Kumuliertes gemischtes Logit-Modell - ein Modell für Boniturdaten aus dem Pflanzenschutz?

Dr. Mareike Kohlmann, GVC/S Scientific Computing

BASF
The Chemical Company

Sommertagung 2011 30.6.2011

AG Landwirtschaftliches Versuchswesen der Dt. Region der IBS

Übersicht

- Aktuelle Projekte mit Ordinaldaten in der BASF bei Scientific Computing
- Datenbeispiel aus dem Unternehmensbereich Pflanzenschutz
- Status Quo: Auswertung der aggregierten Daten mit allgemeinen linearen Modellen (GLS)
- Vorschlag: Kumulative gemischte Logit-Modelle (CLMM) mit Anwendung auf BASF-Agrodaten

Aktuelle Projekte mit Ordinaldaten bei Scientific Computing

- Bewertung von **Schuppenshampoos** in Anwendungsentwicklung
  - Patienten bewerten Effekt auf ordinaler Datenskala
  - Zufällige Effekte durch Heterogenität der Patienten zu Studienbeginn und bzgl. des Behandlungeffekts

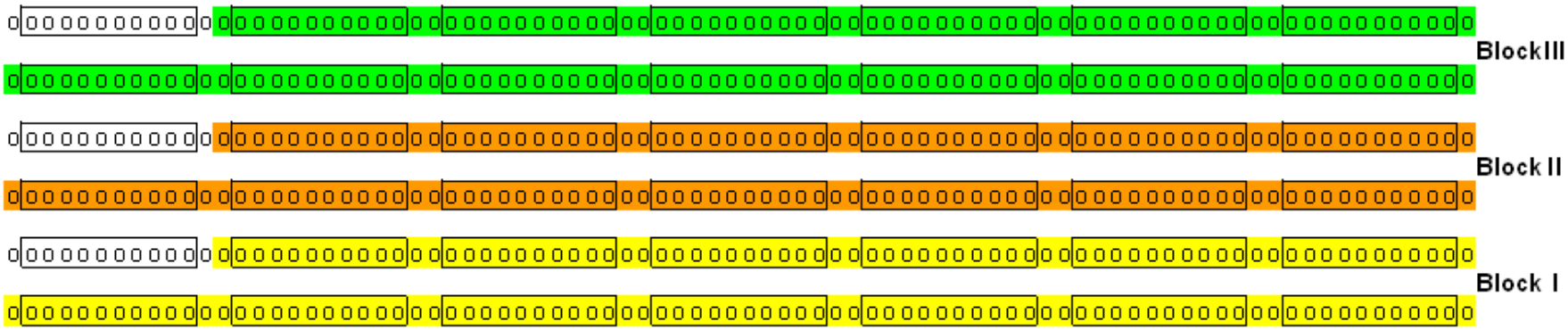
- Bewertung von Spülleistung in Automatic Dishwashing Lab:
  - Laboranten bewerten Spülergebnis auf ordinaler Datenskala
  - Zufälliger Effekt des Bewerters oder der Spülmaschine möglich

Agrodaten:
Mehltau an Weinreben

BASF
The Chemical Company

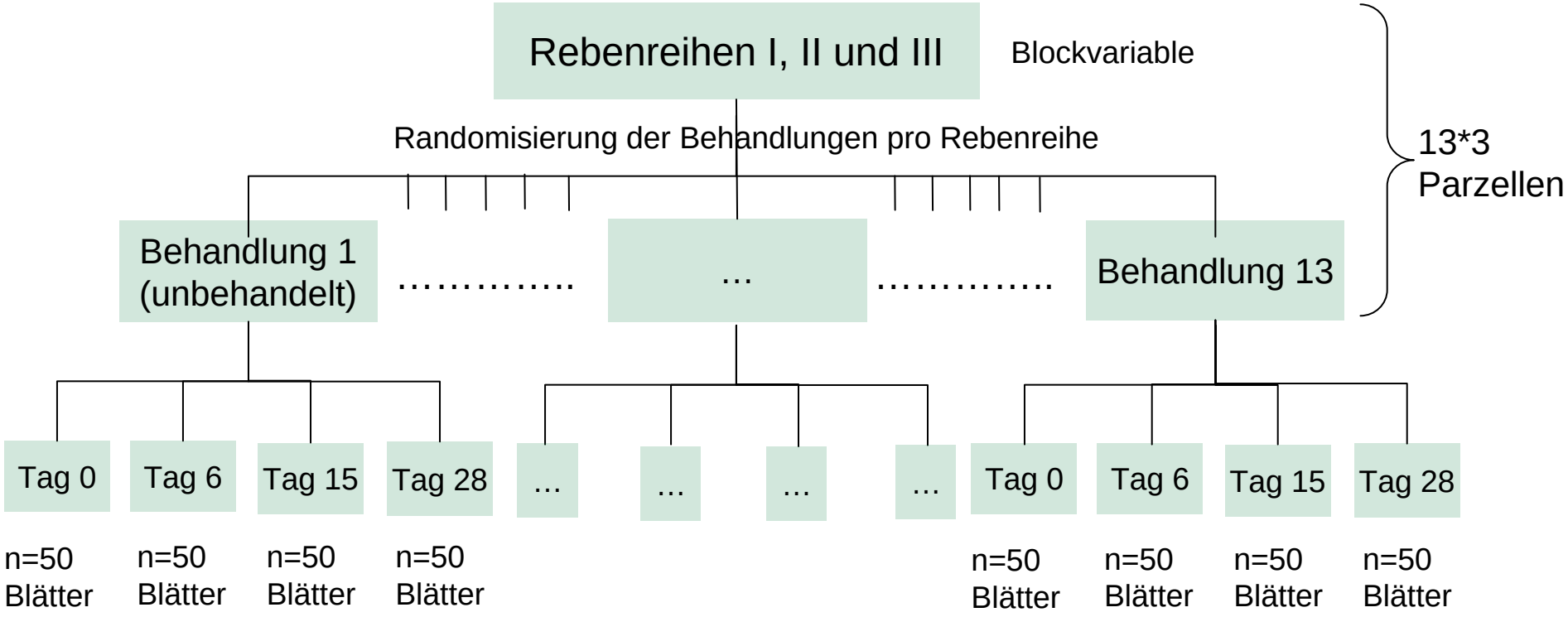
Datendesign:

- Randomisierung der 13 Behandlungen innerhalb des Blocks, je 10 Pflanzen pro Behandlung
- Bonitur des Mehltaus von 5 Blätter pro Weinpflanze
- 4malige Bonitur: an Tag 0, 6, 15 und 28 nach letzter Beh.



