Forschungsmethodik: Interventions- und Evaluationsforschung theoretisch begründen

Vorlesung Biostatistik, WS 2026/2027

Dr. Marcus Wagner, IMISE

Vorlesung VI am 17.11.2026

Inhaltsverzeichnis

9. Allgemeine lineare Modelle

9.6. Auswahl möglicher Kovariaten

10. Zweifaktorielle ANOVA mit Meßwiederholung

10.1. Datenstruktur

10.2. Fragestellungen

10.3. Ablauf, Zerlegung der Varianz

10.4. Globaltests für Interaktion und Haupteffekte

10.5. Kontraste und Querschnittsmodelle

Wiederholung: GLM für Stichproben

• **Lineare Modellgleichung:** $Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + \ ... \ + c_p \cdot K_p + \varepsilon$.

• **Variablen:** *Werte aus einer Stichprobe* $\mathcal{S} \subset \mathcal{P}$ *vom Umfang* $N > (p+1)$

— *"abhängige Variable"* Z *(Zielvariable); stets metrisch*

— *"unabhängige Variablen"* K_1, K_2, ... , K_p *(Kovariaten); metrisch, binär oder kategorial (dann Umkodierung in Dummies)*

— *Interaktionsterme der Gestalt* $c_{j_1,j_2} \cdot K_{j_1} \cdot K_{j_2}$ *(bewirken meist eine Modellkorrektur für bestimmte Subgruppen)*

— *Residuen* ε*, paarweise unabhängig und normalverteilt mit* $\mu = 0$ *und* $\sigma > 0$

• **Globaler Hypothesentest:** *F-Test für* $\widehat{T} = \widehat{\sigma}^2_{\text{erklärt}} \,/\, \widehat{\sigma}^2_{\text{residual}}$ *gegen die Nullhypothese* $c_1 = c_2 = \ ... \ = c_p = 0$ *(kein linearer Gesamtzusammenhang)*

• **Post-hoc-Tests für die Koeffizienten** $\widehat{c}_j$**:** *Zweiseitige* t*-Tests für* $\widehat{T}_j = \widehat{c}_j / \widehat{\sigma}_j$ *gegen die Nullhypothese* $c_j = 0$ *(kein Einfluss von* K_j*), dazu Konfidenzintervalle*

• **Freiheit bei der Modellierung:** *Wie kann man entscheiden, welche Variablen überhaupt ins Modell einbezogen werden?*

Wiederholung: Bewertungskriterien fürs Gesamtmodell

• **Situation:** *Für Variablen Z und K_1, K_2, ... , K_p liegt eine Stichprobe $\mathcal{S}$ vom Umfang N vor. Wir betrachten das GLM $Z \sim \widehat{c}_0 + \sum_{j=1}^{p} \widehat{c}_j \cdot K_j$.*

• **Drei Kriterien:**

1) Signifikanz beim globalen Hypothesentest — absolutes Kriterium; fällt der Globaltest nicht signifikant aus, ist das Modell insgesamt zu verwerfen

2) Bestimmtheitsmaß $\widehat{R}^2$ — absolutes Kriterium; je näher an Eins, um so besser (mißt den vom Modell "erklärten" Anteil der Gesamt-Varianz):

$$\widehat{R}^2 \;=\; \Big(\sum_{i=1}^{N} (Z(i) - \widehat{\mu}_Z)^2 \;-\; \sum_{i=1}^{N} \widehat{\varepsilon}(i)^2 \Big) \Big/ \sum_{i=1}^{N} (Z(i) - \widehat{\mu}_Z)^2 \,.$$

3) Akaike-Informationskriterium AIC — relatives Kriterium (Vergleich mehrerer Modelle zur gleichen Zielgröße); je kleiner, um so besser (mißt den Informationsverlust):

$$AIC \;=\; \ln \Big(\sum_{i=1}^{N} \widehat{\varepsilon}(i)^2 \Big/ N \Big) + 2 \cdot (p+1) \,/\, N \,.$$

Underfit und Overfit

• **Welches ist das bestmögliche Modell zu den vorhandenen Daten?**
— Ein Modell kann immer nur einige wesentliche Züge der objektiven Realität widerspiegeln, aber niemals die gesamte Wirklichkeit!
— Bei der Modellbildung muß man abwägen zwischen Erklärungsgehalt (sollte groß sein) und Komplexität des Modells (sollte gering sein).
— Das Modell sollte nicht nur die vorhandenen Daten erklären, sondern auch Vorhersagen ermöglichen.

• **Underfit:** *Die ins Modell aufgenommenen Kovariaten können die Zielvariable nur unzureichend erklären;* $\widehat{R}^2$ *bleibt zu niedrig.*

• **Overfit:** *Nimmt man zu viele Kovariaten ins Modell auf, so wird die Zielvariable zwar innerhalb der Stichprobe* $\mathcal{S}$ *sehr gut erklärt (hoher* $\widehat{R}^2$*-Wert), aber für Daten außerhalb von* $\mathcal{S}$ *wird das Modell unbrauchbar zur Vorhersage.*

• **Faustregel:** *"Gesundes Verhältnis" zwischen Umfang der Stichprobe und Anzahl der Kovariaten: pro Kovariate sind mindestens 10 Messungen vorhanden*

Vorwärtsselektion

- **Gegeben:** *eine Zielvariable Z und mögliche Kovariaten K_1, K_2, K_3, ... , K_p.*
- **Schritt 1:** *Nimm zu $Z \sim a$ eine Kovariate aus K_1, K_2, ... , K_p hinzu:*

$$Z \sim a + b{\cdot}K_1\,,\;\; Z \sim a + b{\cdot}K_2\,,\;\; Z \sim a + b{\cdot}K_3\,,\;\; \ldots,\;\; Z \sim a + b{\cdot}K_p\,,.$$

und wähle unter den p Modellen jenes mit dem niedrigsten AIC-Wert (bzw. dem höchsten $\widehat{R}^2$-Wert) aus. O.E.d.A. werde das Modell mit K_1 ausgewählt.

- **Schritt 2:** *Nimm zu $Z \sim a + b{\cdot}K_1$ eine Kovariate aus K_2, K_3, ... , K_p hinzu:*

$$Z \sim a + b{\cdot}K_1 + c{\cdot}K_2\,,\;\; Z \sim a + b{\cdot}K_1 + c{\cdot}K_3\,,\;\; \ldots,\;\; Z \sim a + b{\cdot}K_1 + c{\cdot}K_p\,,$$

und wähle unter den $(p-1)$ Modellen wiederum jenes mit dem niedrigsten AIC-Wert (bzw. dem höchsten $\widehat{R}^2$-Wert) aus. O.E.d.A. werde das Modell mit K_2 ausgewählt.

- **Schritt 3:** *Nimm zu $Z \sim a + b \cdot K_1 + c \cdot K_2$ eine Kovariate aus K_3,..., K_p hinzu: usw.*
- **Abbruch:** *Wenn sich AIC für keine der neu hinzugenommenen Kovariaten verkleinert (bzw. $\widehat{R}^2$ sich nicht vergrößert), wird das bisherige Modell akzeptiert.*

In allen Schritten: Modelle ohne signifikanten Globaltest entfallen; wenn eine Kovariate ein mehrstufiger Faktor ist, müssen alle Dummies gleichzeitig einbezogen werden.

