

# Kausalanalyse

## Kapitel 8

# Gliederung

1. Teil 1: Statistik, was soll ich tun?
2. Teil 2: Konfundierung
3. Teil 3: Kollision
4. Teil 4: Die Hintertür schließen
5. Hinweise

# Teil 1

**Statistik, was soll ich tun?**

# Studie A: Östrogen

Was raten Sie dem Arzt? Medikament einnehmen, ja oder nein?

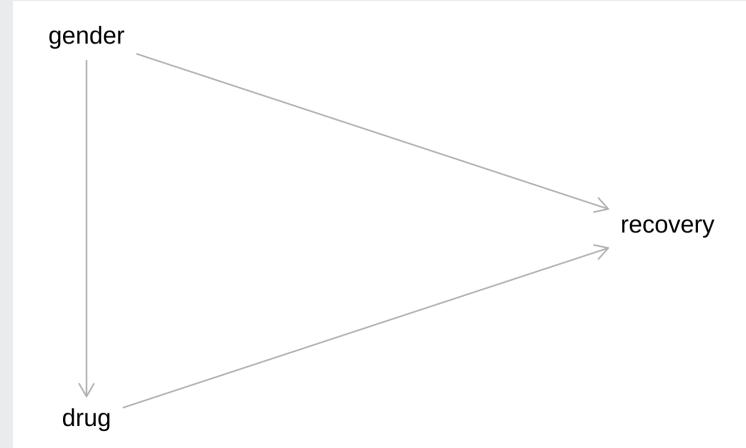
Gruppe	Mit Medikament	Ohne Medikament
Männer	81/87 überlebt (93%)	234/270 überlebt (87%)
Frauen	192/263 überlebt (73%)	55/80 überlebt (69%)
Gesamt	273/350 überlebt (78%)	289/350 überlebt (83%)

- Die Daten stammen aus einer (fiktiven) klinischen Studie,  $n = 700$ , hoher Qualität (Beobachtungsstudie).
- Bei Männern scheint das Medikament zu helfen; bei Frauen auch.
- Aber *insgesamt* (Summe von Frauen und Männern) *nicht*!?
- Was sollen wir den Arzt rate? Soll er das Medikament verschreiben? Vielleicht nur dann, wenn er das Geschlecht kennt?

(Pearl, Glymour, and Jewell, 2016)

# Kausalmodell zur Studie A

- Das Geschlecht (Östrogen) hat einen Einfluss (+) auf Einnahme des Medikaments und auf Heilung (-).
- Das Medikament hat einen Einfluss (+) auf Heilung.
- Betrachtet man die Gesamt-Daten zur Heilung, so ist der Effekt von Geschlecht (Östrogen) und Medikament *vermengt* (konfundiert).



**Betrachtung der Teildaten (stratifiziert pro Gruppe) zeigt den wahren, kausalen Effekt.**

Betrachtung der Gesamtdaten zeigt einen *konfundierten* Effekt: Geschlecht konfundiert den Zusammenhang von Medikament und Heilung.

# Studie B: Blutdruck

Was raten Sie dem Arzt? Medikament einnehmen, ja oder nein?

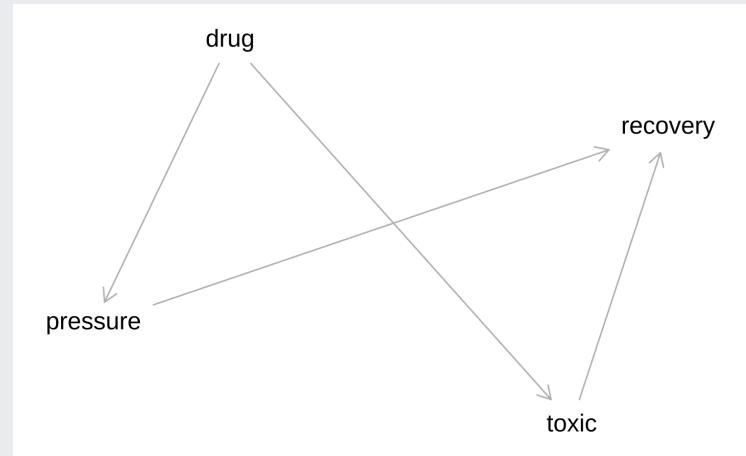
Gruppe	Ohne Medikament	Mit Medikament
geringer Blutdruck	81/87 überlebt (93%)	234/270 überlebt (87%)
hoher Blutdruck	192/263 überlebt (73%)	55/80 überlebt (69%)
Gesamt	273/350 überlebt (78%)	289/350 überlebt (83%)

- Die Daten stammen aus einer (fiktiven) klinischen Studie,  $n = 700$ , hoher Qualität (Beobachtungsstudie).
- Bei geringem Blutdruck scheint das Medikament zu schaden.
- Bei hohem Blutdruck scheint das Medikament auch zu schaden.
- Aber *insgesamt* (Summe über beide Gruppe) *nicht*, da scheint es zu nutzen!?
- Was sollen wir den Arzt raten? Soll er das Medikament verschreiben? Vielleicht nur dann, wenn er den Blutdruck nicht kennt?

(Pearl, Glymour, and Jewell, 2016)

# Kausalmodell zur Studie B

- Das Medikament hat einen (absenkenden) Einfluss auf den Blutdruck.
- Gleichzeitig hat das Medikament einen (toxischen) Effekt auf die Heilung.
- Verringelter Blutdruck hat einen positiven Einfluss auf die Heilung.
- Sucht man innerhalb der Leute mit gesenktem Blutdruck nach Effekten, findet man nur den toxischen Effekt: Gegeben diesen Blutdruck ist das Medikament schädlich aufgrund des toxischen Effekts. Der positive Effekt der Blutdruck-Senkung ist auf diese Art nicht zu sehen.

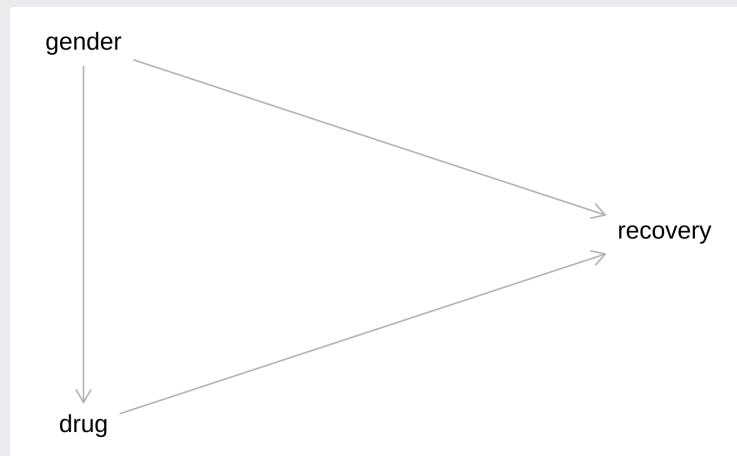


Betrachtung der Teildaten zeigt nur den toxischen Effekt des Medikaments, nicht den nützlichen (Reduktion des Blutdrucks).

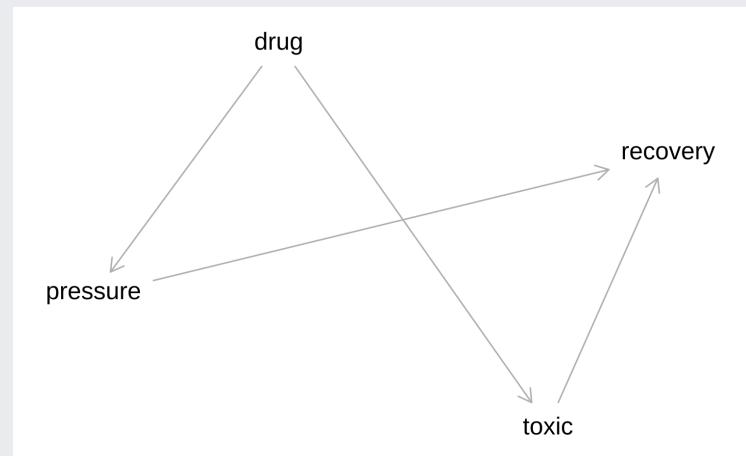
**Betrachtung der Gesamtdaten zeigt den wahren, kausalen Effekt.**

# Studie A und B: Gleiche Daten, unterschiedliches Kausalmodell

Studie A



Studie B



Kausale Interpretation - und damit Entscheidungen für Handlungen - war nur möglich, wenn das Kausalmodell bekannt ist. Die Daten alleine reichen nicht.

# Sorry, Statistik: Du allein schaffst es nicht

**Statistik alleine reicht nicht für  
Kausalschlüsse**



**Statistik plus Theorie erlaubt  
Kausalschlüsse**



- Für Entscheidungen ("Was soll ich tun?") braucht man kausales Wissen.
- Kausales Wissen basiert auf einer Theorie (Kausalmodell) plus Daten.

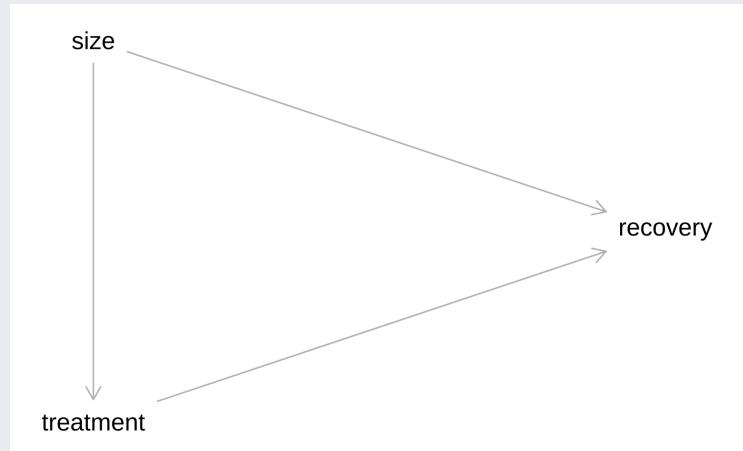
# Studie C: Nierensteine

Nehmen wir an, es gibt zwei Behandlungsvarianten bei Nierensteinen, Behandlung A und B. Ärzte tendieren zu Behandlung A bei großen Steinen (die einen schwereren Verlauf haben); bei kleineren Steinen tendieren die Ärzte zu Behandlung B.

Sollte ein Patient, der nicht weiß, ob sein Nierenstein groß oder klein ist, die Wirksamkeit in der Gesamtpopulation (Gesamtdaten) oder in den stratifizierten Daten (Teildaten nach Steingröße) betrachten, um zu entscheiden, welche Behandlungsvariante er (oder sie) wählt?

# Kausalmodell zur Studie C

- Die Größe der Nierensteine hat einen Einfluss auf die Behandlungsmethode.
- Die Behandlung hat einen Einfluss auf die Heilung.
- Damit gibt es eine Mediation von Größe -> Behandlung -> Heilung.
- Darüberhinaus gibt es noch einen Einfluss von Größe der Nierensteine auf die Heilung.



# Soll ich heiraten?

„Studien zeigen, dass Einkommen und Heiraten (bzw. verheiratete sein) hoch korrelieren. Daher wird sich dein Einkommen erhöhen, wenn du heiratest.“

# Soll ich mich beeilen?

„Studien zeigen, dass Leute, die sich beeilen, zu spät zu ihrer Besprechung kommen. Daher lieber nicht beeilen, oder du kommst zu spät zu deiner Besprechung.“

## Teil 2

### Konfundierung

# Datensatz 'Hauspreise im Saratoga County'

Datenquelle; Beschreibung des Datensatzes

```
d_path <- "https://tinyurl.com/3jn3cc5u"
```

Show 5 entries

Search:

	price	livingArea	bedrooms	waterfront
1	132500	906	2	No
2	181115	1953	3	No
3	109000	1944	4	No
4	155000	1944	3	No
5	86060	840	2	No

Showing 1 to 5 of 1,728 entries

Previous

1

2

3

4

5

...

346

Next

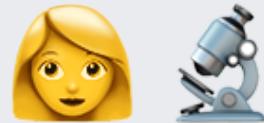
# Immobilienpreise in einer schicken Wohngegend vorhersagen

"Finden Sie den Wert meiner  
Immobilie heraus!  
Die muss viel wert sein!"



Das ist Don, Immobilienmogul,  
Auftraggeber.

