11. Vorlesung Statistik II
Kausale Inferenz mit DAGs und Modellierung wissenschaftlicher Fragestellungen

We are happy to share our materials openly:
The content of these Open Educational Resources by Lehrstuhl für Psychologische Methodenlehre und Diagnostik, Ludwig-Maximilians-Universität München is licensed under CC BY-SA 4.0. The CC Attribution-ShareAlike 4.0 International license means that you can reuse or transform the content of our materials for any purpose as long as you cite our original materials and share your derivatives under the same license.

Lernziele

- Sie kennen die grundlegende Definition von Kausalität in den empirischen Wissenschaften und können einfache kausale Theorien als DAG auszudrücken.
- Sie können für die drei elementaren DAGs (Confounder, Mediator, Collider) entscheiden, ob eine Drittvariable als Kontrollvariable in ein Regressionsmodell aufgenommen werden muss, um einen kausalen Effekt zu schätzen.
- Sie kennen verschiedene Möglichkeiten für eine Variable zu „kontrollieren“ und wissen wie man Experimente als DAG darstellt.
- Sie können alle statistischen Verfahren aus Statistik 1 und 2 als Regressionsmodell darstellen.
- Sie können für verschiedene wissenschaftliche Fragestellungen das passende Regressionsmodell aufstellen und das passende inferenzstatistische Verfahren (Konfidenzintervall oder Hypothesentest) auswählen und anwenden.
- Sie können Regressionsmodelle für verschiedene wissenschaftliche Fragestellungen passend erweitern (z.B. Hinzufügen von weiteren Prädiktoren und Interaktionen).

Kausale Inferenz und Directed Acyclic Graphs (DAG)

Kausalität und Statistik

- In der psychologischen Forschung ist man traditionell sehr vorsichtig damit, Studienergebnisse kausal zu interpretieren:
  - Alle Studierenden lernen in Statistik I „Korrelation ist nicht Kausalität“
  - In Statistik II haben wir Regressionsanalysen immer nur assoziativ interpretiert: „Für eine Person mit einem um 1 Punkt höheren IQ erwarten wir im Mittel ein um β Punkte besseres Abschneiden im Assessmentcenter.“
- *These:* Die meisten psychologischen Fragestellungen sind aber doch an kausalen Aussagen interessiert!

Kausalität und Statistik

- Beispiel aus VL 7: Existiert ein linearer Zusammenhang, also eine Abhängigkeit, zwischen der negativen Selbstbewertung und der Depressionsschwere von Personen?
- Wäre es in diesem Beispiel nicht spannender, zu untersuchen, ob die Stärke der negativen Selbstbewertung die Depressionsschwere einer Person **kausal** beeinflusst; anstatt lediglich den korrelativen Zusammenhang zwischen den beiden Variablen zu betrachten? Daraus ließen sich zumindest für die psychotherapeutische Praxis bedeutsamere Aussagen ableiten.

Kausalität und Statistik II

- Um mit statistischen Modellen wissenschaftliche Erkenntnisse zu generieren, sind zusätzlich zu empirischen Daten theoriebasierte (kausale) Modelle nötig.
- Die kausalen Ursachen, die zur Beobachtung von bestimmten Daten geführt haben, können nicht (alleine) aus den Daten selbst gelernt werden: „No causes in, no causes out."
- „Kausalität steckt nicht in den Daten"

- „Ladder of causation" (nach Judea Pearl): Kausale Inferenz beinhaltet…
  - „Association": (beobachtete) Zusammenhänge zwischen Variablen
  - „Intervention": die Vorhersage der Folge von Interventionen („Was passiert, falls ich das mache?")
  - „Counterfactuals": die Simulation von fehlenden Beobachtungen („Was wäre passiert, wenn ich etwas anderes gemacht hätte?")

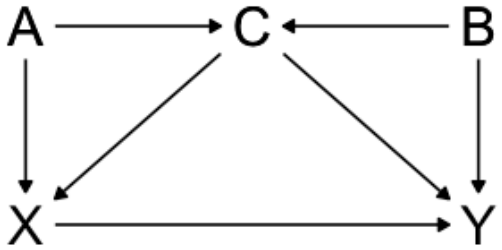
Was ist Kausalität?

- Was ist gemeint mit der Frage, ob eine Variable X einen kausalen Einfluss auf eine andere Variable Y hat? Z.B.:
  - Hat die Intelligenz von Bewerber*innen einen kausalen Einfluss auf den Erfolg im Bewerbungsgespräch?
  - Hat die Gewissenhaftigkeit von Patient*innen einen kausalen Einfluss auf den Therapieerfolg?

- Die (für empirische Wissenschaften) derzeit wichtigste Definition von Kausalität bezieht sich immer auf theoretische Interventionen:
  - Wenn es einen kausalen Einfluss von X auf Y gibt, sollte eine Manipulation von X zu einer Veränderung in Y führen (aber nicht umgekehrt)
  - Es ist irrelevant, ob eine solche Manipulation aus technischen oder ethischen Gründen praktisch durchführbar ist: „Angenommen, wir könnten Gott spielen und X beeinflussen, was passiert dann mit Y?“

Directed Acyclic Graphs (DAGs)

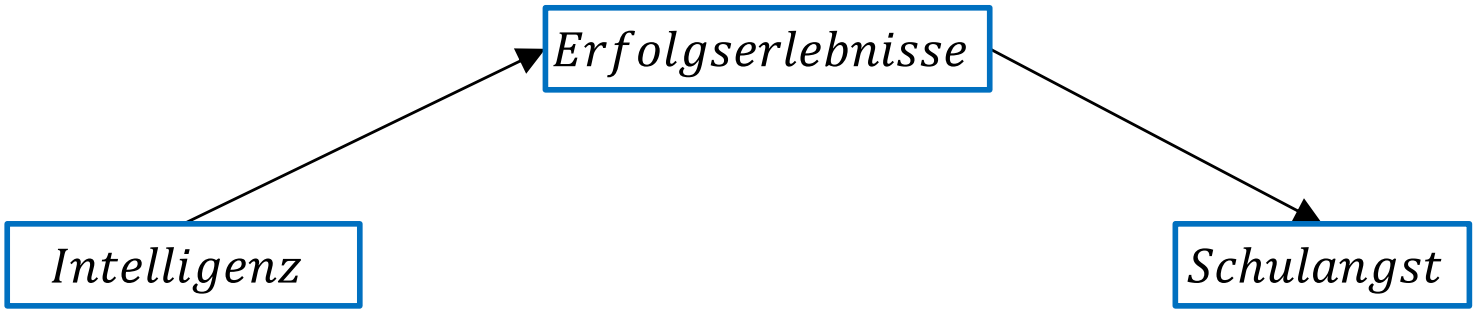
- Ein zunehmend beliebtes Framework, kausale Fragestellungen in der empirischen Wissenschaft zu formalisieren, sind die sogenannten „Directed Acyclic Graphs“ (DAGs).



- DAG als graphische Darstellung einer (abstrakten) Kausaltheorie:
  - Ein Pfeil von einer Variable X auf eine Variable Y bedeutet: X hat einen (direkten) kausalen Effekt auf Y (und nicht umgekehrt)
  - Jeder Pfeil, der nicht eingezeichnet ist, formalisiert die Annahme, dass der entsprechende (direkte) kausale Effekt NICHT vorliegt
- Es gibt ein mathematisches Framework (Do-Calculus), mit dem man allein mithilfe der in einem DAG enthaltenen Kausalannahmen formal beweisen kann, **ob bestimmte Arten von kausalen Aussagen prinzipiell untersuchbar sind, falls der angenommene DAG korrekt ist** (d.h. die darin formalisierten Kausalannahmen zutreffend sind). Dies werden wir nicht im Detail besprechen.

Directed Acyclic Graphs (DAGs)

- DAGs können genutzt werden, um theoretische Kausalannahmen graphisch zu veranschaulichen und potentielle Störfaktoren („Confounder“) aufzudecken. Sie helfen uns zu entscheiden, ob bzw. welche Drittvariablen in das Modell aufgenommen werden sollten und welche nicht.
- DAGs veranschaulichen komplexe kausale Zusammenhänge durch gerichtete Pfeile. Die Richtung der Pfeile gibt dabei die kausale Wirkrichtung an.
- Ein Pfeil kann dabei sowohl gleichgerichtete als auch gegengerichtete Zusammenhänge darstellen:



- Aus DAGs können eindeutige Regeln abgeleitet werden, ob die Aufnahme einer Drittvariablen ins Modell sinnvoll ist oder nicht, je nach Anordnung der Variablen und Pfeile (also je nach theoretischen Annahmen und Überlegungen).

Daten generierende Modelle

- DAGs treffen (zunächst) keine Annahmen über …
  - die Art der enthaltenen Variablen, z.B.:
    - Diskret vs. Kontinuierlich
    - Verteilungsannahme der Variablen
  - die Art des Zusammenhangs zwischen den Variablen, z.B.:
    - Lineare oder nicht lineare Zusammenhänge
    - Nur Haupteffekte oder (eventuell komplexe) Interaktionen

Wiederholung: Welche Prädiktoren sollten also ins Modell und welche nicht?

- In der Literatur wird bei der Aufnahme von Drittvariablen („Kovariate“) als weitere Prädiktoren ins Regressionsmodell davon gesprochen, dass der Einfluss dieser Drittvariablen **statistisch kontrolliert** wird.

  „So, if you feel the urge to “control away” some parts of the associations of interest, it’s quite likely that your estimand is in fact a causal effect […].“ – Julia Rohrer
  https://www.the100.ci/2025/02/17/reviewer-notes-avoid-any-ambiguity-about-analysis-aims/

- Drittvariablen sind Variablen, die uns primär nicht interessieren, die aber womöglich den geschätzten Zusammenhang zwischen AV und interessierender UV beeinflussen können.

- Prinzipiell kann die Kontrolle von solchen Variablen sinnvoll sein, um kausale Fehlschlüsse zu vermeiden. Eine unüberlegte Hinzunahme von Variablen ins Modell kann aber echte Zusammenhänge auch verzerren und Fehlschlüsse erst herbeiführen.

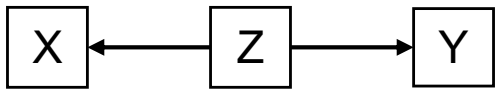
➢ Die Entscheidung, für Drittvariablen zu kontrollieren, diese also in das Modell mitaufzunehmen, sollte deshalb nie leichtfertig geschehen und immer vor dem Hintergrund theoretischer Überlegungen passieren!

➢ **Diese theoretischen Überlegungen können wir mithilfe von DAGs graphisch darstellen. Es ist dann eindeutig, welche Variablen aufgenommen („kontrolliert“) werden müssen und welche nicht (gegeben, dass der DAG stimmt).**

Elementare Bausteine von DAGs

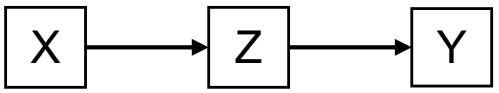
- Alle DAGs bestehen aus nur 3 charakteristischen Kausalstrukturen:

Confounder ("Fork")



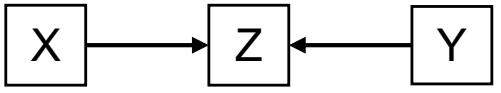
„Der Confounder Z verursacht sowohl die Variable X als auch die Variable Y"

Mediator („Pipe")



„Die Variable X verursacht die Variable Y über den Mediator Z"

Collider

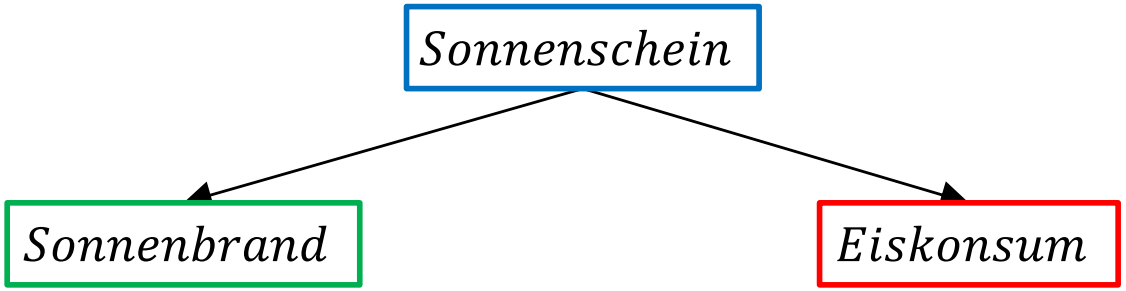


"Sowohl die Variable X als auch die Variable Y verursachen den Collider Z"

- Mit diesen elementaren Bausteinen lassen sich auch beliebig komplizierte Strukturen analysieren.
- Im Folgenden wollen wir immer anhand eines konkreten Beispiels den kausalen Effekt von X auf Y schätzen.
- Wir fragen uns dabei immer, was mit der anderen Variable Z passiert:
  Sollte für die Variable kontrolliert werden oder nicht (das heißt hier im Speziellen, sollte die Variable mit ins Modell aufgenommen werden oder nicht)?

