

Kausale Effekte auf Basis von  
Beobachtungsdaten schätzen

# Organisatorisches

- + Dienstag den 16.05.2023 wird das 4. Projekt (des Gesamtkurses) für jede Gruppe auf Github hochgeladen
- + Mittwoch den 17.05.2023 wird das 4. Projekt besprochen (Vorstellung und Interpretationen)
- + Mittwoch den 24.05.2023 gibt es die Probeklausur im Hörsaal H14 ab 8:30 Uhr. Teil der Vorleistung, mind. 30% der Punkte müssen bestanden werden!
- + Bearbeitungszeit des 4. Projekts: Bis 23.05.2023
- + Ab KW 21 finden die Tutorien statt
- + Mittwoch 24.05.2023 wird das 5. Projekt vorgestellt (nach der Probeklausur)
- + Bearbeitungszeit für das 5. Projekt: Bis 13.06.2023

# Recap zum Experiment

## Interne und Externe Validität:

- + Insbesondere an der internen Validität interessiert
  - + Im Experiment könnte besonders Attrition ein Problem werden
  - + Die Meisten Probleme der internen Validität können durch die Randomisierung im Experiment behoben werden

## Das Experiment konzeptionell:

- + Wie können in Experimenten *kausale* Ergebnisse gewonnen werden?
  - + Vergleich von zwei Gruppen
  - + Ausreichend die Mittelwerte zu vergleichen → gleicher Effekt wie in einer Regression
  - + Bei korrekter Randomisierung muss auf nichts kontrolliert werden (Pfeile zur endogenen Variable wurden gelöscht)

# Recap zum Experiment

## Das Experiment konzeptionell:

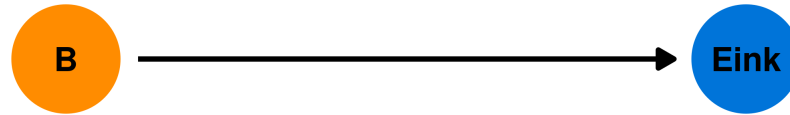
- + Erstellen einer Balancing Tabelle um die Randomisierung zu überprüfen
  - + Gruppen sollten vor dem Treatment nicht signifikant unterschiedlich voneinander sein
  - + Insbesondere hilfreich bei Feldexperimenten und zeitlich lange dauernden Experimenten in denen Attrition Problem sein kann
- + Koeffizienten einer Regression können auch visuell dargestellt werden

## Das Experiment inhaltlich:

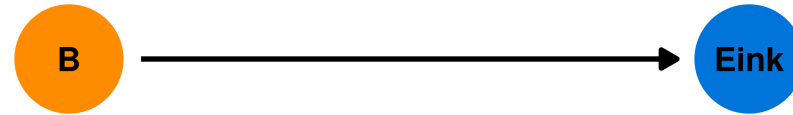
- + Eine psychologische Betreuung bei Wochenbettdepression hat einen signifikanten Effekt auf die Heilungswahrscheinlichkeit der Depressionen
- + Wir sehen auch langfristig (7 Jahre nach dem Experiment) positive Effekte der psychologischen Betreuung (obwohl diese Betreuung nur kurz vor und nach die Geburt stattfand)
- + Die ökonomische Teilhabe von Müttern erhöhte sich durch die psychologische Behandlung signifikant

In der nun folgenden Vorlesungseinheit wollen wir *kausale* Schlüsse aus Beobachtungsdaten ziehen!

Führt mehr Bildung zu höherem Einkommen?



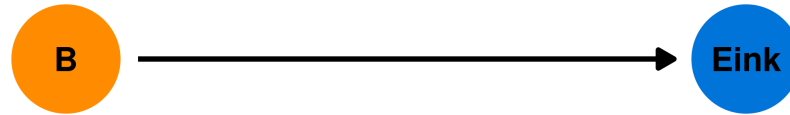
# Führt mehr Bildung zu höherem Einkommen?



Wir können dies mit folgender Gleichung darstellen:

$$\text{Einkommen}_i = \beta_0 + \beta_1 \text{Bildung}_i + \varepsilon_i$$

# Führt mehr Bildung zu höherem Einkommen?

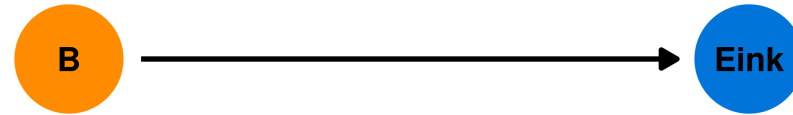


Wir können dies mit folgender Gleichung darstellen:

$$\text{Einkommen}_i = \beta_0 + \beta_1 \text{Bildung}_i + \varepsilon_i$$

Repräsentiert in dieser Regression  $\beta_1$  den *kausalen* Effekt von Bildung auf Einkommen?

# Führt mehr Bildung zu höherem Einkommen?



Wir können dies mit folgender Gleichung darstellen:

$$\text{Einkommen}_i = \beta_0 + \beta_1 \text{Bildung}_i + \varepsilon_i$$

Repräsentiert in dieser Regression  $\beta_1$  den *kausalen* Effekt von Bildung auf Einkommen?

Nein!

- + Ommitted variable Bias
- + Offene *Backdoors*

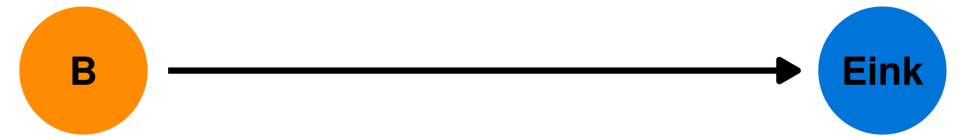
- + Endogenität



# Exogene und endogene Variablen

## Exogene Variablen:

- + Die erklärenden Variablen sind unabhängig von anderen Größen im Modell
- + In einem DAG wäre dies eine Variable in die keine Pfeile führen

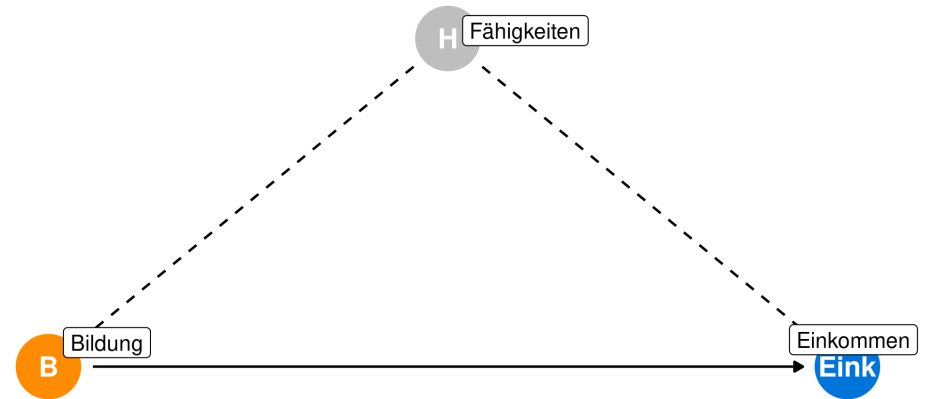


Bildung ist hier *exogen*: Keine Pfeile führen zur Bildung!