Vorwärtsselektion: Beispiel

• **Beispiel 6.1.:** *Zielvariable: Wegstrecke beim 6-Minuten-Gehtest zum Follow-up; fünf mögliche Kovariaten:* K_1 *— Wegstrecke zu Baseline,* K_2 *— BMI,* K_3 *— Alter,* K_4 *— Studienarm und* K_5, K_6 *— NYHA-Klasse II bzw. III vs. I (Daten aus Beispiel 5.5.).*

K1	K2	K3	K4	K5, K6	Anzahl Kovariaten	Signifikanz beim Globaltest	$\tilde{R}^2$	AIC	Auswahl nach AIC
x					1	ja	0.6743	7.8215	Schritt 1
	x				1	nein	0.0230	8.9202	entfällt
		x			1	ja	0.0993	8.8388	
			x		1	nein	0.0021	8.9413	entfällt
				xx	2	ja	0.1682	8.7871	
x	x				2	ja	0.6815	7.8271	
x		x			2	ja	0.6744	7.8493	
x			x		2	ja	0.6744	7.8492	
x				xx	3	ja	0.7105	7.7593	Schritt 2
x	x			xx	4	ja	0.7266	7.7298	Schritt 3
x		x		xx	4	ja	0.7110	7.7855	
x			x	xx	4	ja	0.7115	7.7838	
x	x	x		xx	5	ja	0.7284	7.7510	Abbruch
x	x		x	xx	5	ja	0.7267	7.7574	

• **Ergebnis:** $Z = 228.33 + 0.75 \cdot K_1 - 2.26 \cdot K_2 - 50.93 \cdot K_5 - 32.85 \cdot K_6$

Rückwärtsselektion

- **Gegeben:** *eine Zielvariable Z und mögliche Kovariaten K_1, K_2, K_3, ... , K_p.*
- **Schritt 1:** *Entferne aus dem "vollen" Modell $Z \sim c_0 + \sum_{j=1}^{p} c_j \cdot K_j$ jeweils eine Kovariate und wähle unter den entstehenden Modellen jenes mit dem kleinsten AIC-Wert (bzw. dem höchsten $\widehat{R}^2$-Wert) aus. O.E.d.A. werde K_p entfernt.*
- **Schritt 2:** *Entferne aus dem Modell $Z \sim c_0 + \sum_{j=1}^{p-1} c_j \cdot K_j$ jeweils eine weitere Kovariate und wähle unter den entstehenden Modellen jenes mit dem kleinsten AIC-Wert (bzw. dem höchsten $\widehat{R}^2$-Wert) aus. O.E.d.A. werde K_{p-1} entfernt.*
- **Schritt 3:** *Entferne aus dem Modell $Z \sim c_0 + \sum_{j=1}^{p-2} c_j \cdot K_j$ jeweils eine weitere Kovariate usw.*
- **Abbruch:** *Wenn man keine Kovariate findet, bei deren Entfernung sich AIC verringert (bzw. $\widehat{R}^2$ sich vergrößert), wird das bisherige Modell akzeptiert.*

In allen Schritten: Modelle ohne signifikanten Globaltest entfallen; wenn eine Kovariate ein mehrstufiger Faktor ist, müssen alle Dummies gleichzeitig entfernt werden.

Rückwärtsselektion: Beispiel

• **Nochmals Beispiel 6.1.:** *Zielvariable: Wegstrecke beim Gehtest zum Follow-up; fünf mögliche Kovariaten: K_1 — Wegstrecke zu Baseline, K_2 — BMI, K_3 — Alter, K_4 — Studienarm und K_5, K_6 — NYHA-Klasse II bzw. III vs. I (Daten aus Beispiel 5.5.).*

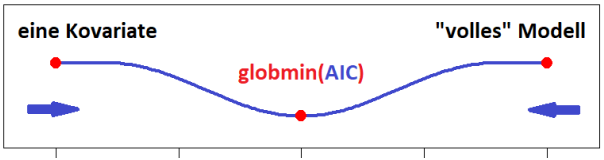
K1	K2	K3	K4	K5, K6	Anzahl Kovariaten	Signifikanz beim Globaltest	R̂²	AIC	Auswahl nach AIC
x	x	x	x	xx	6	ja	0.7285	7.7784	Start
x	x	x	x		4	ja	0.6839	7.8752	
x	x	x		xx	5	ja	0.7284	7.7510	Schritt 1
x	x		x	xx	5	ja	0.7267	7.7574	
x		x	x	xx	5	ja	0.7117	7.8107	
	x	x	x	xx	5	ja	0.3114	8.6449	
x	x	x			3	ja	0.6838	7.8475	
x	x			xx	4	ja	0.7266	7.7298	Schritt 2
x		x		xx	4	ja	0.7110	7.7855	
	x	x		xx	4	ja	0.3112	8.6540	
x	x				2	ja	0.6815	7.8271	Abbruch
x				xx	3	ja	0.7105	7.7593	
	x			xx	3	ja	0.2107	8.7624	

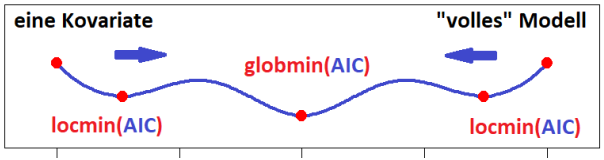
• **Ergebnis:** $Z = 228.33 + 0.75 \cdot K_1 - 2.26 \cdot K_2 - 50.93 \cdot K_5 - 32.85 \cdot K_6$ *(dasselbe)*

• *Auswahl nach $\widehat{R}^2$ hätte stattdessen das "volle" Modell mit allen Kovariaten ergeben!*

Gleichzeitige Betrachtung aller möglichen Modelle

• **Problem:** *Bei der Vorwärts- oder Rückwärtsselektion mit p möglichen Kovariaten beurteilt man nur eine Auswahl von höchstens $p \cdot (p-1)/2$ der insgesamt möglichen $(2^p - 1)$ Modelle. Es gibt keine Garantie, daß man dabei das Modell mit dem insgesamt niedrigsten AIC-Wert (bzw. dem insgesamt besten $\widehat{R}^2$-Wert) findet.*





• **Lösung ("Search for the best subset"):** *Ist ausreichend Zeit bzw. Rechenkapazität vorhanden, so können alle möglichen Modelle berechnet werden. Anschließend identifiziert man das Modell mit dem niedrigsten AIC-Wert (bzw. dem höchsten $\widehat{R}^2$-Wert).*

• **Nochmals Beispiel 6.1.:** *Zielvariable: Wegstrecke beim Gehtest zum Follow-up; fünf mögliche Kovariaten: K_1 — Wegstrecke zu Baseline, K_2 — BMI, K_3 — Alter, K_4 — Studienarm und K_5, K_6 — NYHA-Klasse II bzw. III vs. I (Daten aus Beispiel 5.5.).*

