Wiederholung

Prüfung: 2 Kohorten

Mi 21.01.	401-0643-13S	Statistik II	M. Kalisch	09:00-12:00	211
	Eine Zusammenfassung auf 5 Blättern (maximal DIN-A 4; beliebig beschrieben).				

2 Kohorten alphabetisch nach Nachname sortiert:
Kohorte 1: A bis und mit Müller
Kohorte 2: Müllhaupt bis und mit Z

Prüfung: 2 Kohorten

Mi 21.01.	401-0643-13S	Statistik II	M. Kalisch	09:00-12:00	211
	Eine Zusammenfassung auf 5 Blättern (maximal DIN-A 4; beliebig beschrieben).				

Kohorte 1
8.30 – 8.45: Raum betreten
9 – 12: Prüfung (HG G1)
8.45 – 11.15: Anwesenheitspflicht

Wer früher fertig ist, darf ein fachfremdes Buch/Heft/Ordner aus der Tasche nehmen; Prüfung muss dann aber schon abgegeben sein; KEIN Natel, Compi, etc.

Kohorte 2
11.15 – 12: Schleusenraum (tba) (Anwesenheitspflicht)
12 – 12.15: Transfer zu G1
12.15 – 15.15: Prüfung (HG G1)
Keine weitere Anwesenheitspflicht während Prüfung

Genauere Details folgen noch per mail

Prüfungsumgebung Moodle

- Beispiel (Passwort 1234): https://moodle-app2.let.ethz.ch/mod/quiz/view.php?id=65354
- Single Choice Antworten: Nur eine richtige Antwort
- Numerische Antworten (richtig runden, meist 2 Nachkommastellen) Dezimal-Trennzeichen “,”. Also 6,34 und nicht 6.34

- Sie können R-Studio verwenden. Alle nötigen R Pakete sind installiert, aber noch nicht geladen. Ergebnisse in R-Studio sind für die Korrektoren nicht sichtbar (wie bei einem Taschenrechner).

Genau Form der Prüfung

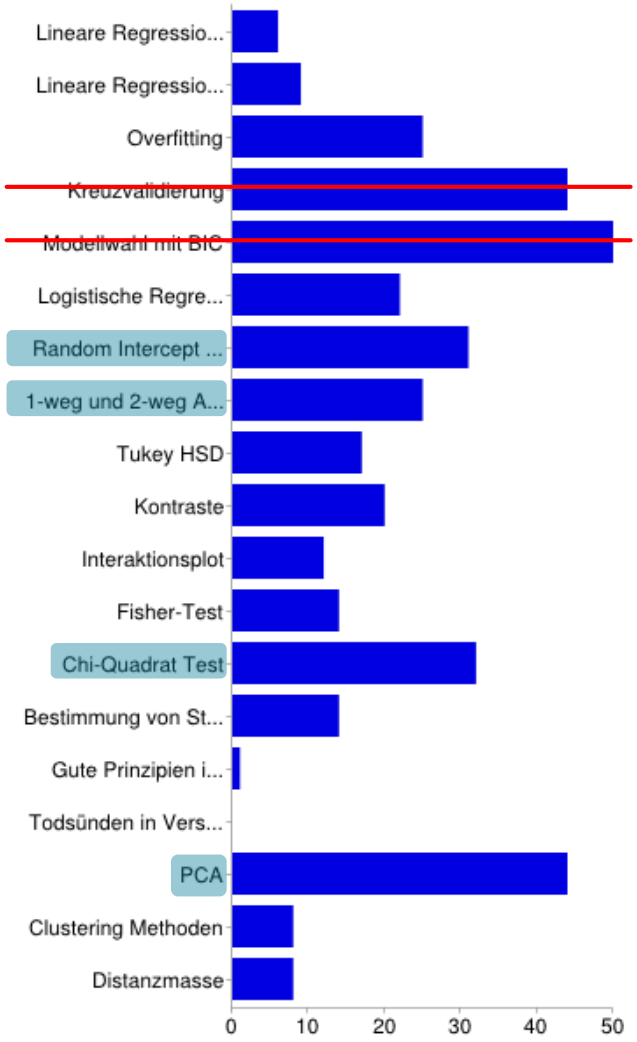
Die Prüfung wird aus zwei Teilen bestehen:

1) 15 MC-Fragen: 13 davon sind sehr ähnlich wie Quizfragen (aber natürlich nicht identisch!); 2 weitere Quizfragen sind etwas schwieriger und verlangen wenige Zeilen in R

2) 3 R-Aufgaben: Das Format ist wie in den Übungen, allerdings werden die Aufgaben 5-10 Unteraufgaben enthalten (also etwas länger sein). Der Inhalt ist sehr ähnlich wie in den Serien, aber natürlich nicht identisch.

Richtwerte für Korrektur:
- Jede (Teil-)Aufgabe gibt 0 oder 1 Punkt (Total also ca. 30-40 Punkte)
- 50% gibt ca. 4, 90% gibt ca. 6

Welche Themen sollen wir in der letzten Vorlesungswoche wiederholen? Deadline: Do, 4.12.14 um 8:00 Uhr



Lineare Regression - Faktoren	6	7%
Lineare Regression - Interaktion	9	10%
Overfitting	25	28%
Kreuzvalidierung	44	50%
Modellwahl mit BIC	50	57%
Logistische Regression	22	25%
Random Intercept and Random Slope Model	31	35%
1-weg und 2-weg ANOVA	25	28%
Tukey HSD	17	19%
Kontraste	20	23%
Interaktionsplot	12	14%
Fisher-Test	14	16%
Chi-Quadrat Test	32	36%
Bestimmung von Stichprobengrösse	14	16%
Gute Prinzipien in Versuchsplanung	1	1%
Todsünden in Versuchsplanung	0	0%
PCA	44	50%
Clustering Methoden	8	9%
Distanzmasse	8	9%

CV und Modellwahl mit BIC kommt nicht in der Prüfung

“Entschärfung” für Prüfung

- V2: KEINE Kreuzvalidierung und KEINE Modellwahl
- Poweranalyse NUR mit Binomialtest
- PCA OHNE Bsp Siebenkampf

