Computergestützte Datenanalyse: DATA-Übung mit R

Tag 4 – 25.07.2025

UNSER PLAN

- **Tag 1**
  - Einführung in R und RStudio
  - „Basics“:
  - Coding Konventionen
  - Objekte, Datenimport & Co
- **Tag 2**
  - Skalenniveau
  - Troubleshooting
  - Datenaufbereitung
  - Datenvisualisierung
  - Deskriptive Statistik
- **Tag 3**
  - Inferenzstatistik I
  - Bivariate Analyse
- **Tag 4**
  - Indexbildung
  - Inferenzstatistik II
  - Abschluss

Genereller Ablauf

- Vier Tage geblockt
- Mischung aus Input- und Übungssessions
- Anwesenheitsabfrage alle 90 Minuten

Heute

- Zwei 15 Minuten Pause
- Eine Mittagspause

REFRESHER

- Welche Datentypen/-klassen gibt es?
- Was ist der Unterschied zwischen Subsetting und Rekodieren?

  Welche Packages, Funktionen?
- Was untersucht der Chi^2-Test?
- Was ist der Unterschied zwischen Cramer‘s V und Pearson‘s R?

REFRESHER

- Was wird hier gerechnet?
- Wie interpretieren wir diesen Output?

```
cor.test(allbus$pt15,
allbus$pt03)
```

```
##
##  Pearson's product-moment correlation
##
## data:  allbus$pt15 and allbus$pt03
## t = 41.087, df = 3323, p-value < 2.2e-16
## alternative hypothesis: true correlation is not e
qual to 0
## 95 percent confidence interval:
##  0.5574159 0.6025157
## sample estimates:
##       cor
## 0.5804107
```

Getting started

1. Rmarkdown file erstellen und einstellen
2. Packages installieren und/ oder aktivieren: rio und tidyverse
3. Import der ALLBUS 2018 Daten
4. In einem **neuem** Dataframe werden die Variablen pt01 bis pt20 gespeichert. Nutzt dafür den Befehl select() und recherchiert, wie das funktioniert

Indexbildung und Reliabilitätsanalyse

Indexbildung

- Manchmal müssen wir, um ein Konstrukt angemessen darzustellen, mehrere Einzelindikatoren zusammenfassen
- Erfassung vieler theoretische Merkmale eines abstrakten Begriffs
  - Beispielsweise setzt sich das Phänomen der **Lebensqualität** aus den Einzelindikatoren Glück, Zufriedenheit und Wohlbefinden zusammen
- Reduktion sozialer Erwünschtheit
  - Beispielsweise Messung rechtsextremer Einstellung
- Reduktion des Messfehlers
  - Testtheorie: Reliabilität definiert als Genauigkeit, „‚mit der eine Skala ein Merkmal misst.“ (Rammstedt, 2010, S. 242)

Voraussetzung

- Alle Variablen müssen in die gleiche Richtung kodiert sein (ggf. rekodieren)
- Beispiel „Big Five“ zur Messung von Persönlichkeit (Rammstedt et al., 2014)
  - „Extraversion“ als eine Dimension der Persönlichkeit, wird über zwei Einzelindikatoren abgefragt
  - Für die Antworten der Befragungsperson steht eine fünfstufige Ratingskala von "trifft überhaupt nicht zu" (1) bis "trifft voll und ganz zu" (5) zur Verfügung.
    - „Ich bin eher zurückhaltend, reserviert.“ (negative Polung)
    - „Ich gehe aus mir heraus, bin gesellig.“ (positive Polung)

Voraussetzung

- Reliabilitätsanalyse: Interne Konsistenz im Antwortverhalten
  - Können wir die Einzelindikatoren zu einem Index zusammenfassen?
- Maßzahl Cronbachs Alpha
  - Ergibt sich aus der Anzahl der Einzelindikatoren und der (durchschnittlichen) Korrelation
  - Nimmt maximal den Wert 1 an
  - Unser Cut-Off-Kriterium > 0.7
  - Alpha steigt allerdings mit der Anzahl der Indikatoren
  - Es kann sein, dass sich Alpha stark verbessert, wenn wir ein Item aus dem Index weglassen (ggf. theoretisch begründen!)

Aufgabe

Korrelationsmatrix:

```
allbus_sub <- allbus %>%
select(px01:px10)
```

Um welche Variablen handelt es sich?

```
cor(allbus_sub,
        use = "pairwise.complete.obs",
        method = "pearson")
```

Was seht ihr?