# Exogene und endogene Variablen

## Endogene Variablen:

- + Die erklärenden Variablen werden von anderen Größen im Modell (mit)bestimmt
- + In einem DAG wäre dies eine Variable in die Pfeile führen



Bildung ist hier *endogen*:  
Die Fähigkeiten (H) → Bildung (B)!

# Exogenität

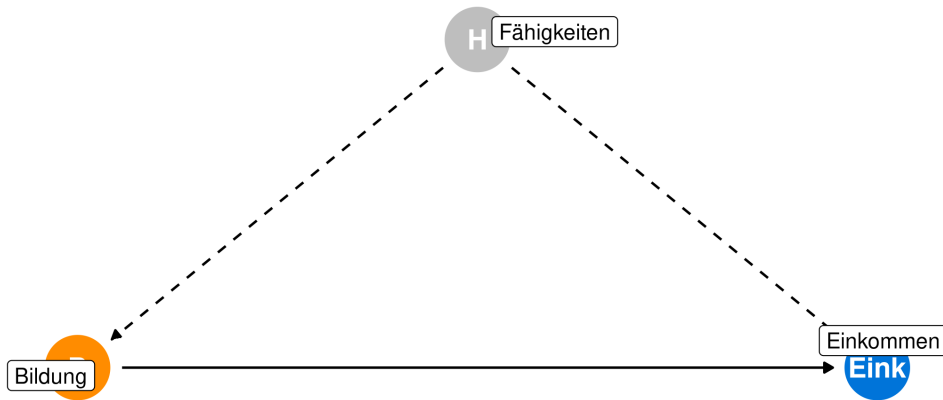
Um den *kausalen* Effekt von Bildung zu extrahieren müsste diese exogen sein.

Exogene Bildung bedeutet: Die Entscheidung für mehr Bildung sollte zufällig getroffen werden (oder zumindest nicht mit den ausgelassenen Variablen korreliert sein).

# Exogenität

Um den *kausalen* Effekt von Bildung zu extrahieren müsste diese exogen sein.

Exogene Bildung bedeutet: Die Entscheidung für mehr Bildung sollte zufällig getroffen werden (oder zumindest nicht mit den ausgelassenen Variablen korreliert sein).



Im DAG ist die Bildung abhängig von der (unbeobachtbaren) Variable "Fähigkeiten":

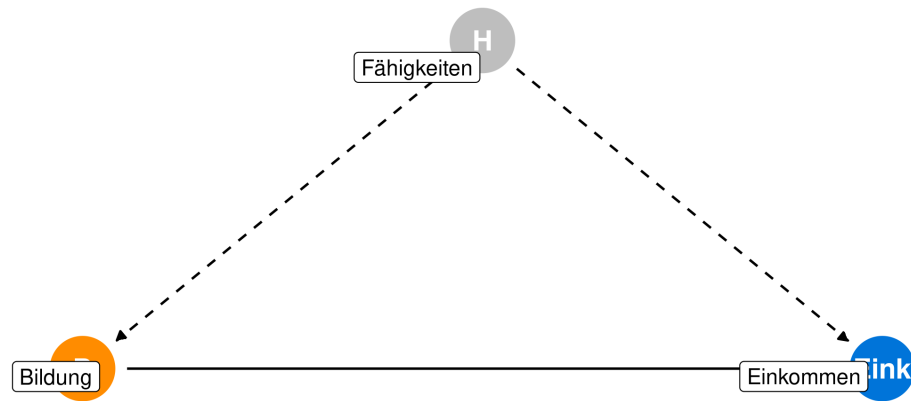
+ Bildung ist *nicht* exogen

Was tun?

# Exogenität

Um den *kausalen* Effekt von Bildung zu extrahieren müsste diese exogen sein.

Exogene Bildung bedeutet: Die Entscheidung für mehr Bildung sollte zufällig getroffen werden (oder zumindest nicht mit den ausgelassenen Variablen korreliert sein).



Im DAG ist die Bildung abhängig von der (unbeobachtbaren) Variable "Fähigkeiten":

+ Bildung ist *nicht* exogen

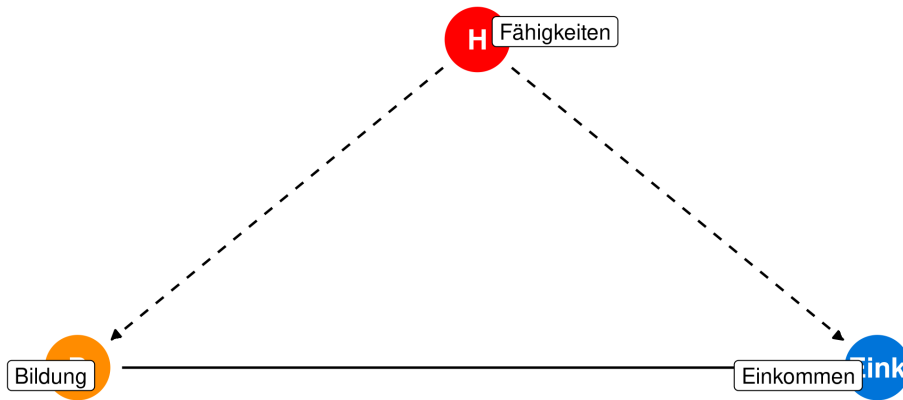
Was tun?

In der Theorie: Für die *backdoor* **Fähigkeiten kontrollieren!**

$$Einkommen_i = \beta_0 + \beta_1 * Bildung_i + \beta_2 * F\u00e4higkeiten_i + \varepsilon_i$$

# Exogenität

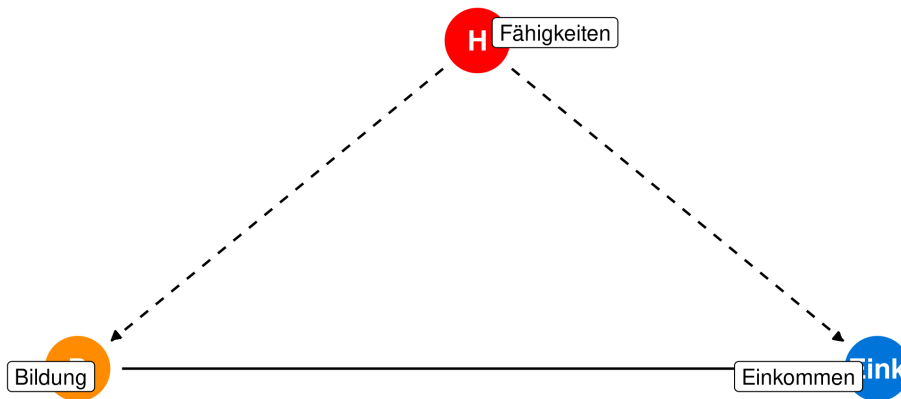
Leider sind die Fähigkeiten für uns nicht beobachtbar, somit können wir die *backdoor* nicht schließen!



