Start:Bayes!

Sebastian Sauer

2022-09-04

Inhaltsverzeichnis

Hinweise

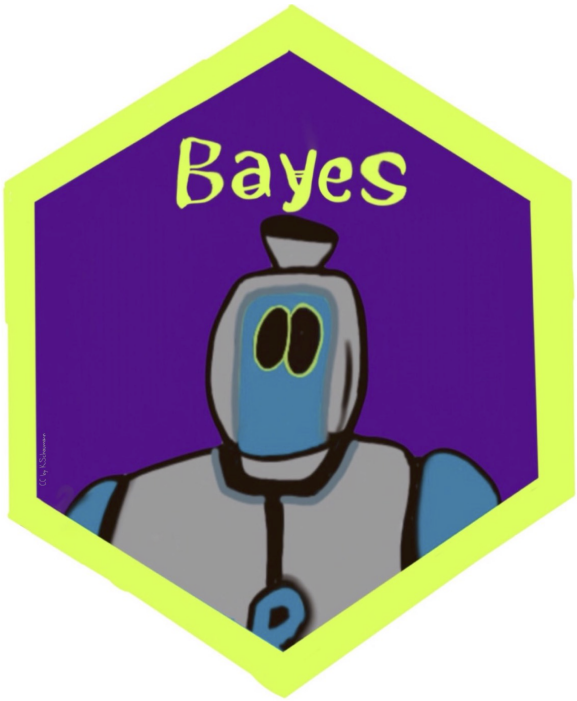


Abbildung 1: Bayes:Start!

Bildquelle: Klara Schaumann

WORK IN PROGRESS

Lernziele

Nach diesem Kurs sollten Sie …

- grundlegende Konzepte der Inferenzstatistik mit Bayes verstehen und mit R anwenden können

- gängige einschlägige Forschungsfragen in statistische Modelle übersetzen und mit R auswerten können
- kausale Forschungsfragen in statistische Modelle übersetzen und prüfen können
- die Güte und Grenze von statistischen Modellen einschätzen können

Voraussetzungen

Um von diesem Kurs am besten zu profitieren, sollten Sie folgendes Wissen mitbringen:

- grundlegende Kenntnisse im Umgang mit R, möglichst auch mit dem tidyverse
- grundlegende Kenntnisse der deskriptiven Statistik
- grundlegende Kenntnis der Regressionsanalyse

Software

- Installieren Sie R und seine Freunde.
- Für die Bayes-Inferenz brauchen Sie[1] zusätzliche Software, was leider etwas Zusatzaufwand erfordert. Lesen Sie hier die Hinweise dazu.
- Installieren Sie die folgende R-Pakete[2]:
  - tidyverse
  - rstanarm
  - easystats
  - weitere Pakete werden im Unterricht bekannt gegeben (es schadet aber nichts, jetzt schon Pakete nach eigenem Ermessen zu installieren)
- R Syntax aus dem Unterricht findet sich im Github-Repo bzw. Ordner zum jeweiligen Semester.

PDF-Version

Von diesem "Webbuch" gibt es hier eine PDF-Version.

Lernhilfen

Videos

Auf dem YouTube-Kanal des Autors finden sich eine Reihe von Videos mit Bezug zum Inhalt dieses Buchs. Besonders diese Playlist passt zu den Inhalten dieses Buchs.

[1] nicht gleich zu Beginn, aber nach 2-3 Wochen
[2] falls Sie die Pakete schon installiert haben, könnten Sie mal in RStudio auf "update.packages" klicken

Online-Zusammenarbeit

Hier finden Sie einige Werkzeuge, die das Online-Zusammenarbeiten vereinfachen:

- Frag-Jetzt-Raum zum anonymen Fragen stellen während des Unterrichts. Der Keycode wird Ihnen bei Bedarf vom Dozenten bereitgestellt.
- Padlet zum einfachen (und anonymen) Hochladen von Arbeitsergebnissen der Studentis im Unterricht. Wir nutzen es als eine Art Pinwand zum Sammeln von Arbeitsbeiträgen. Die Zugangsdaten stellt Ihnen der Dozent bereit.

Modulzeitplan

Nr	Thema	Datum	Kommentar
1	Was ist Inferenz?	3. - 7. Okt. 2022	Erster Termin: 4. Okt. 22, 14.15-15.00
2	Wahrscheinlichkeit	10. - 14. Okt. 22	NA
3	Verteilungen	17. - 21. Okt. 22	NA
4	Globusversuch	24. - 28. Okt. 22	NA
5	Aufhol-Woche	31. Okt. - 4. Nov. 22	Am Di., 1.11. entfällt die Vorlesung. Am Do., 3. 11. e
6	Frag die Post	7. - 11. Nov. 22	Ab diese Woche benötigen wir rstanarm.
NA	NA	14. - 18. Nov. 22	Blockwoche: Kein regulärer Unterricht
7	Gauss-Modelle	21. - 25. Nov. 22	NA
8	Lineare Modelle	28. Nov. - 2. Dez. 22	NA
9	Metrische AV	5. Dez. - 9. Dez. 22	NA
10	Kausalinferenz 1	12. - 16. Dez. 22	NA
11	Kausalinferenz 2	19. - 23. Dez. 22	NA
NA	NA	NA	Jahreswechsel: Kein Unterricht
12	Abschluss	9. Jan. 23 - 13. Jan. 23	NA

Literatur

Pro Thema wird Literatur ausgewiesen.

Technische Details

Dieses Dokument wurde erzeugt am/um 2022-11-30 13:39:57.

```
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.2.1 (2022-06-23)
##  os       macOS Big Sur ... 10.16
```

```
##  system   x86_64, darwin17.0
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       Europe/Berlin
##  date     2022-11-30
##  pandoc   2.19.2 @ /Applications/RStudio.app/Contents/MacOS/quarto/bin/tools/ (via rmar
##
## - Packages ---------------------------------------------------------------
##  package     * version date (UTC) lib source
##  assertthat    0.2.1   2019-03-21 [1] CRAN (R 4.2.0)
##  cellranger    1.1.0   2016-07-27 [1] CRAN (R 4.2.0)
##  cli           3.4.1   2022-09-23 [1] CRAN (R 4.2.0)
##  colorout    * 1.2-2   2022-06-13 [1] local
##  colorspace    2.0-3   2022-02-21 [1] CRAN (R 4.2.0)
##  DBI           1.1.3   2022-06-18 [1] CRAN (R 4.2.0)
##  digest        0.6.30  2022-10-18 [1] CRAN (R 4.2.0)
##  dplyr         1.0.10  2022-09-01 [1] CRAN (R 4.2.0)
##  evaluate      0.17    2022-10-07 [1] CRAN (R 4.2.0)
##  fansi         1.0.3   2022-03-24 [1] CRAN (R 4.2.0)
##  fastmap       1.1.0   2021-01-25 [1] CRAN (R 4.2.0)
##  generics      0.1.3   2022-07-05 [1] CRAN (R 4.2.0)
##  ggplot2       3.4.0   2022-11-04 [1] CRAN (R 4.2.0)
##  glue          1.6.2   2022-02-24 [1] CRAN (R 4.2.0)
##  gt            0.7.0   2022-08-25 [1] CRAN (R 4.2.0)
##  gtable        0.3.1   2022-09-01 [1] CRAN (R 4.2.0)
##  htmltools     0.5.3   2022-07-18 [1] CRAN (R 4.2.0)
##  jsonlite      1.8.3   2022-10-21 [1] CRAN (R 4.2.1)
##  knitr         1.40    2022-08-24 [1] CRAN (R 4.2.0)
##  lifecycle     1.0.3   2022-10-07 [1] CRAN (R 4.2.0)
##  magrittr      2.0.3   2022-03-30 [1] CRAN (R 4.2.0)
##  munsell       0.5.0   2018-06-12 [1] CRAN (R 4.2.0)
##  pillar        1.8.1   2022-08-19 [1] CRAN (R 4.2.0)
##  pkgconfig     2.0.3   2019-09-22 [1] CRAN (R 4.2.0)
##  R6            2.5.1   2021-08-19 [1] CRAN (R 4.2.0)
##  readxl        1.4.1   2022-08-17 [1] CRAN (R 4.2.0)
##  rlang         1.0.6   2022-09-24 [1] CRAN (R 4.2.0)
##  rmarkdown     2.17    2022-10-07 [1] CRAN (R 4.2.0)
##  rstudioapi    0.14    2022-08-22 [1] CRAN (R 4.2.0)
##  scales        1.2.1   2022-08-20 [1] CRAN (R 4.2.0)
##  sessioninfo   1.2.2   2021-12-06 [1] CRAN (R 4.2.0)
##  stringi       1.7.8   2022-07-11 [1] CRAN (R 4.2.0)
##  stringr       1.4.1   2022-08-20 [1] CRAN (R 4.2.0)
##  tibble        3.1.8   2022-07-22 [1] CRAN (R 4.2.0)
##  tidyselect    1.2.0   2022-10-10 [1] CRAN (R 4.2.0)
```

```
##  utf8          1.2.2   2021-07-24 [1] CRAN (R 4.2.0)
##  vctrs         0.5.1   2022-11-16 [1] CRAN (R 4.2.0)
##  withr         2.5.0   2022-03-03 [1] CRAN (R 4.2.0)
##  xfun          0.34    2022-10-18 [1] CRAN (R 4.2.0)
##  yaml          2.3.6   2022-10-18 [1] CRAN (R 4.2.0)
##
##  [1] /Users/sebastiansaueruser/Rlibs
##  [2] /Library/Frameworks/R.framework/Versions/4.2/Resources/library
##
## ------------------------------------------------------------------------------
```

1 Kausalinferenz

1.1 Lernsteuerung

1.1.1 R-Pakete

Für dieses Kapitel benötigen Sie folgende R-Pakete:

```
library(dagitty)
library(tidyverse)
library(rstanarm)
library(easystats)
```

1.1.2 Lernziele

Nach Absolvieren des jeweiligen Kapitel sollen folgende Lernziele erreicht sein.

Sie können …

- rklären, wann eine Kausalaussage gegeben eines DAGs berechtigt ist
- die "Atome" der Kausalität eines DAGs benennen
- "kausale Hintertüren" schließen

1.2 Statistik, was soll ich tun?

1.2.1 Studie A: Östrogen

Mit Blick auf Tabelle 1.1: Was raten Sie dem Arzt? Medikament einnehmen, ja oder nein?

Tabelle 1.1: Daten zur Studie A

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 raten? Soll er das Medikament verschreiben? Vielleicht nur dann, wenn er das Geschlecht kennt Pearl, Glymour, und Jewell (2016)?

1.2.2 Kausalmodell zur Studie A

In Wahrheit sehe die kausale Struktur so aus: 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, confounded). Die kausale Struktur, also welche Variable beeinflusst bzw. nicht, ist in Abbildung 1.1 dargestellt.

