analysis in R

```
> summary(lmer(rate ~ context + (1|subject) + (1|item)))
Linear mixed model fit by REML
Formula: rate ~ context + (1 | subject) + (1 | item)
   AIC   BIC logLik deviance REMLdev
 488.4 506.4 -238.2    475.1   476.4
Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 0.48599  0.69713
 item     (Intercept) 4.13699  2.03396
 Residual             1.06964  1.03424
Number of obs: 150, groups: subject, 10; item, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.4280     0.9472   6.786
contextB      1.6165     0.2068   7.815
contextC      0.6665     0.2068   3.222

Correlation of Fixed Effects:
         (Intr) cntxtB
contextB -0.109
contextC -0.109  0.500
```

Varying slopes

As well as intercepts, slopes can vary:

Level 1: $\text{speech.rate}_i = \alpha + a_{j[i]} + a_{k[i]} + b_{k[i]} \times \text{context}_i + e_i$

Level 2: $a_j = \delta_{subject} + \epsilon^a_{subject}$

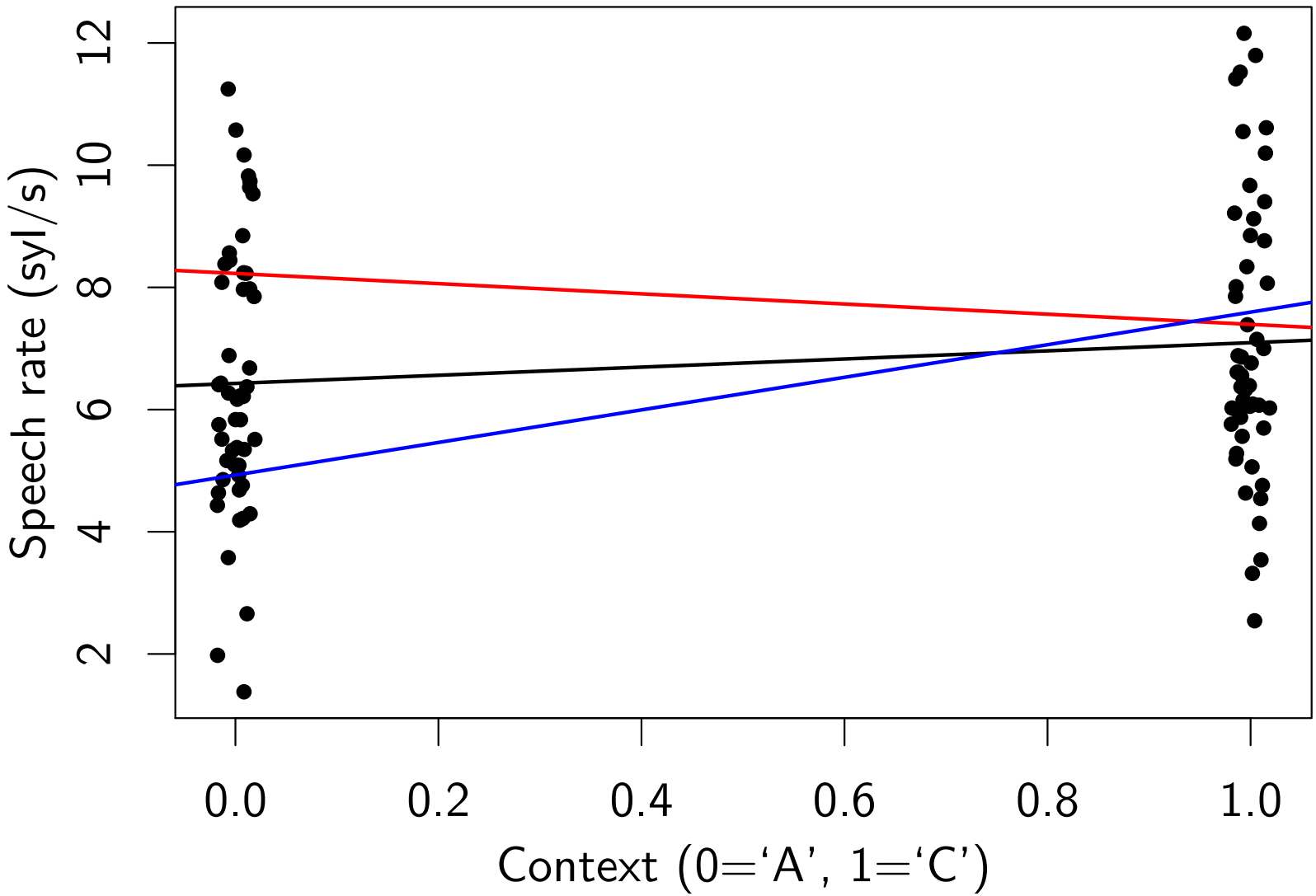
$a_k = \delta_{item} + \epsilon^a_{item}$

$b_k = \mu_{item} + \epsilon^b_{item}$

b_j is varying slope due to item (that is, we have item–context interaction)

μ_a mean value of slope (varied by item)

Varying slopes: visualization



Varying slope: example

```
> summary(lmer(rate ~ context + (1|subject) + (1+context|item)))
Linear mixed model fit by REML
Formula: rate ~ context + (1 | subject) + (1 + context | item)
 AIC   BIC logLik deviance REMLdev
 497 530.1 -237.5    474.2     475
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subject  (Intercept) 0.48791  0.69850
 item     (Intercept) 4.05136  2.01280
          contextB    0.12299  0.35070  -0.210
          contextC    0.01643  0.12818   0.995 -0.110
 Residual             1.04089  1.02024
Number of obs: 150, groups: subject, 10; item, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.4280     0.9380   6.853
contextB      1.6165     0.2574   6.281
contextC      0.6665     0.2119   3.144

Correlation of Fixed Effects:
         (Intr) cntxtB
contextB -0.209
contextC  0.154  0.363
```