$$\text{Einkommen}_i = \beta_0 + \beta_1 \text{Bildung}_i + \beta_2 \text{Fähigkeiten}_i + \varepsilon_i$$

# Exogenität

Leider sind die Fähigkeiten für uns nicht beobachtbar, somit können wir die *backdoor* nicht schließen!



$$\text{Einkommen}_i = \beta_0 + \beta_1 \text{Bildung}_i + \beta_2 \text{Fähigkeiten}_i + \varepsilon_i$$

Da die *backdoor* nicht geschlossen ist wandert der Einfluss der Fähigkeiten in den Fehlerterm:

Der Fehlerterm besteht nun aus:

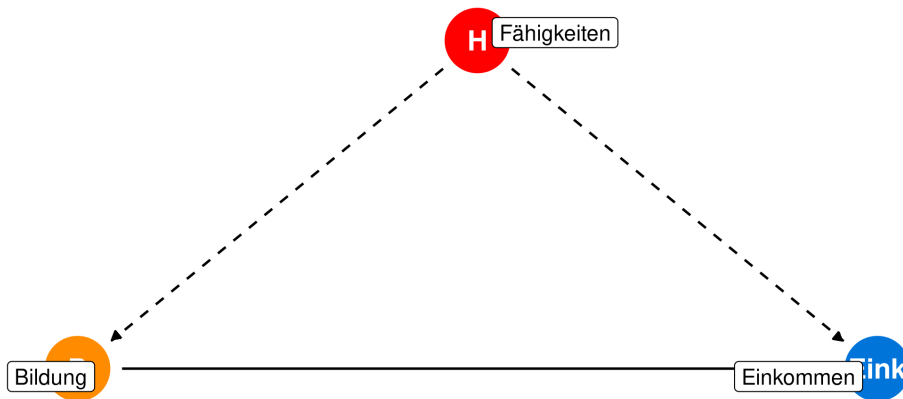
$$\eta_i = \beta_2 \text{Fähigkeiten}_i + \varepsilon_i$$

und damit ist die Bildung mit dem Fehlerterm  $\eta_i$  korreliert:

$$\text{Einkommen}_i = \beta_0 + \beta_1 \text{Bildung}_i + \eta_i$$

# Exogenität

Leider sind die Fähigkeiten für uns nicht beobachtbar, somit können wir die *backdoor* nicht schließen!



$$\text{Einkommen}_i = \beta_0 + \beta_1 \text{Bildung}_i + \beta_2 \text{Fähigkeiten}_i + \varepsilon_i$$

Da die *backdoor* nicht geschlossen ist wandert der Einfluss der Fähigkeiten in den Fehlerterm:

Der Fehlerterm besteht nun aus:

$$\eta_i = \beta_2 \text{Fähigkeiten}_i + \varepsilon_i$$

und damit ist die Bildung mit dem Fehlerterm  $\eta_i$  korreliert:

$$\text{Einkommen}_i = \beta_0 + \beta_1 \text{Bildung}_i + \eta_i$$

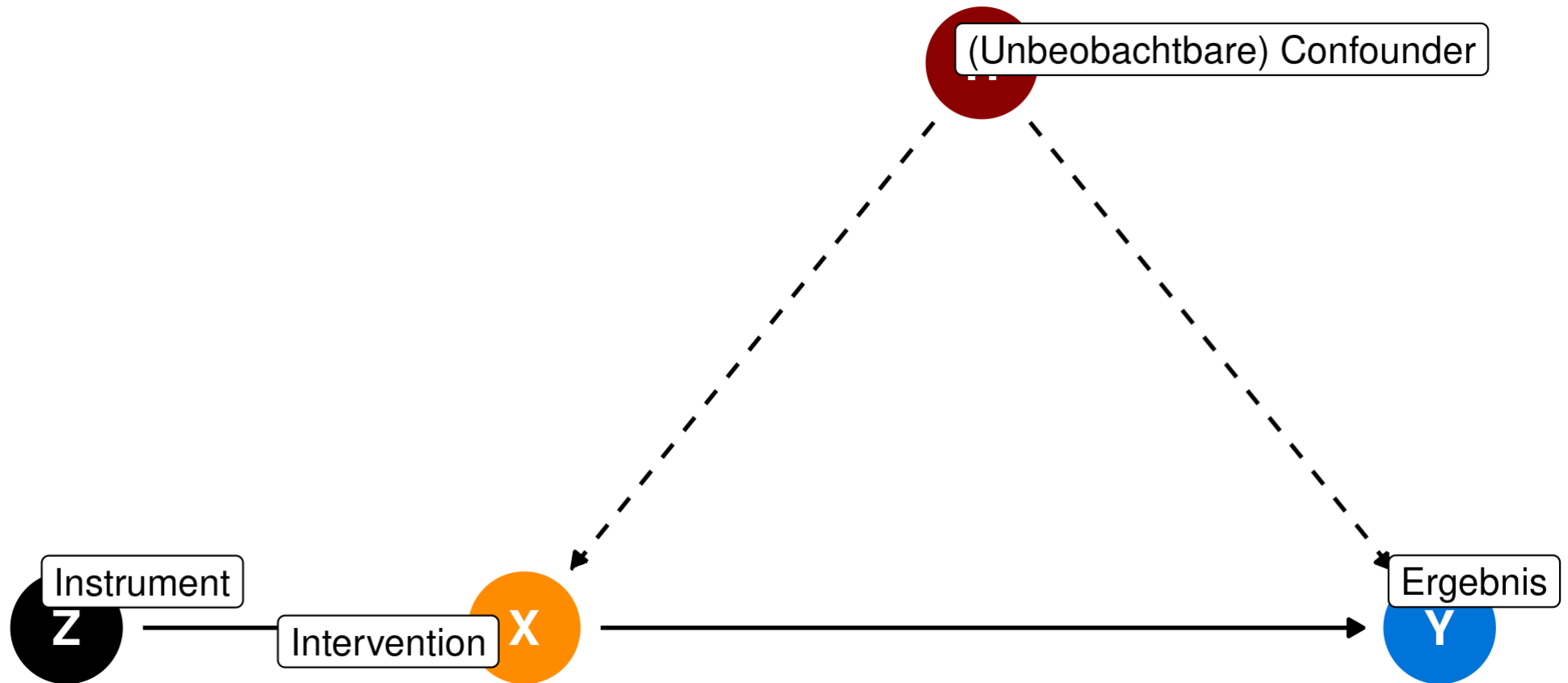
Wie können wir in diesem Fall den Einfluss der Bildung konsistent schätzen?



