

Tutorial 7

Lars Abt

Setup - Pakete laden

```
pacman::p_load(data.table)
```

Daten einlesen um Aufgaben zu bearbeiten

```
# Überprüfen des Working Directories, um die Datei korrekt zu adressieren  
getwd()
```

```
[1] "C:/Users/abtla/Documents/WHK/Angewandte Statistik/Muster/Tutorial 7/code"
```

```
# Sie liegt in input, und wir befinden uns in code  
# daher müssen wir zuerst eine Ebene hoch  
data <- fread("../input/data.csv")
```

```
# Überblick über die Daten  
head(data)
```

	V1	education	effort	wage
	<int>	<num>	<num>	<num>
1:	1	5.766500	3.879049	15.20008
2:	2	6.999380	4.539645	20.26398
3:	3	16.198873	8.117417	23.52689
4:	4	10.017683	5.141017	26.37817
5:	5	5.418465	5.258575	21.38783
6:	6	18.941407	8.430130	23.78392

Wir laden unsere vorbereiteten Daten aus der Datei `data.csv`, die wir von der Website herunterladen können. Dort befinden sich Informationen zu Bildung (`education`), Leistungseinsatz (`effort`) und Gehalt (`wage`).

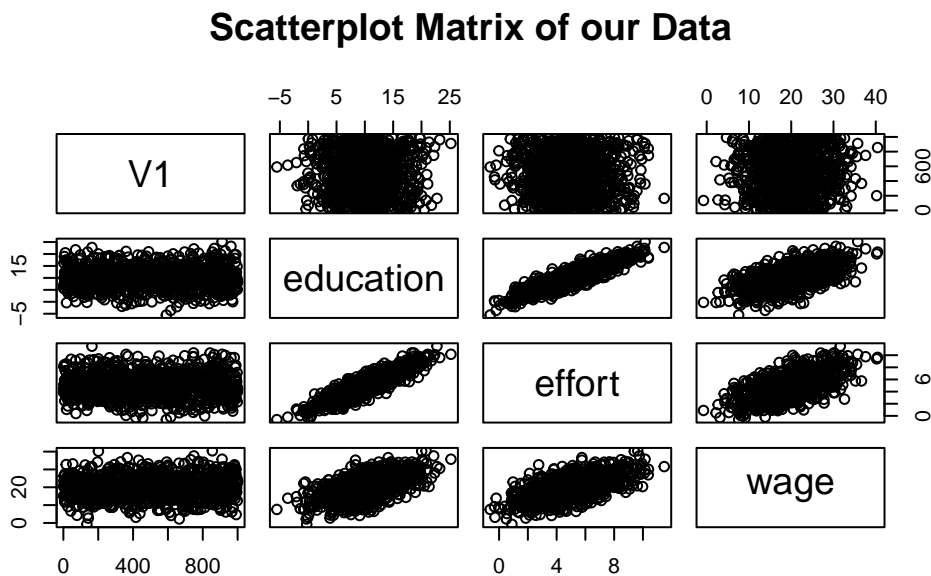
Confounder

Definition (aus der Vorlesung):

“Ein Confounder ist eine Variable, die sowohl mit der unabhängigen als auch der abhängigen Variable korreliert ist. Confounder können die Ergebnisse stark verzerren und sollten daher in der Analyse berücksichtigt werden.”

Unsere Hypothese: Die Leistungsbereitschaft erklärt sowohl das Gehalt einer Person, als auch das Bildungsniveau einer Person. Alternativ kann man aber auch davon ausgehen, dass das Bildungsniveau das Gehalt einer Person erklärt/beeinflusst (was in der Realität mit Sicherheit der Fall ist), oder das Bildungsniveau die Leistungsbereitschaft beeinflusst, quasi eine Art Wissen vom Wert dieser Leistungsbereitschaft. Oder frühe Früchte einer höheren Leistungsbereitschaft. Gehalt kann tendenziell eher weniger als Treiber von Bildungsniveau gesehen werden, womöglich als Motivator für berufliche Weiterbildung, aber vielleicht motivieren sich manche Personen zu mehr Leistung, um ihr (hohes) Gehalt zu rechtfertigen, oder eine Gehaltserhöhung anzustreben.

```
pairs(data, main = "Scatterplot Matrix of our Data")
```



```
# Lineares Modell, wo wir den vermeintlichen Confounder (effort) nicht beachten:  
model_no_confounder <- lm(wage ~ education, data = data)  
summary(model_no_confounder)
```

```
Call:
lm(formula = wage ~ education, data = data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-15.0183  -3.3303   0.1104   3.6243  16.6149
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1528     0.3933   30.89  <2e-16 ***
education     0.7696     0.0353   21.80  <2e-16 ***
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.137 on 998 degrees of freedom
Multiple R-squared:  0.3227,    Adjusted R-squared:  0.322
F-statistic: 475.4 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
# Lineares Modell, wo wir den vermeintlichen Confounder (effort) miteinbeziehen:
model_with_confounder <- lm(wage ~ education + effort, data = data)
summary(model_with_confounder)
```

```
Call:
lm(formula = wage ~ education + effort, data = data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-14.1801  -3.1386  -0.1849   3.2692  16.8933
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.16397     0.42313   24.021  <2e-16 ***
education     0.06877     0.07696    0.894    0.372
effort        1.80874     0.17866   10.124  <2e-16 ***
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.894 on 997 degrees of freedom
Multiple R-squared:  0.3858,    Adjusted R-squared:  0.3846
F-statistic: 313.1 on 2 and 997 DF,  p-value: < 2.2e-16
```