Beispiel: “Search for the best subset”

Nr.	K1	K2	K3	K4	K5/6	$\hat{R}^2$	AIC
00						n/a	n/a
01	x					0.6743	7.8215
02		x				0.0230	8.9202
03			x			0.0993	8.8388
04				x		0.0021	8.9413
05					xx	0.1682	8.7871
06	x	x				0.6815	7.8271
07	x		x			0.6744	7.8493
08	x			x		0.6744	7.8492
09	x				xx	0.7105	7.7593
10		x	x			0.1935	8.7562
11		x		x		0.0292	8.9416
12		x			xx	0.2107	8.7624
13			x	x		0.0995	8.8664
14			x		xx	0.2111	8.7619
15				x	xx	0.1683	8.8147

Nr.	K1	K2	K3	K4	K5/6	$\hat{R}^2$	AIC
16	x	x	x			0.6838	7.8475
17	x	x		x		0.6816	7.8545
18	x	x			xx	0.7266	7.7298
19	x		x	x		0.6744	7.8769
20	x		x		xx	0.7110	7.7855
21	x			x	xx	0.7115	7.7838
22		x	x	x		0.1944	8.7829
23		x	x		xx	0.3112	8.6540
24		x		x	xx	0.2134	8.7837
25			x	x	xx	0.2117	8.7889
26	x	x	x	x		0.6839	7.8752
27	x	x	x		xx	0.7284	7.7510
28	x	x		x	xx	0.7267	7.7574
29	x		x	x	xx	0.7117	7.8107
30		x	x	x	xx	0.3114	8.6449
31	x	x	x	x	xx	0.7285	7.7784

- **Beobachtungen:**

— Die Indikatoren AIC und $\widehat{R}^2$ laufen nicht immer parallel: z.B. # 06 vs. # 16.

— Die vier besten Modelle im Sinne des AIC sind # 18, # 27, # 28 und # 09.

Das “volle” Modell # 31 kommt erst an fünfter Stelle.

— Wegen $N = 72$ sind höchstens 7 Kovariaten sinnvoll.

Beispiel: “Search for the best subset” (II)

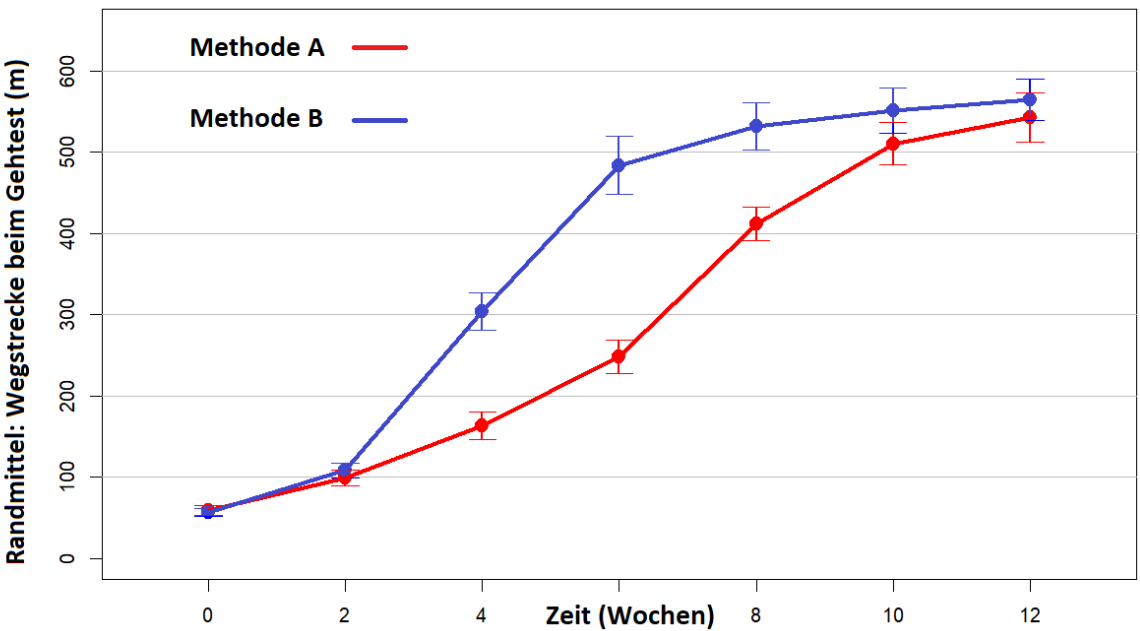
Größe	geschätzter Koeffizient	empirische Standardabw.	Prüfgröße (F- bzw. t-Wert)	p-Wert	95-%-Konfidenzintervall
Gesamtmodell # 18			44.53	< 0.0005	
Intercept	228.33	50.99	4.48	< 0.0005	[126.56 , 330.11]
Wegstrecke BAS	0.75	0.07	11.24	< 0.0005	[0.61 , 0.88]
BMI	-2.26	1.14	-1.99	0.0510	[- 4.54 , 0.01]
NYHA-Klasse: II	-50.93	15.45	-3.30	0.0016	[- 81.77 , -20.09]
NYHA-Klasse: III	-32.85	17.28	-1.90	0.0616	[-67.34 , 1.64]

Größe	geschätzter Koeffizient	empirische Standardabw.	Prüfgröße (F- bzw. t-Wert)	p-Wert	95-%-Konfidenzintervall
Gesamtmodell # 09			55.64	< 0.0005	
Intercept	150.50	33.35	4.51	< 0.0005	[83.95 , 217.05]
Wegstrecke BAS	0.76	0.07	11.29	< 0.0005	[0.62 , 0.90]
NYHA-Klasse: II	-44.28	15.41	-2.87	0.0054	[-75.03 , -13.53]
NYHA-Klasse: III	-26.86	17.38	-1.55	0.1268	[- 61.55 , 7.82]

- *Man könnte sich im Vergleich auch für Modell # 09 entscheiden!*

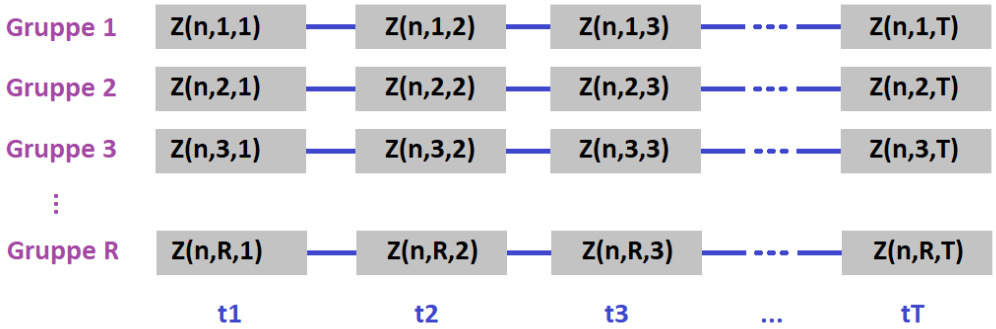
Modelle mit Meßwiederholung

• **Beispiel 6.2.:** *6-Minuten-Gehtests zu Baseline sowie nach 2, 4, 6, 8, 10 und 12 Wochen aktiver Rehabilitation. Wir vergleichen für Gruppen von je 50 Teilnehmern zwei Trainingsmethoden A und B.*