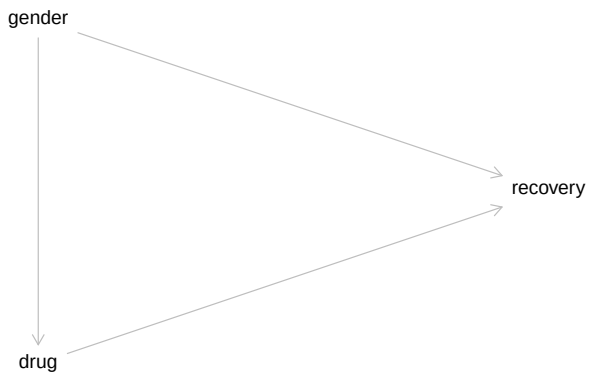


Abbildung 1.1: Zwei direkte Effekte (gender, drug) und ein indirekter Effekt (gender über drug) auf recovery

> **! Wichtig**
>
> Betrachtung der Teildaten (d.h. stratifiziert pro Gruppe) zeigt in diesem Fall den wahren, kausalen Effekt. Stratifizieren ist also in diesem Fall der korrekte, richtige Weg.

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

> **! Wichtig**
>
> Achtung: Das Stratifizieren ist nicht immer und nicht automatisch die richtige Lösung.

1.2.3 Studie B: Blutdruck

Mit Blick auf Tabelle 1.2: Was raten Sie dem Arzt? Medikament einnehmen, ja oder nein?

Tabelle 1.2: Daten zur Wirksamkeit eines Medikaments (Studie B)

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 Blutdrck scheint das Medikamenet 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, und Jewell (2016)

1.2.4 Kausalmodell zur Studie B

Das Medikament hat einen (absenkenden) Einfluss auf den Blutdruck. Gleichzeitig hat das Medikament einen (toxischen) Effekt auf die Heilung. Verringerter 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.

Das Kausalmodell ist in Abbildung 1.2 dargestellt.

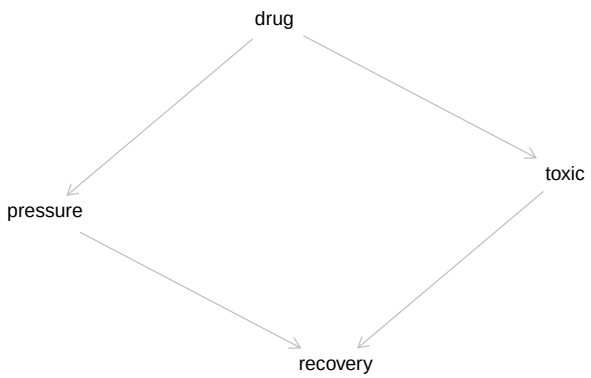


Abbildung 1.2: Drug hat keinen direkten, aber zwei indirekte Effekt auf recovery, einer davon ist heilsam, einer schädlich

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

> **! Wichtig**
>
> Betrachtung der Gesamtdaten zeigt in diesem Fall den wahren, kausalen Effekt. Stratifizieren wäre falsch, da dann nur der toxische Effekt, aber nicht der heilsame Effekt sichtbar wäre.

1.2.5 Studie A und B: Gleiche Daten, unterschiedliches Kausalmodell

Vergleichen Sie die DAGs Abbildung 1.1 und Abbildung 1.2, die die Kausalmodelle der Studien A und B darstellen: Sie sind *unterschiedlich*.

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

1.2.6 Sorry, Statistik: Du allein schaffst es nicht

Statistik alleine reicht nicht für Kausalschlüsse. Statistik plus Theorie erlaubt Kausalschlüsse.

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

1.2.7 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?

1.2.8 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 ("Kette") von Größe → Behandlung → Heilung. Darüber hinaus gibt es noch einen Einfluss von Größe der Nierensteine auf die Heilung.

Das Kausalmodell ist in **?@fig-dag-studie-c** dargestellt.

Sollte man hier `size` kontrollieren, wenn man den Kausaleffekt von `treatment` schätzen möchte?

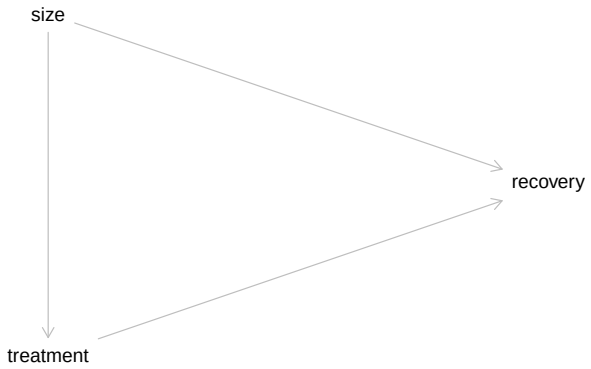


Abbildung 1.3: DAG zur Nierenstein-Studie in zwei Darstellungsformen

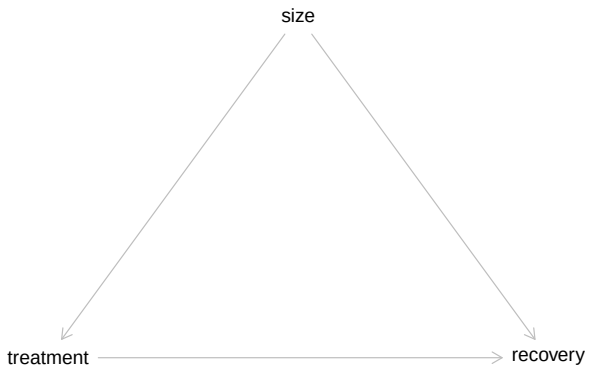


Abbildung 1.4: DAG zur Nierenstein-Studie in zwei Darstellungsformen

1.2.9 Mehr Beispiele

Nehmen Sie Bezug zu folgenden Aussagen:

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

1.3 Konfundierung

1.3.1 Datensatz 'Hauspreise im Saratoga County'

Datenquelle; Beschreibung des Datensatzes

```
d_path <- "https://vincentarelbundock.github.io/Rdatasets/csv/mosaicData/SaratogaHouses
```

1.3.2 Immobilienpreise in einer schicken Wohngegend vorhersagen

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

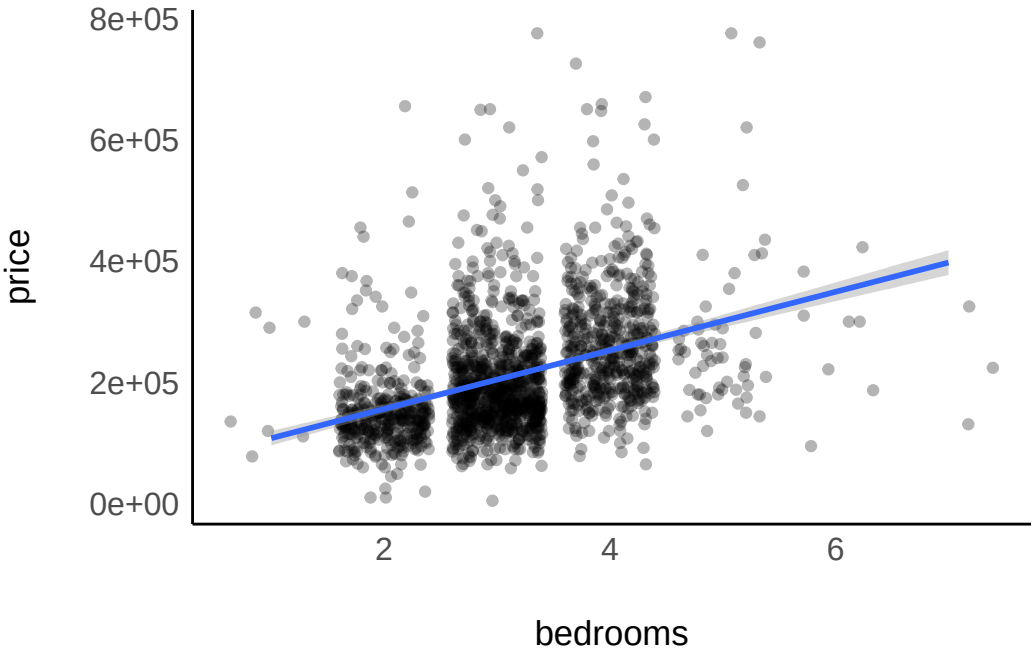
Das ist Don, Immobilienmogul, Auftraggeber.

Das finde ich heraus. Ich mach das wissenschaftlich.

Das ist Angie, Data Scientistin.

1.3.3 Modell 1: Preis als Funktion der Anzahl der Zimmer

"Hey Don! Mehr Zimmer, mehr Kohle!"



1.3.4 Posteriori-Verteilung von Modell 1

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

Zu wenig!

Berechnen wir das Modell:

```
m1 <- stan_glm(price ~ bedrooms,
               refresh = 0,
               data = d)
hdi(m1)
## Highest Density Interval
##
## Parameter   |              95% HDI
## ----------------------------------
## (Intercept) | [43170.53, 76800.33]
## bedrooms    | [43171.46, 53452.42]
```

Mit `estimate_preditioncs` können wir Vorhersagen berechnen (bzw. schätzen; die Vorhersagen sind ja mit Ungewissheit verbunden, daher ist "schätzen" vielleicht das treffendere Wort):

```
dons_house <- tibble(bedrooms = 2)
estimate_prediction(m1, data = dons_house)
## Model-based Prediction
##
## bedrooms | Predicted |       SE |                95% CI
## ---------------------------------------------------------
## 2.00     |  1.55e+05 | 88999.69 | [-27865.52, 3.24e+05]
##
## Variable predicted: price
```

1.3.5 Don hat eine Idee

> "Ich bau eine Mauer! Genial! An die Arbeit, Angie!

Don hofft, durch Verdopplung der Zimmerzahl den doppelten Verkaufspreis zu erzielen. Ob das klappt?

> Das ist keine gute Idee, Don."