Agrodaten:
Mehltau an Weinreben

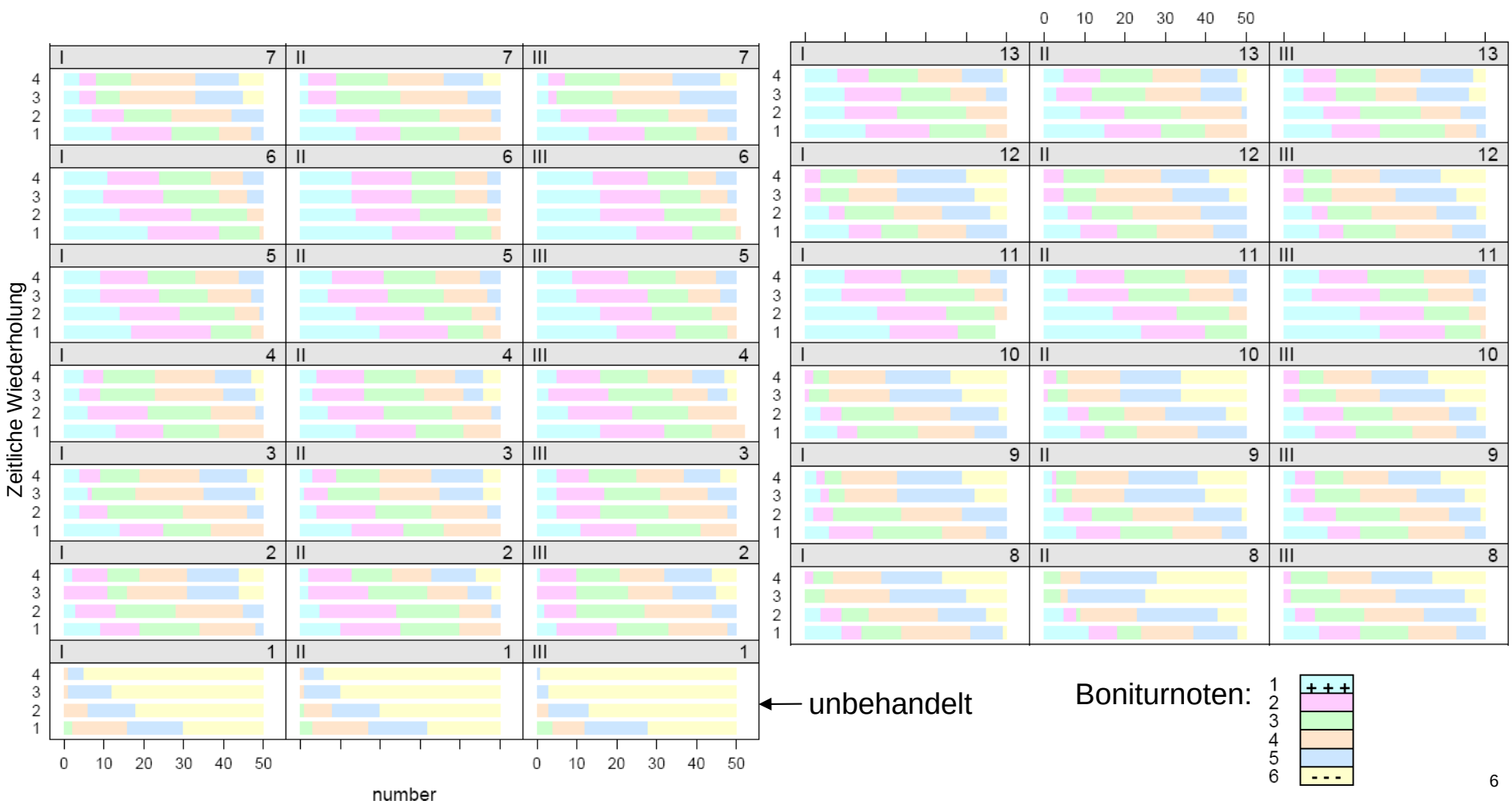
Randomisiertes Blockdesign:



Agrodaten:
Mehltau an Weinreben

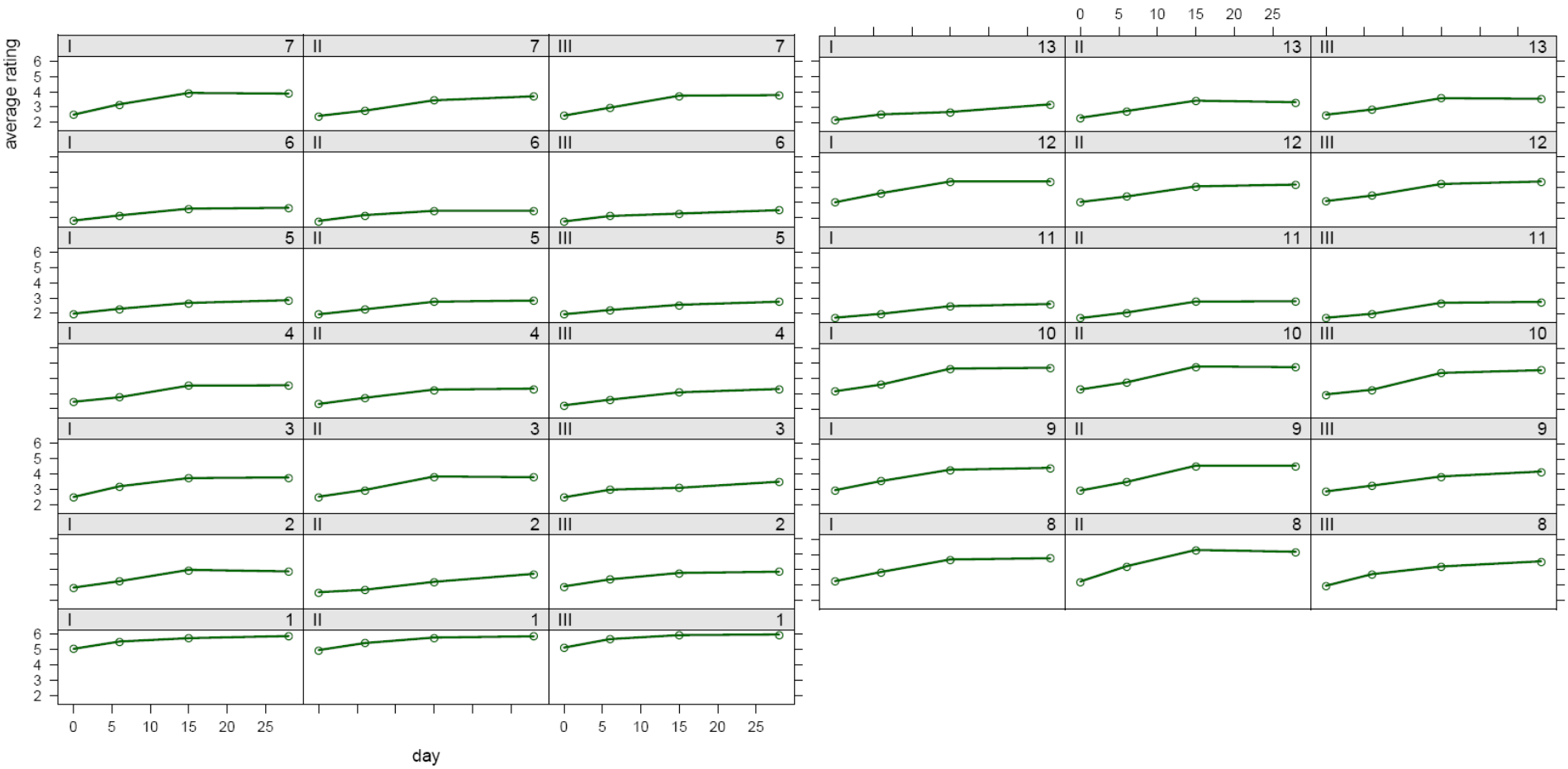
BASF
The Chemical Company

Boniturdaten:



Anwendung auf Agrodaten: Profile der mittleren Bonitur

Neue Zielgröße: Gewichtetes Mittel aus 50 Bonituren pro Tag und Parzelle



Anwendung auf Agrodaten: LM für aggregierte Daten

BASF
The Chemical Company

Bestes lineares Modell:

$$rating_w = \beta_0 + \beta_t\ treat_t + \beta_s\ dayf_s + \beta_u\ treat_t : dayf_s + \varepsilon_w$$

$w:1,..4 \quad t:1,...,12 \quad s:1,2,3 \quad u:1,...,36$

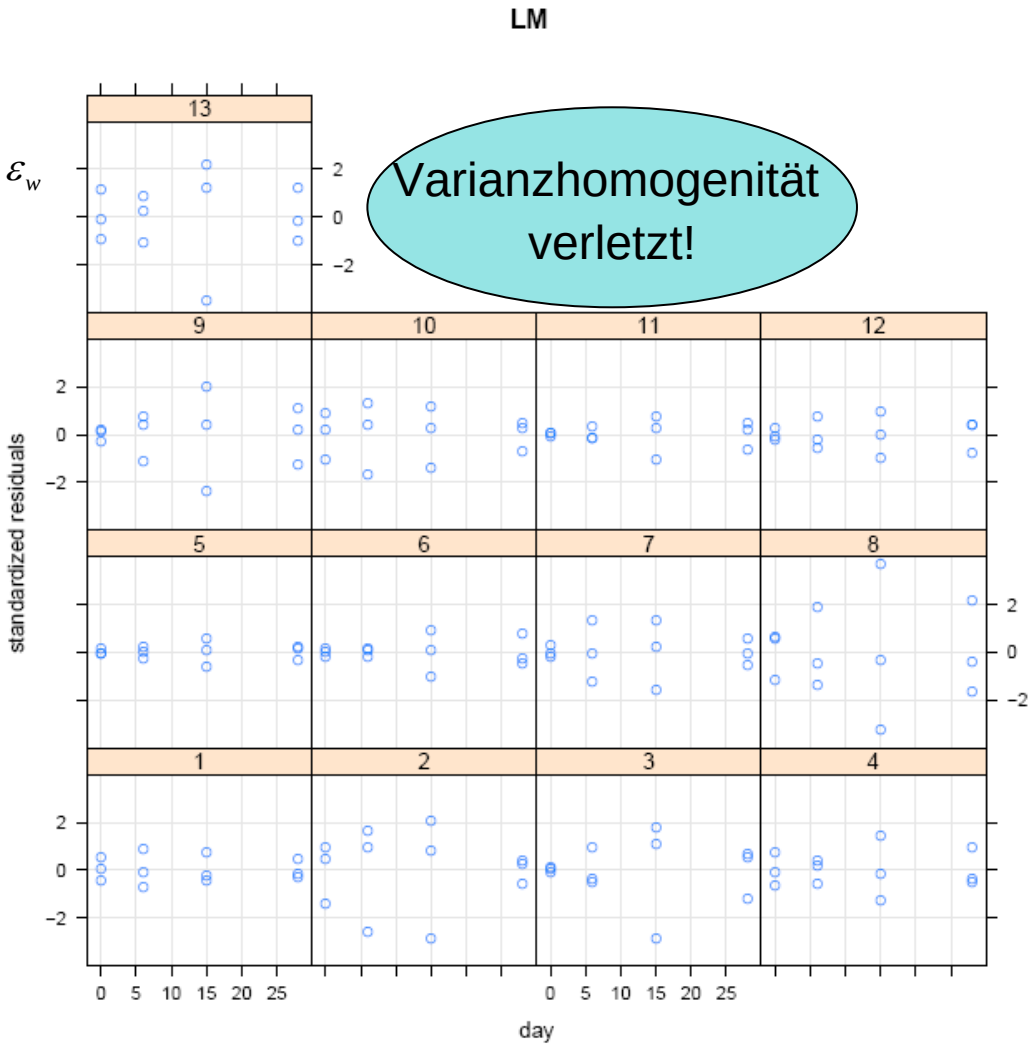
$\varepsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{I})$

