Gemischte Lineare Modelle

Linear Mixed Effect Models

Fritz Günther
fritz.guenther@uni-tuebingen.de

Universität Tübingen

June 9, 2016

Übersicht

- Lineare Modelle allgemein
- Gemischte Lineare Modelle
- Hypothesentests/ Modellvergleiche
- Berichten der Ergebnisse

Literatur

Tutorial:

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf]

Literatur (Einführung):

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Lineare Modelle

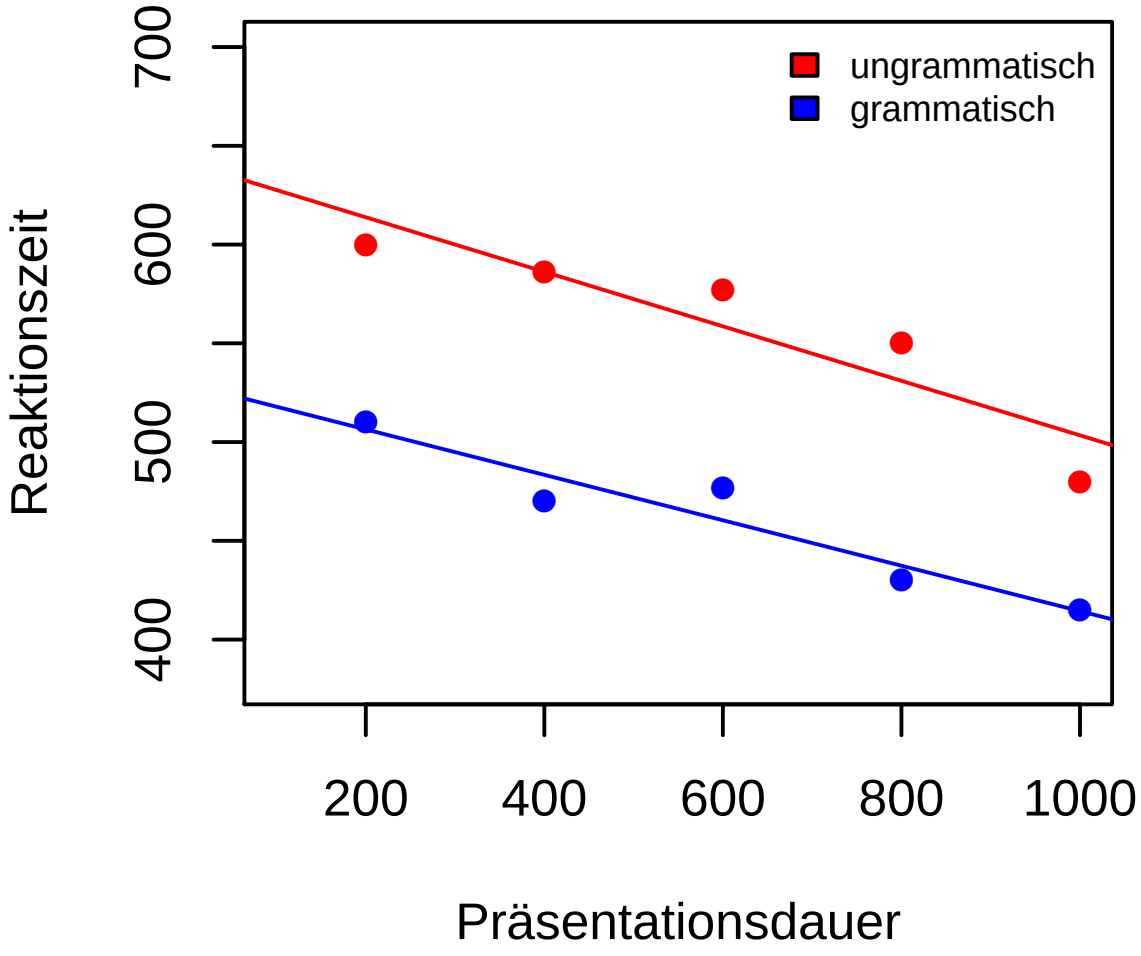
$$Y = a + b + \epsilon$$

Y: Abhängige Variable ("Kriterium")
a: Unabhängige Variable 1 ("Prädiktor 1")
b: Unabhängige Variable 2 ("Prädiktor 2")
ϵ: Zufälliger Fehler

