Mathematische und statistische Methoden I

Dr. Malte Persike

persike@uni-mainz.de
lordsofthebortz.de
twitter.com/methodenlehre
tinyurl.com/gplusmethodenlehre

Sprechstunde jederzeit nach Vereinbarung und nach der Vorlesung

Wallstr. 3, 6. Stock, Raum 06-206

WiSe 2011/2012

Fachbereich Sozialwissenschaften
Psychologisches Institut
Johannes Gutenberg Universität Mainz

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Partialkorrelation

Deutungsmöglichkeiten der einfachen Regression

1. Zufall

2. Direkte Kausalität: $x \rightarrow y$

3. Kausale Drittvariable(n) $x \rightarrow y_1,\; x \rightarrow y_2$

4. Direkte und indirekte Kausalität $x \rightarrow y_1,\; x \rightarrow y_2,\; y_1 \rightarrow y_2$

Partialkor-
relation

Semipartial-
korrelation

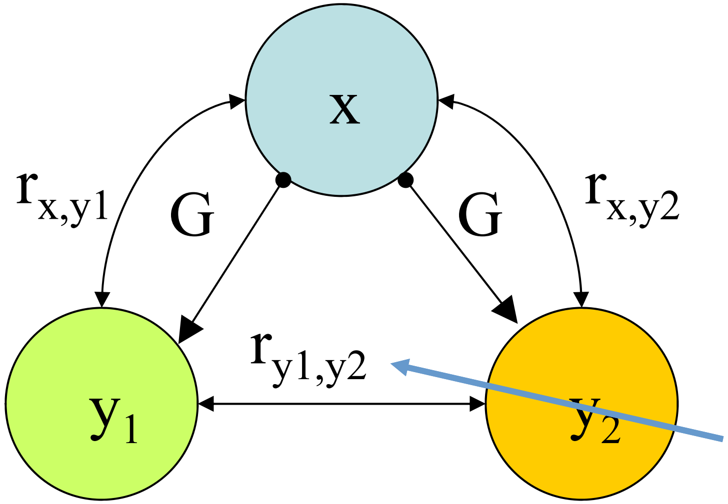
Multiple Parti-
alkorrelation

Partialkorrelation
im Fall zweier korrelierter Variablen

- Definition: Eine **Partialkorrelation** ist die Korrelation zweier Variablen, die vom Effekt anderer Variablen bereinigt wurden.

- Einsatzzweck: Prüfung einer **Kausalvermutung G**

  „Kommt r_{y1y2} dadurch zustande, dass eine Drittvariable x ursächlich auf y_1 und y_2 einwirkt?“



„Schein-
korrelation“

Partialkor-relation

Semipartial-korrelation

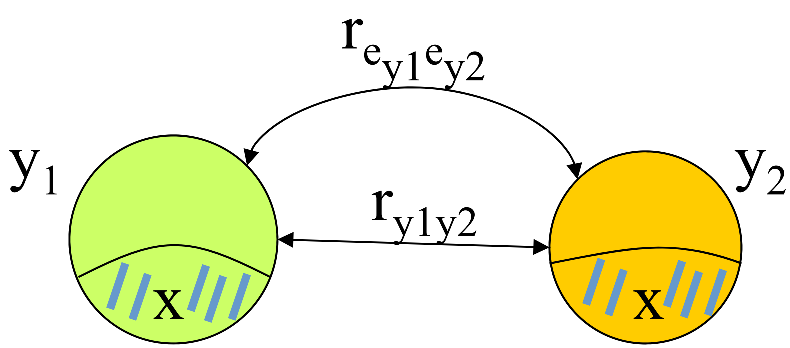
Multiple Parti-alkorrelation

Partialkorrelation

Berechnung und Prüfung

1. Sage y_1 aus x voraus und berechne Residuen e_{y1}
2. Sage y_2 aus x voraus und berechne Residuen e_{y2}
3. Berechne die Korrelation $r_{e_{y1}e_{y2}}$ Schreibe: $r_{y1y2 \cdot x}$

„ohne“



Ist Partialkorrelation **nahe Null**, so beruht die Korrelation r_{y1y2} tatsächlich vor allem auf der Einwirkung von x.