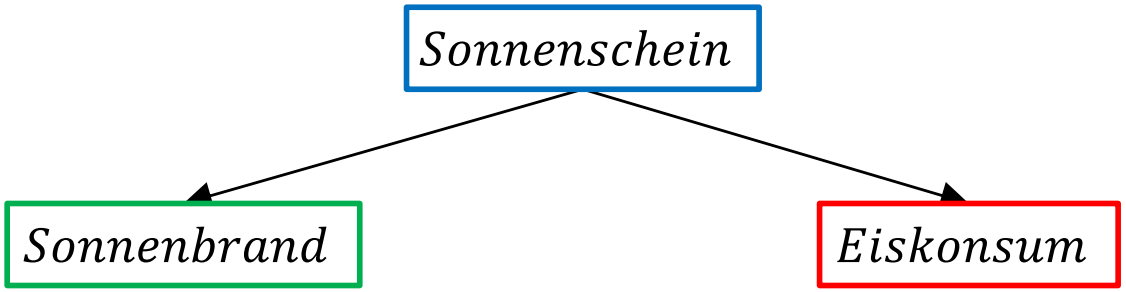
Beispiel 1: Fork – the Confounder

- Wir interessieren uns für den Einfluss von Sonnenbrand (X) einer Person auf ihren Eiskonsum (Y). Als kausales Modell stellen wir auf:



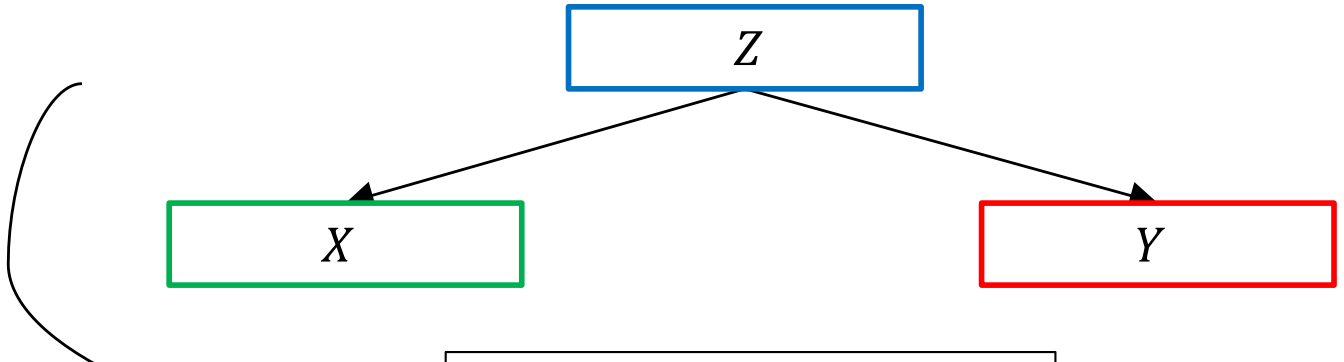
- In einer einfachen linearen Regression (ELR) mit Eiskonsum als AV und Sonnenbrand als UV wird ein Zusammenhang dieser beiden Variablen „entdeckt“, obwohl hier gar kein offensichtlicher (kausaler) Zusammenhang gilt. Im theoretischen (kausalen) Modell existiert kein Pfeil zwischen Sonnenbrand und Eiskonsum.
- Warum zeigt sich also in den Daten dieser Zusammenhang in der ELR?
- In einer multiplen LR (MLR) mit Sonnenschein (Z) als zusätzlichem Prädiktor wird dieser Effekt verschwinden und wir lernen, dass der Sonnenschein hier eigentlich entscheidend sein könnte.

Beispiel 1: Fork – the Confounder



- Anders ausgedrückt: betrachten wir den Zusammenhang zwischen Sonnenbrand und Eiskonsum „alleine“, finden wir einen (positiven) Zusammenhang. Kontrollieren wir für Sonnenschein, betrachten den Zusammenhang also für eine spezifische Ausprägung der Variable Sonnenschein, verschwindet der Zusammenhang zwischen Sonnenbrand und Eiskonsum.
- (Zur Erinnerung: Bei einer MLR ist der Zusammenhang zwischen UV1 und AV über alle Ausprägung von UV2 hinweg gleich)
- Wir könnten also sagen: Sonnenschein erklärt den Zusammenhang zwischen Sonnenbrand und Eiskonsum.
- Dies ist ein klassisches Beispiel für eine „Scheinkorrelation“ (vgl. ÜB 7 Aufg. 2).

Beispiel 1: Fork – the Confounder



Datensimulation des kausalen Modells

```
> n <- 1000
> z <- rnorm(n)
> x <- 1*z + rnorm(n)
> y <- 0*x + 1*z + rnorm(n)
```

Verzerrter Effekt

```
Call: lm(y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03823    0.03932   0.972    0.331
x            0.50619    0.02887  17.533   <2e-16 ***
```

Korrekter Effekt

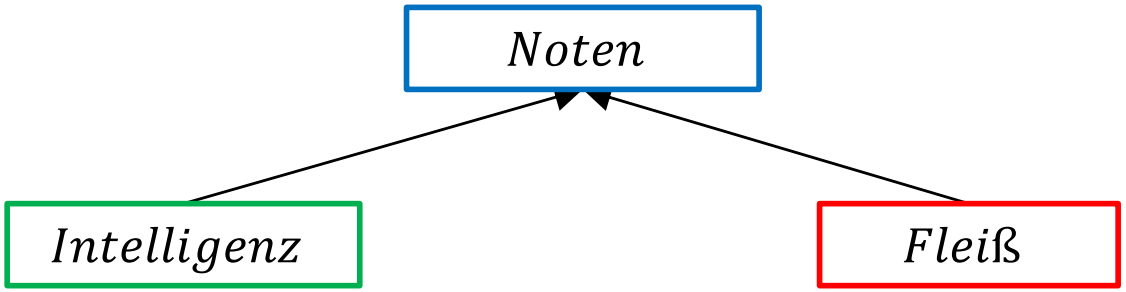
```
Call: lm(y ~ x + z)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02378    0.03174   0.749    0.454
x           -0.02706    0.03279  -0.825    0.409
z            1.01662    0.04396  23.124   <2e-16 ***
```

- Ist eine Variable ein Confounder (wie hier die Variable Z / Sonnenschein), dann sollte sie (wenn möglich) in das Modell als UV mitaufgenommen werden, um den kausalen Effekt von X auf Y zu schätzen.

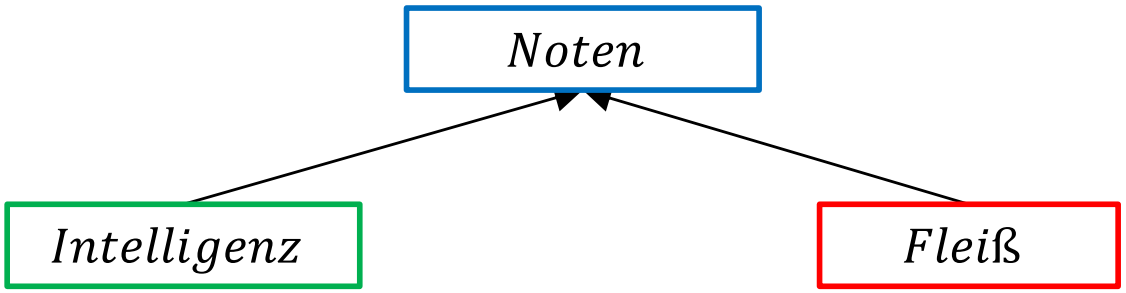
Beispiel 2: Collider

- Wir interessieren uns für den Einfluss von Intelligenz (X) einer Person auf ihren Fleiß (Y). Als kausales Modell stellen wir auf:



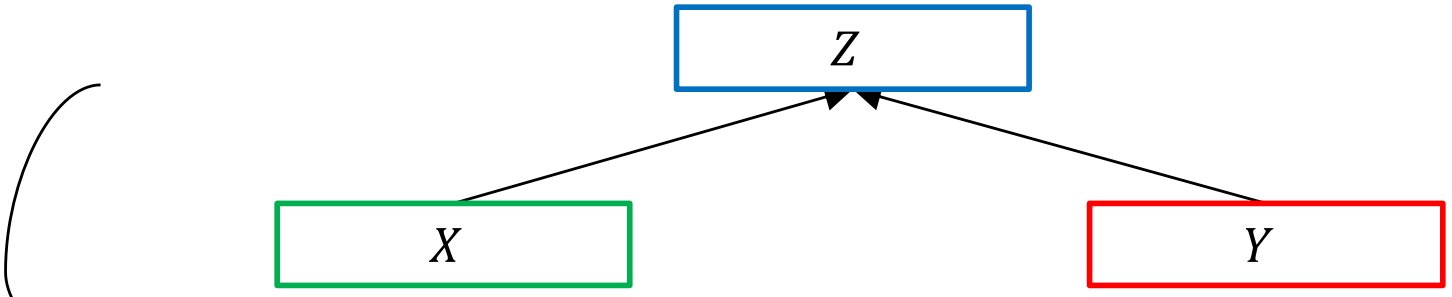
- Muss bei der Suche nach dem (in diesem theoretischen Modell nicht vorhandenen) kausalen Effekt von Intelligenz (UV) auf Fleiß (AV) die Note (Z) als zweite UV mit aufgenommen werden?
- In einer MLR mit beiden Prädiktoren (Intelligenz und Noten) wird plötzlich ein Zusammenhang zwischen Intelligenz und Fleiß auftauchen, der uns in die Irre leiten kann!
- Hier sollte also die Note nicht als Prädiktor ins Modell aufgenommen und eine ELR gerechnet werden, um den kausalen Effekt von Intelligenz auf Fleiß auf den richtigen Wert (= 0) zu schätzen.

Beispiel 2: Collider



- **Begründung**: In diesem Beispiel hat Intelligenz keinen kausalen Effekt auf Fleiß. Kontrollieren wir aber für die Note, welche von beiden Variablen kausal verursacht wird, bedeutet das inhaltlich, dass wir uns den Zusammenhang von Intelligenz und Fleiß unter Konstanthaltung der Noten anschauen. Dadurch entsteht aber ein (nicht-kausaler) Zusammenhang zwischen diesen beiden Variablen.
- **Beispiel**: Wenn wir die Intelligenz von Personen betrachten und dabei für die Note kontrollieren (also z.B. die Intelligenz aller Personen mit einer guten Note), dann können wir in Abhängigkeit der Intelligenz etwas über den Fleiß dieser Personen sagen. Bei einer guten Note und niedrigen Intelligenz müsste dann ja im Mittel der Fleiß höher sein, andernfalls hätten diese Personen (nach unserer theoretischen Annahme) keine gute Note. Wir würden also einen negativen Zusammenhang zwischen Intelligenz und Fleiß finden („explain-away effect“).