Relevanter Code auf unserer Webpage:

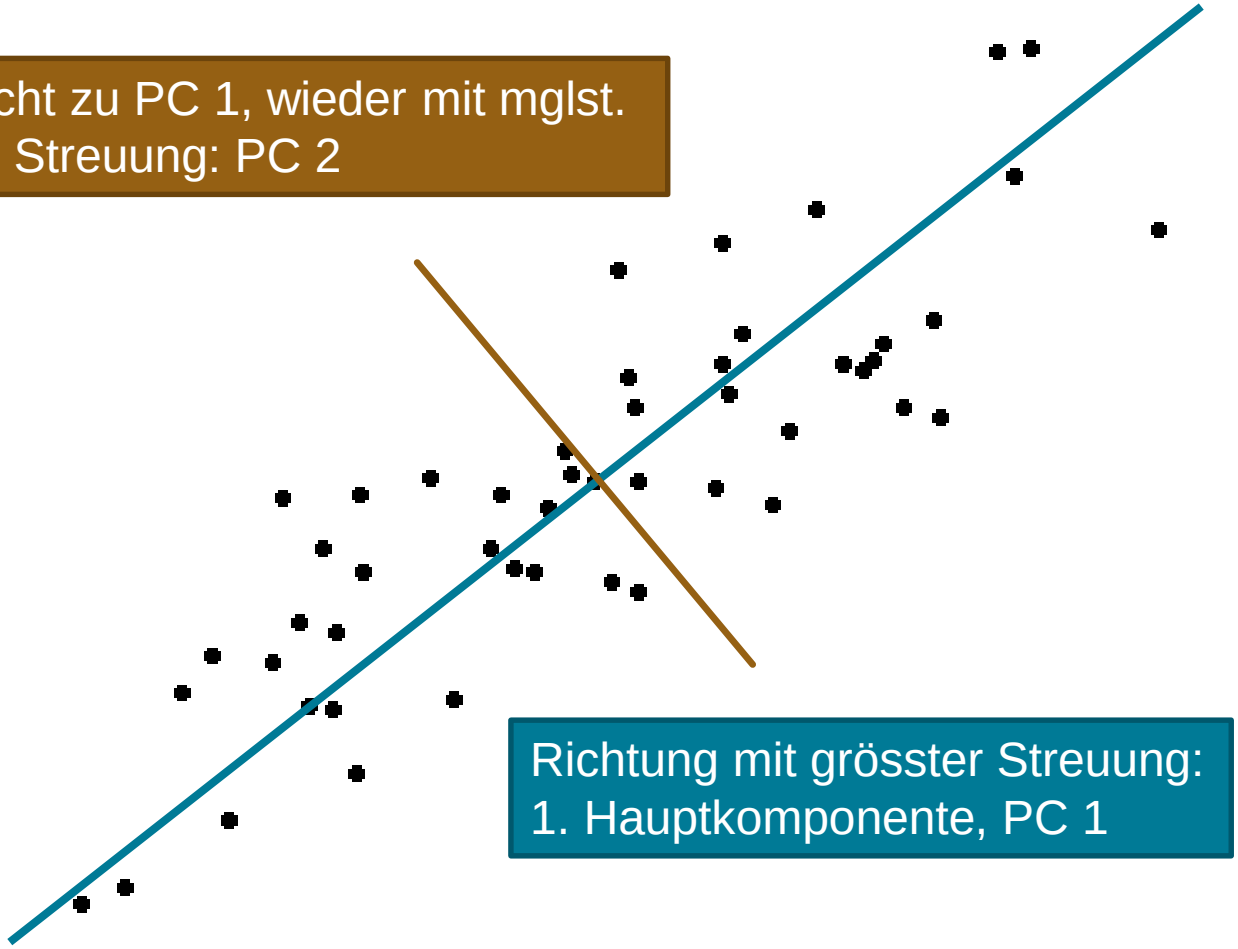
https://stat.ethz.ch/education/semesters/as2014/statistik2/statistik2RCodeVL.R