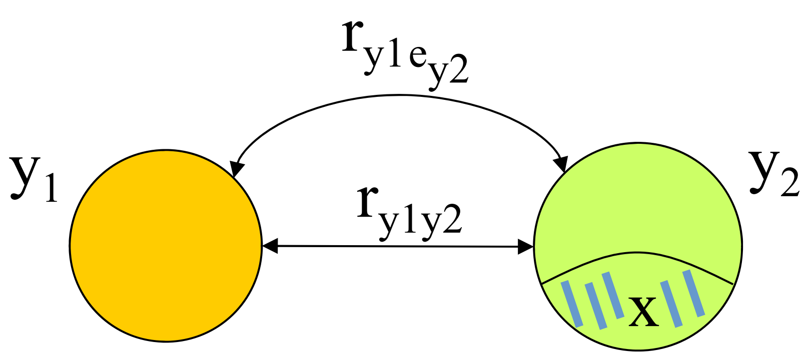
(Prüfung mit Korrelationstest)

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Semipartialkorrelation

im Fall zweier korrelierter Variablen

- Definition: Eine **Semipartialkorrelation** ist die Korrelation zweier Variablen, von denen *eine* vom Effekt einer anderen Variablen bereinigt wurden.
- Einsatzzweck: Prüfung der zusätzlichen Information eines Prädiktors bei der **Erklärung des Kriteriums**
- Die Semipartialkorrelation ist eng verbunden mit der **Nützlichkeit**. Es gilt nämlich $U_{x1} = r^2_{y(x1 \cdot x2)}$



Semipartialkorrelation

Berechnung

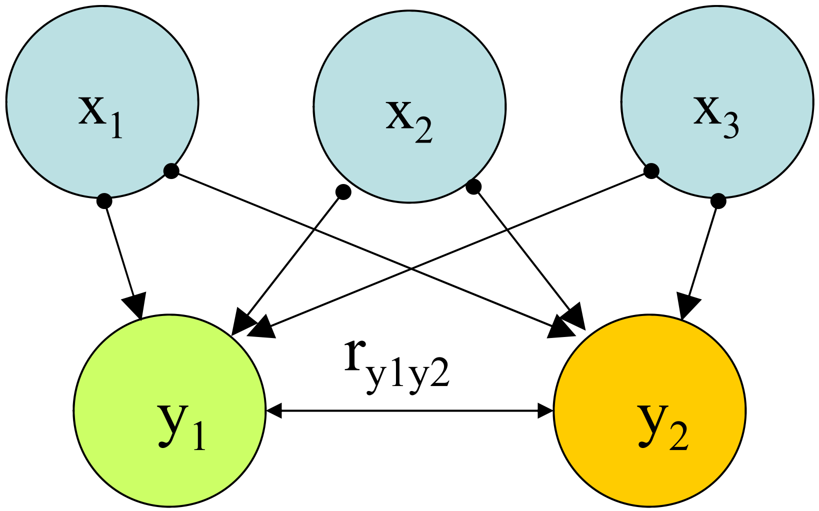
1. Sage y_2 aus x voraus und berechne Residuen e_{y2}
2. Berechne die Korrelation $r_{y_1 e_{y2}}$ Schreibe: $r_{y1(y2 \cdot x)}$
   (analog für Auspartialisierung von x aus $y1$)

„ohne"

(Semi-)Partialkorrelation höherer Ordnung

Prinzip

- Soll der Zusammenhang zwischen zwei Variablen um **mehrere andere Variablen** bereinigt werden, spricht man von (Semi-)Partialkorrelationen höherer Ordnung
- Die Berechnung verläuft analog zu den (Semi-)Partial-korrelationen bei nur einer auszupartialisierenden Variable, nur mithilfe der Multiplen Regression



Partialkor-relation

Semipartial-korrelation

Multiple Parti-alkorrelation

(Semi-)Partialkorrelation höherer Ordnung

Berechnung über multiple Regression

1. Sage y_1 aus den $x_1 \ldots x_k$ voraus und berechne Residuen e_{y1}
2. Sage y_2 aus den $x_1 \ldots x_k$ und berechne Residuen e_{y2}
3. Berechne die Korrelation $r_{e_{y1}e_{y2}}$ $\rightarrow$ $r_{y1y2 \cdot x1 \ldots xk}$
   (Partialkorrelation)

   oder

   Berechne die Korrelation $r_{y_1\, e_{y2}}$ $\rightarrow$ $r_{y1(y2 \cdot x1 \ldots xk)}$
   (Semipartialkorrelation)