Wenn wir unsere Summaries betrachten können wir sehen, dass im ersten Modell der Variable `education` sehr hohe Signifikanz zugeschrieben wird, obwohl sie eigentlich nicht direkt mit `wage` zusammenhängt. Erst wenn wir den Confounder, `effort`, miteinbeziehen, sehen wir, dass `education` nicht mehr signifikant ist. Jetzt simulieren wir noch einmal neue Daten, mit einem neuen Confounder (`education`). Es ist zu beachten, dass es sich hier um künstlich erzeugte Daten ohne reale Grundlage handelt, sodass wir über die tatsächliche Kausalität in der realen Welt keine Aussage treffen können und wollen. Es geht lediglich darum, dass wir die Konzepte verstehen.

```
# Seed setzen um die Ergebnisse reproduzieren zu können
set.seed(34567)

# Anzahl der Datenpunkte festlegen
n <- 1000

# Simulieren der Variable `education_new`
education_new <- 5 + rnorm(n, mean = 0, sd = 2)

# Simulieren der Variable `effort_new` (Confounder), als Funktion unserer Variable `effort`
effort_new <- 2 * education_new + rnorm(n, mean = 0, sd = 2)

# Simulieren der Variable `wage`, welche durch effort beeinflusst wird, und dementsprechend
wage_new <- 10 + 2 * education_new + rnorm(n, mean = 0, sd = 5)

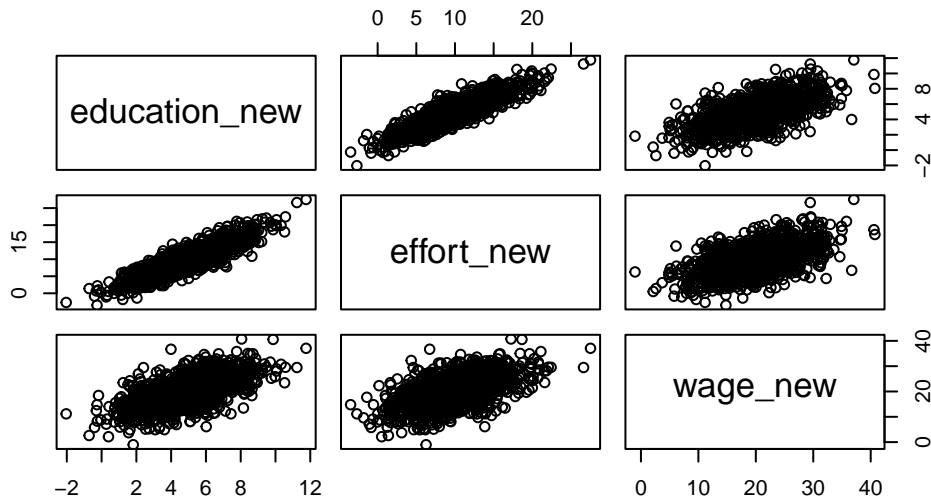
# Dataframe erstellen für unsere Daten
data_new <- data.frame(education_new, effort_new, wage_new)

# Summary unserer simulierten Daten ansehen
summary(data_new)
```

education_new	effort_new	wage_new
Min. :-2.035	Min. :-3.496	Min. :-1.029
1st Qu.: 3.616	1st Qu.: 6.786	1st Qu.:15.350
Median : 5.064	Median :10.018	Median :20.019
Mean : 4.984	Mean : 9.984	Mean :20.024
3rd Qu.: 6.449	3rd Qu.:13.164	3rd Qu.:24.519
Max. :11.759	Max. :27.524	Max. :40.733

```
# Scatterplot ansehen, um die Korrelationen zu visualisieren
pairs(data_new, main = "Scatterplot Matrix of Adjusted Data")
```

Scatterplot Matrix of Adjusted Data



```
# Lineares Modell, wo wir den Confounder (education) nicht beachten:  
model_no_confounder_new <- lm(wage_new ~ effort_new, data = data_new)  
summary(model_no_confounder_new)
```

Call:

```
lm(formula = wage_new ~ effort_new, data = data_new)
```

Residuals:

Min	1Q	Median	3Q	Max
-18.1781	-3.6308	-0.2947	3.7944	19.1779

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	12.31033	0.40411	30.46	<2e-16 ***
effort_new	0.77257	0.03682	20.98	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.304 on 998 degrees of freedom

Multiple R-squared: 0.306, Adjusted R-squared: 0.3054

F-statistic: 440.1 on 1 and 998 DF, p-value: < 2.2e-16

```
# Lineares Modell, wo wir den Confounder (education) miteinbeziehen:
model_with_confounder_new <- lm(wage_new ~ effort_new + education_new, data = data_new)
summary(model_with_confounder_new)
```