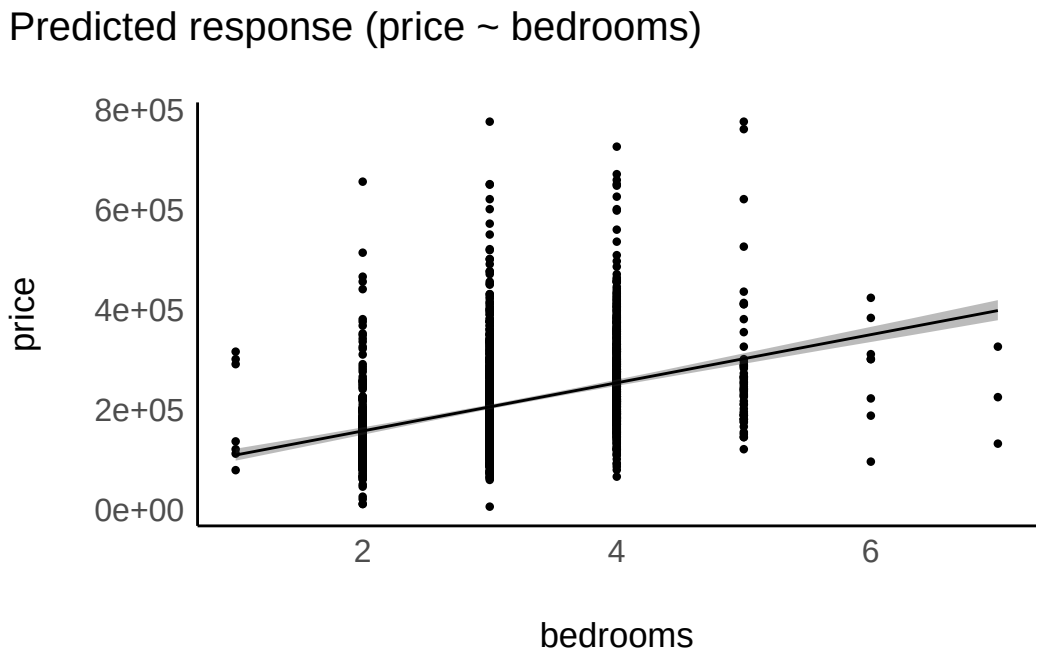
Berechnen wir die Vorhersagen für Dons neues Haus (mit den durch Mauern halbierten Zimmern).

```
dons_new_house <- tibble(bedrooms = 4)
estimate_prediction(m1, dons_new_house)
## Model-based Prediction
```

```
##
## bedrooms | Predicted |       SE |                95% CI
## -----------------------------------------------------
## 4.00     |  2.52e+05 | 89606.38 | [66925.37, 4.19e+05]
##
## Variable predicted: price
```

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

Volltreffer! Jetzt verdien ich 100 Tausend mehr! Ich bin der Größte!



> **Hinweis**
>
> Zur Erinnerung: "4e+05" ist die Kurzform der wissenschaftlichen Schreibweise und bedeutet: $4 \cdot 100000 = 4 \cdot 10^5 = 400000$

1.3.6 R-Funktionen, um Beobachtungen vorhersagen

`estimate_prediction(m1, dons_new_house)` erstellt *Vorhersageintervalle*, berücksichtigt also *zwei Quellen* von Ungewissheit:

- Ungewissheiten in den Parametern (Modellkoeffizienten, $\beta_0, \beta_1, ...$)
- Ungewissheit im "Strukturmodell": Wenn also z.B. in unserem Modell ein wichtiger Prädiktor fehlt, so kann die Vorhersagen nicht präzise sein. Fehler im Strukturmodell schlagen sich in breiten Schätzintervallen (bedingt durch ein großes σ) nieder.

`estimate_expectation(m1, dons_new_house)` erstellt *Konfidenzintervalle*. berücksichtigt also nur *eine Quelle* von Ungewissheit:

- Ungewissheiten in den Parametern (Modellkoeffizienten, $\beta_0, \beta_1, ...$)

Die Schätzbereiche sind in dem Fall deutlich kleiner:

```
estimate_expectation(m1, dons_new_house)
## Model-based Expectation
##
## bedrooms | Predicted |      SE |               95% CI
## -----------------------------------------------------
## 4.00     |  2.53e+05 | 3173.22 | [2.46e+05, 2.59e+05]
##
## Variable predicted: price
```

1.3.7 Modell 2: `price ~ bedrooms + livingArea`

Berechnen wir das Modell `m2: price ~ bedrooms + livingArea`.

```
m2 <- stan_glm(price ~ bedrooms + livingArea, data = d, refresh = 0)

hdi(m2)
## Highest Density Interval
##
## Parameter   |                95% HDI
## ------------------------------------
## (Intercept) | [ 23946.99, 49850.00]
## bedrooms    | [-19313.12, -8877.84]
## livingArea  | [   118.65,   132.09]
```

Was sind die Vorhersgaen des Modell?

```
estimate_prediction(m2, data = tibble(bedrooms = 4, livingArea = 1200))
## Model-based Prediction
##
## bedrooms | livingArea | Predicted |       SE |               95% CI
## ---------------------------------------------------------------------
## 4.00     |    1200.00 |  1.29e+05 | 68425.54 | [-8414.60, 2.56e+05]
##
## Variable predicted: price
```

Andere, aber ähnliche Frage: Wieviel Haus kostet ein Haus mit sagen wir 4 Zimmer *gemittelt* über die verschiedenen Größen von `livingArea`? Stellen Sie sich alle Häuser mit 4 Zimmern vor (also mit verschiedenen Wohnflächen). Wir möchten nur wissen, was so ein Haus "im

Mittel" kostet. Wir möchten also die Mittelwerte pro `bedroom` schätzen, gemittelt für jeden Wert von `bedroom` über `livingArea`:

```
estimate_means(m2, at = "bedrooms", length = 7)
## Estimated Marginal Means
##
## bedrooms |     Mean |               95% CI
## ----------------------------------------
## 1.00     | 2.43e+05 | [2.31e+05, 2.54e+05]
## 2.00     | 2.28e+05 | [2.21e+05, 2.35e+05]
## 3.00     | 2.14e+05 | [2.11e+05, 2.17e+05]
## 4.00     | 2.00e+05 | [1.95e+05, 2.05e+05]
## 5.00     | 1.86e+05 | [1.76e+05, 1.96e+05]
## 6.00     | 1.72e+05 | [1.56e+05, 1.87e+05]
## 7.00     | 1.57e+05 | [1.37e+05, 1.77e+05]
##
## Marginal means estimated at bedrooms
```

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

"Verringert!? Weniger Geld?! Oh nein!"

1.3.8 Die Zimmerzahl ist negativ mit dem Preis korreliert

… wenn man die Wohnfläche (Quadratmeter) kontrolliert.

"Ne-Ga-Tiv!"

Hauspreis stratifizieren

Quellcode

1.4 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 β 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, eher umgangssprachlich, 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, $x_2, x_3, ...$ 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.

1.4.1 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.

Wozu ist sie dann gut?!

! Wichtig

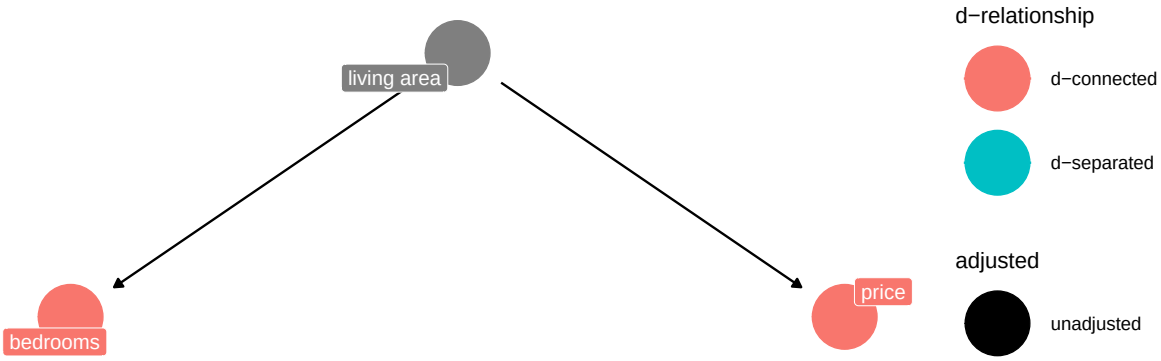
In Beobachtungsstudien hilft nur ein (korrektes) Kausalmodell. Ohne Kausalmodell ist es nutzlos, die Regressionskoeffizienten (oder eine andere Statistik) zur Erklärung der Ursachen heranzuziehen.

1.5 Welches Modell richtig ist, kann die Statistik nicht sagen

Often people want statistical modeling to do things that statical modeling cannot do. For example, we'd like to know wheter an effect is "real" or rather spurios. Unfortunately, modeling merely quantifies uncertainty in the precise way that the model understands the problem. Usually answers to lage 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 apparaent accuracy of their predictions. Rounds of model criticism and revision embody 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

1.5.1 Kausalmodell für Konfundierung, `km1`



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

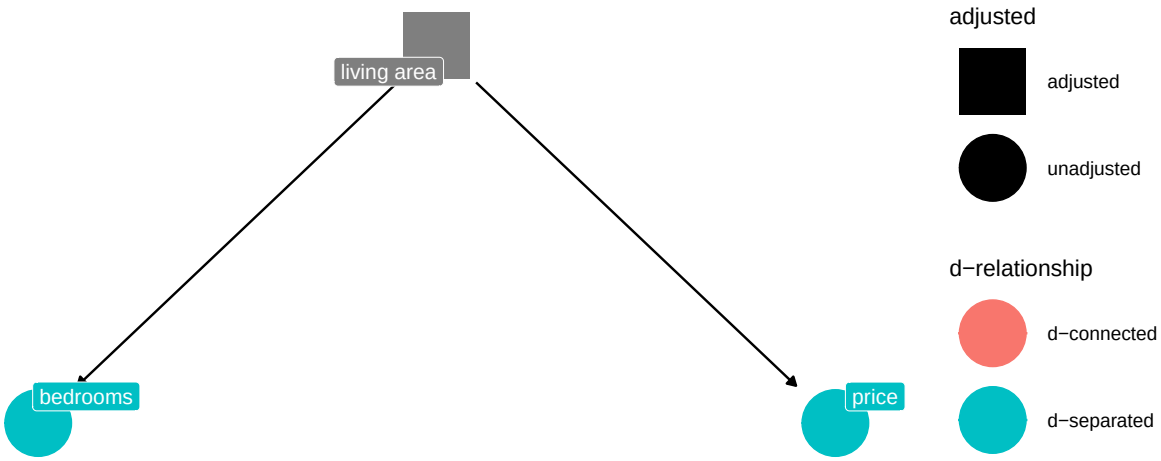
Eine Scheinkorrelation ist ein Zusammenhang, der *nicht* auf eine kausalen Einfluss beruht.

`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.

1.5.2 `m2` kontrolliert die Konfundierungsvariable `livingArea`

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

> Was tun wir jetzt bloß?! Oh jeh!
>
> Wir müssen die Konfundierungsvariable kontrollieren.