Zweifaktorielle ANOVA: Datenstruktur

- **Ziel:** *Modell soll Abhängigkeit der Daten berücksichtigen, die für jede Einzelperson ("Subjekt") zu den verschiedenen Zeitpunkten t_1, t_2, t_3, t_4, ... erhoben wurden.*
- $Z(n, r, t)$ — *Messung von Z für Teilnehmer n in Gruppe r zum Zeitpunkt t*

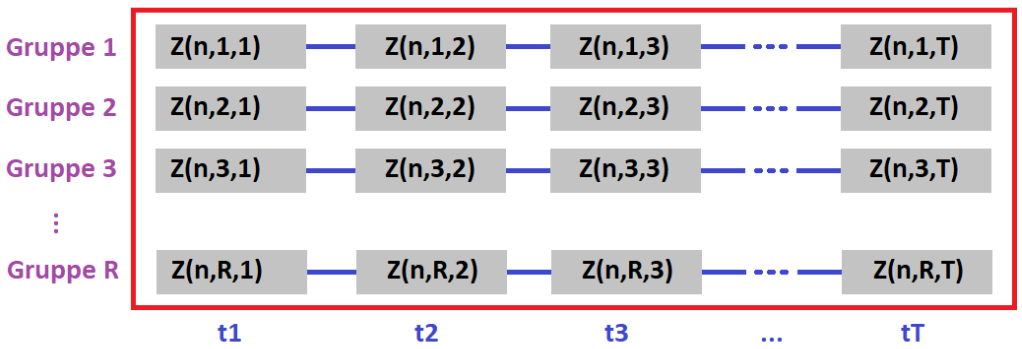


- **Faktoren:** *Man unterscheidet Zwischensubjektfaktoren (Studienarm, Geschlecht, ...) (Index r) und Innersubjektfaktoren (meist Zeitpunkt der Messung) (Index t).*

Ein Modell — viele Fragen

Zweifaktorielle ANOVA: man stellt verschiedene Fragen an den gleichen Datensatz!

- **Interaktionseffekt als wichtigste Frage:** *Unterscheiden sich die Gruppen in der Veränderung der Zielvariablen über die Zeit?*
- **Haupteffekt Zeit:** *Verändert sich die Zielvariable im Verlauf der Zeit im Mittel über alle Gruppen?*
- **Haupteffekt Gruppe:** *Unterscheidet sich die Zielvariable zwischen den Gruppen im Mittel über alle Zeitpunkte?*

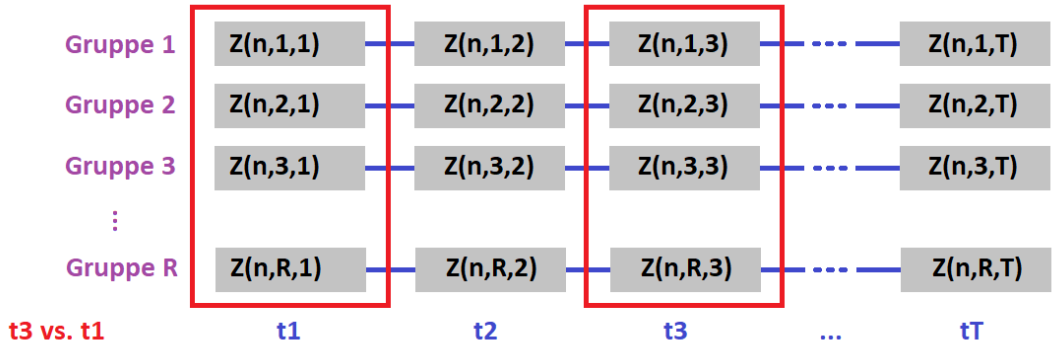


Ein Modell — viele Fragen: Kontraste

Einfache Kontraste: *Wir analysieren nur einen Teil des Datensatzes und vergleichen jeweils die Daten zu den Zeitpunkten t_2, t_3, ... , t_T mit den Daten vom Zeitpunkt t_1.*

• **Einfacher Innersubjektkontrast Zeit : Gruppe:** *Unterscheiden sich die Gruppen in der Veränderung der Zielvariablen im Vergleich von t_2 vs. t_1, t_3 vs. t_1, ... , t_T vs. t_1?*

• **Einfacher Innersubjektkontrast Zeit:** *Verändert sich die Zielvariable im Mittel über alle Gruppen im Vergleich von t_2 vs. t_1, t_3 vs. t_1, ... , t_T vs. t_1?*

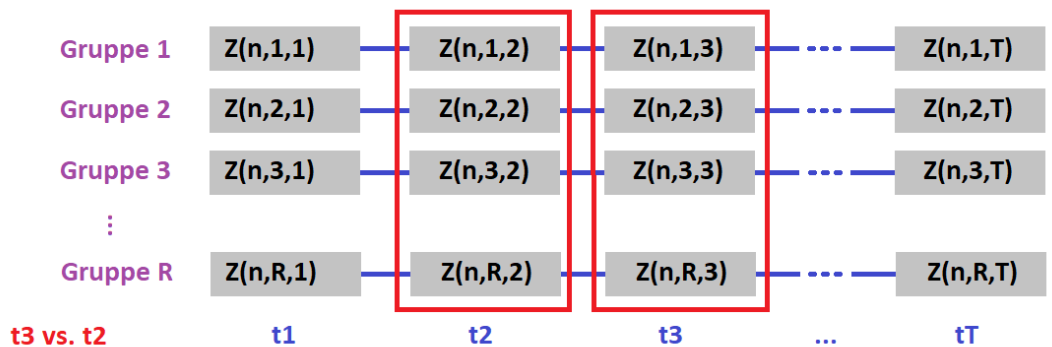


Ein Modell — viele Fragen: Kontraste (II)

Wiederholte Kontraste: *Wir analysieren nur einen Teil des Datensatzes und vergleichen die Daten zu den Zeitpunkten t_2, t_3, ... , t_T mit den Daten zum jeweils unmittelbar vorhergehenden Zeitpunkt.*

- **Wiederholter Innersubjektkontrast Zeit : Gruppe:** *Unterscheiden sich die Gruppen in der Veränderung der Zielvariablen im Vergleich von t_2 vs. t_1, t_3 vs. t_2, ... , t_T vs. t_{T-1}?*

- **Wiederholter Innersubjektkontrast Zeit:** *Verändert sich die Zielvariable im Mittel über alle Gruppen im Vergleich von t_2 vs. t_1, t_3 vs. t_2, ... , t_T vs. t_{T-1}?*