Wiederholung

- PCA
- Chi-Quadrat Test
- RSRI
- ANOVA

PCA: Intuition

Senkrecht zu PC 1, wieder mit mglst. grosser Streuung: PC 2

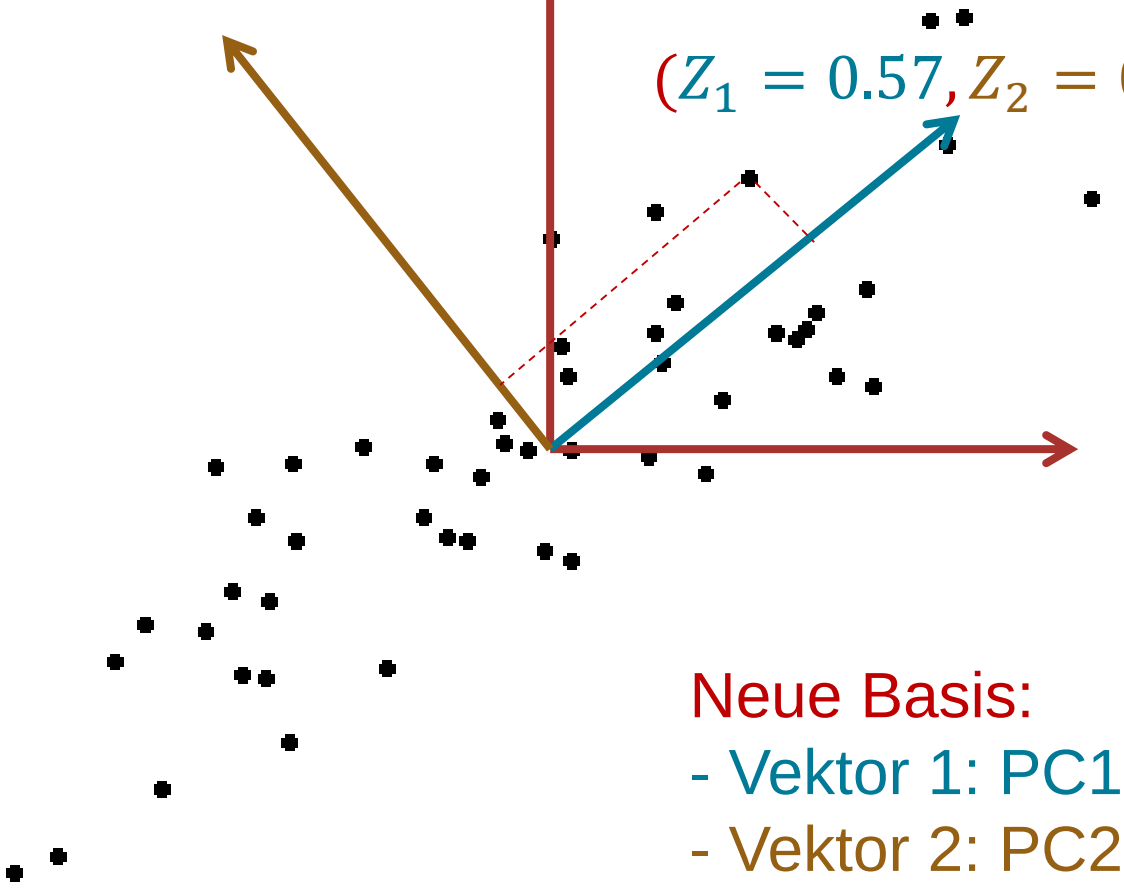


Richtung mit grösster Streuung: 1. Hauptkomponente, PC 1

	Koord. 1	Koord. 2
Std. Basis	$X_1 = 0.3$	$X_2 = 0.5$
PC Basis	$Z_1 = 0.57$	$Z_2 = 0.14$

"scores"

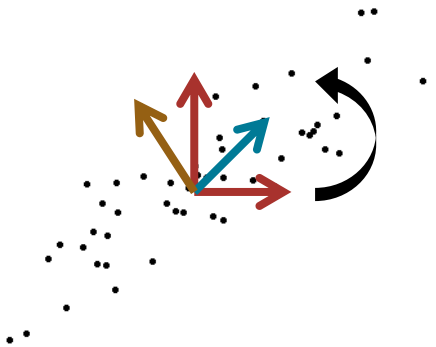
PCA: Basiswechsel



Neue Basis:
- Vektor 1: PC1
- Vektor 2: PC2

PCA: Basiswechsel mit Linearer Algebra

	Koord. 1	Koord. 2
Std. Basis	$X_1 = 0.3$	$X_2 = 0.5$
PC Basis	$Z_1 = 0.57$	$Z_2 = 0.14$



- Standard Basis und PC Basis sind je eine Orthonormal Basis (Achsen senkrecht, Länge 1)
- Basiswechsel: Rotationsmatrix Φ
- Spalten der Rotationsmatrix sind *loadings*:

$$\Phi = \begin{matrix} \text{PC1} & \text{PC2} \\ \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} & \begin{matrix} \text{X1} \\ \text{X2} \end{matrix} \end{matrix}$$

- Basiswechsel mit Rotationsmatrix ist einfach:
  Φ: Von PC Basis nach Standardbasis
  Φ^{-1}: Von Standardbasis nach PC Basis

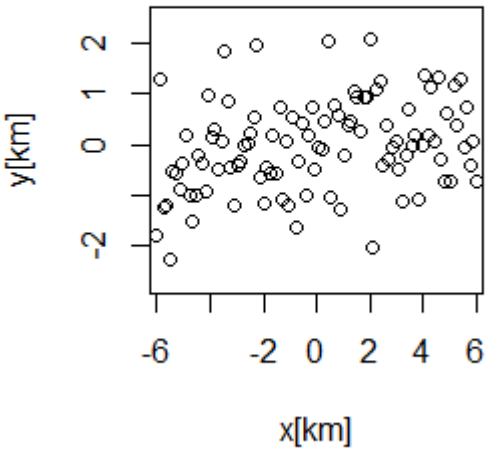
Bzgl. Std.basis

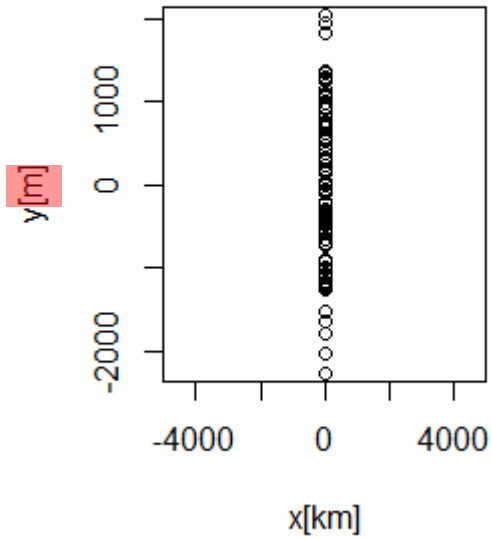
$$\Phi^{-1} = \begin{matrix} \text{X1} & \text{X2} \\ \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} & \begin{matrix} \text{PC1} \\ \text{PC2} \end{matrix} \end{matrix} ; Z = \Phi^{-1} * X = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} * \begin{pmatrix} 0.3 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.57 \\ 0.14 \end{pmatrix}$$

Bzgl. PC Basis "scores"

To scale or not to scale …

Messungen auf einer Landkarte (z.B. Bodenschätze)





Welche Einheiten ?