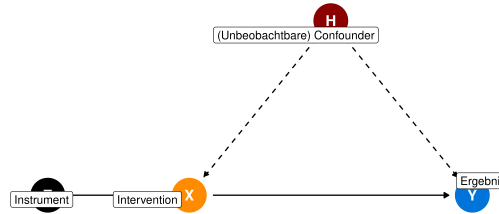
# Instrumentalvariablenschätzung

# Instrumentalvariablenschätzung

Was ist ein Instrument?



# Prinzip der Instrumentalvariablen-schätzung



**Hintergrund:**  $X$  sollte exogen sein um *kausal* interpretiert werden zu können

**Ziel der Instrumentalvariablen:** Exogene Variation von  $X$  finden, welche dann *kausal* interpretiert werden kann

Sie können sich dies vorstellen als Gegenteil davon auf eine Variable zu kontrollieren:

- + Wir erklären  $X$  und  $Y$  mit der Variablen  $Z$ , aber anstatt uns auf den Teil zu konzentrieren, welcher nicht durch  $Z$  erklärt werden kann, nehmen wir **nur den Anteil der durch  $Z$  erklärt wird!**
- + Anstatt zu sagen "du bist auf einer *backdoor*, ich schließe dich" sagen wir "du hast keine *backdoor*! Ich will, dass mein  $X$  genau so sein soll wie du! Ich nehme nur den Part von  $X$ , welcher von dir erklärt wird!"
- + Dadurch nutzen wir nur noch die exogene Variation in  $X$ , welche durch  $Z$  erklärt wird

# Prinzip der Instrumentalvariablenschätzung

## Folge:

- + Wir nutzen nicht mehr die komplette Information unseres Datensatzes, sondern nur noch einen Teil, d.h. wir benötigen mehr Beobachtungen um Effekte messen zu können
- + Diese ungenauere Schätzung des Effekts drückt sich in der Regression als größerer Standardfehler des Schätzers aus

## Schwierigkeiten:

- + Es kann sehr schwer sein zu argumentieren, warum ein Instrument keine *backdoor* hat
- + In wenigen Fällen haben wir tatsächlich randomisierte Instrumente
- + Manchmal müssen wir zusätzliche Kontrollvariablen aufnehmen um Instrumente zu rechtfertigen
- + Instrumentalvariablen müssen sauber ökonomisch begründet sein/werden
- + Instrumentalvariablen sind auf den ersten Blick oft seltsam und ergeben erst durch den Kontext Sinn

# Was macht ein Instrument aus?

## Relevanz:

- + Die Instrumentalvariable muss mit der/den endogenen Variable/n korreliert sein.

## Ausschließlichkeit:

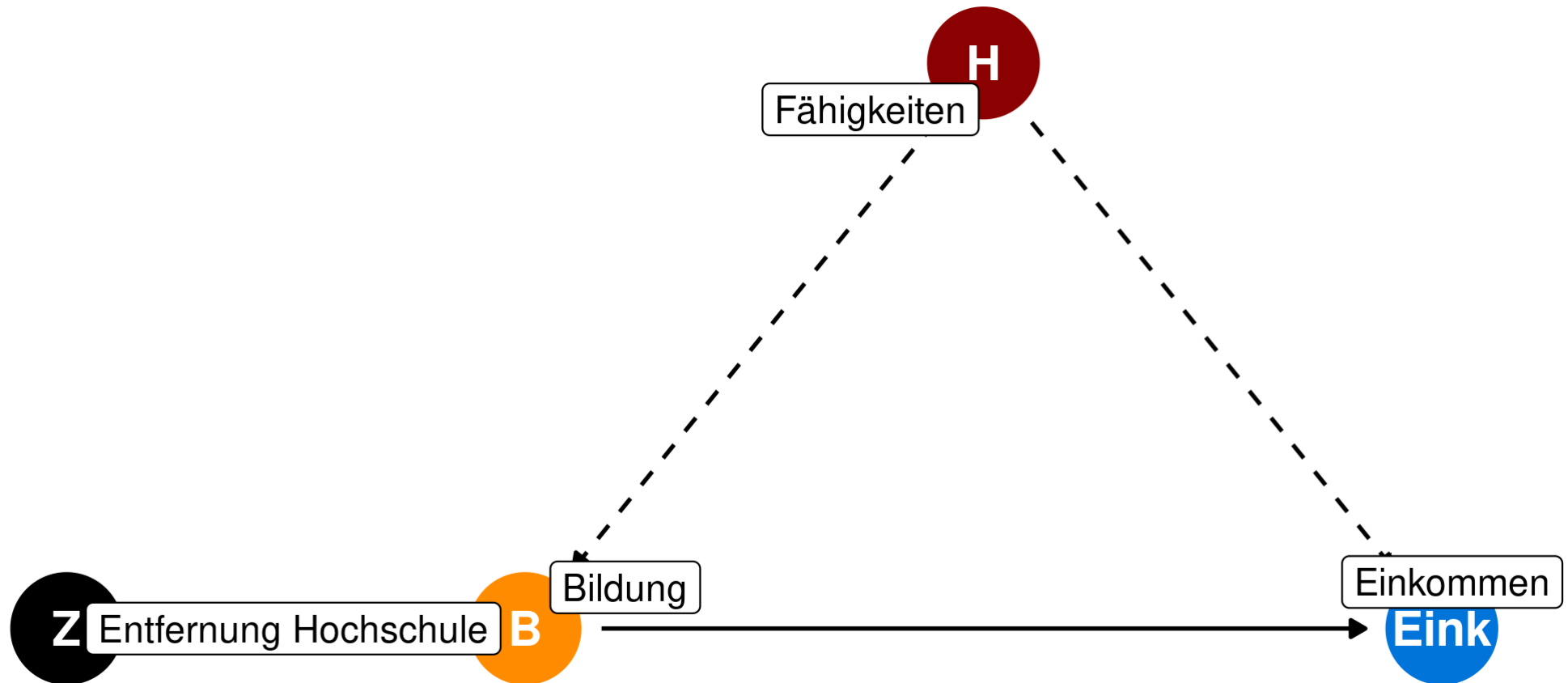
- + Die Instrumentalvariable beeinflusst die exogene Variable nicht direkt, sondern **ausschließlich** über die endogene Variable

## Exogenität:

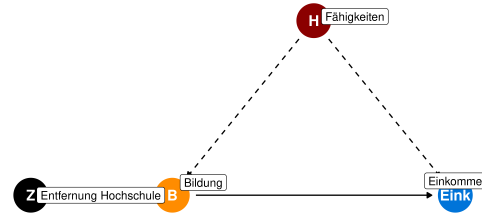
- + Die Instrumentalvariable ist nicht mit den ausgelassenen Variablen (omitted variables) korreliert