Call:

```
lm(formula = wage_new ~ effort_new + education_new, data = data_new)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.9409	-3.4960	-0.1717	3.5204	18.6077

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.230214	0.420250	24.34	<2e-16 ***
effort_new	-0.003048	0.075544	-0.04	0.968
education_new	1.971232	0.170672	11.55	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.984 on 997 degrees of freedom

Multiple R-squared: 0.3879, Adjusted R-squared: 0.3867

F-statistic: 316 on 2 and 997 DF, p-value: < 2.2e-16

Wir sehen also, dass wir den Confounder auch anders aufbauen können. Wichtig ist, dass wir genau hinschauen, wie korrelierte Konstrukte zusammenhängen und wie wir sie in unseren Modellen berücksichtigen. Es gilt die Phrase: *“Correlation does not imply causation.* (zu Deutsch: *“Korrelation impliziert keine Kausalität”*).

Kausale Graphen

Hier möchten wir die Zusammenhänge zwischen den angesprochenen Punkten diskutieren.

- Kriminalität und Anzahl der Kirchen könnten korrelieren, da große Städte/Siedlungen, welche bereits langer Zeit eine gewisse Größe haben, auch über einen gewissen Reichtum, eine gewisse wirtschaftliche Kraft verfügen. Dies macht sie attraktiv für kriminelle Vereinigungen, aber auch für große Religionsgemeinschaften.

- Bewegung und Fitness können angetrieben werden durch einen gesunden Lebensstil. Wer einen gesunden Lebensstil pflegt, der bewegt sich viel, der versucht einen hohen Fitnesslevel zu halten, um sich seine Gesundheit möglichst lange zu erhalten. Nicht unbedingt ein Confounder, da viel Bewegung mit Sicherheit einen positiven Einfluss auf die Fitness eines Menschen hat!
- Bio-Lebensmittel und eine Autismus-Diagnose können mit Self-Care erklärt werden. Wer viel Wert auf seine Gesundheit legt, kauft sich eher die gesünderen, aber teureren Lebensmittel, um sich selbst etwas Gutes zu tun. Genauso werden mögliche anderweitige gesundheitliche Einschränkungen und Probleme frühzeitig abgeklärt, und bis zur eindeutigen Diagnose verfolgt, um idealen Umgang damit zu gewährleisten.
- Sitzgurte und das Überleben bei Autounfällen können mit einem Sicherheitsbedürfnis erklärt werden. Menschen, welche keinen großen Wert auf Sicherheit legen, sehen wohl kaum die Notwendigkeit für Sitzgurte. Ebenso wird die Gefahr von schnellem Fahren wohl von dieser Personengruppe besonders unterschätzt. Nichtsdestotrotz haben Sitzgurte natürlich einen positiven Effekt auf die Überlebenschancen bei Autounfällen, da sie den Körper im Auto halten und somit Verletzungen durch das Herausschleudern verhindern.

Kausale Effekte