Cronbachs Alpha

- Wir nutzen die Funktion alpha() aus dem Psych-Package

```
Reliability analysis
Call: alpha(x = df[trust], check.keys = TRUE)

 raw_alpha std.alpha G6(smc) average_r S/N    ase mean   sd median_r
       0.9       0.9    0.91       0.4 8.8 0.0025  4.2 0.94     0.38

 Reliability if an item is dropped:
     raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
pt01      0.90      0.90    0.92      0.43 8.9   0.0025 0.016  0.39
pt02      0.89      0.89    0.91      0.40 8.1   0.0027 0.017  0.37
pt03      0.88      0.88    0.90      0.39 7.7   0.0029 0.015  0.36
pt04      0.90      0.90    0.91      0.42 8.6   0.0026 0.017  0.39
pt08      0.89      0.89    0.91      0.40 7.9   0.0028 0.018  0.37
pt09      0.90      0.89    0.91      0.41 8.5   0.0026 0.016  0.39
pt10      0.89      0.89    0.90      0.41 8.2   0.0027 0.017  0.37
pt11      0.89      0.89    0.91      0.41 8.5   0.0026 0.018  0.38
pt12      0.88      0.88    0.90      0.39 7.6   0.0029 0.015  0.36
pt14      0.89      0.89    0.91      0.41 8.3   0.0026 0.018  0.38
pt15      0.89      0.89    0.91      0.39 7.8   0.0028 0.016  0.37
pt19      0.89      0.88    0.89      0.39 7.7   0.0028 0.013  0.37
pt20      0.89      0.89    0.90      0.39 7.7   0.0028 0.013  0.37
```

Aufgabe

Korrelationsmatrix:

```
allbus_sub <- allbus %>%
select(px01:px10)
```

Um welche Variablen handelt es sich?

```
cor(allbus_sub,
        use = "pairwise.complete.obs",
        method = "pearson")
```

Was seht ihr?

Aufgabe

Korrelationsmatrix:

install.packages(“psych“)

library(psych)

alpha(allbus_sub)

Was seht ihr?

raw_alpha <dbl>	std.alpha <dbl>	G6(smc) <dbl>	average_r <dbl>	S/N <dbl>	ase <dbl>	mean <dbl>	sd <dbl>	median_r <dbl>
0.8176307	0.8176685	0.8364781	0.3096076	4.484517	0.004531764	2.253593	0.6997273	0.3046075

	raw_alpha <dbl>	std.alpha <dbl>	G6(smc) <dbl>	average_r <dbl>	S/N <dbl>	alpha se <dbl>	var.r <dbl>	med.r <dbl>
px01	0.8179796	0.8199486	0.8247519	0.3359879	4.553970	0.004544392	0.01240284	0.3182501
px02	0.8064941	0.8076461	0.8136710	0.3181173	4.198751	0.004793231	0.01703049	0.3141702
px03	0.8118608	0.8119047	0.8300114	0.3241444	4.316454	0.004704338	0.01869410	0.3141702
px04	0.7947232	0.7934270	0.8113661	0.2991147	3.840903	0.005106118	0.01963501	0.2861827
px05	0.8070169	0.8065751	0.8252992	0.3166268	4.169964	0.004811441	0.01953043	0.3075234
px06	0.7889937	0.7902014	0.8119407	0.2950286	3.766476	0.005362700	0.02031510	0.2745536
px07	0.7943422	0.7946189	0.8163511	0.3006449	3.868998	0.005153922	0.01971158	0.3041859
px08	0.7892032	0.7891120	0.8025894	0.2936664	3.741854	0.005272843	0.01640828	0.2997131
px09	0.7910886	0.7897658	0.8021216	0.2944829	3.756600	0.005214541	0.01621413	0.2997131
px10	0.8077998	0.8077501	0.8287183	0.3182625	4.201562	0.004824466	0.02018835	0.3015367

Indexbildung

- Sind die Variablen richtig kodiert und Cronbachs Alpha entsprechend hoch, können wir unseren Index bilden (zumindest aus einer ersten methodischen Perspektive)
    - Mittelwertindex
    - Additiver Index

Mittelwertindex

- Der Mittelwertindex wird errechnet, in dem der Mittelwert aus den Einzelitems der Skala gebildet wird.
  - Mit rowMeans()
  - Oder mit mutate

```
allbus$? <- rowMeans(allbus_sub, na.rm = FALSE)
```

```
allbus <- allbus %>%
 mutate(?  = rowMeans(across(px01:px10), na.rm = FALSE))
```

Additiver Index

- Der additive Index wird errechnet, in dem die Summe aus den Einzelitems der Skala gebildet wird.
  - Mit rowSums()
  - Oder mit mutate()

```
allbus$? <- rowSums(allbus_sub, na.rm = FALSE)
```

```
allbus <- allbus %>%
 mutate(?  = rowSums(across(px01:px10), na.rm = FALSE))
```

Weiterführende Methoden

- Explorative Faktorenanalyse (EFA)
- konfirmatorische Faktorenanalyse (CFA)
- Mit lavaan-Package in R

Aufgabe

- Ein Index für Vertrauen in Institutionen?
- Können Sie einen Index zum allgemeinen Vertrauen erstellen?
- Verbessert sich die Reliabilität, wenn Sie ein Item weglassen?
- Mittelwert und Additiven Index erstellen!
- Plus: Handelt es sich bei "Vertrauen" um ein eindimensionales Konzept? Und könnte es mit der Anzahl an Variablen zu Problemen mit Cronbach's Alpha kommen?

