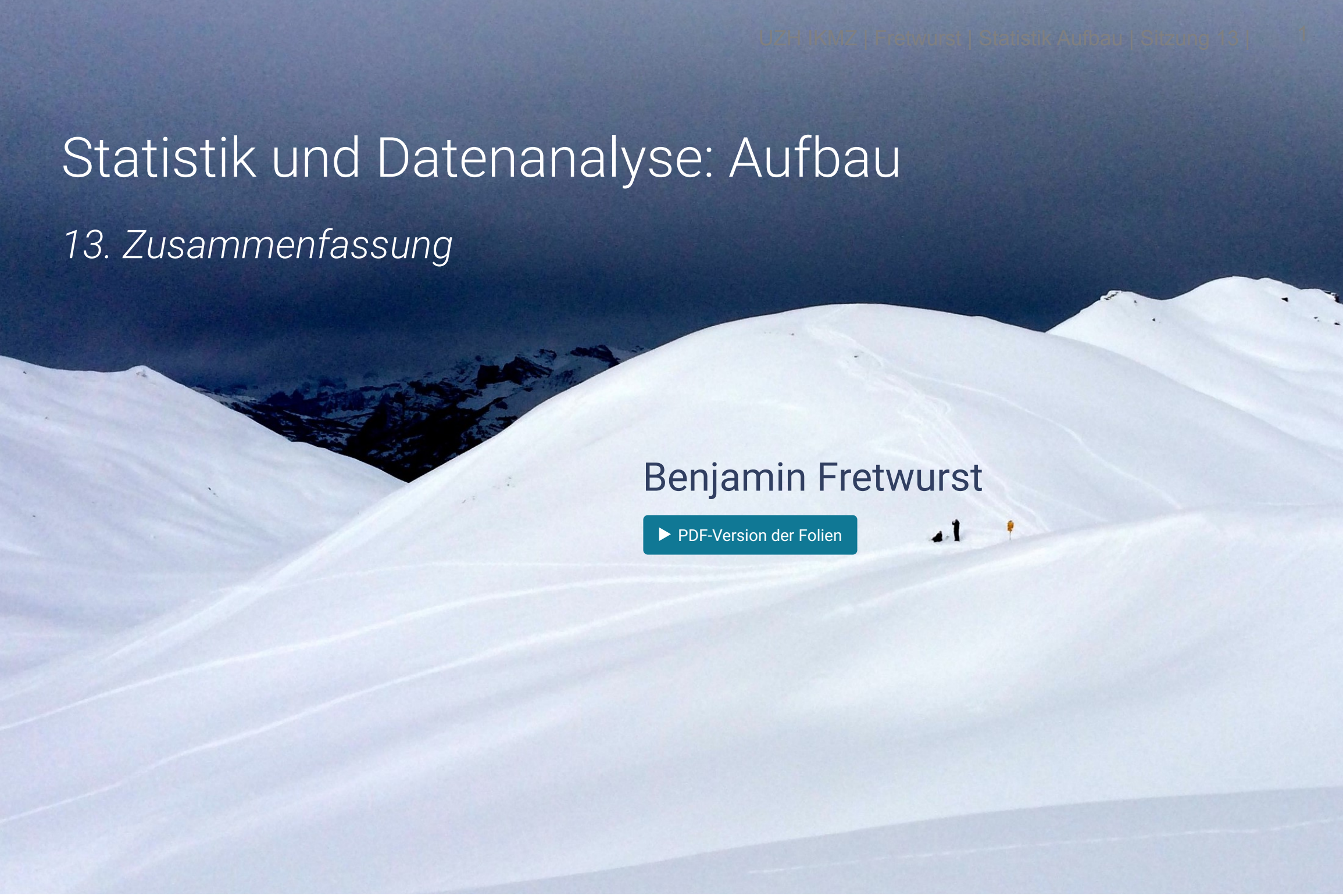
Statistik und Datenanalyse: Aufbau

13. Zusammenfassung



Benjamin Fretwurst

▶ PDF-Version der Folien

Inhalt

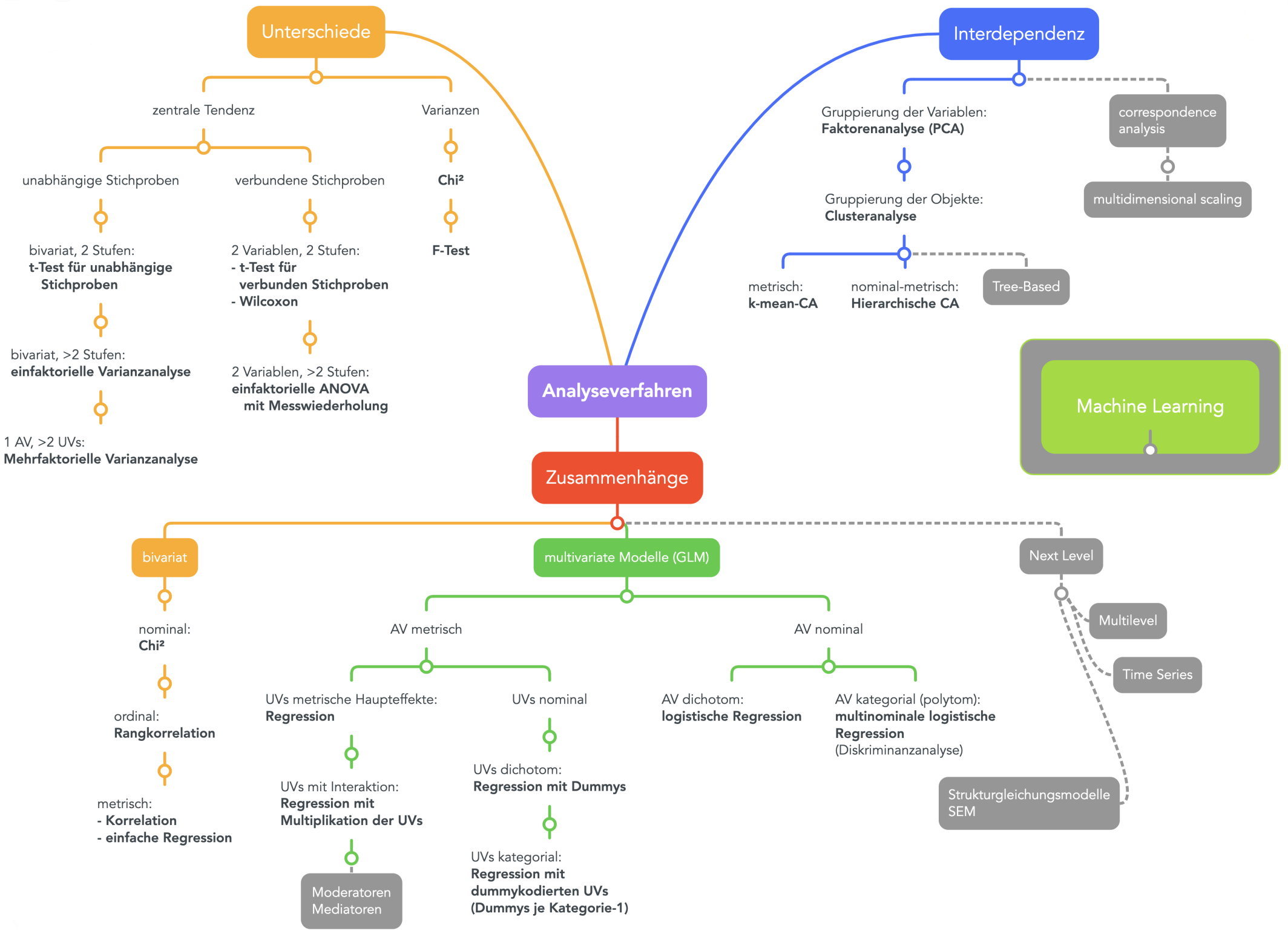
Orga

Photo: Lisa Vogel

Lernziele

Zusammenfassung

- Regression und ihre Voraussetzungen
- Kategoriale UV und Interaktion
- Logistische Regression
- Faktorenanalyse
- Clusteranalyse


Unterschiede
zentrale Tendenz
Varianzen
unabhängige Stichproben
verbundene Stichproben
Chi²
bivariat, 2 Stufen:
t-Test für unabhängige Stichproben
2 Variablen, 2 Stufen:
- t-Test für verbunden Stichproben
- Wilcoxon
F-Test
bivariat, >2 Stufen:
einfaktorielle Varianzanalyse
2 Variablen, >2 Stufen:
einfaktorielle ANOVA mit Messwiederholung
1 AV, >2 UVs:
Mehrfaktorielle Varianzanalyse
Interdependenz
Gruppierung der Variablen:
Faktorenanalyse (PCA)
correspondence analysis
multidimensional scaling
Gruppierung der Objekte:
Clusteranalyse
metrisch:
k-mean-CA
nominal-metrisch:
Hierarchische CA
Tree-Based
Machine Learning
Analyseverfahren
Zusammenhänge
bivariat
multivariate Modelle (GLM)
Next Level
nominal:
Chi²
ordinal:
Rangkorrelation
metrisch:
- Korrelation
- einfache Regression
AV metrisch
AV nominal
Multilevel
Time Series
UVs metrische Haupteffekte:
Regression
UVs nominal
AV dichotom:
logistische Regression
AV kategorial (polytom):
multinominale logistische Regression
(Diskriminanzanalyse)
UVs mit Interaktion:
Regression mit Multiplikation der UVs
UVs dichotom:
Regression mit Dummys
Strukturgleichungsmodelle SEM
Moderatoren
Mediatoren
UVs kategorial:
Regression mit dummykodierten UVs
(Dummys je Kategorie-1)

1 Bivariate Statistik

Kovarianz einer Variablen mit sich selbst ist deren Varianz. Die Kovarianz zweiter z-transformierter Variablen ist die Korrelation r.

$$cov = \frac{1}{n}\sum_{i}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Table 1: Varianz-Kovarianz-Tabelle Persönlichkeitsskala

	PS01_01	PS01_02	PS01_03	PS01_04	PS01_05
PS01_01	1.17	0.89	0.97	0.9	0.96
PS01_02	0.89	1.38	0.94	0.91	0.88
PS01_03	0.97	0.94	1.55	1.26	1.16
PS01_04	0.9	0.91	1.26	1.49	1.15
PS01_05	0.96	0.88	1.16	1.15	1.55

Was steht in der Diagonalen? Was, wenn es Korrelationen wären?

2 Regression – bivariat

Die Idee vom Modell

Modellidee

Für das Ergebnis der Datenerhebung wird ein Modell entworfen, das Zusammenhänge einfach darstellt. Da das Modell nie zu 100% das Ergebnis treffen wird, bleibt ein Rest, den wir Modellfehler oder einfach Fehler nennen.

Grundmodell

Ergebnis = (Modell) + Fehler

Beispiel

Mittelwert von $x_i = \overline{x} + Fehler_i$ (die Abweichungen vom Mittelwert)

2.1 Die bivariate Regressionsgleichung

$$\text{GG: } Y_i = \beta_1 + \beta_2 X_{i2} + U_i$$
$$\text{Stichprobe: } Y_i = b_1 + b_2 X_{i2} + e_i$$

Modell und Schätzung

Das Regressionsmodell für die Zusammenhänge in der GG wird durch die Berechnung der b's in der Stichprobe geschätzt. Die «Variablen» (Y und X) sind fix. Es bleiben nur die b's zu schätzen, von denen die Lage der Regressionsgerade abhängt und damit die Fehler (Errors) aka Residuen e_i. Subtrahiert man in der Formel oben $b_1 + b_2 X_{i2}$, erhält man:
$e_i = Y_i - (b_1 - b_2 X_{i2})$, woraus sich b_1 und b_2 ableiten lassen:

$$b_1 = \overline{Y} - b_2 \overline{X}_2$$
$$b_2 = r_{Y2} \frac{S_Y}{S_2}$$

2.2 Der Regressionskoeffizient b

Regressionskoeffizient b (aka Steigungskoeffizient) und r

$$b = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

$$r_{YX} = b \cdot \frac{s_X}{s_Y}$$

2.3 Der standardisierte Regressionskoeffizient

Der standardisierte Regressionskoeffizient BETA aka b*

$$BETA = b \cdot \frac{s_X}{s_Y} = r_{YX}$$

Beschreibung

Die standardisierten Regressionskoffizienten geben einen Zusammenhang in Standardabweichungen an: Wenn x um eine Standardabweichung grösser ist, um wie viele Standardabweichungen ist dann y grösser (oder kleiner, wenn negativ)?

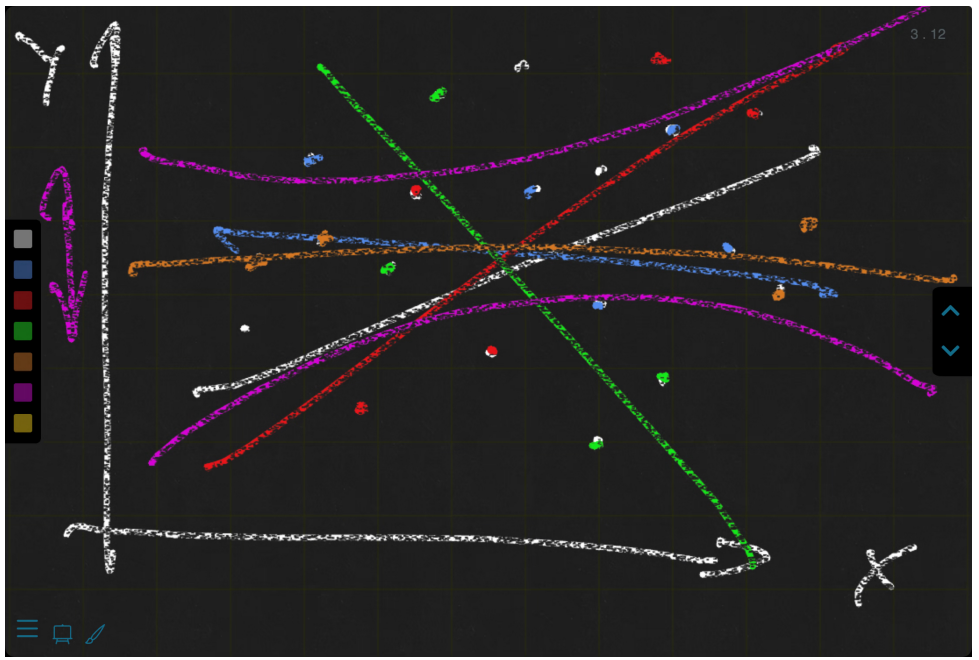
Wie Korrelationen bzw. partielle Korrelationen

Die BETAs sind den Korrelationen sehr ähnlich: +1 ist ein perfekter positiver Zusammenhang, 0 kein Zusammenhang und -1 ein perfekter negativer Zusammenhang. Interpretieren würde ich ab 0.1, wenn sie signifikant sind.

Standardfehler der b’s

$$se_b^2 = \frac{s_e^2}{n \cdot s_2^2} \text{ mit } s_e^2 = \frac{1}{n-3} \sum e_i^2$$

Die Standardfehler der b’s sind (bei sehr vielen Ziehungen) die «durchschnittliche» Abweichung der b’s von dem wahren Wert β.
Standardfehler kann man auch für die standardisierten Regressionskoeffizienten (BETA) berechnen.