Continuous predictors

Level 1: $$\text{speech.rate}_i = \alpha + a_{j[i]} + a_{k[i]} + b_t \times \textbf{time}_i + b_{k[i]} \times \text{context}_i + e_i$$

Level 2: $$a_j = \delta_{subject} + \epsilon^a_{subject}$$
$$a_k = \delta_{item} + \epsilon^a_{item}$$
$$b_k = \mu_{item} + \epsilon^b_{item}$$

Varying slopes without varying intercepts are also possible, but rarely needed in practice.

Continuous predictors: example

```
> summary(lmer(rate ~ context + time + (1|subject) + (1+context|item)))
Linear mixed model fit by REML
Formula: rate ~ context + time + (1 | subject) + (1 + context | item)
   AIC   BIC logLik deviance REMLdev
 491.1 527.2 -233.6    459.9   467.1
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 subject  (Intercept) 0.5140446 0.716969
 item     (Intercept) 4.2347270 2.057845
          contextB    0.0662157 0.257324 -0.461
          contextC    0.0021566 0.046439  0.974 -0.247
 Residual             0.9501461 0.974754
Number of obs: 150, groups: subject, 10; item, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)  7.10359    0.97366   7.296
contextB     1.50092    0.22835   6.573
contextC     0.53312    0.19908   2.678
time        -0.07407    0.01920  -3.858

Correlation of Fixed Effects:
         (Intr) cntxtB cntxtC
contextB -0.329
contextC -0.033  0.428
time     -0.180  0.131  0.174
```

Predictors at higher levels

Level 1: $$\text{speech.rate}_i = \alpha + a_{j[i]} + a_{k[i]} + b_t \times \text{time}_i + b_{k[i]} \times \text{context}_i + e_i$$

Level 2: $$a_j = \delta_{\text{subject}} + \beta \times \text{age} + \epsilon^a_{\text{subject}}$$

$$a_k = \delta_{\text{item}} + \epsilon^a_{\text{item}}$$

$$b_k = \mu_{\text{item}} + \epsilon^b_{\text{item}}$$

In R:

```
rate ~ context + time + (1|subject+age) + (1+context|item))
```

More multilevel scenarios (1)

We want to test the effect of a new language learning method, where

Level 1:	Students in classrooms	test scores, gender ...
Level 2:	Classrooms in schools	teacher's attitude ...
Level 3:	Schools in cities/districts	average income, city size ...
Level 4:	Cities in countries	native languages, SLA policies ...

Model fit

- In (complex) multilevel models, the standard inferences from least-squares regression are not available.
- The typical practice (like any multiple regression model) is to find the best fit (e.g., guided by AIC).
- In most cases, simulation based inference and model checks are the only option.
- If you really need your p-values, there are tools to calculate 'simulation based' p-values (e.g., `pvals.fnc()` from `languageR` package by Baayen).

Summary

- Multilevel (or mixed-effect) models are a generalization of linear models where some of the parameters are considered random variables.
- We model parameters as 'random' where we have a systematic variation that we can explain by input variables.
- Fixed parameters are those where inferential uncertainty is assumed to be random.
- The multilevel modeling provides a solution to language-as-a-fixed-effect fallacy, but it is applicable to a wider range of problems.

Where to go from here

- We have only covered examples of least-squares regression, multilevel models can also be fitted for any 'generalized linear model'.
- Once you have taken the path of specifying 'random' parameters, you can embrace it, and use Bayesian inference.

Recommended reading:

- Gelman and Hill (2007): if you are serious about multilevel modeling, worth your reading time: well written, precise, accessible.
- Baayen (2008): especially takes a linguistic perspective.

References

Baayen, R Harald (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Clark, Herbert H (1973). "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research". In: *Journal of Verbal Learning and Verbal Behavior* 12, pp. 335–359.

Gelman, Andrew and Jennifer Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Raaijmakers, Jeroen G. W., Joseph M. C. Schrijnemakers, and Frans Gremmen (1999). "How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and Alternative Solutions". In: *Journal of Memory and Language* 41, pp. 416–426.

Clark's solution to the language-as-a-mixed-effect fallacy

The solution offered Clark 1973 for the language-as-a-mixed-effect fallacy:

$$\min F'(i,j) = \frac{F_1 F_2}{F_1 + F_2}$$

Where $F_1(n, n_1)$ is the F value due to subjects, $F_2(n, n_2$ is the F-value due to items (e.g., words or sentences). The associated degrees of freedom i will be equal to number of observations (note that n_1 is the DF due to number of subjects, and n_2 is the DF due to number of items) n, and j can be calculated using,

$$j = \frac{(F_1 + F_2)^2}{\frac{F_1^2}{n_2} + \frac{F_2^2}{n_1}}$$

See Clark (1973) for the details and the derivations.

Fixed and random effects

The predictors whose coefficients are modeled as random variables are called 'random effects', and the ones with constant coefficients are called 'fixed effects'.
Question is which variables shoud you model as random, and which ones should be modeled as fixed. Commonly expected* guidlines are, model a variable as

fixed if all values (or all values that the researcher is interested) are represented in the sample.

random if sample contains only part of the vlaues that the research aims to generalize.

* See Gelman and Hill (2007, p.245) for a discussion.