```
# Kausale Effekte an Hand von Beispieldaten anschauen
cities_df <- data.frame(
  city = c("Bielefeld", "Berlin", "Budapest", "Barcelona", "Bologna", "Bremen", "Bordeaux", "Buenos Aires", "Cairo", "Chicago", "Copenhagen", "Dallas", "Denver", "Detroit", "Frankfurt", "Hamburg", "Hongkong", "London", "Los Angeles", "Madrid", "Manila", "Miami", "Moskau", "New York", "Paris", "Peking", "Prag", "Rom", "San Francisco", "Shanghai", "Stockholm", "Tokio", "Wien", "Zürich"),
  unemployment_no_intervention = c(5.2, 6.3, 4.3, 8.1, 5.2, 4, 7.5, 3.4, 5.1, 6.8, 4.5, 7.2, 3.8, 5.5, 6.1, 4.9, 8.5, 3.2, 6.5, 5.8, 4.7, 7.9, 3.5, 6.2, 4.1, 5.3, 7.1, 3.9, 6.4, 5.6, 4.4, 8.2, 3.7, 6.7, 5.4, 4.6, 7.3, 3.6, 6.9, 5.7, 4.8, 8.3, 3.3, 6.6, 5.2, 4.5, 7.4, 3.8, 6.3, 5.1, 4.9, 8.1, 3.5, 6.8, 5.5, 4.7, 7.6, 3.9, 6.4, 5.3, 4.6, 8.4, 3.7, 6.7, 5.6, 4.8, 7.8, 3.6, 6.5, 5.4, 4.9, 8.2, 3.8, 6.9, 5.7, 4.7, 7.5, 3.9, 6.6, 5.5, 4.8, 8.3, 3.7, 6.8, 5.6, 4.9, 7.9, 3.8, 6.7, 5.7, 4.8, 8.4, 3.9, 6.9, 5.8, 5.0, 8.5, 4.0, 7.0, 5.9, 5.1, 8.6, 4.1, 7.1, 6.0, 5.2, 8.7, 4.2, 7.2, 6.1, 5.3, 8.8, 4.3, 7.3, 6.2, 5.4, 8.9, 4.4, 7.4, 6.3, 5.5, 9.0, 4.5, 7.5, 6.4, 5.6, 9.1, 4.6, 7.6, 6.5, 5.7, 9.2, 4.7, 7.7, 6.6, 5.8, 9.3, 4.8, 7.8, 6.7, 5.9, 9.4, 4.9, 7.9, 6.8, 6.0, 9.5, 5.0, 8.0, 6.9, 6.1, 9.6, 5.1, 8.1, 7.0, 6.2, 9.7, 5.2, 8.2, 7.1, 6.3, 9.8, 5.3, 8.3, 7.2, 6.4, 9.9, 5.4, 8.4, 7.3, 6.5, 10.0, 5.5, 8.5, 7.4, 6.6, 10.1, 5.6, 8.6, 7.5, 6.7, 10.2, 5.7, 8.7, 7.6, 6.8, 10.3, 5.8, 8.8, 7.7, 6.9, 10.4, 5.9, 8.9, 7.8, 7.0, 10.5, 6.0, 9.0, 7.9, 7.1, 10.6, 6.1, 9.1, 8.0, 7.2, 10.7, 6.2, 9.2, 8.1, 7.3, 10.8, 6.3, 9.3, 8.2, 7.4, 10.9, 6.4, 9.4, 8.3, 7.5, 11.0, 6.5, 9.5, 8.4, 7.6, 11.1, 6.6, 9.6, 8.5, 7.7, 11.2, 6.7, 9.7, 8.6, 7.8, 11.3, 6.8, 9.8, 8.7, 7.9, 11.4, 6.9, 9.9, 8.8, 8.0, 11.5, 7.0, 10.0, 8.9, 8.1, 11.6, 7.1, 10.1, 9.0, 8.2, 11.7, 7.2, 10.2, 9.1, 8.3, 11.8, 7.3, 10.3, 9.2, 8.4, 11.9, 7.4, 10.4, 9.3, 8.5, 12.0, 7.5, 10.5, 9.4, 8.6, 12.1, 7.6, 10.6, 9.5, 8.7, 12.2, 7.7, 10.7, 9.6, 8.8, 12.3, 7.8, 10.8, 9.7, 8.9, 12.4, 7.9, 10.9, 9.8, 9.0, 12.5, 8.0, 11.0, 9.9, 9.1, 12.6, 8.1, 11.1, 10.0, 9.2, 12.7, 8.2, 11.2, 10.1, 9.3, 12.8, 8.3, 11.3, 10.2, 9.4, 12.9, 8.4, 11.4, 10.3, 9.5, 13.0, 8.5, 11.5, 10.4, 9.6, 13.1, 8.6, 11.6, 10.5, 9.7, 13.2, 8.7, 11.7, 10.6, 9.8, 13.3, 8.8, 11.8, 10.7, 9.9, 13.4, 8.9, 11.9, 10.8, 10.0, 13.5, 9.0, 12.0, 10.9, 10.1, 13.6, 9.1, 12.1, 11.0, 10.2, 13.7, 9.2, 12.2, 11.1, 10.3, 13.8, 9.3, 12.3, 11.2, 10.4, 13.9, 9.4, 12.4, 11.3, 10.5, 14.0, 9.5, 12.5, 11.4, 10.6, 14.1, 9.6, 12.6, 11.5, 10.7, 14.2, 9.7, 12.7, 11.6, 10.8, 14.3, 9.8, 12.8, 11.7, 10.9, 14.4, 9.9, 12.9, 11.8, 11.0, 14.5, 10.0, 13.0, 11.9, 11.1, 14.6, 10.1, 13.1, 12.0, 11.2, 14.7, 10.2, 13.2, 12.1, 11.3, 14.8, 10.3, 13.3, 12.2, 11.4, 14.9, 10.4, 13.4, 12.3, 11.5, 15.0, 10.5, 13.5, 12.4, 11.6, 15.1, 10.6, 13.6, 12.5, 11.7, 15.2, 10.7, 13.7, 12.6, 11.8, 15.3, 10.8, 13.8, 12.7, 11.9, 15.4, 10.9, 13.9, 12.8, 12.0, 15.5, 11.0, 14.0, 12.9, 12.1, 15.6, 11.1, 14.1, 13.0, 12.2, 15.7, 11.2, 14.2, 13.1, 12.3, 15.8, 11.3, 14.3, 13.2, 12.4, 15.9, 11.4, 14.4, 13.3, 12.5, 16.0, 11.5, 14.5, 13.4, 12.6, 16.1, 11.6, 14.6, 13.5, 12.7, 16.2, 11.7, 14.7, 13.6, 12.8, 16.3, 11.8, 14.8, 13.7, 