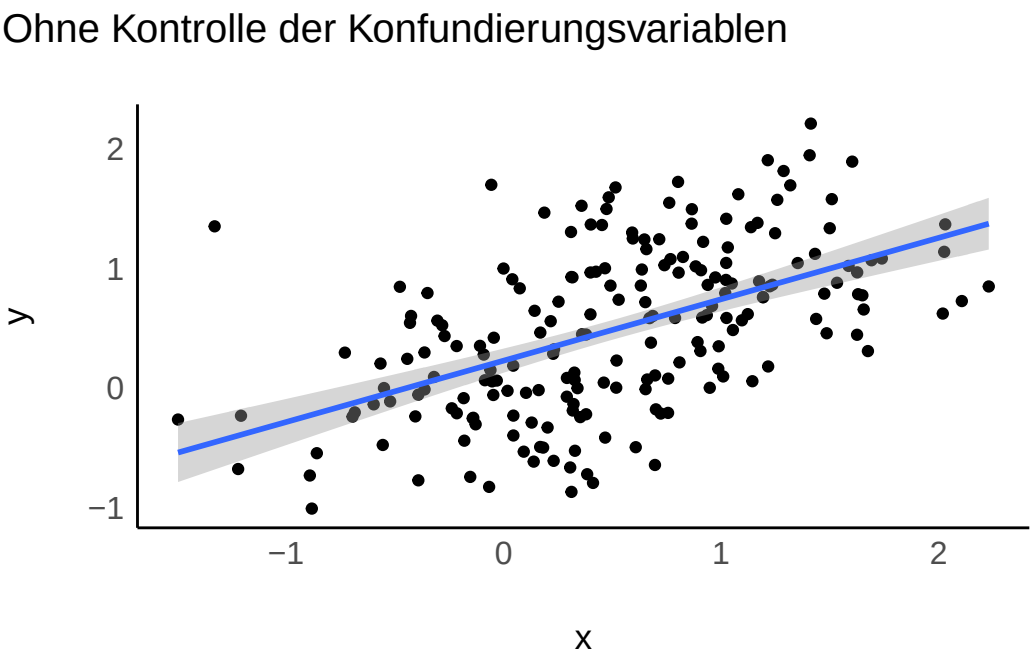


Durch das Kontrollieren ("adjustieren"), sind `bedrooms` und `price` nicht mehr korreliert, nicht mehr `d_connected`, sondern jetzt `d_separeted`.

1.5.3 Konfundierer kontrollieren

1. Ohne Kontrollieren der Konfundierungsvariablen

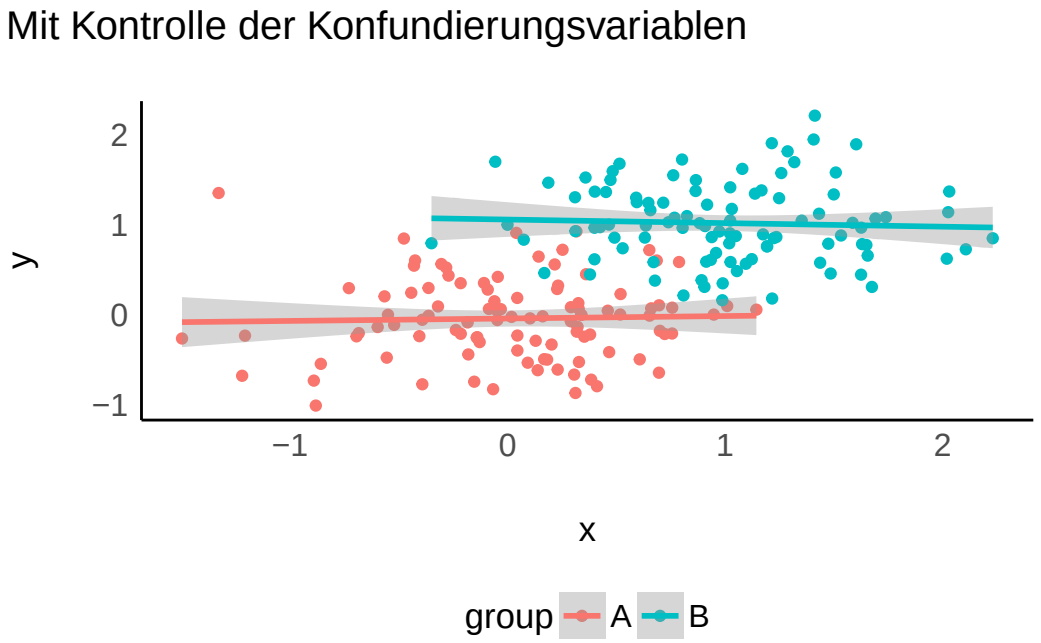
Regressionsmodell: `y ~ x`



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

M2. it Kontrollieren der Konfundierungsvariablen

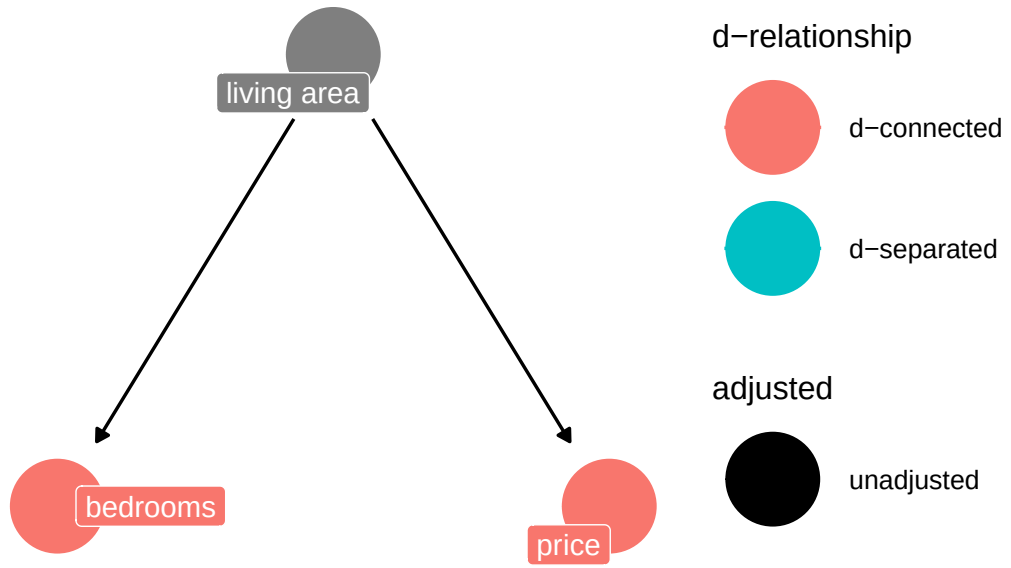
Regressionsmodell: `y ~ x + group`



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

Quellcode

1.5.4 `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 Kausalmodell `km1` vereinbar.

1.5.5 Kausalmodell 2, `km2`

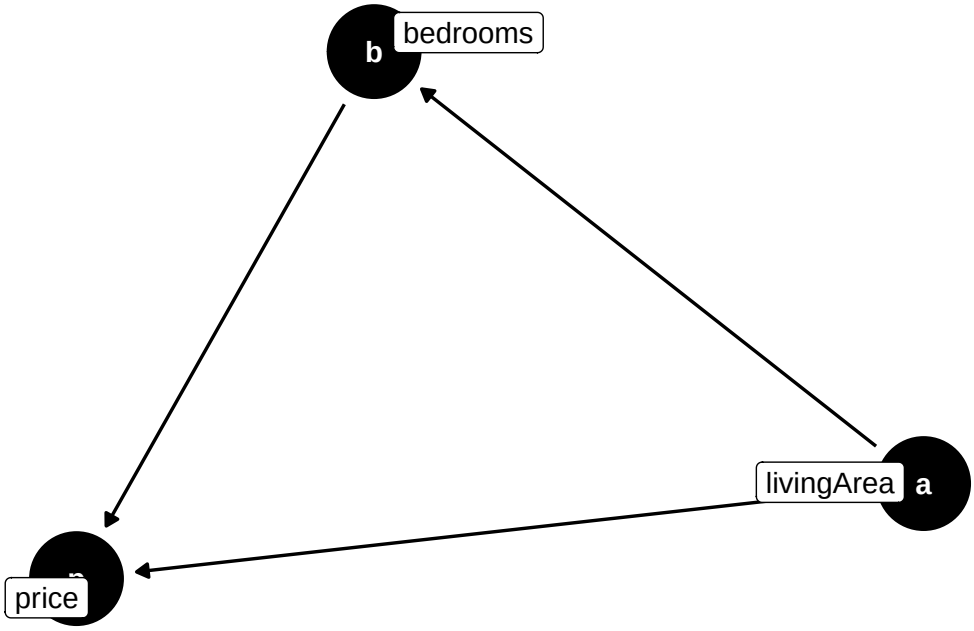
Unser Modell `m2` sagt uns, dass beide Prädiktoren jeweils einen eigenen Beitrag zur Erklärung der AV haben.

Daher könnte das folgende Kausalmodell, `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*.



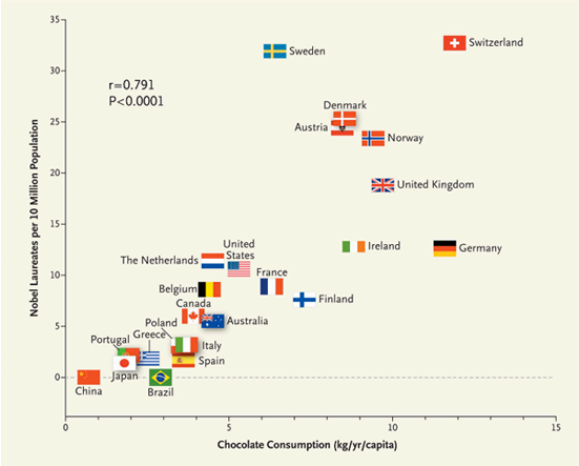
1.5.6 Schoki macht Nobelpreis! (?)

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

> **!** Wichtig
>
> Korrelation ungleich Kausation!

1.5.7 Kausalmodell für die Schoki-Studie

Der "Schoki-DAG" in Abbildung 1.5 zeigt den DAG für das Schokoloaden-Nobelpreis-Modell.



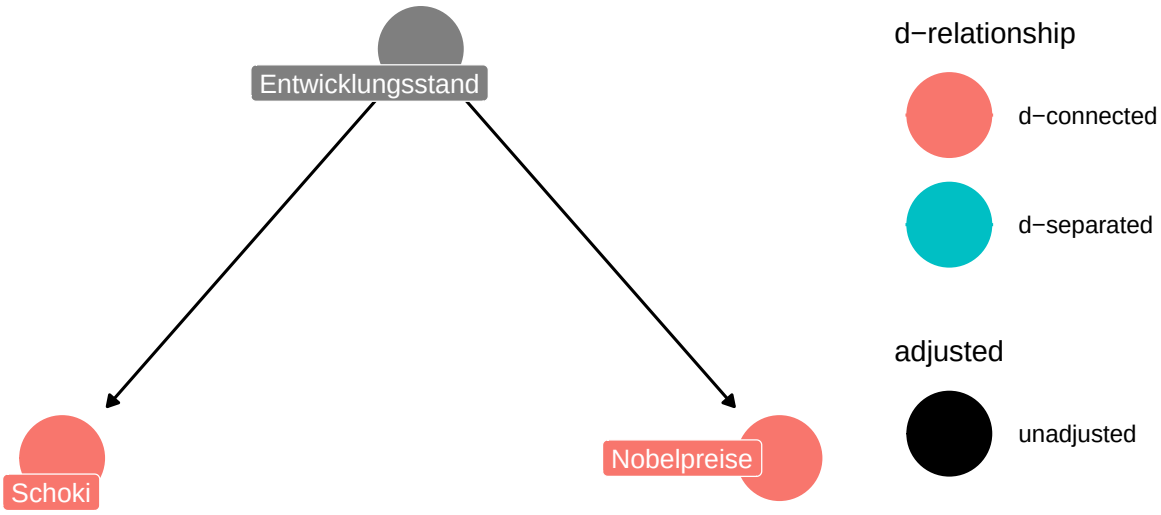


Abbildung 1.5: Macht Schokolade Nobelpreise?

