Workshop

„On-Farm-Experimente“ 23./24.11.2011 Kassel

Entwicklung des Auswertungsmodells

Joachim Spilke

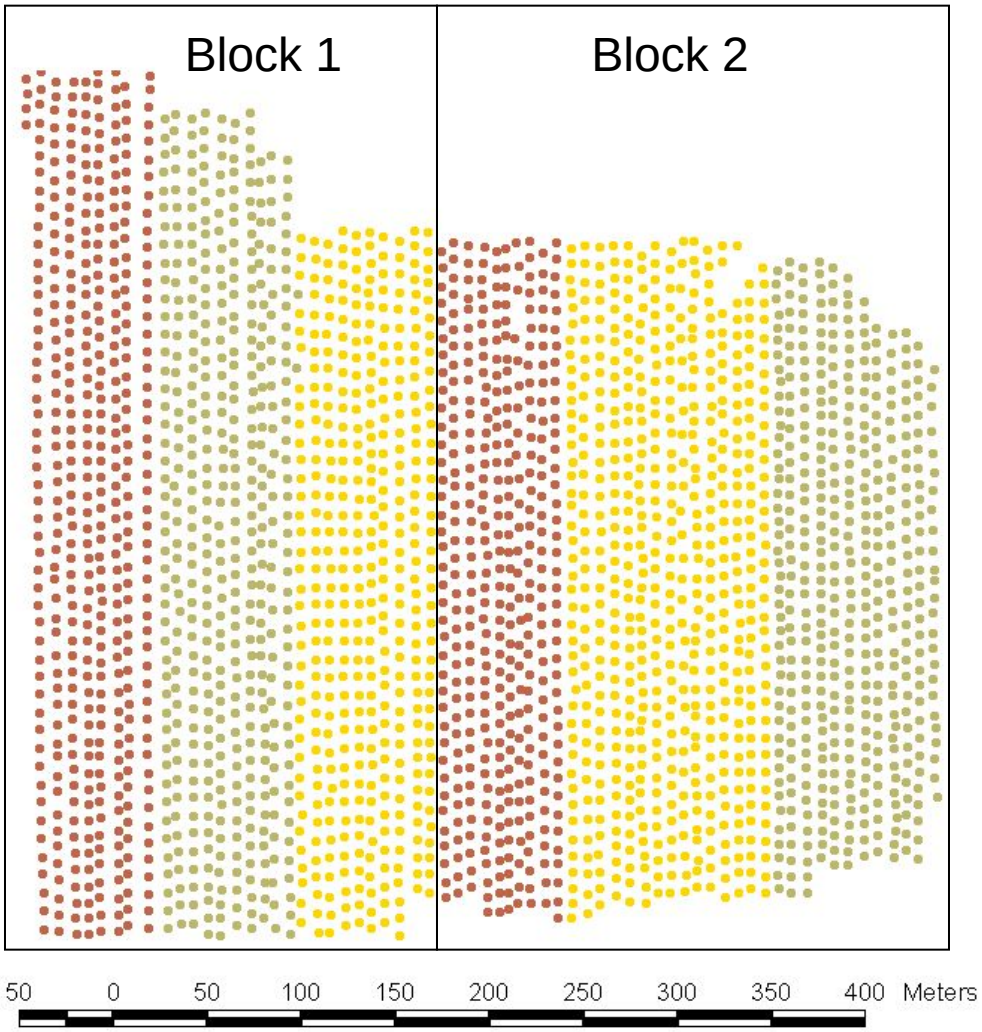
Martin-Luther-Universität Halle-Wittenberg, Institut für Agrar- und Ernährungswissenschaften

Arbeitsgruppe Biometrie und Agrarinformatik

Modellwahl

- Bei Nutzung von Versuchsanlagen ist im Allgemeinen das Auswertungsmodell festgelegt
- Auf Randomisation basierende Versuchsanlagen müssen auch Ausgangspunkt für OFE sein!
- Prüffaktoren und Designfaktoren damit festgelegt
- **Problemstellung:**
  Welcher zusätzlichen festen und zufälligen Faktoren (Störgrößen) sind zu beachten?
  Wie sollten sie in das Modell einbezogen werden?

Demonstrationsbeispiel (reduzierter Datensatz!)



Demonstrationsbeispiel

1. Auswertungsansatz:

   Nutzung der Parzellenmittelwerte und **Auswertung als Blockanlage**

PROC MIXED DATA=parz_mean_1;
  CLASS variante block;
  MODEL parz_mean = variante block;
  LSMEANS variante /PDIFF;
RUN;

Demonstrationsbeispiel

1. Auswertungsansatz:

   Nutzung der Parzellenmittelwerte und Auswertung als Blockanlage

Variante	LSMean	SE
k	94.27	1.0152
n	101.29	1.0152
s	102.17	1.0152

			Differenz	SE	DF	t Value	Pr > |t|
k	-	n	-7.0118	1.4357	2	-4.88	0.0395
k	-	s	-7.8931	1.4357	2	-5.50	0.0315
n	-	s	-0.8813	1.4357	2	-0.61	0.6019

Demonstrationsbeispiel

2. Auswertungsansatz:

   Nutzung der Einzelwerte

– Übergang zu diesem Ansatz hat Konsequenzen für die Wahl des Auswertungsmodells!
– führt „fast zwingend“ zu einem linearen gemischten Modell

Modellwahl

- Feste Effekte („Erwartungswertstruktur“)
  - „Kandidateneffekte“ im Beispiel:
    (Festlegung erfolgt aus sachlogischer Sicht!)
    - Mähdrescher (qualitativ) (Fahrspur nicht verfügbar!)
    - Leitfähigkeit (quantitativ)
    - räumlicher Trend (quantitativ)
- Zufällige Effekte und deren Struktur („Kovarianzstruktur“)

Modellwahl

- Das „wahre“ Modell ist bei praktischen Anwendungen stets unbekannt!
- Nur das „beste“ Modell aus den untersuchten Kandidatenmodellen ist identifizierbar
- „All models are wrong but some are helpful“ (George Box)

Modellwahl

- Modellwahl aus statistischer Sicht ist eine Abwägung
- Balance zwischen Unter- und Überanpassung gesucht
- Abhängigkeit vom Stichprobenumfang!