12.9, 16.4, 11.9, 14.9, 13.8, 13.0, 16.5, 12.0, 15.0, 13.9, 13.1, 16.6, 12.1, 15.1, 14.0, 13.2, 16.7, 12.2, 15.2, 14.1, 13.3, 16.8, 12.3, 15.3, 14.2, 13.4, 16.9, 12.4, 15.4, 14.3, 13.5, 17.0, 12.5, 15.5, 14.4, 13.6, 17.1, 12.6, 15.6, 14.5, 13.7, 17.2, 12.7, 15.7, 14.6, 13.8, 17.3, 12.8, 15.8, 14.7, 13.9, 17.4, 12.9, 15.9, 14.8, 14.0, 17.5, 13.0, 16.0, 14.9, 14.1, 17.6, 13.1, 16.1, 15.0, 14.2, 17.7, 13.2, 16.2, 15.1, 14.3, 17.8, 13.3, 16.3, 15.2, 14.4, 17.9, 13.4, 16.4, 15.3, 14.5, 18.0, 13.5, 16.5, 15.4, 14.6, 18.1, 13.6, 16.6, 15.5, 14.7, 18.2, 13.7, 16.7, 15.6, 14.8, 18.3, 13.8, 16.8, 15.7, 14.9, 18.4, 13.9, 16.9, 15.8, 15.0, 18.5, 14.0, 17.0, 15.9, 15.1, 18.6, 14.1, 17.1, 16.0, 15.2, 18.7, 14.2, 17.2, 16.1, 15.3, 18.8, 14.3, 17.3, 16.2, 15.4, 18.9, 14.4, 17.4, 16.3, 15.5, 19.0, 14.5, 17.5, 16.4, 15.6, 19.1, 14.6, 17.6, 16.5, 15.7, 19.2, 14.7, 17.7, 16.6, 15.8, 19.3, 14.8, 17.8, 16.7, 15.9, 19.4, 14.9, 17.9, 16.8, 16.0, 19.5, 15.0, 18.0, 16.9, 16.1, 19.6, 15.1, 18.1, 17.0, 16.2, 19.7, 15.2, 18.2, 17.1, 16.3, 19.8, 15.3, 18.3, 17.2, 16.4, 19.9, 15.4, 18.4, 17.3, 16.5, 20.0, 15.5, 18.5, 17.4, 16.6, 20.1, 15.6, 18.6, 17.5, 16.7, 20.2, 15.7, 18.7, 17.6, 16.8, 20.3, 15.8, 18.8, 17.7, 16.9, 20.4, 15.9, 18.9, 17.8, 17.0, 20.5, 16.0, 19.0, 17.9, 17.1, 20.6, 16.1, 19.1, 18.0, 17.2, 20.7, 16.2, 19.2, 18.1, 17.3, 20.8, 16.3, 19.3, 18.2, 17.4, 20.9, 16.4, 19.4, 18.3, 17.5, 21.0, 16.5, 19.5, 18.4, 17.6, 21.1, 16.6, 19.6, 18.5, 17.7, 21.2, 16.7, 19.7, 18.6, 17.8, 21.3, 16.8, 19.8, 18.7, 17.9, 21.4, 16.9, 19.9, 18.8, 18.0, 21.5, 17.0, 20.0, 18.9, 18.1, 21.6, 17.1, 20.1, 19.0, 18.2, 21.7, 17.2, 20.2, 19.1, 18.3, 21.8, 17.3, 20.3, 19.2, 18.4, 21.9, 17.4, 20.4, 19.3, 18.5, 22.0, 17.5, 20.5, 19.4, 18.6, 22.1, 17.6, 20.6, 19.5, 18.7, 22.2, 17.7, 20.7, 19.6, 18.8, 22.3, 17.8, 20.8, 19.7, 18.9, 22.4, 17.9, 20.9, 19.8, 19.0, 22.5, 18.0, 21.0, 19.9, 19.1, 22.6, 18.1, 21.1, 20.0, 19.2, 22.7, 18.2, 21.2, 20.1, 19.3, 22.8, 18.3, 21.3, 20.2, 19.4, 22.9, 18.4, 21.4, 20.3, 19.5, 23.0, 18.5, 21.5, 20.4, 19.6, 23.1, 18.6, 21.6, 20.5, 19.7, 23.2, 18.7, 21.7, 20.6, 19.8, 23.3, 18.8, 21.8, 20.7, 19.9, 23.4, 18.9, 21.9, 20.8, 20.0, 23.5, 19.0, 22.0, 20.9, 20.1, 23.6, 19.1, 22.1, 21.0, 20.2, 23.7, 19.2, 22.2, 21.1, 20.3, 23.8, 19.3, 22.3, 21.2, 20.4, 23.9, 19.4, 22.4, 21.3, 20.5, 24.0, 19.5, 22.5, 21.4, 20.6, 24.1, 19.6, 22.6, 21.5, 20.7, 24.2, 19.7, 22.7, 21.6, 20.8, 24.3, 19.8, 22.8, 21.7, 20.9, 24.4, 19.9, 22.9, 21.8, 21.0, 24.5, 20.0, 23.0, 21.9, 21.1, 24.6, 20.1, 23.1, 22.0, 21.2, 24.7, 20.2, 23.2, 22.1, 21.3, 24.8, 20.3, 23.3, 22.2, 21.4, 24.9, 20.4, 23.4, 22.3, 21.5, 25.0, 20.5, 23.5, 22.4, 21.6, 25.1, 20.6, 23.6, 22.5, 21.7, 25.2, 20.7, 23.7, 22.6, 21.8, 25.3, 20.8, 23.8, 22.7, 21.9, 25.4, 20.9, 23.9, 22.8, 22.0, 25.5, 21.0, 24.0, 22.9, 22.1, 25.6, 21.1, 24.1, 23.0, 22.2, 25.7, 21.2, 24.2, 23.1, 22.3, 25.8, 21.3, 24.3, 23.2, 22.4, 25.9, 21.4, 24.4, 23.3, 22.5, 26.0, 21.5, 24.5, 23.4, 22.6, 26.1, 21.6, 24.6, 23.5, 22.7, 26.2, 21.7, 24.7, 23.6, 22.8, 26.3, 21.8, 24.8, 23.7, 22.9, 26.4, 21.9, 24.9, 23.8, 23.0, 26.5, 22.0, 25.0, 23.9, 23.1, 26.6, 22.1, 25.1, 24.0, 23.2, 26.7, 22.2, 25.2, 24.1, 23.3, 26.8, 22.3, 25.3, 24.2, 23.4, 