Varianzanalyse

Ein- / Mehrfaktorielle ANOVA (ANalysis Of VAriance)

- **Ziel**: Analyse des Einflusses **unabhängiger Variablen** (UVn) auf eine **abhängige Variable** (AV)
- Während bei der Regression der Fokus auf den **Zusammenhängen** zwischen UVn und AV liegt, soll die ANOVA gemessene **Unterschiede** in der AV erklären.
- **Beispiele:** Wie gut wirken unterschiedliche Therapieformen bei spezifischen Phobien bei verschiedenen Geschlechtern? Wie verändert sich die Schulleistung von Kindern über mehrere Nachhilfesitzungen hinweg?
- Generell versucht die ANOVA immer, Unterschiede **zwischen Gruppen** in der Höhe der AV aufzuzeigen.

Einführung

Lineares Modell

Voraus-
setzungen

Varianzanalyse

Ein- / Mehrfaktorielle ANOVA (ANalysis Of VAriance)

- Die AV muss dabei **stetig** sein (intervallskaliert) die UVn sind i.d.R **nominal-** oder **ordinalskaliert**
- Die UVn werden im Rahmen der ANOVA auch als **Treatments** oder **Faktoren** bezeichnet, die Ausprägungen eines Treatments als **Treatmentstufen** oder **Faktorstufen**
- Nach der Anzahl der Treatments unterscheidet man die einfaktorielle, zweifaktorielle oder allgemein mehrfaktorielle ANOVA

Varianzanalyse

Ein- / Mehrfaktorielle ANOVA (ANalysis Of VAriance)

- **Beispiel:** Experiment mit einem 3stufigen Treatment A
- Man habe an 3 Personengruppen (definiert durch die Treatmentstufen A_1 bis A_3) die Werte einer AV erhoben.

Vp-Nr.	A_1	A_2	A_3
1	9	6	10
2	14	8	11
3	12	13	7
…	…	…	…
n-1	17	19	14
n	15	11	12

Varianzanalyse

Ein- / Mehrfaktorielle ANOVA (ANalysis Of VAriance)

- **Beispiel:** Nun nehme man das Geschlecht als „Treatment" B mit zwei Stufen hinzu

- Man hat nun 6 Personengruppen (oder „Zellen")

	Vp-Nr.	A_1	A_2	A_3
B_1	1	9	6	10
	…	…	…	…
	n_1	12	13	7
B_2	1	17	19	14
	…	…	…	…
	n_2	15	11	12

Varianzanalyse

Ein- / Mehrfaktorielle ANOVA (ANalysis Of VAriance)

- Die abhängige Variable sei X genannt.
- Mit der korrekten Indizierung lässt sich die Tabelle dann schreiben als

	Vp-Nr.	A_1	A_2	A_3
B_1	1	$x_{1,1,1} = 9$	$x_{1,2,1} = 6$	$x_{1,3,1} = 10$
	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	n_1	$x_{n_1,1,1} = 12$	$x_{n_1,2,1} = 13$	$x_{n_1,3,1} = 7$
B_2	1	$x_{1,1,2} = 17$	$x_{1,2,2} = 19$	$x_{1,3,2} = 14$
	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	n_2	$x_{n_2,1,2} = 15$	$x_{n_2,2,2} = 11$	$x_{n_2,3,2} = 12$