Modellwahl

- Zweischrittstrategie (pragmatischer Ansatz!)
  - Optimierung Erwartungswertstruktur (ML-Methode)
  - Optimierung Kovarianzstruktur (REML-Methode)
- Kontrolle der Erwartungswertstruktur durch Analyse der OLS-Residuen auf Verzerrungen (Rückversicherung 1)
- Kontrolle der Kovarianzstruktur durch Vergleich der Varianz der OLS-Residuen und geschätzter Varianz-Kovarianzfunktion (Rückversicherung 2)

Modellwahl

- Zweischrittstrategie (pragmatischer Ansatz!)
  - Optimierung Erwartungswertstruktur (ML-Methode)
  - Optimierung Kovarianzstruktur (REML-Methode)
- Vorteil dieser Strategie:
  - weniger Varianten zu untersuchen
  - OLS-Residuen erlauben eine Kontrolle auf Verzerrungen und Hinweis auf die vorliegende Kovarianzstruktur
- Nachteil dieser Strategie:
  - ggf. „Nachjustierung“ der einbezogenen festen Effekte im Auswertungsmodell erforderlich

Modellwahl der festen Effekte

– Optimierung Erwartungswertstruktur (Nutzung Likelihood der ML-Methode)

$$AIC_{ML} = -2\log L(\hat{\theta}) + 2\cdot(p_X + q)$$

$$AICC_{ML} = -2\log L(\hat{\theta}) + \frac{2n(p_X + q)}{n - (p_X + q) - 1}$$

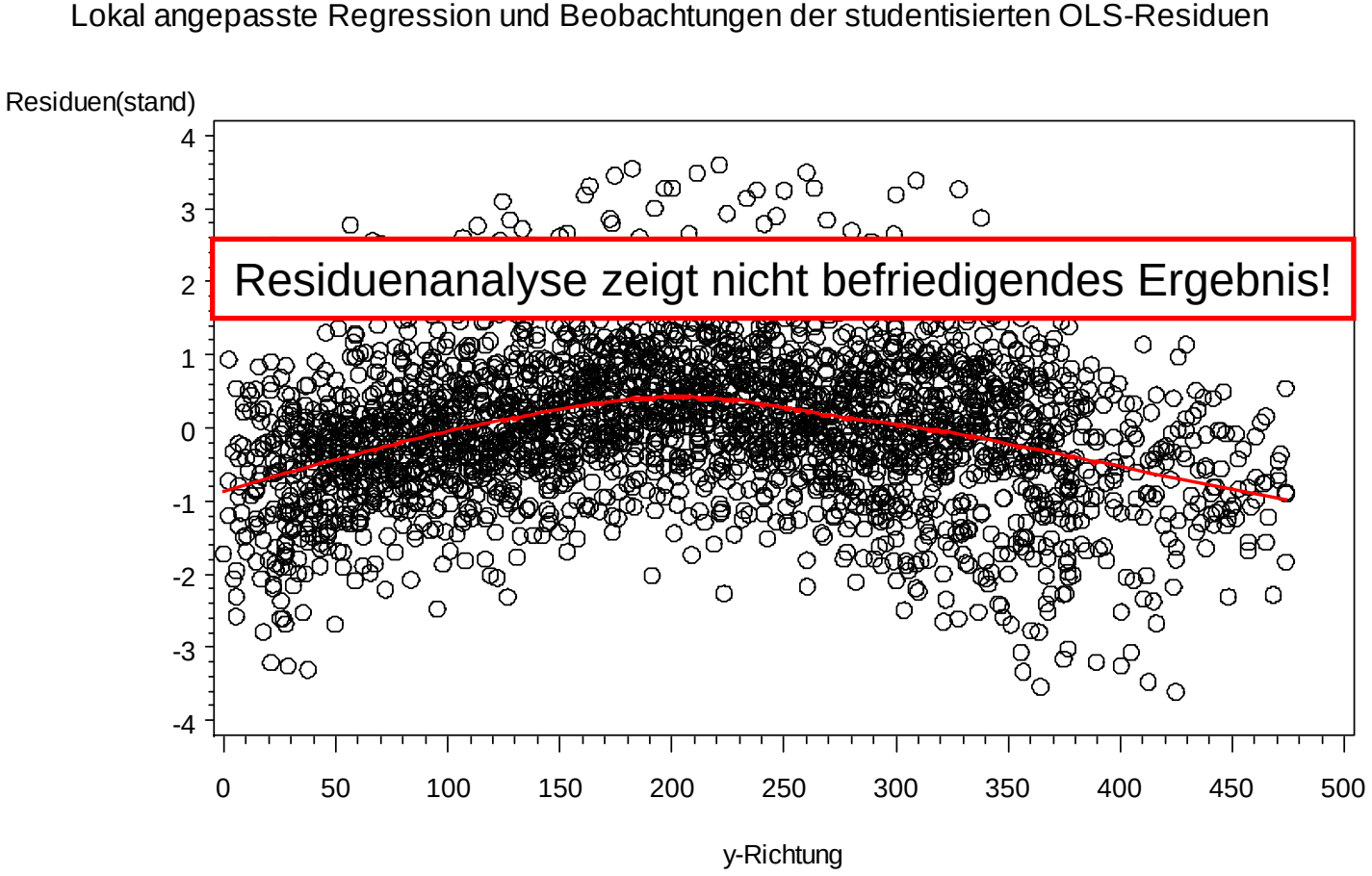
$$BIC_{ML} = -2\log L(\hat{\theta}) + (p_X + q)\log(n)\,.$$