26.9, 22.4, 25.4, 24.3, 23.5, 27.0, 22.5, 25.5, 24.4, 23.6, 27.1, 22.6, 25.6, 24.5, 23.7, 27.2, 22.7, 25.7, 24.6, 23.8, 27.3, 22.8, 25.8, 24.7, 23.9, 27.4, 22.9, 25.9, 24.8, 24.0, 27.5, 23.0, 26.0, 24.9, 24.1, 27.6, 23.1, 26.1, 25.0, 24.2, 27.7, 23.2, 26.2, 25.1, 24.3, 27.8, 23.3, 26.3, 25.2, 24.4, 27.9, 23.4, 26.4, 25.3, 24.5, 28.0, 23.5, 26.5, 25.4, 24.6, 28.1, 23.6, 26.6, 25.5, 24.7, 28.2, 23.7, 26.7, 25.6, 24.8, 28.3, 23.8, 26.8, 25.7, 24.9, 28.4, 23.9, 26.9, 25.8, 25.0, 28.5, 24.0, 27.0, 25.9, 25.1, 28.6, 24.1, 27.1, 26.0, 25.2, 28.7, 24.2, 27.2, 26.1, 25.3, 28.8, 24.3, 27.3, 26.2, 25.4, 28.9, 24.4, 27.4, 26.3, 25.5, 29.0, 24.5, 27.5, 26.4, 25.6, 29.1, 24.6, 27.6, 26.5, 25.7, 29.2, 24.7, 27.7, 26.6, 25.8, 29.3, 24.8, 27.8, 26.7, 25.9, 29.4, 24.9, 27.9, 26.8, 26.0, 29.5, 25.0, 28.0, 26.9, 26.1, 29.6, 25.1, 28.1, 27.0, 26.2, 29.7, 25.2, 28.2, 27.1, 26.3, 29.8, 25.3, 28.3, 27.2, 26.4, 29.9, 25.4, 28.4, 27.3, 26.5, 30.0, 25.5, 28.5, 27.4, 26.6, 30.1, 25.6, 28.6, 27.5, 26.7, 30.2, 25.7, 28.7, 27.6, 26.8, 30.3, 25.8, 28.8, 27.7, 26.9, 30.4, 25.9, 28.9, 27.8, 27.0, 30.5, 26.0, 29.0, 27.9, 27.1, 30.6, 26.1, 29.1, 28.0, 27.2, 30.7, 26.2, 29.2, 28.1, 27.3, 30.8, 26.3, 29.3, 28.2, 27.4, 30.9, 26.4, 29.4, 28.3, 27.5, 31.0, 26.5, 29.5, 28.4, 27.6, 31.1, 26.6, 29.6, 28.5, 27.7, 31.2, 26.7, 29.7, 28.6, 27.8, 31.3, 26.8, 29.8, 28.7, 27.9, 31.4, 26.9, 29.9, 28.8, 28.0, 31.5, 27.0, 30.0, 28.9, 28.1, 31.6, 27.1, 30.1, 29.0, 28.2, 31.7, 27.2, 30.2, 29.1, 28.3, 31.8, 27.3, 30.3, 29.2, 28.4, 31.9, 27.4, 30.4, 29.3, 28.5, 32.0, 27.5, 30.5, 29.4, 28.6, 32.1, 27.6, 30.6, 29.5, 28.7, 32.2, 27.7, 30.7, 29.6, 28.8, 32.3, 27.8, 30.8, 29.7, 28.9, 32.4, 27.9, 30.9, 29.8, 29.0, 32.5, 28.0, 31.0, 29.9, 29.1, 32.6, 28.1, 31.1, 30.0, 29.2, 32.7, 28.2, 31.2, 30.1, 29.3, 32.8, 28.3, 31.3, 30.2, 29.4, 32.9, 28.4, 31.4, 30.3, 29.5, 33.0, 28.5, 31.5, 30.4, 29.6, 33.1, 28.6, 31.6, 30.5, 29.7, 33.2, 28.7, 31.7, 30.6, 29.8, 33.3, 28.8, 31.8, 30.7, 29.9, 33.4, 28.9, 31.9, 30.8, 30.0, 33.5, 29.0, 32.0, 30.9, 30.1, 33.6, 29.1, 32.1, 31.0, 30.2, 33.7, 29.2, 32.2, 31.1, 30.3, 33.8, 29.3, 32.3, 31.2, 30.4, 33.9, 29.4, 32.4, 31.3, 30.5, 34.0, 29.5, 32.5, 31.4, 30.6, 34.1, 29.6, 32.6, 31.5, 30.7, 34.2, 29.7, 32.7, 31.6, 30.8, 34.3, 29.8, 32.8, 31.7, 30.9, 34.4, 29.9, 32.9, 31.8, 31.0, 34.5, 30.0, 33.0, 31.9, 31.1, 34.6, 30.1, 33.1, 32.0, 31.2, 34.7, 30.2, 33.2, 32.1, 31.3, 34.8, 30.3, 33.3, 32.2, 31.4, 34.9, 30.4, 33.4, 32.3, 31.5, 35.0, 30.5, 33.5, 32.4, 31.6, 35.1, 30.6, 33.6, 32.5, 31.7, 35.2, 30.7, 33.7, 32.6, 31.8, 35.3, 30.8, 33.8, 32.7, 31.9, 35.4, 30.9, 33.9, 32.8, 32.0, 35.5, 31.0, 34.0, 32.9, 32.1, 35.6, 31.1, 34.1, 33.0, 32.2, 35.7, 31.2, 34.2, 33.1, 32.3, 35.8, 31.3, 34.3, 33.2, 32.4, 35.9, 31.4, 34.4, 33.3, 32.5, 36.0, 31.5, 34.5, 33.4, 32.6, 36.1, 31.6, 34.6, 33.5, 32.7, 36.2, 31.7, 34.7, 33.6, 32.8, 36.3, 31.8, 34.8, 33.7, 32.9, 36.4, 31.9, 34.9, 33.8, 33.0, 36.5, 32.0, 35.0, 33.9, 33.1, 36.6, 32.1, 35.1, 34.0, 33.2, 36.7, 32.2, 35.2, 34.1, 33.3, 36.8, 32.3, 35.3, 34.2, 33.4, 36.9, 32.4, 35.4, 34.3, 33.5, 37.0, 32.5, 35.5, 34.4, 33.6, 37.1, 32.6, 35.6, 34.5, 33.7, 37.2, 32.7, 35.7, 34.6, 33.8, 37.3, 32.8, 35.8, 34.7, 33.9, 37.4, 