Beispiel:

$$\text{Reaktionszeit} = \text{Präsentationsdauer} + \text{Grammatikalität} + \epsilon$$

Beispieldaten

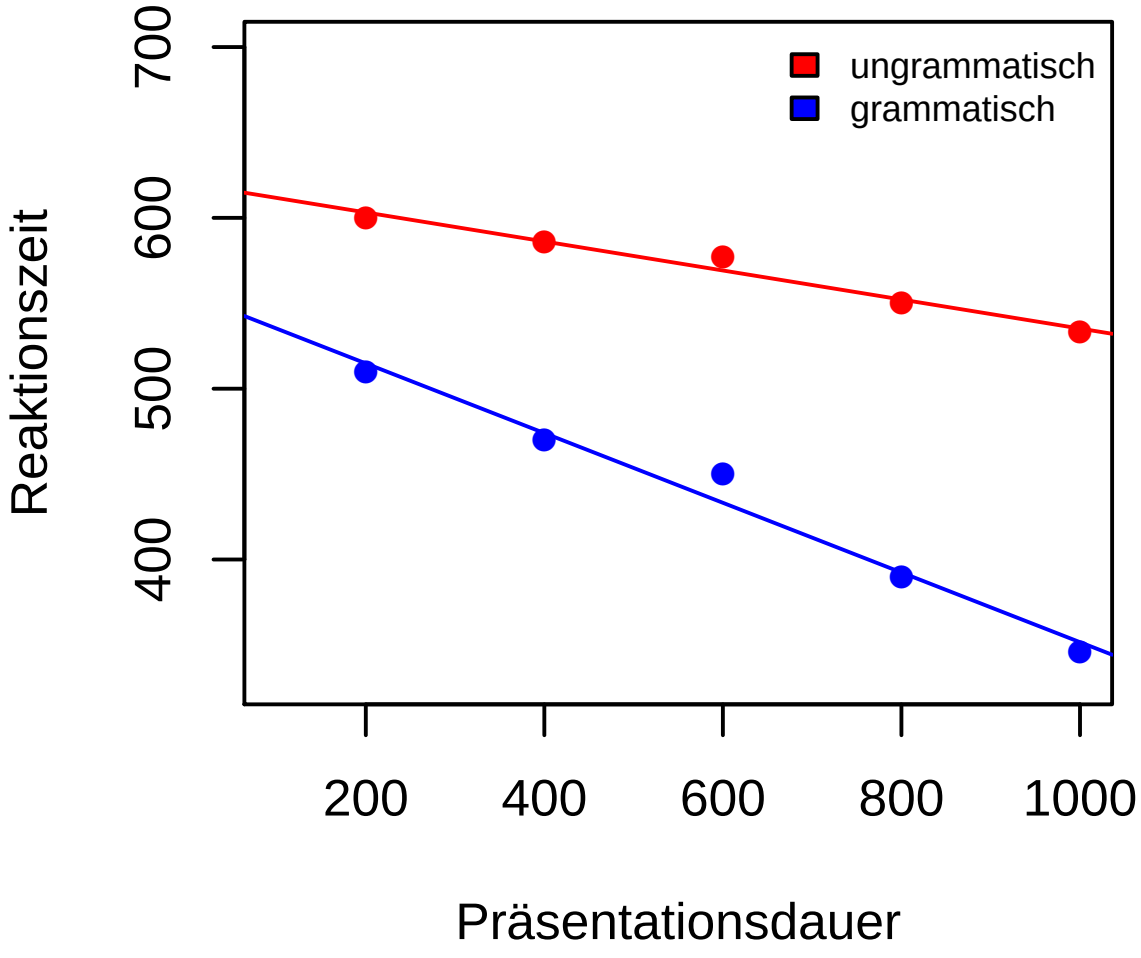


Lineare Modelle: Interaktionen

$$Y = a + b + ab + \epsilon$$

Y: Abhängige Variable ("Kriterium")
a: Unabhängige Variable 1 ("Prädiktor 1")
b: Unabhängige Variable 2 ("Prädiktor 2")
ab: Interaktionseffekt beider Prädiktoren
ϵ: Zufälliger Fehler

Beispieldaten mit Interaktion



Feste vs zufällige Effekte

Feste Effekte: Erhobene Faktorstufen sind *Vollerhebung* der interessierenden Faktorstufen

Beispiele: Präsentationsdauer, Präsentation eines grammatischen vs ungrammatischen Satzes

$\Longrightarrow$ Keine Generalisierung nötig

Zufällige Effekte: Erhobene Faktorstufen sind *Teilstichprobe* der interessierenden Faktorstufen

Beispiele: Versuchspersonen (VPs), Items

$\Longrightarrow$ Generalisierung erwünscht

Grundproblem

Verschiedene VPs sind i.A. unterschiedlich schnell
$\Longrightarrow$ Zufällige Effekte für VPs im Modell (F_1 ANOVA)

$$Y = a + b + ab + (1|subject) + \epsilon$$

Verschiedene Items werden i.A. unterschiedlich schnell bearbeitet
$\Longrightarrow$ Zufällige Effekte für Items im Modell (F_2 ANOVA)

$$Y = a + b + ab + (1|item) + \epsilon$$

Grundproblem

Gemischte Lineare Modelle

$$Y = a + b + ab + (1|subject) + (1|item) + \epsilon$$

Wieso nicht gleich?

- Schätzung der Modelle war lange sehr aufwendig
- Implementiert "an sich" keine Signifikanztests

Berechnen eines LMEM

- Das Paket `lme4` installieren:
  `install.packages("lme4")`
- Das Paket `lme4` laden:
  `library(lme4)`