REFRESHER

- Warum wollen wir einen Index bilden?
- Was sind Voraussetzungen?

Inferenzstatistik

Inferenzstatistik

- Tests zur Überprüfung von Unterschiedshypothesen
  - T-Test (für unabhängige Stichproben)
- Korrelationen
  - Pearson‘s r, Kendall‘s tau und Spearman‘s Rho
- Lineare Regression
  - Diagnostik
  - Multiple lineare Regression

Statistischer Test auf Gruppenunterschiede

T-Test (für unabhängige Stichproben)

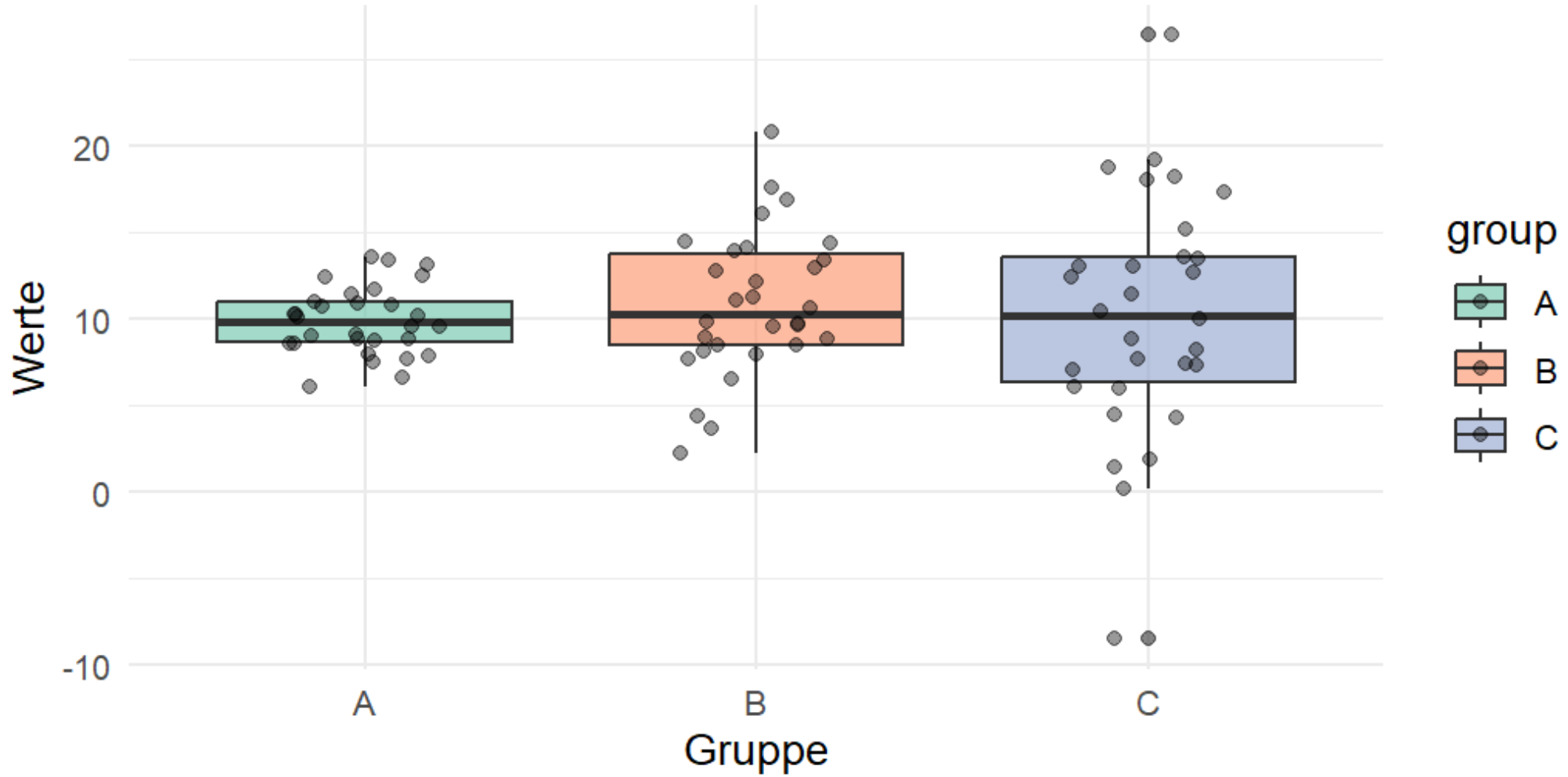
- Gibt es einen Unterschied zwischen (genau) zwei Gruppen?
- Gruppierungsvariable ist dichotom/Testvariable ist intervallskaliert
- Test der Nullhypothese: Es besteht kein Unterschied (shortcut: t-Wert +/- 2)

```
t.test(allbus$pt20 ~ allbus$westost, na.rm=TRUE)
##
##  Welch Two Sample t-test
##
## data:  allbus$pt20 by allbus$westost
## t = 6.9701, df = 1882.3, p-value = 4.365e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2748059 0.4900062
## sample estimates:
## mean in group Westdeutschland  mean in group Ostdeutschland
##                      3.712366                      3.329960
```