-> Erweiterung zu allgemeinem linearen Modell (GLS)

$\varepsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{V})$

$\mathrm{V} = \mathrm{diag}(\mathrm{v}_1^2, .., \mathrm{v}_4^2)$

$\mathrm{Hier}: \mathrm{V} = \mathrm{diag}(\mathrm{v}_1^2, \mathrm{v}_1^2, \mathrm{v}_2^2, \mathrm{v}_1^2)$



Anwendung auf Agrodaten: GLS für aggregierte Daten

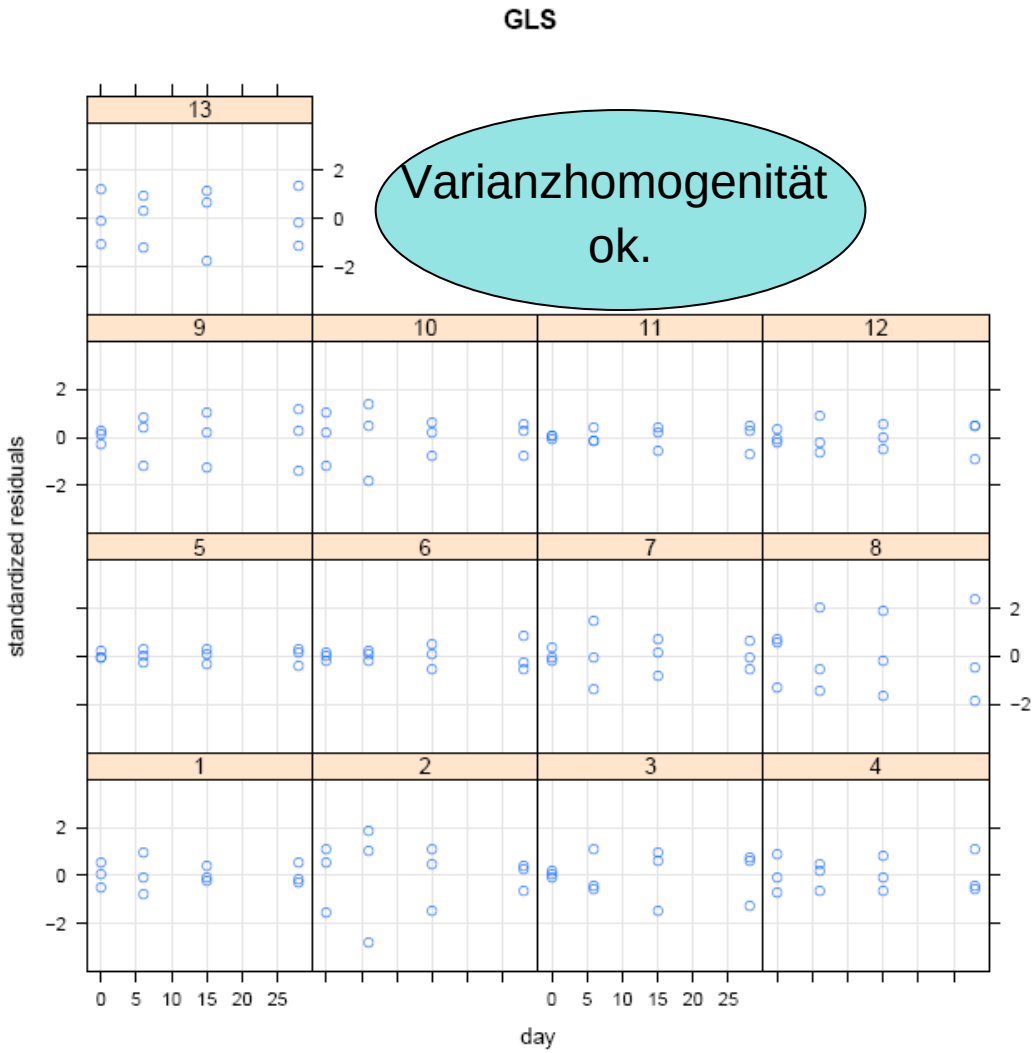
BASF
The Chemical Company

Schätzung der Varianzkomponenten:

v1: 0.15 [0.13;0.17]

v2 (Tag 15): 2.10 [1.52;2.84] * v1

Reduktion des BICs von LM auf GLS um mehr als 20!