Einführung

Lineares Modell

Voraus-setzungen

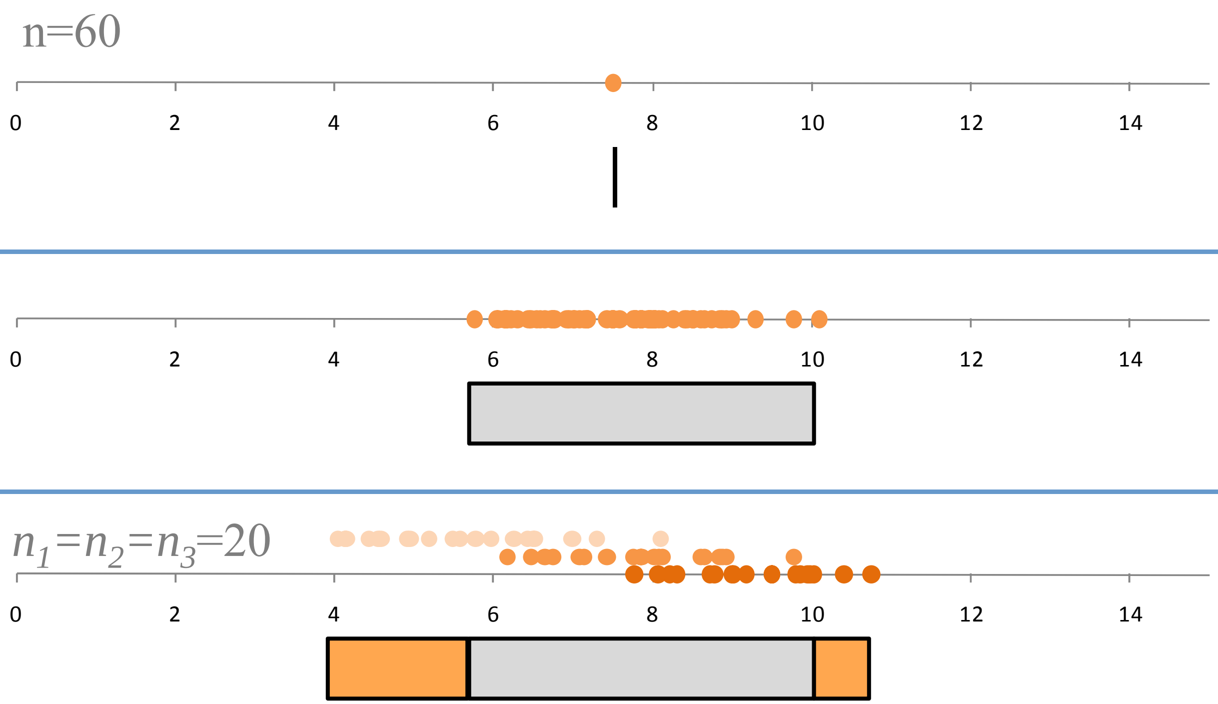
Varianzanalyse

Ein- / Mehrfaktorielle ANOVA (ANalysis Of VAriance)

- Die ANOVA geht davon aus, dass am Zustandekommen jedes Messwertes $x_{i,j,k}$ ($i = 1 \ldots n$, $j = 1 \ldots p$, $k = 1 \ldots q$ mit $p=3$, $q=2$) mehrere Komponenten beteiligt sind:

    1. **Populationswert**, von allen Personen geteilt
       Dieser ist für alle Personen konstant
    2. **Effekt der Treatmentstufe** j des Treatments A
       Dieser ist für jede Person in Gruppe A_j konstant
    3. **Effekt der Treatmentstufe** k des Treatments B
       Dieser ist für jede Person in Gruppe B_k konstant
    4. **Interaktionseffekt** jeder spezifischen Kombination von A_j und B_k
       Dieser ist für jede Person in Zelle $A_j\,B_k$ konstant
    5. **Zufallsfehler**, der Unterschiede zwischen Personen ausmacht, die nicht auf A oder B zurückgehen
       Dieser ist individuell für jede Person

- Wie die Regression geht die ANOVA von einem einfachen **linearen Modell** aus.
- Sie nimmt an, dass der Messwert einer Person auf einer beliebigen AV additiv aus **systematischen Komponenten** und einer **Fehlerkomponente** besteht