Streuung von Regressionsgeraden

Varianzzerlegung der Gesamtmodellgüte R^2

$$Y_i = \overline{Y} + e_i$$

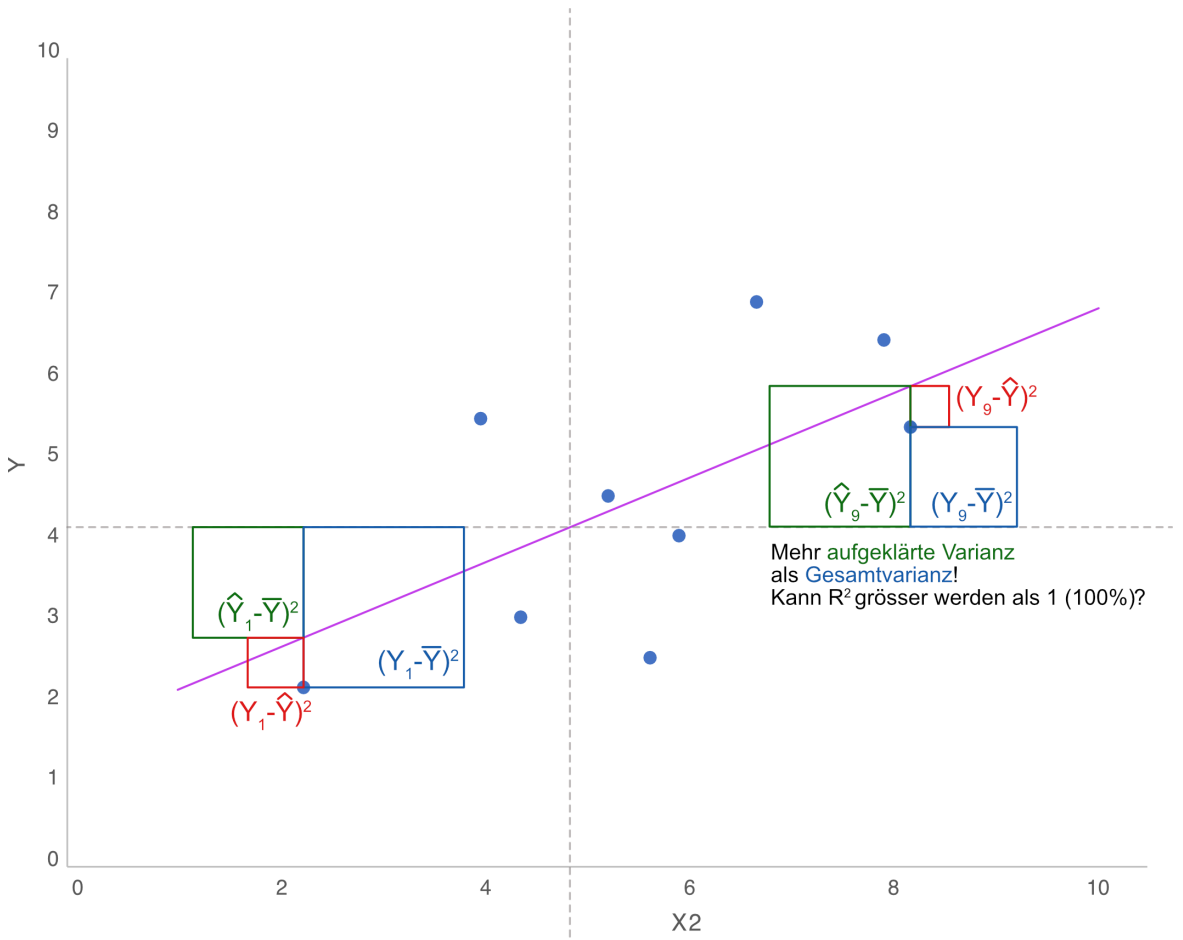
$$Y_i = b_1 + b_2 X_i + e_i$$

$$\hat{Y}_i = b_1 + b_2 X_i$$

$$Y_i = \hat{Y}_i + e_i$$

$$SS_T = SS_R + SS_M$$

$$R^2 = \frac{SS_M}{SS_T}$$



2.4 Das Bestimmtheitsmass R^2

R^2 ist Varianzaufklärung

Das Bestimmtheitsmass R^2 gibt an, wie viel Variant der AV durch die UV's aufgeklärt werden konnte. R^2 geht von 0 bis 1, bzw., wenn in Prozent ausgedrückt, von 0% bis 100%.

$R^2 = \frac{SS_M}{SS_T}$

- Anteil der erklärten Varianz der AV durch die UVs.
- SS_T: Summe der quadrierten Abweichungen **T**otal für die AV (Y).
- SS_M: Summe der quadrierten Abweichungen des **M**odells ($\hat{\mathbf{Y}}$)
- $R^2 = \frac{\text{aufgeklärte Varianz}}{\text{Gesamtvarianz}}$

F-Test (R^2)

Gibt an, ob durch das Modell insgesamt überzufällig gut Varianz aufgeklärt wurde. Also, ob die Nullhypothese zurückgewiesen werden kann, dass die AV nicht durch sämtliche UVs zusammen (oder mind. eine UV) im Modell erklärt werden kann.

$R^2_{adj.}$

Formel für das korrgigierte $R^2_{adj.}$

$R^2_{adj.} = R^2 \cdot \frac{n-k-1}{n-1}$ bei kleinen Stichproben (wobei k die Anzahl UVs ist).

Funktion und Anwendung des $R^2_{adj.}$

Das $R^2_{adj.}$ korrigiert (adjustiert) für die Anzahl der UVs im Modell im Verhältnis zum Stichprobenumfang, um zu verhindern, dass zu viel «zufällig» aufgeklärte Varianz überinterpretiert wird. Bei grösseren Stichproben ($\frac{n-k-1}{n-1}$ ist schon bei 101 Fällen und 10 UVs mit 0.9 fast eins) mit mehreren hundert Fällen spielt die Korrektur keine Rolle mehr und man nehme ruhig und besser das normale R^2.

Kennwerte von Regressionsanalysen – Signifikanz

t-Werte der b's oder standardisierten Regressionskoeffizienten (BETA)

Umrechnung der b's in t-Werte, die sich (bei gegebenem Stichprobenumfang bzw. den Degrees of Freedom) unter der Annahme der Nullhypothese ergeben. Sie sind innerhalb einer Regressionsanalyse vergleichbar. Sie sind für die b's und BETAS identisch.

p-Werte der b's bzw. BETAS (p oder sig.)

Geben die Wahrscheinlichkeit an, dass ein in einer Stichprobe gefundenes b zustandekommt, obwohl die Nullhypothese gilt. Ist auch für die b's und BETAS identisch. Bei $p < .05$ sprechen wir von einem von 0 signifikant verschiedenen b, wenn das Signifikanzniveau bei 95% liegt (5% Irrtumswahrscheinlichkeit).

2.5 Beispieloutput einer Regression

```r
```{r Regression}
Regression <- lm(Mischief ~ Cloak, data = Invisibility)

olsrr::ols_regress(Regression)
```
```

```
                         Model Summary
---------------------------------------------------------------
R                        0.343       RMSE                1.787
R-Squared                0.118       Coef. Var          40.845
Adj. R-Squared           0.078       MSE                 3.193
Pred R-Squared          -0.050       MAE                 1.333
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
-----------------------------------------------------------------
               Sum of
              Squares        DF    Mean Square      F        Sig.
-----------------------------------------------------------------
Regression      9.375         1          9.375    2.936    0.1007
Residual       70.250        22          3.193
Total          79.625        23  ← die Fallzah n
-----------------------------------------------------------------

                                 Parameter Estimates                                        CI
-------------------------------------------------------------------------------------------------
      model      ~~Beta~~ B   Std. Error    Std. Beta      t       Sig      lower     upper
-------------------------------------------------------------------------------------------------
(Intercept)     3.750          0.516                 7.270    0.000     2.680     4.820
      Cloak     1.250          0.730         0.343   1.713    0.101    -0.263     2.763
-------------------------------------------------------------------------------------------------
```

3 Regression multivariat

Regressionsgleichung mit 2 UVs

$$\text{GG: } Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i$$
$$\text{n: } Y_i = b_1 + b_2 X_{i2} + b_3 X_{i3} + e_i$$

Regressionsgleichung mit k UVs

$$\text{GG: } Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} \ldots + \beta_k X_{ik} + U_i$$
$$\text{n: } Y_i = b_1 + b_2 X_{i2} + b_3 X_{i3} \ldots + b_k X_{ik} + e_i$$

Transformationen

Die X_i bis X_k können direkte lineare UVs sein, oder durch Lineartransformation linearisierte UVs bei kurvilinearen Zusammenhängen.

Gängige Lineartransformationen sind die Quadrierung einer UV (mit vorheriger Zentrierung), reziproke Transformation (1/X), logarithmische Transformation (log(x)) oder exponentielle Transformation (log(y) der AV).

3.1 OLS

Grundidee OLS

Wir suchen die b's. Die gesuchten b's sollen eine Regressionsgerade ergeben, die «optimal» in der Punktwolke der gemessenen Werte liegt. Wir suchen also die b's, die die kleinsten quadrierten Abweichungen zwischen den vom Modell vorhergesagten und den gemessenen Werten ergibt. Das «Prinzip der kleinsten Quadrate» wird als OLS bezeichnet (Ordenary Least Squares).

$$\sum_{i=1}^{n} e_i^2 \rightarrow minimal$$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \rightarrow minimal$$

3.2 Die Formel für b nach OLS

$$b_2 = (r_{Y2} - r_{23}r_{Y3})\frac{1}{1 - R^2_{2.3}}\frac{S_y}{S_2}$$

In Worten

Der Anstieg der «Regressionsgeraden» für X_2 ergibt sich aus der Korrelation r_{Y2} zwischen X_2 und Y, die um den vermittelten Zusammenhang über die Drittvariable, also das Produkt aus r_{23} und r_{Y3} reduziert wird. Der Rest sind Korrekturen damit, wie stark X_2 von den übrigen Variablen erklärt wird $\frac{1}{1-R^2_{2.3}}$ und quasi die Umkehr der Standardisierung $\frac{S_Y}{S_2}$.

4 BLUE – Voraussetzungen von OLS

«Linear Estimator»

Die «Linear Estimator» sind die b's, also b_1, b_2, … b_k. Geschätzt wird linear, aber die UVs können transformiert sein.

«Unbiased»

«Unbiased» bedeutet, dass wir unverzerrte Schätzer, also unverzerrte b's haben wollen. Die b's scätzen ihre βs unverzerrt, wenn die Streuung der b's um die wahren βs herum liegen (man sagt auch: «erwartungstreu»).

«Best» bezeichnet die Effizienz der Schätzer b

Die besten Schätzer erhalten wir, wenn die Standardfehler der b's (se_b) minimal sind.

«Unbiased», also Unverzerrtheit der b's

Die Variablen (X und Y) müssen fix sein

Wir müssen also davon ausgehen, dass die erhobenen Variablen bei einer nächsten Ziehung nicht ganz anders aussehen würden.

im U_i darf keine relevante Einflussgrösse vergessen worden sein

Es darf im Unbekannten U_i keine Variable stecken, die mit den UV's korreliert. Der Erwartungswert dieser Covarianz muss 0 sein: $E(C_{2U}) = 0 = E(C_{3U})$.

Modellspezifikation

Wir sollten aus der Theorie und in der Operationalisierung keine Variable vergessen, die mit den UVs zusammenhängt! Theoriearbeit besteht in der Suche nach der vollen Modellspezifikation! Die perfekte Modellspezifikation wäre das Ende der Forschung zu einem fixen Phänomen.

4.1 Multikollinearität – Grund und Problem multipler Regression

Definition

Multikollinearität bedeutet, dass die Varianz einer Variablen durch eine oder mehrere übrige UVs teilweise aufgeklärt wird.

herausgerechnete Erklärungskraft

Wird einer Variablen viel Erklärungsvarianz ($R_{2.34...}$) weggerechnet, dann hat sie kaum noch welche, um die AV zu erklären.

Wann ein Problem

- Der Grund für Regressionsanalysen
- Problem hoher Multikollinearität (TOL < .5)
- Standardfehler ➪ Schätzqualität schlecht (VIF > 2)

Steigende Fehlerstreuung bei Multikollinearität

Fehlervarianz von b_2

$$s^2_{b_2} = \frac{s^2_e}{n} \cdot \frac{1}{V_2} \cdot \frac{1}{1-R^2_{2.34...}}$$

Die Fehlerstreuung des Regressionskoeffizienten b ist proportional zur Streuung der Fehler e_i und umgekehrt proportional zur Fallzahl n sowie zur Varianz V_2 (bzw. s^2_2 von X_2) und zu Multikollinearität bzw. Toleranz ***TOL*** = $1 - R^2_{2.34...}$.

Toleranz ist die exklusive Varianz einer UV

$$TOL_{b_2} = 1 - R^2_{2.34...}$$

Toleranz ist der Prozentsatz Varianz, der nicht durch die übrigen UVs rausgerechnet wird.

Der Varianz-Inflation-Factor VIF

$$VIF_{b_2} = \frac{1}{(1-R^2_{2.34...})} = \frac{1}{TOL_{b_2}}$$

4.2 Linearität

Wenn keine Linearität vorliegt, werden die Effekte nicht richtig geschätzt. Häufig sind die Residuen nicht homoskedastisch. Lösen kann man das Problem durch Lineartransformation.

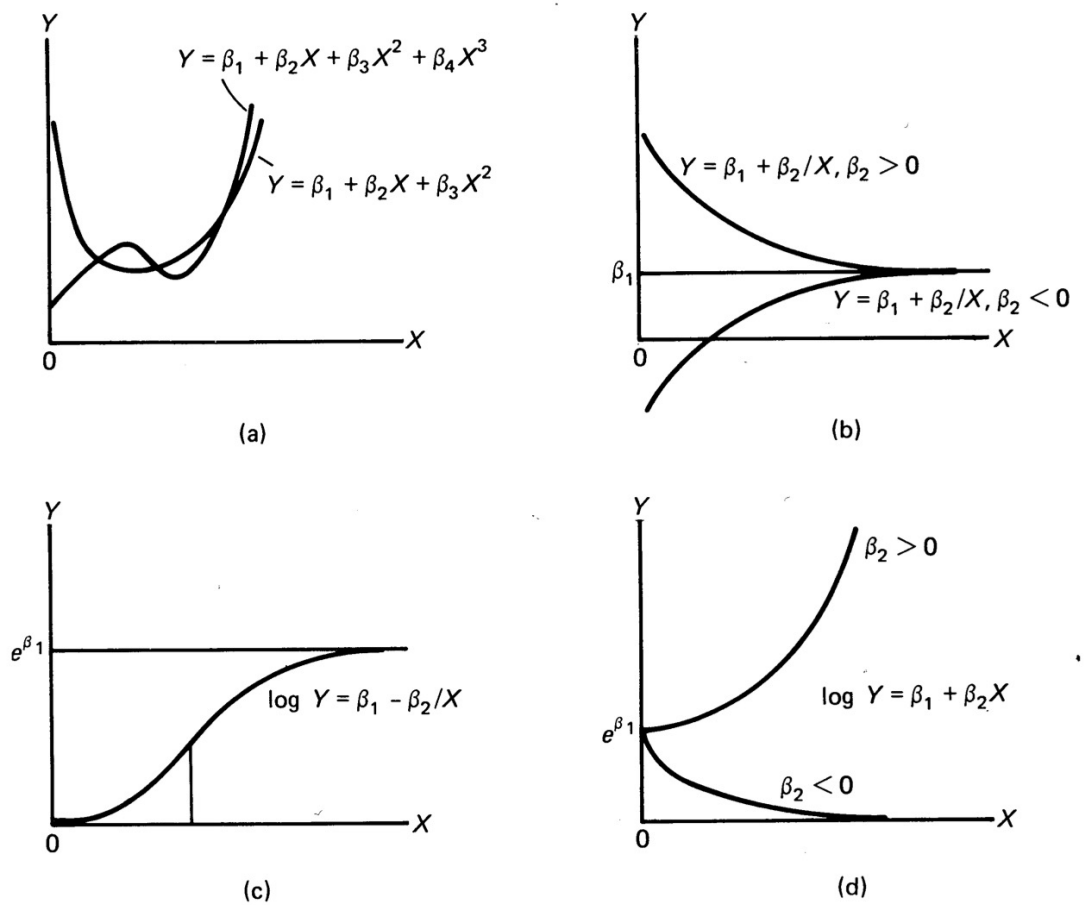
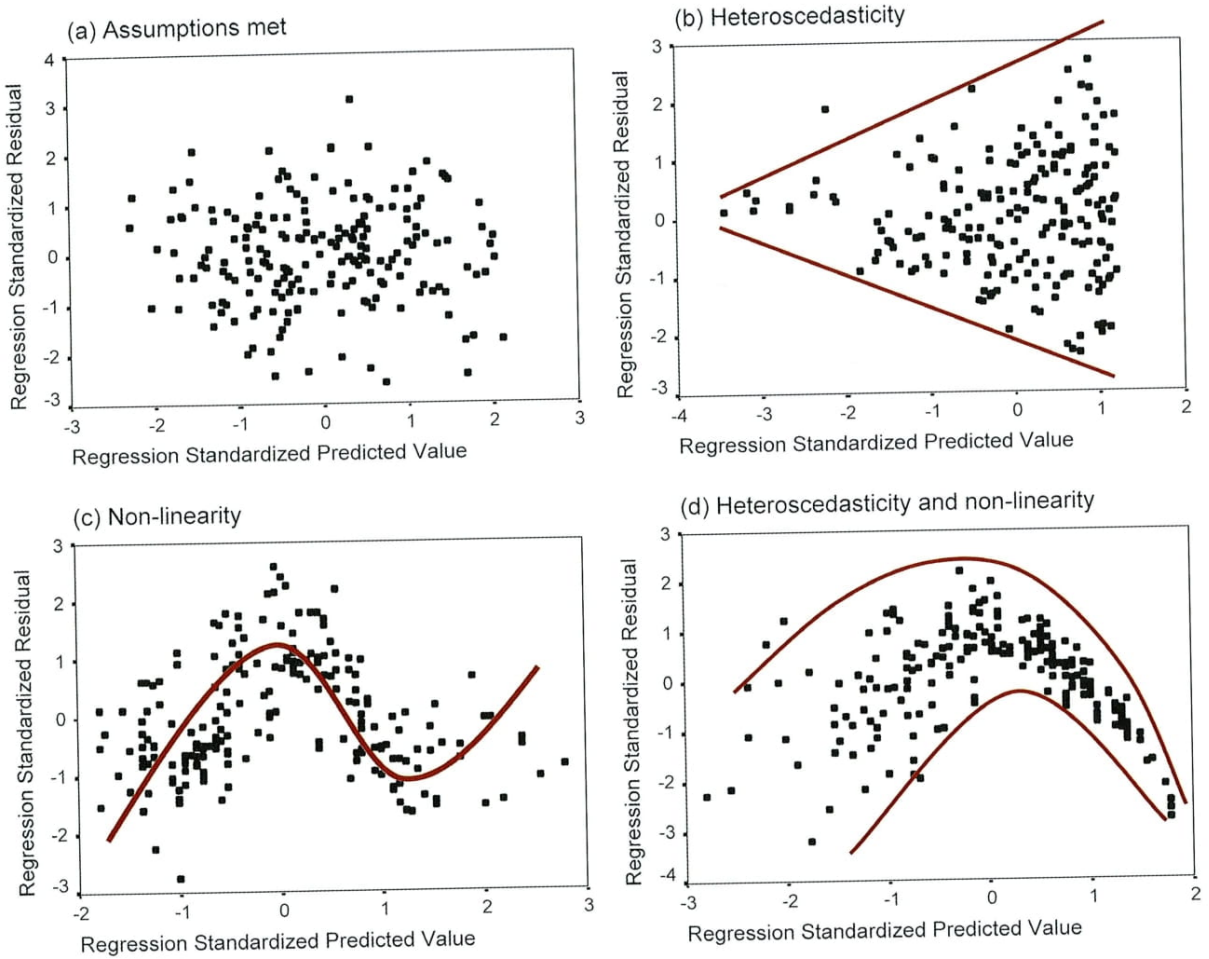


FIGURE 4.2 Alternative relationships: (a) polynomial, (b) reciprocal. (c) log–reciprocal, (d) semilog.