T-Test Voraussetzungen/ Ablauf

- Annahmen eines T-Tests
  - Test auf Varianzhomogenität (Levene-Test)
  - Per Default berechnet R den Welch‘s t-Test, der keine Varianzhomogenität voraussetzt
  - Literatur: Why Psychologists Should by Default Use Welch’s t-test Instead of Student’s t-test

Überprüfung der Varianzhomogenität



T-Test Voraussetzungen/ Ablauf

- Annahmen eines T-Tests
  - Test auf Varianzhomogenität (Levene-Test)
  - Per Default berechnet R den Welch‘s t-Test, der keine Varianzhomogenität voraussetzt
  - Literatur: Why Psychologists Should by Default Use Welch’s t-test Instead of Student’s t-test
- Test auf Normalverteilung
  - Bei Verletzung der Normalverteilung ggf. Wilcoxon/Mann-Whitney Test

Aufgabe

- Unterschiede zwischen Personen aus West/-Ostdeutschland beim Vertrauen in das Europäische Parlament (pt20)

```
allbus$westost <- factor(allbus$eastwest, labels =
c("Westdeutschland", "Ostdeutschland"))

table(allbus$westost)


tapply(allbus$pt20, allbus$westost, mean, na.rm=TRUE)
```

Aufgabe

- Levene Test, um Varianzhomogenität zu testen

```
install.packages(“car)
library(car)

class(allbus$pt20) #checken, ob die Variable numerisch ist
leveneTest(allbus$pt20, allbus$westost)
```

Aufgabe

- T-Test berechnen

```
t.test(allbus$pt20 ~allbus$westost, na.rm = TRUE)
t.test(allbus$pt20 ~allbus$westost, na.rm = TRUE, var.equal = FALSE)
t.test(allbus$pt20 ~allbus$westost, na.rm = TRUE, var.equal =TRUE)
```

t.test(allbus$pt20 ~allbus$westost, na.rm = TRUE)

```
        Welch Two Sample t-test

data:  allbus$pt20 by allbus$westost
t = 6.9701, df = 1882.3, p-value = 4.365e-12
alternative hypothesis: true difference in means between group Westdeutschland and group
Ostdeutschland is not equal to 0
95 percent confidence interval:
 0.2748059 0.4900062
sample estimates:
mean in group Westdeutschland  mean in group Ostdeutschland
                     3.712366                      3.329960
```

Ein statistischer Bericht der Auswertung könnte zusammenfassend folgendermaßen lauten:

Personen in Ost- und Westdeutschland unterscheiden sich signifikant voneinander in ihrem Vertrauen in das europäische Parlament, $t(1882.3)=6.971$, $p<.05$. Westdeutsche ($M=3.71$) weisen dabei ein höheres Vertrauen auf als Ostdeutsche ($M=3.33$).

Optional: Standardvarianz mitangeben

Aufgabe

- Gibt es Gruppenunterschiede zwischen binärer Geschlechtkodierung und Vertrauen in Bundestag?
- Welche Variablen?
- Mittelwerte anschauen
- Levene Test
- T-Test durchführen
- Report schreiben