Ergebnis für verschiedene „Kandidatenmodelle“ (Auszug)

Modell (feste Effekte)	**p**	**-2 LL**	**AIC**	**BIC**	$\hat{\sigma}_e^2$
variante block ($p=p_X+q = 4 + 1$)	5	17800.4	17810.4	17839.7	53.8
variante block md	7	17770.2	17784.2	17825.2	52.5
variante block md ec25	8	17661.5	17677.5	17724.5	50.3
variante block md ec25 ec25*ec25	9	17643.2	17661.2	17714.0	49.5
...					
variante block md ec25 ec25*ec25 y	10	17206.3	17226.3	17284.9	42.2
variante block md ec25 ec25*ec25 y x	11	17187.6	17209.6	17274.1	42.0

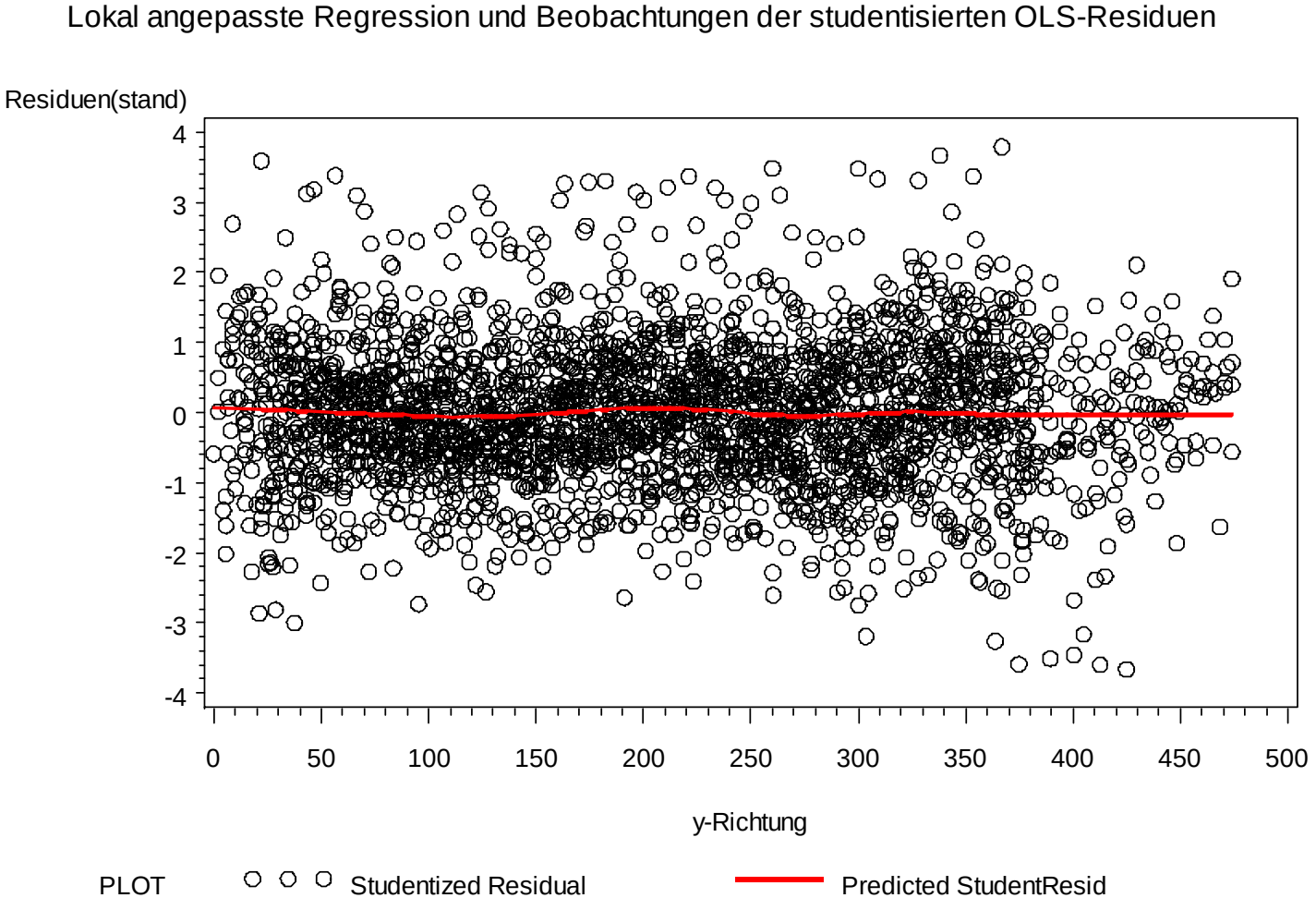
Zwischenergebnis für:

yield = variante md block ec25 ec25*ec25 x y



Modellerweiterung erforderlich:

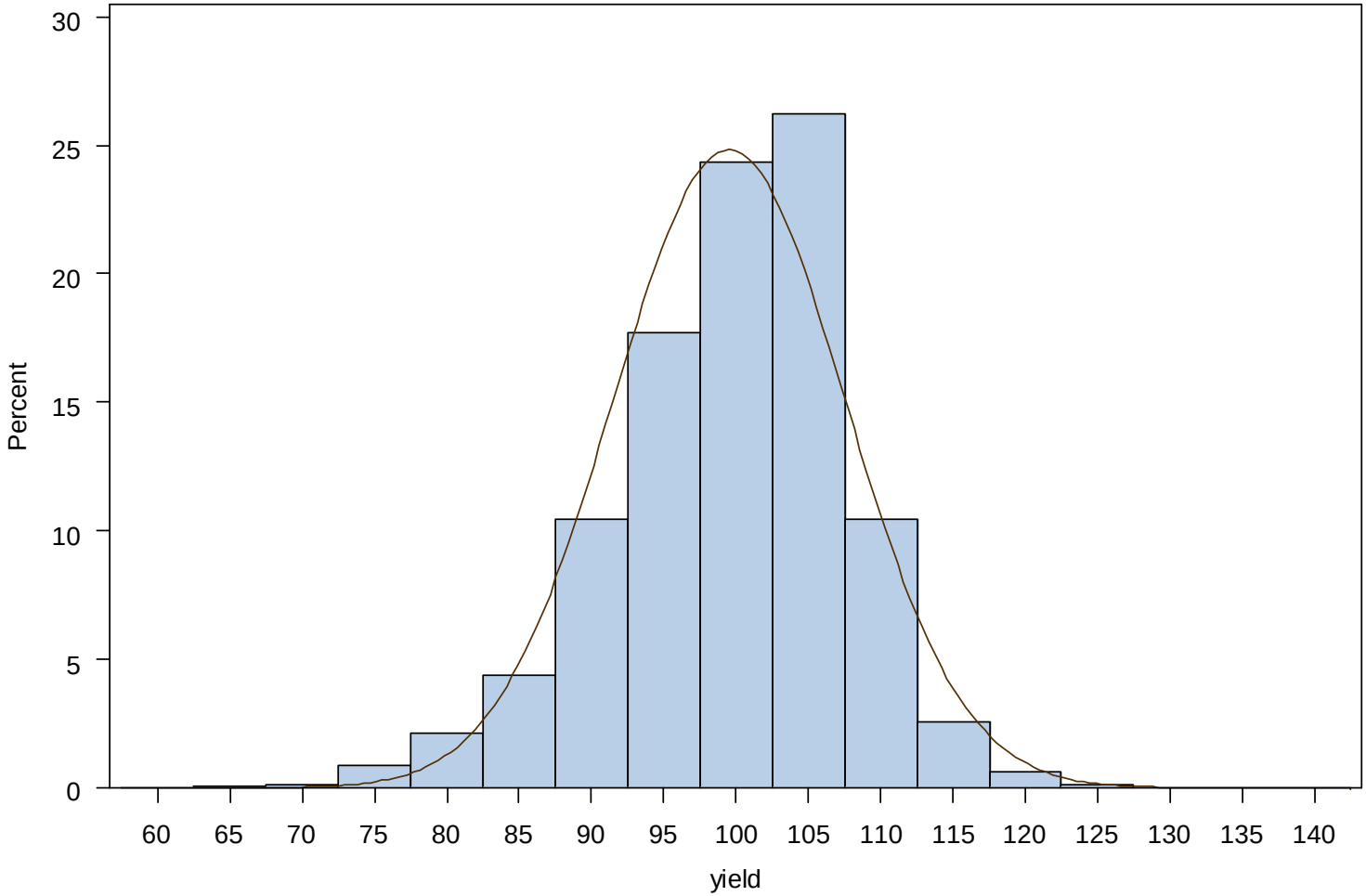
yield = variante md block ec25 ec25*ec25 x y y^2