1.5.8 Dons Kausalmodell, `km3`

So sieht Dons Kausalmodell aus, s. Abbildung 1.6.

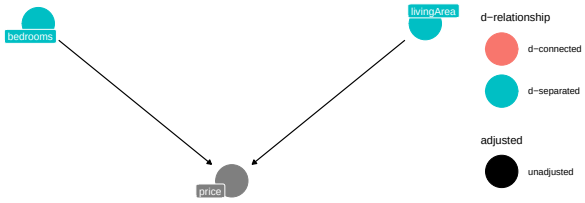


Abbildung 1.6: Dons Kausalmodell

> Ich glaube aber an mein Kausalmodell. Mein Kausalmodell ist das größte! Alle anderen Kausalmodelle sind ein Disaster!"

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

Rechne doch selber, die Korrelation aus, Don:

```
d %>%
  summarise(cor(bedrooms, livingArea))
## # A tibble: 1 x 1
##   `cor(bedrooms, livingArea)`
##                         <dbl>
## 1                       0.656
```

1.5.9 Unabhängigkeiten laut `km1`

`b`: bedrooms, `p`: price, `a` area (living area), s. Abbildung 1.7.

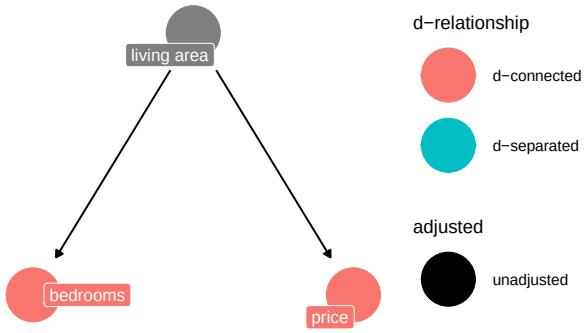


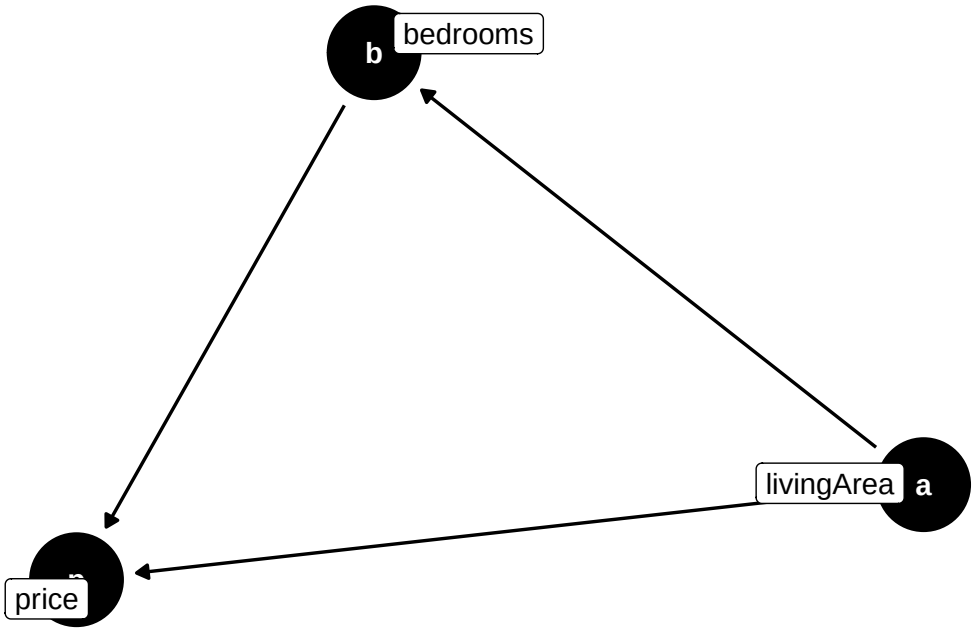
Abbildung 1.7: **?(caption)**

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

Passt nicht zu den Daten/zum Modell!

1.5.10 Unabhängigkeiten laut `km2`

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

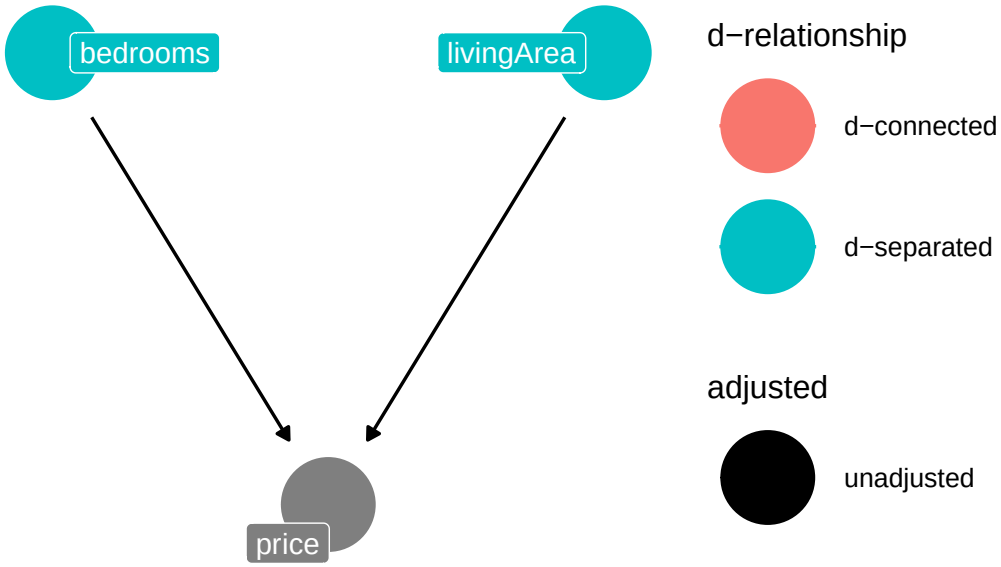


keine Unabhängigkeiten

Passt zu den Daten/zum Modell

1.5.11 Unabhängigkeiten laut `km3`

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



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

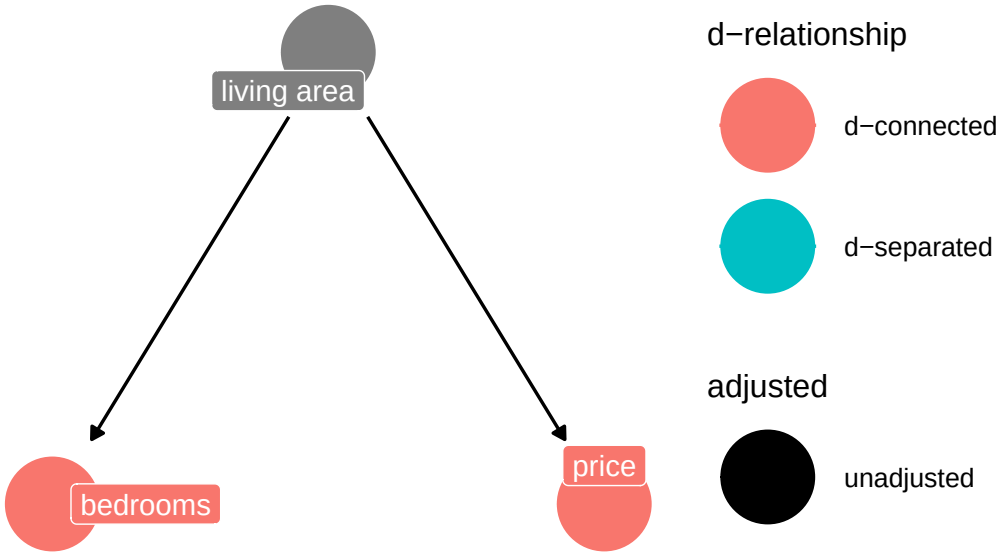
Passt nicht zu den Daten/zum Modell!

1.6 DAGs: Directed Acyclic Graphs

Was sind DAGs?

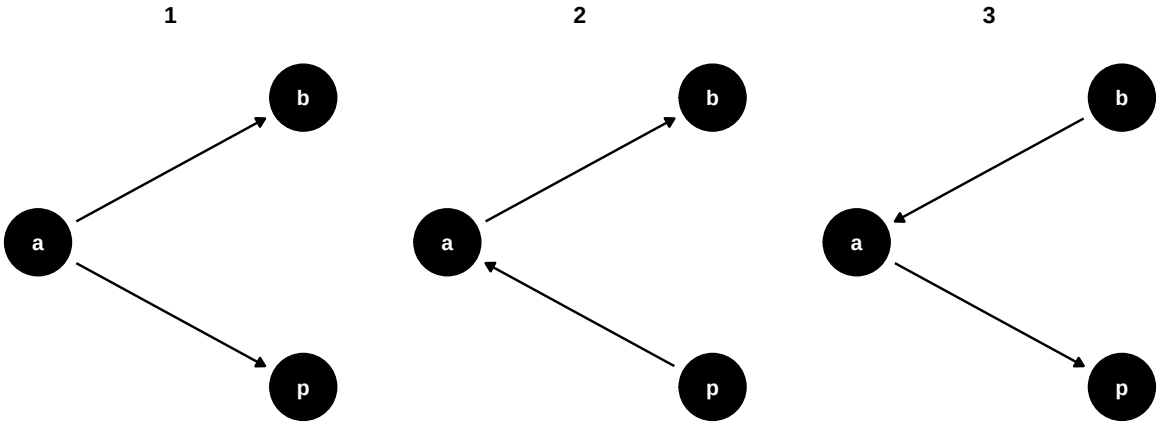
- DAGs sind eine bestimmte Art von Graphen zur Analyse von Kausalstrukturen.
- Ein *Graph* besteht aus Knoten (Variablen) und Kanten (Linien), die die Knoten verbinden.
- DAGs sind *gerichtet*; die Pfeile zeigen immer in eine Richtung (und zwar von Ursache zu Wirkung).
- DAGs sind *azyklisch*; die Wirkung 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.

1.6.1 DAG von `km1`



1.6.2 Leider passen potenziell viele DAGs zu einer Datenlage

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

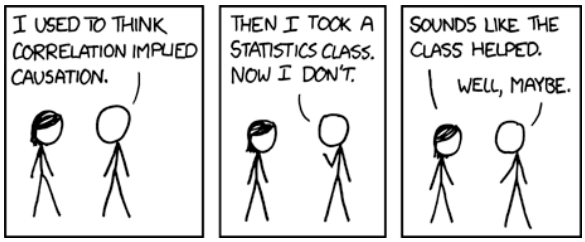


1.6.3 Was ist eigentlich eine Ursache?

Etwas verursachen kann man auch (ziemlich hochtrabend) als "Kausation" verwenden.