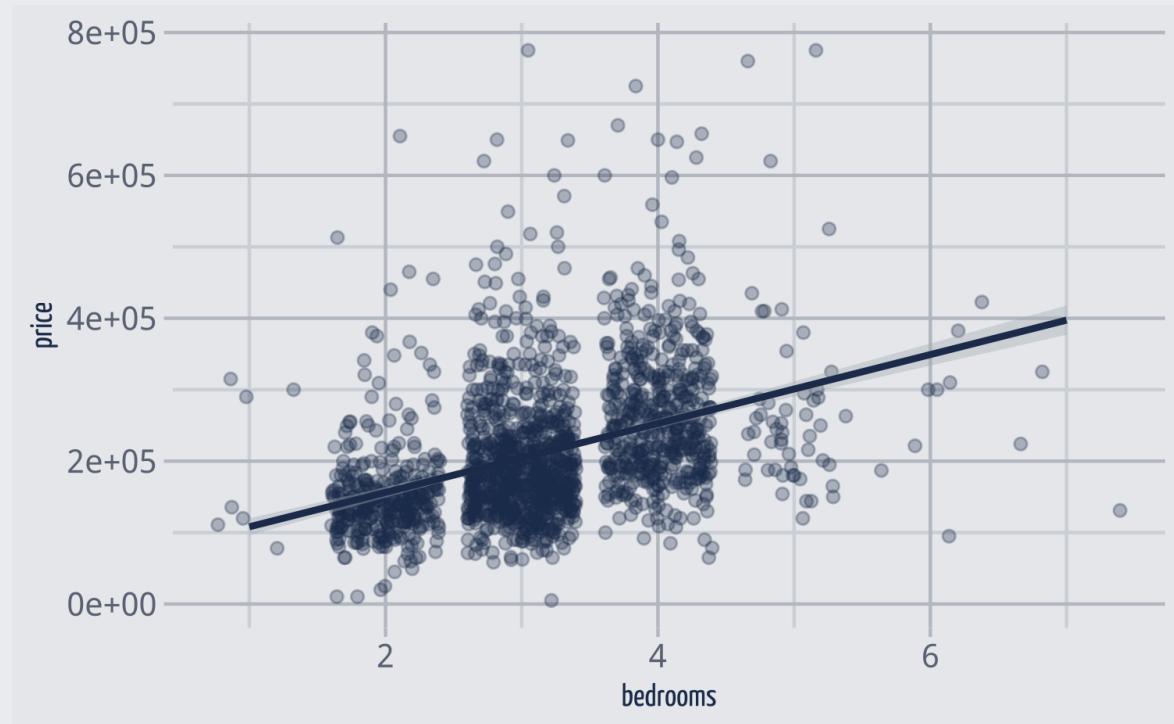
"Das finde ich heraus.  
Ich mach das wissenschaftlich."



Das ist Angie, Data Scientistin.

# Modell 1: Preis als Funktion der Anzahl der Zimmer

"Hey Don! Mehr Zimmer, mehr Kohle!"



# Posteriori-Verteilung von Modell 1

"Jedes Zimmer  
mehr ist knapp  
50 Tausend  
wert. Dein Haus  
hat einen Wert  
von etwa 150  
Tausend."



"Zu wenig! 😞"



```
m1 <- stan_glm(price ~ bedrooms,  
                  refresh = 0,  
                  data = d)  
coef(m1)
```

```
## (Intercept)    bedrooms  
##      59738.82     48256.34
```

```
dons_house <- tibble(bedrooms = 2)  
mean(posterior_predict(m1, dons_house))
```

```
## [1] 156390.7
```

# Don hat eine Idee

"Ich bau eine Mauer!  
In jedes Zimmer!  
Genial!  
An die Arbeit, Angie!"



"Das ist keine gute Idee, Don."



```
dons_new_house <- tibble(bedrooms = 4)  
mean(posterior_predict(m1, dons_new_house))
```

```
## [1] 251585.4
```

Mit 4 statt 2 Schlafzimmer steigt der Wert auf 250k, laut m1.

"Volltreffer! Jetzt verdien ich 100 Tausend mehr! 😎"



# R-Funktionen, um Beobachtungen vorhersagen

- `posterior_predict()`: ([Hilfeseite](#))
  - Was macht der Befehl? Zieht Stichproben aus der PPV.
  - Wozu braucht man den Befehl?
    - Um neue Beobachtungen vorherzusagen; falls man z.B. an einem Vorhersage-Wettbewerb teilnimmt 😎.
    - Um die Modellgüte zu prüfen: sagt unser Modell den Datensatz gut vorher?
- `predictive_intervals()`: ([Hilfeseite](#))
  - Was macht der Befehl? Berechnet Perzentilintervalle auf Basis der PPV.
  - Wozu braucht man den Befehl? Um die Ungenauigkeit (Ungewissheit) des Modells einzuschätzen.
- `predictive_error()`: ([Hilfeseite](#))
  - Was macht der Befehl: Berechnet die Vorhersagefehler (Residuen):  
 $r = y - \hat{y}$ .
  - Wozu braucht man den Befehl? Ein anderer Blickt auf die Ungewissheit des Modells.

# Modell 2: Preis als Funktion von Zimmerzahl und von Quadratmetern

Modell 2 hat schlechte Nachrichten für Don.

```
m2 <- stan_glm(price ~ bedrooms + livingArea, data = d)
```

```
coef(m2)
```

```
## (Intercept)    bedrooms   livingArea  
##  36934.3098 -14208.5306     125.5246
```

```
mean(posterior_predict(m2, newdata = tibble(bedrooms = 4, livingArea = :
```

```
## [1] 127322.6
```

"Die Zimmer zu halbieren, hat den Wert des Hauses *verringert*, Don!"



"Halbiert!? Weniger Geld?! Oh nein!"



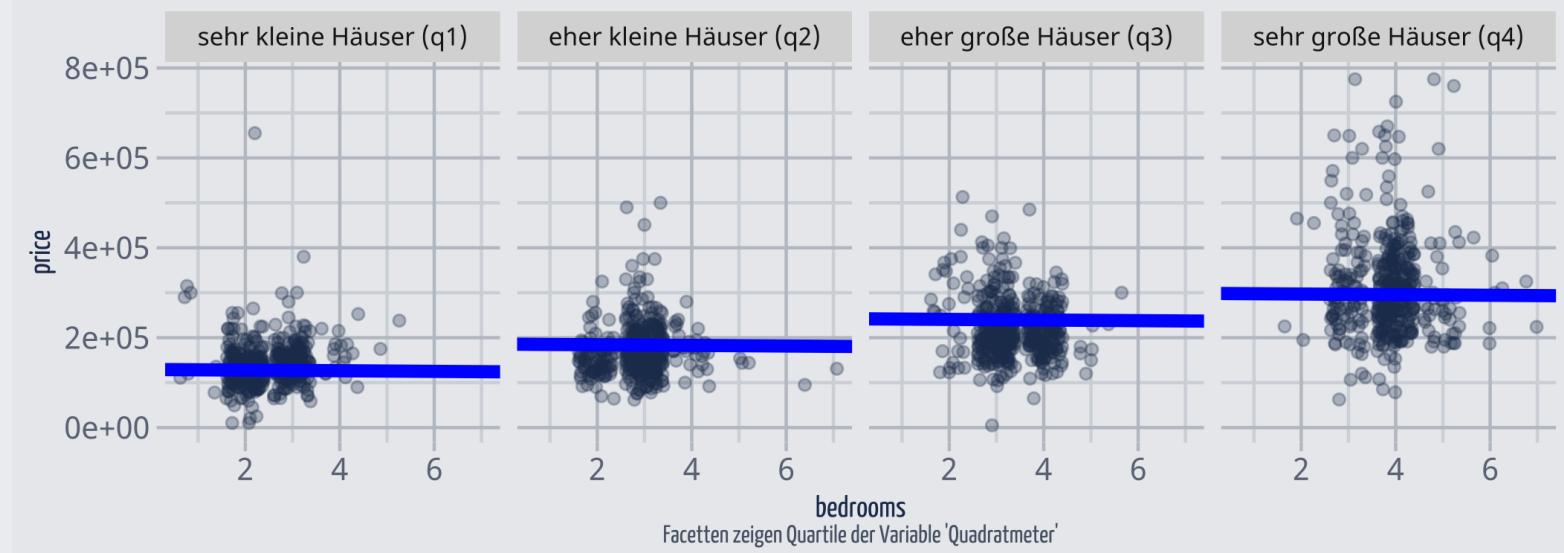
# Die Zimmerzahl ist negativ mit dem Preis korreliert

... wenn man die Wohnfläche (Quadratmeter) kontrolliert.

Ne-Ga-Tiv!



Hauspreis als Funktion von Zimmerzahl und Quadratmeter



Quellcode

# Kontrollieren von Variablen

💡 Durch das Aufnehmen von Prädiktoren in die multiple Regression werden die Prädiktoren *kontrolliert* (adjustiert, konditioniert):

- Die Koeffizienten einer multiplen Regression zeigen den Zusammenhang  $\beta$  des einen Prädiktors mit  $y$ , wenn man den (oder die) anderen Prädiktoren statistisch *konstant hält*.
- Man nennt die Koeffizienten einer multiplen Regression daher auch *parzielle Regressionskoeffizienten*. Manchmal spricht man auch vom "Netto-Effekt" eines Prädiktors, oder davon, dass ein Prädiktor "bereinigt" wurde vom (linearen) Einfluss der anderen Prädiktoren auf  $y$ .
- Damit kann man die Regressionskoeffizienten so interpretieren, dass Sie den Effekt des Prädiktors  $x_1$  auf  $y$  anzeigen *unabhängig* vom Effekt der anderen Prädiktoren,  $x_2, x_3, \dots$  auf  $y$
- Man kann sich dieses Konstanthalten vorstellen als eine Aufteilung in Gruppen: Der Effekt eines Prädiktors  $x_1$  wird für jede Ausprägung (Gruppe) des Prädiktors  $x_2$  berechnet.

# Das Hinzufügen von Prädiktoren kann die Gewichte der übrigen Prädiktoren ändern

"Aber welche und wie viele Prädiktoren soll ich denn jetzt in mein Modell aufnehmen?!"

Und welches Modell ist jetzt richtig?!"



"Leider kann die Statistik keine Antwort darauf geben."



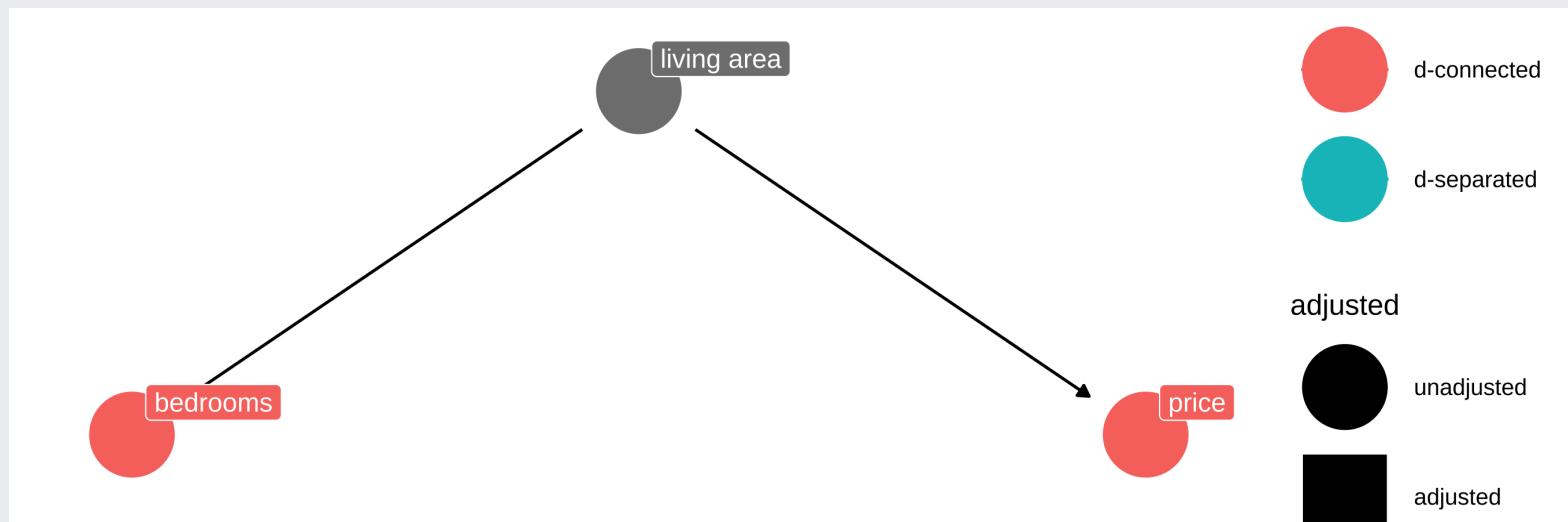
## In Beobachtungsstudien hilft nur ein (korrektes) Kausalmodell

# Welches Modell richtig ist, kann die Statistik nicht sagen

Often people want statistical modeling to do things that statical modeling cannot do. For exmaple, we'd like to know whether an effect is "real" or rather spurious. Unfortunately, modeling merely quantifies uncertainty in the precise way that the model understands the problem. Usually answer to large world questions about truth and causation depend upon information not included in the model. For example, any observed correlation between an outcome and predictor could be eliminated or reversed once another predictor is added to the model. But if we cannot think of the right variable, we might never notice. Therefore all statical models are vulnerable to and demand critique, regardless of the precision of their estimates and apparent accuracy of their predictions. Rounds of model criticism and revision emojiify the real tests of scientific hypotheses. A true hypothesis will pass and fail many statistical "tests" on its way to acceptance.