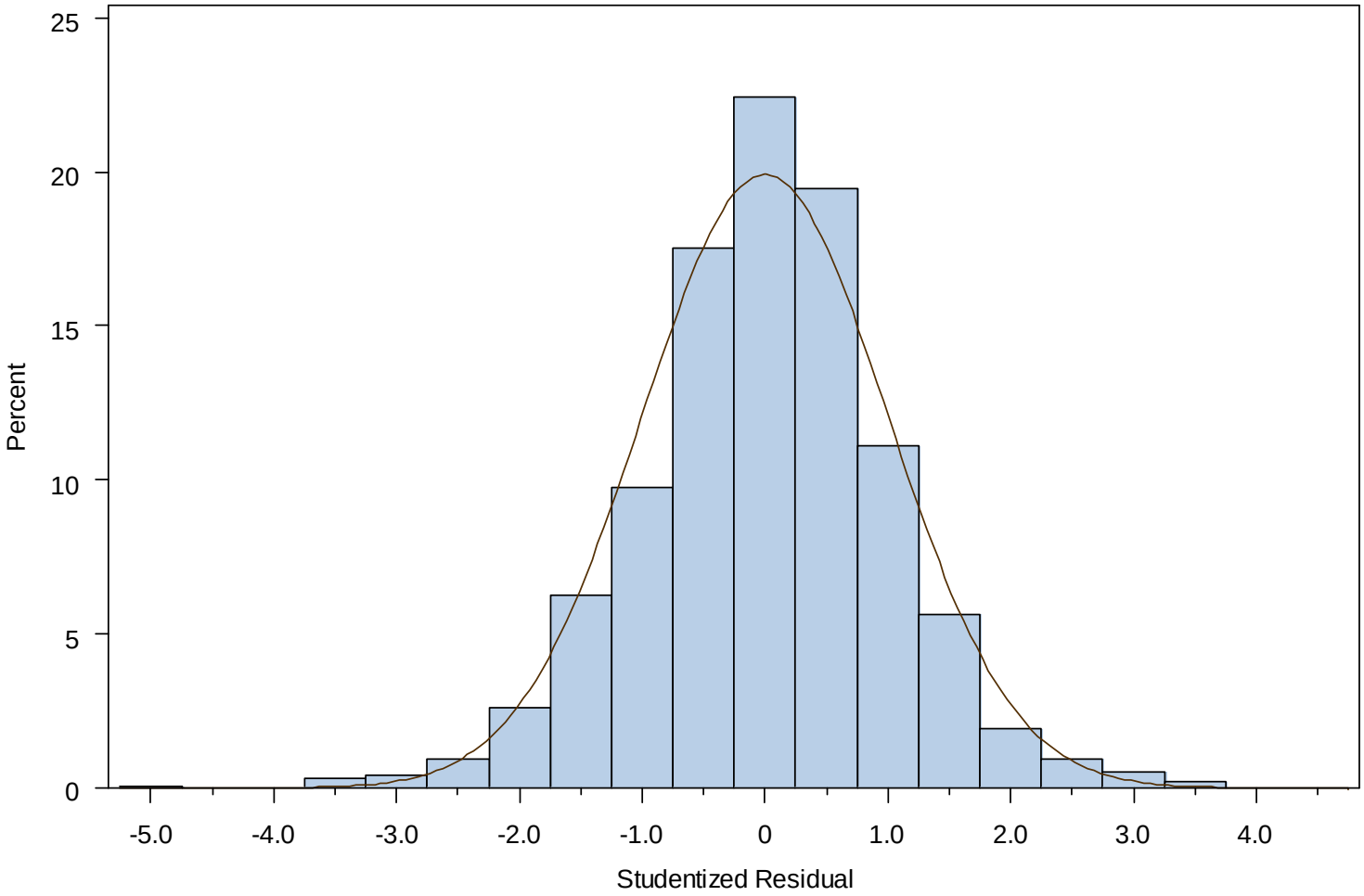
Ergebnis für feste Effekte
(nach Modellerweiterung!)

Modell (feste Effekte)	**p**	**-2 LL**	**AIC**	**BIC**	$\hat{\sigma}_e^2$
variante block ($p=p_X+q = 4 + 1$)	5	17800.4	17810.4	17839.7	53.8
variante block md	7	17770.2	17784.2	17825.2	52.5
variante block md ec25	8	17661.5	17677.5	17724.5	50.3
variante block md ec25 ec25*ec25	9	17643.2	17661.2	17714.0	49.5
...					
variante block md ec25 ec25*ec25 y	10	17206.3	17226.3	17284.9	42.2
variante block md ec25 ec25*ec25 y x	11	17187.6	17209.6	17274.1	42.0
variante block md ec25 ec25*ec25 y y*y x	12	16819.5	16843.5	16913.9	36.5