Lineartransformationen nichtlinearer Zusammenhänge

4.3 Heteroskedastizität



Ursachen für Heteroskedastizität

Heteroskedastizitätsproblem und -lösung

Probleme

- Die Residuen hängen mit X zusammen.
- Standardfehler der b verzerrt
- nichtlineare Zusammenhänge unerkannt

Lösungen

- Gibt es!
- Generalized least Squares (GLS)
- Kurvilineare Schätzungen

4.4 Annahmen zur Residualverteilung

Normalverteilung und Unabhängigkeit der Residuen

Schaut man sich visuell an. Wenn sie stark verletzt ist (z.B. bimodal) oder extrem schief, dann andere Methode.

Unabhängigkeit der Fehler

Die Fehler können nur voneinander abhängig sein, bei zeitlich geordneten Erhebungen, also Zeitreihenanalysen. Das braucht uns also erstmal nicht kümmern.

5 Dummys & Kategoriale

Kategoriale umkodieren

Kategoriale Variablen können in Dummys umkodiert (`case_match`) werden. Für jede Ausprägung der Kategorialen wird eine Dummy angelegt mit 1, wenn die jeweile Ausprägung zutrifft und 0, wenn nicht.

Kategoriale	D_A	D_B	D_C
A	1	0	0
A	1	0	0
B	0	1	0
C	0	0	1
A	1	0	0
B	0	1	0
A	1	0	0

die letzte Dummy ergibt sich

Wenn eine kategoriale 3 Ausprägungen daraus 3 Dummys gebaut werden, ergibt sich die letzte Dummy aus den ersten beiden!

Referenzkategorie weglassen

Wenn wir Dummys für eine Kategoriale in eine Regression aufnehmen wollen, müssen wir immer eine Kategorie weglassen, die wir dann die Referenzkategorie nennen.

Als Gleichung mit einer Dummy für polytom

Regressionsgleichung

$$
\begin{aligned}
Fem\_Mean\_IDX &= b_1 + b_2 \cdot weiblich + b_3 \cdot nonbinär + e \\
Fem\_Mean\_IDX &= b_1 + e \quad \text{wenn weiblich} = 0 \text{ und nonbinär} = 0 \text{ (m)} \\
&= b_1 + b_2 + e \quad \text{wenn weiblich} = 1 \\
&= b_1 + b_3 + e \quad \text{wenn nonbinär} = 1
\end{aligned}
$$

In Worten

Wenn Dummyvariablen für die unterschiedlichen Ausprägugnen einer polytomen Variablen stehen, steht die Konstante für die Referenzkategorie (muss es geben, da sonst perfekte Multikollinearität herrscht). Die übrigen b's stehen für die Mittelwertdifferenz zwischen den anderen Ausprägungen und der Referenzkategorie und werden mit t-Test (der Regression) auf Signifikanz getestet.

5.1 Regression mit einer Dummy und einer metrischen UV

Die Gleichung

$$Y = b_1 + b_2 D_2 + b_3 X_3 + e$$

In Worten

Wenn ein Dummyvariable D (zB “gender”) mit b2 in eine Regressionsgleichung eingeführt wird, dann ergeben sich zwei Parralelen mit dem Abstand von b_2.

5.2 Regression mit polytomer UV (zwei Dummys) und metrischer

	Stat Mean IDX		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.07	1.95 – 4.20	**<0.001**
PS Mean IDX	-0.32	-0.66 – 0.03	0.073
weiblich	0.97	0.45 – 1.48	**<0.001**
nonbinär	0.13	-0.87 – 1.14	0.793
Observations	166		
R^2 / R^2 adjusted	0.113 / 0.097		

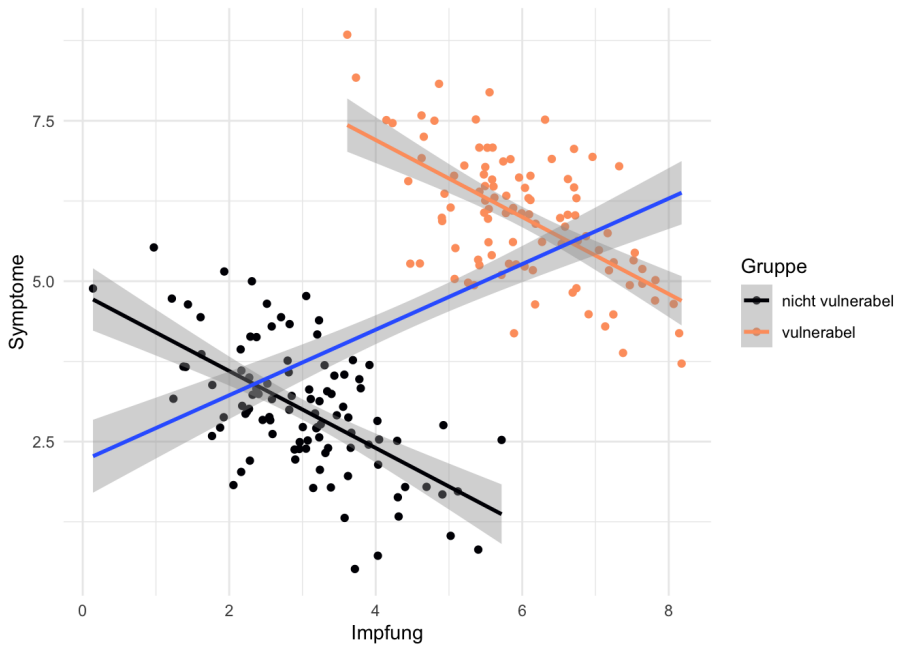
5.3 Beispiel: Simpsons Paradox

Hintegrund

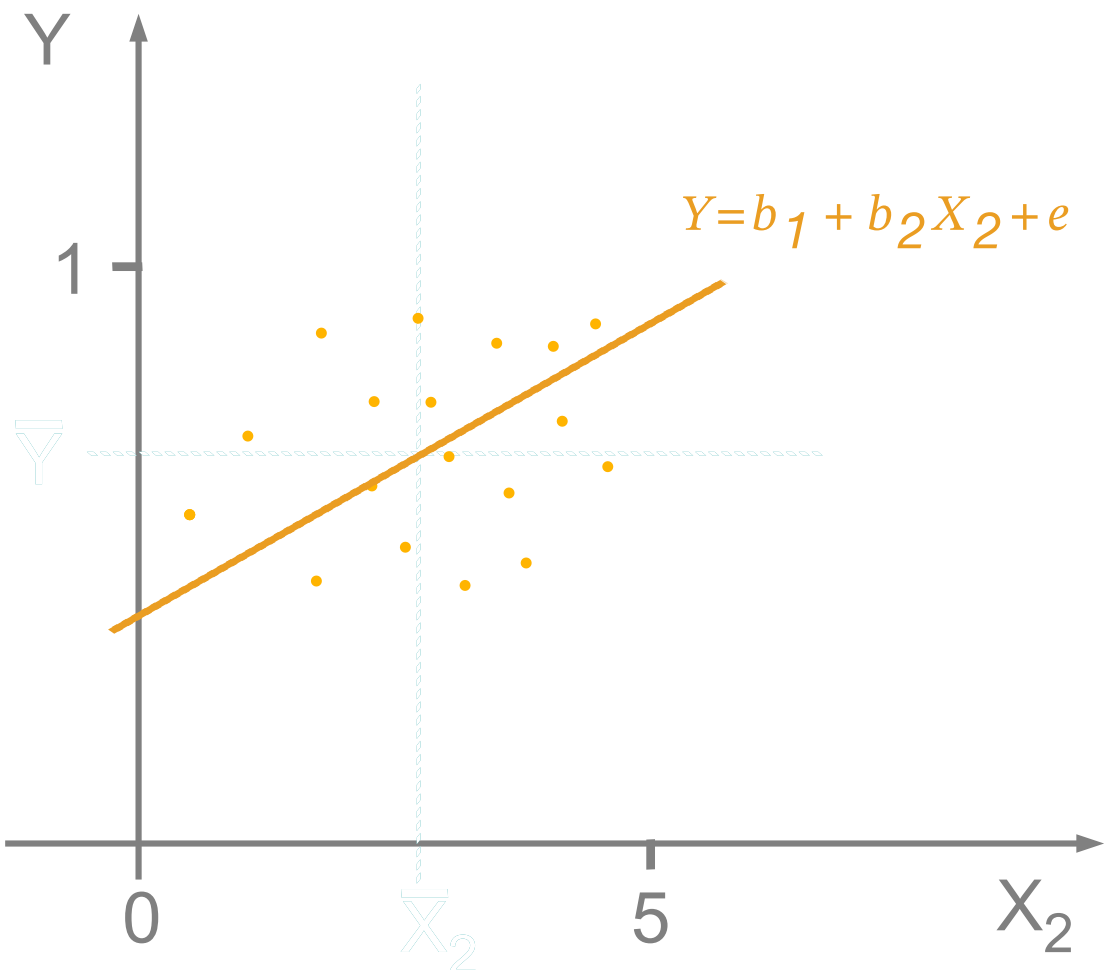
Beim Simpsons-Paradox haben gibt es mehrere Gruppen den gleichen Anstieg, aber in der UV und der AV unterschiedliche Level. Berücksichtigt man die Gruppen nicht, wird ein falscher Zusammenhang geschätzt. #Unterspezfikation

Bespiel Impfung

Oft haben Gruppen von geimpften Personen eine höhere Sterblichkeit betreffend der Krankheit, gegen die sie sich impfen lassen. Das liegt aber daran, dass sich vulnerable Personen eher impfen lassen.



6 Interaktionen mit Slope-Dummy

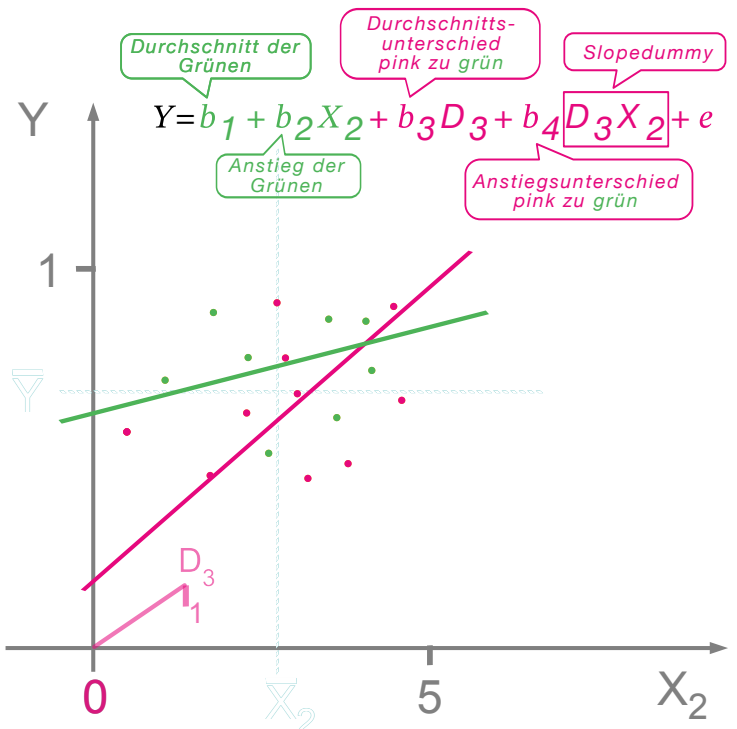


6.1 Interaktion mit Slope-Dummy (X_2 zentriert)

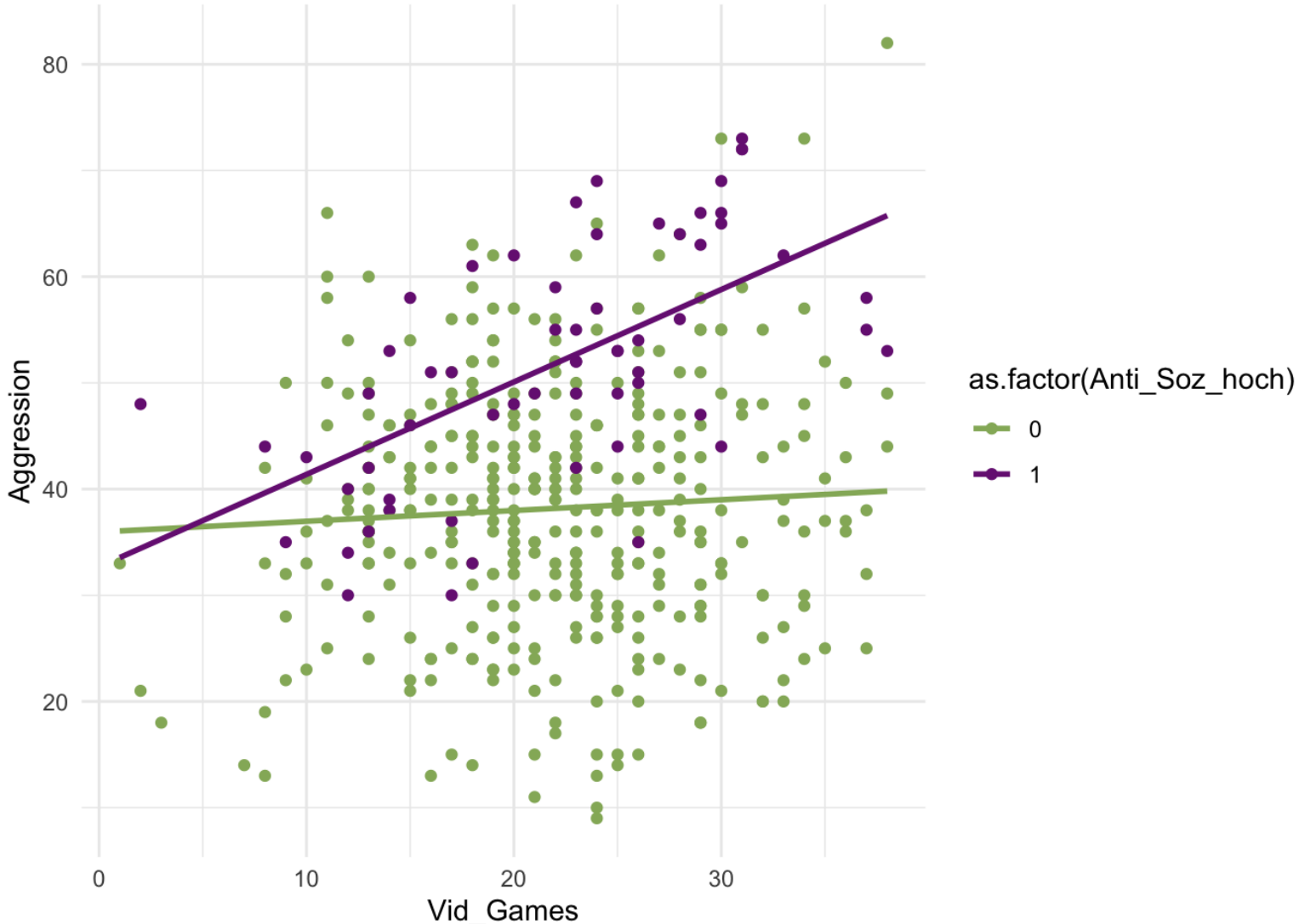
Regression mit Slope-Dummy

$$Y = b_1 + b_2 X_2 + b_3 D_3 + b_4 D_3 X_2 + e$$

$$\begin{aligned}
Y &= b_1 + e && | X_2 = 0, D_3 = 0 \\
Y &= b_1 + b_3 + e && | X_2 = 0, D_3 = 1 \\
Y &= b_1 + b_2 X_2 + e && | D_3 = 0 \\
Y &= b_1 + b_2 X_2 + b_3 + b_4 X_2 && | D_3 = 1
\end{aligned}$$