(McElreath, 2020, S. 139)

# Kausalmmodell für Konfundierung, km1



Wenn dieses Kausalmmodell stimmt, findet man eine *Scheinkorrelation* zwischen price und bedrooms.

Eine Scheinkorrelation ist ein Zusammenhang, der *nicht* auf eine kausalen Einfluss beruht.

d\_connected heißt, dass die betreffenden Variablen "verbunden" sind durch einen gerichteten (d wie directed) Pfad, durch den die Assoziation (Korrelation) wie durch einen Fluss fließt . d\_separated heißt, dass sie nicht d\_connected sind.

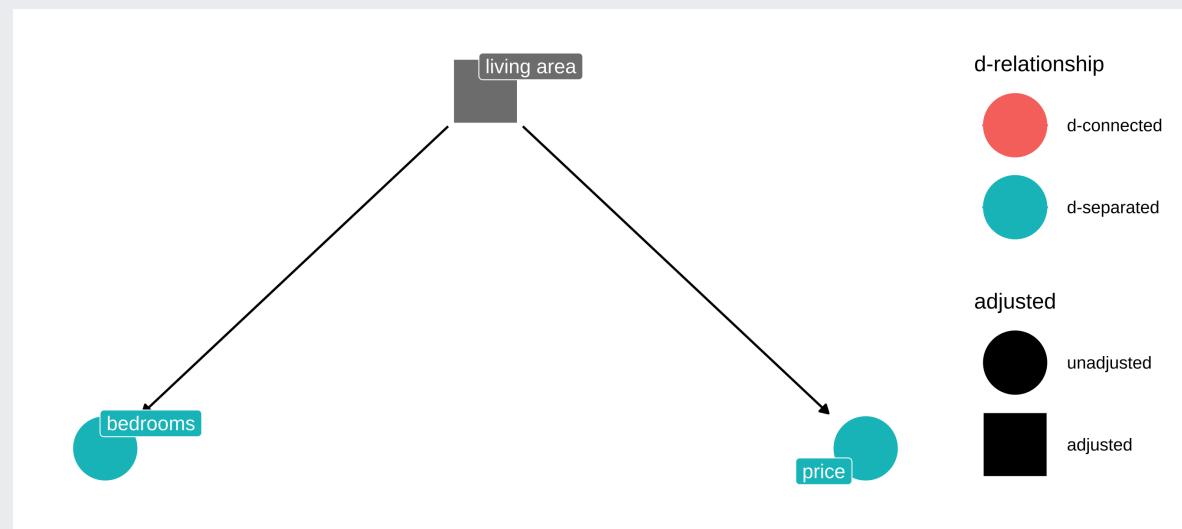
# m2 kontrolliert die Konfundierungsvariable livingArea

Wenn das Kausalmmodell stimmt, dann zeigt m2 den kausalen Effekt von livingArea.

"Was tun wir jetzt bloß?!"



"Wir müssen die Konfundierungsvariable kontrollieren."

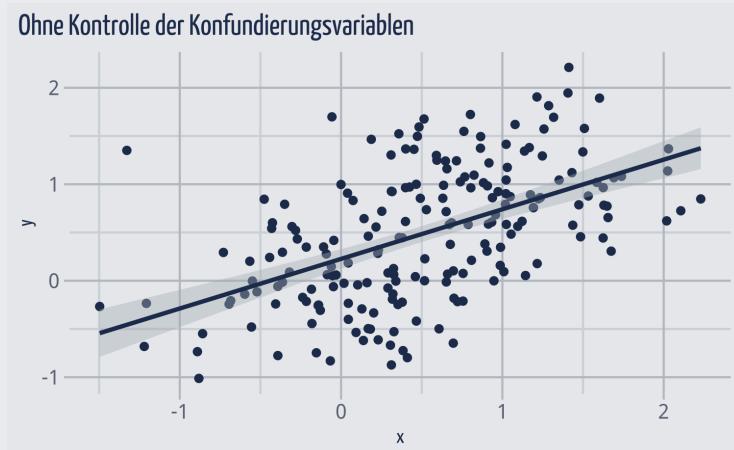


Durch das Kontrollieren ("adjustieren"), sind bedrooms und price nicht mehr korreliert, nicht mehr d\_connected, sondern jetzt d\_separated.

# Konfundierer kontrollieren

Ohne Kontrollieren der Konfundierungsvariablen

Regressionsmodell:  $y \sim x$

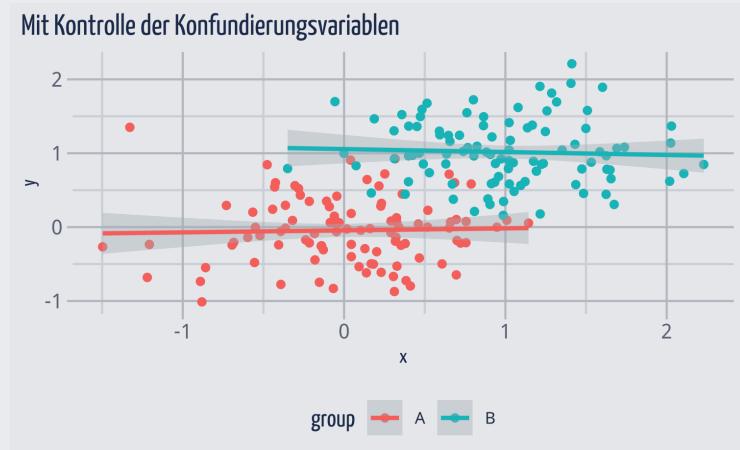


Es wird (fälschlich) eine Korrelation zwischen x und y angezeigt:  
Scheinkorrelation.

Quellcode

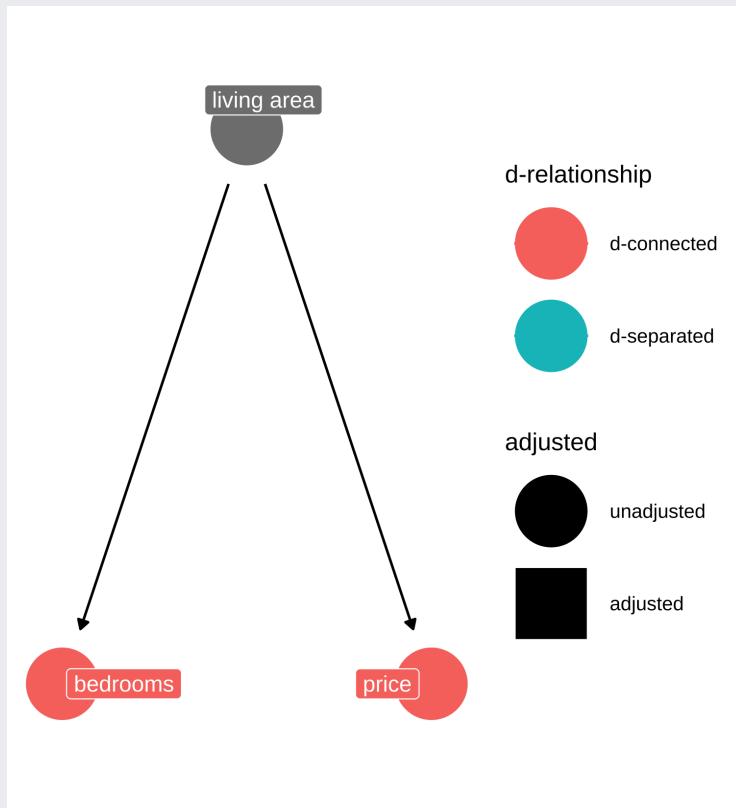
Mit Kontrollieren der Konfundierungsvariablen

Regressionsmodell:  $y \sim x + \text{group}$



Es wird korrekt gezeigt, dass es keine Korrelation zwischen x und y gibt, wenn group kontrolliert wird.

# **m1 und m2 passen nicht zu den Daten, wenn km1 stimmt**

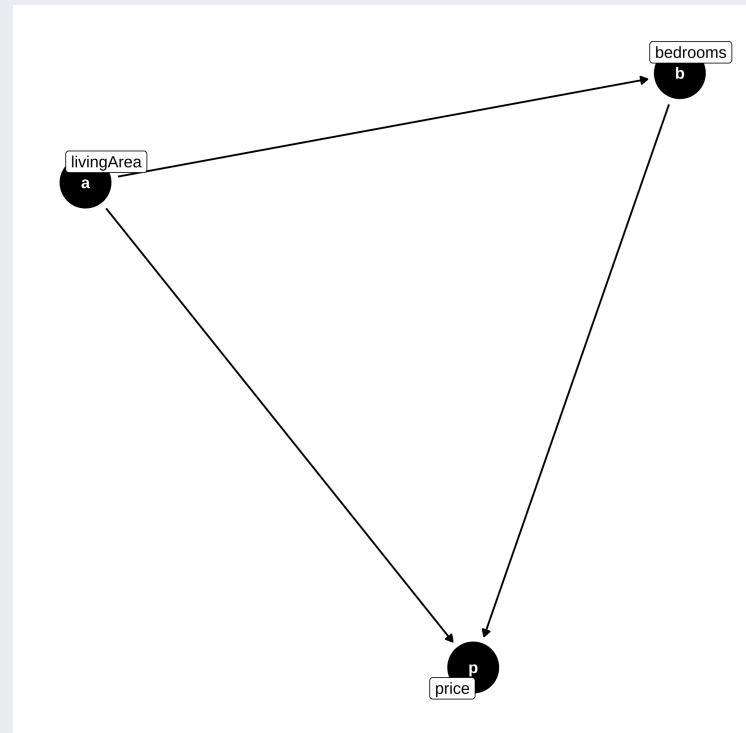


- Laut km1 dürfte es keine Assoziation (Korrelation) zwischen bedrooms und price geben, wenn man livingArea kontrolliert.
- Es gibt aber noch eine Assoziation zwischen bedrooms und price geben, wenn man livingArea kontrolliert.
- Daher sind sowohl m1 und m2 nicht mit dem Kausalmmodell km1 vereinbar.

# Kausalmmodell 2, km2

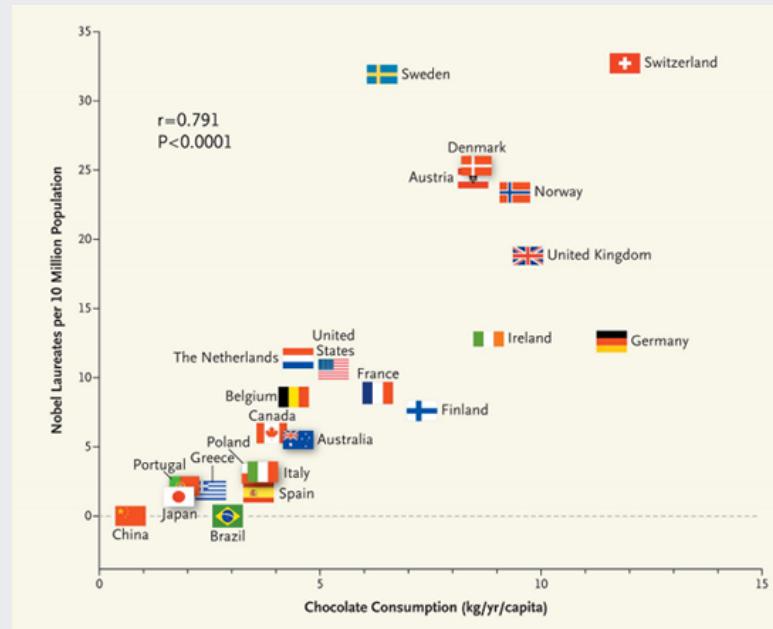
Unser Modell m2 sagt uns, dass beide Prädiktoren jeweils einen eigenen Beitrag zur Erklärung der AV haben.

