Einführung in die Datenanalyse mit R: Datenvisualisierung

Marco Wähner


Gebäude 37.03.03.14


marco.waehner@hhu.de


Sprechstunde nach Vereinbarung

Inferenzstatistik

- Tests zur Überprüfung von Unterschiedshypothesen
  - T-Test (für unabhängige Stichproben; Kapitel 8)
- Korrelationen
  - Pearson‘s r (Kapitel 10)
  - Visualisierung von Korrelationen
  - Funktion zu Korrelationen
- Lineare Regression
  - Multiple lineare Regression (Kapitel 12; weiterführendes Kapitel u.a. zum Bestimmtheitsmaß R-Quadrat)

Disclaimer: Sowohl die Analysemethoden als auch unterschiedliche (statistische) Voraussetzungen können wir nur punktuell thematisieren.

T-Test

- 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

- Annahmen eines T-Tests (u.a. Kapitel 8.2.4)
  - Test auf Varianzhomogenität (Levene-Test)
    - Per Default berechnet R den Welch‘s t-Test, der kein Varianzhomogenität voraussetzt (liegt Varianzhomogenität vor, dann wird der Student‘s t-Test verwendet)
  - Test auf Normalverteilung über Shapiro-Wilk Test
    - Bei Verletzung der Normalverteilung ggf. Wilcoxon/Mann-Whitney Test

Korrelation

- Zusammenhangsmaße zwischen Variablen, u.a.
  - Pearson‘s r, Kendall‘s tau und Spearman‘s Rho

- Pearson‘s r
  - Normiertes Zusammenhangsmaß (+/- 1) zwischen zwei metrischen Variablen
  - Spezialfall: Punktbiseriale Korrelation zwischen einer metrischen und einer dichotomen Variablen
  - Beispiel: Korrelation zwischen Vertrauen in politische Parteien und Vertrauen in den Bundestag

```
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 equal to 0
## 95 percent confidence interval:
```

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 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)
  - lm()-Funktion:

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

Lineare Regression

- 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
```

- 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

Praxis

Evaluation

Kontakt

Vielen Dank!

Marco Wähner

 Gebäude 37.03.03.14

 marco.waehner@hhu.de

 Sprechstunde nach Vereinbarung