Beispiel 2: Collider



Datensimulation des kausalen Modells

```
> n <- 1000
> x <- rnorm(n)
> y <- 0*x + rnorm(n)
> z <- 1*x + 1*y + rnorm(n)
```

Verzerrter Effekt

```
Call: lm(y ~ x + z)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001578   0.022284  -0.071    0.944
x           -0.531703   0.027384 -19.417   <2e-16 ***
z            0.536929   0.015840  33.897   <2e-16 ***
```

Korrekter Effekt

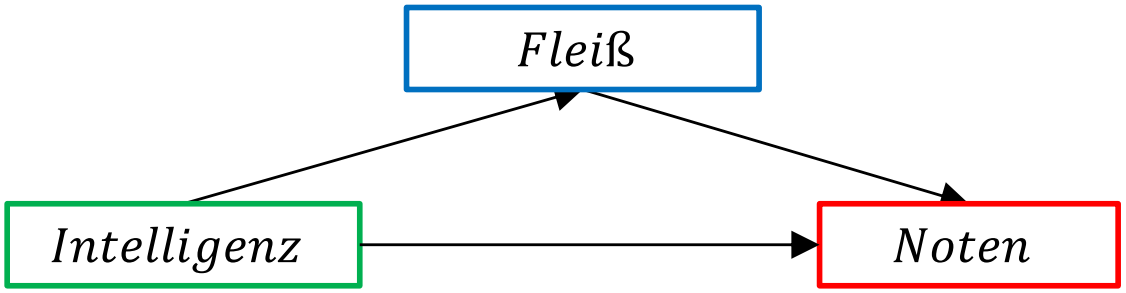
```
lm(y ~ x)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001448   0.032677  -0.044    0.965
x            0.004991   0.032763   0.152    0.879
```

- Ist eine Variable ein Collider (wie hier die Variable Z / Noten), dann sollte sie (wenn möglich) NICHT in das Modell als UV mitaufgenommen werden, um den kausalen Effekt von X auf Y zu schätzen.

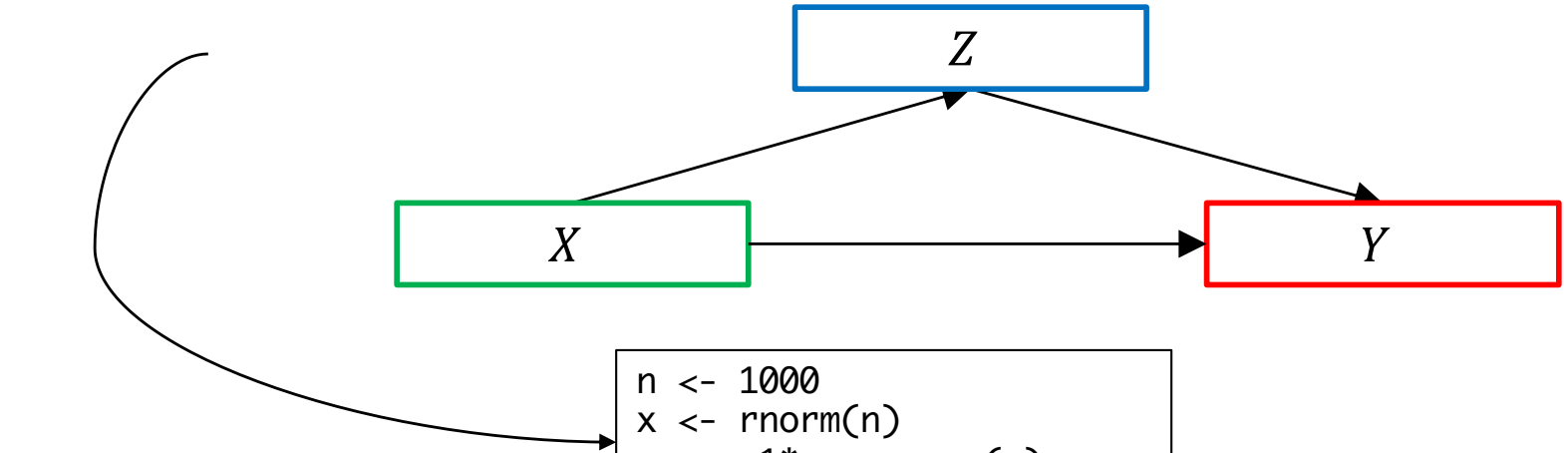
Beispiel 3: Pipe (Mediation)

- Wir interessieren uns für den Einfluss von Intelligenz einer Person auf ihre Noten. Als kausales Modell stellen wir auf:



- Wir gehen davon aus, dass Intelligenz zu besseren Noten führt (= positiver kausaler Zusammenhang). Gleichzeitig verhindert Intelligenz die Drittvariable Fleiß (= negativer kausaler Zusammenhang), obwohl Fleiß selbst wieder zu besseren Noten führen würde.
- In einer ELR würde nun der gesamte (totale) Einfluss von Intelligenz auf Noten geschätzt werden, also inklusive des gegenläufigen Effekts, der über den indirekten Pfad der Variable Fleiß führt.
- In einer MLR bekommen wir eine bessere Schätzung des einzelnen (direkten) Einflüsses.

Beispiel 3: Pipe (Mediation)



Datensimulation des kausalen Modells

```
n <- 1000
x <- rnorm(n)
z <- -1*x + rnorm(n)
y <- 1*x + 1*z + rnorm(n)
```

Totaler Effekt

```
Call: lm(y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03143    0.04333  -0.725    0.468
x           -0.03880    0.04291  -0.904    0.366
```

Direkter Effekt

```
Call: lm(y ~ x + z)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03073    0.03183  -0.966    0.334
x            0.90425    0.04513  20.035   <2e-16 ***
z            0.91394    0.03130  29.195   <2e-16 ***
```

- Ist eine Variable ein Mediator (wie hier die Variable Z / Fleiß), dann hängt es von der Fragestellung ab, ob sie in das Modell als UV mitaufgenommen werden sollte.

Zusammenfassung

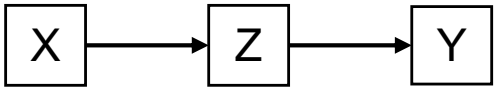
- Alle DAGs bestehen aus nur 3 charakteristischen Kausalstrukturen:

Confounder ("Fork")

„Der Confounder Z verursacht sowohl die Variable X als auch die Variable Y"

Wenn möglich, sollten Confounder mit in das statistische Modell aufgenommen werden!

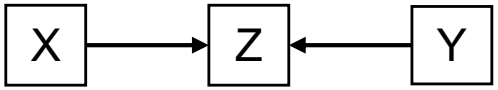
Mediator („Pipe")



„Die Variable X verursacht die Variable Y über den Mediator Z"

Es ist abhängig davon, welcher kausale Effekt von X auf Y von Interesse ist, ob ein Mediator ins Modell aufgenommen werden soll!

Collider



"Sowohl die Variable X als auch die Variable Y verursachen den Collider Z"

Wenn möglich, sollten Collider **nicht** mit in das statistische Modell aufgenommen werden!

- Mit diesen Regeln lässt sich aus jedem noch so komplexen DAG ableiten, welche Variablen für welche Fragestellung nötig sind (siehe auch http://www.dagitty.net/dags.html).

Zusammenfassung

- **Ob ein DAG zutreffend ist, ist wieder eine inhaltliche Diskussion. Ist der DAG eine falsche Repräsentation der Realität, dann sind auch die daraus abgeleiteten notwendigen Prädiktoren im Modell unter Umständen falsch.**
- Das ist allerdings kein Argument gegen die Verwendung von DAGs, sondern eher eine Kritik an ungenau formulierten psychologischen Theorien:

➢ Eine gute Theorie sollte es uns ermöglichen, die Zusammenhänge zwischen den Variablen, die die Theorie umfasst, sinnvoll darzustellen.

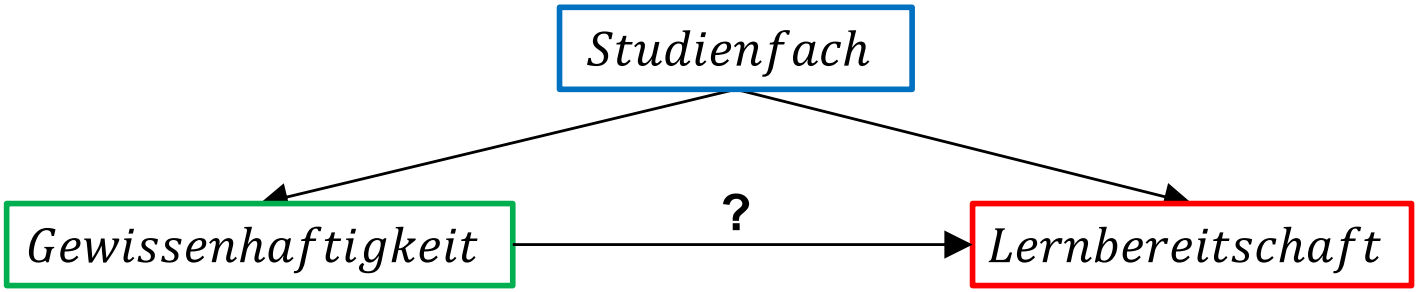
➢ Annahmen über Zusammenhänge zwischen Variablen und deren Rollen im Modell verschwinden nicht einfach, nur, weil wir diese Annahmen nicht offenlegen (z.B. in einem DAG).

➢ DAGs führen keine neuen Annahmen ein, sie legen nur unsere Annahmen offen.

➢ Sobald wir ein Modell in R (oder einer anderen Statistiksoftware) schätzen wollen und die entsprechende Syntax eingeben (z.B. „Intelligenz ~ Fleiß + Noten“), haben wir bereits implizit Annahmen über Zusammenhänge und Rollen von Variablen getroffen!

Gewollte Kontrolle von Drittvariablen

- Was meinen wir mit „sollte wenn möglich (nicht) in das Modell aufgenommen werden“ / „es sollte wenn möglich (nicht) für diese Variable kontrolliert werden“?
- Man kann auf mehrere Arten für Drittvariablen kontrollieren.
- Beispiel: Wir interessieren uns für den kausalen Effekt von Gewissenhaftigkeit auf Lernbereitschaft und vermuten aber, dass das Studienfach (BWL vs. Psychologie) ein Confounder ist.
- Der DAG ist also:



Gewollte Kontrolle von Drittvariablen

Wir können für den theoretisch angenommenen Confounder Studienfach auf drei Arten kontrollieren:

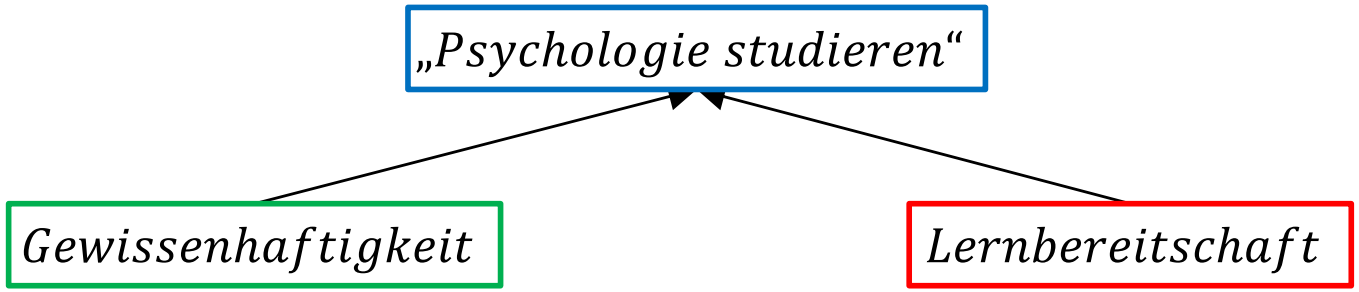
1. **Stratifizierung**: Das Modell wird für alle Gruppen einer Variable einzeln gerechnet.
   Beispiel: Um für Studienfach zu kontrollieren, wird eine ELR mit Gewissenhaftigkeit als UV und Lernbereitschaft als AV in der Gruppe der BWL-Studierenden und dann in der Gruppe der Psychologie-Studierenden gerechnet.
   Problem: schwierig bzw. nicht möglich, wenn der Confounder viele Ausprägungen hat (z.B. weitere Studienfächer) oder metrisch ist (z.B. Alter)

2. **Aufnahme der Drittvariable in das Modell**: Es wird eine MLR mit Gewissenhaftigkeit und Studienfach als UVs und Lernbereitschaft als AV gerechnet.
   Problem: Es wird nur dann vollständig für den Einfluss der Drittvariable kontrolliert, wenn der angenommene funktionale Zusammenhang stimmt (z.B. ist dieser in einer M**L**R immer **linear**).