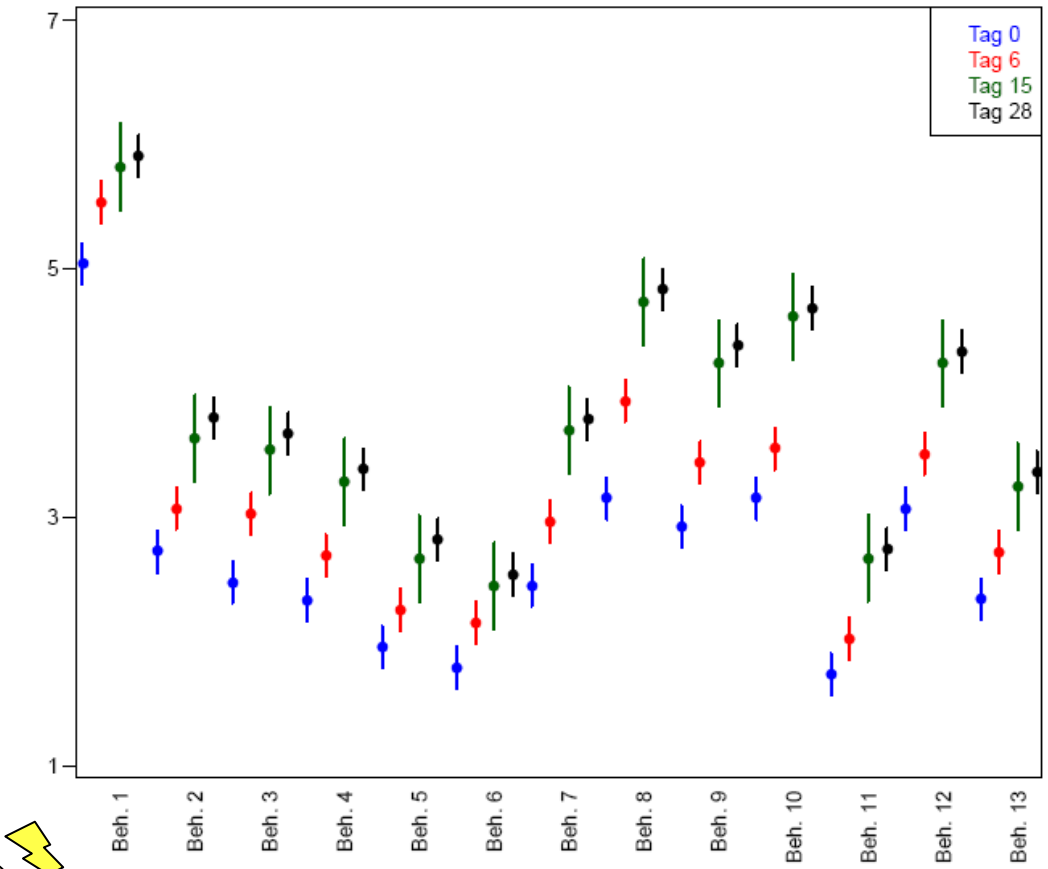


Anwendung auf Agrodaten: GLS für aggregierte Daten

BASF
The Chemical Company

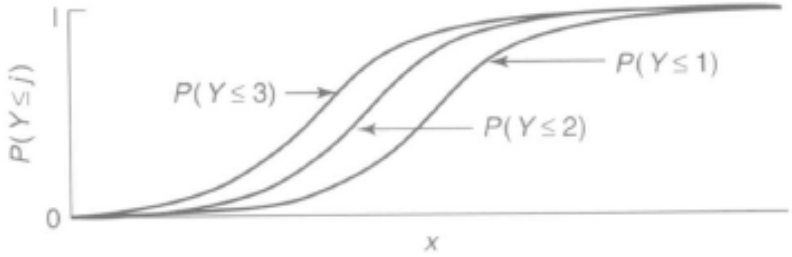
Behandlungseffekte

Beh.	Effekt	95% CI			
6 Tag 0	1,8	[1,6 ; 2,0]	2 Tag 0	2,7	[2,6 ; 2,9]
6 Tag 6	2,2	[2,0 ; 2,3]	2 Tag 6	3,1	[2,9 ; 3,2]
6 Tag 15	2,4	[2,1 ; 2,8]	2 Tag 15	3,6	[3,3 ; 4,0]
6 Tag 28	2,5	[2,4 ; 2,7]	2 Tag 28	3,8	[3,6 ; 4,0]
11 Tag 0	1,7	[1,6 ; 1,9]	7 Tag 0	2,5	[2,3 ; 2,6]
11 Tag 6	2,0	[1,9 ; 2,2]	7 Tag 6	3,0	[2,8 ; 3,1]
11 Tag 15	2,7	[2,3 ; 3,0]	7 Tag 15	3,7	[3,4 ; 4,0]
11 Tag 28	2,7	[2,6 ; 2,9]	7 Tag 28	3,8	[3,6 ; 4,0]
5 Tag 0	2,0	[1,8 ; 2,1]	9 Tag 0	2,9	[2,8 ; 3,1]
5 Tag 6	2,3	[2,1 ; 2,4]	9 Tag 6	3,4	[3,3 ; 3,6]
5 Tag 15	2,7	[2,3 ; 3,0]	9 Tag 15	4,2	[3,9 ; 4,6]
5 Tag 28	2,8	[2,7 ; 3,0]	9 Tag 28	4,4	[4,2 ; 4,6]
4 Tag 0	2,3	[2,2 ; 2,5]	12 Tag 0	3,1	[2,9 ; 3,2]
4 Tag 6	2,7	[2,5 ; 2,9]	12 Tag 6	3,5	[3,3 ; 3,7]
4 Tag 15	3,3	[2,9 ; 3,6]	12 Tag 15	4,2	[3,9 ; 4,6]
4 Tag 28	3,4	[3,2 ; 3,6]	12 Tag 28	4,3	[4,2 ; 4,5]
13 Tag 0	2,3	[2,2 ; 2,5]	8 Tag 0	3,2	[3,0 ; 3,3]
13 Tag 6	2,7	[2,6 ; 2,9]	8 Tag 6	3,9	[3,8 ; 4,1]
13 Tag 15	3,2	[2,9 ; 3,6]	8 Tag 15	4,7	[4,4 ; 5,1]
13 Tag 28	3,4	[3,2 ; 3,5]	8 Tag 28	4,8	[4,7 ; 5,0]
3 Tag 0	2,5	[2,3 ; 2,6]	10 Tag 0	3,2	[3,0 ; 3,3]
3 Tag 6	3,0	[2,9 ; 3,2]	10 Tag 6	3,6	[3,4 ; 3,7]
3 Tag 15	3,5	[3,2 ; 3,9]	10 Tag 15	4,6	[4,3 ; 5,0]
3 Tag 28	3,7	[3,5 ; 3,8]	10 Tag 28	4,7	[4,5 ; 4,8]
			1 Tag 0	5,0	[4,9 ; 5,2]
			1 Tag 6	5,5	[5,4 ; 5,7]
			1 Tag 15	5,8	[5,5 ; 6,2]
			1 Tag 28	5,9	[5,7 ; 6,1]



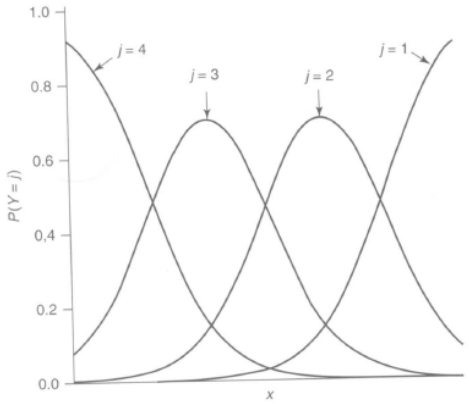
Kumulative Modelle für Ordinaldaten

Modellierung von kumulativen Logits:

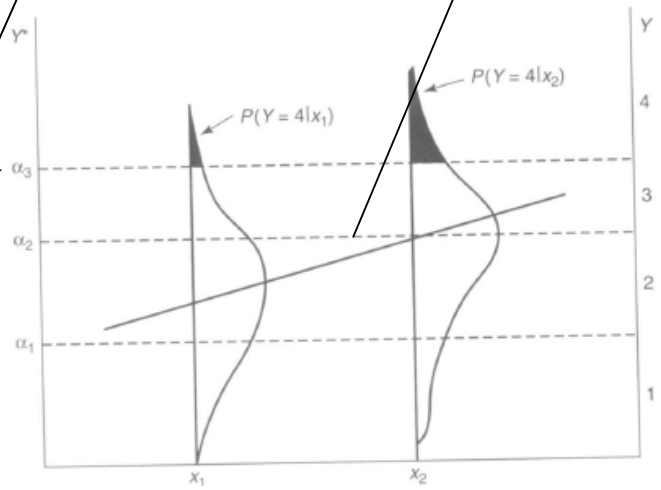


$$\text{logit}[P(Y_i \le j) \mid \mathbf{x}_i] = \alpha_j - (\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots)$$

entsprechende Kategorie-spezifische Wahrscheinlichkeiten:



Schwellen-wert



bei negativem Effekt von x_2 auf Mittelwert

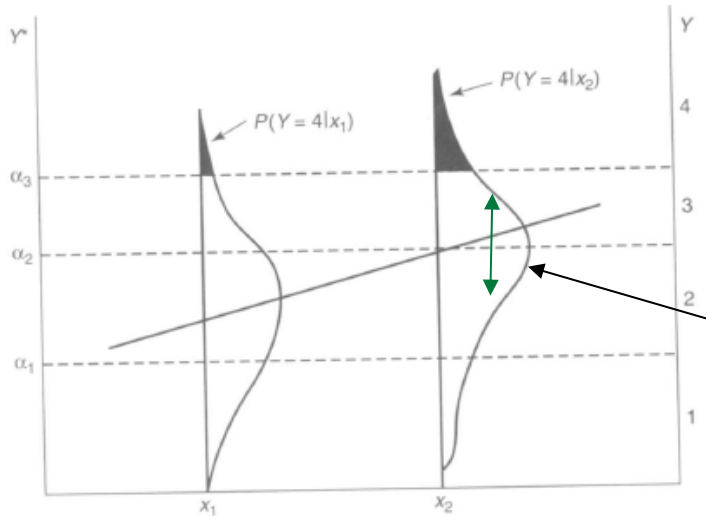
Kumulative gemischte Modelle für Ordinaldaten (CLMM)

Es gilt für Beobachtung t in Parzelle i

Zufälliger Parzelleneffekt

$$u_i \sim N(0, \sigma_u^2)$$

$$\text{logit}[P(Y_{it} \le j) | \mathbf{x_{it}}] = \frac{\alpha_j - (u_i + \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots)}{\exp(\gamma x_T)}$$