> **Hinweis**
>
> Weiß man, was die Wirkung W einer Handlung H (Intervention) ist, so hat man H als Ursache von W erkannt.

McElreath (2020)



Quelle und Erklärung

1.6.4 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.

1.7 Kollision

1.7.1 Kein Zusammenhang von Intelligenz und Schönheit (?)

Gott ist gerecht (?)

Zumindest findet sich in folgenden Daten kein Zusammenhang von Intelligenz (`talent`) und Schönheit (`looks`), wie Abbildung 1.8 illustriert. Für geringe Intelligenzwerte gibt es eine breites Spektrum von Schönheitswerten und für hohe Intelligenzwerte sieht es genauso aus.

Gott ist gerecht (?)

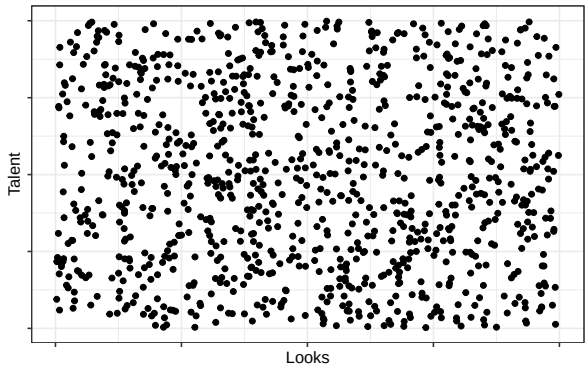


Abbildung 1.8: Kein Zusammenhang von Intelligenz und Schönheit in den Daten

1.7.2 Aber Ihre Dates sind entweder schlau oder schön

Seltsamerweise beobachten Sie, dass die Menschen, die Sie daten (Ihre Dates), entweder schön sind oder schlau - aber seltens beides gleichzeitig (schade), s. Abbildung 1.9.

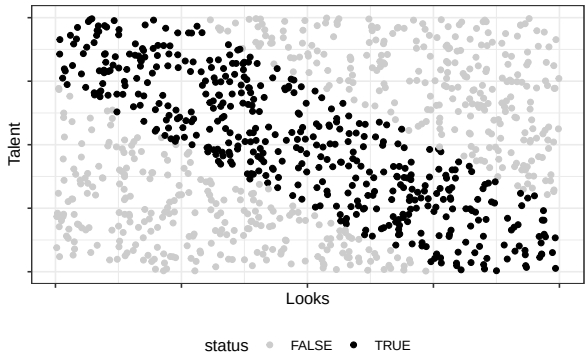


Abbildung 1.9: Ihre Datingpartner sind komischerweise entweder schlau oder schön (aber nicht beides), zumindest in der Tendenz.

Wie kann das sein?

1.8 DAG zur Rettung

Der DAG in Abbildung 1.10 bietet eine rettende Erklärung.

In ähnlicher Weise, s. Abbildung 1.11.

1.8.1 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).

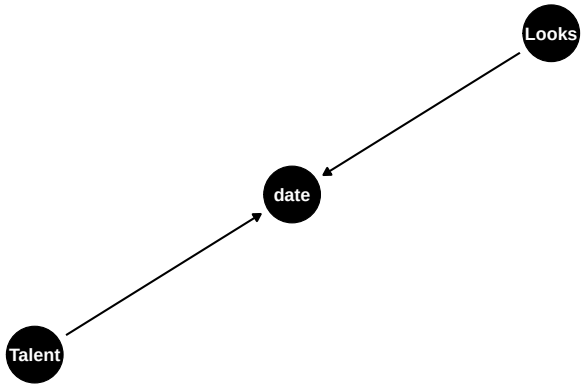


Abbildung 1.10: Date als gemeinsame Wirkung von Schönheit und Intelligenz. Stratifiziert man die gemeinsame Wirkung (dates), so kommt es zu einer Scheinkorrelation zwischen Schönheit und Intelligenz.

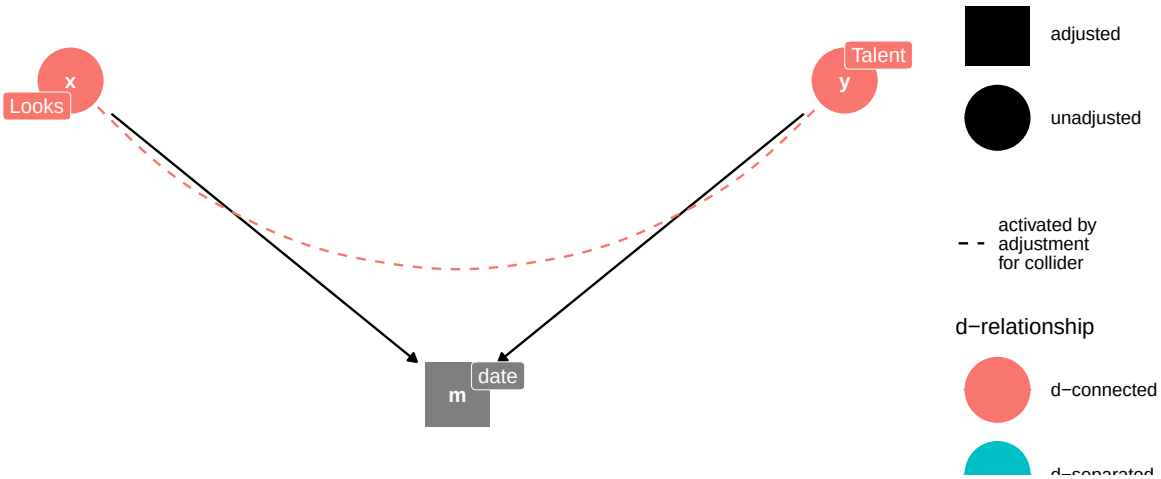


Abbildung 1.11: Durch Kontrolle der gemeinsamen Wirkung entsteht eine Scheinkorrelation zwischen den Ursachen

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, s. Abbildung 1.10, vgl. Rohrer (2018).

> **! Wichtig**
>
> 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.

1.8.2 Einfaches Beispiel zur Kollision

In der Zeitung *Glitzer* werden nur folgende Menschen gezeigt:

- Schöne Menschen
- Reiche Menschen

ehen wir davon aus, dass Schönheit und Reichtum unabhängig voneinander sind.

Wenn ich Ihnen sage, dass Don nicht schön ist, aber in der Glitzer häufig auftaucht, was lernen wir dann über seine finanzielle Situation?[1]

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

1.8.3 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: $X \not\perp\!\!\!\perp Y \mid Z$.

1.8.4 Durch Kontrollieren entsteht eine Verzerrung bei der Kollision

Abbildung 1.10 zeigt: Durch Kontrollieren entsteht eine Kollision, eine Scheinkorrelation zwischen den Ursacxhen.

Kontrollieren kann z.B. bedeuten:

- *Stratifizieren*: Aufteilen von `date` in zwei Gruppen und dann Analyse des Zusammenhangs von `talent` und `looks` in jeder Teilgruppe von `date`

[1] Don muss reich sein.

- *Kontrollieren mit Regressioin*: Durch Aufnahme von `date` als Prädiktor in eine Regression zusätzlich zu `looks` mit `talent` als Prädikotr

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", die Korrelation kann "fließen" - was sie hier nicht soll, denn es handelt sich um Scheinkorrelation.

Das Kontrollieren von `date` geht zumeist durch Bilden einer Auswahl einer Teilgruppe von sich.

1.8.5 IQ, Fleiss und Eignung fürs Studium

Sagen wir, über die *Eignung* für ein Studium würden nur (die individuellen Ausprägungen) von Intelligenz (IQ) und Fleiss entscheiden, s. den DAG in Abbildung 1.12.

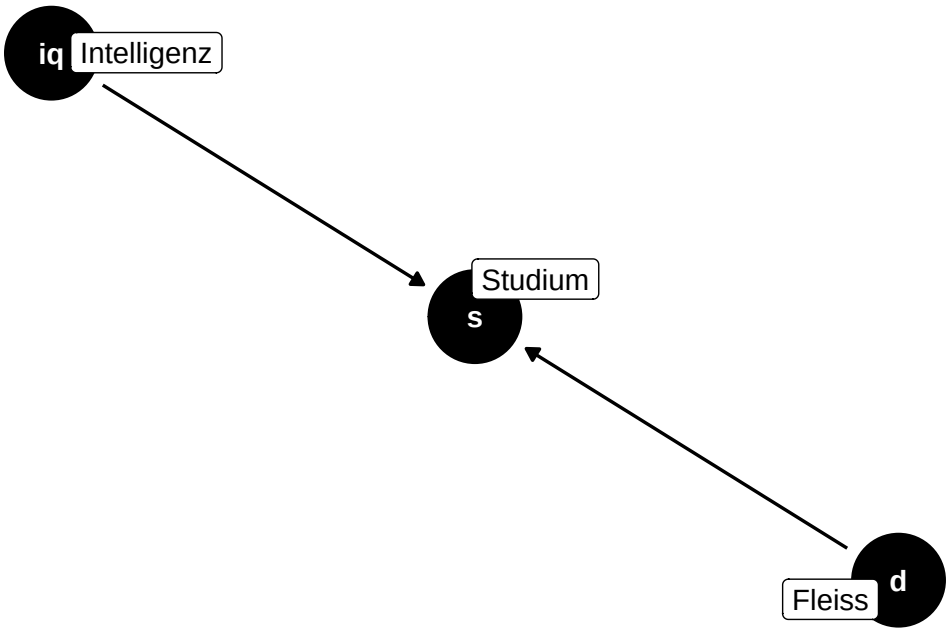


Abbildung 1.12: Kollisionsstruktur im Dag zur Studiumseignung

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
```

```r
d_eignung <-
tibble(
  iq = rnorm(N),  # normalverteilt mit MW=0, sd=1
  fleiss = rnorm(N),
  glueck = rnorm(N, mean = 0, sd = .1),
  eignung = 1/2 * iq + 1/2 * fleiss + glueck,
  # nur wer geeignet ist, studiert (in unserem Modell):
  studium = ifelse(eignung > 0, 1, 0)
  )
```

Laut unserem Modell setzt sich Eignung zur Hälfte aus Intelligenz und zur Hälfte aus Fleiss zusammen, plus etwas Glück.

1.8.6 Schlagzeile "Schlauheit macht Studentis faul!"

Eine Studie untersucht den Zusammenhang von Intelligenz (iq) und Fleiß (f) bei Studentis (s).

Ergebnis: Ein *negativer* Zusammenhang!?

Berechnen wir das "Eignungsmodell", aber nur mit Studis (`studium == 1`):