Wiederholung: Korrelation

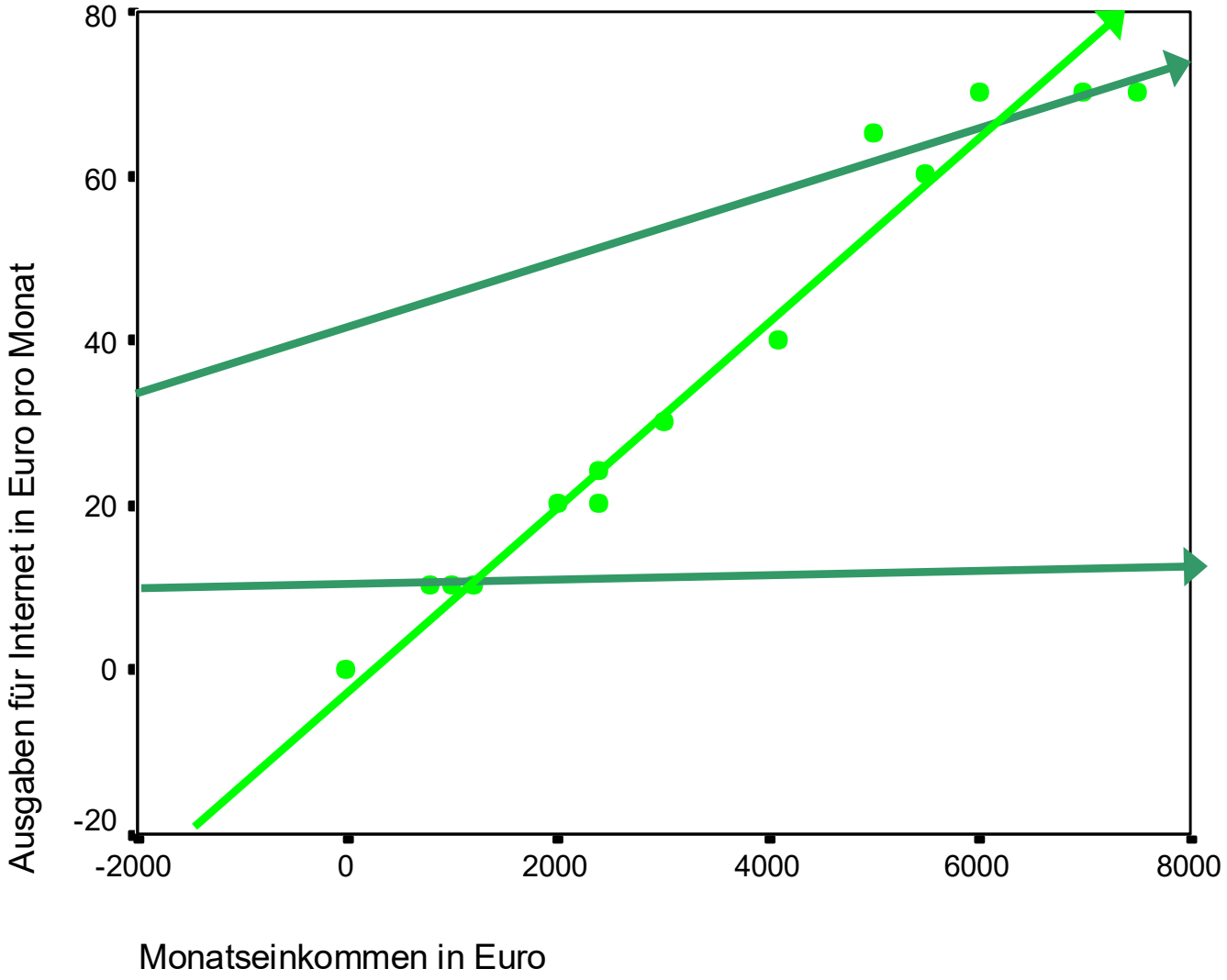
Drei wesentliche Punkte zur Interpretation der Korrelation

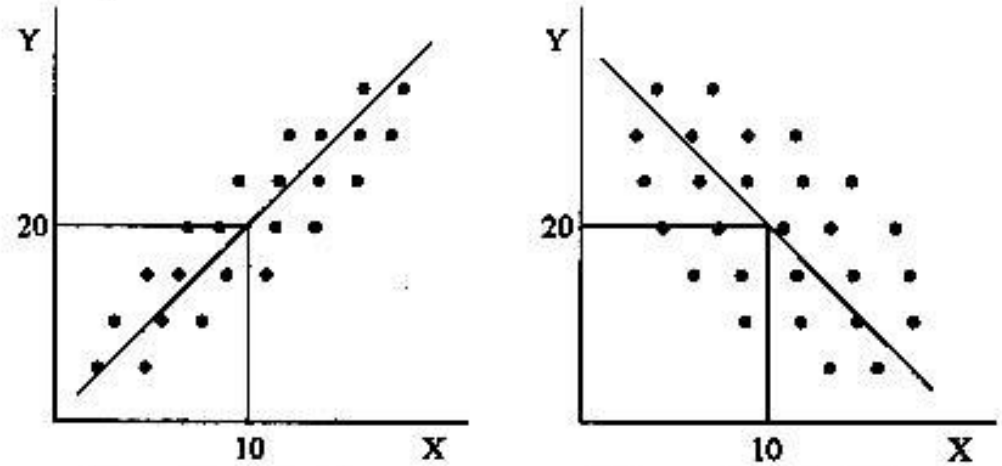
1. Wie stark eine Variable mit der anderen Variable zusammenhängt, wird in der Höhe des Koeffizienten angegeben. Je höher dieser Wert ist, desto stärker wird eine Variable durch die andere Variable bestimmt. Der Wert kann zwischen 0 und +-1 liegen
2. Richtung des Zusammenhangs: Hier kommt es auf das Vorzeichen des Koeffizienten an. Bei einem + sprechen wir von einem positiven Zusammenhang, während wir bei einem - von einem negativen Zusammenhang sprechen.
3. Signifikanztest: Hier wird überprüft, ob wir auch in der Grundgesamtheit von einem Zusammenhang zwischen den beiden Variablen ausgehen können. Wenn der p-value kleiner als 0.05 ist, können wir davon ausgehen, dass ein Zusammenhang zwischen den beiden Variablen auch in der Grundgesamtheit vorliegt