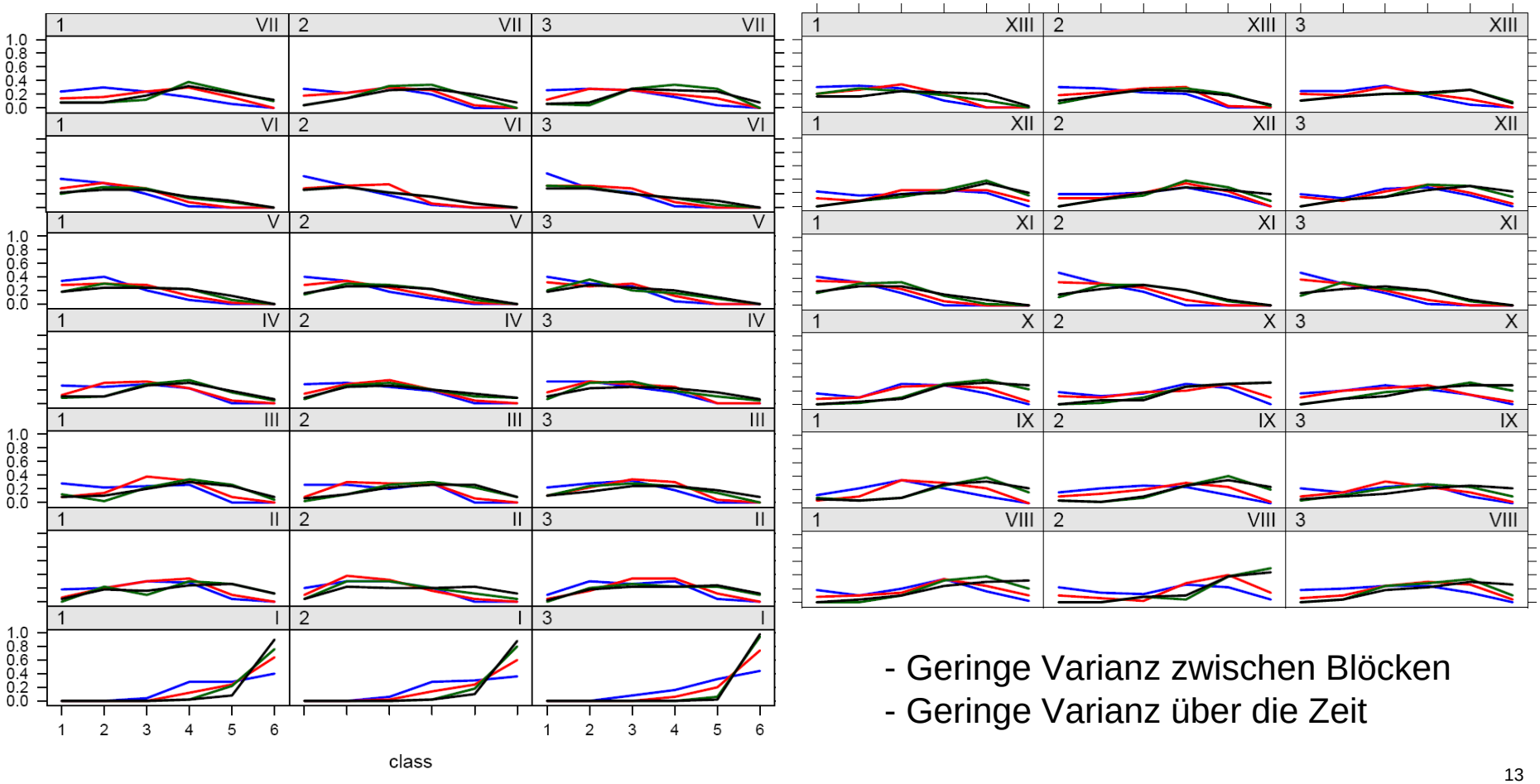
Dispersionseffekt für Kovariable x_T

Hier für x_2

Anwendung auf Agrodaten: CLMM

BASF
The Chemical Company

Beobachtete Anteile an Tag 0, 6, 15 und 28:



- Geringe Varianz zwischen Blöcken
- Geringe Varianz über die Zeit

Anwendung auf Agrodaten: CLMM

BASF
The Chemical Company

Modellbildung

	Schwellenwerte			Feste Effekte	Varianzkomponente	Zufälliger Effekt	BIC
	flexibel	symmetrisch	äquidistant	Beh., Zeitpunkt u. Interaktion	Tag 28	Parzelle	
Modell 1	x			x			23532
Modell 2	x			x	x		23503
Modell 3	x			x		x	23498
Modell 4	x			x	x	x	23467
Modell 5		x		x	x	x	23545
Modell 6			x	x	x	x	23560

R Code für Modell 4

```
require(ordinal)

clmm(bonitur ~ treatf*dayf,
   random=cellf, data=rawdata,
   weights=number,
   scale=~day28,link="logistic",
   Hess=TRUE, method="nlminb",
   threshold="flexible", nAGQ=10)
```

Ergebnisse von Modell 4

Schwellenwerte 1|2: -5.2 2|3: -3.8
3|4: -2.6 4|5: -1.2 5|6: 0.6

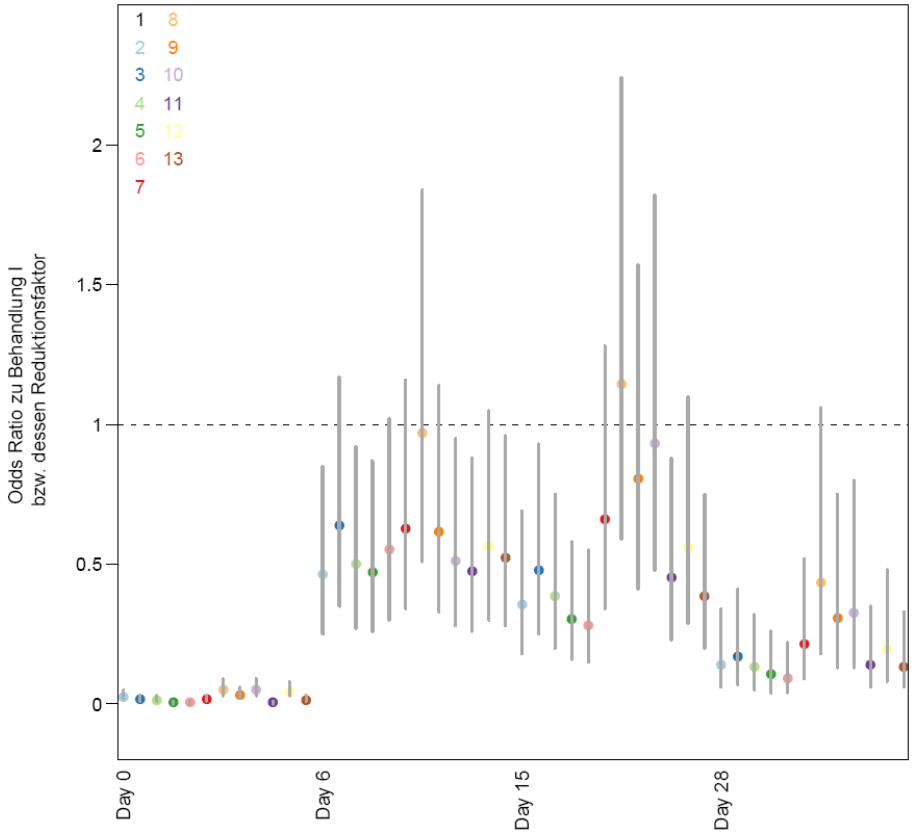
$s_{Parzelle}$: 0.21 [0.15;0.29]