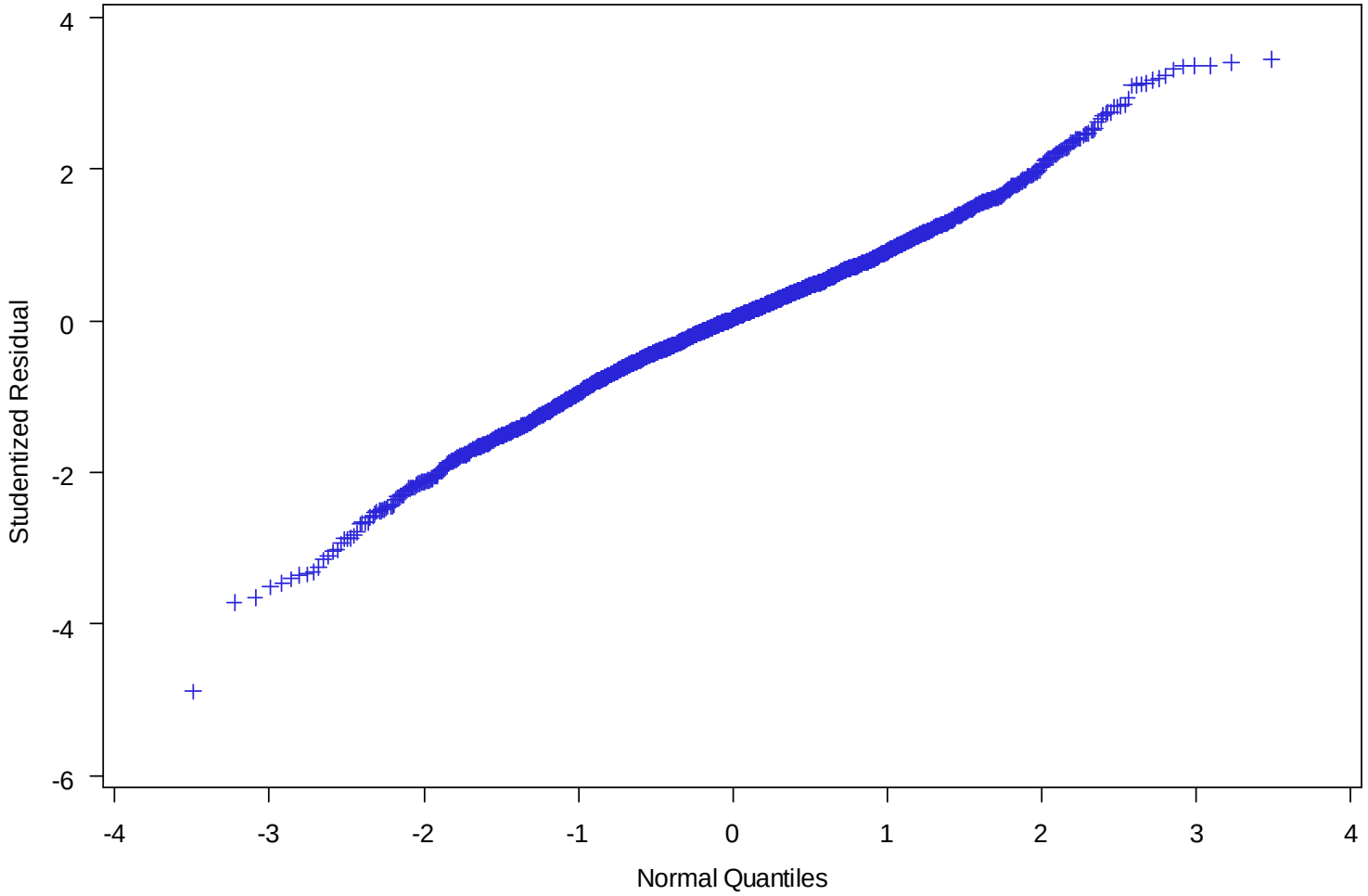
Verteilung der Beobachtungen



Verteilung der standardisierten Residuen für das Arbeitsmodell



QQ-Plot der Residuen für das Arbeitsmodell

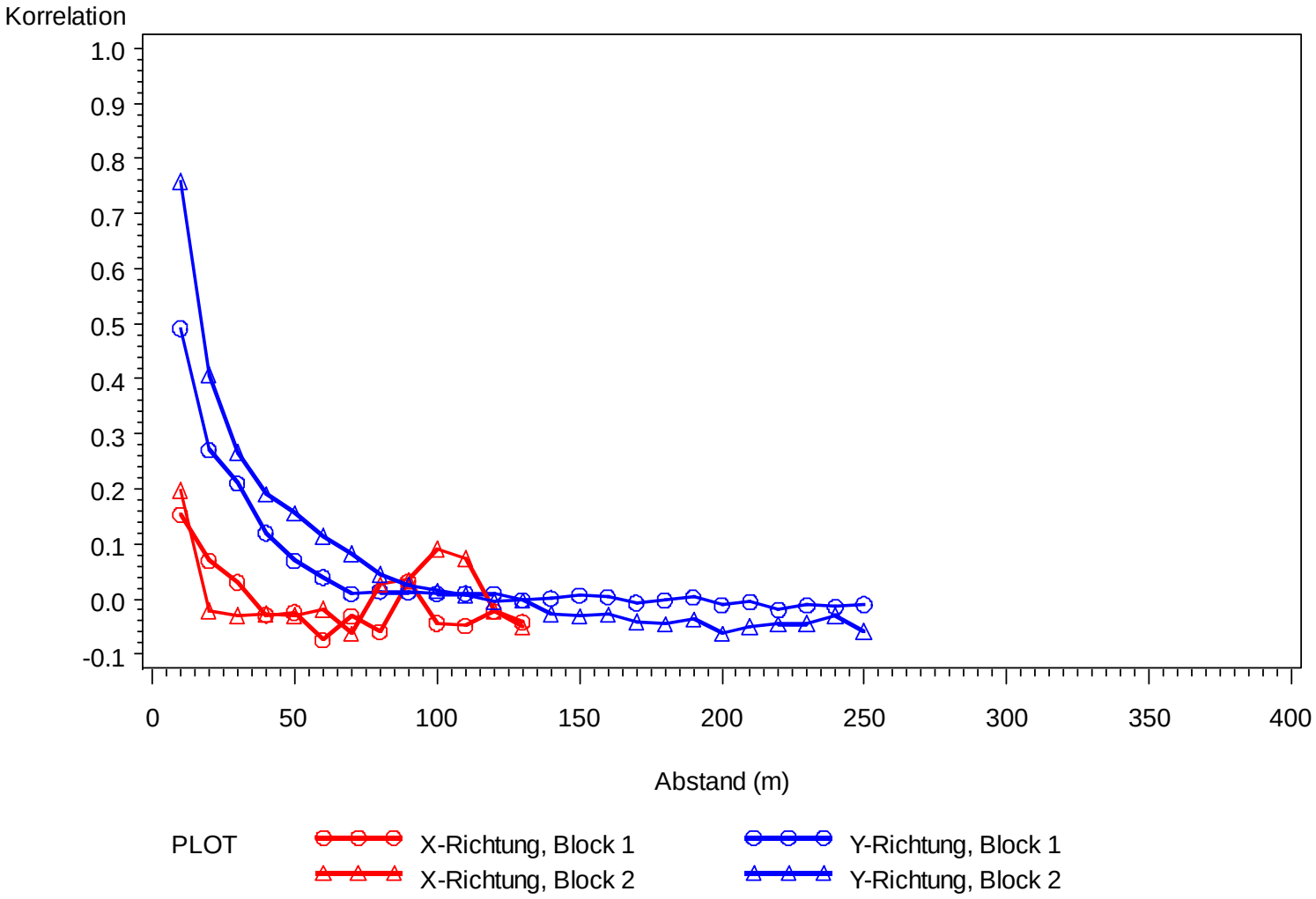


Zwischenbilanz

- „Arbeitsmodell“ für die festen Effekte festgelegt
- Nutzung dieses Modells als Auswertungsmodell würde bedeuten: Annahme unabhängiger Beobachtungen! (u.a.: FG = N-p_X =2614-11 !!!)
- <u>Völlig unrealistisch</u> im vorliegenden Sachzusammenhang!