Beispiel 1: Interpretation der PCs

```
> pr.out$rotation
                PC1        PC2
Murder   -0.5358995  0.4181809
Assault  -0.5831836  0.1879856
UrbanPop -0.2781909 -0.8728062
Rape     -0.5434321 -0.1673186
```

- PC 1 ist gross, wenn v.a. Murder, Assault und Rape klein sind
  → PC 1 spiegelt “Verbrechen” wieder
- PC 2 ist gross, wenn UrbanPop klein ist
  → PC 2 spiegelt “Verstädterung” wieder

R

- Ab Zeile 750

Wiederholung

- PCA
- Chi-Quadrat Test
- RSRI
- ANOVA

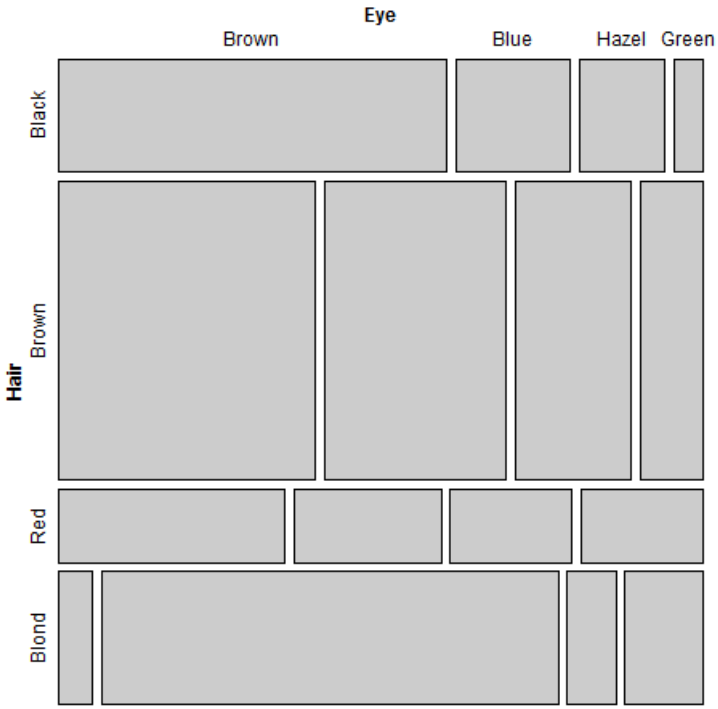
Chi-Quadrat Test: Spalten und Zeilen unabhängig?

- Haar- und Augenfarbe (R: ?HairEyeColor)

Hair / Eye	Brown	Blue	Hazel	Green	Total
Black	68	20	15	5	108
Brown	119	84	54	29	286
Red	26	17	14	14	71
Blond	7	94	10	16	127
Total	220	215	93	64	592

- Mögliche Fragen:
  - Visualisierung (v.a. wenn mehr als 2 Kategorien)
  - Abhängigkeit? Wo?

Visualisierung kategorischer Daten: Mosaic Plot



Hair / Eye	Brown	Blue	Hazel	Green	Total
Black	68	20	15	5	108
Brown	119	84	54	29	286
Red	26	17	14	14	71
Blond	7	94	10	16	127

Fläche proportional zu Tabelleneintrag

Chi-Quadrat Test

"observed values"
$O_{ij} = n_{ij}$

	A=1	…	A=n	Total
B=1	n_{11}		n_{1n}	n_{1*}
…				
B=m	n_{m1}		n_{mn}	n_{m*}
Total	n_{*1}		n_{*n}	n

H_0: A, B sind unabhängig

$$P(A = i \cap B = j) = P(A = i) * P(B = j) \approx \hat{P}(A = i) * \hat{P}(B = j) = \frac{n_{*i}}{n} * \frac{n_{j*}}{n}$$

Erwartungswert der Zelle falls H_0 stimmt: $E_{ij} = n * \frac{n_{*i}}{n} * \frac{n_{j*}}{n} = \frac{n_{*i}\, n_{j*}}{n}$

Chi-Quadrat Test

	A=1	…	A=n	Total
B=1	n_{11}		n_{1n}	n_{1*}
…				
B=m	n_{m1}		n_{mn}	n_{m*}
Total	n_{*1}		n_{*n}	n

Wie verschieden sind beobachtete und erwartete Werte?
Verbreitet: **Pearson Chi-Quadrat Statistik**

$$X^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij}^2$$

Pearson Residuen

$$R_{ij} = \frac{O_{ij}-E_{ij}}{\sqrt{E_{ij}}}$$

Beitrag jeder Zelle zur Modellabweichung

Falls H_0 stimmt, folgt X^2 einer Chi-Quadrat Verteilung mit (I-1)(J-1) Freiheitsgraden (*falls n gross – s. nächste slide*).
→ p-Werte