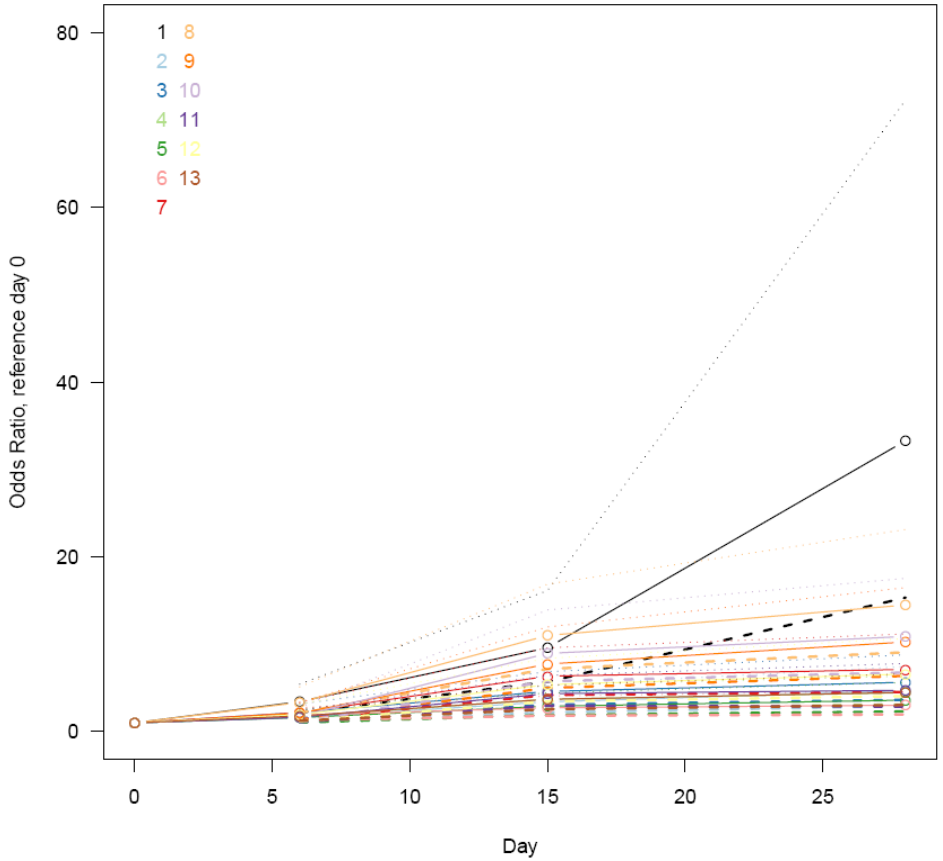
Dispersionsparameter für Tag 28:
1.17 [1.12;1.24]

Anwendung auf Agrodaten: CLMM, Modell 4

BASF
The Chemical Company

Behandlungseffekte





Anwendung auf Agrodaten: Vegleich der 2 Modellierungsansätze

- Modellstruktur:

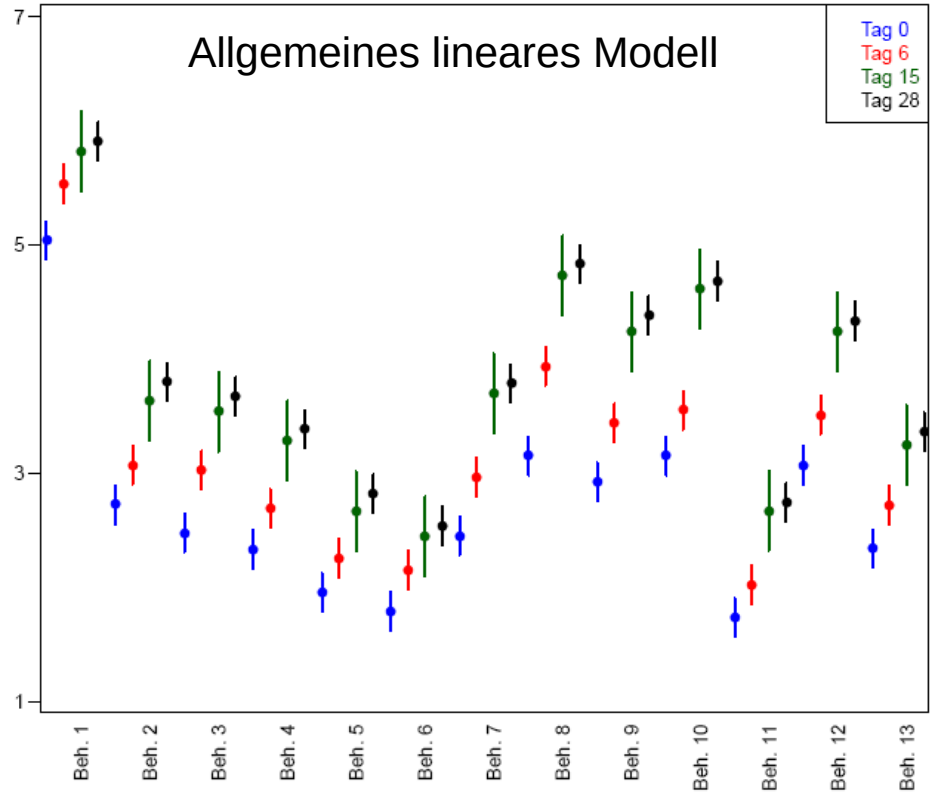
  - Bei beiden Modellen:
    Effekt durch Behandlung und über die Zeit, mit Interaktion

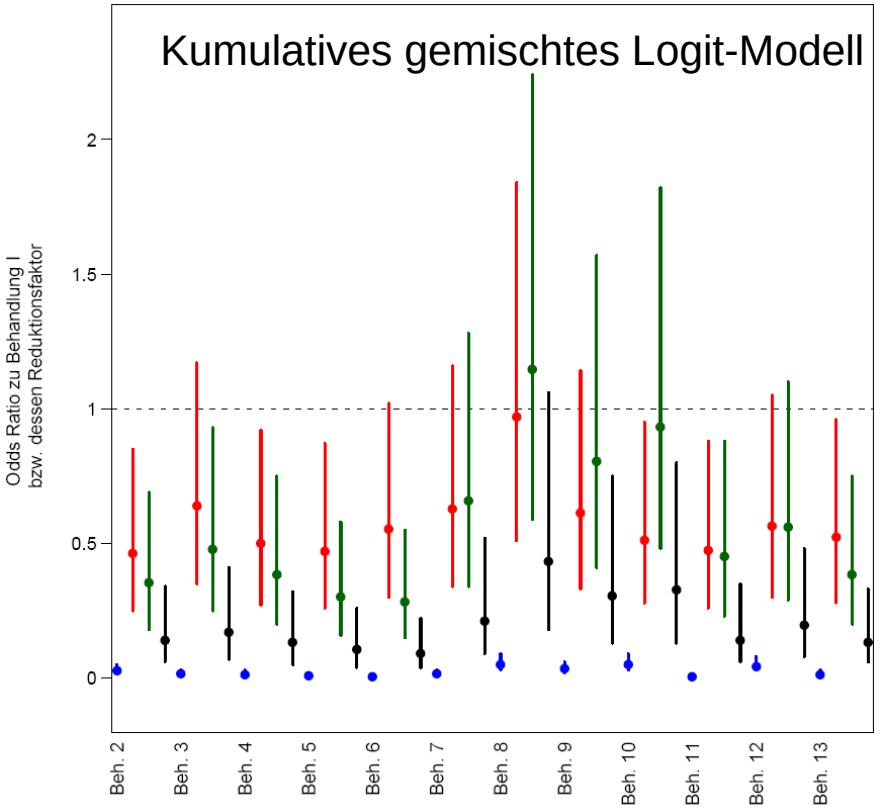
  - Varianzheterogenität bei Tag 15 für GLS, bei Tag 28 für CLMM

  - Zufälliger Parzelleneffekt in CLMM

Anwendung auf Agrodaten: Vegleich der 2 Modellierungsansätze

BASF
The Chemical Company





-> Sehr ähnliches Ranking bzgl. Behandlungseffekten, Behandlungsunterschiede bei ordinalem Datenniveau jedoch nicht signifikant

Literatur

- Übersichtsliteratur

  Agresti, A. (2010): Analysis of Ordinal Categorical Data, *Wiley*.

- ANOVA und ordinale Regressionmodelle

  Piepho, H.-P. (1997): Schwellenwertmodelle mit festen und zufälligen Effekten für Boniturdaten aus landwirtschaftlichen Versuchen, *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, 28, 183-195.

  Piepho, H.-P., Kalka, E. (2003): Threshold models with fixed and random effects for ordered categorical data, *Food Quality and Preference* 14, 343-357.