32.9, 35.9, 34.8, 34.0, 37.5, 33.0, 36.0, 34.9, 34.1, 37.6, 33.1, 36.1, 35.0, 34.2, 37.7, 33.2, 36.2, 35.1, 34.3, 37.8, 33.3, 36.3, 35.2, 34.4, 37.9, 33.4, 36.4, 35.3, 34.5, 38.0, 33.5, 36.5, 35.4, 34.6, 38.1, 33.6, 36.6, 35.5, 34.7, 38.2, 33.7, 36.7, 35.6, 34.8, 38.3, 33.8, 36.8, 35.7, 34.9, 38.4, 33.9, 36.9, 35.8, 35.0, 38.5, 34.0, 37.0, 35.9, 35.1, 38.6, 34.1, 37.1, 36.0, 35.2, 38.7, 34.2, 37.2, 36.1, 35.3, 38.8, 34.3, 37.3, 36.2, 35.4, 38.9, 34.4, 37.4, 36.3, 35.5, 39.0, 34.5, 37.5, 36.4, 35.6, 39.1, 34.6, 37.6, 36.5, 35.7, 39.2, 34.7, 37.7, 36.6, 35.8, 39.3, 34.8, 37.8, 36.7, 35.9, 39.4, 34.9, 37.9, 36.8, 36.0, 39.5, 35.0, 38.0, 36.9, 36.1, 39.6, 35.1, 38.1, 37.0, 36.2, 39.7, 35.2, 38.2, 37.1, 36.3, 39.8, 35.3, 38.3, 37.2, 36.4, 39.9, 35.4, 38.4, 37.3, 36.5, 40.0, 35.5, 38.5, 37.4, 36.6, 40.1, 35.6, 38.6, 37.5, 36.7, 40.2, 35.7, 38.7, 37.6, 36.8, 40.3, 35.8, 38.8, 37.7, 36.9, 40.4, 35.9, 38.9, 37.8, 37.0, 40.5, 36.0, 39.0, 37.9, 37.1, 40.6, 36.1, 39.1, 38.0, 37.2, 40.7, 36.2, 39.2, 38.1, 37.3, 40.8, 36.3, 39.3, 38.2, 37.4, 40.9, 36.4, 39.4, 38.3, 37.5, 41.0, 36.5, 39.5, 38.4, 37.6, 41.1, 36.6, 39.6, 38.5, 37.7, 41.2, 36.7, 39.7, 38.6, 37.8, 41.3, 36.8, 39.8, 38.7, 37.9, 41.4, 36.9, 39.9, 38.8, 38.0, 41.5, 37.0, 40.0, 38.9, 38.1, 41.6, 37.1, 40.1, 39.0, 38.2, 41.7, 37.2, 40.2, 39.1, 38.3, 41.8, 37.3, 40.3, 39.2, 38.4, 41.9, 37.4, 40.4, 39.3, 38.5, 42.0, 37.5, 40.5, 39.4, 38.6, 42.1, 37.6, 40.6, 39.5, 38.7, 42.2, 37.7, 40.7, 39.6, 38.8, 42.3, 37.8, 40.8, 39.7, 38.9, 42.4, 37.9, 40.9, 39.8, 39.0, 42.5, 38.0, 41.0, 39.9, 39.1, 42.6, 38.1, 41.1, 40.0, 39.2, 42.7, 38.2, 41.2, 40.1, 39.3, 42.8, 38.3, 41.3, 40.2, 39.4, 42.9, 38.4, 41.4, 40.3, 39.5, 43.0, 38.5, 41.5, 40.4, 39.6, 43.1, 38.6, 41.6, 40.5, 39.7, 43.2, 38.7, 41.7, 40.6, 39.8, 43.3, 38.8, 41.8, 40.7, 39.9, 43.4, 38.9, 41.9, 40.8, 40.0, 43.5, 39.0, 42.0, 40.9, 40.1, 43.6, 39.1, 42.1, 41.0, 40.2, 43.7, 39.2, 42.2, 41.1, 40.3, 43.8, 39.3, 42.3, 41.2, 40.4, 43.9, 39.4, 42.4, 41.3, 40.5, 44.0, 39.5, 42.5, 41.4, 40.6, 44.1, 39.6, 42.6, 41.5, 40.7, 44.2, 39.7, 42.7, 41.6, 40.8, 44.3, 39.8, 42.8, 41.7, 40.9, 44.4, 39.9, 42.9, 41.8, 41.0, 44.5, 40.0, 43.0, 41.9, 41.1, 44.6, 40.1, 43.1, 42.0, 41.2, 44.7, 40.2, 43.2, 42.1, 41.3, 44.8, 40.3, 43.3, 42.2, 41.4, 44.9, 40.4, 43.4, 42.3, 41.5, 45.0, 40.5, 43.5, 42.4, 41.6, 45.1, 40.6, 43.6, 42.5, 41.7, 45.2, 40.7, 43.7, 42.6, 41.8, 45.3, 40.8, 43.8, 42.7, 41.9, 45.4, 40.9, 43.9, 42.8, 42.0, 45.5, 41.0, 44.0, 42.9, 42.1, 45.6, 41.1, 44.1, 43.0, 42.2, 45.7, 41.2, 44.2, 43.1, 42.3, 45.8, 41.3, 44.3, 43.2, 42.4, 45.9, 41.4, 44.4, 43.3, 42.5, 46.0, 41.5, 44.5, 43.4, 42.6, 46.1, 41.6, 44.6, 43.5, 42.7, 46.2, 41.7, 44.7, 43.6, 42.8, 46.3, 41.8, 44.8, 43.7, 42.9, 46.4, 41.9, 44.9, 43.8, 43.0, 46.5, 42.0, 45.0, 43.9, 43.1, 46.6, 42.1, 45.1, 44.0, 43.2, 46.7, 42.2, 45.2, 44.1, 43.3, 46.8, 42.3, 45.3, 44.2, 43.4, 46.9, 4
```

Form machbar, da wir nur ein Bielefeld haben, und dort eine Intervention veranstalten können, oder eben nicht. Und keine zwei identischen Städte miteinander vergleichen können, wie bei einem Videospiel wo wir vom selben Speicherpunkt verschiedene Routen ausprobieren und die beste aussuchen können.