- Daher könnte das folgende Kausalmmodell, km2 besser passen.
- In diesem Modell gibt es eine Wirkkette:  $a \rightarrow b \rightarrow p$ .
- Insgesamt gibt es zwei Kausaleinflüsse von a auf p:
  - $a \rightarrow p$
  - $a \rightarrow b \rightarrow p$
- Man nennt die mittlere Variable einer Wirkkette auch einen *Mediator* und den Pfad von der UV (a) über den Mediator (b) zur AV (p) auch *Mediation*.



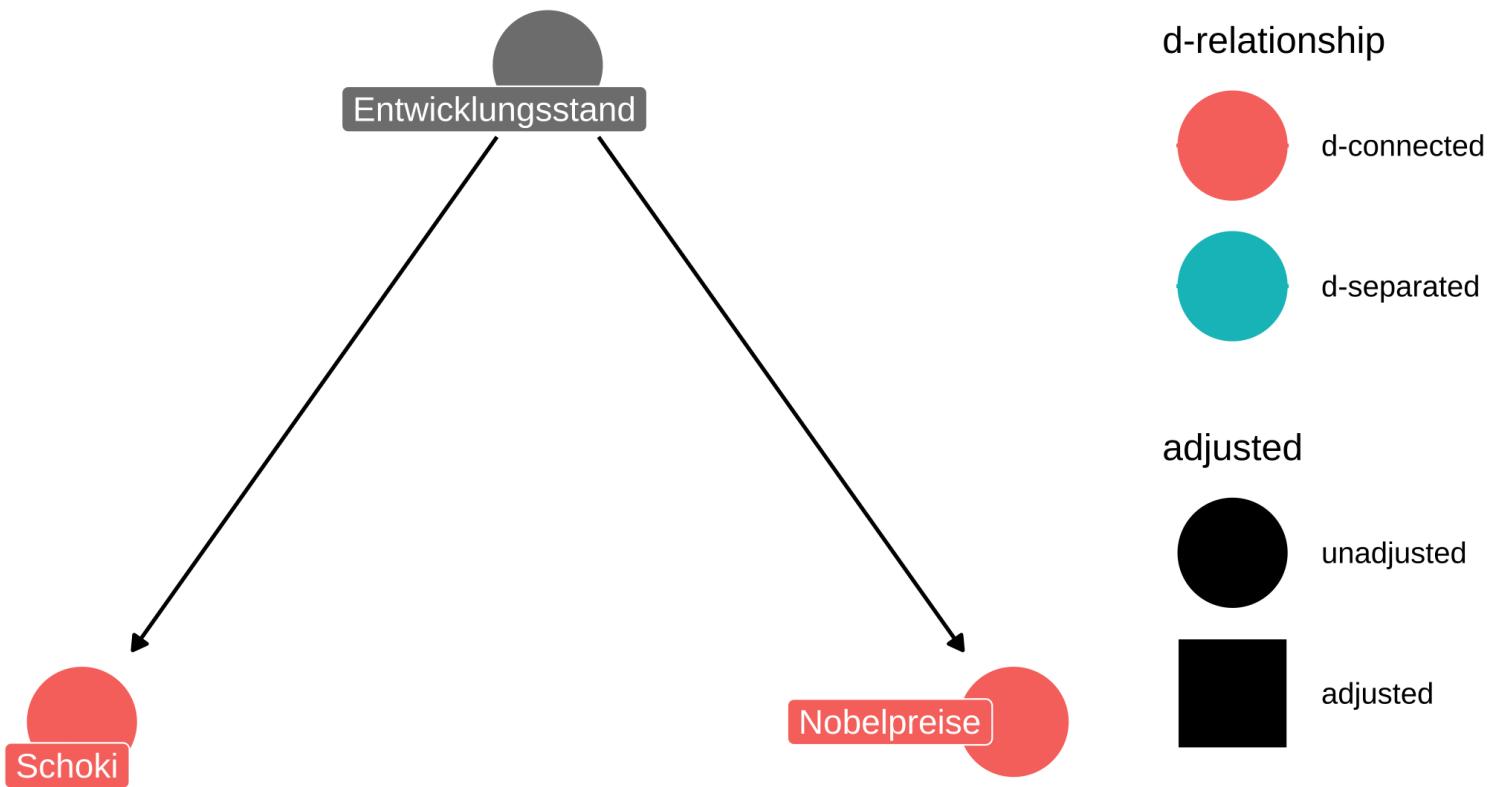
# Schoki macht Nobelpreis! (?)

Eine Studie fand eine starke Korrelation,  $r = 0.79$  zwischen (Höhe des) Schokoladenkonsums eines Landes und (Anzahl der) Nobelpreise eines Landes (Messerli, 2012).

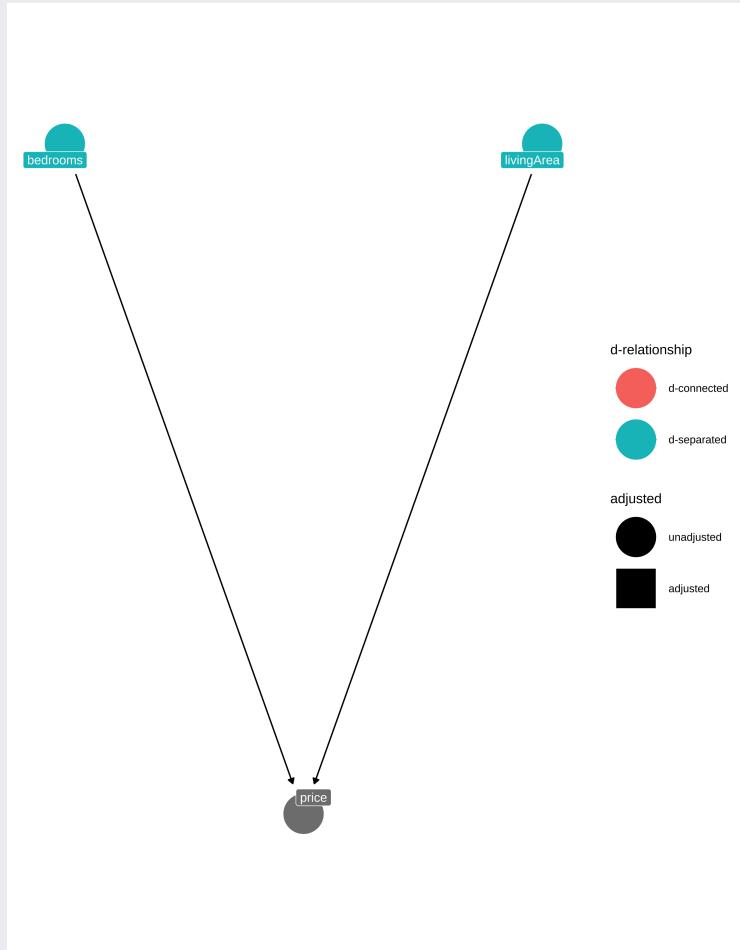


💣 Korrelation ungleich Kausation!

# Kausalmodell für die Schoki-Studie



# Dons Kausalmodell, km3



"Ich glaube aber an dieses Kausalmodell. Der Experte bin ich!"



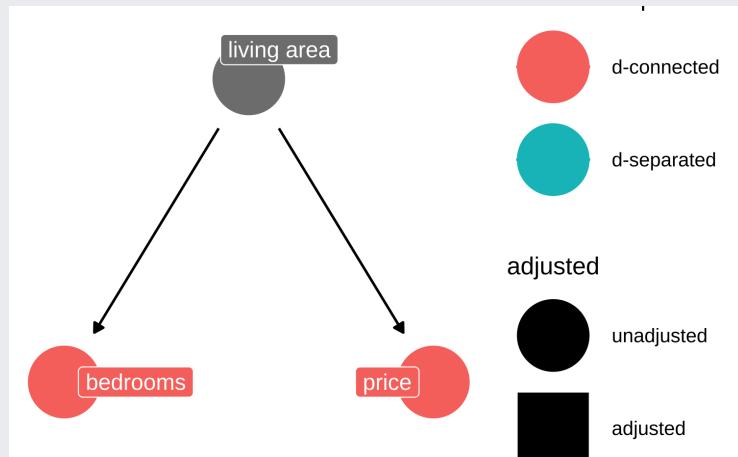
"Don, nach deinem Kausalmodell müssten `bedrooms` und `livingArea` unkorreliert sein. Sind sie aber nicht."

```
## # A tibble: 1 × 1
##   `cor(bedrooms, livingArea)`<dbl>
## 1 0.656
```



# Unabhängigkeiten laut km1

b: bedrooms, p: price, a area (living area)

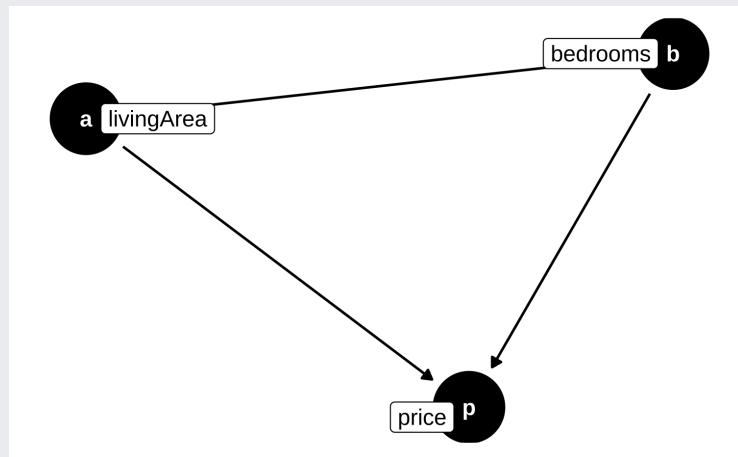


$b \perp\!\!\!\perp p \mid a$ : bedrooms sind unabhängig von price, wenn man livingArea kontrolliert.

⚠️ Passt nicht zu den Daten/zum Modell

# Unabhängigkeiten laut km2

b: bedrooms, p: price, a area (living area)

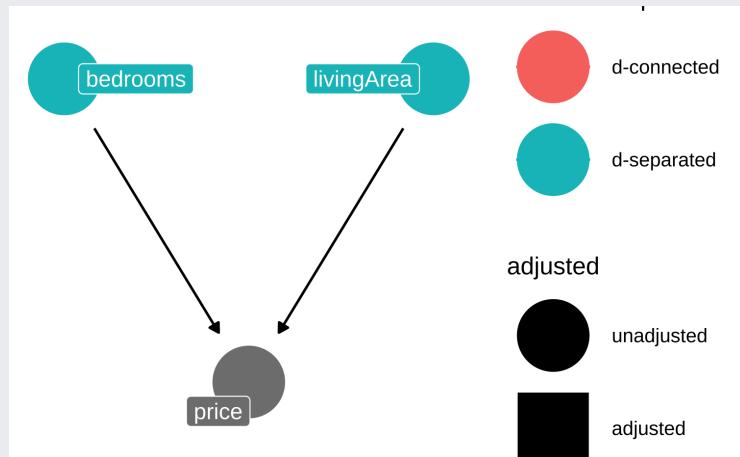


keine Unabhängigkeiten

? Passt zu den Daten/zum Modell

# Unabhängigkeiten laut km3

b: bedrooms, p: price, a area (living area)



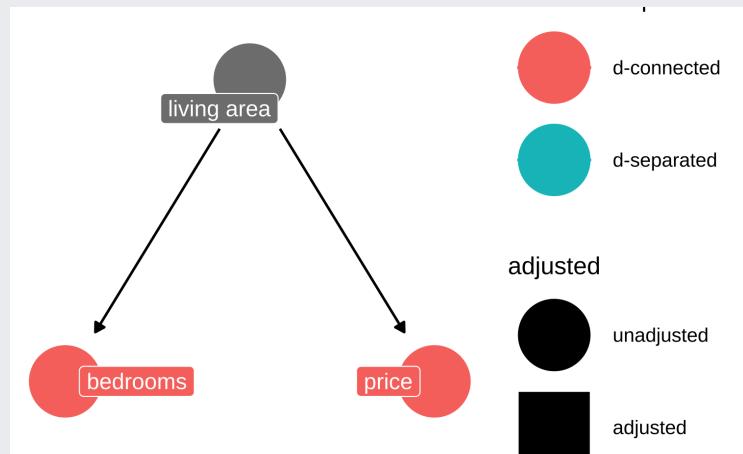
$b \perp\!\!\!\perp a$ : bedrooms sind unabhängig von livingArea (a)