- Weiterführend: Bayes-Schätzung mit Agro-Daten

  Lee, A. C.-L. (2009): Random Effects Models for Ordinal Data, Dissertation, verfügbar unter https://researchspace.auckland.ac.nz/bitstream/handle/2292/4544/02whole.pdf?sequence=4

Diskussion

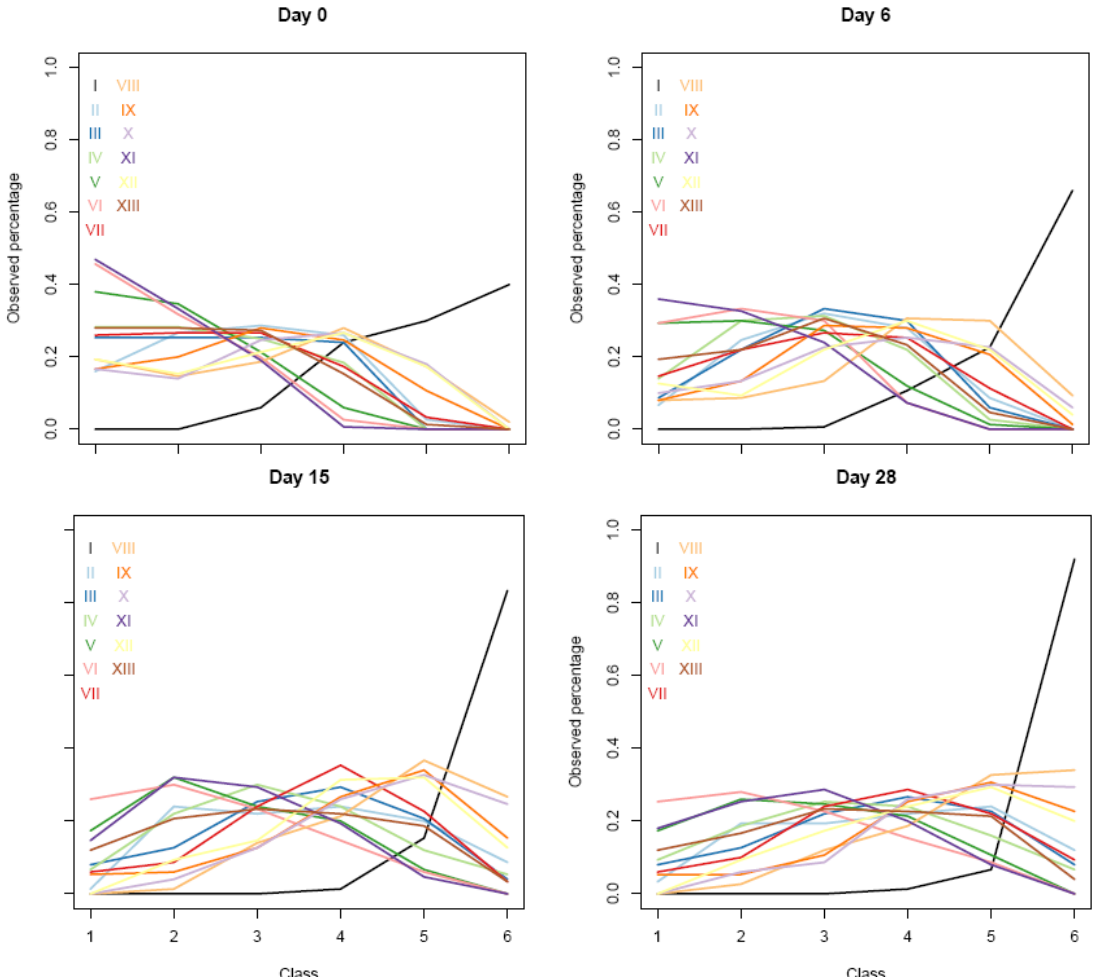
1) Erfahrungen des Publikums mit kumulativem gemischtem Logit-Modell:
   - Modellierung
   - Im Vergleich zu Analysen mit aggregierten Daten

2) Blockeffekt trotz Nichtsignifikanz im Modell belassen?

Appendix

- Beobachtete kumulierte Logits
- Exploration: Lineares Modell pro Parzelle für aggregierte Daten
- CLMM: Tabelle mit Behandlungseffekte
- CLMM: BLUPs des zufälligen Parzellen-Effekts
- Offene Modellierungsaspekte für CLMMs

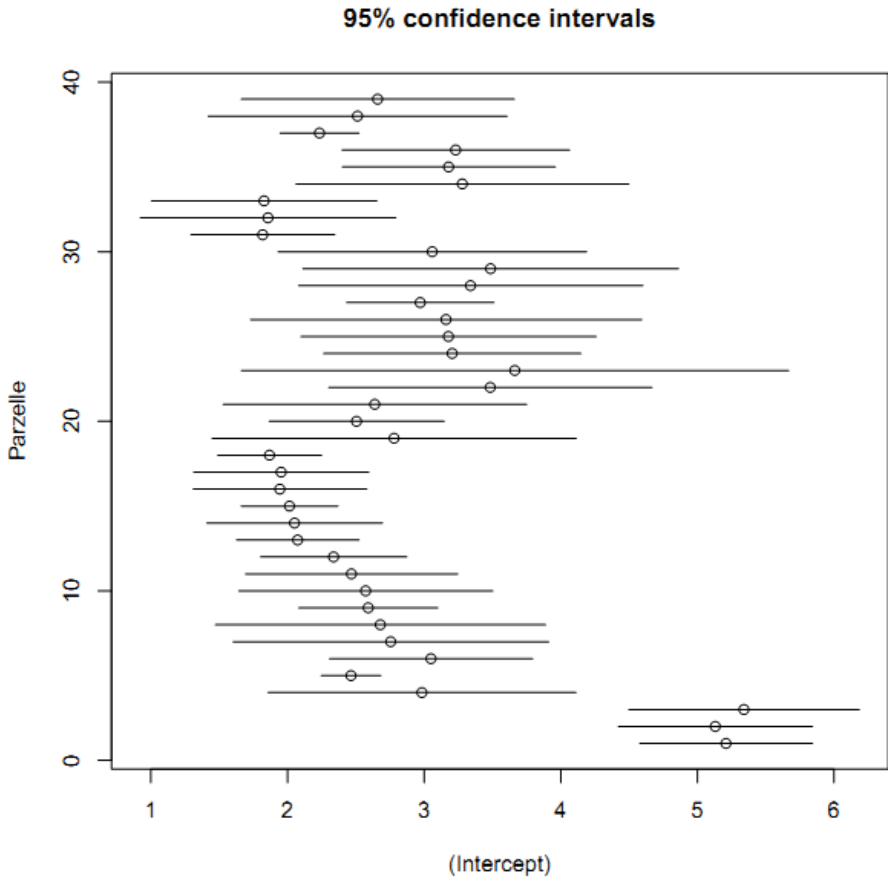
Anwendung auf Agrodaten: Beobachtete kumulierte Logits

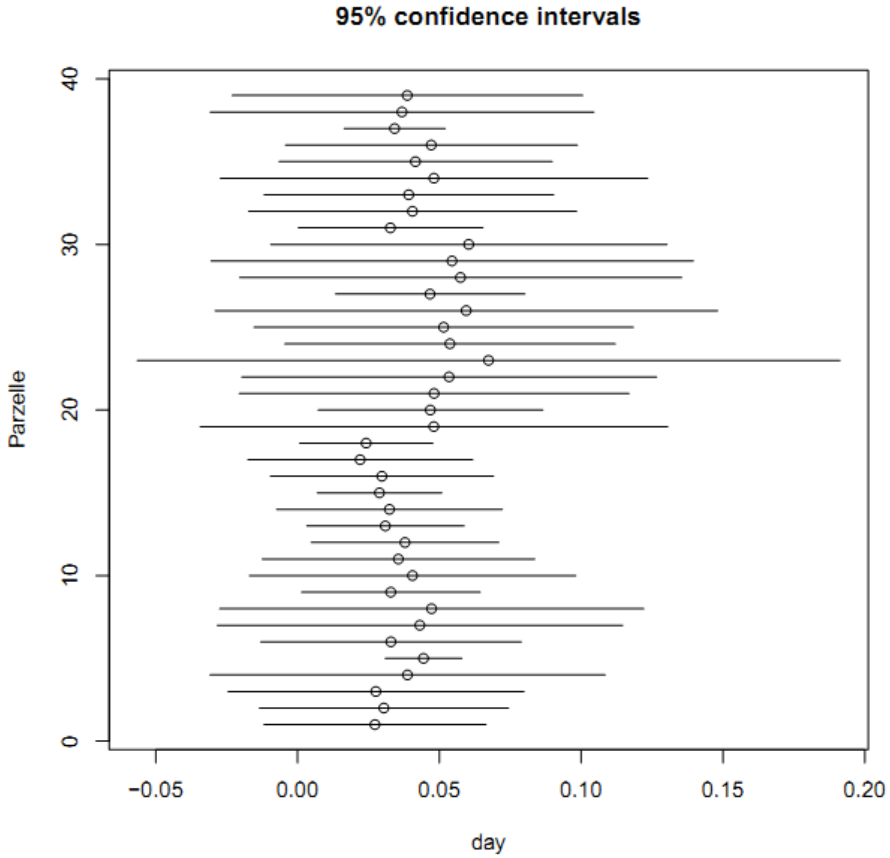


Klasse 1 = Befallsfrei
Klasse 2 = < 5% Befall
Klasse 3 = 5-10% Befall
Klasse 4 = 11 - 15% Befall
Klasse 5 = 26-50% Befall
Klasse 6 = > 50% Befall