# Instrumentalvariablenschätzung (Beispiel)

Was wäre ein mögliches Instrument für den Effekt der Bildung auf das Einkommen



# Was macht ein Instrument aus?



Bedingungen für ein valides Instrument:

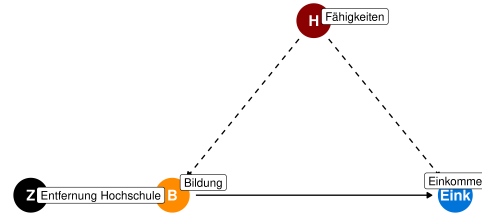
## Relevanz:

- + Die Instrumentalvariable muss mit der/den endogenen Variable/n korreliert sein.
- +  $Z \rightarrow B$ ;  $\text{Cor}(Z, B) \neq 0$

## Ausschließlichkeit:

- + Die Instrumentalvariable beeinflusst die exogene Variable nicht direkt, sondern **ausschließlich** über die endogene Variable
- +  $Z \rightarrow B \rightarrow \text{Eink}$ ;  $Z \not\rightarrow \text{Eink}$ ;  $\text{Cor}(Z, \text{Eink}|B) = 0$

# Was macht ein Instrument aus?



## Bedingungen für ein valides Instrument:

### Relevanz:

- + Die Instrumentalvariable muss mit der/den endogenen Variable/n korreliert sein.
- +  $Z \rightarrow B$ ;  $\text{Cor}(Z, B) \neq 0$

### Ausschließlichkeit:

- + Die Instrumentalvariable beeinflusst die exogene Variable nicht direkt, sondern **ausschließlich** über die endogene Variable
- +  $Z \rightarrow B \rightarrow \text{Eink}$ ;  $Z \not\rightarrow \text{Eink}$ ;  $\text{Cor}(Z, \text{Eink}|B) = 0$

## Wie können wir die Bedingungen testen?

### Relevanz:

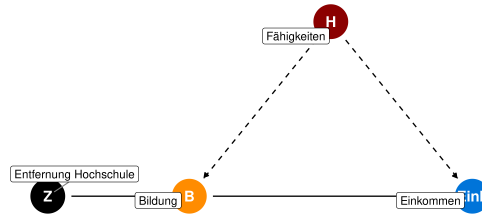
- + Kann mittels F-Test getestet werden

### Ausschließlichkeit:

- + Bei einem Instrument: Nur argumentativ, kann nicht getestet werden!
- + Bei mehreren Instrumenten: Kann mittels Sargan-Hansen-Test getestet werden (allerdings ist dieser nicht besonders zuverlässig).



# Was macht ein Instrument aus?

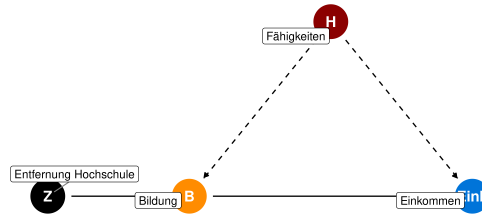


Bedingungen für ein valides Instrument:

## Exogenität:

- + Die Instrumentalvariable ist nicht mit den ausgelassenen Variablen (omitted variables) korreliert
- + Das bedeutet: Die Instrumentalvariable ist nicht mit dem Fehlerterm korreliert
- +  $H \not\rightarrow Z; \text{Cor}(Z, H) = 0$

# Was macht ein Instrument aus?



Bedingungen für ein valides Instrument:

**Exogenität:**

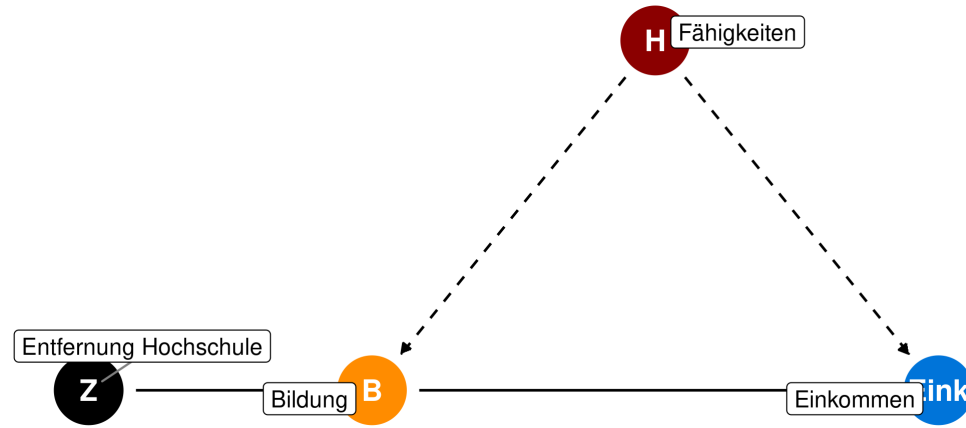
- + Die Instrumentalvariable ist nicht mit den ausgelassenen Variablen (omitted variables) korreliert
- + Das bedeutet: Die Instrumentalvariable ist nicht mit dem Fehlerterm korreliert
- +  $H \not\rightarrow Z; \text{Cor}(Z, H) = 0$

Wie können wir die Bedingungen testen?

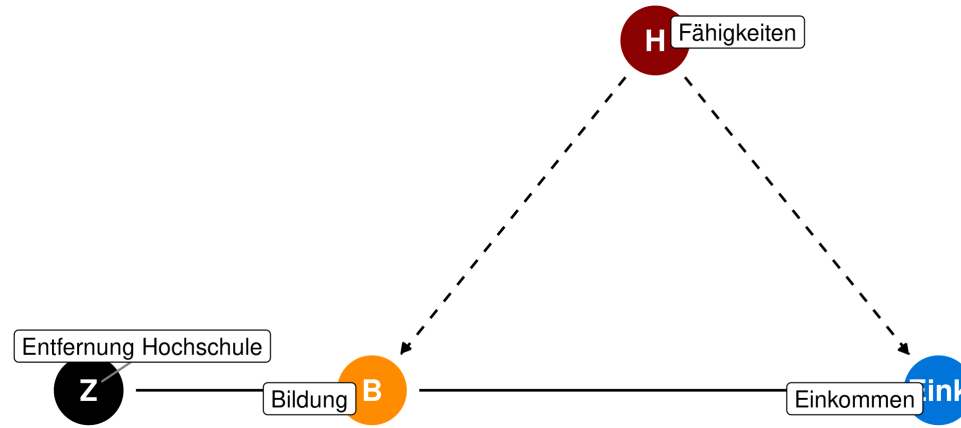
**Exogenität:**

- + Kann nicht getestet werden. Argumentativ auf ökonomischer Basis.