⚠️ Passt nicht zu den Daten/zum Modell

# DAGs: Directed Acyclic Graphs

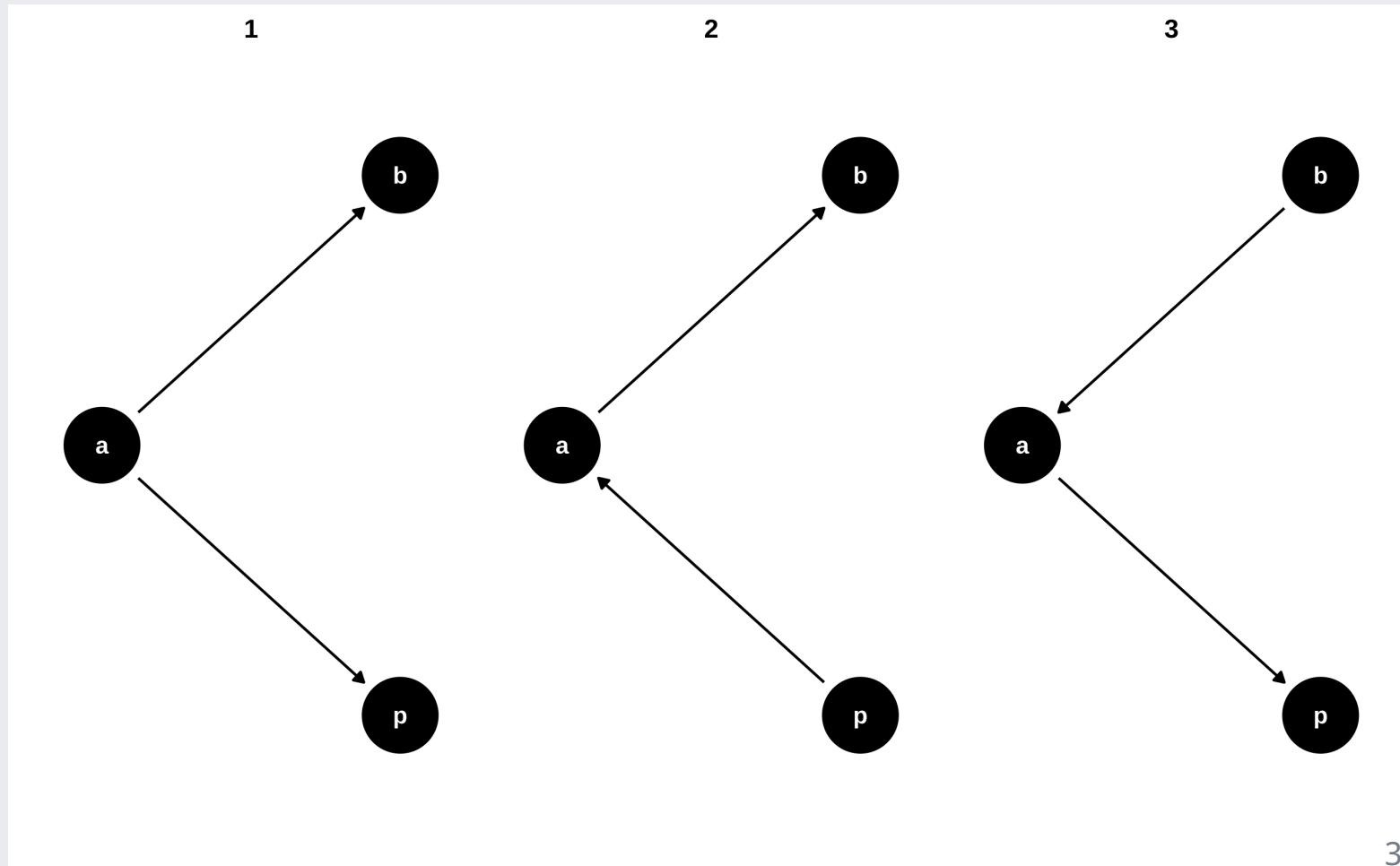
- DAGs sind eine bestimmte Art von Graphen zur Analyse von Kausalstrukturen.
- Ein *Graph* besteht aus Knoten (Variablen) und Kanten (Linien), die die Knoten verbinden.
- DAGs sind *gerichtet*; die Pfeile zeigen immer in eine Richtung (und zwar von Ursache zu Wirkung).
- DAGs sind *azyklisch*; die Wirkung eines Knoten darf nicht wieder auf ihn zurückführen.
- Ein *Pfad* ist ein Weg durch den DAG, von Knoten zu Knoten über die Kanten, unabhängig von der Pfeilrichtung.

## DAG von km1



# Leider passen potenziell viele DAGs zu einer Datenlage

b: bedrooms, p: price, a area (living area)

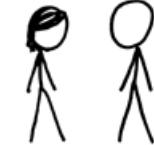


# Was ist eigentlich Kausation?

Weiβ man, was die Wirkung  $W$  einer Handlung  $H$  (Intervention) ist, so hat man  $H$  als Ursache von  $W$  erkannt.

(McElreath, 2020)

I USED TO THINK  
CORRELATION IMPLIED  
CAUSATION.



THEN I TOOK A  
STATISTICS CLASS.  
NOW I DON'T.



SOUNDS LIKE THE  
CLASS HELPED.  
WELL, MAYBE.



Quelle und Erklärung

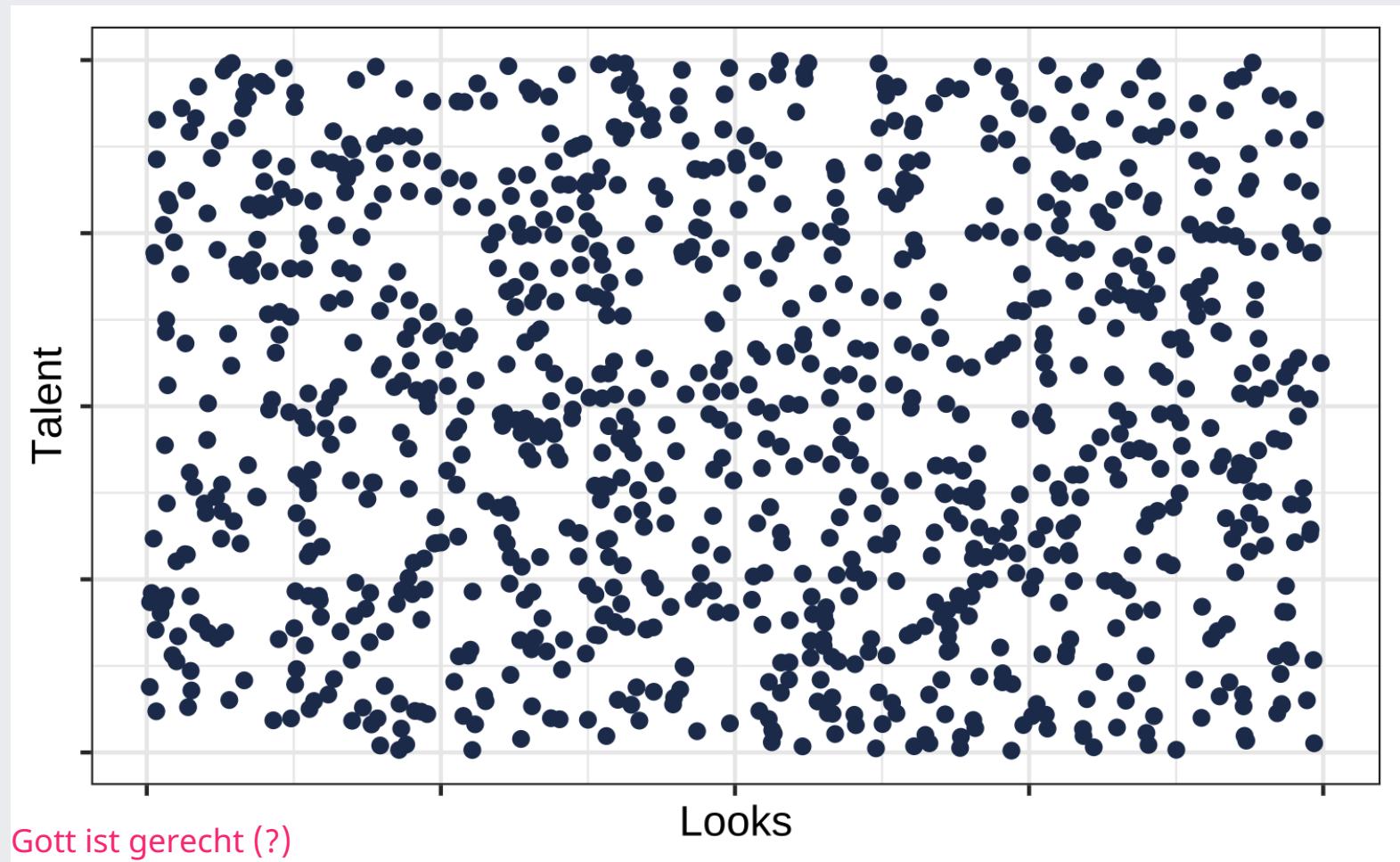
# Fazit

- Sind zwei Variablen korreliert (abhängig, assoziiert), so kann es dafür zwei Gründe geben:
  - Kausaler Zusammenhang
  - Nichtkausaler Zusammenhang ("Scheinkorrelation")
- Eine mögliche Ursache einer Scheinkorrelation ist Konfundierung.
- Konfundierung kann man entdecken, indem man die angenommene Konfundierungsvariable kontrolliert (adjustiert), z.B. indem man ihn als Prädiktor in eine Regression aufnimmt.
- Ist die Annahme einer Konfundierung korrekt, so löst sich der Scheinzusammenhang nach dem Adjustieren auf.
- Löst sich der Scheinzusammenhang nicht auf, sondern drehen sich die Vorzeichen der Zusammenhänge nach Adjustieren um, so spricht man einem *Simpson Paradox*.
- Die Daten alleine können nie sagen, welches Kausalmodell der Fall ist in einer Beobachtungsstudie. Fachwissen (inhaltliches wissenschaftliches Wissen) ist nötig, um DAGs auszuschließen.