Empirische räumliche Korrelation für Residuen des Arbeitsmodells

(je Block und Richtung)



Modellwahl zufällige Effekte und deren Kovarianzstruktur

– Optimierung Kovarianzstruktur
(Nutzung Restricted Likelihood der REML-Methode)

$$AIC_{REML} = -2\log_R L(\hat{\theta}) + 2 \cdot q$$

$$AICC_{REML} = -2\log_R L(\hat{\theta}) + \frac{2nq}{n-q-1}$$

$$BIC_{REML} = -2\log_R L + q\log(n)^*$$

*n: falls „random“ : Anzahl Stufen des ersten zufälligen Effekts
falls „repeated“: Anzahl Stufen des ersten Subjekts

Modellwahl zufällige Effekte

```
PROC MIXED METHOD=REML
  CLASS variante block md;
  MODEL yield = variante block md ec25 ec25*ec25 x y y*y;
  LSMEANS variante / PDIFF;
RUN;
```

Variante	LSMean	SE
k	96.0	0.3282
n	100.6	0.2064
s	101.7	0.2335

?

			Differenz	SE	DF	t Value	Pr > |t|
k	-	n	-4.55	0.4252	2603	-10.70	<.0001
k	-	s	-5.67	0.4659	2603	-12.17	<.0001
n	-	s	-1.12	0.2890	2603	-3.89	0.001

Ergebnis für zufällige Effekte

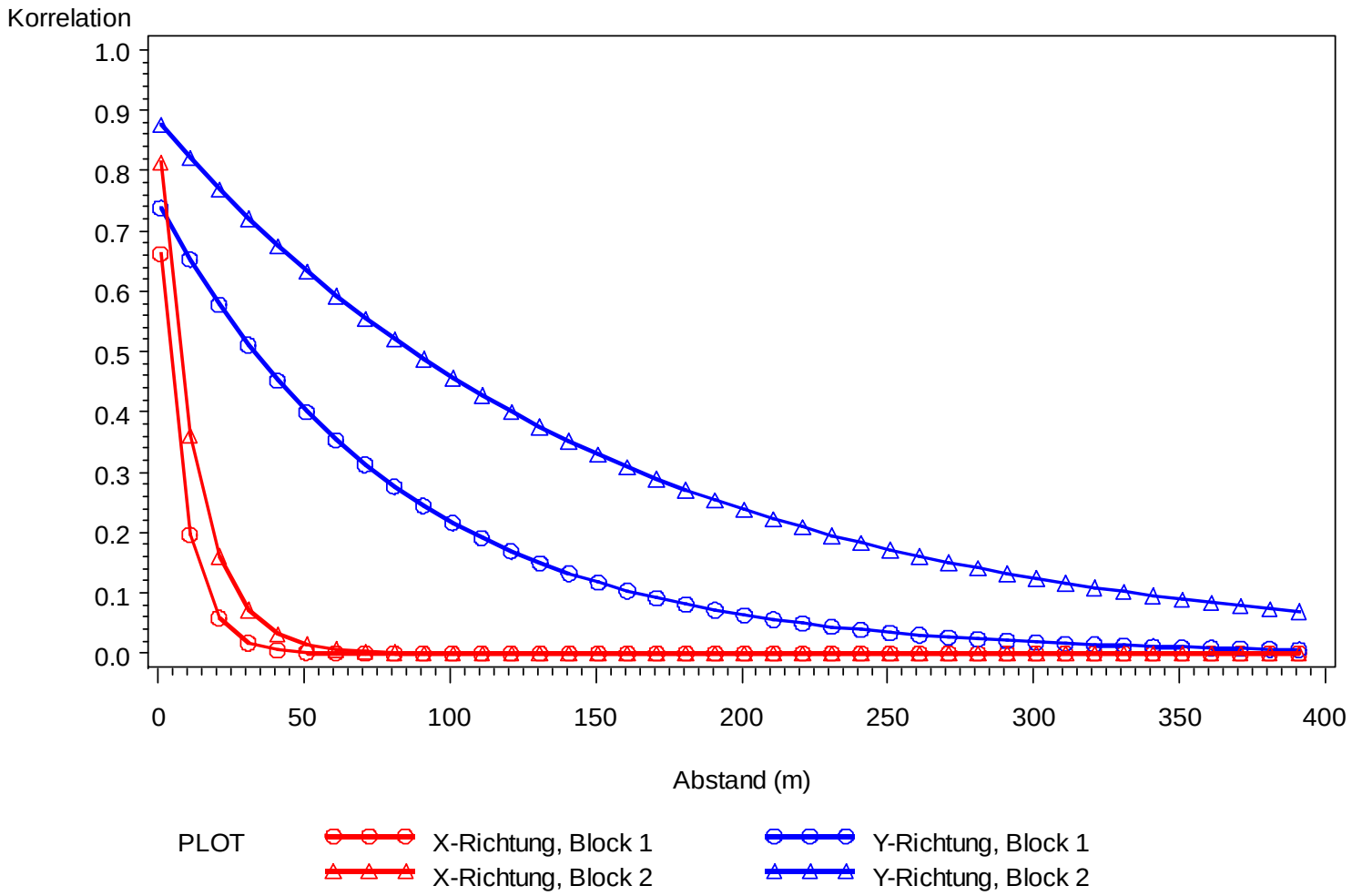
Modell für die festen Effekte: variante block md ec25 ec25*ec25 x y y*y	**q**	**-2 RLL**	**AIC**	**BIC**
Modell für die zufälligen Effekte:				
rest (**Bezugsmodell!)**	1	16888.0	16890.0	16895.9
RANDOM variante*block rest (identische Kovarianz innerhalb Parzelle!)	2	16816.7	16820.7	16818.1
...				
RANDOM variante*block (?) REPEATED / SUB=block TYPE=(POWA) (x y) LOCAL	4	15021.7	15029.7	15024.4
RANDOM variante*block (?) REPEATED / SUB=block GROUP=block TYPE=(POWA) (x y) LOCAL	7	14953.3	14967.3	14958.2
RANDOM variante*block REPEATED / SUB=variante*block TYPE=(POWA) (x y) LOCAL	Keine Lösung!			
RANDOM variante*block REPEATED / SUB=intercept TYPE=(POWA) (x y) LOCAL	5	14969.8	14979.8	14975.3