6.2 Interaktion Videogames und antisoziales Verhalten



6.3 Regression (unzentriert)

▶ Code

		Agression	
Predictors	*Estimates*	*std. Beta*	*p*
(Intercept)	35.95	-0.00	**<0.001**
Video Games(Hours per week)	0.10	0.11	0.238
Anti Soz hoch	-3.28	0.37	0.501
Vid_Games:Anti_Soz_hoch	0.77	0.15	**<0.001**
Observations	442		
R^2 / R^2 adjusted	0.183 / 0.177		

Variables	Tolerance	VIF
Vid_Games	0.832	1.202
Anti_Soz_hoch	0.105	9.526
Vid_Games:Anti_Soz_hoch	0.102	9.759

6.4 Regression nach Zentrierung

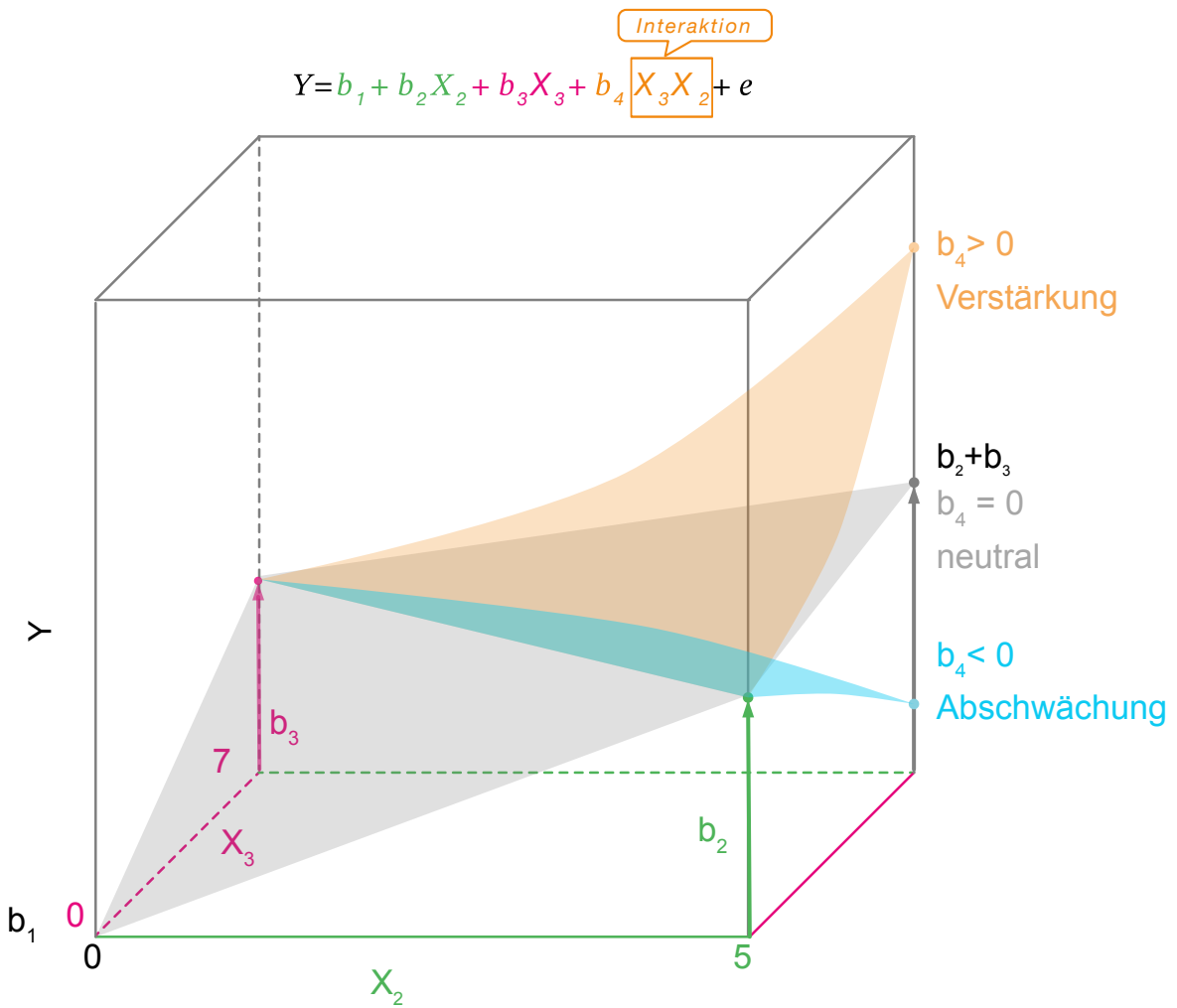
▶ Code

		Agression	
Predictors	*Estimates*	*std. Beta*	*p*
(Intercept)	38.17	-0.00	**<0.00**
Video Games(Hours per week)	0.10	0.11	0.238
Anti Soz hoch	13.52	0.37	**<0.00**
Vid_Games:Anti_Soz_hoch	0.77	0.15	**<0.00**
Observations	442		
R^2 / R^2 adjusted	0.183 / 0.177		

Variables	Tolerance	VIF
Vid_Games	0.832	1.202
Anti_Soz_hoch	0.999	1.001
Vid_Games:Anti_Soz_hoch	0.831	1.203

7 Interaktion zweier metrischer Variablen

Interaktion zweier metrischer Variablen



Regression mit Interaktion (zentrierter X)

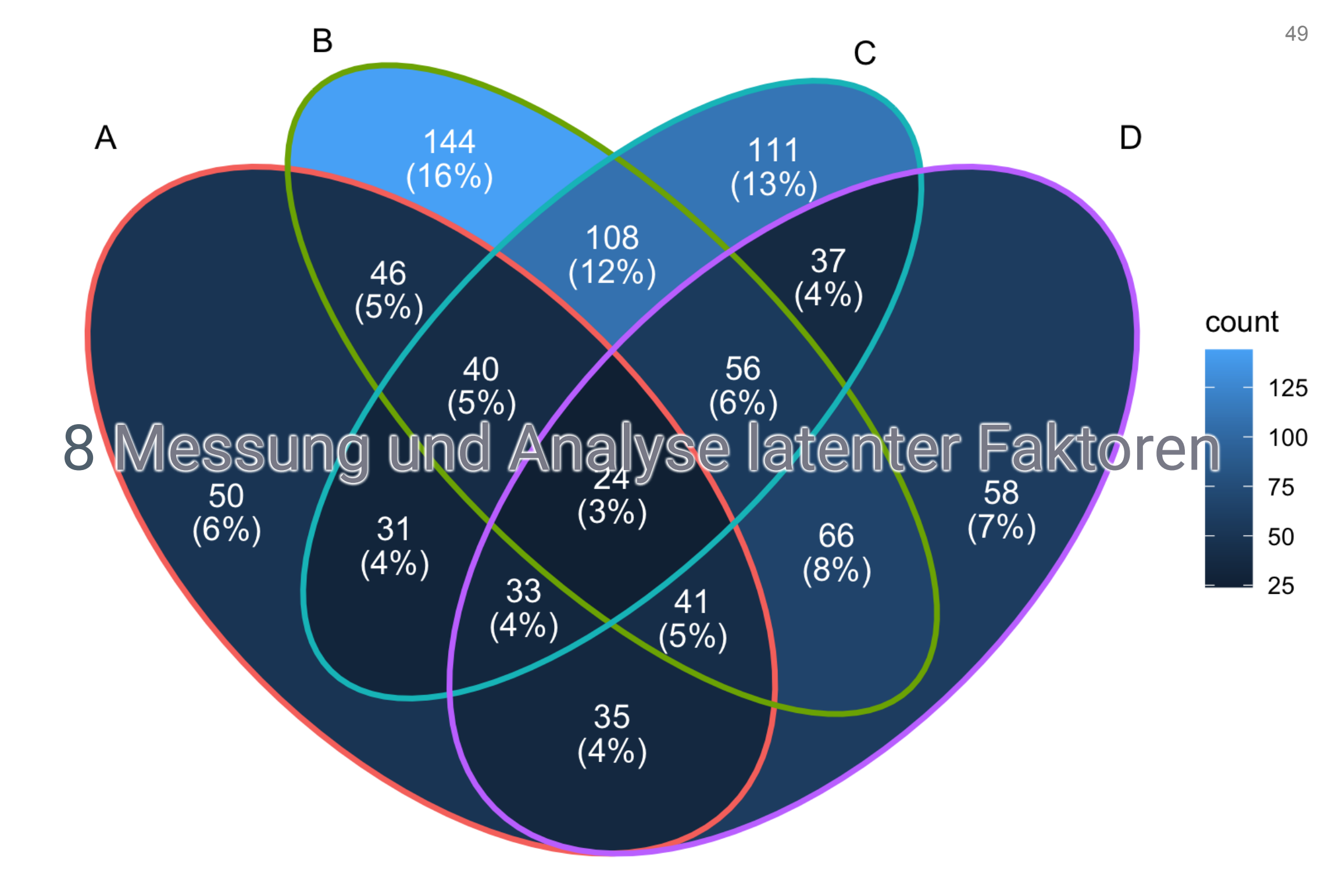
$$Y = b_1 + b_2X_2 + b_3X_3 + b_4X_2X_3 + e$$

- im Modell sind die Haupteffekte und die Interaktion zweier metrischer
- das b_1 ist der Schnittpunkt mit der Y-Achse (X=0) wieder
- das b_2 ist der Anstieg von X_2
- das b_3 ist der Anstieg von X_3
- das b_4 ist die Interaktion
  - ist $b_4 > 0$ Verstärkung
  - ist $b_4 < 0$ Abschwächung
  - ist $b_4 = 0$ neutral

7.1 Regression mit Faktoren und Interaktionen in R

▶ Code

	Stat Mean IDX					
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*p*	std. p
(Intercept)	3.98	0.16	2.90 – 5.06	-0.00 – 0.33	**<0.001**	0.056
PS Mean IDX	-0.30	-0.13	-0.68 – 0.08	-0.29 – 0.04	0.127	0.127
gender [männlich]	0.65	-0.66	-2.39 – 3.70	-1.06 – -0.27	0.673	**0.001**
gender [non-binär]	-1.89	-1.41	-3.70 – -0.08	-2.77 – -0.06	**0.041**	**0.041**
gender [is doch völlig egal]	-3.58	-0.37	-11.48 – 4.32	-1.16 – 0.41	0.372	0.348
PS Mean IDX × gender [männlich]	-0.55	-0.23	-1.57 – 0.47	-0.67 – 0.20	0.292	0.292
PS Mean IDX × gender [is doch völlig egal]	1.10	0.47	-1.67 – 3.86	-0.71 – 1.64	0.435	0.435
Observations	166					
R^2 / R^2 adjusted	0.134 / 0.101					


A
B
C
D
144
(16%)
111
(13%)
108
(12%)
46
(5%)
37
(4%)
40
(5%)
56
(6%)
50
(6%)
24
(3%)
58
(7%)
31
(4%)
66
(8%)
33
(4%)
41
(5%)
35
(4%)
count
125
100
75
50
25

Was geht? … mit Faktorenanalysen!

- Mit Faktorenanalysen können latente Einflüsse explorativ gefunden werden.
- Die Messung latenter Konstrukte kann (konfirmatorisch) geprüft werden.
- Mit Faktorenanalysen können Indices gebaut werden.
- Wenn UVs in Regressionsmodellen hoch multikollinear sind, können sie zu unkorrelierten Faktoren zusammengefasst werden.

Fragen, die die FA beantwortet

Wie viel geht bei der Dimensionsreduktion durch die Faktoren verloren, bzw. was wird abgebildet?

Mit der Gesamtlösung kann man schauen, welchen Anteil der Varianz aller Faktoren durch die Faktorlösung abgebildet wird.

Wie gut werden die Variablen durch die Faktoren abgebildet?

Die Kommunalitäten und «Uniqueness» geben an, wie gut jede Variable durch die gebildeten Faktoren repräsentiert werden.

Was bedeuten die Faktoren?

Faktorladungen geben die Korrelationen der Faktoren mit jeder Variable an. Also welche Faktoren, welche Variablen repräsentieren? Dadurch kann den Faktoren ein Sinn und ein Name gegeben werden.

8.1 Vorgehen der PCA und Faktorenanalyse

1. Prüfen, ob ein Set an Variablen für eine Faktorenanalyse geeignet ist

- Korrelationsanalyse
- KMO

2. Feststellen, wie viele latente Faktoren extrahiert werden sollten

- Scree-Plot
- Parallelanalyse

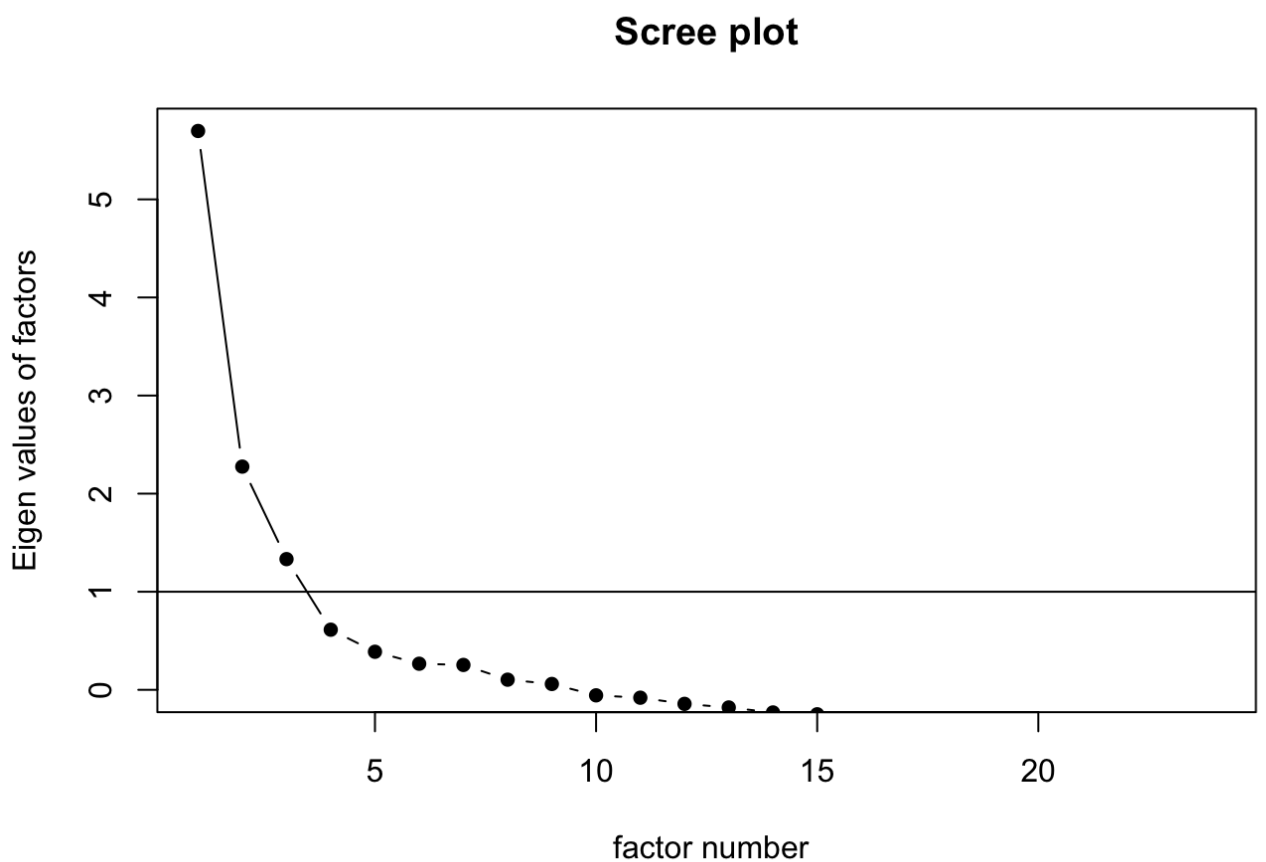
3. Interpretation der Faktoren

- Berechnung der Faktorladungen
- Identifikation der Faktoren (Interpretation)

Scree Plot der Eigenwerte (eigen values)

Die «factor number» über der 1-Linie («Eigen values» > 1) ist eine Empfehlung für die Anzahl an Faktoren, bei denen jeder Faktor mehr Varianz (Eigenwert > 1) auf sich vereint als die ursprünglichen Dimensionen.

- In R psych::scree()



8.2 Faktorrotation

Unrotiert

Beim Verfahren der Faktorenanalyse wird erst ein Faktor in die Variablen gelegt, der alle am besten erklärt. Dann kommt der zweite und optimiert den Rest der Varianz usw. Das ergibt ein Ungleichgewicht zwischen den Faktoren. Darum wird rotiert.