# Instrumentalvariablenschätzung

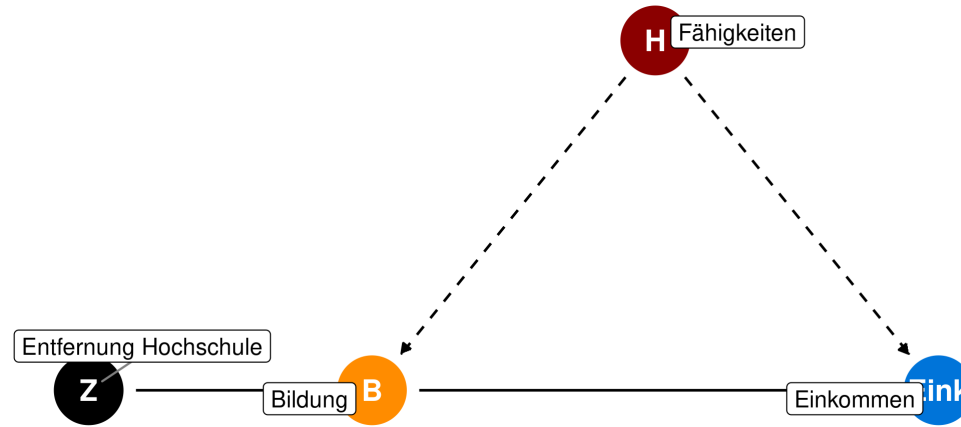


# Instrumentalvariablenschätzung



**Relevanz:** Durch die Nähe zur Universität kann Bildung zu niedrigeren Kosten erworben werden, d.h. es ist wahrscheinlicher, dass diese Personen mehr in Bildung investieren ✓

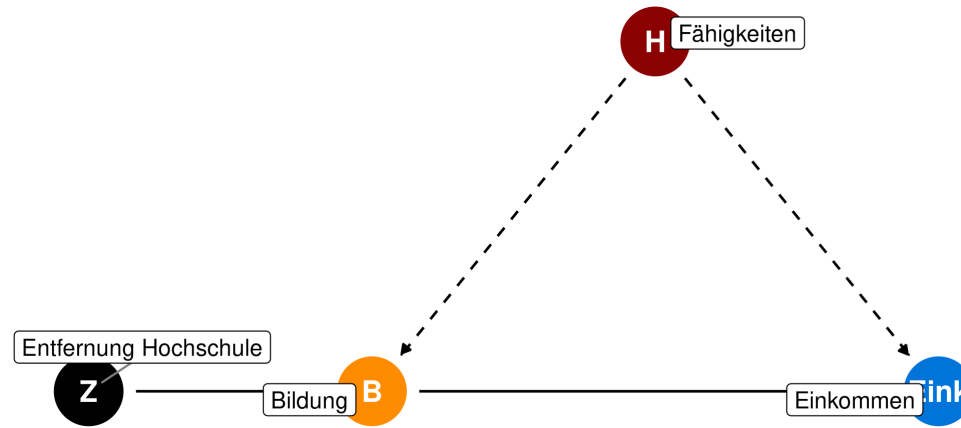
# Instrumentalvariablenschätzung



**Relevanz:** Durch die Nähe zur Universität kann Bildung zu niedrigeren Kosten erworben werden, d.h. es ist wahrscheinlicher, dass diese Personen mehr in Bildung investieren ✓

**Ausschließlichkeit:** Die Nähe zur Universität hat keinen direkten Einfluss auf das Einkommen ✓

# Instrumentalvariablenschätzung

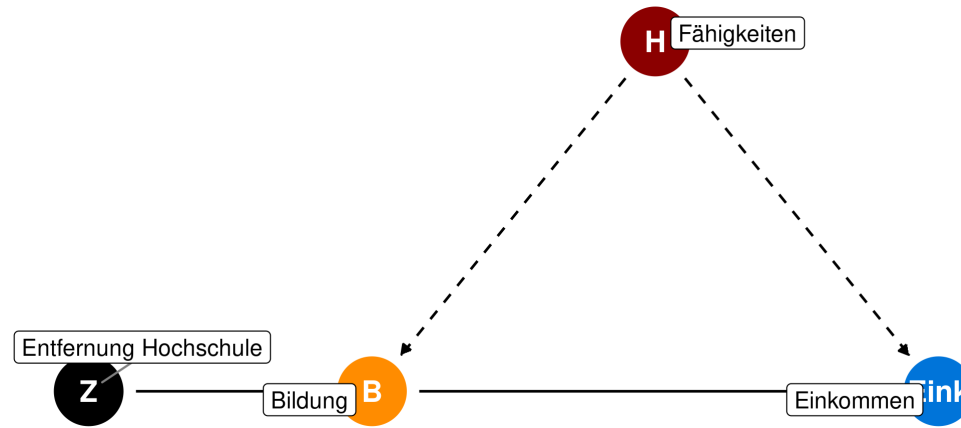


**Relevanz:** Durch die Nähe zur Universität kann Bildung zu niedrigeren Kosten erworben werden, d.h. es ist wahrscheinlicher, dass diese Personen mehr in Bildung investieren ✓

**Ausschließlichkeit:** Die Nähe zur Universität hat keinen direkten Einfluss auf das Einkommen ✓

**Exogenität:** Individuelle Fähigkeiten sind nicht abhängig von der Nähe zu einer Universität ✓

# Instrumentalvariablenschätzung



**Relevanz:** Durch die Nähe zur Universität kann Bildung zu niedrigeren Kosten erworben werden, d.h. es ist wahrscheinlicher, dass diese Personen mehr in Bildung investieren ✓

**Ausschließlichkeit:** Die Nähe zur Universität hat keinen direkten Einfluss auf das Einkommen ✓

**Exogenität:** Individuelle Fähigkeiten sind nicht abhängig von der Nähe zu einer Universität ✓

**Ausschließlichkeit und Exogenität sind sehr schwer zu zeigen/argumentieren!**

Am Besten sind hier tatsächlich randomisierte Zuteilungen als Instrument zu nehmen.

# Instrumentalvariablenschätzung - empirische Analyse

Im folgenden wollen wir die Nähe des Wohnorts zur einer Universität als Instrumentalvariable verwenden. Hierfür nutzen wir die Daten von Card (1995), welche im `wooldridge` Paket in R verfügbar sind:

Card, David. 1995. "Aspects of Labour Economics: Essays in Honour of John Vanderkamp." In. University of Toronto Press.

[NBER Working Paper finden Sie hier](#)



# Instrumentalvariablenschätzung - empirische Analyse

Im folgenden wollen wir die Nähe des Wohnorts zur einer Universität als Instrumentalvariable verwenden. Hierfür nutzen wir die Daten von Card (1995), welche im `wooldridge` Paket in R verfügbar sind:

Card, David. 1995. "Aspects of Labour Economics: Essays in Honour of John Vanderkamp." In. University of Toronto Press.

[NBER Working Paper finden Sie hier](#)

Insbesondere nutzen wir folgende Variablen:

Variablenname	Erklärung
lwage	Jährliches Einkommen (logarithmiert)
educ	Bildungsjahre
nearc4	Lebt nahe einer Universität (=1) oder weiter entfernt (=0)

Die Daten stammen aus einer Umfrage in den USA im Jahr 1976 mit 3010 Männern.