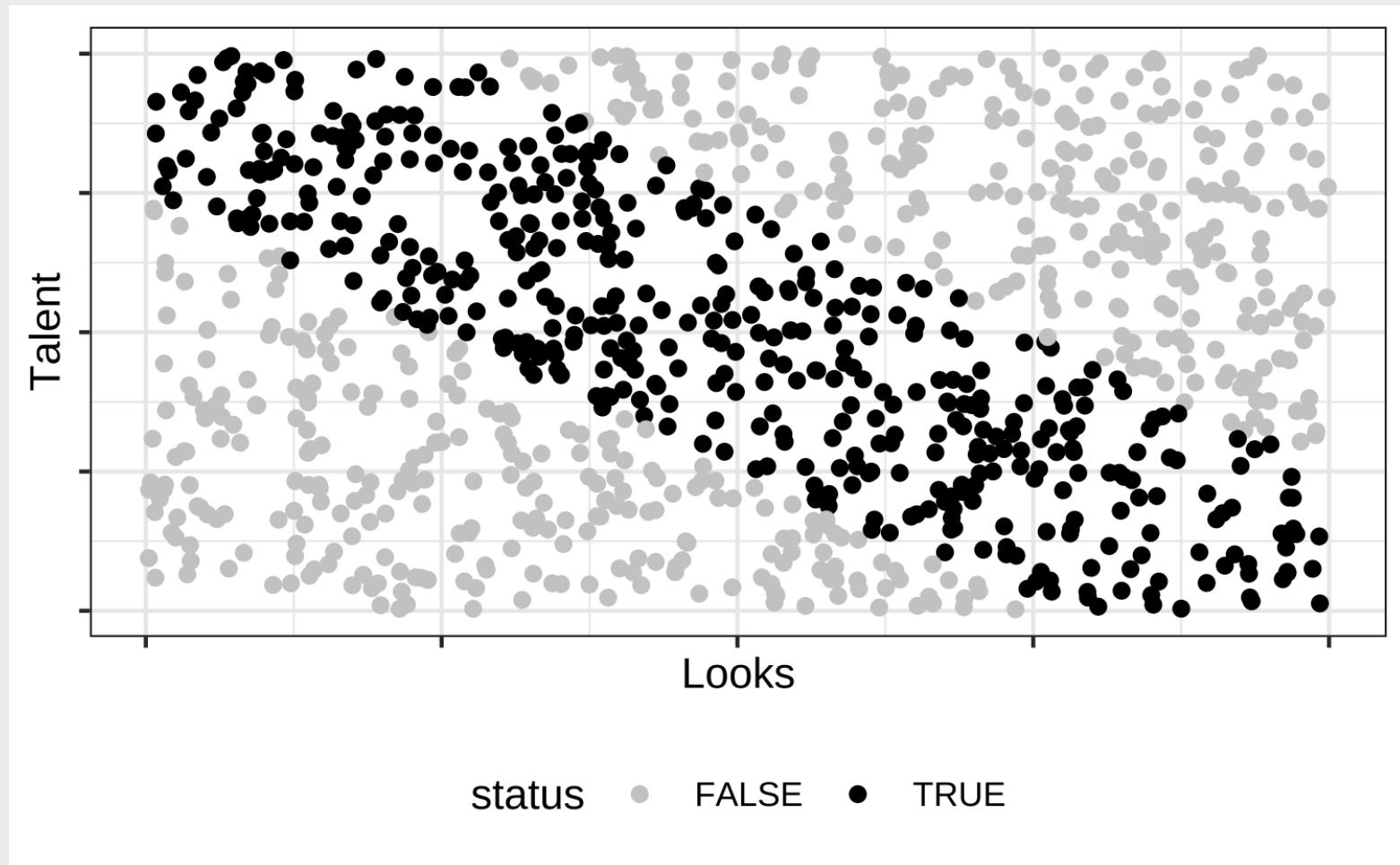
# Teil 3

## Kollision

# Kein Zusammenhang von Intelligenz und Schönheit



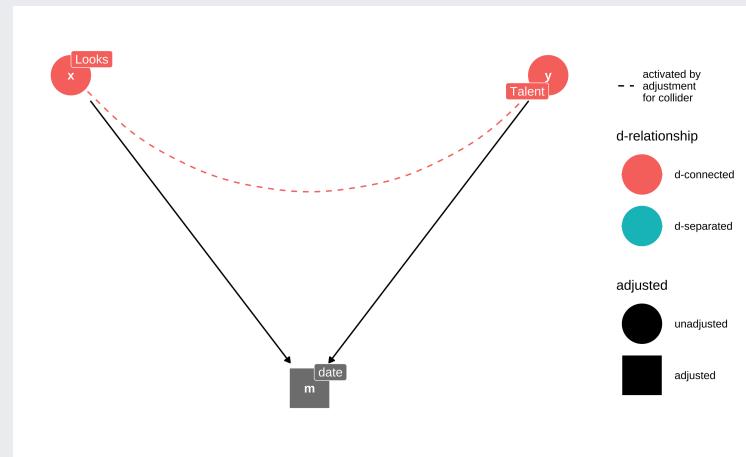
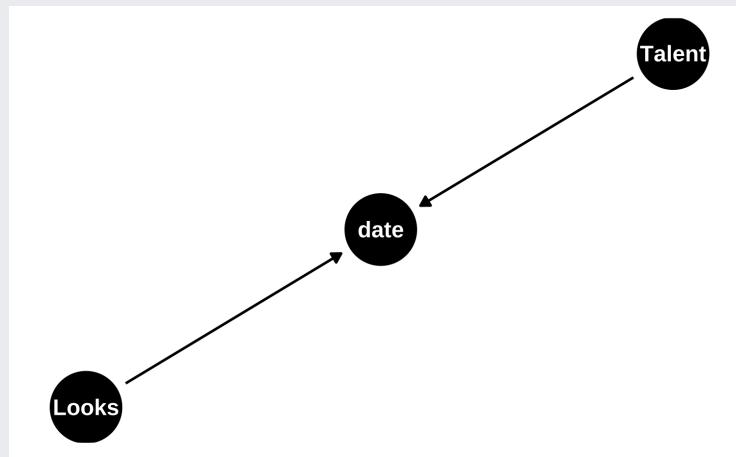
# Aber Ihre Dates sind entweder schlau oder schön



Wie kann das sein?

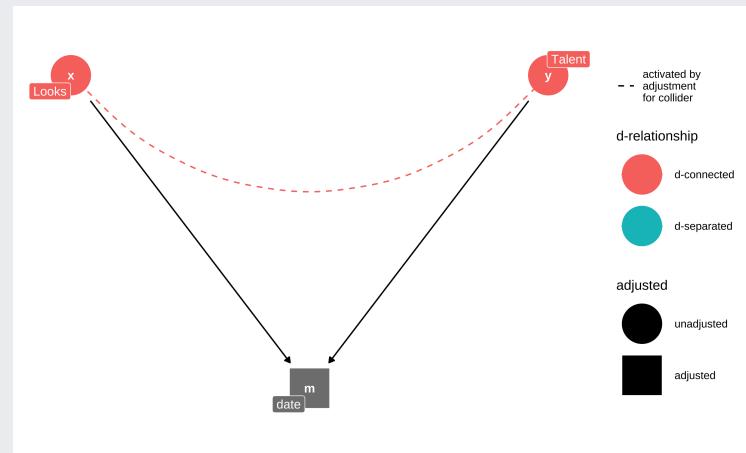
# DAG zur Rettung

Dieser DAG bietet eine rettende Erklärung:



# Was ist eine Kollision?

- Als *Kollision* (Kollisionsverzerrung, Auswahlverzerrung, engl. collider) bezeichnet man einen DAG, bei dem eine Wirkung zwei Ursachen hat (eine gemeinsame Wirkung zweier Ursachen).
- Kontrolliert man die *Wirkung*  $m$ , so entsteht eine Scheinkorrelation zwischen den Ursachen  $x$  und  $y$ .
- Kontrolliert man die Wirkung nicht, so entsteht keine Scheinkorrelation zwischen den Ursachen.



Vgl. Rohrer (2018).

Man kann also zu viele oder falsche Prädiktoren einer Regression hinzufügen, so dass die Koeffizienten nicht die kausalen Effekte zeigen, sondern durch Scheinkorrelation verzerrte Werte.

# Einfaches Beispiel zur Kollision

- In der Zeitung *Glitzer* werden nur folgende Menschen gezeigt:
  - Schöne Menschen
  - Reiche Menschen
- Gehen wir davon aus, dass Schönheit und Reichtum unabhängig voneinander sind.
- Wenn ich Ihnen sage, dass Don nicht schön ist, was lernen wir dann über seine finanzielle Situation?

"Ich bin schön, unglaublich schön, und groß, großartig, tolle Gene!!!"



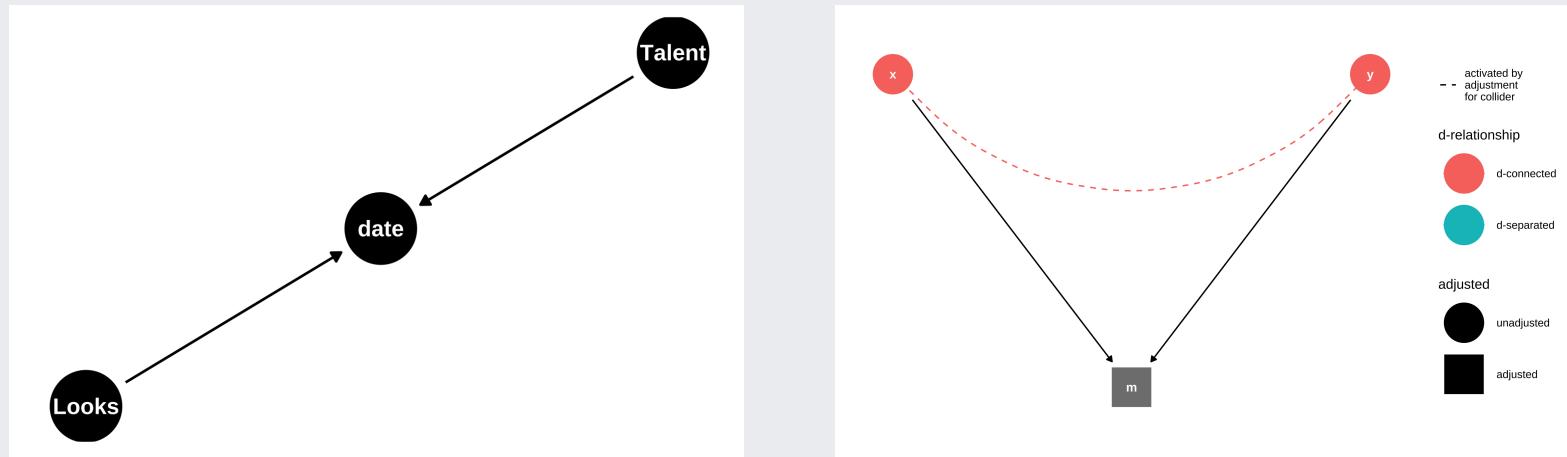
# Noch ein einfaches Beispiel zur Kollision

"So langsam  
check ich's!"



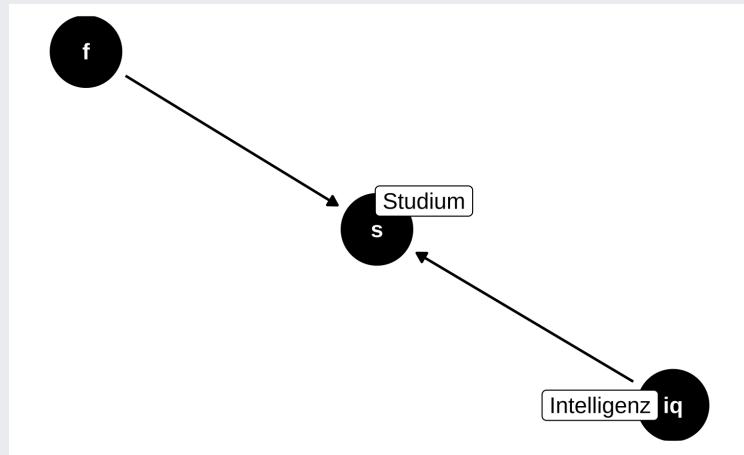
- Sei  $Z = X + Y$ , wobei  $X$  und  $Y$  unabhängig sind.
- Wenn ich Ihnen sage,  $X = 3$ , lernen Sie nichts über  $Y$ , da die beiden Variablen unabhängig sind
- Aber: Wenn ich Ihnen zuerst sage,  $Z = 10$ , und dann sage,  $X = 3$ , wissen Sie sofort, was  $Y$  ist ( $Y = 7$ ).
- Also:  $X$  und  $Y$  sind abhängig – gegeben  $Z$ :  $X \not\perp\!\!\!\perp Y | Z$ .

# Durch Kontrolle entsteht eine Verzerrung bei der Kollision



- Ohne Kontrolle von date entsteht keine Scheinkorrelation zwischen Looks und Talent. Der Pfad ("Fluss") von Looks über date nach Talent ist blockiert.
- Kontrolliert man date, so *öffnet* sich der Pfad Looks->date-> Talent und die Scheinkorrelation entsteht: Der Pfad ist nicht mehr blockiert.
- Das Kontrollieren von date geht zumeist durch Bilden einer Auswahl einer Teilgruppe von sich.

# IQ, Fleiss und Eignung fürs Studium



Bei positiver eignung wird ein Studium aufgenommen (studium = 1) ansonsten nicht (studium = 0).

Quelle

eignung (fürs Studium) sei definiert als die Summe von iq und fleiss, plus etwas Glück:

```
set.seed(42)  # Reproduzierbarkeit
N <- 1e03

d_eignung <-
tibble(
  iq = rnorm(N),
  fleiss = rnorm(N),
  glueck = rnorm(N, 0, sd = .1),
  eignung =
    1/2 * iq + 1/2 * fleiss +
    glueck,
  studium = ifelse(eignung > 0,
                  1, 0)
)
```

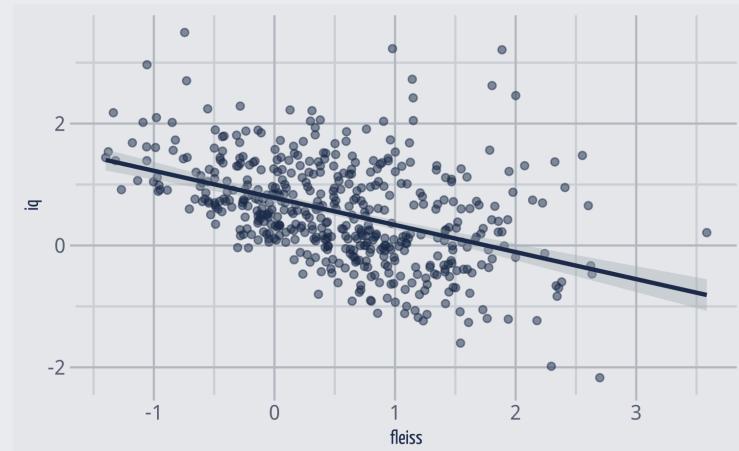
# Schlagzeile "Schlauheit macht faul!"