R

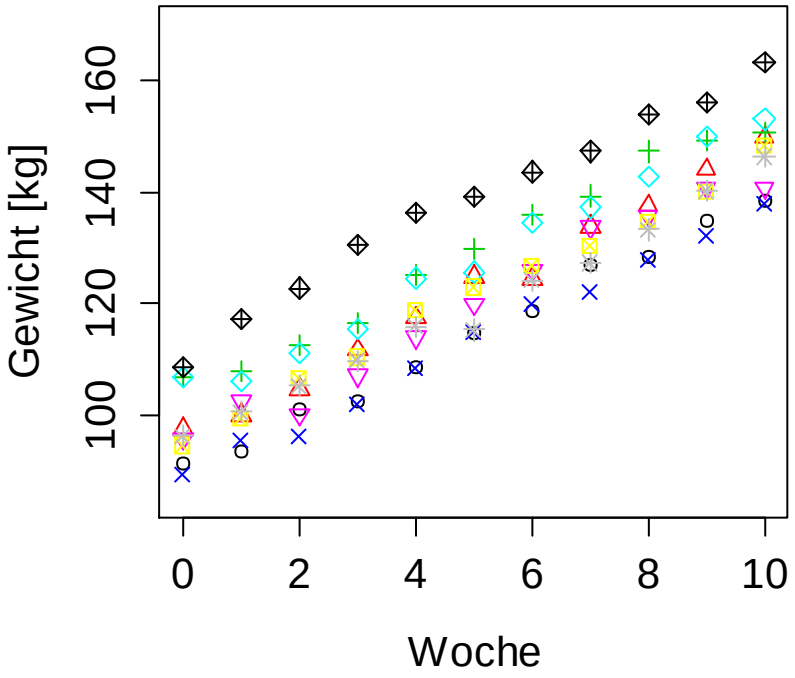
- Ab Zeile 411

Wiederholung

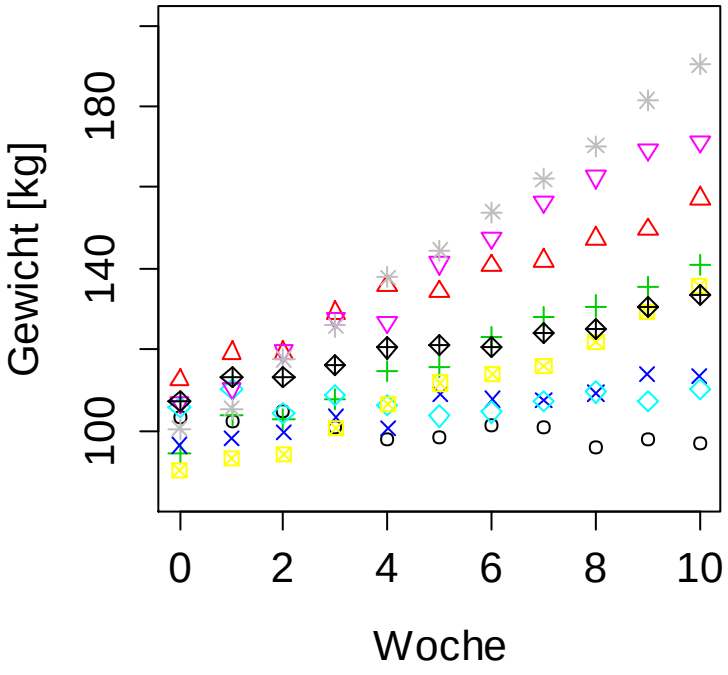
- PCA
- Chi-Quadrat Test
- RSRI
- ANOVA

Zurück zum Krafttraining

Jede Person hat eine unterschiedliche Kraft zu Beginn



Unterschiedliche Kraft zu Beginn
&
Spricht unterschiedlich auf Training an



Wiederholte Messungen 3/3: Random Slope and Random Intercept (RIRS)