3. **Kontrolle durch Forschungsdesign**: Durch Erhebung der Stichprobe in bestimmten Populationen kann für Drittvariablen kontrolliert werden.
   Beispiel: Um Zusammenhänge zwischen dem Verhalten der Eltern und der Entwicklung von Kindern zu untersuchen, werden nur Familien mit adoptierten Kindern betrachtet, um für genetische Einflüsse der Eltern zu kontrollieren.

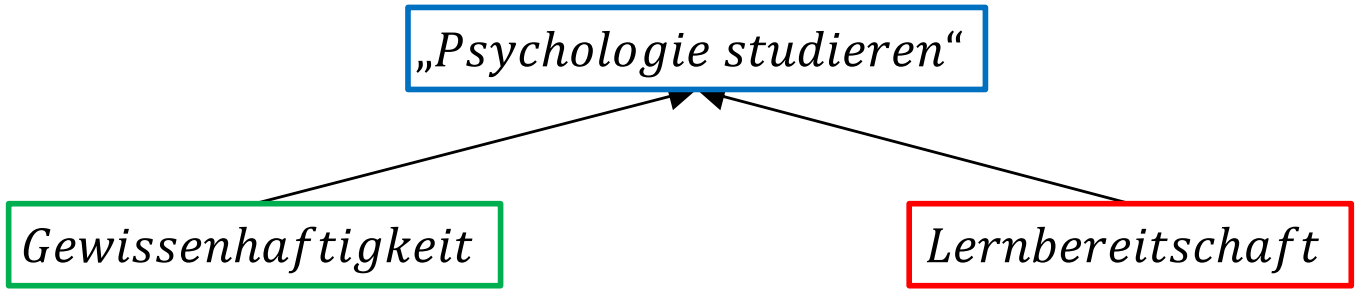
Ungewollte Kontrolle von Drittvariablen

- Vor allem der letzte Punkt, Kontrolle durch Forschungsdesign, kann zu einer ungewollten Kontrolle von Variablen führen (indem sich Collider in das Modell bzw. die Untersuchung „einschleichen“ → Collider-Bias)
- Beispiel: Wir erheben unsere Stichprobe ausschließlich in der Population der Psychologie-Studierenden und untersuchen den kausalen Effekt von Gewissenhaftigkeit auf Lernbereitschaft.
- Es ist denkbar, dass sowohl die Gewissenhaftigkeit als auch die Lernbereitschaft einen kausalen Effekt darauf haben, einen Studienplatz in Psychologie zu erhalten.



Ungewollte Kontrolle von Drittvariablen

- “Psychologie studieren“ ist hier also ein Collider!
- Da nur Psychologie-Studierende in unserer Stichprobe sind, entspricht das einer Kontrolle der Variable „Psychologie studieren“, also einer Kontrolle des Colliders. (da wir ja nur Psychologie-Studierende betrachten)
- Ergebnisse einer ELR zwischen Gewissenhaftigkeit und Lernbereitschaft sind also unter Umständen verzehrt, obwohl wir die Drittvariable „Psychologie studieren“ nicht einmal miterhoben haben.
- Es ist in diesem Fall also **nicht** möglich, **nicht** für den Collider zu kontrollieren!

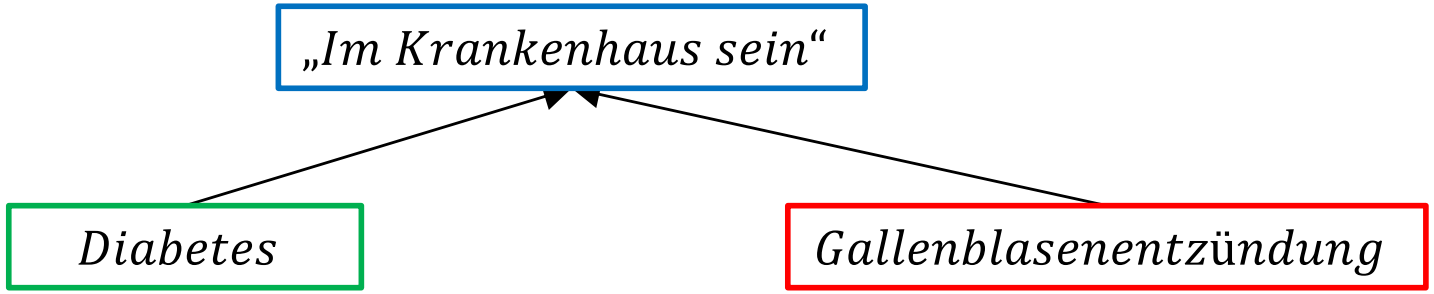


Ungewollte Kontrolle von Drittvariablen

- Collider-Bias tritt also immer dann auf, wenn wir
  - einen Collider mit in das statistische Modell aufnehmen (z.B. wie auf Folie 17) oder
  - durch spezielle Selektion der Stichprobe aus einer Population (ungewollt) für eine Variable kontrollieren (z.B. wie auf Folie 24)
- In der Literatur nennt man diese zweite Form von Collider Bias auch „Berkson‘s paradox“ (das Prinzip ist ähnlich zum „explain-away effect“ auf Folie 16).

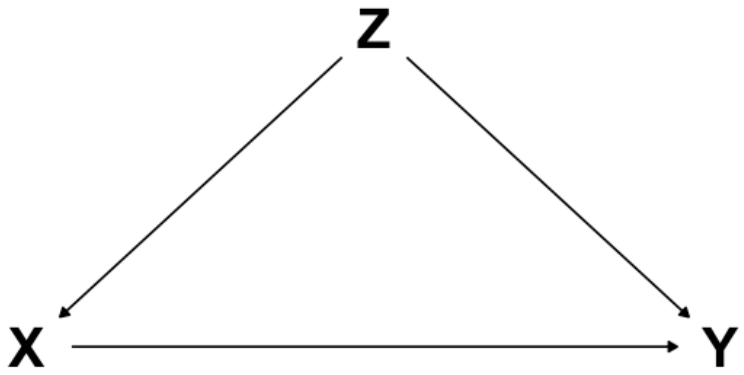
Ungewollte Kontrolle von Drittvariablen

- Originalbeispiel von Berkson:
  - Untersuchung der Risikofaktoren einer Krankheit mithilfe einer Stichprobe von Personen, **die aktuell im Krankenhaus sind**
  - Beispiel: Ist Diabetes ein Risikofaktor für Gallenblasenentzündungen?
  - Eine Person im Krankenhaus **ohne Diabetes** hat eine **höhere Wahrscheinlichkeit eine Gallenblasenentzündung** zu haben (verglichen mit einer Person aus der Gesamtpopulation außerhalb)
  - Grund: Die Person, die im Krankenhaus ist, muss einen anderen Grund als Diabetes (der potentiell eine Gallenblasenentzündung verursacht) für die Einlieferung haben. Sonst wäre sie nicht im Krankenhaus.

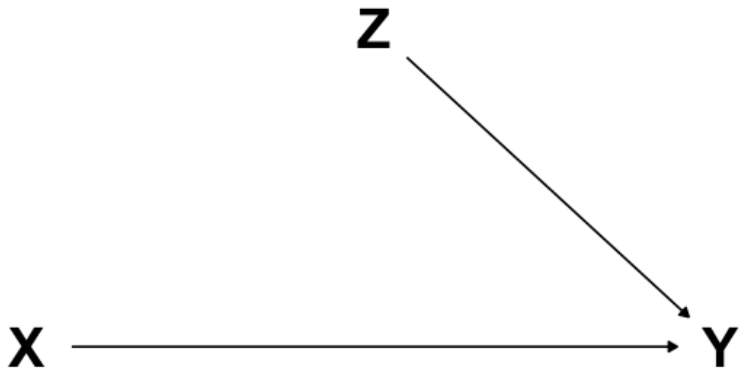


DAG Darstellung von Experimenten

- In der Psychologie wird eine kausale Interpretation eines Effekts von X auf Y häufig nur dann akzeptiert, wenn die Variable X in einem Experiment zufällig randomisiert wurde (z.B. RCT in der klinischen Psychologie, A/B Test in der Werbepsychologie). Wie lässt sich ein randomisiertes Experiment als DAG darstellen?
- Angenommen man geht von dem folgenden DAG (links) aus und interessiert sich für den kausalen Effekt von X auf Y. Eine echte Randomisierung von X (im Rahmen eines Experiments) führt dazu, dass alle Pfeile auf X gelöscht werden können, da die randomisierte Variable per Design nicht mehr von den anderen Variablen beeinflusst sein kann.
- Ein experimentelles Design ist also (vereinfacht gesagt) die gleichzeitige Kontrolle von allen möglichen Confoundern.



Ohne Randomisierung



Randomisierung von X

Zusammenfassung

- Kausale Aussagen basieren immer auf kausalen Annahmen, die inhaltlich-theoretischer Natur sind – nicht statistischer Natur! (auch in Experimenten)
- Directed Acyclic Graphs können genutzt werden, um diese Annahmen graphisch darzustellen.
- Die Frage, ob für eine Drittvariable kontrolliert werden muss, ist erneut eine inhaltliche Frage. Es hängt davon ab, ob die Drittvariable ein Confounder, Mediator oder Collider ist.
- Die Kontrolle von Drittvariablen ist auf mehrere Arten möglich: Stratifizierung, Aufnahme in das Modell oder per Design.
- Vor allem bei möglichen Collider-Drittvariablen ist es wichtig zu beachten, dass nicht ungewollt für sie kontrolliert wird, zum Beispiel durch „falsche“(d.h., selektive) Stichprobenziehung.

Weitere Infos zu DAGs

Weiterführende Literatur zu diesem Thema:

- Cinelli, C., Forney, A., & Pearl, J. (2022). A Crash Course in Good and Bad Controls. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221099552
- Elwert, F. (2013). Graphical Causal Models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 245–273). Dordrecht: Springer Netherlands
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. 2nd Edition. Chapman and Hall/CRC. **Oder in seiner Online Vorlesung auf Youtube (u.a. Lecture 6: https://youtu.be/UpP-_mBvECI)**
- **Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, 27–42. https://doi.org/10.1177/2515245917745629**

Modellierung wissenschaftlicher Fragestellungen

Fazit Statistik 1 und 2

- Wir haben nun die nötigen statistischen Hilfsmittel, um für unterschiedliche Konfigurationen von AV und UV(s) aus Stichprobendaten Schlüsse auf Gegebenheiten in einer Population zu ziehen (= statistische Inferenz)
- Dabei unterscheiden wir konzeptuell zwischen
  1. der Schätzung von Parametern oder Parameterkombinationen mit **Konfidenzintervallen**
  2. der **Testung von Nullhypothesen** über Parameter oder Parameterkombinationen
- Je nach inhaltlicher Fragestellung sind bestimmte (aber nicht unbedingt alle) Parameter (oder Kombinationen) eines Modells für uns von besonderem Interesse.
- Je nach inhaltlicher Fragestellung kann es für die Interpretation von Parametern (bzw. deren Hypothesentests) hilfreich oder sogar notwendig sein, Variablen im Modell zu
  - **zentrieren** (um der Ausprägung 0 eine Bedeutung zu verleihen), oder sogar zu
  - **z-standardisieren** (um zusätzlich unabhängig von verwendeten Einheiten einer Variable zu sein)

Fazit Regressionsmodelle

- Mit den in Statistik 2 besprochenen Regressionsmethoden ist eine große Vielfalt an verschiedenen Forschungsfragen möglich.
- AV (Kriterium):
  - Stetig (lineare Regression) oder
  - Binär (logistische Regression)
- UVs (Prädiktoren):
  - Eine (einfache Regression) oder mehrere (multiple Regression) stetige UVs
  - Eine (einfache Regression) oder mehrere (multiple Regression) diskrete (kategoriale) UVs mit zwei oder mehr Ausprägungen
  - Interaktionen zwischen zwei oder mehr stetigen und/oder diskreten (kategorialen) UVs
  - Beliebige Kombinationen der oben genannten Möglichkeiten
- Stetige AVs und UVs können jeweils unstandardisiert, standardisiert oder zentriert verwendet werden. Auch andere Transformationen sind möglich.