- Das Modell schätzen:

```
model <- lmer(RT ~ Gramm + PresT + Gramm:PresT +
                   (1 |VP) + (1 |Item), dat)
```

oder

```
model <- lmer(RT ~ Gramm*PresT +
                   (1 |VP) + (1 |Item), dat)
```

Hands On: Ergebnisse anschauen

Wichtig: Alle Faktoren als Faktoren definieren

```
dat$PresT <- as.factor(dat$PresT)
```

Dies ändert den Output des Modells

```
model <- lmer(RT ~ Gramm*PresT +
                   (1 |VP) + (1 |Item), dat)
```

Hands On: Ergebnisse anschauen

```
> summary(model)
Linear mixed model fit by REML ['lmerMod']
Formula: RT ~ Gramm + PresT + Gramm:PresT + (1 | VP) + (1 | Item)
   Data: dat

REML criterion at convergence: 56473.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9564 -0.6864 -0.0037  0.6789  3.7196

Random effects:
 Groups   Name        Variance  Std.Dev.
 VP       (Intercept) 1.043e-20 1.021e-10
 Item     (Intercept) 6.575e-13 8.108e-07
 Residual             3.993e+02 1.998e+01
Number of obs: 6400, groups:  VP, 32; Item, 20

Fixed effects:
                     Estimate Std. Error t value
(Intercept)           599.4039     0.7899   758.8
Grammungramm          -70.7890     1.1171   -63.4
PresT400              -10.0899     1.1171    -9.0
PresT600              -19.3762     1.1171   -17.3
PresT800              -29.5120     1.1171   -26.4
PresT1000             -38.2906     1.1171   -34.3
Grammungramm:PresT400  -8.1590     1.5798    -5.2
Grammungramm:PresT600 -20.0520     1.5798   -12.7
Grammungramm:PresT800 -28.9892     1.5798   -18.3
Grammungramm:PresT1000 -40.8810     1.5798   -25.9
```