Lineare Regression

Lineare Regression

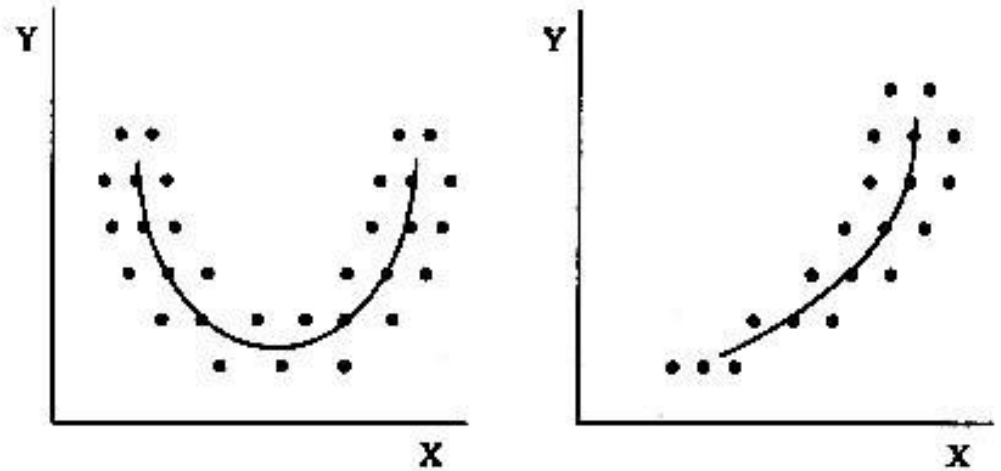
- Lineare Regression als Verfahren zur Schätzung des Einflusses einer (oder mehrerer) Variable(n) auf eine abhängige (metrische) Variable
  - Inwieweit kann ein Merkmal auf andere Merkmale „zurückgeführt“ werden
  - In den SoWi wohl am häufigsten verwendete Analyseverfahren
  - Typische Forschungsfrage: Wie stark ist der Einfluss der Berufserfahrung auf das Einkommen? Welche Faktoren beeinflussen die Lebenszufriedenheit? Hat eine Zunahme des Umweltwissens eine Veränderung des Umweltverhaltens zur Folge? (Beispiele aus Wolf und Best, 2010)


80
60
40
20
0
-20
Ausgaben für Internet in Euro pro Monat
-2000
0
2000
4000
6000
8000
Monatseinkommen in Euro



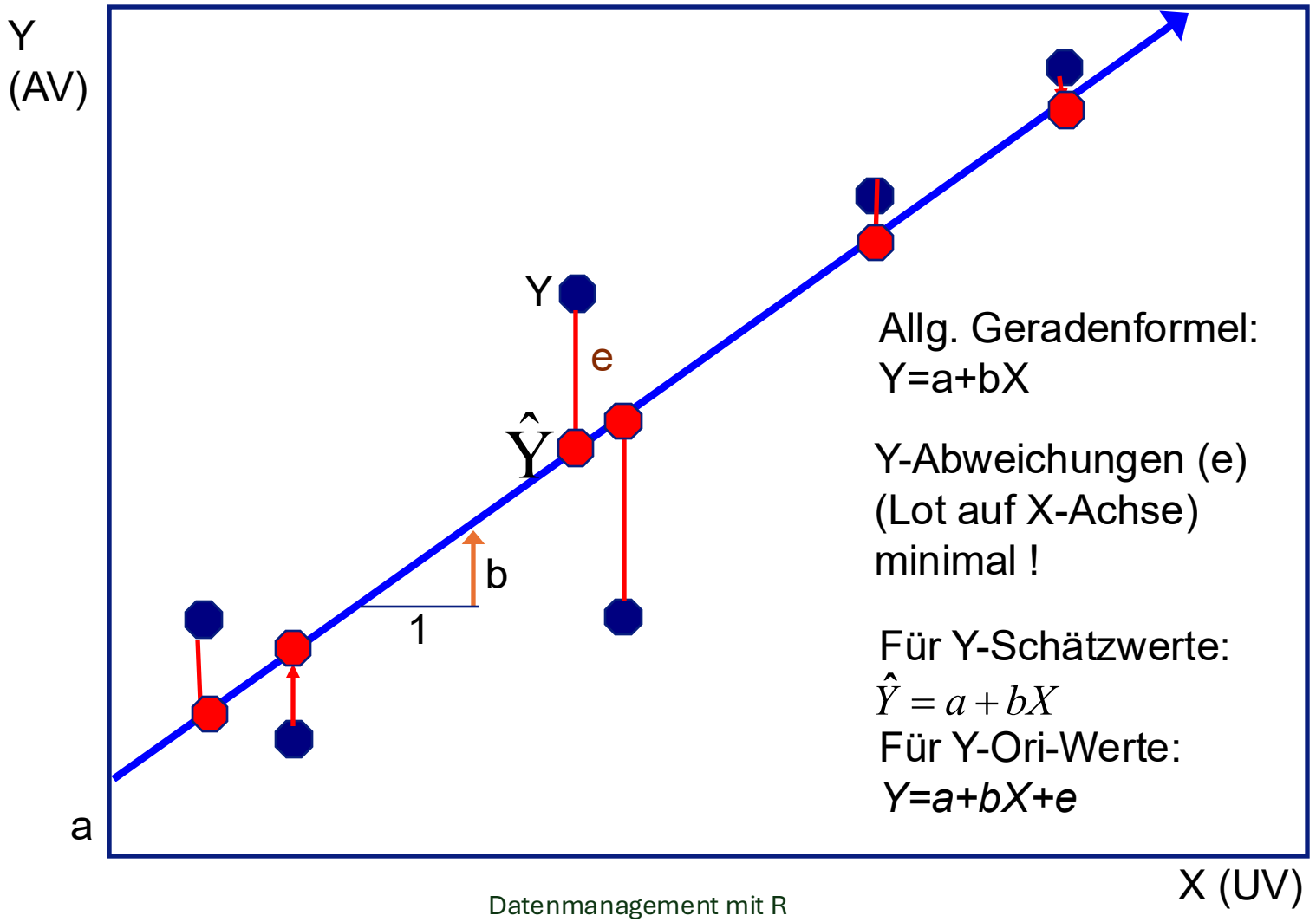
(a) positive Beziehung (b) negative Beziehung

