Tut_7

Kausale Inferenz und Confounder

1.

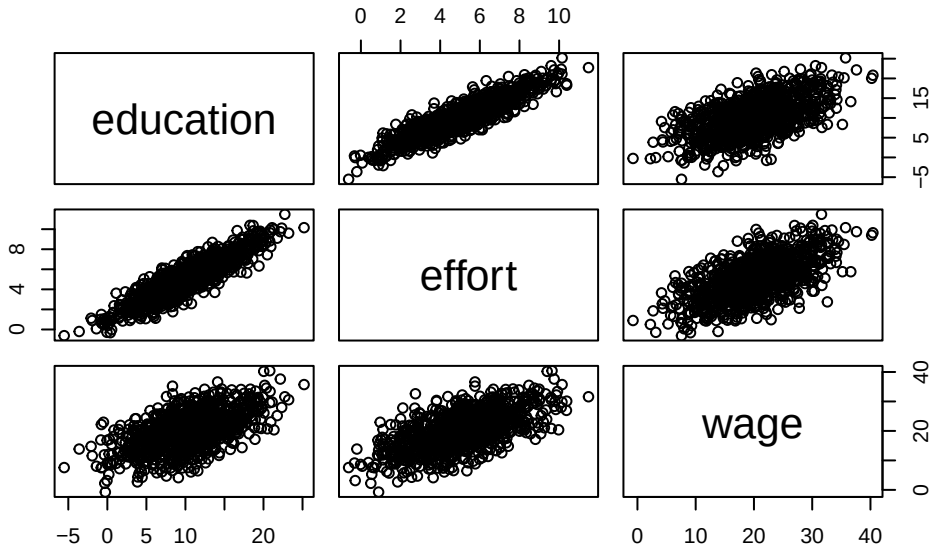
Laden Sie den Datensatz (https://statistik.julianhinz.com/tutorials/tutorial8/data.csv), der simulierte Daten zu Gehalt, Bildung und Leistungseinsatz enthält.

```
data <- read.csv("data.csv", header = TRUE)
data <- data[c(2,3,4)]
attach(data)

head(data)
```

```
  education   effort     wage
1  5.766500 3.879049 15.20008
2  6.999380 4.539645 20.26398
3 16.198873 8.117417 23.52689
4 10.017683 5.141017 26.37817
5  5.418465 5.258575 21.38783
6 18.941407 8.430130 23.78392
```

```
pairs(data)
```



2.

Stellen Sie eine These auf über die Beziehung der Daten und illustrieren diese in einem kausalen Graphen.

“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.”

Effort beeinflusst sowohl wage, als auch education.

3.

Testen Sie nun diese Theorie. Welche Variable ist ein Confounder? Wie haben Sie diese Variable identifiziert?

```
mod_conf0 <- lm(wage ~ education)
summary(mod_conf0)
```

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

```
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
```

```r
mod_conf1 <- lm(wage ~ education + effort)
summary(mod_conf1)
```

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

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 educaction 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.

Kausale Effekte

Gegeben sind die folgenden wahren Daten:

```
cities_df <- data.frame(
  city = c("Bielefeld", "Berlin", "Budapest", "Barcelona", "Bologna", "Bremen",
           "Bordeaux", "Brüssel"),
  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)
)
```

1.

Bestimmen Sie hier den wahren kausalen Effekt der Politikmaßnahme auf die Arbeitslosigkeit für jede Stadt.

$Y_i(1) - Y_i(0)$

```
cities_df$causal_effect <- cities_df$unemployment_intervention -
  cities_df$unemployment_no_intervention

cities_df$causal_effect
```

```
[1] -1 -1 -1 -1 -1 -1 -1 -1
```

2.

Was ist der durchschnittliche kausale Effekt?

$E[Y_i(1)] - E[Y_i(0)]$

```
mean(cities_df$causal_effect)
```

```
[1] -1
```

Die Maßnahmen führen zu einer kausalen Reduktion der Arbeitslosigkeit um 1%.

3.

Schätzen Sie den Effekt der Politikmaßnahme auf die Arbeitslosigkeit basierend auf zufälliger Zuweisung. Warum weicht dieser Effekt von dem vorherigen ab? Wir sammeln ihre Ergebnisse und berechnen den durchschnittlichen kausalen Effekt, den Sie geschätzt haben.

```
set.seed(2025)

sample_amount <- sample(1:8, 1)
nas <- sample(1:8, sample_amount)

cities_df$unemployment_intervention[nas] <- NA
cities_df$unemployment_no_intervention[!(1:8 %in% nas)] <- NA

ate <- mean(cities_df$unemployment_intervention, na.rm = TRUE) -
  mean(cities_df$unemployment_no_intervention, na.rm = TRUE)

ate
```

```
[1] -0.8933333
```

In diesem zufälligen Beispiel an Zuweisungen erhalten wir einen kausalen Effekt, der nicht exakt -1 ist. Dies ist zu erwarten, da wir eine Schätzung anhand einer kleinen Stichprobe durchgeführt haben. Außerdem wurden die NAs zufällig verteilt.

4.

Nun sind Sie die perfekte Politikerin/der perfekte Politiker und Sie wollen die Arbeitslosigkeit insbesondere in den Städten mit einer hohen Arbeitslosigkeit reduzieren (5%). Welche Städte würden Sie auswählen? Wie ändern sich die Effekte? Was bedeutet das für Menschen in der Wissenschaft, die Politikmaßnahmen evaluieren wollen?

```
cities_df <- data.frame(
  city = c("Bielefeld", "Berlin", "Budapest", "Barcelona", "Bologna", "Bremen",
           "Bordeaux", "Brüssel"),
  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)
)

cities_greater_5 <- which(cities_df$unemployment_no_intervention > 5)
```

```r
cities_df$unemployment_no_intervention[cities_greater_5] <- NA
cities_df$unemployment_intervention[!(1:8 %in% cities_greater_5)] <- NA

ate <- mean(cities_df$unemployment_intervention, na.rm = TRUE) -
  mean(cities_df$unemployment_no_intervention, na.rm = TRUE)

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.

5.

Erweitern Sie die Daten um weitere Städte und berechnen Sie den Effekt der Politikmaßnahme erneut unter zufalliger Zuweisung.

```r
set.seed(2025)

additional_cities <- make.unique(c(LETTERS,
                                   outer(LETTERS, LETTERS, paste0)))
additional_cities <- additional_cities[1:100]
additional_cities_df <- data.frame(
  city = additional_cities,
  unemployment_no_intervention = rnorm(100, mean = 5, sd = 1)
)

additional_cities_df$unemployment_intervention <-
  additional_cities_df$unemployment_no_intervention - 1

additional_cities_df$causal_effect <-
  additional_cities_df$unemployment_intervention -
  additional_cities_df$unemployment_no_intervention

mean_causal_effect <- mean(additional_cities_df$causal_effect)
```

```r
sample_amount <- sample(1:100 , 1)
nas_to_assign <- sample(1:100, sample_amount)

additional_cities_df$unemployment_intervention[nas_to_assign] <- NA
additional_cities_df$unemployment_no_intervention[!(1:100 %in% nas_to_assign)] <- NA

ate <- mean(additional_cities_df$unemployment_intervention, na.rm = TRUE) -
  mean(additional_cities_df$unemployment_no_intervention, na.rm = TRUE)

ate
```

```
[1] -1.124023
```