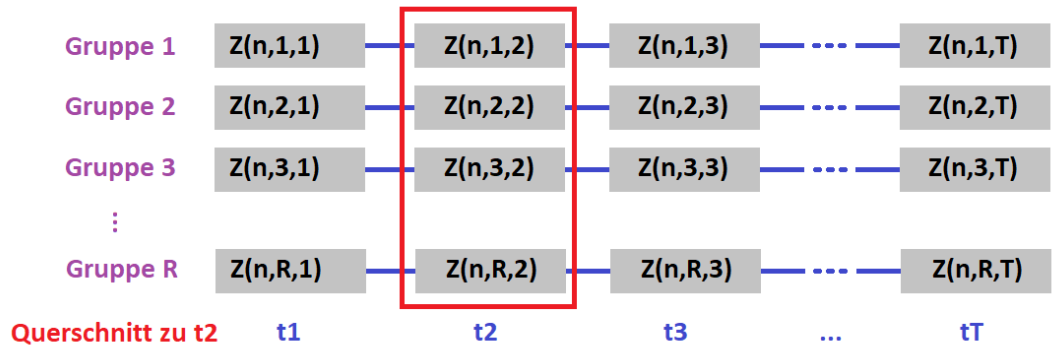


Ein Modell — viele Fragen: Querschnittsmodelle

- **Lineares Querschnittsmodell:** *Wir analysieren nur einen Teil des Datensatzes, nämlich die Daten zu einem einzigen Meßzeitpunkt. Für diese Daten stellen wir ein GLM mit Z als Zielvariable auf. Die einzige Kovariate ist die Gruppenzugehörigkeit (kategorial, R-stufig — also entstehen $(R-1)$ Dummy-Variablen).*

- **Beispiel: Querschnitts-GLM zu** $t = t_2$ *mit der Modellgleichung*

$$Z(\,t_2\,) \sim c_0 + c_1 \cdot [\,\text{Gruppe 2 ?}\,] + c_2 \cdot [\,\text{Gruppe 3 ?}\,] + \ldots + c_{R-1} \cdot [\,\text{Gruppe R ?}\,]$$



Zweifaktorielle ANOVA: Ablauf

- **Zweifaktorielle ANOVA: mehrstufiges Verfahren**

Schritt 0: Vorarbeiten

— *Datenstruktur und Voraussetzungen überprüfen*

— *Mauchly-Test durchführen, Korrekturfaktor ε für die Freiheitsgrade bestimmen*

Schritt 1: Drei Globaltests

— *Zerlegung der Gesamt-Stichprobenvarianz, Aufbau der Prüfgrößen*

— *Globaltest I: Ist der Interaktionseffekt "Zeit : Gruppe" signifikant?*

— *Globaltest II: Ist der Haupteffekt "Zeit" signifikant?*

— *Globaltest III: Ist der Haupteffekt "Gruppe" signifikant?*

Schritt 2: Post-hoc-Analysen

— *vorgeplante Kontrastanalysen*

— *Querschnittsmodelle und Parameterschätzungen*

Zweifaktorielle ANOVA: technische Voraussetzungen

• **Allgemeine Voraussetzungen:**

— Die abhängige Variable ist in allen Gruppen und zu allen Zeitpunkten normalverteilt.
— Es gibt keine fehlenden Werte (sehr restriktiv!).

• **Voraussetzungen zum Zwischensubjektfaktor** *(Gruppe):*

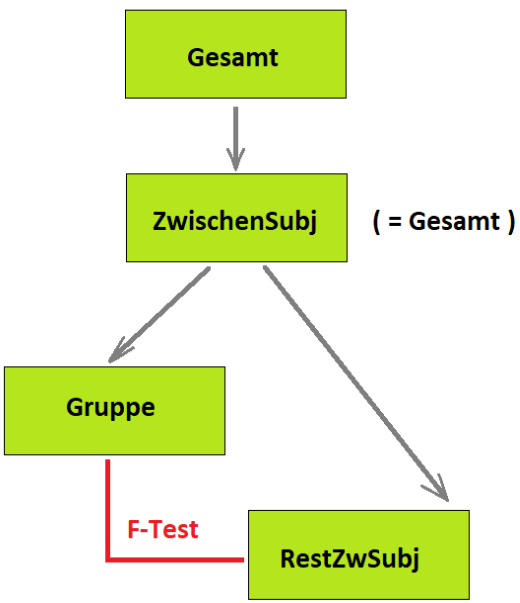
— Der Zwischensubjektfaktor teilt die Probanden in R unabhängige Gruppen auf, die alle denselben Umfang N haben (es gibt also insgesamt $N \cdot R$ Teilnehmer).
— Die Varianzen der abhängigen Variable stimmen gruppenweise (annähernd) überein.

• **Voraussetzungen zum Innersubjektfaktor** *(Zeit):*

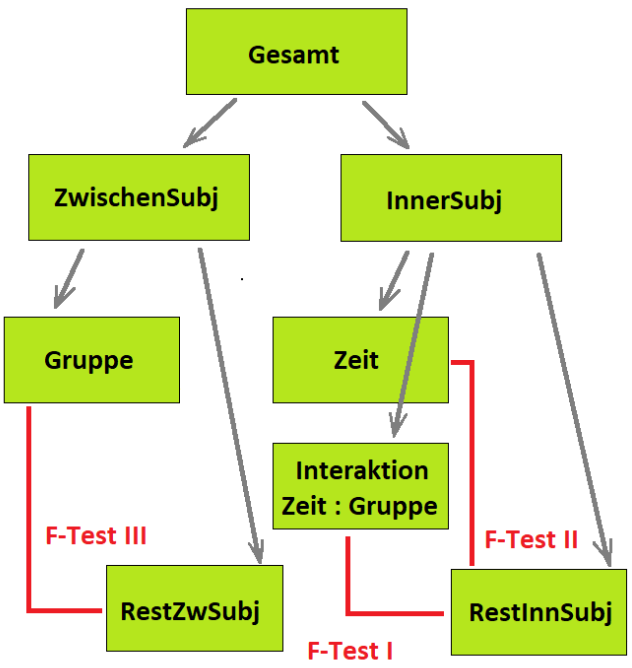
— Die Varianzen der abhängigen Variable für alle Teilnehmer stimmen zu den verschiedenen Zeitpunkten (annähernd) überein.
— Die auf die verschiedenen Zeitpunkte bezogenen Kovarianzmatrizen $\widehat{C}(r)$ der abhängigen Variable für die einzelnen Gruppen stimmen (annähernd) überein.
— Die auf die verschiedenen Zeitpunkte bezogene Kovarianzmatrix $\widehat{C}$ der abhängigen Variable für alle Teilnehmer ist "sphärisch" (globale Varianz-Homogenität).