```r
m_eignung <-
  stan_glm(iq ~ fleiss, data = d_eignung %>%  filter(studium == 1), refresh = 0)

hdi(m_eignung)
## Highest Density Interval
##
## Parameter   |        95% HDI
## ----------------------------
## (Intercept) | [ 0.70,  0.86]
## fleiss      | [-0.53, -0.36]
```

Abbildung 1.13 zeigt das Modell und die Daten.

IQ ist *nicht* unabhängig von Fleiß in unseren Daten, sondern abhängig.

Nichtwissenschaftliche Berichte, etwa in einigen Medien, greifen gerne Befunde über Zusammenhänge auf und interpretieren die Zusammenhänge - oft vorschnell - als kausal.[2]

[2] Ehrlicherweise muss man zugeben, dass auch wissenschaftliche Berichte Daten über Zusammenhänge gerne kausal interpretieren, oft vorschnell.

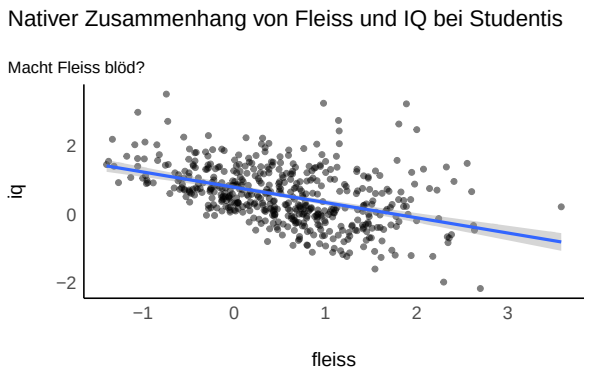


Abbildung 1.13: Der Zusammenhang von Fleiss und IQ

1.8.7 Kollisionsverzerrung nur bei Stratifizierung

Nur durch das Stratifizieren (Aufteilen in Subgruppen, Kontrollieren, Adjustieren) tritt die Scheinkorrelation auf, s. Abbildung 1.14.

> **Hinweis**
>
> *Ohne* Stratifizierung tritt *keine* Scheinkorrelation auf. *Mit* Stratifizierung tritt Scheinkorrelation auf.

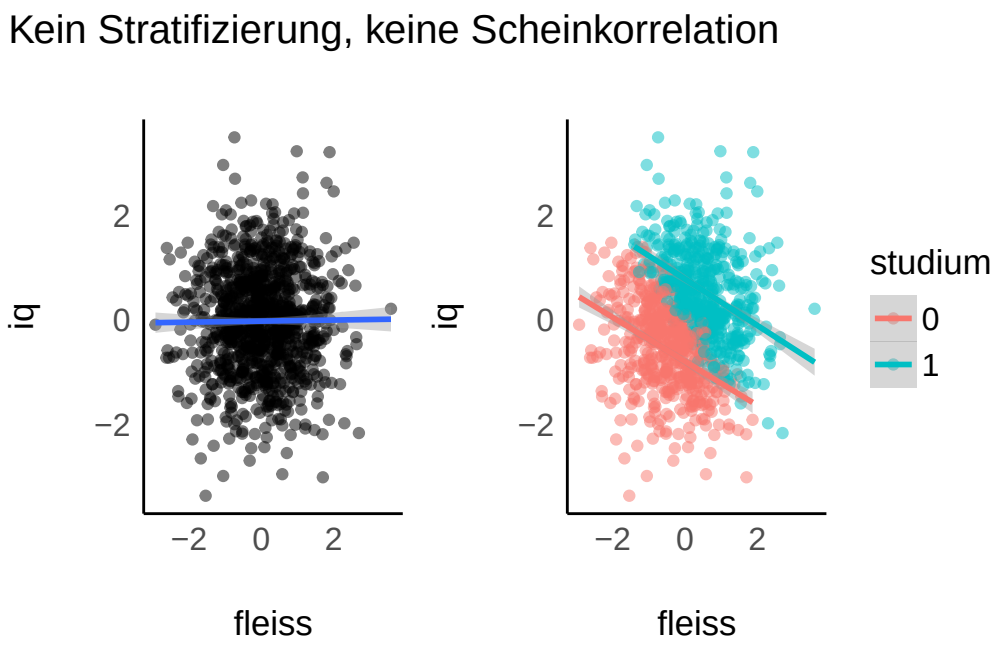


Abbildung 1.14: Stratifizierung und Scheinkorrelation

Wildes Kontrollieren einer Variablen - Aufnehmen in die Regression - kann genausog ut schaden wie nützen.

Nur Kenntnis des DAGs verrät die richtige Entscheidung: ob man eine Variable kontrolliert oder nicht.

> **Hinweis**
>
> Nimmt man eine Variable als zweiten Prädiktor auf, so "kontrolliert" man diese Variable. Das Regressionsgewicht des ersten Prädiktors wird "bereinigt" um den Einfluss des zweiten Prädiktors; insofern ist der zweite Prädiktor dann "kontrolliert".

1.8.8 Einfluss von Großeltern und Eltern auf Kinder

Wir wollen hier 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`: $G \to E \to K$
- direkter Effekt von `E` auf `K`: $E \to K$
- direkter Effekt von `G` auf `K`: $G \to K$

Wir sind v.a. interessiert an $G \to K$, dem *direkten kausalen* Effekt von Großeltern auf ihre Enkel, s. Abbildung 1.15, Kurz (2021).

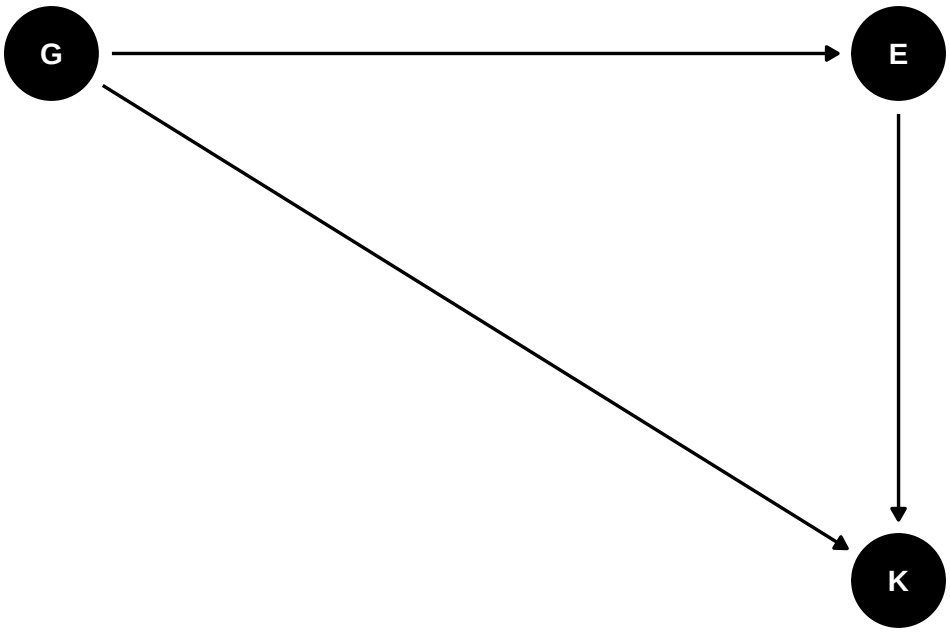


Abbildung 1.15: **?(caption)**

Aber was ist, wenn wir vielleicht eine *u*nbekannte Variable übersehen haben? (S. nächster Abschnitt)

1.9 Vertiefung

VERTIEFUNG

1.9.1 Der Gespenster-DAG

Es gibt "unheilbare" DAGs, nennen wir sie "Gespenster-DAGs", in denen es nicht möglich ist, einen (unverzerrten) Kausaleffekt zu bestimmen, s. Abbildung 1.16. Letztlich sagt uns der DAG bzw. unsere Analyse zum DAG: "Deine Theorie ist nicht gut, zurück an den Schreibtisch und denk noch mal gut nach. Oder sammele mehr Daten."

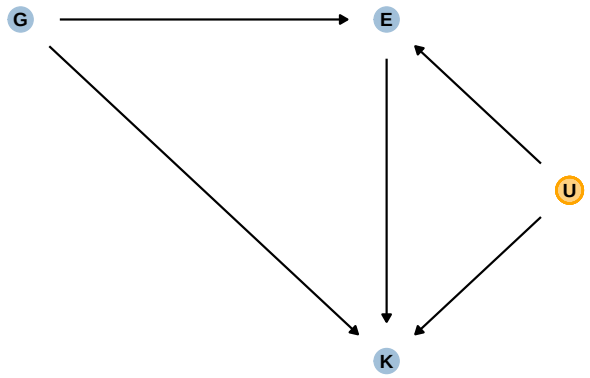


Abbildung 1.16: Der Gespenster-DAG: Eine Identifikation der Kausaleffekt ist nicht (vollständig) möglich.

- `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 in diesem Fall chancenlos. Wir müssen diesen DAG verloren geben, McElreath (2020), S. 180.

1.10 Die Hintertür schließen

1.10.1 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`

Das Kausalmodell ist in Abbildung 1.17 dargestellt.

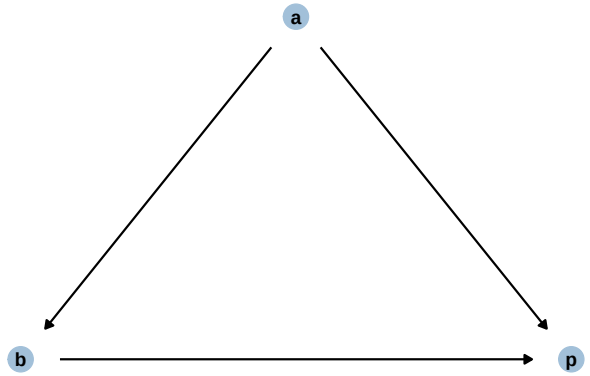


Abbildung 1.17: Der Preis wird sowohl von der Zimmerzahl als auch der Wohnfläche beeinflusst

Im Regressionsmodell `p ~ 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 nur den zweiten Pfad und wir würden `b` ändern, so würde sich `p` nicht ändern.

1.10.2 Gute Experimente zeigen den echten kausalen Effekt

Abbildung 1.18 zeigt eine erfreuliche Situation: Die "Hintertür" zu unserer UV (Zimmerzahl) ist geschlossen!

Ist die Hintertür geschlossen - führen also keine Pfeile in unserer UV - so kann eine Konfundierung ausgeschlossen werden.

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

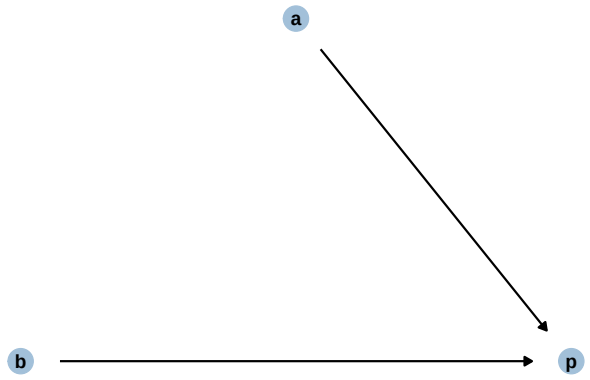


Abbildung 1.18: Unverzerrte Schätzung des kausalen Effekts unserer UV (Zimmerzahl). Das Regressionsgewicht ist hier der unverzerrte Kausaleffekt. Es spielt keine Rolle, ob der andere Prädiktor im Modell enthalten ist. Da die beiden Prädiktoren unkorreliert sind, hat die Aufnahme des einen Prädiktors keinen Einfluss auf das Regressionsgewicht des anderen.