Eine Studie untersucht den Zusammenhang von Intelligenz und Fleiß bei Studenten.

Ergebnis: Ein negativer Zusammenhang.

```
m_eignung <-  
  stan_glm(  
    iq ~ fleiss,  
    data = d_eignung %>%  
      filter(studium == 1))
```

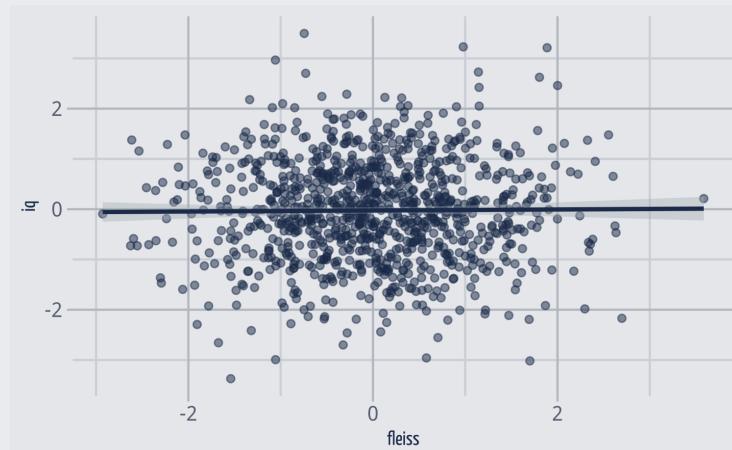
```
## (Intercept)      fleiss  
##   0.7806146   -0.4428830
```



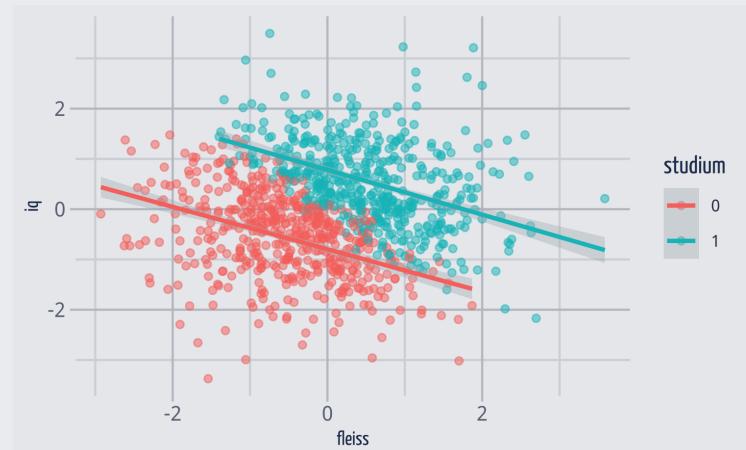
# Kollisionsverzerrung nur bei Stratifizierung

Nur durch das Stratifizieren (Aufteilen in Subgruppen, Kontrollieren, Adjustieren) tritt die Scheinkorrelation auf.

Ohne Stratifizierung tritt keine Scheinkorrelation auf



Mit Stratifizierung tritt Scheinkorrelation auf

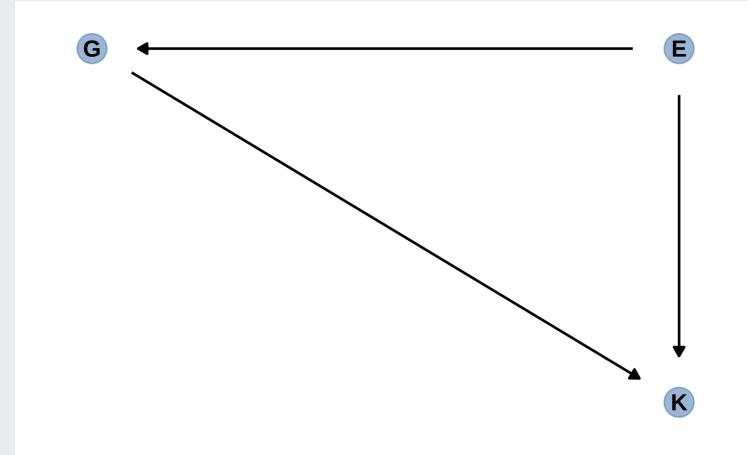


Kontrollieren einer Variablen - Aufnehmen in die Regression - kann genausogut schaden wie nützen.

Nur Kenntnis des DAGs verrät die richtige Entscheidung.

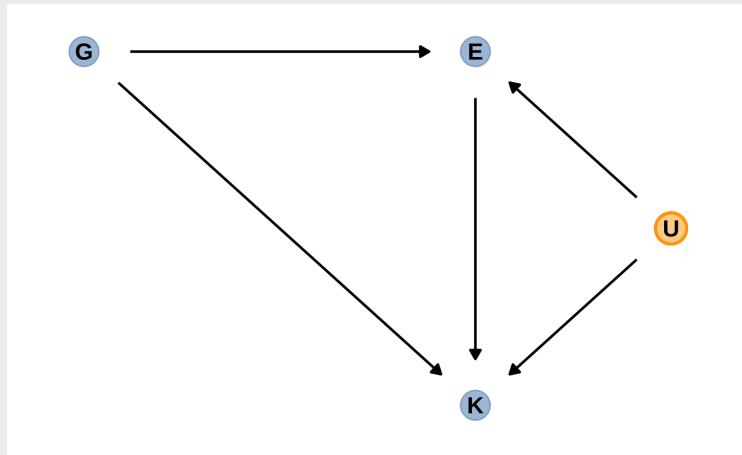
# Einfluss von Großeltern und Eltern auf Kinder

- Wir wollen den (kausalen) Einfluss der Eltern E und Großeltern G auf den Bildungserfolg der Kinder K. untersuchen.
- Wir nehmen folgende Effekte an:
  - indirekter Effekt von G auf K:  
 $G \rightarrow E \rightarrow K$
  - direkter Effekt von E auf K:  
 $E \rightarrow K$
  - direkter Effekt von G auf K:  
 $G \rightarrow K$



R-Syntax stammt von [Kurz \(2021\)](#).

# Der Gespenster-DAG



- U könnte ein ungemessener Einfluss sein, der auf E und K wirkt, etwa Nachbarschaft.
- Die Großeltern wohnen woanders (in Spanien), daher wirkt die Nachbarschaft der Eltern und Kinder nicht auf sie.
- E ist sowohl für G als auch für U eine Wirkung, also eine Kollisionsvariable auf diesem Pfad.
- Wenn wir E kontrollieren, wird es den Pfad  $G \rightarrow K$  verzerren, auch wenn wir niemals U messen.

Die Sache ist chancenlos. Wir müssen den DAG verloren geben. 🧟

(McElreath, 2020, S. 180)

## Teil 4

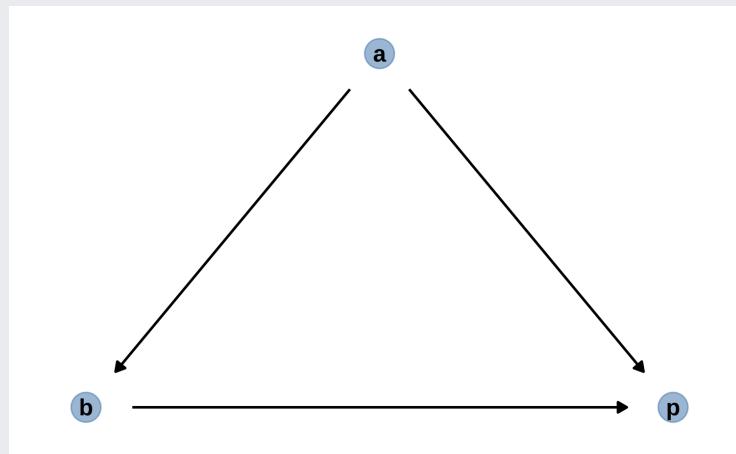
Die Hintertür schließen

# Zur Erinnerung: Konfundierung

Forschungsfrage: Wie groß ist der (kausale) Einfluss der Schlafzimmerzahl auf den Verkaufspreis des Hauses?

a: livingArea, b: bedrooms, p: prize

UV: b, AV: p

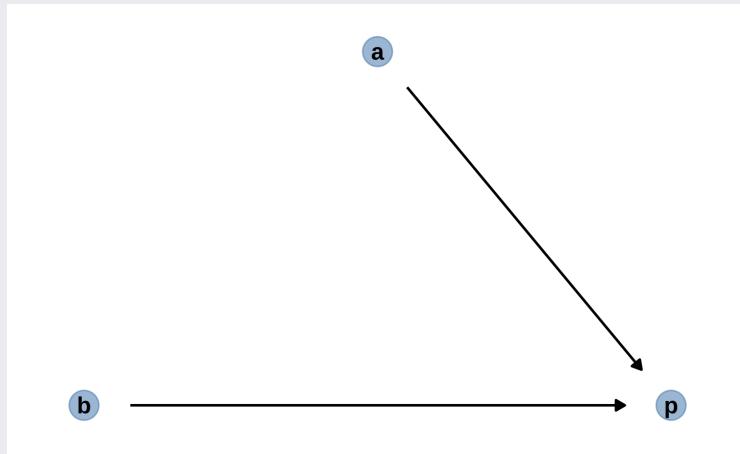


- Im Regressionsmodell  $p \sim b$  wird der kausale Effekt verzerrt sein durch die Konfundierung mit a.
- Der Grund für die Konfundierung sind die zwei Pfade zwischen b und p:

1.  $b \rightarrow p$
2.  $b \rightarrow a \rightarrow p$

- Beide Pfade erzeugen (statistische) Assoziation zwischen b und p.
- Aber nur der erste Pfad ist kausal; der zweite ist nichtkausal.
- Gäbe es nur den zweiten Pfad und wir würden b ändern, so würde sich p nicht ändern.

# Gute Experimente zeigen den echten kausalen Effekt



- Die "Hintertür" der UV (b) ist jetzt zu!
- Der einzige verbleibende, erste Pfad ist der kausale Pfad und die Assoziation zwischen b und p ist jetzt komplett kausal.

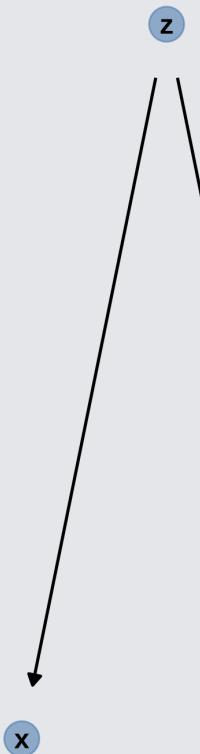
- Eine berühmte Lösung, den kausalen Pfad zu isolieren, ist ein (randomisiertes, kontrolliertes) Experiment.
- Wenn wir den Häusern zufällig (randomisiert) eine Anzahl von Schlafzimmern (b) zuweisen könnten (unabhängig von ihrer Quadratmeterzahl, a), würde sich der Graph so ändern.
- Das Experiment *entfernt* den Einfluss von a auf b.
- Wenn wir selber die Werte von b einstellen im Rahmen des Experiments, so kann a keine Wirkung auf b haben.
- Damit wird der zweite Pfad,  $b \rightarrow a \rightarrow p$  geschlossen ("blockiert").