Auswertungsmodell

```
PROC MIXED METHOD=REML
  CLASS variante block md;
  MODEL yield = variante block md ec25 ec25*ec25  y y*y
                /DDFM=KR (FIRSTORDER);
  REPEATED / SUB=block GROUP=block TYPE=SP(POWA)  (x y) LOCAL;
  LSMEANS variante / PDIFF;
RUN;
```

Geschätzte räumliche Korrelation

(je Block und Richtung, Power-Funktion)



Geschätzte räumliche Korrelation

- Gute Widerspiegelung der Anisotropie und Unterschiede zwischen den Blocks
- Korrelation fällt gegenüber dem empirischen Variogramm / Kovariogramm langsamer ab

Ergebnisse Auswertungsmodell

Variante	LSMean	SE
k	95.3	1.3094
n	100.6	1.2053
s	100.1	1.2253

			Differenz	SE	DF	t Value	Pr > |t|
k	-	n	-4.78	1.5543	34.6	-3.07	0.004
k	-	s	-4.83	1.5106	33.2	-3.20	0.003
n	-	s	-0.06	1.4403	40.1	-0.04	0.969

Vergleich von Modell I und Auswertungsmodell

Variante	LSMean	SE
k	96.0	0.3282
n	100.6	0.2064
s	101.7	0.2335

			Differenz	SE	DF	t Value	Pr > |t|
k	-	n	-4.55	0.4252	2603	-10.70	<.0001
k	-	s	-5.67	0.4659	2603	-12.17	<.0001
n	-	s	-1.12	0.2890	2603	-3.89	0.001

Vergleich von Modell I und Auswertungsmodell

Variante	LSMean	SE
k	96.0	0.3282
n	100.6	0.2064
s	101.7	0.2335

			Differenz	SE	DF	t Value	Pr > |t|
k	-	n	-4.55	0.4252	2603	-10.70	<.0001
k	-	s	-5.67	0.4659	2603	-12.17	<.0001
n	-	s	-1.12	0.2890	2603	-3.89	0.001

Vergleich von Modell I und Auswertungsmodell

Variante	LSMean	SE
k	96.0	0.3282
n	100.6	0.2064
s	101.7	0.2335

			Differenz	SE	DF	t Value	Pr > |t|
k	-	n	-4.55	0.4252	2603	-10.70	<.0001
k	-	s	-5.67	0.4659	2603	-12.17	<.0001
n	-	s	-1.12	0.2890	2603	-3.89	0.001

Variante	LSMean	SE
k	95.29	1.3094
n	100.07	1.2053
s	100.13	1.2253

			Differenz	SE	DF	t Value	Pr > |t|
k	-	n	-4.78	1.5543	34.6	-3.07	0.004
k	-	s	-4.83	1.5106	33.2	-3.20	0.003
n	-	s	-0.06	1.4403	40.1	-0.04	0.969

Vergleich Auswertung auf Basis von Mittel- und Einzelwerten

Variante	LSMean	SE	DF	Intervallbreite(0.95)
k	94.28	1.0152	2	± 4.53
n	101.29	1.0152	2	± 4.53
s	102.20	1.0152	2	± 4.53

Variante	LSMean	SE	DF	Intervallbreite(0.95)
k	95.29	1.3094	40.6	± 2.64
n	100.07	1.2053	38.9	± 2.44
s	100.13	1.2253	40.8	± 2.47

Vergleich Auswertung auf Basis von Mittel- und Einzelwerten

			Differenz	SE	DF	GD(0.05)|
k	-	n	-7.01	1.4357	2	6.18
k	-	s	-7.89	1.4357	2	6.18
n	-	s	-0.88	1.4357	2	6.18

			Differenz	SE	DF	GD(0.05)
k	-	n	-4.78	1.5543	34.6	3.16
k	-	s	-4.83	1.5106	33.2	3.07
n	-	s	-0.06	1.4403	40.1	2.91

Schlussbemerkungen (1)

- Erfolg der Modellwahl ist nicht an „Signifikanz“ oder „Nichtsignifikanz“ zu messen - sonst wäre immer Modell I zu bevorzugen !
- Erfolg der Modellwahl ist an der Identifizierung eines den Daten adäquaten Auswertungsmodells zu messen – das kann mit einer Erhöhung der Standardfehler, Abnahme der Freiheitsgrade verbunden sein und kann zu Verlust von Signifikanz führen !

Schlussbemerkungen (2)

- Sorgfältige Modellwahl für OFE zwingend
- Analytische Kriterien sind wertvolles Hilfsmittel – nicht mehr!
- Rückversicherung durch Analyse der Residuen und der geschätzten Varianz-Kovarianzfunktion unverzichtbar!
- Bei der Auswertung von OFE verbietet sich jeder Automatismus !