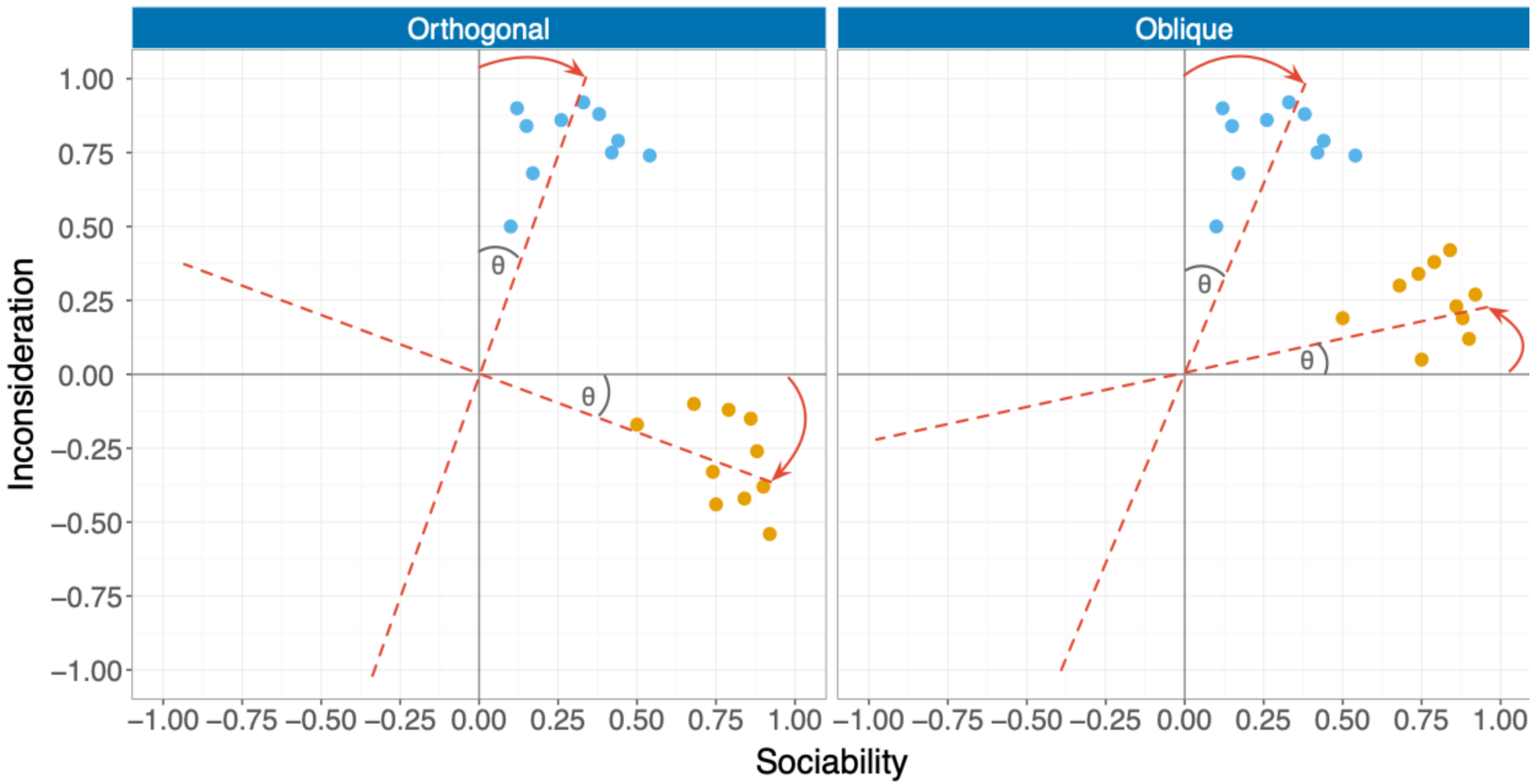
Orthogonale und oblique Rotation

Faktoren werden rotiert, damit jeder optimal Varianz erklärt. Wenn orthogonal rotiert wird, sind die Faktoren 100% unkorreliert. Nach obliquer Rotation sind die Faktoren leicht korreliert, geben aber die Variablen besser wieder.

Faktorrotation

Bei orthogonaler Rotation sind die Faktoren unkorreliert.
Orthogonal ist klarer interpretierbar …

Bei der obliquen dürfen sie leicht korrelieren.
… oblique gibt realistischere Ergebnisse.



Faktorladungen

$$MR1 = b_r raq_6 + b_2 raq_{18} + b_3 raq_{13} + b_4 raq_7 + b_5 raq_{10} + b_6 raq_{15} + ... \quad (1)$$

$$MR2 = b_r raq_{09} + b_2 raq_{23} + b_3 raq_{19} + b_4 raq_{22} + b_5 raq_{02} \quad (2)$$

- Die Bs sind die Faktorladungen.
- Faktorladungen geben an, welches Gewicht (Bedeutung) die einzelnen Variablen für den jeweiligen Faktor haben.
- Jeder Faktor wird anhand der Variablen mit den höchsten Ladungen auf diesem Faktor interpretiert.

Faktorladungen RAQ

Faktorladungen sind die Korrelationen der Variablen mit den Faktoren (MR1 bis MR3).

	Variable	MR1	MR2	MR3	Complexity	Uniqueness
1	R101_07	0.74			1.02	0.4
2	R101_15	0.74			1.21	0.39
3	R101_13	0.71			1.03	0.43
4	R101_14	0.66			1.22	0.49
5	R101_06	0.66		-0.3	1.49	0.57
6	R101_02	0.55			1.09	0.65
7	R101_12	0.5	0.33		1.88	0.61
8	R101_18	0.42			1.54	0.69
9	R101_10	0.35			1.15	0.85
10	R101_08		0.75		1	0.45
11	R101_11		0.66		1.24	0.57
12	R101_09		0.61		1.11	0.6
13	R101_05	0.35	0.55		1.72	0.47

	Variable	MR1	MR2	MR3	Complexity	Uniqueness
14	R101_22		0.53		1.15	0.66
15	R101_01	0.34	0.51		1.75	0.54
16	R101_24		-0.45	0.4	2.06	0.65
17	R101_03		-0.35		1.24	0.87
18	R101_23		-0.32		2.27	0.81
19	R101_21			0.75	1.02	0.39
20	R101_04			0.7	1.15	0.55
21	R101_20			0.65	1.21	0.43
22	R101_19			0.47	1.27	0.71
23	R101_17			0.44	1.72	0.65
24	R101_16	0.3		0.3	2.96	0.63

:::
:::

Variableneignung – Kommunalitäten & Uniqueness

Kommunalitäten

Die Kommunalität einer Variable ist der Varianzanteil, den sie mit den extrahierten Faktoren teilt. Kommunalitäten unter .4 sind eher dürftig.

Uniquness = 1 - Kommunalität

Uniqueness

Die Uniqueness-Werte drücken aus, wie hoch der Varianzanteil ist, der **nicht** durch die Faktorenlösung erklärt werden konnte. Werte über .6 sind eher dürftig.

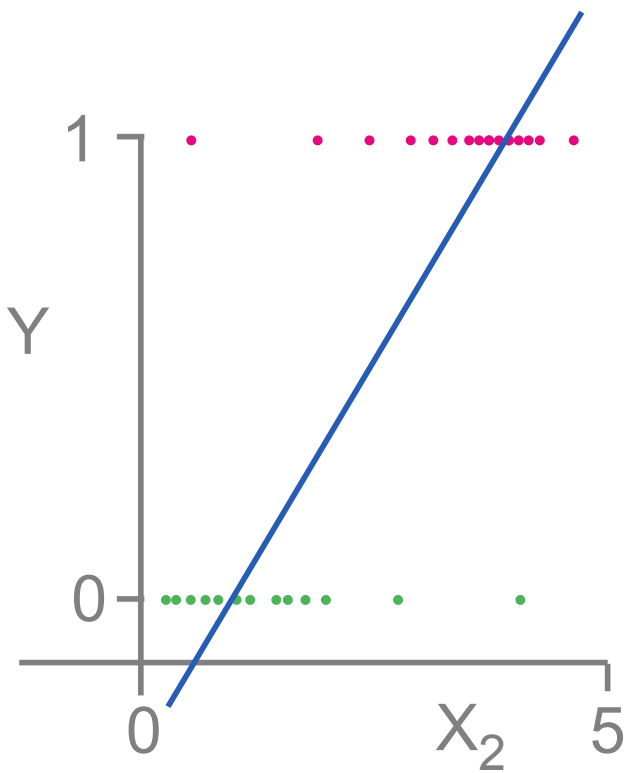
Complexity

Die Komplexität je Variable gibt an, ob es Mehrfachladungen auf einer Variable gibt. Wenn sie 1 ist, dann ist das Ergebnis eindeutig, wenn sie nahe 2 ist, dann laden zwei Faktoren auf dieser Variable.

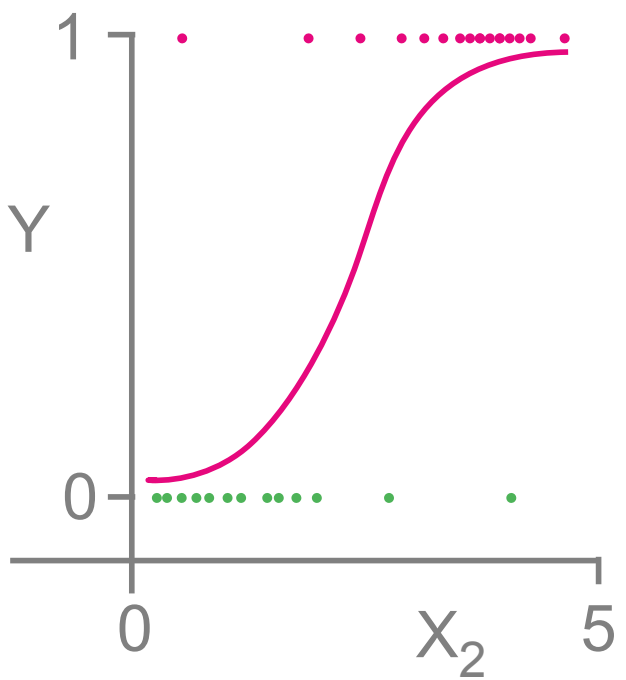
9 Logistische Regression

Lineare vs. logistische Regression

Lineare Regression



Logistische Regression



Lineare Regression vs. Logistische

9.1 Interpretation der binär logistischen Regression

- Die B's haben die gleiche Bedeutung wie bei linearer Regression, beziehen sich aber auf die logarithmierte AV. Das Vorzeichen ist interpretierbar (positiv/negativ, wenn signifikant).
- Die Interpretation der Zusammenhänge läuft am besten über die Exp(B) = «Odds Ratio» (OR). OR geben an, wie stark sich die (Wett)Quote für AV = 1 ändert, wenn die die UV um eine Einheit grösser wird.
- Ist OR > als 1, steigt die Quote. Ist OR < 1, sinkt sie. Das CI der OR schliesst 1 ein, wenn der Effekt nicht signifikant ist.
- Die H0 der OR geht von 1 aus. Die Werte gehen von $\frac{1}{\infty}$ bis ∞.

9.2 Odds und Odds-Ratio

Odds

$$Odds = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y_{\text{trifft ein}})}{1-P(Y_{\text{trifft ein}})}$$

Odds Ratio

$$OR = Exp(B) = e^{\beta} = \frac{\text{Odds nach dem Ansteig von x um eine Einheit}}{\text{Odds vor dem Anstieg von x um eine Einheit}} = \frac{Odds_{nach}}{Odds_{vor}}$$

$$Odds_{nach} = Exp(B) \cdot Odds_{vor}$$

B (Regressionskoeffizient)	**Exp(B)** (Odds Ratio)		**P(y=1)**
B > 0	$e^B > 1$	nimmt um den Faktor Exp(B) zu	Zunahme
B = 0	$e^B = 1$	bleibt gleich	bleibt gleich
B < 0	$e^B < 1$	sinkt um den Faktor Exp(B)	Abnahme

RAQ erklärt Interesse an Maschine Learning

LogReg: Aversion gegen Computer erklärt Interesse an ML

Characteristic	OR[1]	95% CI[1]	p-value	VIF[1]
ML1	0.74	0.53, 1.02	0.065	1.0
ML2	0.54	0.34, 0.83	0.007	1.7
ML3	1.74	1.12, 2.81	0.018	1.7

[1] OR = Odds Ratio, CI = Confidence Interval, VIF = Variance Inflation Factor

Null deviance = 230; Null df = 165; Log-likelihood = -109; AIC = 226; BIC = 238; Deviance = 218; Residual df = 162; No. Obs. = 166

LM: Aversion gegen Computer erklärt Interesse an ML

Characteristic	Beta	95% CI[1]	p-value	VIF[1]
ML1	-0.07	-0.15, 0.00	0.066	1.0
ML2	-0.14	-0.24, -0.04	0.005	1.6
ML3	0.13	0.02, 0.23	0.016	1.6

[1] CI = Confidence Interval, VIF = Variance Inflation Factor

R^2 = 0.071; Adjusted R^2 = 0.054; Sigma = 0.488; Statistic = 4.13; p-value = 0.007; df = 3; Log-likelihood = -114; AIC = 239; BIC = 254; Deviance = 38.5; Residual df = 162; No. Obs. = 166

9.3 Multinominale Regression

Multinomial bedeutet, dass es eine kategoriale (multinomiale) AV gibt. Im Prinzip werden mehrere binäre logistische Regressionen durchgeführt und zusätzlich ausgegeben, wie gut das Gesamtmodel ist. Es kommen die Fit-Kennungen AIC und BIC dazu.

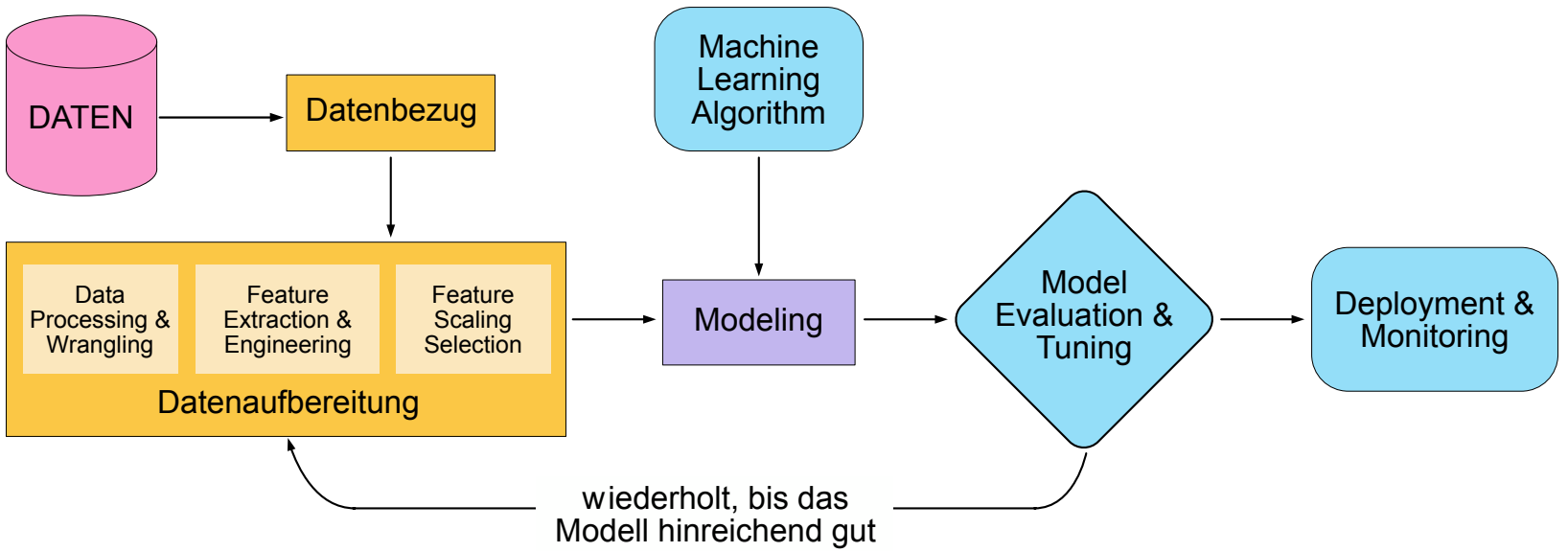
Wir schauen uns nächste Woche Outputs an.

Supervised ML

Trainieren ➪ Testen ➪ Anwenden

Die verfügbaren Daten werden in einen grösseren und einen kleineren Teil getrennt:

1. Trainingsdaten, an denen das Modell trainiert wird (Modelling)
2. Testdaten, an denen das Modell evaluiert wird.



10 Clusteranalyse

10.1 Problemstellung und Vorgehen

Problemstellung

Wie können Fälle in einem Datensatz nach einer oder mehreren Variablen gruppiert werden?

Grundsätzliches Vorgehen

Wir suchen Gruppen (Cluster) von Fällen, die sich untereinander so stark wie möglich ähneln (homoge Cluster) und so stark von den anderen Gruppen unterscheiden wie möglich. Es geht also um Segmentierung anhand von Mustern in den Daten – Clusteranalyse ist Mustererkennung.