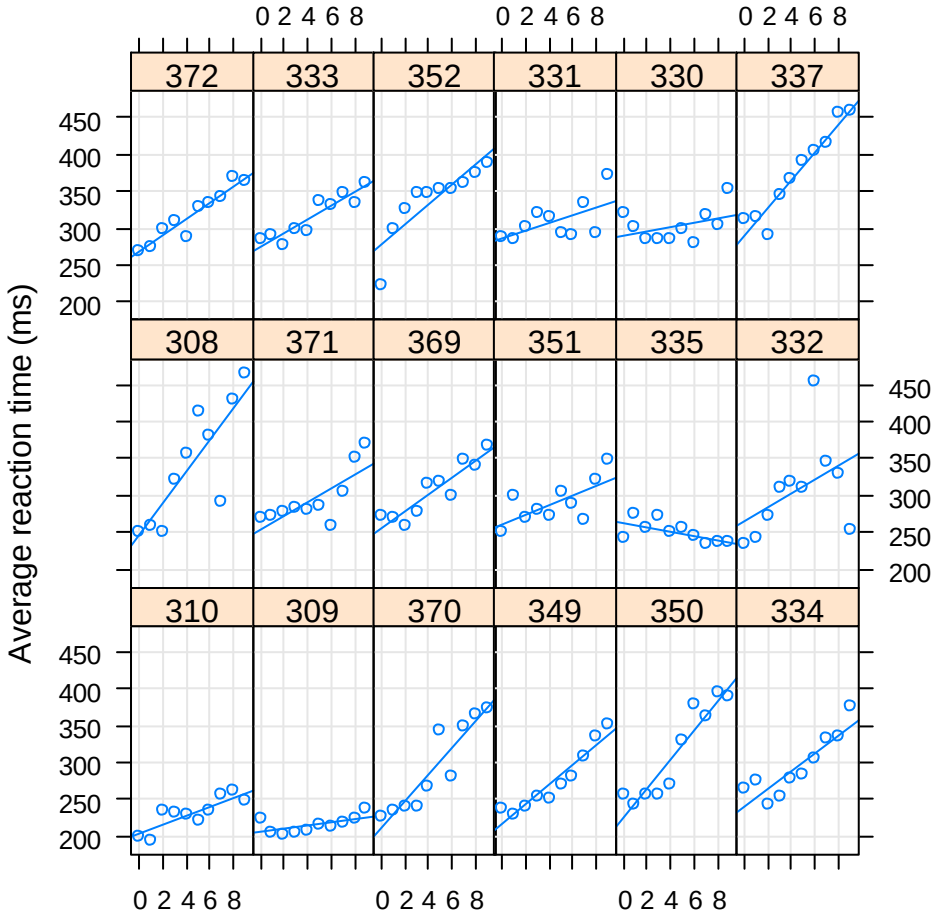
i: Person
j: Woche

- Möglichkeit 2: Mixed Effects Model

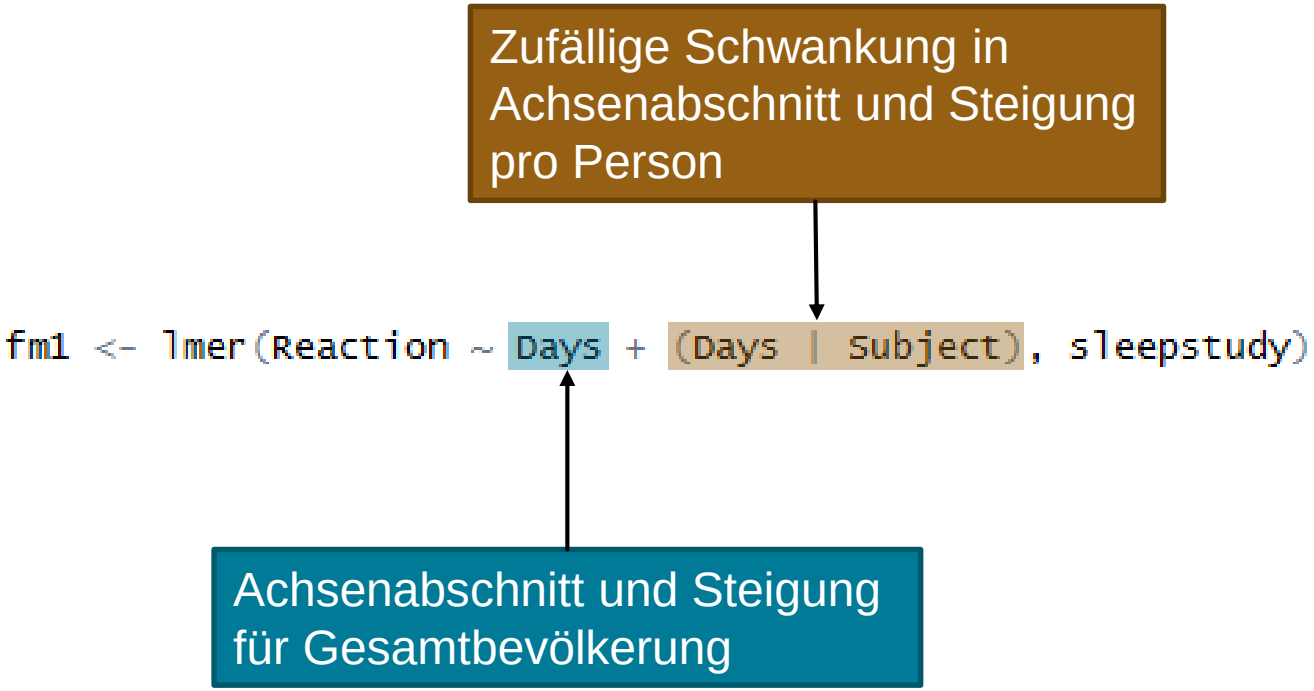
$$y_{ij} = (\beta_0 + u_{1,i}) + (\beta_1 + u_{2,i})x_j + \varepsilon_{ij},$$
$$\varepsilon_{ij} \sim N(0, \sigma^2)\ i.i.d$$
$$u_{1,i} \sim N(0, \sigma_1^2), u_{2,i} \sim N(0, \sigma_2^2), cor(u_1, u_2) = \rho$$

- Schätze: $\beta_0, \beta_1, \sigma, \sigma_1, \sigma_2, \rho$

Reaktionszeit - Überblick



RIRS Modell in R: Input



RIRS Modell in R: Output

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.09   24.740
          Days         35.07    5.922   0.07
 Residual             654.94   25.592
Number of obs: 180, groups:  Subject, 18

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  251.405      6.825  16.998  36.838  < 2e-16
Days          10.467      1.546  16.995   6.771 3.27e-06
```

$$y_{ij} = (251.4 + u_{1,i}) + (10.5 + u_{2,i})x_j + \varepsilon_{ij},$$
$$\varepsilon_{ij} \sim N(0, 25.6^2)\ i.i.d$$
$$u_{1,i} \sim N(0, 24.7^2), u_{2,i} \sim N(0, 5.9^2), cor(u_1, u_2) = 0.07$$

R

- Ab Zeile 269

Wiederholung

- PCA
- Chi-Quadrat Test
- RSRI
- ANOVA

ANOVA - Idee

- ANOVA 1: Zwei Medikamente zur Blutdrucksenkung und Placebo (Faktor). Gibt es einen sign. Unterschied in der Wirkung (kontinuierlich)?

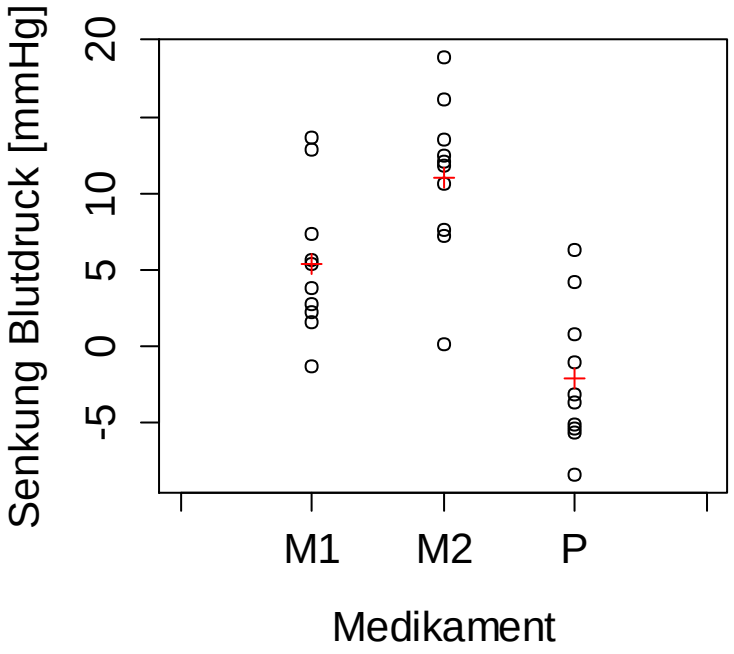
$$Y \sim X + \varepsilon$$ 1-weg ANOVA

- ANOVA 2: Zwei Medikamente zur Blutdrucksenkung, Placebo (Faktor) und Geschlecht (Faktor). Gibt es einen sign. Unterschied in der Wirkung (kontinuierlich) (evtl. geschlechterspezifisch)?

$$Y \sim X1 + X2 + \varepsilon$$ 2-weg ANOVA

Beispiel in R: ANOVA-Tabelle

$g = 3, p = 10$



$SS_B = 872.3$
$SS_W = 642.1$

$$F = \frac{436.1}{23.8} = 18.34$$