$$x = \mu$$

Von allen geteilter Populationswert

= Streuung von μ

$$x = \mu + e$$

plus Fehler

= Fehlerstreuung

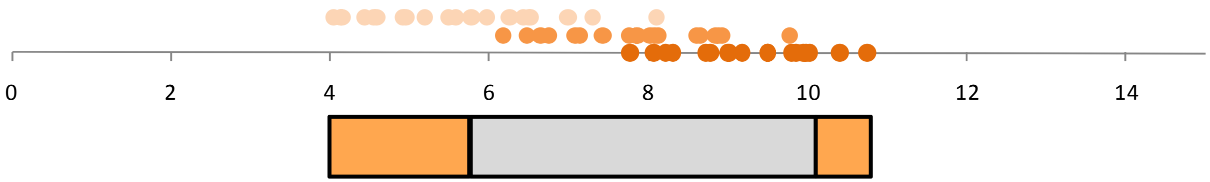
$$x = \mu + a + e$$

plus Effekt A

= Treatment -streuung A

- Wie die Regression geht die ANOVA von einem einfachen **linearen Modell** aus.

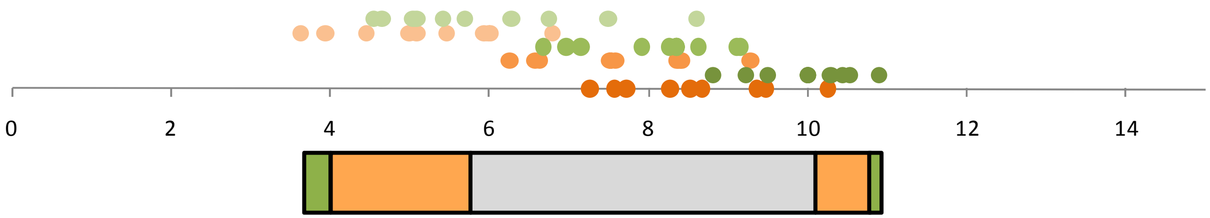
- Sie nimmt an, dass der Messwert einer Person auf einer beliebigen AV additiv aus **systematischen Komponenten** und einer **Fehlerkomponente** besteht



$$x = \mu + a + e$$

plus Effekt A

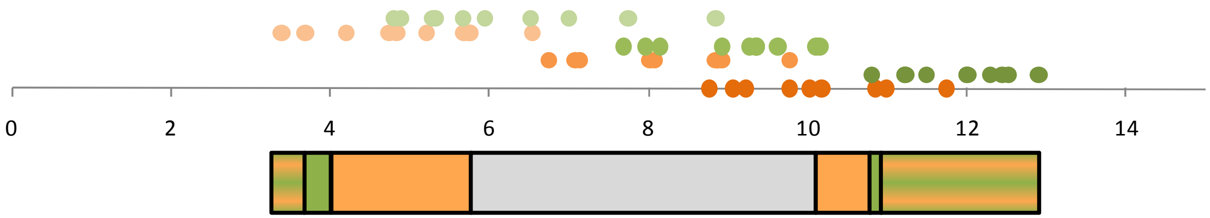
= Treatment -streuung A



$$x = \mu + a + b + e$$

plus Effekt B

= Treatment -streuung B



$$x = \mu + a + b + a \times b + e$$

= Treatment -streuung AB

plus Effekt AB

Einführung

Lineares Modell

Voraus-setzungen

Varianzanalyse

Ein- / Mehrfaktorielle ANOVA (ANalysis Of VAriance)

- Der einfaktoriellen Varianzanalyse liegt also formal ein sehr einfaches lineares Modell zugrunde

$$\boxed{x_{ijk} = \mu + a_j + b_k + a_j b_k + e_{ij}}$$

x_{ijk} = Messwert der Person i unter Treatmentstufen j und k

μ = Populationswert aller $\sum n_{jk} = N$ Personen

a_j = Effekt der Treatmentstufe j

b_k = Effekt der Treatmentstufe k

$a_j b_k$ = Interaktionseffekt der Treatmentstufen j und k

e_{ijk} = Fehler der Person i unter Treatmentstufen j und k

- **Merke:** Der Messwert einer beliebigen aller N Person setzt sich zusammen aus einem **Populationswert**, den **Treatmenteffekten** & einem individuellen **Messfehler**.