```
# Seed setzen für Replizierbarkeit
set.seed(2345)
# nicht set.seed(8) verwenden

# Zufällige Anzahl an Treatments generieren
sample_amount <- sample(1:8 , 1)
sample_amount
```

```
[1] 3
```

```
nas_to_assign <- sample(1:8, sample_amount)
nas_to_assign
```

```
[1] 6 7 3
```

```
# NAs (Treatment vorhanden oder nicht) zuweisen
cities_df$unemployment_intervention[nas_to_assign] <- NA
cities_df$unemployment_no_intervention[!(1:8 %in% nas_to_assign)] <- NA

# Basierend darauf möchten wir jetzt erneut den kausalen Effekt berechnen
ate <- mean(cities_df$unemployment_intervention, na.rm = TRUE) - mean(cities_df$unemployment.
print(ate)
```

```
[1] -0.6266667
```

In diesem zufälligen Beispiel an Zuweisungen erhalten wir einen kausalen Effekt, der nicht exakt -1 ist. Das liegt daran, dass wir die NAs zufällig zugewiesen haben. In der Realität ist es natürlich nicht möglich, dass die Zuweisung von Interventionen zufällig erfolgt. In der Regel wird dies durch ein Experiment oder eine gezielte Zuweisung durchgeführt. Mit einer anderen Anzahl an NAs und einer anderen Zuweisung von diesen, erhalten wir immer wieder andere Ergebnisse, dementsprechend sind diese immer mit Vorsicht zu genießen. Zudem sehen wir, dass der kausale Effekt in meinem Beispiel mehr als doppelt so stark geschätzt wird, als er tatsächlich ist (da wir ihn im Vorfeld selbst definiert haben). Dies liegt nicht zuletzt an der sehr kleinen Stichprobe.

Durch `set.seed(8)` werden sogar exakt 8 NAs generiert, was dazu führt, dass wir gar kein Treatment beobachten können. Gerne selbst einmal ausprobieren!

```

# Wir erstellen noch einmal unseren Dataframe, da wir ihn zuvor verändert habe
cities_df <- data.frame(
  city = c("Bielefeld", "Berlin", "Budapest", "Barcelona", "Bologna", "Bremen", "Bordeaux", "
  unemployment_no_intervention = c(5.2, 6.3, 4.3, 8.1, 5.2, 4, 7.5, 3.4),
  unemployment_intervention = c(4.2, 5.3, 3.3, 7.1, 4.2, 3, 6.5, 2.4)
)

# Einträge mit einer Arbeitslosigkeit von mehr als 5%, wo wir als "perfekte Politiker" intervenieren
cities_above_5 <- c(1,2,4,5,7)

# Intervention darauf basierend zuweisen
cities_df$unemployment_no_intervention[cities_above_5] <- NA
cities_df$unemployment_intervention[!(1:8 %in% cities_above_5)] <- NA

# Kausale Effekte, basierend auf dieser Zuordnung, berechnen
ate <- mean(cities_df$unemployment_intervention, na.rm = TRUE) - mean(cities_df$unemployment_no_intervention, na.rm = TRUE)
print(ate)

```

```
[1] 1.56
```

In diesem Beispiel erhalten wir nun einen Effekt von 1,56, also schätzen wir gar einen Anstieg der Arbeitslosigkeit als kausalen Effekt des Treatments. Dies ist nicht übereinstimmend mit der Realität, und lässt uns sehen, dass wir bei der Beurteilung von kausalen Effekten immer sehr vorsichtig sein müssen. Eine mögliche Lösung, um solche Schwierigkeiten zu umgehen, ist der *differences-in-differences*-Ansatz, den wir in der nächsten Übung näher untersuchen werden.

Weitere Städte, künstliche Erweiterung des Datensatzes

```

# Seed setzen für Replizierbarkeit
set.seed(26)

# Wir generieren mehr Städte, ich nenne sie A bis Z, mit Arbeitslosigkeit gestreut um 5%
additional_cities <- LETTERS[1:26]
additional_cities_df <- data.frame(
  city = additional_cities,
  unemployment_no_intervention = rnorm(26, mean = 5, sd = 1)
)

```

```

# Wir ordnen diesen Städten jetzt den Interventionswert -1 zu
additional_cities_df$unemployment_intervention <- additional_cities_df$unemployment_no_intervention

# Und berechnen diesbezüglich wieder den kausalen Effekt, den wir bereits kennen
additional_cities_df$causal_effect <- additional_cities_df$unemployment_intervention - additional_cities_df$unemployment_no_intervention

# Test auf erfolgreiche Zuweisung des Effekts
mean_causal_effect <- mean(additional_cities_df$causal_effect)
print(mean_causal_effect)

```

```
[1] -1
```

```

# Wir weisen wieder zufällig NAs zu
sample_amount <- sample(1:26 , 1)
nas_to_assign <- sample(1:26, sample_amount)

# Und ordnen diese NAs wieder zu
additional_cities_df$unemployment_intervention[nas_to_assign] <- NA
additional_cities_df$unemployment_no_intervention[!(1:8 %in% nas_to_assign)] <- NA

# Und berechnen anschließend wieder unseren kausalen Effekt
ate <- mean(additional_cities_df$unemployment_intervention, na.rm = TRUE) - mean(additional_cities_df$unemployment_no_intervention, na.rm = TRUE)
print(ate)

```

```
[1] -0.3228017
```

Hier sehen wir jetzt einen kausalen Effekt von -0.3228017. Dieser ist auch ein ganzes Stück entfernt von unserem wahren Wert, jedoch scheint er, bei einer einmaligen Simulation, schon besser zu passen, als er es in unserer kleineren Stichprobe tat. Dies zeigt, dass größere Stichproben und mehr Daten uns helfen können, bessere Schätzungen zu erhalten.

Bonus: Spurious Correlations

Die Website von Tyler Vigen ist eine fantastische Quelle für viele Korrelationen, welche keinen kausalen Zusammenhang aufweisen: <https://tylervigen.com/spurious-correlations> Hier findet ihr viele interessante Korrelationen, vielleicht findet ihr ja auch eine, die euch besonders überrascht, oder einen spannenden Fun-Fact für die nächste Party. Ich persönlich bin großer Fan der Scheidungsrate, welche hochgradig mit dem Margarine-Preis korreliert, oder die Scheidungsrate im Vereinigten Königreich, welche sehr stark mit der Veröffentlichungsrate von Disney Filmen korreliert!