```
> fm <- aov(y ~ g, data = df)
> summary(fm)
            Df Sum Sq Mean Sq F value   Pr(>F)
g            2  872.3   436.1   18.34 9.32e-06 ***
Residuals   27  642.1    23.8
```

$g - 1 = 2$
$g*(p-1)=27$

$$MS_B = \frac{872.3}{2} = 436.1$$
$$MS_W = \frac{642.1}{27} = 23.8$$

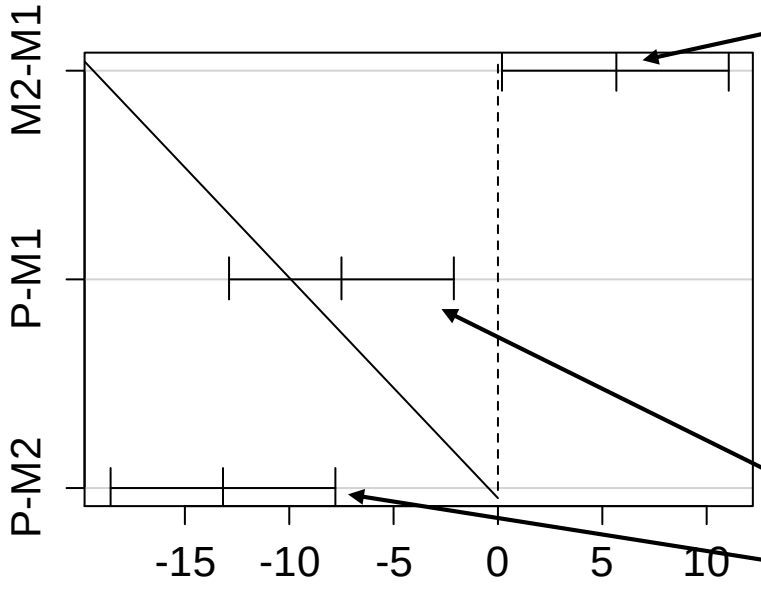
Für 1-weg und 2-weg ANOVA besprochen

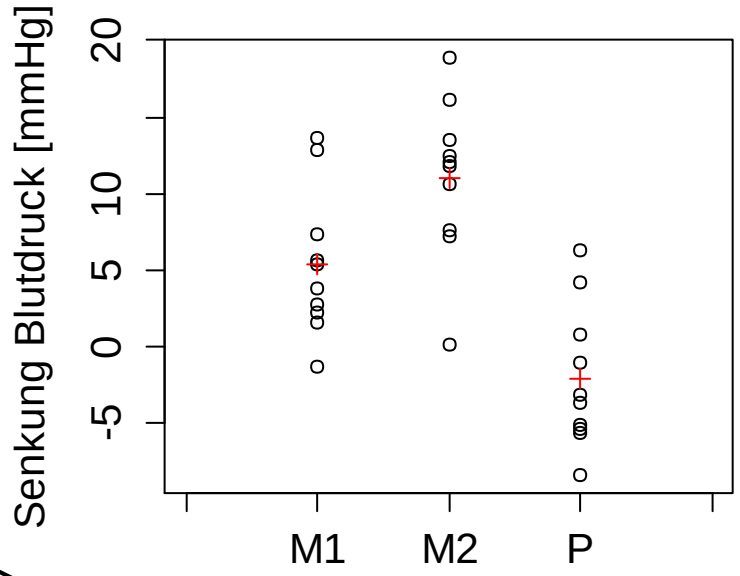
Beispiel in R: TukeyHSD

```
TukeyHSD(fm)
plot(TukeyHSD(fm))
```

M2 ist deutlich wirksamer als M1

95% family-wise confidence level





M1 und M2 sind deutlich wirksamer als Placebo

Konstraste – Bsp 1: Paarweise Vergleiche

(Alternative zu TukeyHSD)

$$\underset{K}{\begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}} * \underset{\mu}{\begin{pmatrix} \mu_{M1} \\ \mu_{M2} \\ \mu_P \end{pmatrix}} = \underset{m}{\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}} \quad \Leftrightarrow \quad \begin{aligned} \mu_{M2} - \mu_{M1} &= 0 \\ \mu_P - \mu_{M1} &= 0 \\ \mu_P - \mu_{M2} &= 0 \end{aligned}$$

Kontraste nur für 1-weg ANOVA besprochen

Kontraste – Bsp 1: R

- Funktion ‘glht’ (General Linear Hypotheses Test) im package ‘multcomp’

```
Fit: aov(formula = y ~ g, data = df)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
M2-M1 == 0    5.670      2.181   2.600  0.03853 *
P-M1 == 0    -7.496      2.181  -3.437  0.00533 **
P-M2 == 0   -13.166      2.181  -6.037  < 0.001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Approx. 95%-VI für Unterschied M1 vs. M2:
$5.67 \pm 2 * 2.181$

Kontraste – Bsp 2: Gruppe der Medikamente vs. Placebo

$$\begin{pmatrix} 0.5 & 0.5 & -1 \\ -1 & 1 & 0 \end{pmatrix} * \begin{pmatrix} \mu_{M1} \\ \mu_{M2} \\ \mu_P \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$\Longleftrightarrow$