Zugehörigkeit des Verfahrens

Die Clusteranalyse gehört zu den explorativen Verfahren.

Im Kontext von ML wird sie als «unsupervised learning» behandelt.

10.2 Voraussetzungen

Nicht zu viele fehlende Werte

Zu viele fehlende Werte verfälschen die Clusterbildung.

Skalenniveau

Das Skalenniveau spielt keine Rolle. Die Algorithmen können Cluster anhand von metrischen, dichotomenn oder kategorialen (mehrere Kategorien) Variablen extrahieren.

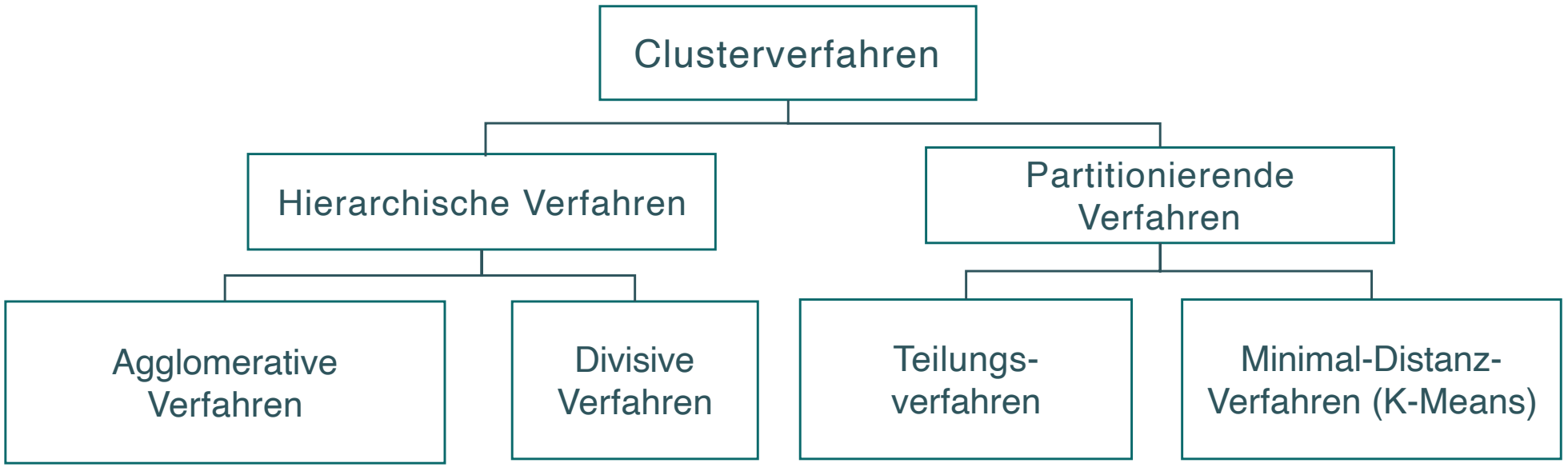
Fallzahl

Es braucht relativ grosse Stichproben. Bei kleinen Stichproben sind die Clusteranalysen recht ungenau.

Ähnliche Skalierung der Variablen

Haben die Variablen sehr unterschiedliche Skalierungen (zB Geschlecht dichotom und Alter in Jahren), dann ist eine vorherige z-Transformation der Variablen sinnvoll.

Systematik der Clusteranalyseverfahren



11 Hierarchische Clusteranalyse

Ablauf

1. Jeder Fall ist sein eigenes Cluster (2468 Befragte → 2468 Cluster)
2. Für jede Paarung werden die Ähnlichkeitsmasse berechnet
3. Die beiden Fälle geringster Distanz werden zusammengelegt
4. Es werden wieder die Abstände zwischen dem neuen Cluster und den übrigen berechnet
5. Schritt 3 und 4 so lange wiederholen bis alle Objekte in einem Cluster sind

11.1 Vorteile und Nachteile

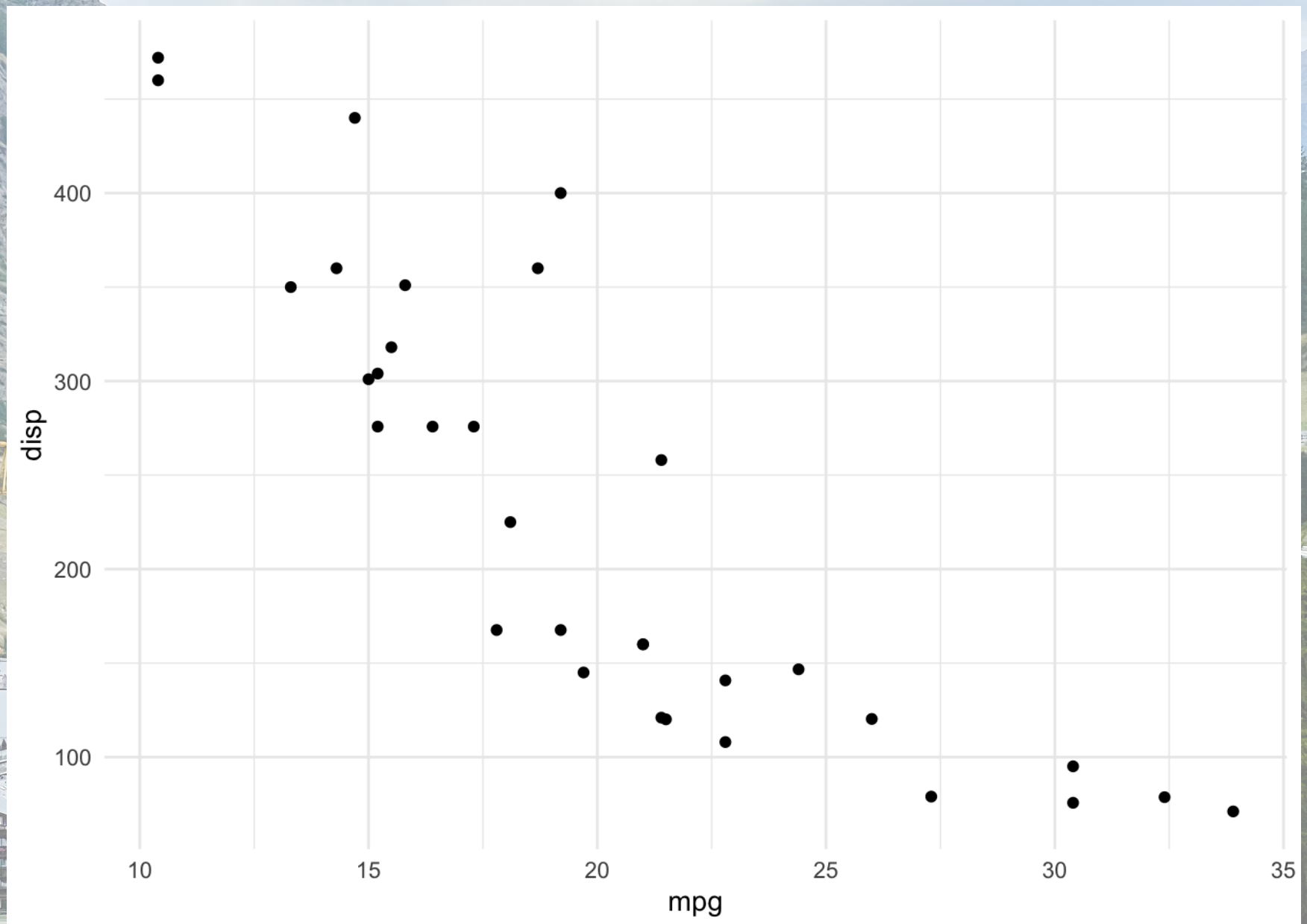
Vorteile

- Kann auch mit kategorialen Variablen und ordinalen gerechnet werden.
- Einfach identifizierbare Ausreisser
- Kann gut angepasst werden

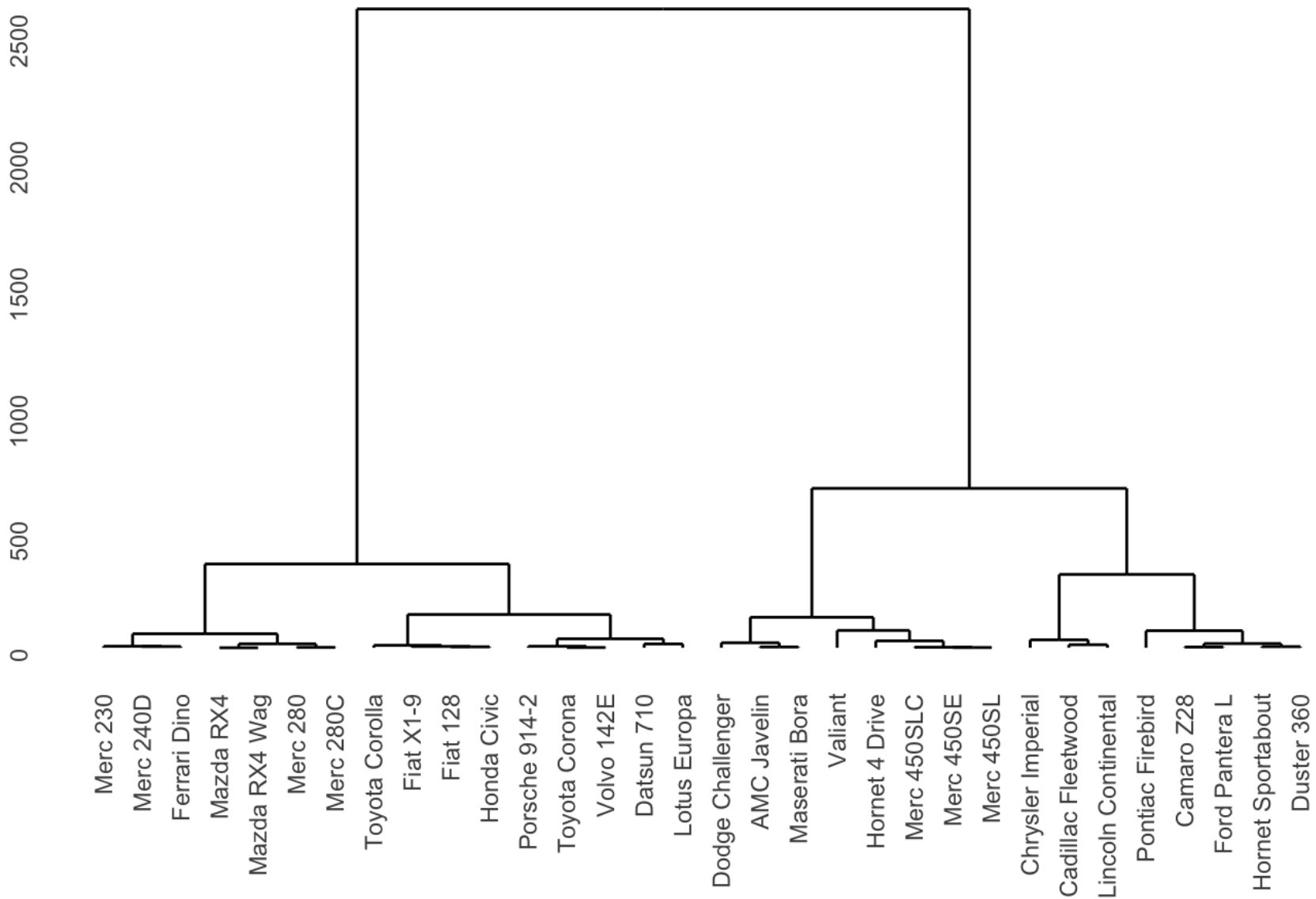
Nachteile

- Es müssen immer jeweils paarweise Distanzen auf jeder Stufe berechnet werden
- Sehr rechenaufwendig und damit viel langsamer als k-means-Cluster
- Viele Masse und Kennwerte