Übersicht inferenzstatistischer Modelle in Statistik 1 & 2

Modell	AV	UV
Einstichproben t-Test	stetig (Mittelwert)	-
Zweistichproben t-Test	stetig (Mittelwert)	diskret (2 Ausprägungen)
Binomialtest	diskret/binär (Häufigkeit)	-
ANOVA (einfaktoriell)	stetig (Mittelwert)	diskret (2+ Ausprägungen)
ANOVA (mehrfaktoriell)	stetig (Mittelwert)	mehrere diskrete (mit je 2+ Ausprägungen)
Lineare Regression	stetig (Mittelwert)	(keine,) eine oder mehrere diskrete (mit je 2+ Auspr.) oder stetige
Logistische Regression	diskret/binär (Häufigkeit)	(keine,) eine oder mehrere diskrete (mit je 2+ Auspr.) oder stetige

Statistische Verfahren aus Statistik 1 als Regressionsmodelle

- Alle Statistischen Verfahren aus Statistik 1 und 2 lassen sich auch als Regressionsmodelle darstellen (und noch viele mehr).
- Das ist vor allem dann sinnvoll, wenn man die Modelle erweitern will, um kompliziertere Forschungsfragen zu beantworten, für die einfachen Verfahren nicht ausreichen.
- Wir werden uns nun einige wichtige Beispiele ansehen, wie die Verfahren aus Statistik 1 und 2 als Regressionsmodelle dargestellt werden können.

Korrelation

Korrelation vs. Regression

Auch zur in Vorlesung 3 von Statistik 1 vorgestellten Methode der Korrelation zur Beschreibung des Zusammenhangs zweier stetiger Variablen gibt es mit der Effektstärke β_z einer ELR eine Möglichkeit zur statistischen Inferenz:

Korrelation:

```
> cor(Sportdaten$Liegestuetz_prae, Sportdaten$Standweitsp_prae)
[1] 0.1790058
```

Regression mit z-Standardisierter AV und UV:

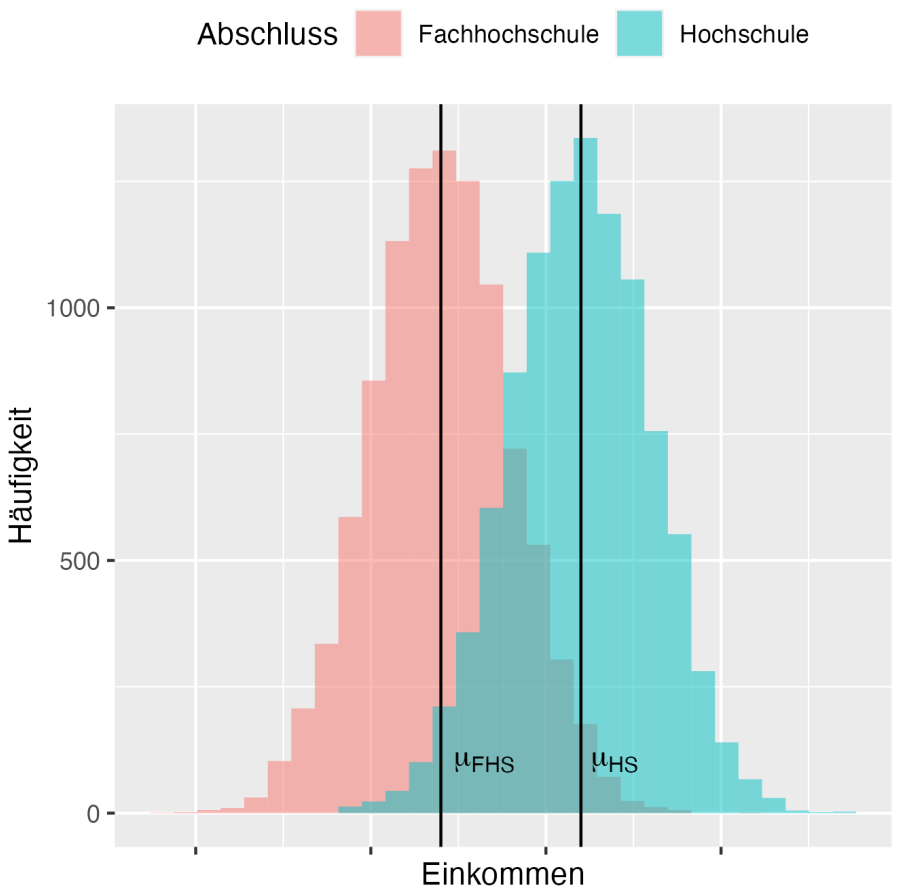
```
> r <- lm(scale(Liegestuetz_prae) ~ 0 + scale(Standweitsp_prae),
Sportdaten)
> summary(r)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
scale(Standweitsp_prae)    0.1790     0.1796   0.997    0.327
```

Bemerkung: Weil wir für β_z auch ein Konfidenzintervall berechnen können, kennen wir damit auch eine einfache Möglichkeit für die Intervallschätzung von Korrelationen.

Eine stetige AV, eine diskrete UV mit 2 Ausprägungen

t-Test (2 Stichproben, unabh.) vs. Regressionsmodell



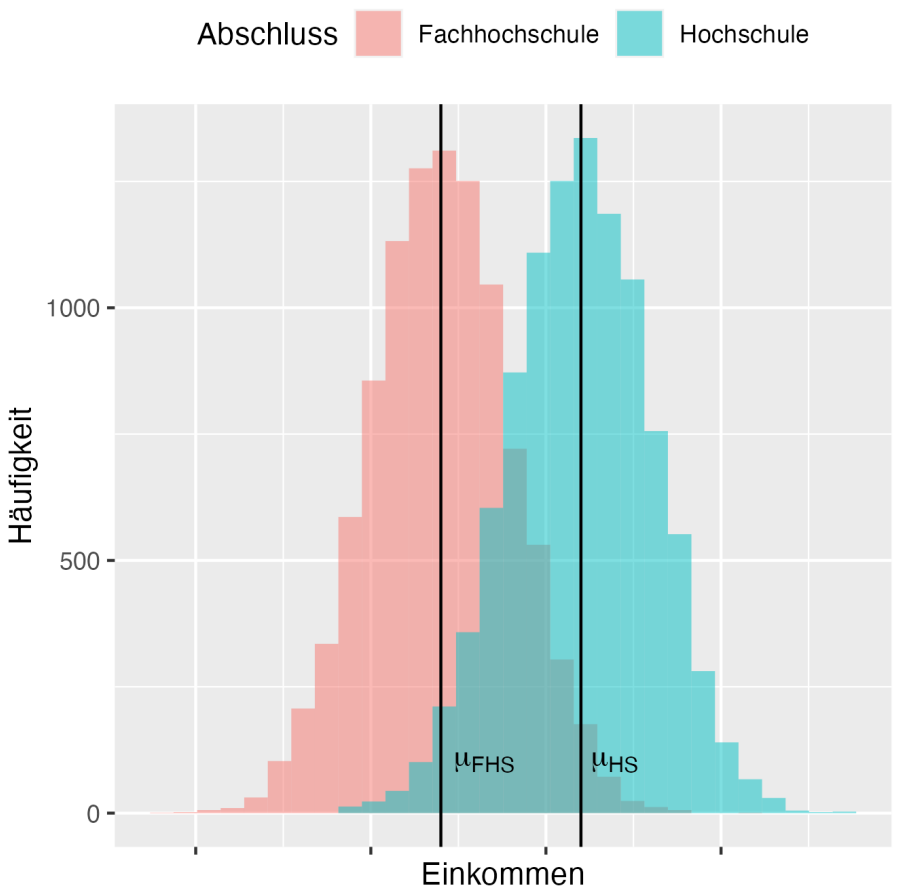
- Fragestellung in Anlehnung an Aufgabe 6/7 des Übungsblattes *Regressionsmodelle mit diskreten Prädiktoren*:

  *Sie interessieren sich dafür, **ob** sich das durchschnittliche Einkommen von Personen mit Hochschulabschluss und Personen mit Fachhochschulabschluss unterscheidet:*

- Hypothesen:

$$H_0: \mu_{FHS} - \mu_{HS} = 0$$
$$H_1: \mu_{FHS} - \mu_{HS} \neq 0$$

t-Test (2 Stichproben, unabh.) vs. Regressionsmodell



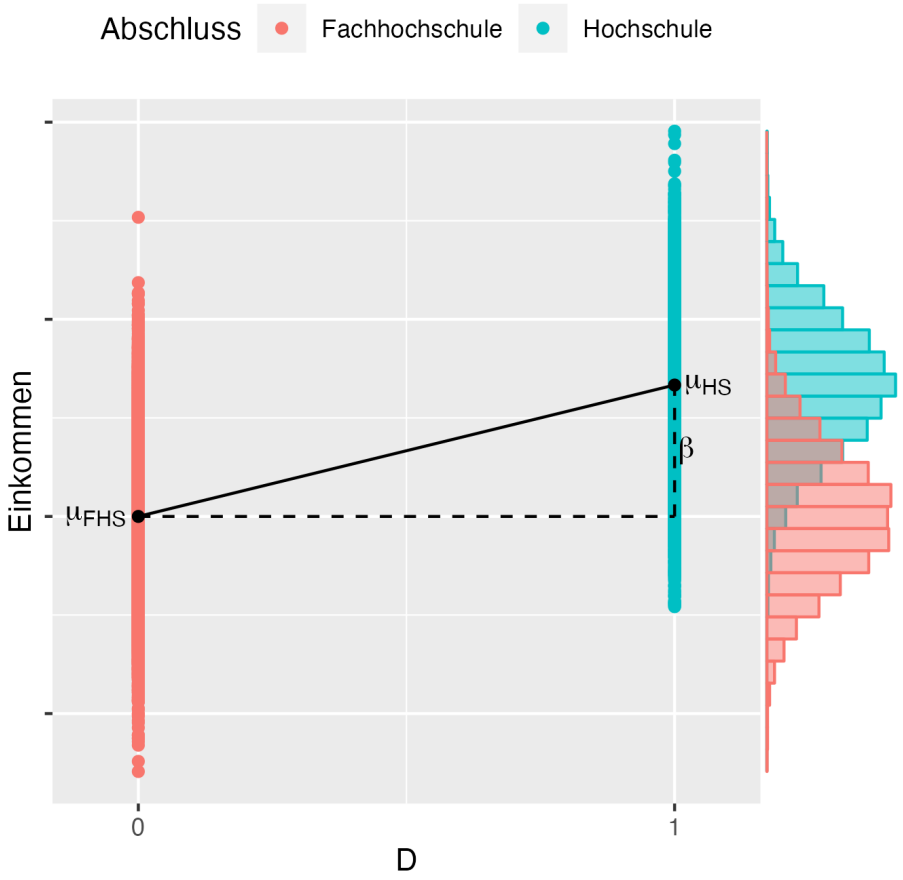
Testen der Hypothesen mit einem T-Test:

```
> t.test(Einkommen ~ Abschluss, einkommen, var.equal = TRUE, paired = FALSE)

        Two Sample t-test

data:  Einkommen by Abschluss
t = -10.623, df = 298, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Fachhochschule and group Hochschule is not equal to 0
```

t-Test (2 Stichproben, unabh.) vs. Regressionsmodell



- Modellierung der Fragestellung mit einem Dummy-Regressionsmodell

$$Y_i = \alpha + \beta \cdot D_i + \epsilon_i$$

- Hypothesen:

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

- *Hinweis*:

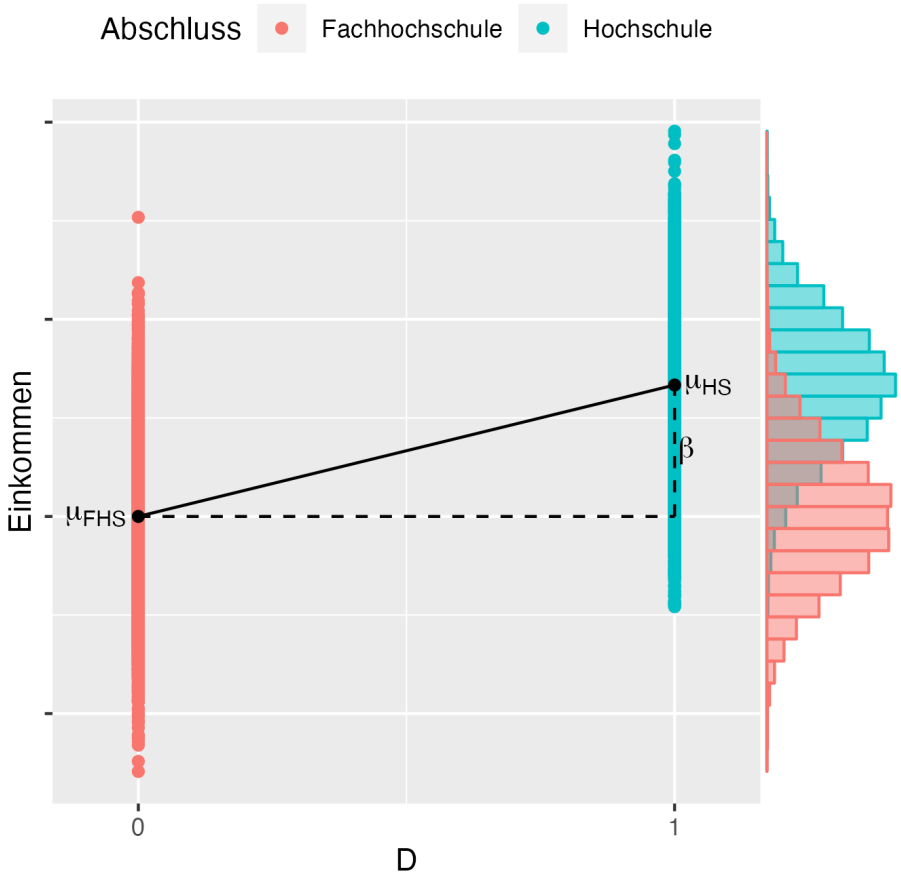
$$\beta \mathrel{\widehat{=}} \mu_{HS} - \mu_{FHS}$$

während im t-Test

$$\mu_{FHS} - \mu_{HS}$$

getestet wird.

t-Test (2 Stichproben, unabh.) vs. Regressionsmodell



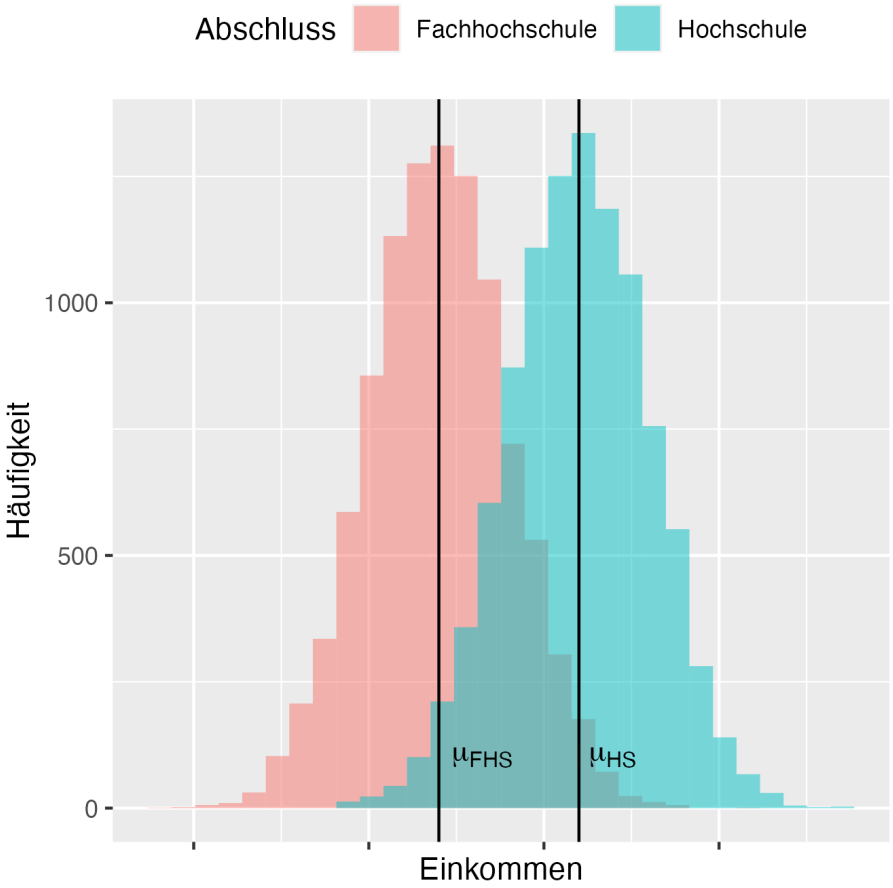
Testung der Hypothesen mit einem Dummy-Regressionsmodell

```
> r <- lm(Einkommen ~ Abschluss, einkommen)
> summary(r)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    35837       1788   20.04   <2e-16
AbschlussHS    26861       2529   10.62   <2e-16

Residual standard error: 21900 on 298 degrees of
freedom
```

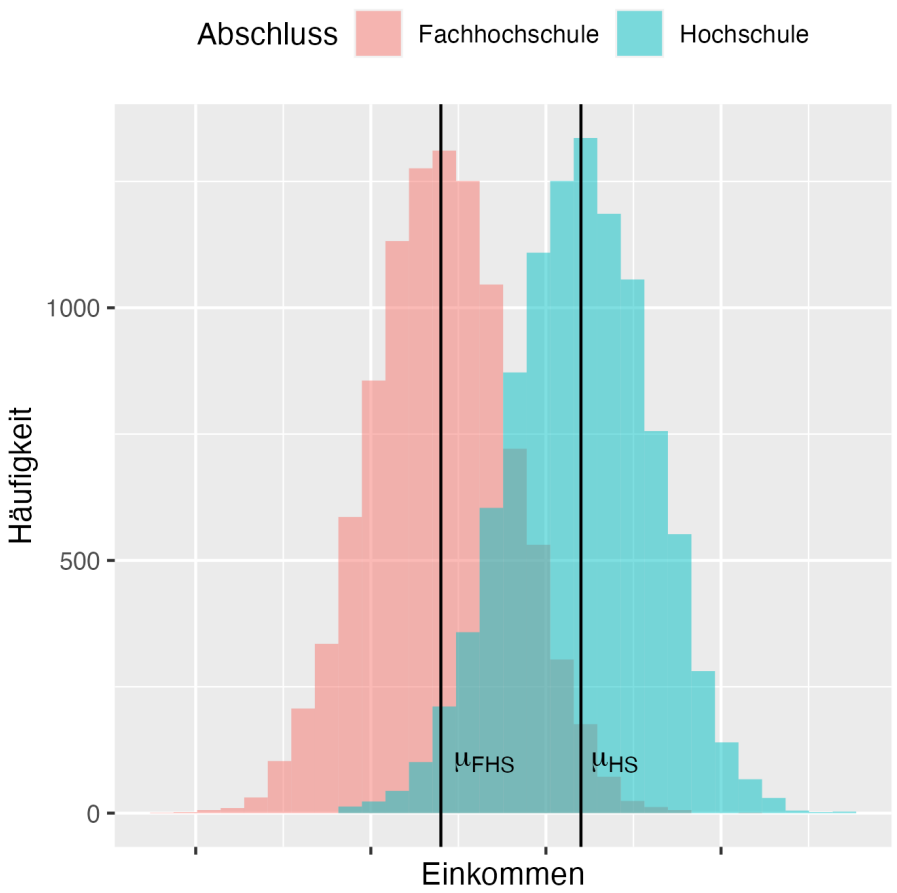
KI für $\mu_1 - \mu_2$ vs. Regressionsmodell



- Fragestellung in Anlehnung an Aufgabe 6/7 des Übungsblattes *Regressionsmodelle mit diskreten Prädiktoren*:

  *Sie interessieren sich dafür, **wie stark** sich das durchschnittliche Einkommen von Personen mit Hochschulabschluss und Personen mit Fachhochschulabschluss unterscheidet.*

KI für $\mu_1 - \mu_2$ vs. Regressionsmodell



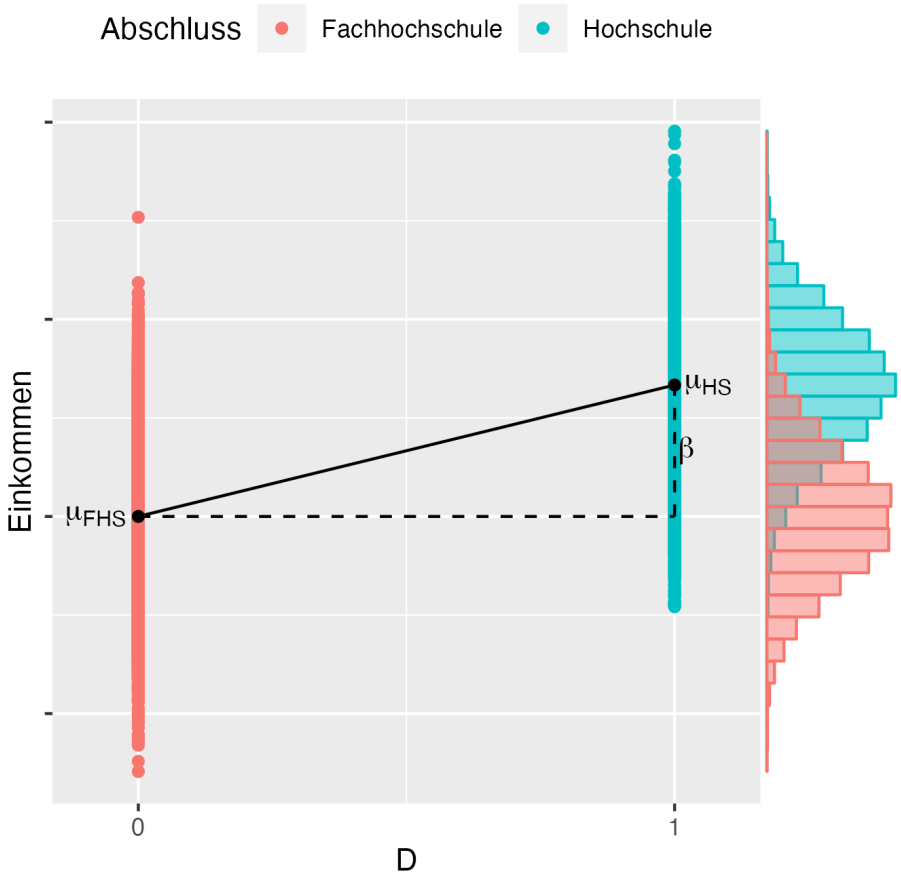
KI für $\mu_{FHS} - \mu_{HS}$ mit den Methoden aus Statistik 1:

```
> t.test(Einkommen ~ Abschluss,
einkommen, var.equal = TRUE, paired =
FALSE)

        Two Sample t-test

95 percent confidence interval:
 -31837.19 -21884.59
```