Lineare Beziehungen



(c) u-förmige Beziehung (d) j-förmige Beziehung

Kurvilineare Beziehungen


Y
(AV)
Y
e
Ŷ
b
1
a
X (UV)
Allg. Geradenformel:
Y=a+bX
Y-Abweichungen (e)
(Lot auf X-Achse)
minimal !
Für Y-Schätzwerte:
$\hat{Y} = a + bX$
Für Y-Ori-Werte:
Y=a+bX+e

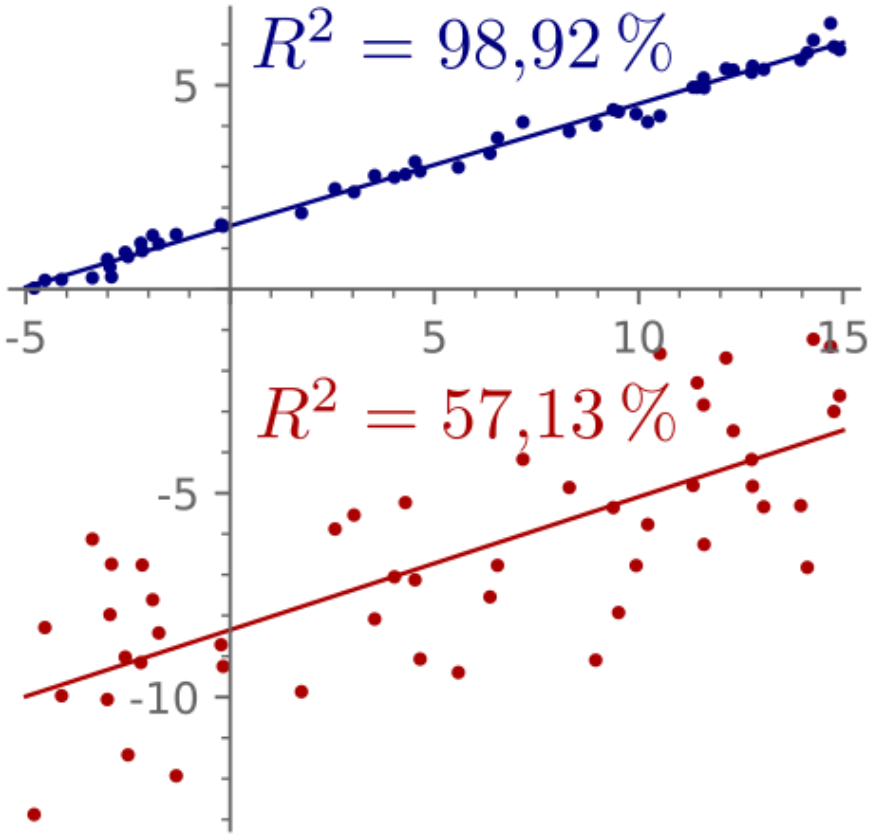
- Die Regressionsgerade wird mit der allgemeinen Geradenformel beschrieben:
  - a = Achsenabschnitt
  - b = Steigungskoeffizient
    (Steigung der Gerade pro X- Einheit)
- Die Y-Punkte werden dadurch nicht exakt abgebildet, sondern nur geschätzt (mit Fehlern).

Der Schätzfehler heisst e (das Residuum).