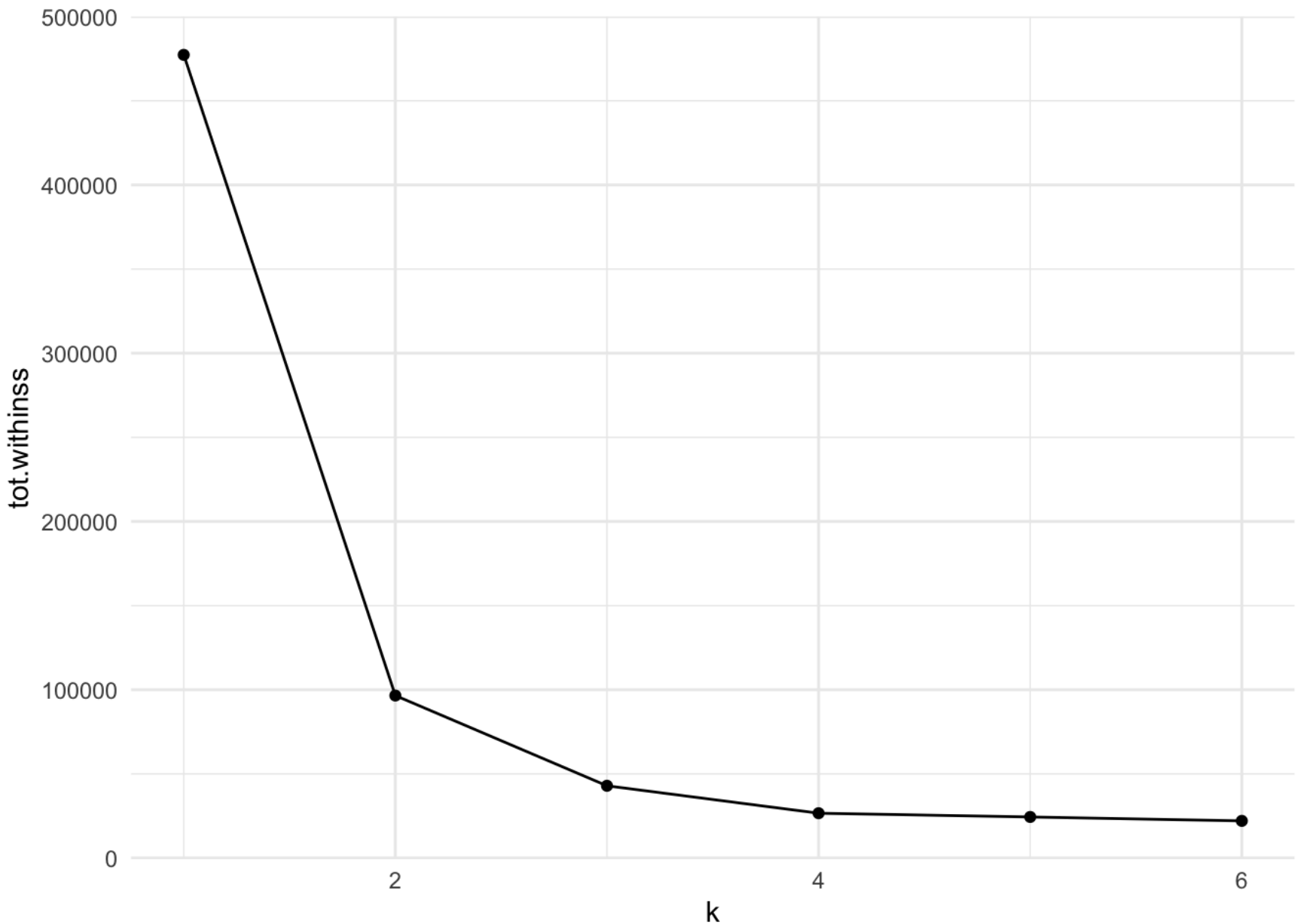
Hierarchische Clusteranalyse in R



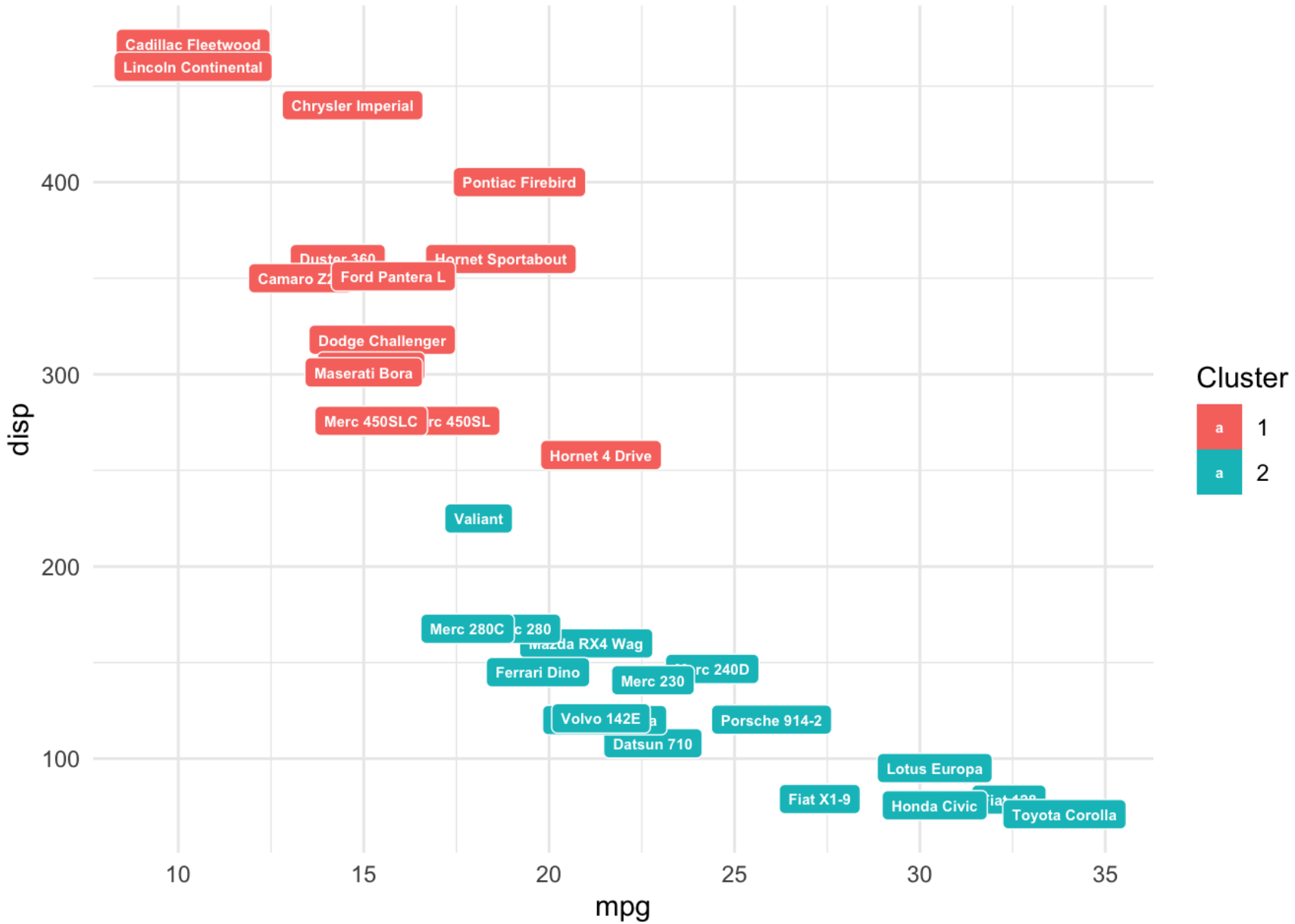
Dendrogramm



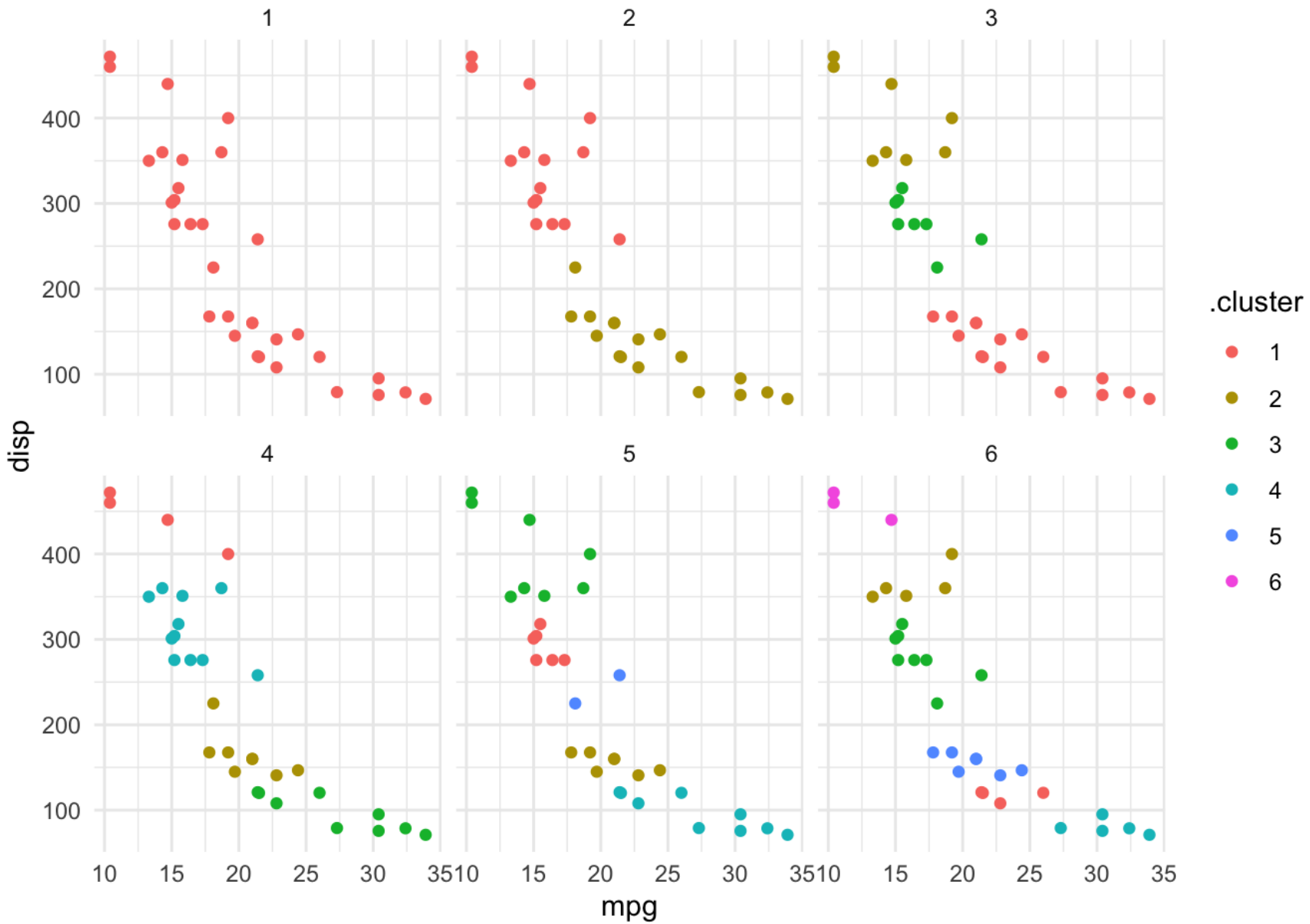
Screeplot der Quadratsumme innerhalb der Cluster



Cluster


Cadillac Fleetwood
Lincoln Continental
Chrysler Imperial
Pontiac Firebird
Hornet Sportabout
Ford Pantera L
Dodge Challenger
Maserati Bora
Merc 450SLC
Hornet 4 Drive
Valiant
Merc 280C
Ferrari Dino
Merc 230
Volvo 142E
Porsche 914-2
Datsun 710
Lotus Europa
Fiat X1-9
Honda Civic
Toyota Corolla
400
300
200
100
disp
10
15
20
25
30
35
mpg
Cluster
1
2

Vergleich nach Clustergrössen



12 K-Mean-Clustering

Ablauf

0. Clustervariablen festlegen
1. Bestimmung der Anzahl Cluster
2. zufällige Startpartition der Cluster anlegen
3. nächste Elemente der Cluster bestimmen
4. Clusterzentren neu ausrichten
5. nächste Elemente der Cluster bestimmen
6. **iterativ Clusterzentren immer wieder neu ausrichten (4.) und zugehörige Elemente bestimmen (5.)**
7. wenn sich nichts mehr tut, enden
8. Interpretation der Clusterlösung

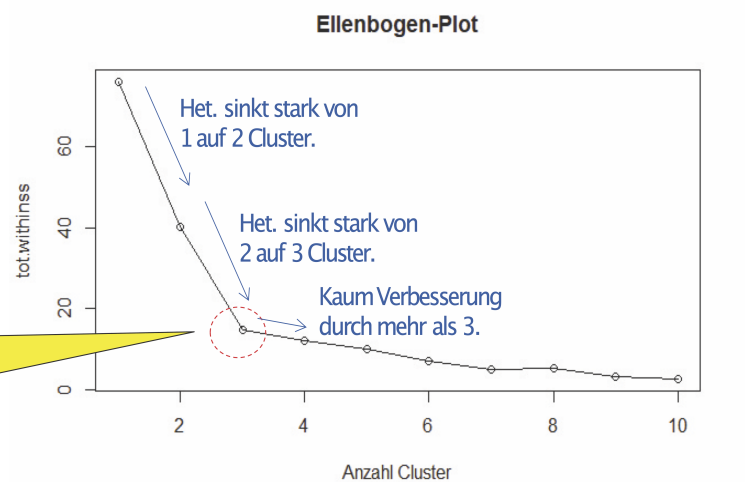
12.1 Voraussetzung

- Die Variablen für die Clusteranalyse müssen metrisch skaliert sein. Kategoriale und ordinale Variablen (mit Informationsverlust) können aber in Dummys umgewandelt werden.
- Die Variablen sollten ähnliche Standardabweichungen haben, können dafür aber z-transformiert werden oder Faktoren einer FA sein (die sind z-transformiert)
- Die Variablen sollten nicht zu stark korrelieren. Bei höheren Korrelationen bietet sich eine vorherige FA an.

12.2 Clusterzahl bestimmen

Es werden k-mean-Clusteranalysen für 1-viele durchgeführt und dann der Knick (Ellbogen) gesucht.

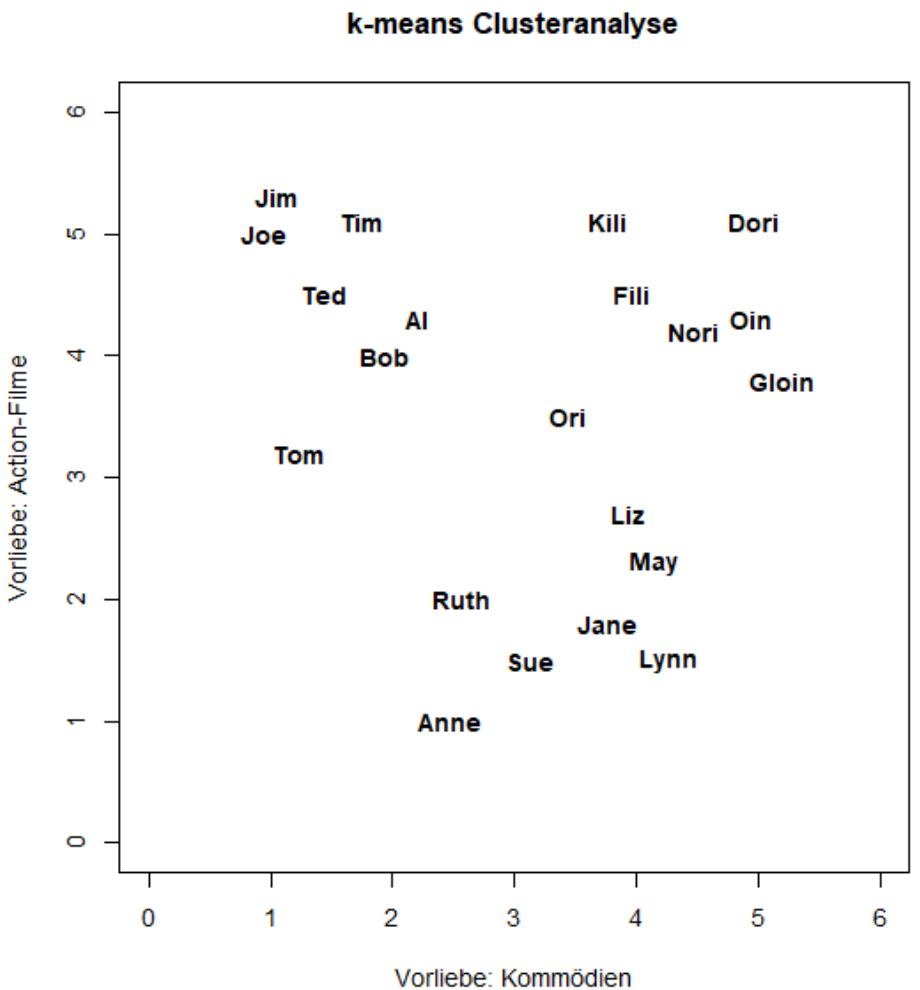
Hier ist der Ellenbogen. Also gibt es **genau drei Cluster**. Die Lösung, die oben schrittweise gefunden wurde, ist also die beste Clusterlösung.
Achtung: Die korrekte Anzahl ist genau der Knick. Nicht wie beim Scree–Kriterium!



Gütemass der Lösung als R^2

R^2 als $\frac{between_{SS}}{total_{SS}}$ wie Varianzaufklärung.

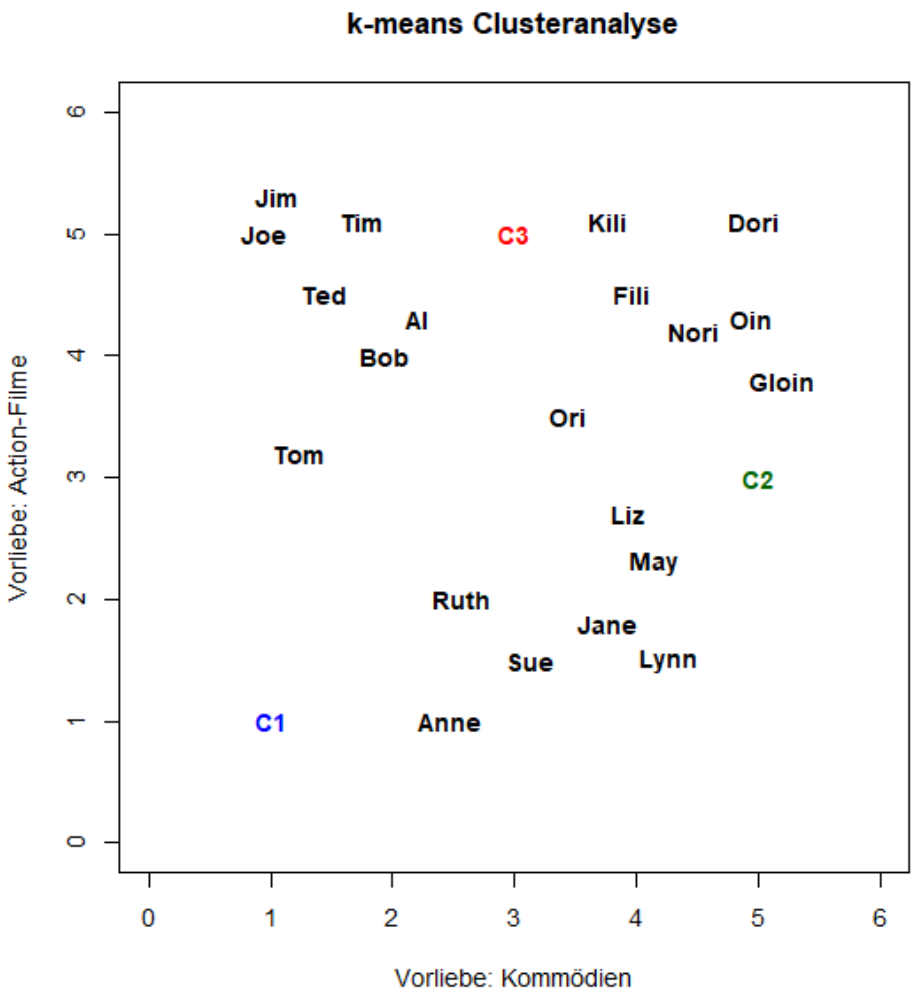
12.3 K-Means-Clusteranalyse Iterationen



Ausgangslage für die Cluster-Analyse

- 21 Elemente, die bezüglich zwei metrischen Merkmalen geclustert werden sollen.
- Für jedes Element wurde jedes Merkmal gemessen, so können die Fälle auf einem Koordinatensystem eingetragen werden.
- Als Distanzmass wird die euklidische Distanz verwendet. Das ist genau die Distanz, die man von Auge sieht.