Eine berühmte Lösung, den kausalen Pfad zu isolieren, ist ein (randomsiertes, 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").

1.10.3 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…".

$b \perp\!\!\!\perp p \mid a$

1.10.4 Die vier Atome der Kausalanalyse

Abbildung 1.19 stellt die vier "Atome" der Kausalinferenz dar. Mehr gibt es nicht! Kennen Sie diese vier Grundbausteine, so können Sie jedes beliebige Kausalsystem (DAG) entschlüsseln.

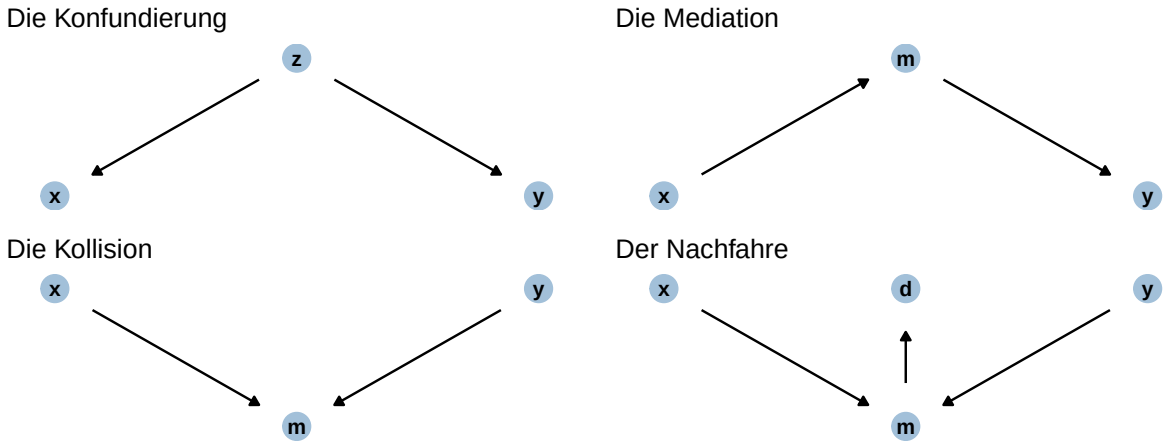


Abbildung 1.19: Die vier Atome der Kausalinferenz

1.10.5 Mediation

Die *Mediation* (Wirkkette, Rohr, Kette, chain) beschreibt Pfade, in der die Kanten gleiche Wirkrichtung haben: $x \to m \to y$. Anders gesagt: Eine Mediation, auch "Kette" oder "Wirkkette" genannt, ist eine Kausalabfolge der Art $x \to m \to y$, s. Abbildung 1.20. Die Variable in der Mitte m der Kette wird auch *Mediator* genannt, weil sei die Wirkung von X auf Y "vermittelt" oder übeträgt. Die Erforschung von Mediation spielt eine recht wichtige Rolle in einigen Wissenschaften, wie der Psychologie.

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).

1.11 Der Nachfahre

Ein *Nachfahre* (descendent) ist eine Variable die von einer anderen Variable beeinflusst wird, s. fig-dag-nachfahre. 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.

Die Mediation

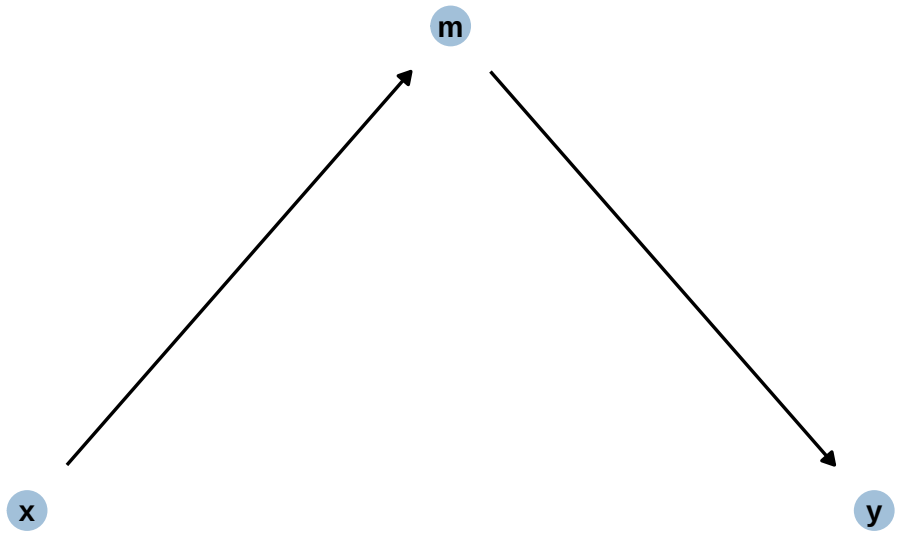


Abbildung 1.20: Das Kausalmodell der Mediation.

Der Nachfahre

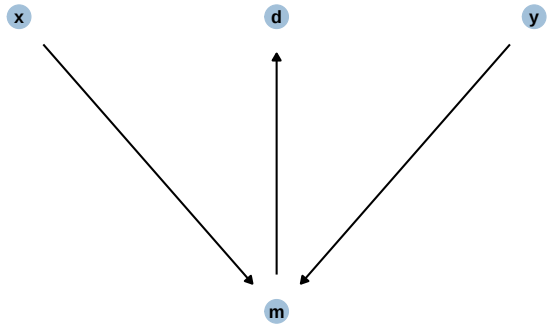


Abbildung 1.21: Ein Nachfahre verhält sich ähnlich wie sein Vorfahre...

1.11.1 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 mann kontrollieren muss, um den Pfad zu schließen (falls möglich).

1.12 Schließen Sie die Hintertür (wenn möglich)!, `bsp1`

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

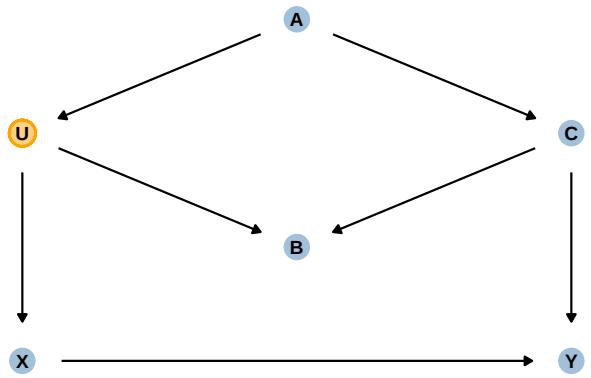


Abbildung 1.22: Puh, ein schon recht komplizierter DAG

Es gibt zwei Hintertürpfade in Abbildung 1.22:

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.

1.12.1 Schließen Sie die Hintertür (wenn möglich)!, `bsp2`

S. DAG in Abbildung 1.23: UV: W, AV: D

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

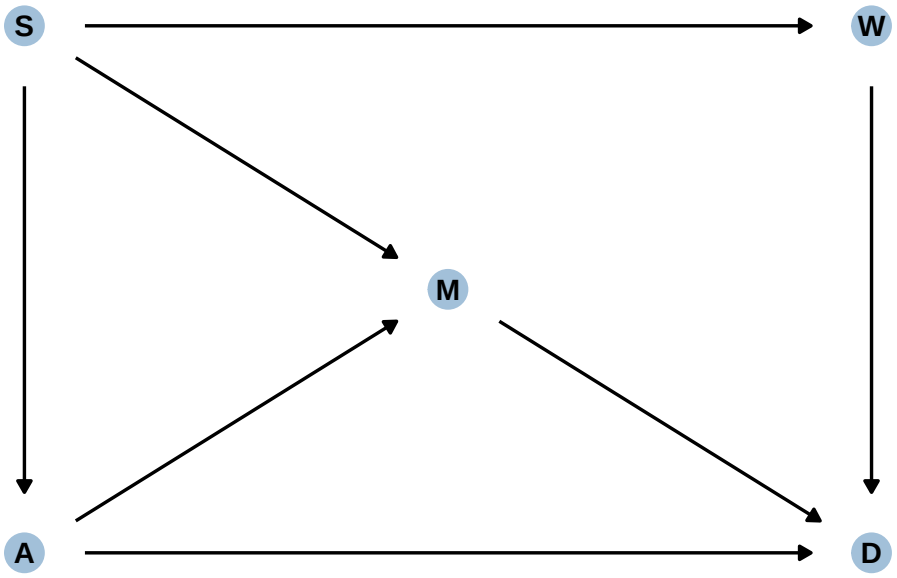


Abbildung 1.23: Welche Variablen muss man kontrollieren, um den Effekt von W auf D zu bestimmen?

- entweder A und M
- oder S

Mehr Infos

Details finden sich bei McElreath (2020) oder Kurz (2021), ,S. 188.

1.12.2 Implizierte bedingte Unabhängigkeiten von `bsp2`

Ein Graph ohne `Us` ist eine starke - oft zu starke (unrealistisch optimistische) - Annahme. Auch wenn die Daten nicht sagen können, welcher DAG der richtige ist, können wir zumindest lernen, welcher DAG falsch ist. Die vom Modell implizierten bedingten Unabhängigkeiten geben uns Möglichkeiten, zu prüfen, ob wir einen DAG verwerfen (ausschließen) können. 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
```

1.12.3 Fazit

Wie (und sogar ob) Sie statistische Ergebnisse (z.B. eines Regressionsmodells) interpretieren können, hängt von der *epistemologischen Zielrichting* 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. Kenntnis der "kausalen Atome" ist Voraussetzung zur Ableitung von Kausalschlüsse in Beobachtungsstudien.

Kurz, A. Solomon. 2021. *Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition*. https://bookdown.org/content/4857/.

McElreath, Richard. 2020. *Statistical rethinking: a Bayesian course with examples in R and Stan*. 2. Aufl. CRC texts in statistical science. Boca Raton: Taylor; Francis, CRC Press.

Messerli, Franz H. 2012. „Chocolate Consumption, Cognitive Function, and Nobel Laureates". *New England Journal of Medicine* 367 (16): 1562–64. https://doi.org/10.1056/NEJMon1211064.

Pearl, Judea, Madelyn Glymour, und Nicholas P. Jewell. 2016. *Causal inference in statistics: a primer*. Chichester, West Sussex: Wiley.

Rohrer, Julia M. 2018. „Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data". *Advances in Methods and Practices in Psychological Science* 1 (1): 27–42. https://doi.org/10.1177/2515245917745629.