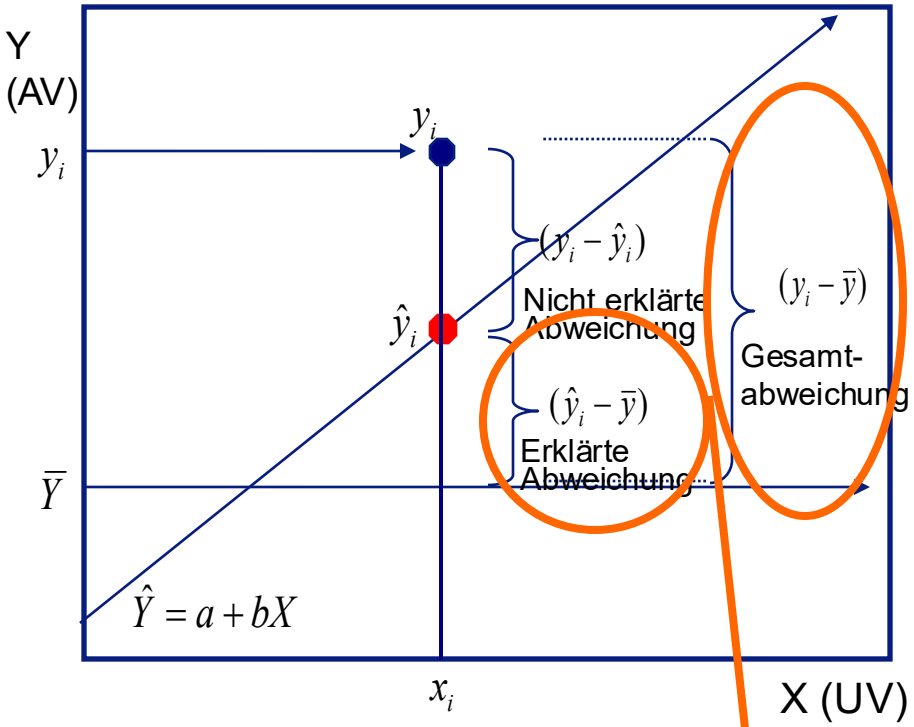
$$Y = a + bX_i + e_i$$

mit $$e_i = Y_i - \hat{Y}_i$$

Bestimmtheitsmaß



https://de.wikipedia.org/wiki/Bestimmtheitsma%C3%9F#Rechenbeispiel



- Gesamtvariation: $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- erklärte Variation: (Regressionsvariation) $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- nicht erklärte Variation: (Residualvariation, wird minimiert) $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Erklärte Varianz in %:
(Bestimmtheitsmass oder auch Determinationskoeffizient)

$$R^2 = \frac{\text{erklärte Varianz}}{\text{Gesamtvarianz}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Lineare Regression mit R

```
model <- lm(y ~ x, data=df)
summary(model)
```

Lineare Regression

model <- lm(pt12 ~ age, data = allbus)
summary(model)

- Beispiel einer bivariaten linearen Regression
  - Vertrauen in die Bundesregierung zurückgeführt auf das (metrische) Alter

```
#Regression mit Alter
model1 <- lm(pt12 ~ age, data = allbus)
summary(model1)
##
## Call:
## lm(formula = pt12 ~ age, data = allbus)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.01796 -0.98122  0.02298  1.02088  3.05552
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.03686    0.07763  51.998   <2e-16 ***
## age         -0.00105    0.00142  -0.739     0.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.462 on 3427 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.0001595,  Adjusted R-squared:  -0.0001322
## F-statistic: 0.5467 on 1 and 3427 DF,  p-value: 0.4597
```

Multiple Regressionen

Typische Forschungsfragen

- Welchen Einfluss haben sozioökonomische und demografische Merkmale auf das Nettoeinkommen der Befragten? Gibt es Einkommensunterschiede zwischen Ost- und Westdeutschland? (vereinfachtes Beispiel angelehnt an Wolf & Best, 2010)
  - Wir filtern die Daten: nur „berufstätige“ Befragte
  - Abhängige Variable: individuelles Nettoeinkommen
  - Unabhängige Variablen: Alter, Geschlecht, Bildung (kategorisiert) und Ost-/Westdeutschland

Multiple Regressionen in R

```
model <- lm(AV ~ UV1 + UV2 + UV3... , data = allbus)
summary(model)

allbus$ost <- as.factor(allbus$ostwest)
table(allbus$ost)
table(allbus$ost, allbus$eastwest)

model2 <- lm(pt12 ~ ost, data = allbus)
summary(model2)

model3 <- lm(pt12 ~ age + ost, data = allbus)
summary(model3)
```

Beta Koeffizienten

```
install.packages("lm.beta“)
library(lm.beta)


model <- lm(y ~ x1 + x2, data = yourdata)
lm.beta(model)

lm.beta(model3)
```

Regressionsdiagnostik

- Mit plot(model)
  - Normalverteilung der Residuen über einen Q-Q-Plot
  - Homoskedastizität der Residuen
  - Übersicht möglicher einflussreicher Ausreißer auf die Regressionsgerade
- Mit vif(model)
  - Multikollinearität
- Mit bptest(model)
  - Heteroskedastizität
  - Wenn der p-Wert nahe 0 liegt, kann die Nullhypothese der Homoskedastizität abgelehnt wird.

BN

- Abgabetermin besprechen?
- Beispiel

WIEDERHOLUNG

FRAGEN, UNKLARHEITEN, FEEDBACK?

VIELEN DANK ☺

ahrabhi.kathirgamalingam@cais-research.de