Zerlegung der Gesamt-Varianz

- **Einfaktorielle ANOVA** *zum Vergleich*



- **Zweifaktorielle ANOVA**



Zerlegung der Gesamt-Varianz (II)

Zähler	*Nenner (df)*	*emp. Varianz*	*Prüfgröße*
SS_{Gesamt}	$(N \cdot R \cdot T) - 1$	$\widehat{\sigma}^2_{Gesamt}$	
SS_{ZwSubj}	$(N \cdot R) - 1$	$\widehat{\sigma}^2_{ZwSubj}$	
SS_{Gruppe}	$R - 1$	$\widehat{\sigma}^2_{Gruppe}$	$\widehat{\sigma}^2_{Gruppe} / \widehat{\sigma}^2_{RestZwSubj} = \widehat{T}_{III}$
$SS_{RestZwSubj}$	$(N-1) \cdot R$	$\widehat{\sigma}^2_{RestZwSubj}$	
$SS_{InnSubj}$	$N \cdot R \cdot (T-1)$	$\widehat{\sigma}^2_{InnSubj}$	
SS_{Zeit}	$(T-1)$	$\widehat{\sigma}^2_{Zeit}$	$\widehat{\sigma}^2_{Zeit} / \widehat{\sigma}^2_{RestInnSubj} = \widehat{T}_{II}$
$SS_{Zeit:Gruppe}$	$(R-1) \cdot (T-1)$	$\widehat{\sigma}^2_{Zeit:Gruppe}$	$\widehat{\sigma}^2_{Zeit:Gruppe} / \widehat{\sigma}^2_{RestInnSubj}$
$SS_{RestInnSubj}$	$(N-1) \cdot R \cdot (T-1)$	$\widehat{\sigma}^2_{RestInnSubj}$	$= \widehat{T}_{I}$

- $\widehat{T}_{I} = 0$, *falls* $SS_{Zeit} + SS_{Gruppe} = SS_{Zellen}$ *(siehe nächste Seite)*
- $\widehat{T}_{II} = 0$, *falls* $\widehat{\mu}(t_1) = \widehat{\mu}(t_2) = \ ... \ = \widehat{\mu}(t_T)$ *(Gruppenmittel für jeden Zeitpunkt)*
- $\widehat{T}_{III} = 0$, *falls* $\widehat{\mu}(r_1) = \widehat{\mu}(r_2) = \ ... \ = \widehat{\mu}(r_R)$ *(Zeitmittel für jede Gruppe)*

Zerlegung der Gesamt-Varianz: Sums of Squares

- *Teilnehmer pro Gruppe:* $1 \leqslant n \leqslant N$, *Gruppen:* $1 \leqslant r \leqslant R$, *Zeitpunkte:* $1 \leqslant t \leqslant T$

$$SS_{\textit{Gesamt}} = \sum_n \sum_r \sum_t \big(Z(n,r,t) - \widehat{\mu} \big)^2 , \quad \widehat{\mu} = \frac{1}{N \cdot R \cdot T} \cdot \sum_n \sum_r \sum_t Z(n,r,t) ,$$

$$SS_{\textit{ZwSubj}} = T \cdot \sum_n \sum_r \big(\widehat{\mu}(n,r) - \widehat{\mu} \big)^2 , \qquad \widehat{\mu}(n,r) = \frac{1}{T} \cdot \sum_t Z(n,r,t) \quad (N \cdot R \textit{ Stück}),$$

$$SS_{\textit{Gruppe}} = N \cdot T \cdot \sum_r \big(\widehat{\mu}(r) - \widehat{\mu} \big)^2 , \qquad \widehat{\mu}(r) = \frac{1}{N \cdot T} \cdot \sum_n \sum_t Z(n,r,t) \qquad (R \textit{ Stück}),$$

$$SS_{\textit{RestZwSubj}} = SS_{\textit{ZwSubj}} - SS_{\textit{Gruppe}} ,$$

$$SS_{\textit{InnSubj}} = SS_{\textit{Gesamt}} - SS_{\textit{ZwSubj}} ,$$

$$SS_{\textit{Zeit}} = N \cdot R \cdot \sum_t \big(\widehat{\mu}(t) - \widehat{\mu} \big)^2 , \qquad \widehat{\mu}(t) = \frac{1}{N \cdot R} \cdot \sum_n \sum_r Z(n,r,t) \qquad (T \textit{ Stück}),$$

$$SS_{\textit{Zellen}} = N \cdot \sum_r \sum_t \big(\widehat{\mu}(r,t) - \widehat{\mu} \big)^2 , \qquad \widehat{\mu}(r,t) = \frac{1}{N} \cdot \sum_r \sum_t Z(n,r,t) \ (R \cdot T \textit{ Stück}),$$

$$SS_{\textit{Zeit:Gruppe}} = SS_{\textit{Zellen}} - SS_{\textit{Zeit}} - SS_{\textit{Gruppe}} ,$$

$$SS_{\textit{RestInnSubj}} = SS_{\textit{InnSubj}} - SS_{\textit{Zeit}} - SS_{\textit{Zeit:Gruppe}} .$$

Zerlegung der Varianz am Beispiel

• **Anwendung auf Beispiel 6.2.** *mit* $N = 50$, $R = 2$, $T = 7$*:*

Bezeichnung Zähler	*Wert Zähler*	*Nenner (df)*	*empirische Varianz*
SS_{Gesamt}	31 174 896.7	699	$\widehat{\sigma}^2_{Gesamt} = 44\,599.3$
SS_{ZwSubj}	3 445 409.5	99	$\widehat{\sigma}^2_{ZwSubj} = 34\,802.1$
SS_{Gruppe}	1 142 592.8	1	$\widehat{\sigma}^2_{Gruppe} = 1\,142\,592.8$
$SS_{RestZwSubj}$	2 302 816.7	98	$\widehat{\sigma}^2_{RestZwSubj} = 23\,498.1$
$SS_{InnSubj}$	27 729 487.1	600	$\widehat{\sigma}^2_{InnSubj} = 46\,215.8$
SS_{Zeit}	24 594 729.2	6	$\widehat{\sigma}^2_{Zeit} = 4\,099\,122.0$
$SS_{Zeit:Gruppe}$	1 155 839.9	6	$\widehat{\sigma}^2_{Zeit:Gruppe} = 192\,640.0$
$SS_{RestInnSubj}$	1 978 918.0	588	$\widehat{\sigma}^2_{RestInnSubj} = 3\,365.5$

• **Werte der drei Prüfgrößen:**

$\widehat{T}_I = 57.24$, $\widehat{T}_{II} = 1217.98$, $\widehat{T}_{III} = 48.62$

Mathematischer Hintergrund: Verteilung der Prüfgrößen

• **Lehrsatz 6.3.:** *Sei $R \geqslant 2$ und $T \geqslant 2$. Eine metrische Variable Z sei in $\mathcal{P}_1$, ... , $\mathcal{P}_R$ zu allen Zeitpunkten t_1, ... , t_T normalverteilt. Ziehen wir immer wieder R paarweise unabhängige Sets von je T verbundenen Stichproben $\big(\mathcal{S}_1(t_1), \mathcal{S}_1(t_2), \ldots, \mathcal{S}_1(t_T)\big) \subset \mathcal{P}_1$, ... , $\big(\mathcal{S}_R(t_1), \mathcal{S}_R(t_2), \ldots, \mathcal{S}_R(t_T)\big) \subset \mathcal{P}_R$ mit stets demselben Umfang N, so gilt:*

1) Unter den Voraussetzungen von p. 6-18 sind die drei Prüfgrößen $\widehat{T}_I$, $\widehat{T}_{II}$ und $\widehat{T}_{III}$ jeweils Fisher-verteilt (mit Freiheitsgraden df_1 und df_2 laut Tabelle).