KI für $\mu_1 - \mu_2$ vs. Regressionsmodell



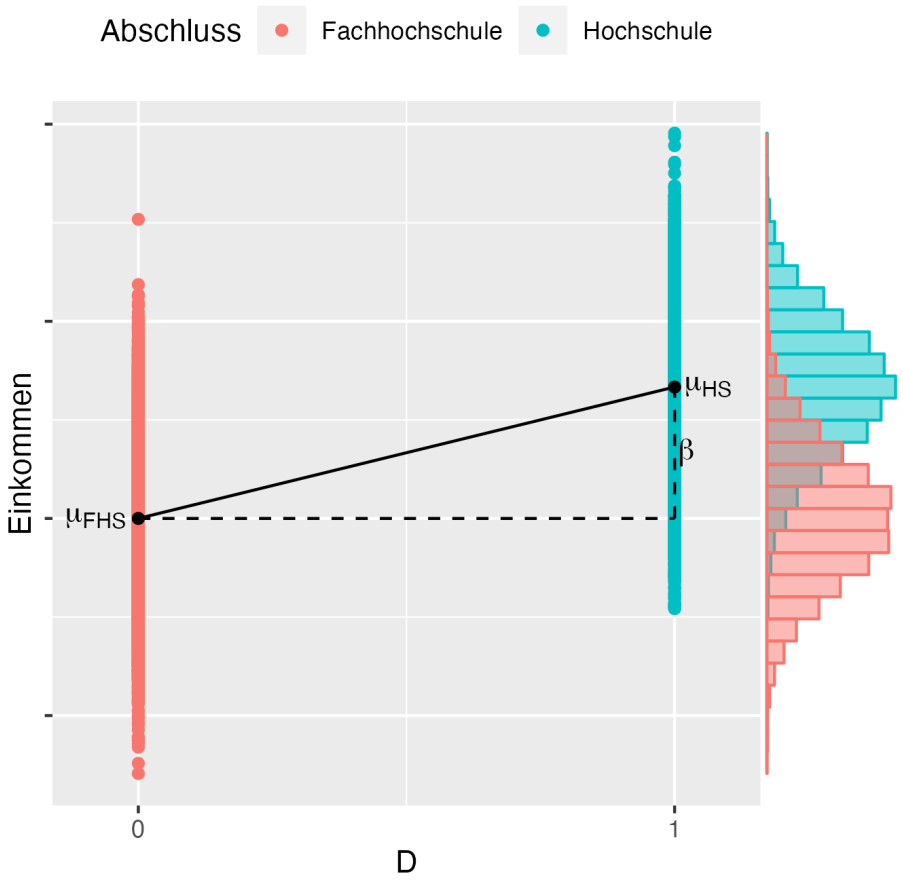
- Modellierung der Fragestellung mit einem Dummy-Regressionsmodell
- Modellgleichung:

$$Y_i = \alpha + \beta \cdot D_i + \epsilon_i$$

- KI für β:

```
> confint(r, par = "AbschlussHochschule")
                          2.5 %   97.5 %
AbschlussHochschule 21884.59 31837.19
```

Kausale Interpretation von $\mu_1 - \mu_2$ (bzw. β)



- Unter welchen Annahmen dürften wir $\mu_1 - \mu_2$ bzw. β als kausalen Effekt des Abschlusses auf das Einkommen interpretieren?

- Wenn wir annehmen würden, dass es entweder 1) **keine Confounder** gibt, die sich kausal sowohl auf den Abschluss als auch das Einkommen auswirken *(sehr unplausibel)* oder 2) der Abschluss in einem Experiment **randomisiert** wurde *(praktisch so gut wie unmöglich)*, dürften wir $\mu_1 - \mu_2$ (bzw. β) als kausalen Effekt interpretieren.

Eine stetige AV, eine diskrete, mehrstufige UV

Einfaktorielle ANOVA

- Aufgabe in Anlehnung an Übungsblatt 4, Aufgabe 1

  a) Sie wollen untersuchen, ob die Farbe des Klausurpapiers (rot, grün oder blau) einen Einfluss auf die (stetige) Prüfungsleistung von Studentinnen hat. (Omnibustest)
  b) Sie haben die Hypothese, dass die Prüfungsleistung bei Klausuren mit rotem Klausurpapier schlechter ist als bei Klausuren mit grünem Klausurpapier.

- Modellgleichung der ANOVA:

$$Y_{ij} = \mu_j + \epsilon_{ij}, \qquad \text{mit j} = \text{rot, grün, blau}$$

- Hypothesen:

  a) $H_0: \mu_{rot} = \mu_{grün} = \mu_{blau}$; $H_1: \mu_j \neq \mu_k$ für mindestens ein Paar j, k

  b) $H_0: \mu_{rot} \geq \mu_{grün}$; $H_1: \mu_{rot} < \mu_{grün}$

Einfaktorielle ANOVA

Hypothesen:

a) $H_0: \mu_{rot} = \mu_{grün} = \mu_{blau}$; $H_1: \mu_j \neq \mu_k$ für mindestens ein Paar j, k

```
> anova <- aov(Pruefungsleistung ~ Farbe, farbe)
> summary(anova)
             Df Sum Sq Mean Sq F value   Pr(>F)
Farbe         2   2933  1466.4    11.6 2.11e-05 ***
Residuals   147  18589   126.5
```

b) $H_0: \mu_{rot} \geq \mu_{grün}$; $H_1: \mu_{rot} < \mu_{grün}$

```
> hypB <- glht(anova, linfct = mcp(Farbe = "rot - gruen >= 0"))
> summary(hypB)

       Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts
Linear Hypotheses:
                  Estimate Std. Error t value Pr(<t)
rot - gruen >= 0    9.380      2.249   4.171      1
```

Einfaktorielle ANOVA als Regression mit zwei Dummy-Variablen

- Modellgleichung der Regression (Farbe = rot ist die Referenzkategorie):

$$Y_i = \alpha + \beta_1 D_1 + \beta_2 D_2 + \epsilon_i$$

$$D_1 = \begin{cases} 1, \text{falls Farbe} = grün \\ 0, \text{sonst.} \end{cases}$$

$$D_2 = \begin{cases} 1, \text{falls Farbe} = blau \\ 0, \text{sonst.} \end{cases}$$

- $\mu_{rot} = E(Y_i|D_1 = 0, D_2 = 0) = \alpha$

  $\mu_{grün} = E(Y_i|D_1 = 1, D_2 = 0) = \alpha + \beta_1$

  $\mu_{blau} = E(Y_i|D_1 = 0, D_2 = 1) = \alpha + \beta_2$

➢ $\mu_{grün} - \mu_{rot} = \alpha + \beta_1 - \alpha = \beta_1$

➢ Hypothesen:

a) $H_0: \beta_j = 0 \; für \; alle \; j; \;\; H_1: \beta_j \neq 0$ für mindestens ein j

b) $H_0: \beta_1 \leq 0; \;\; H_1: \beta_1 > 0$

Einfaktorielle ANOVA als Regression mit zwei Dummy-Variablen

a) $H_0: \beta_j = 0$ *für alle* j; $H_1: \beta_j \neq 0$ für mindestens ein j

```
> r <- lm(Pruefungsleistung ~ Farbe, farbe)
> summary(r)

Residual standard error: 11.25 on 147 degrees of freedom
Multiple R-squared:  0.1363,   Adjusted R-squared:  0.1245
F-statistic:  11.6 on 2 and 147 DF,  p-value: 2.107e-05
```

b) $H_0: \beta_1 \leq 0$; $H_1: \beta_1 > 0$

```
> summary(glht(r, linfct = "Farbegruen <= 0"))

      Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = Pruefungsleistung ~ Farbe, data = farbe)

Linear Hypotheses:
                 Estimate Std. Error t value Pr(>t)
Farbegruen <= 0   -9.380      2.249  -4.171      1
(Adjusted p values reported -- single-step method)
```

Kausale Interpretation von Hypothese b)

$$H_0: \beta_1 \leq 0; \;\; H_1: \beta_1 > 0$$

```
> summary(glht(r, linfct = "Farbegruen <= 0"))

       Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = Pruefungsleistung ~ Farbe, data = farbe)

Linear Hypotheses:
                 Estimate Std. Error t value Pr(>t)
Farbegruen <= 0   -9.380      2.249  -4.171      1
(Adjusted p values reported -- single-step method)
```

- Unter welchen Annahmen handelt es sich hier um einen Test des kausalen Effekts der Papierfarbe rot (im Vergleich zu grün) auf die Prüfungsleistung?

- Wenn wir annehmen würden, dass es entweder 1) **keine Confounder** gibt, die sich kausal sowohl auf die Papierfarbe als auch die Prüfungsleistung auswirken *(je nach Kontext eventuell plausibel)* oder 2) die Farbe in einem Experiment **randomisiert** wurde *(wäre praktisch umsetzbar)*, dürften wir den Hypothesentest kausal interpretieren.

Eine stetige AV, keine UV (nur eine Stichprobe)

KI für μ vs. Regressionsmodell

Siehe Statistik 1, Übungsblatt "Intervallschätzung II", Aufgabe 6a:

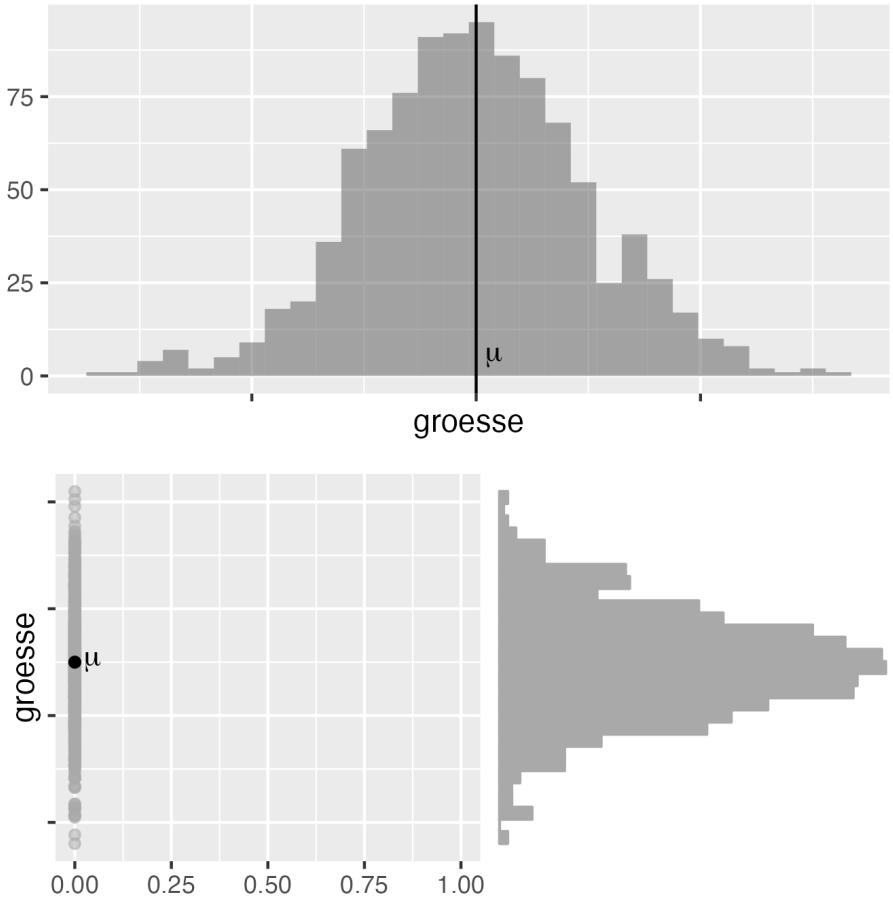
- *Der Datensatz enthält die Körpergröße in cm und die Augenfarbe von n = 1000 Personen, die zufällig aus einer Population gezogen wurden (einfache Zufallsstichprobe). Sie können davon ausgehen, dass das Histogramm der Körpergröße in der Population durch die Wahrscheinlichkeitsdichtefunktion einer Normalverteilung approximiert werden kann.*
  Berechnen Sie ein 0.95-Konfidenzintervall für die durchschnittliche Körpergröße in der Population.