Varianzanalyse

Einfaktorielle ANOVA (ANalysis Of VAriance)

- Der beobachtete Messwert jeder Person ist also in drei Komponenten **zerlegt**
- Damit lässt sich die Tabelle schreiben als

	Vp-Nr.	A_1	…	A_p
B_1	1	$a_1+b_1+a_1b_1+e_{1,1,1}$	…	$a_p+b_1+a_pb_1+e_{1,p,1}$
	…	…	…	…
	n_1	$a_1+b_1+a_1b_1+e_{n1,1,1}$	…	$a_p+b_1+a_pb_1+e_{n1,p,1}$
B_q	1	$a_1+b_q+a_1b_q+e_{1,1,q}$	…	$a_p+b_q+a_pb_q+e_{1,p,q}$
	…	…	…	…
	n_2	$a_1+b_q+a_1b_q+e_{n2,1,q}$	…	$a_p+b_q+a_pb_q+e_{n2,p,q}$

Varianzanalyse

Einfaktorielle ANOVA (ANalysis Of VAriance)

- Die zentrale **Forschungsfrage**, die von der ANOVA beantwortet werden soll, lautet nun:

Gilt $\boxed{a_j = a_{j'}}$ für alle

oder $\boxed{a_j \neq a_{j'}}$ für mindestens eine

der paarweisen Kombinationen von j und j'.

- **In Worten:** Gibt es mindestens eine Treatmentstufe von A, die auf den Messwert der Versuchspersonen anders wirkt als die übrigen Treatmentstufen von A?

Varianzanalyse

Einfaktorielle ANOVA (ANalysis Of VAriance)

- Die zentrale **Forschungsfrage**, die von der ANOVA beantwortet werden soll, lautet nun:

  Gilt $\boxed{b_k = b_{k'}}$ für alle

  oder $\boxed{b_k \neq b_{k'}}$ für mindestens eine

  der paarweisen Kombinationen von k und k'.

- **In Worten:** Gibt es mindestens eine Treatmentstufe von B, die auf den Messwert der Versuchspersonen anders wirkt als die übrigen Treatmentstufen von B?

Einführung

Lineares Modell

Voraus-
setzungen

Varianzanalyse

Einfaktorielle ANOVA (ANalysis Of VAriance)

- Die zentrale **Forschungsfrage**, die von der ANOVA beantwortet werden soll, lautet nun:

Gilt $\boxed{a_j b_k = a_{j'} b_{k'}}$ für alle

oder $\boxed{a_j b_k \neq a_{j'} b_{k'}}$ für mindestens

eine der paarweisen Kombinationen von jk & $j'k'$.

- **In Worten:** Gibt es mindestens eine Interaktion von A und B, die auf den Messwert der Versuchspersonen anders wirkt als die übrigen Kombinationen von A und B?

Varianzanalyse

Ein- / Mehrfaktorielle ANOVA - Annahmen

- Die Beobachtungen müssen **unabhängig** sein, d.h. sie müssen von verschiedenen Merkmalsträgern stammen
- Die **Fehler** in jeder Treatmentstufe sollten **normalverteilt** sein mit einem erwarteten Mittelwert von 0.

  $$\boxed{\bar{e}_{jk} = 0}$$

- Die Fehlervarianzen in jeder der p Treatmentstufen sollen erwartet (nicht numerisch) gleich sein (**Homoskedastizität**).

  $$\boxed{s^2_{e_{1,1}} = s^2_{e_{1,2}} = \ldots = s^2_{e_{p,q}}}$$

- Treatmenteffekt und Fehler müssen **additiv** sein, d.h. die Fehler dürfen nicht mit den Erwartungswerten der Treatmentstufen korrelieren.

Varianzanalyse

Einfaktorielle ANOVA – Prüfung der Annahmen

- Die **Unabhängigkeit der Beobachtungen** wird i.d.R. „begründet angenommen“ (educated guess).
- Zur Prüfung der **Normalverteilungsannahme** der Fehler (i.e. Zellresiduen) wird der Kolmogoroff-Smirnov Test verwendet
- Zur Prüfung der **Homogenität der Fehlervarianzen** wird zumeist der Bartlett Test, seltener der Levene Test bzw. der F-Test verwendet.
- Die **Unabhängigkeit der Treatmenteffekte und Messfehler** kann über einen Korrelationstest von Zellmittelwerten und Varianzen geprüft werden (selten praktiziert)