2) Wenn die "Sphärizitäts"-Voraussetzung für C verletzt ist, bleibt Aussage 1) richtig, aber die Freiheitsgrade für die Fisher-Verteilungen von $\widehat{T}_I$, $\widehat{T}_{II}$ müssen korrigiert werden. Dabei ist nach der Regel auf p. 6-25 entweder $\varepsilon = \widehat{\varepsilon}_{G-G}$ oder $\varepsilon = \widehat{\varepsilon}_{H-F}$ zu wählen.

	$\widehat{T}_I$	$\widehat{T}_{II}$	$\widehat{T}_{III}$
df_1 df_2	$(R-1)\cdot(T-1)$ $(N-1)\cdot R\cdot(T-1)$	$(T-1)$ $(N-1)\cdot R\cdot(T-1)$	$(R-1)$ $(N-1)\cdot R$
df_1, *korr.* df_2, *korr.*	$\varepsilon\cdot(R-1)\cdot(T-1)$ $\varepsilon\cdot(N-1)\cdot R\cdot(T-1)$	$\varepsilon\cdot(T-1)$ $\varepsilon\cdot(N-1)\cdot R\cdot(T-1)$	*wie oben* *wie oben*

Hintergrund (II): "Sphärizität" und Mauchly-Test

• $Z(n,r,t)$ — *Messung von Z für Teilnehmer n in Gruppe r zum Zeitpunkt t*

• *Für je zwei verschiedene Zeitpunkte t_i, t_j bilden wir die empirische Standardabweichung der Differenzen*

$$\widehat{\sigma}_{t_i - t_j} = \sqrt{\frac{1}{N \cdot R - 1} \cdot \sum_n \sum_r \Big(\big(Z(n,r,t_i) - Z(n,r,t_j) \big) - \big(\widehat{\mu}(t_i) - \widehat{\mu}(t_j) \big) \Big)^2}$$

$$\text{mit} \quad \widehat{\mu}(t_i) = \frac{1}{N \cdot R} \cdot \sum_n \sum_r Z(n,r,t_i)\,, \quad \widehat{\mu}(t_j) = \frac{1}{N \cdot R} \cdot \sum_n \sum_r Z(n,r,t_j)\,.$$

• **Sphärizität** *bedeutet, daß alle empirischen Standardabweichungen der Differenzen gleich sind:* $\widehat{\sigma}_{t_2-t_1} = \widehat{\sigma}_{t_3-t_1} = \widehat{\sigma}_{t_4-t_1} = \ldots = \widehat{\sigma}_{t_3-t_2} = \widehat{\sigma}_{t_4-t_2} = \ldots = \widehat{\sigma}_{t_4-t_3} = \ldots$ *Dies kann mit Hilfe der Kovarianzmatrix $\widehat{C}$ für alle Teilnehmer überprüft werden.*

• **Der Mauchly-Test** *erfolgt gegen die Nullhypothese, daß für die Kovarianzmatrix C in der Gesamtpopulation $\mathcal{P}$ Sphärizität vorliegt. Fällt er signifikant aus, so müssen später die Freiheitsgrade für die globalen Prüfgrößen korrigiert werden.*

Hintergrund (III): Mauchly-Test und Korrekturfaktor

• **Mauchly-Test im Beispiel:** *Für Beispiel 6.2. ergibt sich* $\widehat{W}_{trf} = 512.55$ *als transformierte Prüfgröße.* $\widehat{W}_{trf}$ *ist annähernd* χ^2*-verteilt mit df*$= 20$ *Freiheitsgraden, damit erhalten wir* $p < 0.0005$*, und der Test ist z.B. zu* $\alpha = 0.05$ *signifikant.*

• **Korrekturfaktor nach Greenhouse-Geisser:** $\widehat{\varepsilon}_{G-G}$ *wird aus den Daten der Kovarianzmatrix* C *für alle Teilnehmer und der Anzahl* T *der Zeitpunkte berechnet. Es gilt* $1/(T-1) \leqslant \widehat{\varepsilon}_{G-G} \leqslant 1$*; je kleiner* $\widehat{\varepsilon}_{G-G}$ *ausfällt, desto stärker ist die Sphärizität verletzt. Im Beispiel:* $\widehat{\varepsilon}_{G-G} = 0.452$*.*

• **Korrekturfaktor nach Huynh-Feldt:** $\widehat{\varepsilon}_{H-F}$ *wird aus* $\widehat{\varepsilon}_{G-G}$ *sowie den Anzahlen* N*,* R *und* T *der Teilnehmer pro Gruppe, der Gruppen und der Zeitpunkte berechnet. Je kleiner* $\widehat{\varepsilon}_{H-F}$ *ausfällt, desto stärker ist die Sphärizität verletzt. Im Beispiel:* $\widehat{\varepsilon}_{H-F} = 0.466$*.*

• **Auswahl des Korrekturfaktors:** *Solange* $\widehat{\varepsilon}_{G-G} < 0.75$ *ausfällt, wird* $\widehat{\varepsilon}_{G-G}$ *als Korrekturfaktor beibehalten; falls* $\widehat{\varepsilon}_{G-G} \geqslant 0.75$*, wird stattdessen* $\widehat{\varepsilon}_{H-F}$ *benutzt.*

• *Details zur Berechnung von* $\widehat{W}_{trf}$*,* $\widehat{\varepsilon}_{G-G}$ *und* $\widehat{\varepsilon}_{H-F}$ *stehen in den Anmerkungen.*

Drei globale Hypothesentests

• **F-Test I (für den Interaktionseffekt):** *Unterscheiden sich die beiden Trainingsmethoden in der Veränderung der Gehstrecken über die Zeit?*

— *Test gegen* $\mathcal{H}_0$*: "*$SS_{Zellen} = SS_{Zeit} + SS_{Gruppe}$*" zum Niveau* $\alpha = 0.05$

— *Prüfgröße:* $\widehat{T}_I = \widehat{\sigma}^2_{Zeit:Gruppe} \,/\, \widehat{\sigma}^2_{RestInnSubj} = 57.24$

Korrektur	Korrekturfaktor	df1	df2	Q_{95}(F, df1, df2)	Prüfgröße	p-Wert
keine	1.000	6	588	2.11	57.24	< 0.0005
Greenhouse-Geisser	0.452	2.712	265.776	2.73	57.24	< 0.0005
Huynh-Feldt	0.466	2.796	274.008	2.70	57.24	< 0.0005

• **F-Test II (für den Haupteffekt Zeit):** *Verändern sich die mittleren Gehstrecken unabhängig von Methode A oder B im Verlauf der Zeit?*