Hypothesentests und Signifikanzen

t-Werte bieten eine (grobe!) Faustregel:
Einfluss ist vorhanden bei $t > 2$

Hypothesentest und Signifikanzen

Ist das R-Paket `lmerTest` installiert und geladen, so gibt
`summary(model)`
auch p-Werte aus

Freiheitsgrade sind hier per Satterthwaite-Approximation geschätzt

Hypothesentests und Signifikanzen

Wie kommt man ohne Approximation an Hypothesentests auf Signifikanz?
Wie testet man den Effekt einer ganzen Variable, nicht nur den einzelner Stufen?

Antwort: **Likelihood-Ratio-Tests**

Hierfür benötigen wir noch drei Konzepte:

- Geschachtelte Modelle (Nested Models)
- Modellpassung/ Likelihood
- Modellvergleiche

Hierarchisch Geschachtelte Modelle

Zwei Modelle sind *hierarchisch geschachtelte Modelle* genau dann, wenn ein Modell ein Spezialfall des anderen Modells ist

bzw.

Zwei Modelle sind *hierarchisch geschachtelte Modelle* genau dann, wenn ein Modell alle Parameter des anderen Modells enthält und noch mehr

Beispiel:

(1) $Y = a + b \qquad + (1|subject) + (1|item) + \epsilon$

(2) $Y = a + b + \mathbf{ab} + (1|subject) + (1|item) + \epsilon$

Hier ist (1) das *einfachere* Modell und (2) das *komplexere*, da (1) weniger Parameter enthält

Modellpassung/ Likelihood

Likelihood ist definiert als

$$L(Parameter) = P(Daten|Parameter)$$

Es wird genau jenes Parameterset als Modellparameter geschätzt, das das Auftreten der Daten am wahrscheinlichsten macht (Maximum-Likelihood-Schätzung)

Beispiel: $p_{Niete} = 0.9$ bei 18 Nieten und 2 Gewinnlosen

Generell: Je höher die Likelihood, desto besser beschreibt ein Modell die Daten

Modellvergleiche

Geschachtelte Modelle können anhand ihrer Likelihood miteinander verglichen werden (Likelihood-Ratio-Test)

Die Likelihood des komplexeren Modells ist **immer** größer (oder gleich) der des einfacheren

Aber: Ist sie signifikant größer?

Modellvergleiche

Trade-Off
(vgl. Occam's Razor: *"Entia non sunt multiplicanda praeter necessitatem"*)

Nutzen: Passung des Modells (Likelihood)
Kosten: Zusätzliche Parameter im Modell

Der Nutzen muss die Kosten rechtfertigen!
(Der Likelihood-Ratio-Test implementiert dieses Prinzip)

Hands On: Modellvergleiche

Start: *Nullmodel*

```
m0 <- lmer(RT ~ (1 |VP) + (1 |Item), dat, REML = F)
```

Test auf Signifikanz für Grammatikalität:

```
m1 <- lmer(RT ~ Gramm +
                (1 |VP) + (1 |Item), dat, REML = F)
anova(m0,m1)
```

Test auf Signifikanz für Präsentationsdauer:

```
m2 <- lmer(RT ~ PresT +
                (1 |VP) + (1 |Item), dat, REML = F)
anova(m0,m2)
```

Hands On: Ergebnisse anschauen

- Die Ergebnisse:

```
> anova(m1,m0)
Data: dat
Models:
m0: RT ~ 1 + (1 | VP) + (1 | Item)
m1: RT ~ Gramm + (1 | VP) + (1 | Item)
   Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
m0  4 69257 69284 -34625    69249
m1  5 61604 61638 -30797    61594 7655.5      1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
> anova(m2,m0)
Data: dat
Models:
m0: RT ~ 1 + (1 | VP) + (1 | Item)
m2: RT ~ PresT + (1 | VP) + (1 | Item)
   Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
m0  4 69257 69284 -34625    69249
m2  5 68233 68267 -34112    68223 1026.3      1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hands On: Modellvergleiche

Test auf Interaktion:

```
m3 <- lmer(RT ~ Gramm + PresT +
                (1 |VP) + (1 |Item), dat, REML = F)
m4 <- lmer(RT ~ Gramm + PresT + Gramm:PresT
                (1 |VP) + (1 |Item), dat, REML = F)
anova(m4,m3)
```

```
> anova(m4,m3)
Data: dat
Models:
m3: RT ~ Gramm + PresT + (1 | VP) + (1 | Item)
m4: RT ~ Gramm + PresT + Gramm:PresT + (1 | VP) + (1 | Item)
   Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
m3  6 57296 57336 -28642    57284
m4  7 56505 56552 -28245    56491 793.02      1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Reihenfolge der Tests

- Der Interaktionsparameter `Gramm:PresT` ist nur dann sinnvoll, wenn das Modell schon die Parameter `Gramm` und `PresT` enthält!
- Eine Interaktion höherer Ordnung benötigt immer alle "niedrigeren" Parameter
- Beispiel: Dreifachinteraktion `a:b:c` benötigt notwendig auch folgende Parameter im Modell: `a, b, c, a:b, b:c, a:c`

Kovariaten

Gemischte Modelle erlauben einfaches Einfügen von Kovariaten in das Modell

Beispiel:

$$\begin{aligned} \text{RT} \sim \text{Gramm} + \text{PresT} + \text{Gramm:PresT} \\ + \text{Satzlänge} + \text{Muttersprache} \\ + (1 \mid \text{VP}) + (1 \mid \text{Item}) \end{aligned}$$

Hypothesentests sind auch mit Kovariaten mglich. Für Test auf Interaktion vergleiche das obige Modell mit

$$\begin{aligned} \text{RT} \sim \text{Gramm} + \text{PresT} + \\ + \text{Satzlänge} + \text{Muttersprache} \\ + (1 \mid \text{VP}) + (1 \mid \text{Item}) \end{aligned}$$

Berichten der Ergebnisse - Ein Beispiel

"We used the *lme4* package (Bates, Maechler & Bolker, 2014) for R (R Core Team, 2014) to perform a linear mixed effects analysis for the influence of grammaticality stimulus duration on reaction times. As fixed effects, we entered grammaticality and stimulus duration into the model. As random effects, we entered random intercepts for subjects as well as items.
We tested for the significance of our fixed effects by performing likelihood ratio tests of the full model with the effect in question against the model without the effect in question."

Berichten der Ergebnisse - Ein Beispiel

"The analysis yielded a significant effect of grammaticality ($\chi^2(1) = 7655.5, p < .001$) as well as an additional effect of stimulus duration ($\chi^2(1) = 4310, p < .001$).
Furthermore, we found a significant interaction between both variables ($\chi^2(1) = 793.02, p < .001$). The model parameters of the final model (containing both main effects and interaction effect) and their confidence intervals are shown in Table 1."