# Hintertür schließen auch ohne Experimente

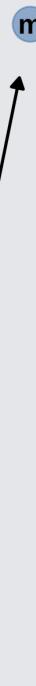
- Konfundierende Pfade zu blockieren zwischen der UV und der AV nennt man auch *die Hintertür schließen* (backdoor criterion).
- Wir wollen die Hintertüre schließen, da wir sonst nicht den wahren, kausalen Effekt bestimmen können.
- Zum Glück gibt es neben Experimenten noch andere Wege, die Hintertür zu schließen, wie die Konfundierungsvariable  $a$  in eine Regression mit aufzunehmen.
- Warum blockt das Kontrollieren von  $a$  den Pfad  $b \leftarrow a \rightarrow p$ ?
- Stellen Sie sich den Pfad als eigenen Modell vor.
- Sobald Sie  $a$  kennen, bringt Ihnen Kenntnis über  $b$  kein zusätzliches Wissen über  $p$ .
- Wissen Sie hingegen nichts über  $a$ , lernen Sie bei Kenntnis von  $b$  auch etwas über  $p$ .
- Konditionieren ist wie "gegeben, dass Sie  $a$  schon kennen...".
- $b \perp\!\!\!\perp p | a$

# Die vier Atome der Kausalanalyse

Die Konfundierung



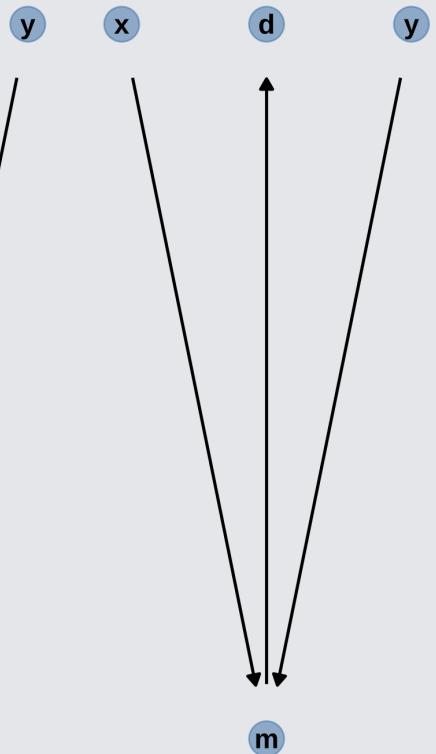
Die Mediation



Die Kollision

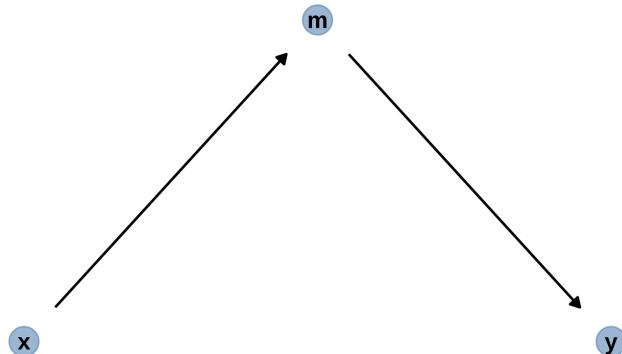


Der Nachfahre



# Mediation

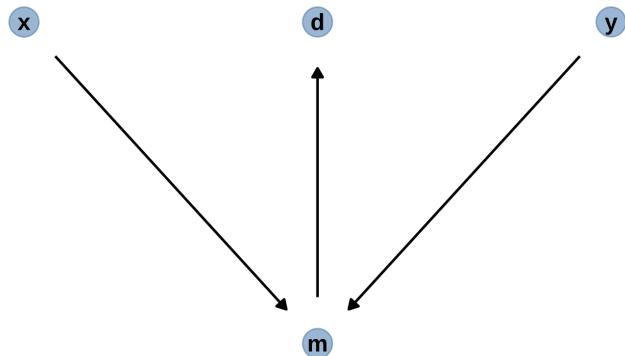
Die Mediation



- Die *Mediation* (Wirkkette, Rohr, Kette) beschreibt Pfade, in der die Kanten gleiche Wirkrichtung haben:  $x \rightarrow m \rightarrow y$ .
- Ohne Kontrollieren ist der Pfad offen: Die Assoziation "fließt" den Pfad entlang (in beide Richtungen).
- Kontrollieren blockt (schließt) die Kette (genau wie bei der Gabel).

# Der Nachfahre

Der Nachfahre



- Ein *Nachfahre* (descendent) ist eine Variable die von einer anderen Variable beeinflusst wird.
- Kontrolliert man einen Nachfahren d, so kontrolliert man damit zum Teil den Vorfahren (die Ursache), m.
- Der Grund ist, dass d Information beinhaltet über m.
- Hier wird das Kontrollieren von d den Pfad von x nach y teilweise öffnen, da m eine Kollisionsvariable ist.

# Kochrezept zur Analyse von DAGs

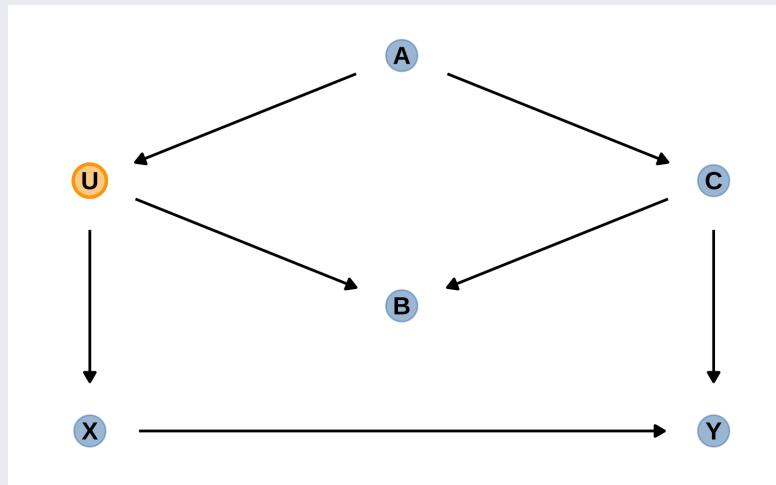
Wie kompliziert ein DAG auch aussehen mag, er ist immer aus diesen vier Atomen aufgebaut.

Hier ist ein Rezept, das garantiert, dass Sie welche Variablen Sie kontrollieren sollten und welche nicht:

1. Listen Sie alle Pfade von UV (X) zu AV (Y) auf.
2. Beurteilen Sie jeden Pfad, ob er gerade geschlossen oder geöffnet ist.
3. Beurteilen Sie für jeden Pfad, ob er ein Hintertürpfad ist (Hintertürpfade haben einen Pfeil, der zur UV führt).
4. Wenn es geöffnete Hinterpfade gibt, prüfen Sie, welche Variablen man kontrollieren muss, um den Pfad zu schließen (falls möglich).

# Schließen Sie die Hintertür (wenn möglich)!, bsp1

UV:  $X$ , AV:  $Y$ , drei Covariaten ( $A$ ,  $B$ ,  $C$ ) und ein ungemessene Variable,  $U$



Es gibt zwei Hintertürpfade:

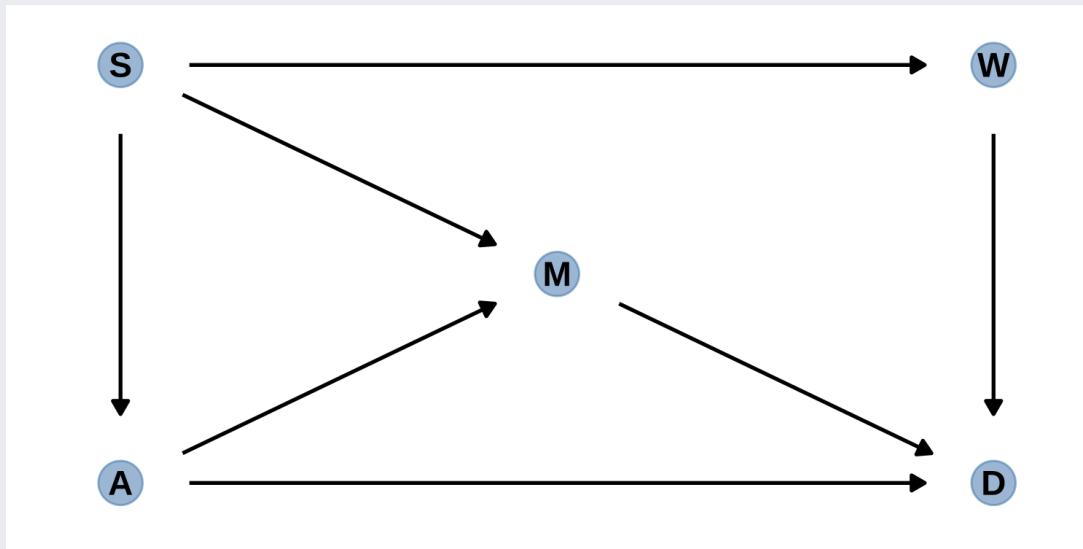
1.  $X \leftarrow U \leftarrow A \rightarrow C \rightarrow Y$ , offen
2.  $X \leftarrow U \rightarrow B \leftarrow C \rightarrow Y$ , geschlossen

Kontrollieren von  $A$  oder (auch)  $C$  schließt die offene Hintertür.

(McElreath, 2020; Kurz, 2021), s.S. 186.

# Schließen Sie die Hintertür (wenn möglich)!, bsp2

UV:  $W$ , AV:  $D$



Kontrollieren Sie diese Variablen, um die offenen Hintertüren zu schließen:

- entweder  $A$  und  $M$
- oder  $S$

Mehr Infos

(McElreath, 2020; Kurz, 2021), s.S. 188.

# Implizierte bedingte Unabhängigkeiten von bsp2

- Ein Graph ohne Us ist eine starke, oft zu starke (unrealistisch optimistische) Annahme.
- Auch wenn die Daten aus nicht sagen können, welcher DAG der richtige ist, können wir zumindest lernen, welcher DAG falsch ist.
- Einige testbare Modellimplikationen sind vom Modell implizierte bedingte Unabhängigkeiten.
- Bedingten Unabhängigkeit zwischen zwei Variablen sind Variablen, die nicht assoziiert (also stochastisch unabhängig) sind, wenn wir eine bestimmte Menge an Drittvariablen kontrollieren.
- bsp2 impliziert folgende bedingte Unabhängigkeiten:

```
## A _||_ W | S  
## D _|||_ S | A, M, W  
## M _|||_ W | S
```

# Fazit

- Wie (und sogar ob) Sie statistische Ergebnisse (z.B. eines Regressionsmodells) interpretieren können, hängt von der *epistemologischen Zielrichtung* der Forschungsfrage ab:
  - Bei *deskriptiven* Forschungsfragen können die Ergebnisse (z.B. Regressionskoeffizienten) direkt interpretiert werden. Z.B. "Der Unterschied zwischen beiden Gruppen beträgt etwa ...". Allerdings ist eine kausale Interpretation nicht zulässig.
  - Bei *prognostischen* Fragestellungen spielen die Modellkoeffizienten keine Rolle, stattdessen geht es um vorhergesagten Werte,  $\hat{y}_i$ , z.B. auf Basis der PPV. Kausalaussagen sind zwar nicht möglich, aber auch nicht von Interesse.
  - Bei *kausalen* Forschungsfragen dürfen die Modellkoeffizienten nur auf Basis eines Kausalmodells (DAG) oder eines (gut gemachten) Experiments interpretiert werden.
- Modellkoeffizienten ändern sich (oft), wenn man Prädiktoren zum Modell hinzufügt oder wegnimmt.
- Entgegen der verbreiteten Annahme ist es falsch, möglichst viele Prädiktoren in das Modell aufzunehmen, wenn das Ziel eine Kausalaussage ist.
- Kenntniss der "kausalen Atome" ist Voraussetzung zur Ableitung von Kausalschlüsse in Beobachtungsstudien.

# Hinweise

# Zu diesem Skript

- Dieses Skript bezieht sich auf folgende **Lehrbücher**:
  - Regression and other stories
- Dieses Skript wurde erstellt am 2021-12-04 17:36:39
- Lizenz: **MIT-Lizenz**
- Autor: Sebastian Sauer.
- Um diese HTML-Folien korrekt darzustellen, ist eine Internet-Verbindung nötig.
- Mit der Taste ? bekommt man eine Hilfe über Shortcuts.
- Wenn Sie die Endung .html in der URL mit .pdf ersetzen, bekommen Sie die PDF-Version der Datei.
- Alternativ können Sie im Browser Chrome die Folien als PDF drucken (klicken Sie auf den entsprechenden Menüpunkt).
- Den Quellcode der Skripte finden Sie **hier**.
- Eine PDF-Version kann erzeugt werden, indem man im Chrome-Browser die Webseite druckt (Drucken als PDF).



Homepage

# Literatur

Diese R-Pakete wurden verwendet.

Kurz, A. S. (2021). *Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition*.

McElreath, R. (2020). *Statistical rethinking: a Bayesian course with examples in R and Stan*. 2nd ed. CRC texts in statistical science. Taylor and Francis, CRC Press.

Messerli, F. H. (2012). "Chocolate Consumption, Cognitive Function, and Nobel Laureates". In: *New England Journal of Medicine* 367.16, pp. 1562-1564. DOI: [10.1056/NEJMon1211064](https://doi.org/10.1056/NEJMMon1211064).

Pearl, J., M. Glymour, and N. P. Jewell (2016). *Causal inference in statistics: a primer*. Wiley. 136 pp.

Rohrer, J. M. (2018). "Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data". In: *Advances in Methods and Practices in Psychological Science* 1.1, pp. 27-42. DOI: [10.1177/2515245917745629](https://doi.org/10.1177/2515245917745629).