Anwendung auf Agrodaten: L(M)M für aggregierte Daten

Exploration: Lineares Modell pro Parzelle





-> Variabilität zu Studienbeginn durch Behandlung,

jedoch nicht über die Zeit

Anwendung auf Agrodaten: CLMM, Modell 4

BASF
The Chemical Company

Behandlungseffekte

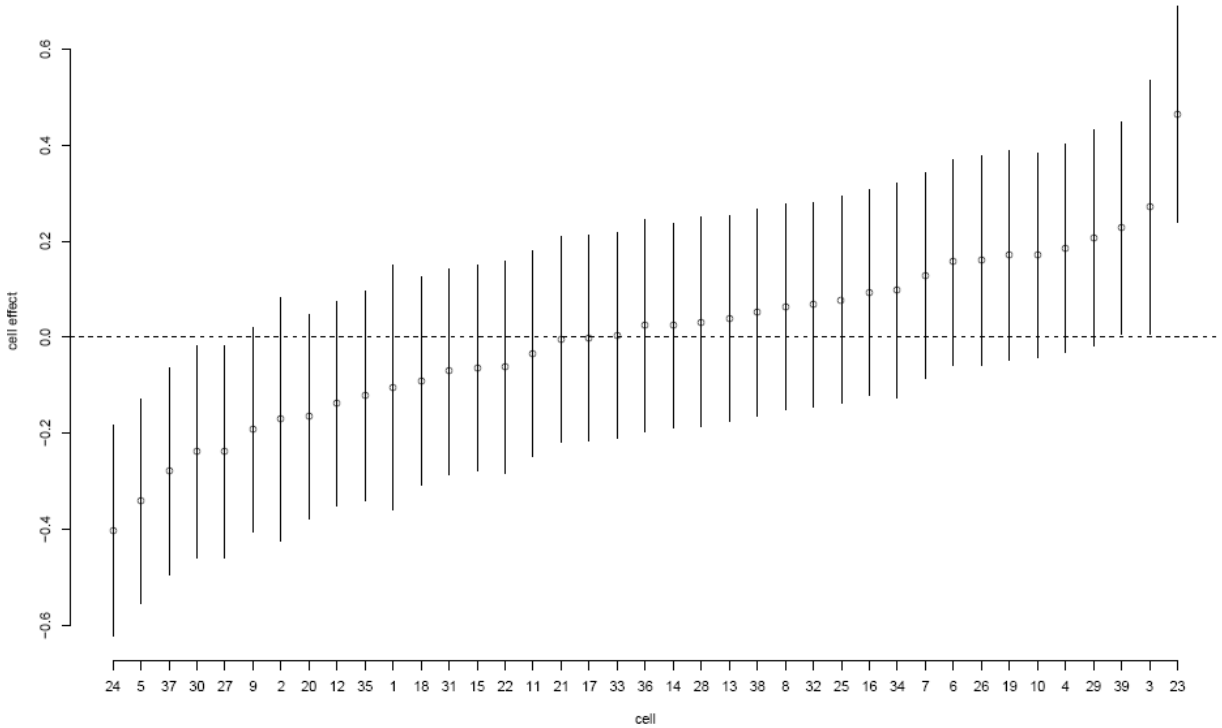
Behandlung	Effekt	lowerCI	upperCI	Effekt2	lowerCI	upperCI
1 Tag 0	1,00			1,00		
2 Tag 0	0,03	0,02	0,05	1,00		
3 Tag 0	0,02	0,01	0,03	1,00		
4 Tag 0	0,02	0,01	0,03	1,00		
5 Tag 0	0,01	0	0,01	1,00		
6 Tag 0	0,01	0	0,01	1,00		
7 Tag 0	0,02	0,01	0,03	1,00		
8 Tag 0	0,05	0,03	0,09	1,00		
9 Tag 0	0,04	0,02	0,06	1,00		
10 Tag 0	0,05	0,03	0,09	1,00		
11 Tag 0	0,01	0	0,01	1,00		
12 Tag 0	0,05	0,03	0,08	1,00		
13 Tag 0	0,02	0,01	0,03	1,00		
1 Tag 6	3,43	2,17	5,43	3,43	2,17	5,43
2 Tag 6	0,47	0,25	0,85	1,6	1,06	2,4
3 Tag 6	0,64	0,35	1,17	2,19	1,52	3,18
4 Tag 6	0,50	0,27	0,92	1,72	1,22	2,44
5 Tag 6	0,47	0,26	0,87	1,62	1,05	2,5
6 Tag 6	0,55	0,3	1,02	1,9	1,31	2,75
7 Tag 6	0,63	0,34	1,16	2,16	1,41	3,33
8 Tag 6	0,97	0,51	1,84	3,34	2,18	5,12
9 Tag 6	0,62	0,33	1,14	2,11	1,34	3,33
10 Tag 6	0,51	0,28	0,95	1,76	1,12	2,77
11 Tag 6	0,48	0,26	0,88	1,64	1,1	2,43
12 Tag 6	0,56	0,3	1,05	1,94	1,2	3,13
13 Tag 6	0,52	0,28	0,96	1,8	1,24	2,61
1 Tag 15	9,59	5,7	16,15	9,59	5,7	16,15
2 Tag 15	0,36	0,18	0,69	3,41	2,21	5,27
3 Tag 15	0,48	0,25	0,93	4,59	3,07	6,86
4 Tag 15	0,39	0,2	0,75	3,71	2,6	5,3
5 Tag 15	0,30	0,16	0,58	2,9	1,9	4,42
6 Tag 15	0,28	0,15	0,55	2,72	1,82	4,07
7 Tag 15	0,66	0,34	1,28	6,34	4,19	9,59
8 Tag 15	1,15	0,59	2,24	11	7,17	16,88
9 Tag 15	0,81	0,41	1,57	7,73	4,99	11,97
10 Tag 15	0,93	0,48	1,82	8,95	5,76	13,92
11 Tag 15	0,45	0,23	0,88	4,35	2,95	6,42
12 Tag 15	0,56	0,29	1,1	5,38	3,45	8,39
13 Tag 15	0,39	0,2	0,75	3,69	2,56	5,33
1 Tag 28	33,29	15,34	72,27	33,29	15,34	72,27
2 Tag 28	0,14	0,06	0,34	4,66	2,92	7,45
3 Tag 28	0,17	0,07	0,41	5,65	3,63	8,78
4 Tag 28	0,13	0,05	0,32	4,44	3,1	6,36
5 Tag 28	0,11	0,04	0,26	3,6	2,31	5,62
6 Tag 28	0,09	0,04	0,22	3,02	1,96	4,67
7 Tag 28	0,21	0,09	0,52	7,12	4,54	11,17
8 Tag 28	0,44	0,18	1,06	14,49	9,08	23,13
9 Tag 28	0,31	0,13	0,75	10,26	6,38	16,48
10 Tag 28	0,33	0,13	0,8	10,89	6,76	17,54
11 Tag 28	0,14	0,06	0,35	4,72	2,86	7,79
12 Tag 28	0,20	0,08	0,48	6,58	4,06	10,65
13 Tag 28	0,14	0,06	0,33	4,49	3,05	6,61

Anwendung auf Agrodaten: CLMM, Modell 4

BLUPs des zufälligen Effekts:

- 39 Parzellen mit $N(0,\sigma^2)$ mit $\hat{\sigma} = 0.21$ [0.15;0.29],

  impliziert eine Intra-Parzellenkorrelation von 0.01



Offene Modellierungsaspekte bei CLMM

- Weitere Modellerweiterungen möglich

  - andere Verteilungsannahmen für kumulative Logits

    (z.B. extreme minimal oder maximum value)

  - keine Parametrisierung über Schwellenwertmodell

    (z.B. adjacent categories, continuation-ratio Logits)

  - komplexere Struktur der zufälligen Effekte

  - multivariate Zielgröße