— *Test gegen* $\mathcal{H}_0$*: "*$\mu(t_1) = \mu(t_2) = \ ... \ = \mu(t_7)$*" zum Niveau* $\alpha = 0.05$

— *Prüfgröße:* $\widehat{T}_{II} = \widehat{\sigma}^2_{Zeit} \,/\, \widehat{\sigma}^2_{RestInnSubj} = 1217.98$

Korrektur	Korrekturfaktor	df1	df2	Q_{95}(F, df1, df2)	Prüfgröße	p-Wert
keine	1.000	6	588	2.11	1217.98	< 0.0005
Greenhouse-Geisser	0.452	2.712	265.776	2.73	1217.98	< 0.0005
Huynh-Feldt	0.466	2.796	274.008	2.70	1217.98	< 0.0005

Drei globale Hypothesentests (II)

• **F-Test III (für den Haupteffekt Gruppe):** *Unterscheiden sich die Gehstrecken für die beiden Trainingsmethoden im Mittel über alle Zeitpunkte?*
— *Test gegen* $\mathcal{H}_0$*: "*$\mu(r_1) = \mu(r_2)$*" zum Niveau* $\alpha = 0.05$
— *Prüfgröße:* $\widehat{T}_{III} = \widehat{\sigma}^2_{Gruppe} \,/\, \widehat{\sigma}^2_{RestZwSubj} = 48.62$

Korrektur	Korrekturfaktor	df1	df2	Q_{95}(F, df1, df2)	Prüfgröße	p-Wert
keine	1.000	1	98	3.94	48.62	< 0.0005

• **Effektmaß:** *Das "partielle Eta-Quadrat" zu* $\widehat{T}_I = \widehat{\sigma}^2_{Zeit:Gruppe} \,/\, \widehat{\sigma}^2_{RestInnSubj}$ *ergibt sich als*

$$\eta^2 = 1 - \frac{SS_{RestInnSubj}}{SS_{Zeit:Gruppe} + SS_{RestInnSubj}} = 0.369\,.$$

Analog zu R^2 *ist der Effekt umso stärker, je näher* η^2 *an Eins liegt.*

• **Fazit:** *Die Wegstrecke beim 6-Minuten-Gehtest hängt sowohl von der Trainingsmethode als auch vom Zeitpunkt ab (signifikanter Interaktionseffekt Zeit : Gruppe mit* $\widehat{F}(2.71, 265.78) = 57.24$, $\widehat{\varepsilon}_{G-G} = 0.452$, $\eta^2 = 0.369$ *und* $p < 0.0005$*).*

Einfache Innersubjektkontraste

• **Zeit : Gruppe:** *Die Veränderungen der Gehstrecken zwischen den Gruppen unterscheiden sich bis 8 Wochen signifikant (5-%-Niveau), aber nicht mehr zu 10 und 12 Wochen.*

Bezeichnung	Wert Zähler	Nenner (df)	Prüfgröße	p-Wert	partielles η^2
t2 vs. t1	1794.0	1	8.73	0.004	0.082
t3 vs. t1	256184.8	1	146.15	< 0.0005	0.599
t4 vs. t1	709478.7	1	167.59	< 0.0005	0.631
t5 vs. t1	187761.9	1	58.47	< 0.0005	0.374
t6 vs. t1	23328.0	1	5.68	0.019	0.055
t7 vs. t1	7442.0	1	1.61	0.207	0.016

• **Zeit:** *Die Gehstrecken sind zu allen sechs FUP's gegenüber BAS signifikant erhöht.*

Bezeichnung	Wert Zähler	Nenner (df)	Prüfgröße	p-Wert	partielles η^2
t2 vs. t1	103558.0	1	503.74	< 0.0005	0.837
t3 vs. t1	1536153.9	1	876.37	< 0.0005	0.899
t4 vs. t1	4726582.6	1	1116.47	< 0.0005	0.919
t5 vs. t1	8544151.2	1	2660.72	< 0.0005	0.964
t6 vs. t1	11165647.7	1	2716.32	< 0.0005	0.965
t7 vs. t1	12264123.4	1	2654.67	< 0.0005	0.964

Querschnittsmodelle

• **Lineare Querschnittsmodelle:** *Für jeden Zeitpunkt t berechnen wir ein GLM $Z(t) \sim c_0(t) + c_1(t) \cdot [\text{Gruppe B ?}] + \varepsilon$ (da der Zwischensubjektfaktor in unserem Beispiel nur 2 Stufen hat, handelt es sich hier um eine einfache lineare Regression).*

Zeitpunkt	Abhängige Variable	Koeffizient	geschätzter Wert	Standard-abweichung	p-Wert	Untergrenze 95-%-K.I.	Obergrenze 95-%-K.I.
Baseline (*)	Z(t1)	Intercept	59.38	2.76	< 0.0005	53.90	64.86
		Methode: B	-2.56	3.90	0.513	-10.30	5.18
2 Wochen (*)	Z(t2)	Intercept	98.90	4.70	< 0.0005	89.58	108.22
		Methode: B	9.42	6.64	0.159	-3.75	22.59
4 Wochen	Z(t3)	Intercept	163.08	10.06	< 0.0005	143.12	183.04
		Methode: B	140.60	14.23	< 0.0005	112.37	168.83
6 Wochen	Z(t4)	Intercept	247.72	14.49	< 0.0005	218.97	276.47
		Methode: B	235.68	20.49	< 0.0005	195.03	276.33
8 Wochen	Z(t5)	Intercept	411.48	12.52	< 0.0005	386.64	436.32
		Methode: B	120.00	17.70	< 0.0005	84.88	155.12
10 Wochen	Z(t6)	Intercept	510.34	13.42	< 0.0005	483.72	536.96
		Methode: B	40.64	18.96	0.035	3.02	78.26
12 Wochen (*)	Z(t7)	Intercept	542.44	14.01	< 0.0005	514.64	570.24
		Methode: B	21.84	19.82	0.273	-17.48	61.16

Parameterschätzungen

• **Frage 1:** *Wie schätzt das Modell die Wegstrecke nach 4 Wochen für Methode A (inkl. 95-%-K.I.)?*

4 Wochen	Z(t3)	Intercept	163.08	10.06	< 0.0005	143.12	183.04
		Methode: B	140.60	14.23	< 0.0005	112.37	168.83

— *Antwort: Etwa* $163\,m\,[143\,m, 183\,m]$.

• **Frage 2:** *Wie schätzt das Modell die Wegstrecke nach 4 Wochen für Methode B?*

— *Antwort: Etwa* $(163 + 141)\,m = 304\,m$.

• **Frage 3:** *Wie groß ist der mittlere Unterschied zwischen beiden Gruppen nach 10 Wochen (inkl. 95-%-K.I.)?*

10 Wochen	Z(t6)	Intercept	510.34	13.42	< 0.0005	483.72	536.96
		Methode: B	40.64	18.96	0.035	3.02	78.26

— *Antwort: Etwa* $41\,m\,[3\,m, 78\,m]$.