# Testen der Relevanz

Zuerst sollten wir testen ob unser Instrument relevant ist.

Konkret: Hat das Instrument einen Einfluss auf die endogene Variable ( $Z \rightarrow B$ )

# Testen der Relevanz

Zuerst sollten wir testen ob unser Instrument relevant ist.

Konkret: Hat das Instrument einen Einfluss auf die endogene Variable ( $Z \rightarrow B$ )

```
# A tibble: 2 × 5
  term      estimate std.error statistic  p.value
<chr>    <dbl>    <dbl>    <dbl>    <dbl>
1 (Intercept)  12.7    0.0856    148.    0
2 nearc4      0.829   0.104     7.99 1.84e-15
```

Wir sehen einen signifikanten Effekt von der Entfernung zur Hochschule (*nearc4*) auf die Bildung (*educ*) → **Relevant**

Weiterhin sollten wir prüfen, ob das Instrument valide ist, dies prüfen wir mit der F-Statistik.

# Testen der Relevanz

Zuerst sollten wir testen ob unser Instrument relevant ist.

Konkret: Hat das Instrument einen Einfluss auf die endogene Variable ( $Z \rightarrow B$ )

```
# A tibble: 2 × 5
  term      estimate std.error statistic  p.value
<chr>    <dbl>    <dbl>    <dbl>    <dbl>
1 (Intercept) 12.7      0.0856    148.     0
2 nearc4      0.829     0.104     7.99 1.84e-15
```

Wir sehen einen signifikanten Effekt von der Entfernung zur Hochschule (*nearc4*) auf die Bildung (*educ*) → Relevant

Weiterhin sollten wir prüfen, ob das Instrument valide ist, dies prüfen wir mit der F-Statistik.

```
glance(first_stage_basic)
```

```
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic p.value  df logLik  AIC  BIC
  <dbl>    <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.0208 0.0205 2.65    63.9 1.84e-15 1 -7203. 14411. 14429.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Die F-Statistik (*statistic*) in unserem Modell liegt bei 63,9.

Die F-Statistik in der *First-Stage* sollten größer als 10 sein. Als Faustregel gilt: Bei Werten unter 10 haben wir es mit einem *schwachen* Instrument zu tun.

# Ausschließlichkeit und Exogenität

Beeinflusst die Nähe zur Universität das Einkommen **ausschließlich** über die Bildung?

Oder gibt es außerdem einen direkten Effekt der "Nähe zur Universität" auf das Einkommen?

Was wären die potentiellen Kanäle von Z -> Eink?

# Ausschließlichkeit und Exogenität

Beeinflusst die Nähe zur Universität das Einkommen **ausschließlich** über die Bildung?

Oder gibt es außerdem einen direkten Effekt der "Nähe zur Universität" auf das Einkommen?

Was wären die potentiellen Kanäle von Z -> Eink?

- + Wie steht es bspw. um die regionalen Gegebenheiten?
  - + Eine dünn besiedelte Region hat weniger Universitäten, aber auch weniger potentielle Arbeitgeber und dadurch potentiell geringere Einkommen

# Ausschließlichkeit und Exogenität

Beeinflusst die Nähe zur Universität das Einkommen **ausschließlich** über die Bildung?

Oder gibt es außerdem einen direkten Effekt der "Nähe zur Universität" auf das Einkommen?

Was wären die potentiellen Kanäle von Z -> Eink?

- + Wie steht es bspw. um die regionalen Gegebenheiten?
  - + Eine dünn besiedelte Region hat weniger Universitäten, aber auch weniger potentielle Arbeitgeber und dadurch potentiell geringere Einkommen

Wir können die Exogenitätsannahme unserer Instrumentalvariable abschwächen, indem wir zusätzliche Kontrollvariablen in unser Modell aufnehmen

- + Dadurch ist unsere Exogenitätsannahme der Instrumentalvariable *bedingt* auf die Kontrollvariablen

# Ausschließlichkeit und Exogenität

Wir nehmen die folgenden Kontrollvariablen mit in unser Modell auf:

Variablenname	Erklärung
smsa	Lebt in einer dicht besiedelten Region (=1) oder nicht (=0)
exper	Erfahrung
expersq	Erfahrung <sup>2</sup>
south	Lebt im Süden der USA (=1), oder nicht (=0)

$$Einkommen_i = \beta_0 + \beta_1 Bildung_i + \beta_2 X_{Kontrollvariablen_i} + \eta_i$$



# Ausschließlichkeit und Exogenität

Relevanz der Instrumentalvariable *bedingt* auf die Kontrollvariablen:

```
first_stage_c <- lm(educ ~ nearc4 + smsa + ex  
tidy(first_stage_c)
```

```
# A tibble: 6 × 5  
  term      estimate std.error statistic  
  <chr>      <dbl>    <dbl>    <dbl>  
1 (Intercept) 16.7      0.180     92.7  0  
2 nearc4      0.346     0.0842     4.11  4  
3 smsa        0.364     0.0866     4.21  2  
4 exper      -0.426     0.0344    -12.4  1  
5 expersq     0.000977  0.00168    0.580  5  
6 south      -0.583     0.0764    -7.63  3
```

```
glance(first_stage_c)
```

```
# A tibble: 1 × 12  
  r.squared adj.r.squared sigma statistic p.v  
  <dbl>      <dbl> <dbl>    <dbl>    <dbl> <dbl>  
1 0.452      0.452 1.98     496.     0.000977  
# i 3 more variables: deviance <dbl>, df.resi
```

**Hier erhalten wir eine F-Statistik von 496!** Diese F-Statistik ist jedoch nicht ganz korrekt. Hier empfiehlt es sich immer auf den "Weak instruments" Test zu achten (wir gehen etwas später darauf ein).

# Two-stage least squares (2SLS) (händisch)

**Ziel:** Den *exogenen Teil* der Bildung finden mit Hilfe des Instruments und diesen *exogenen Teil* für die Schätzung nutzen:

## First stage

$$\hat{Bildung}_i = \gamma_0 + \gamma_1 * Nähe.zur.Uni_i + \gamma_2 * X_{Kontrollvariablen_i} + \nu_i$$

$\hat{Bildung}$  ist der exogene Part der Bildung, unabhängig von den Fähigkeiten (und anderen Einflussgrößen)!

## Second stage

$$Einkommen = \beta_0 + \beta_1 * \hat{Bildung}_i + \beta_2 * X_{Kontrollvariablen_i} + \varepsilon_i$$

# Two-stage least squares (2SLS) (händisch)

**Ziel:** Den *exogenen Teil* der Bildung finden mit Hilfe des Instruments und diesen *exogenen Teil* für die Schätzung nutzen:

## First stage

$$\hat{Bildung}_i = \gamma_0 + \gamma_1 * Nähe.zur.Uni_i + \gamma_2 * X_{Kontrollvariablen_i} + \nu_i$$

$\hat{Bildung}$  ist der exogene Part der Bildung, unabhängig von den Fähigkeiten (und anderen Einflussgrößen)!

## Second stage

$$Einkommen = \beta_0 + \beta_1 * \hat{Bildung}_i + \beta_2 * X_{Kontrollvariablen_i} + \varepsilon_i$$

Erinnern Sie sich noch an unsere Berechnung der **Relevanz** des Instruments ( `first_stage_c` )?