➢ `t.test(daten$groesse, conf.level = 0.95)`

```
        One Sample t-test

data:  daten$groesse
t = 538.18, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 169.1151 170.3529
```

KI für μ vs. Regressionsmodell



- Im Einstichprobenfall haben wir gar keine UV und wir haben gelernt, dass selbst die ELR eine UV braucht.
- Es ist aber tatsächlich möglich, ein Regressionsmodell ohne UV und nur mit einem Achsenabschnitt zu betrachten (sog. Intercept-only Modell):

```
> r <- lm(groesse ~ 1, Daten)
> confint(r)
                2.5 %    97.5 %
(Intercept) 169.1151 170.3529
```

Eine diskrete, binäre AV, keine UV (nur eine Stichprobe)

Binomialtest und KI für π vs. Regressionsmodell

Siehe Statistik 1, Übungsblatt "Intervallschätzung II", Aufgabe 6b:

- *Der Datensatz enthält die Körpergröße in cm und die Augenfarbe von n = 1000 Personen, die zufällig aus einer Population gezogen wurden (einfache Zufallsstichprobe). Sie können davon ausgehen, dass das Histogramm der Körpergröße in der Population durch die Wahrscheinlichkeitsdichtefunktion einer Normalverteilung approximiert werden kann.*
  Berechnen Sie ein 0.95-Konfidenzintervall für die relative Häufigkeit der Augenfarbe Braun in der Population. Interpretieren Sie dieses.

```
> table(Daten$augenfarbe)
 blau braun gruen
  310   354   336
> DescTools::BinomCI(354, 1000, method = 'wald', conf.level =
0.95)
       est    lwr.ci    upr.ci
[1,] 0.354 0.3243608 0.3836392
```

Binomialtest und KI für π vs. Regressionsmodell

- Wir können nun auch in der logistischen Regression die UV gänzlich weg lassen und nur einen einzelnen Parameter α schätzen.

- Diesen bzw. sein KI müssen wir dann aber noch wegen $P(Y_i = 1) = \frac{e^{\alpha}}{1+e^{\alpha}}$ in Wahrscheinlichkeiten umrechnen:

```
> r <- glm(braun ~ 1, Daten,  family = 'binomial')
> confint(r)
     2.5 %      97.5 %
-0.7319761 -0.4726674

> ki <- confint(r)

> exp(ki) / (1 + exp(ki))
    2.5 %     97.5 %
0.3247612 0.3839851
```

Erweiterung der Modelle durch die Möglichkeiten der Regression

Erweiterte Fragestellungen I

- Frage: Gibt es Unterschiede im Zusammenhang zwischen der Erfahrung des Therapeuten/der Therapeutin (in Jahren) und der stetigen Depressionsschwere nach der Therapie **je nach Therapieform** (KVT, TT oder ET)?
- Modellgleichung:

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_{KVT,i} + \beta_3 \cdot D_{TT,i} + \beta_4 \cdot \left(X_i \cdot D_{KVT,i}\right) + \beta_5\left(X_i \cdot D_{TT,i}\right) + \varepsilon_i$$

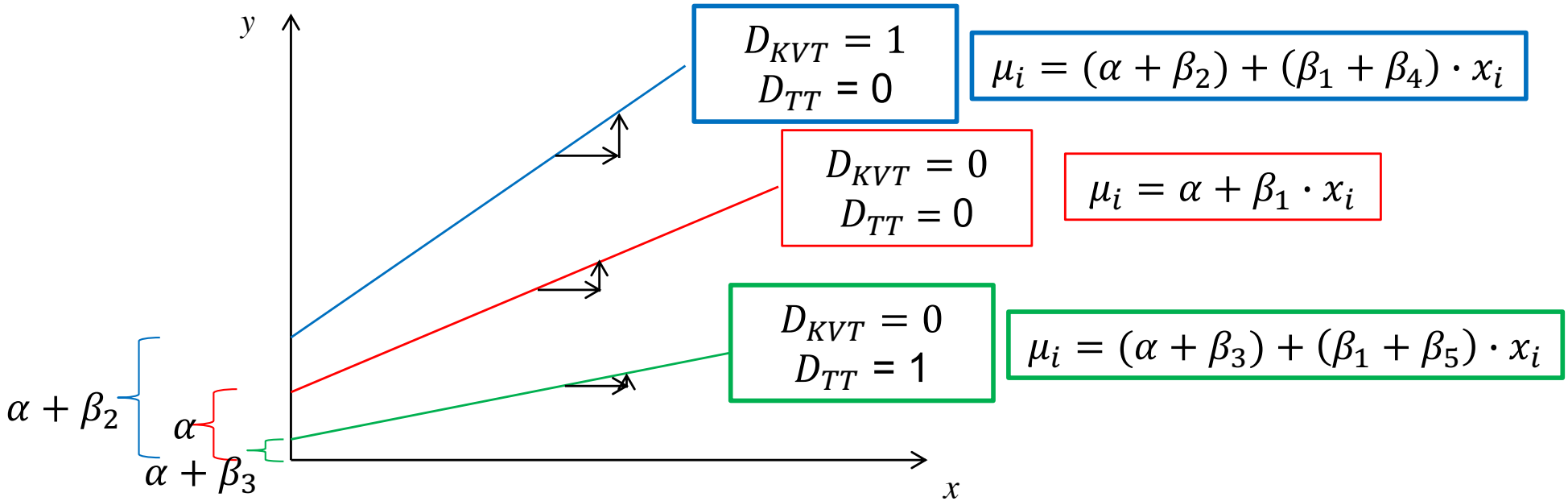
mit ET als Referenzkategorie und zwei Dummy-Variablen, die kennzeichnen ob eine Person i mit KVT, bzw. TT therapiert wurde.

- Hypothesen:

$H_0: \beta_4 = 0$ und $\beta_5 = 0$ und $\beta_5 - \beta_4 = 0$

$H_1: \beta_4 \neq 0$ oder $\beta_5 \neq 0$ oder $\beta_5 - \beta_4 \neq 0$

Erweiterte Fragestellungen I



- Der Modellparameter β_4 quantifiziert den Unterschied in der Steigung zwischen KVT und ET, während sich β_5 auf den Unterschied in der Steigung zwischen TT und ET bezieht.
- β_2, bzw. β_3 quantifiziert den therapiebezogenen Unterschied in der durchschnittlichen Depressionsschwere von Personen, die von einem Therapeuten/ einer Therapeutin ohne Berufserfahrung therapiert wurden

Erweiterte Fragestellungen II

- Frage: Ändert sich das Einkommen mit steigender Intelligenz positiver für Menschen mit Hochschulabschluss und ohne Migrationshintergrund als für Menschen mit Migrationshintergrund und ohne Hochschulabschluss?
- Modellgleichung (es wird komplex):

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 D_{1,i} + \beta_3 D_{2,i} + \beta_4 \left(D_{1,i} \cdot D_{2,i}\right) + \beta_5 \left(X_i \cdot D_{1,i}\right) + \beta_6 \left(X_i \cdot D_{2,i}\right) + \beta_7 \left(X_i \cdot D_{1,i} \cdot D_{2,i}\right) + \varepsilon_i$$

  mit zwei Dummy-Variablen: D_1 kennzeichnet ob eine Person i einen Hochschulabschluss besitzt (Referenzkategorie: kein Hochschulabschluss) und D_2 gibt an, ob eine Person einen Migrationshintergrund besitzt (Referenzkat.: kein Migrationshintergrund).
- Hypothesen:

$$H_0: \beta_5 \leq \beta_6$$

$$H_1: \beta_5 > \beta_6$$

Erweiterte Fragestellungen II

Modellgleichung (es wird komplex):

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 D_{1,i} + \beta_3 D_{2,i} + \beta_4 (D_{1,i} \cdot D_{2,i}) + \beta_5 (X_i \cdot D_{1,i}) + \beta_6 (X_i \cdot D_{2,i}) + \beta_7 (X_i \cdot D_{1,i} \cdot D_{2,i}) + \varepsilon_i$$

- Zusammenhang Einkommen (Y_i) mit Intelligenz (X_i) für Personen ohne Hochschulabschluss ($D_1 = 0$) und ohne Migrationshintergrund ($D_2 = 0$):

$$Y_i = \alpha + \beta_1 \cdot X_i + \varepsilon_i$$

- ➢ Zusammenhang Einkommen (Y_i) mit Intelligenz (X_i) für Personen mit Hochschulabschluss ($D_1 = 1$) und ohne Migrationshintergrund ($D_2 = 0$):

$$Y_i = \alpha + \beta_2 + (\beta_1 + \beta_5) \cdot X_i + \varepsilon_i$$

- ➢ Zusammenhang Einkommen (Y_i) mit Intelligenz (X_i) für Personen ohne Hochschulabschluss ($D_1 = 0$) und mit Migrationshintergrund ($D_2 = 1$):

$$Y_i = \alpha + \beta_3 + (\beta_1 + \beta_6) \cdot X_i + \varepsilon_i$$

- Zusammenhang Einkommen (Y_i) mit Intelligenz (X_i) für Personen mit Hochschulabschluss ($D_1 = 1$) und mit Migrationshintergrund ($D_2 = 1$):

$$Y_i = (\alpha + \beta_2 + \beta_3 + \beta_4) + (\beta_1 + \beta_5 + \beta_6 + \beta_7) \cdot X_i + \varepsilon_i$$

Erweiterte Fragestellungen III

- Frage: Hat das Einkommen einen Einfluss auf den Zusammenhang zwischen der (stetigen) Arbeitszufriedenheit und der Wahrscheinlichkeit zu kündigen?
- Modellgleichung ($Y_i = 1$ bedeutet Person i hat gekündigt):

$$P(Y_i = 1|x_{i1}, x_{i2}) = \frac{e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3(x_{i1}x_{i2})}}{1 + e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3(x_{i1}x_{i2})}}$$

- X_{i1} ist die Arbeitszufriedenheit einer zufällig gezogenen Person i und X_{i2} gibt das Einkommen einer zufällig gezogenen Person i an. In VL 9 (Folie 59ff.) haben wir Interaktionen zwischen stetigen Prädiktoren eingeführt (inklusive Interpretation als Moderatorvariable). Dies können wir jetzt auf die logistische Regression übertragen (Einkommen wird als Moderator betrachtet).
- Hypothesen:

$$H_0: \beta_3 = 0$$

$$H_1: \beta_3 \neq 0$$

Exkurs: Noch mehr Erweiterungen

- Die eingeführten Regressionsmodelle (sowohl im Rahmen der linearen als auch im Rahmen der logistischen Regression) lassen sich beliebig um Prädiktoren und Interaktionsterme erweitern.
- Beispiel: Lineare Regression mit zwei stetigen Variablen und einer Dummy-Variable sowie den beiden Interaktionstermen zwischen der Dummy-Variable und den jeweiligen stetigen Prädiktorvariablen:

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot D_i + \beta_4 \cdot (X_{i1} \cdot D_i) + \beta_5 (X_{i2} \cdot D_i) + \varepsilon_i$$

- Oder ergänzt um die Interaktion der stetigen Variablen und der Dreifachinteraktion mit der Dummy-Variable:

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot D_i + \beta_4 \cdot (X_{i1} \cdot D_i) + \beta_5 (X_{i2} \cdot D_i) + \beta_6 \cdot (X_{i1} \cdot X_{i2}) + \beta_7 \cdot (X_{i1} \cdot X_{i2} \cdot D_i) + \varepsilon_i$$

- Oder das entsprechende Beispiel mit binärer AV in der logistischen Regression:

$$P(Y_i = 1 | x_{i1}, x_{i2}, d_i) = \frac{e^{\alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot d_i + \beta_4 \cdot (x_{i1} \cdot d_i) + \beta_5 (x_{i2} \cdot d_i)}}{1 + e^{\alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot d_i + \beta_4 \cdot (x_{i1} \cdot d_i) + \beta_5 (x_{i2} \cdot d_i)}}$$

Exkurs: Noch mehr Erweiterungen

- Die eingeführten Regressionsmodelle (sowohl im Rahmen der linearen als auch im Rahmen der logistischen Regression) lassen sich beliebig um Prädiktoren und Interaktionsterme erweitern.
- Theoretisch lassen sich auch transformierte Prädiktorvariablen in der Regressionsanalyse ergänzen. Zum Beispiel könnte eine quadrierte Variable im Modell genutzt werden um einen möglichen quadratischen Zusammenhang abzubilden:

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \varepsilon_i$$

- Die Beispiele auf den letzten Folien zeigen, dass lineare Modelle sehr flexibel erweitert werden können, dann allerdings auch immer schwieriger zu interpretieren sind (insbesondere, wenn viele Interaktionen im Modell sind).
- Deshalb sollten theoretische Überlegung bei der Modellbildung immer berücksichtigt werden und Prädiktoren nicht wahllos hinzugefügt werden (siehe auch erster Teil dieser VL zu kausaler Inferenz).