Medikamente vs. Placebo

$$0.5 * \mu_{M1} + 0.5 * \mu_{M2} - \mu_P = 0$$

$$\mu_{M2} - \mu_{M1} = 0$$

Medikamente untereinander

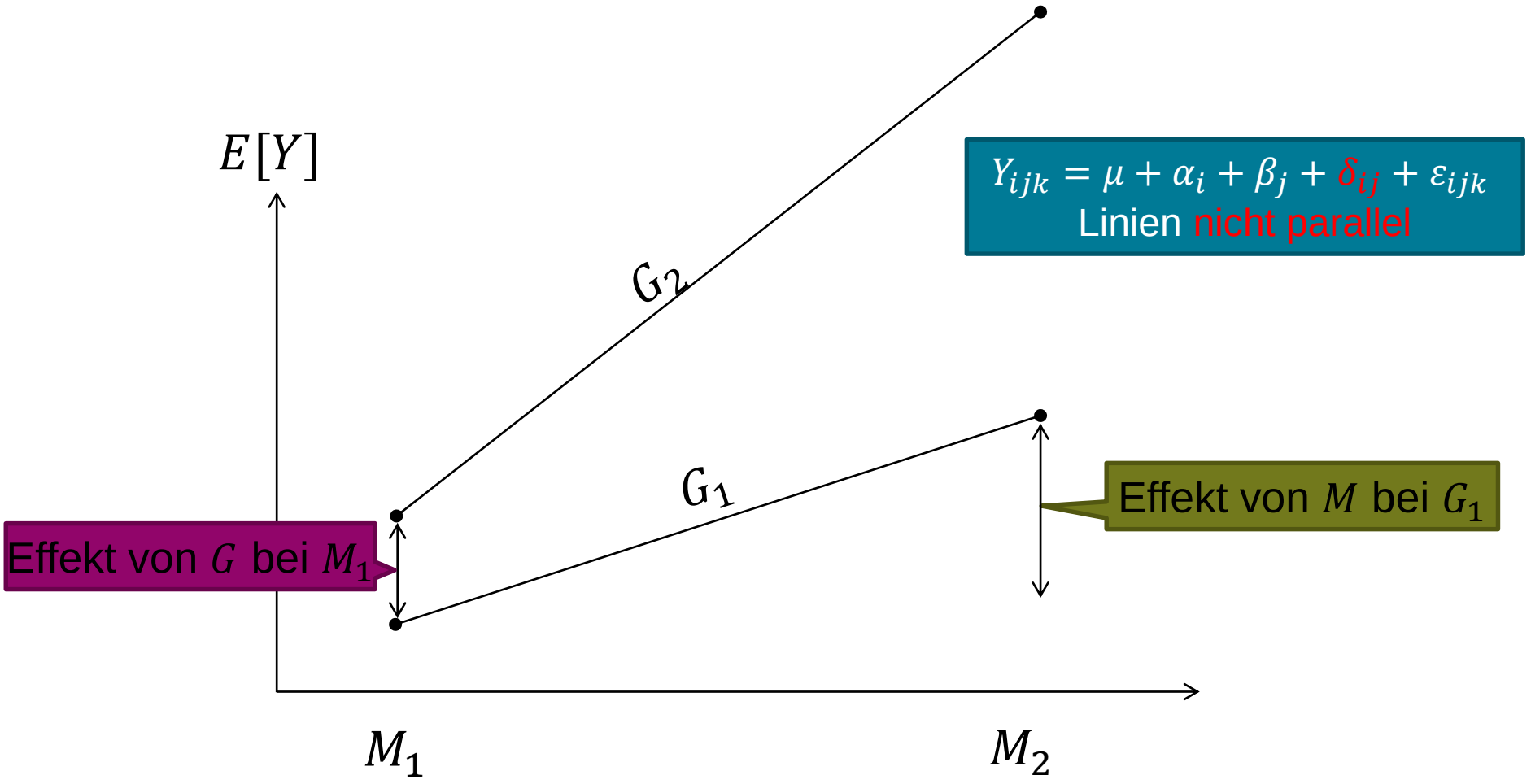
Kontraste – Bsp 2: R

```
  Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: aov(formula = y ~ g, data = df)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
M-P == 0     10.331      1.889    5.47 1.73e-05 ***
M2-M1 == 0    5.670      2.181    2.60   0.0294 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Die Medikamente sind deutlich wirksamer als Placebo

M2 ist deutlich wirksamer als M1

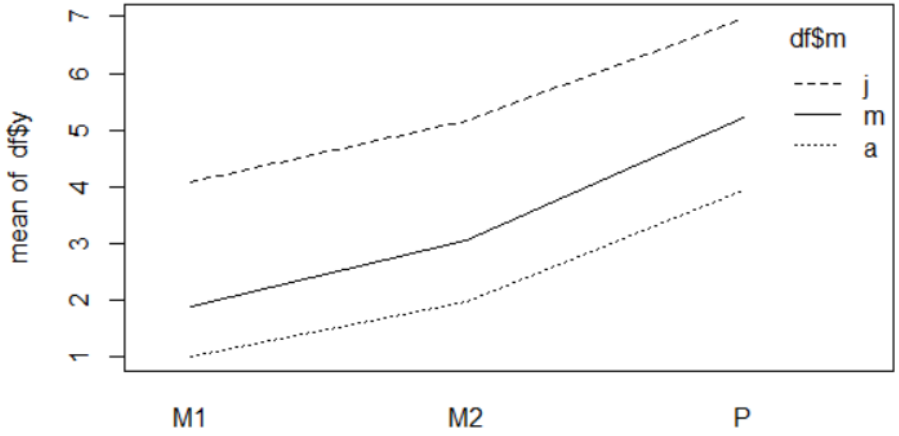
Nur 2-weg ANOVA: Modell-Visualisierung: Mit Interaktion



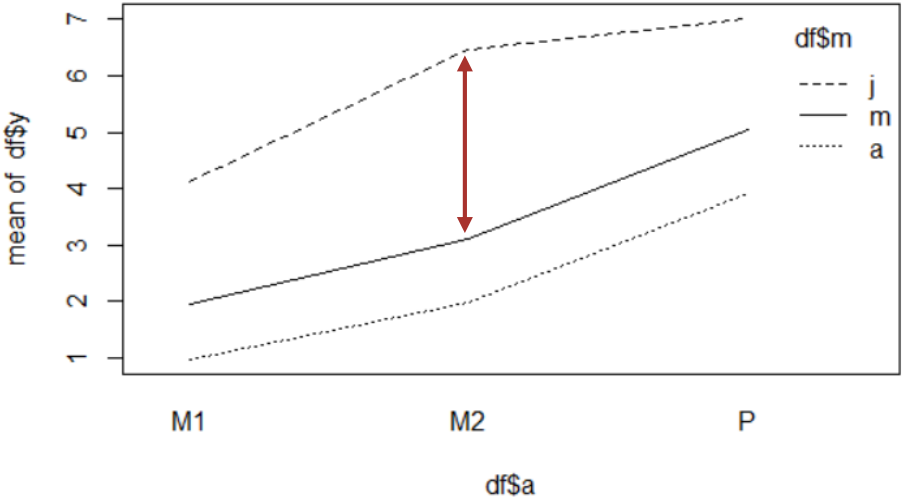
$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$
Linien nicht parallel

Mehr als zwei Faktorstufen (Bsp: Empfinden nach Schmerzmittel)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
a	2	1436.2	718.1	670.657	<2e-16
m	2	1486.2	743.1	693.971	<2e-16
a:m	4	6.7	1.7	1.567	0.181
Residuals	891	954.1	1.1		



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
a	2	1347.8	673.9	705.24	<2e-16
m	2	2034.8	1017.4	1064.69	<2e-16
a:m	4	79.9	20.0	20.91	<2e-16
Residuals	891	851.4	1.0		



R

- Ab Zeile 307