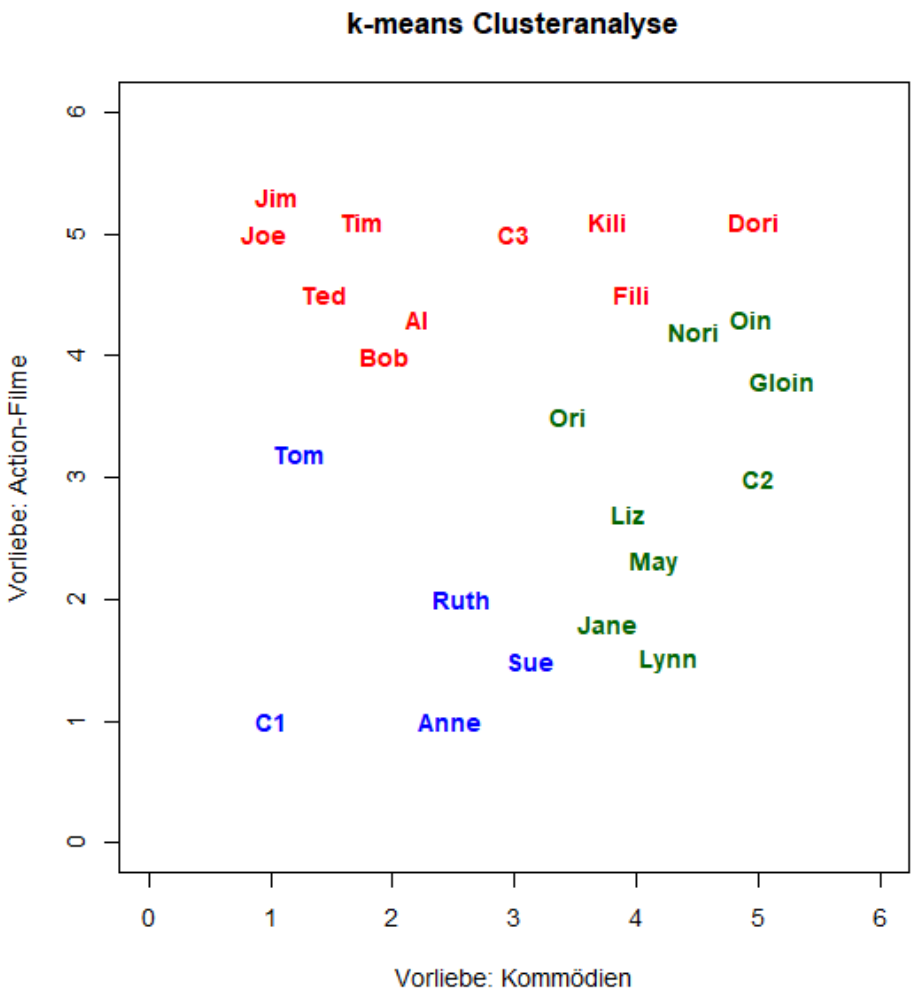
K-Means-Cluster-Algorithmus



Schritt 1: Cluster-Zentren

- Es wird eine feste Anzahl (k) von Clusterzentren definiert, die irgendwo zufällig verteilt werden
- Die Cluster-Zentren müssen nicht in der Nähe der tatsächlichen Cluster sein.
- Hier heissen die Zentren: C1, C2 und C3 und sind absichtlich ausserhalb der von Auge sichtbaren Cluster gesetzt.

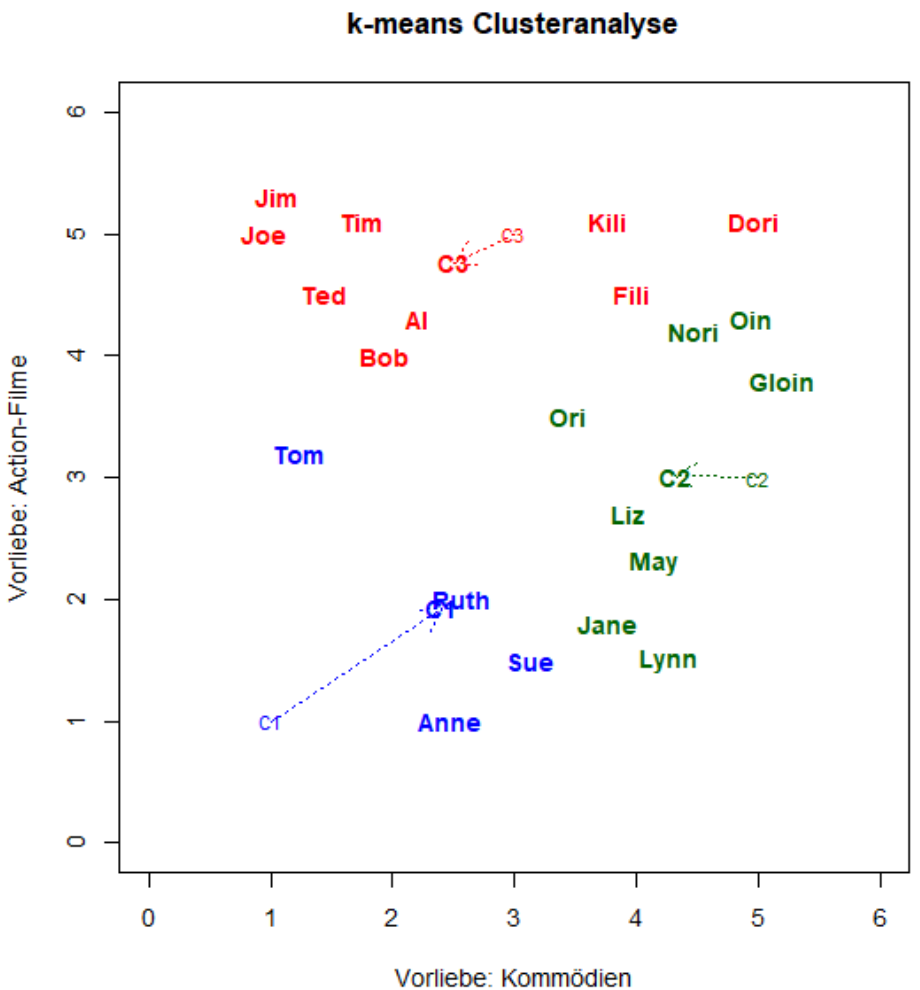
K-Means-Cluster-Algorithmus



Schritt 2: Elemente zuordnen

- Jedes Element wird dem Cluster zugeordnet, zu dessen Zentrum es den geringsten Abstand hat.
- Dies ist die erste Lösung der k-Means Analyse (schlechte Lösung)
- Die Cluster-Zentren liegen nun aber nicht im Zentrum der Cluster

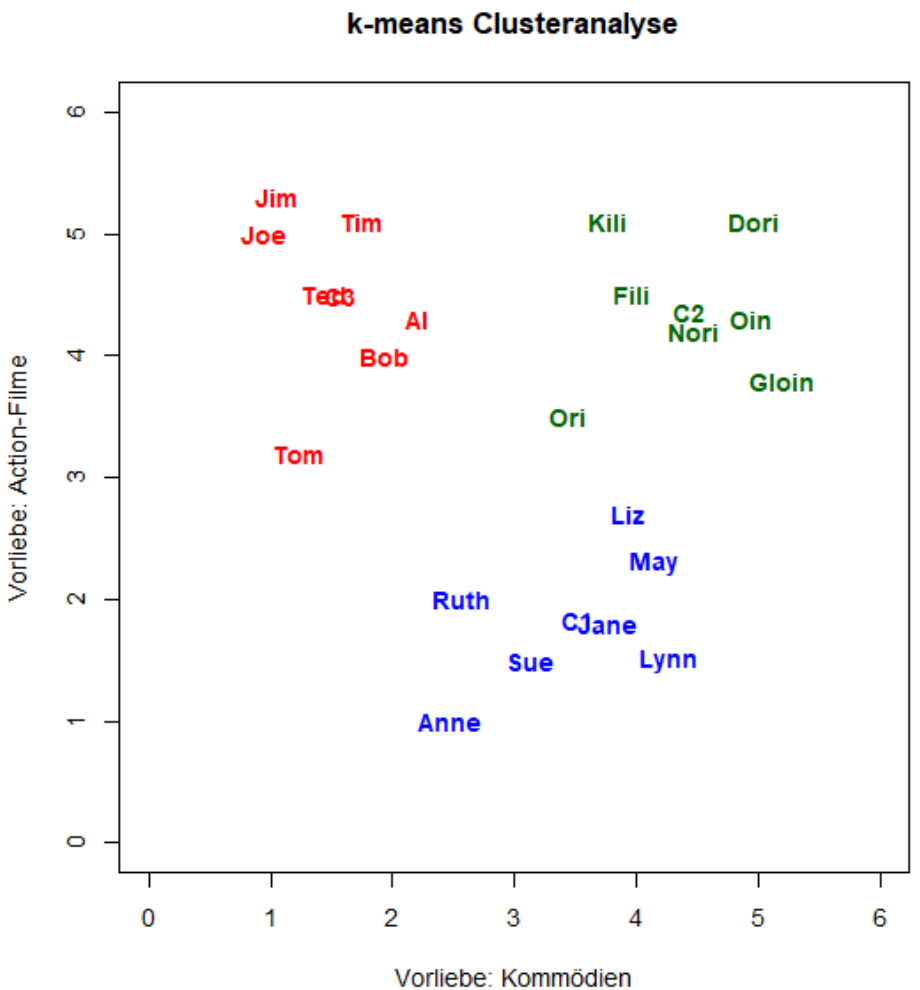
K-Means-Cluster-Algorithmus



Schritt 3: Zentren zentrieren

- Die tatsächlichen Zentren der Cluster werden berechnet, indem man den Mittelpunkt aller Elemente berechnet, die zu jedem Cluster gehören.
- Die Cluster-Zentren werden dort hin verschoben.
- Nun haben sie sich aber näher an einige Elemente bewegt und sich von anderen entfernt

K-Means-Cluster-Algorithmus



nochmals Schritt 2: Elemente zuordnen

- Es gibt keine Änderungen mehr. Alle Elemente bleiben in den Clustern, denen sie zugeordnet waren.
- Jetzt stimmen auch die Zentren
- Endlösung ist erreicht: Die drei Zentren liegen genau im Zentrum der Elemente, die ihnen am nähesten sind.

Der k-means Algorithmus

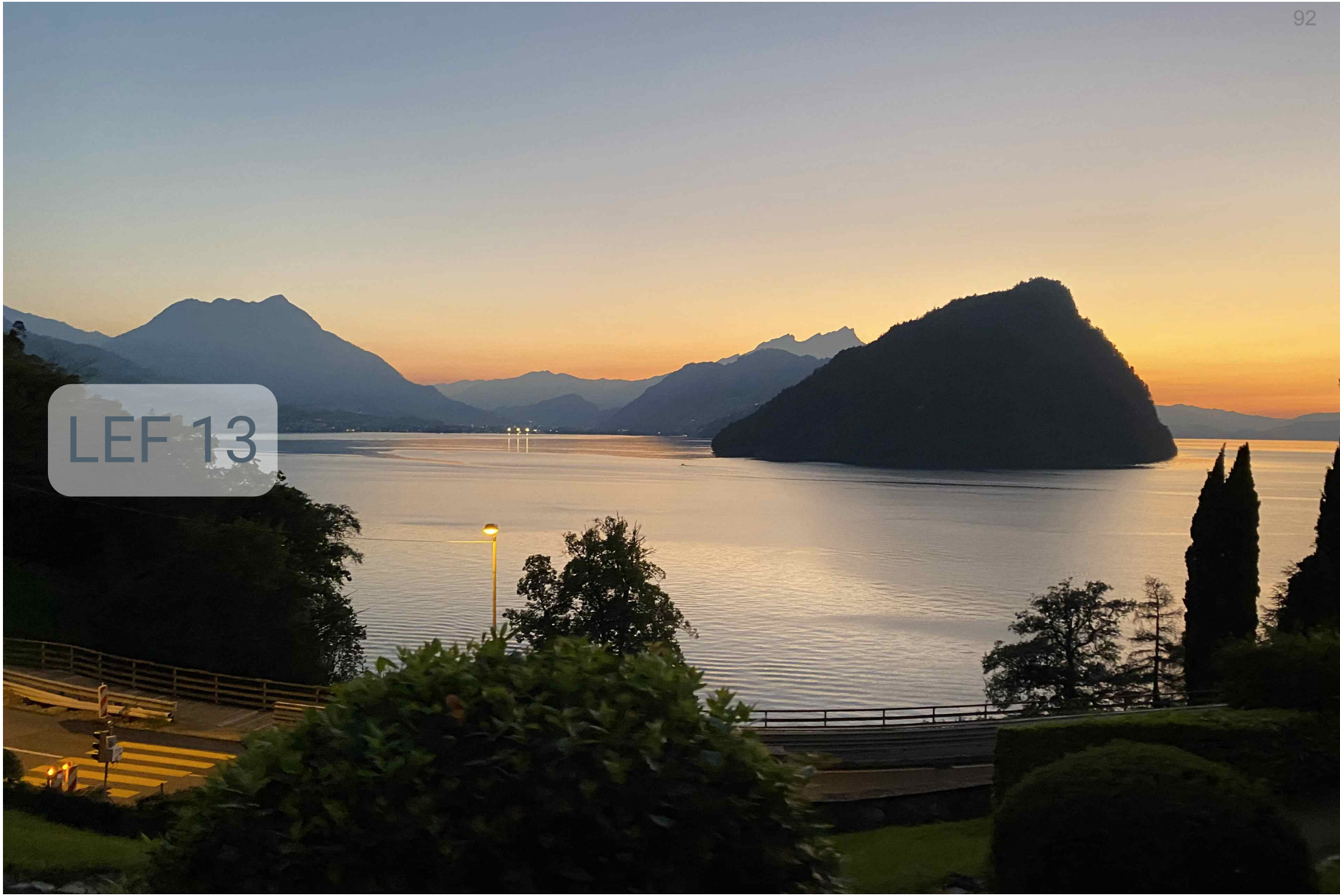
In Worten

Am Anfang die Anzahl (k) der Cluster festlegen. Dann werden Clusterzentren zufällig in den Variablenraum gelegt. Dann wird jeder Fall seinem nächsten Clusterzentrum zugeordnet. Dann das Clusterzentrum in den Mittelpunkt seines Clusters verlegt. Wieder werden alle Fälle ihren jeweils nächsten Clustern zugeordnet. Dann wieder Verschiebund der Clusterzentren usw. bis kein Fall mehr sein Cluster wechselt und keine Verschiebung der Clusterzentren mehr stattfindet.

Gütemass ist die Summe quadrierter Clusterzentrenabweichungen.

Ausblick

Im nächsten Input gehe ich die Lernerfolgsfragen durch. Das finden Sie in Vorbereitung auch schon in Kapitel 14.


LEF 13

Essayfragen 13

1. Wie in der Grafik **?@sec-modellguete** zu sehen war, kann die aufgeklärte Varianz auch mal grösser sein als Gesamtvarianz. Kann dann R^2 grösser sein als 1? Begründen Sie Ihre Antwort.
2. Wenn wir eine Regressionsanalyse mit einer UV haben mit gegebenem s_2^2 und unveränderbarer Fehlerstreuung s^2. Was können wir tun, wenn wir denken, dass das b in der Stichprobe eine relevante Grösse hat, aber nicht signifikant ist?

MC-Fragen 13

MC 13.1.

MC 13.1: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Clusteranalysen dienen der Zusammenfassung von Fällen in Gruppen.
○	○	Mit Clusteranalysen werden Variablen zusammengefasst.
○	○	Mit der Clusteranalyse werden möglichst unähnliche Fälle in Clustern zusammengefasst.
○	○	Die Clusteranalyse ist ein exploratives Verfahren.

Punkte: 0

MC 13.2.

MC 13.2: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Die Clusteranalyse wird im ML-Kontext als «unsupervised learning» behandelt.
○	○	Ziel der Clusteranalyse ist eine Reduktion der Fälle auf wenige untereinander homegene Cluster bei möglichst geringer Homogenität innerhalb der Cluster.
○	○	Häufig werden Clusteranalysen vor Faktorenanalysen durchgeführt, damit letztere besser fitten.
○	○	Damit die Variablen nicht hoch korrelieren, werden vor CAs oft FAs durchgeführt.

Punkte: 0

MC 13.3.

MC 13.3: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Clusteranalysen sind vor allem für die Analyse fehlender Werte geeignet.
○	○	Es gibt Clusteranalyseverfahren, die mit allen möglichen Skalenniveaus gut klarkommen.
○	○	Clusteranalysen funktionieren besonders gut mit kleinen Fallzahlen.
○	○	Es kann immer höchstens so viele Cluster geben, wie es Variablen gibt.

Punkte: 0

MC 13.4.

MC 13.4: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Proximitätsmasse sind Ähnlichkeits- beziehungsweise Distanzmasse.
○	○	Euklidische Distanzen können nur für metrische Variablen festgestellt werden.
○	○	Für eine Clusteranalyse werden immer mindestens so viele Variablen benötigt, wie man Cluster extrahieren will.
○	○	Manche Clusteralgorithmen nehmen als Ähnlichkeitsmass auch die Korrelation.

Punkte: 0

MC 13.5.

MC 13.5: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	k-means-Cluster kann nur auf metrische (und Dummys) angewendet werden.
○	○	Bei k-means-Cluster sind ähnliche Standardabweichungen der Variablen erwünscht.
○	○	Bei k-means-cluster müssen hohe Korrelationen zwischen den Variablen vorliegen.
○	○	Mit dem Ellbogenkriterium werden Clusterzentren einander zugeordnet.

Punkte: 0

MC 13.6.

MC 13.6: Mal was zu R?

richtig	falsch	Aussagen
○	○	Clusteranalysen haben eher deskriptiven Analysegehalt.
○	○	Clusteranalysen werden häufig eingesetzt um Typologien zu bilden.
○	○	Hierarchische Clusteranalysen sind so rechenaufwendig, dass selbst moderne Rechner eher verrottet sind als sie mit der Analyse fertig werden.
○	○	Die Clusterzentren der k-means-Cluster korrelieren hoch miteinander.

Punkte: 0

Insgesamt 0 von 12 Punkten, was 0% und etwa einer 1 entspricht.