Fortführung

- Messwiederholungen - Random Effect Structures

Messwiederholungen

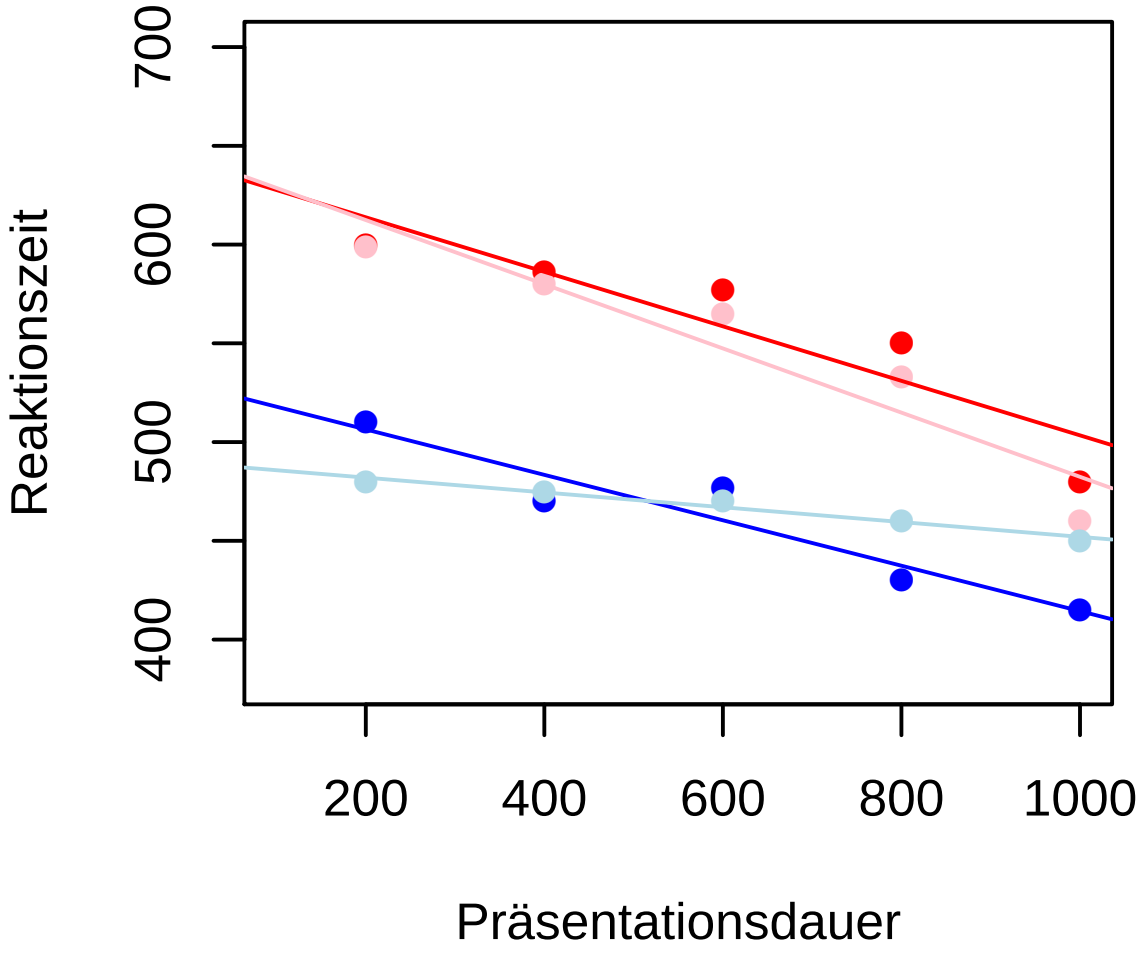
Bisher haben wir nur *Random Intercepts* betrachtet:

Für jede VP bzw. jedes Item wird ein bestimmter konstanter Einfluss auf die RTs angenommen
(zB kann VP3 generell 50ms langsamer sein als der Durchschnitt)

In den Beispieldaten ist aber jede VP (und jedes Item) in jeder Experimentalbedingung (vollständige Messwiederholung)

$\Longrightarrow$ Was, wenn die Bedingungen für verschiedene VPs unterschiedlich starken Einfluss haben?

Messwiederholungen



Messwiederholungen

Da bei Messwiederholungen dieser Fall nicht ausgeschlossen werden kann, sollten Random Slopes ins Modell mit aufgenommen werden (Barr et al. 2013)

Dies entspricht individuellen Steigungen für jede VP bzw jedes Item, auf dem eine Messwiederholung stattfindet

Das vollständige Modell für die Beispieldaten sieht also folgendermaßen aus:

$$\text{RT} \sim \text{Gramm*PresT} + (\text{Gramm*PresT} \,|\text{VP}) + (\text{Gramm*PresT} \,|\text{Item})$$

Messwiederholungen: Between- und Within-Designs

Within: VPs und Items

$$\text{RT} \sim \text{Gramm*PresT} + (\text{Gramm*PresT} \,|\text{VP}) + (\text{Gramm*PresT} \,|\text{Item})$$

Within: VPs, Between: Items

$$\text{RT} \sim \text{Gramm*PresT} + (\text{Gramm*PresT} \,|\text{VP}) + (1 \,|\text{Item})$$

Within: Items, Between: VPs

$$\text{RT} \sim \text{Gramm*PresT} + (1 \,|\text{VP}) + (\text{Gramm*Pres} \,|\text{Item})$$

Between: VPs und Items

$$\text{RT} \sim \text{Gramm*PresT} + (1 \,|\text{VP}) + (1 \,|\text{Item})$$

Messwiederholungen: Between- und Within-Designs

Die Random Effect Structure spiegelt also direkt das Experimentaldesign wieder!

Beispiel: Das Material besteht aus semantisch sinnlosen und sinnvollen Sätzen, wobei keine Minimalpaare möglich sind. Man kann also nicht annehmen, dass es einen Satz als sinnlose und sinnvolle Variante gibt. Jede Person sieht jeden Satz des Materials. Dabei sind die Hälfte der VPs L1-Sprecher, die andere Hälfte L2-Sprecher.

Was für Random Slopes sollten daher ins Modell aufgenommen werden?

Messwiederholungen: Between- und Within-Designs

Antwort:

RT $\sim$ Sinn*Sprache + (Sinn |VP) + (Sprache |Item)

Jedes Item wird von L1- und L2- Sprechern bearbeitet, liefert also hier Werte für beide Bedingungen von Sprache
Jede VP bearbeitet sinnlose und sinnvolle Sätze, liefert also hier Werte für beide Bedingungen von Sinn

Messwiederholungen: Hypothesentests

Ein konvergierendes Modell ist in jedem Fall wichtiger als eine vollständige Random Effect Structure!

Was, wenn das Modell nicht konvergiert?

Vereinfachung der Random Effect Structure auf ein noch zu rechtfertigendes Format (durch inhaltliche Punkte oder durch forward- oder backward selection).

Messwiederholungen: Hypothesentests

Beispiel: Hypothesentest auf Interaktion zweier Faktoren
Beide Faktoren haben within-subjects Messwiederholung
Nur Faktor A hat within-subjects Messwiederholung

```
m0 <- lmer(RT ~ a + b +
                (a*b |VP) + (a |Item), dat, REML = F)
m1 <- lmer(RT ~ a + b + a:b
                (a*b |VP) + (a |Item), dat, REML = F)
anova(m0,m1,test="Chisq")
```

Binäre Kriteriumsvariablen

Binäre Kriteriumsvariablen

Beispiel: Beurteilung von VPs über Korrektheit von Sätzen
Drei Prädiktoren a, b, c

Daten:

VP	Item	a	b	c	Answer
1	1	1	13	1	1
1	2	1	14	2	0
1	3	1	8	1	1
1	4	1	7	2	1

Binäre Kriteriumsvariablen

Theoretischer Hintergrund:

- Lineare Modelle setzen *stetige* Kriterien voraus, nicht kategoriale
- Über die *Wahrscheinlichkeit* des Beobachtens einer Kategorie lassen sich jedoch stetige Variablen erzeugen (z.B. sogenannte *Logits*)
- Diese können durch lineare Modelle vorhergesagt werden

Binäre Kriteriumsvariablen

Umsetzung in R mit `glmer`:

```
model <- glmer(Answer ~ a*b*c + (1 |VP) + (1 |Item),
                        dat, family = "binomial")
```

Ausführliche Anleitung unter:
http://www.ats.ucla.edu/stat/r/dae/melogit.htm