```
first_stage_c <- lm(educ ~ nearc4 + smsa + exper + expersq + south, data = card)
```

Das ist die Regression, welche wir hier **First stage** nennen!

Wie berechnen wir die **Second stage**?

# Two-stage least squares (2SLS) (händisch)

Wir nutzen die gefitteten Werte aus unserer **first stage** und fügen diese als **bildung\_hat** unserem Datensatz `card` hinzu:

```
first_stage_c <- lm(educ ~ nearc4 + smsa + exper + expersq + south, data = card)

card <- card %>%
  mutate(bildung_hat = first_stage_c$fitted.values)

card %>%
  select(educ, nearc4, smsa, exper, south, bildung_hat) %>%
  head()
```

	educ	nearc4	smsa	exper	south	bildung_hat
1	7	0	1	16	0	10.48202
2	12	0	1	9	0	13.29188
3	12	0	1	16	0	10.48202
4	11	1	1	10	0	13.23025
5	12	1	1	16	0	10.82766
6	12	1	1	8	0	14.04675

# Two-stage least squares (2SLS) (händisch)

Anschließend berechnen wir die **second stage** mit diesen gefitteten Werten für jede Person ( $Bildung_i$ ):

```
second_stage <- lm(lwage ~ bildung_hat + smsa + exper + expersq + south, data = card)
tidy(second_stage)
```

```
# A tibble: 6 × 5
  term          estimate std.error statistic  p.value
<chr>         <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  3.70      0.858      4.32 1.63e- 5
2 bildung_hat  0.135     0.0508     2.66 7.78e- 3
3 smsa         0.125     0.0298     4.20 2.75e- 5
4 exper        0.107     0.0228     4.68 3.01e- 6
5 expersq      -0.00226  0.000355   -6.35 2.45e-10
6 south        -0.141     0.0360     -3.92 9.06e- 5
```

# Two-stage least squares (2SLS) (automatisch)

In R gibt es im Paket `AER` die Möglichkeit eine solche 2SLS automatisch durchzuführen

## Vorteil:

- + Schneller als von Hand
- + Die Standardfehler werden direkt korrigiert!
- + Umfangreiche Diagnostik nach der Regression

## Syntax:

```
ivreg(Y ~ 2nd stage | first stage, data)
```

```
iv_2sls <- ivreg(lwage ~ educ + smsa + exper  
                 nearc4 + smsa + exp  
                 data = card)  
  
tidy(iv_2sls)
```

```
# A tibble: 6 × 5  
  term      estimate std.error statistic p  
  <chr>      <dbl>      <dbl>      <dbl> <dbl>  
1 (Intercept)  3.70        0.821        4.51 6.  
2 educ         0.135       0.0487        2.78 5.  
3 smsa         0.125       0.0285        4.39 1.  
4 exper        0.107       0.0218        4.89 1.  
5 expersq     -0.00226    0.000340     -6.64 3.  
6 south       -0.141      0.0344       -4.10 4.
```

# Two-stage least squares (2SLS) (händisch)

Erhalten wir unterschiedliche Werte für die Regression mittels OLS vs. IV (händisch/automatisch)?

# Two-stage least squares (2SLS) (händisch)

Erhalten wir unterschiedliche Werte für die Regression mittels OLS vs. IV (händisch/automatisch)?

Log(Einkommen) auf Bildung regressiert

	OLS	2SLS händisch	2SLS automatisch
	(1)	(2)	(3)
Bildung	0.082 [0.073, 0.091]		0.135 [0.010, 0.261]
Bildung_Dach		0.135 [0.004, 0.266]	
Num.Obs.	3010	3010	3010
R2	0.263	0.132	0.205
R2 Adj.	0.262	0.130	0.204

- + Die Bildung hat einen deutlich stärkeren Einfluss auf das Einkommen mit der IV Regression
  - + In der OLS Regression steigt das Einkommen um 8,2% für jedes zusätzliche Jahr an Schulbildung
  - + In der 2SLS Regression hat die Bildung einen deutlich höheren Einfluss! (13,5%)
- + Dieses Ergebnis überrascht:
  - + Die Fähigkeiten einer Person sollten ihr Einkommen positiv beeinflussen
  - + Die Fähigkeiten einer Person sollten ihre Bildungsentscheidung positiv beeinflussen
  - + Problem der ausgelassenen Variablen (omitted variable bias), d.h. Endogenität!



# OLS vs. IV

Eigentlich würden wir erwarten, dass der Effekt der Bildung auf das Einkommen **kleiner** ist, wenn wir auf die Fähigkeiten kontrollieren.

Warum erhalten wir größere anstatt kleinere Werte für die Bildung im IV?

# OLS vs. IV

Eigentlich würden wir erwarten, dass der Effekt der Bildung auf das Einkommen **kleiner** ist, wenn wir auf die Fähigkeiten kontrollieren.

Warum erhalten wir größere anstatt kleinere Werte für die Bildung im IV?

**Messfehler:**

- + Eventuell wird die Bildung nicht richtig erfasst (falsche Angaben der Personen)
- + Dadurch wird der Effekt der Bildung in der OLS Regression unterschätzt
- + Persönliche Einschätzung: Eher unwahrscheinlich

# OLS vs. IV

Für welche Personen ist es relevant das eine Universität nahe dem Wohnort liegt um in mehr Bildung zu investieren?

Manche Schüler werden immer zur Uni gehen, egal ob gerade eine in der Nähe ist, andere werden nie gehen, auch wenn eine Uni da wäre. Jedoch könnte auch eine Gruppe von Schülern vorhanden sein, welche nur zur Uni gehen, wenn sie in der Nähe einer Uni wohnen.

Welche Schüler sind dies?

# OLS vs. IV

Für welche Personen ist es relevant das eine Universität nahe dem Wohnort liegt um in mehr Bildung zu investieren?

Manche Schüler werden immer zur Uni gehen, egal ob gerade eine in der Nähe ist, andere werden nie gehen, auch wenn eine Uni da wäre. Jedoch könnte auch eine Gruppe von Schülern vorhanden sein, welche nur zur Uni gehen, wenn sie in der Nähe einer Uni wohnen.

Welche Schüler sind dies?

+ Schüler die zuhause wohnen, dadurch pendeln können und keine Miete zahlen müssen!

Es könnte gut sein, dass der Effekt von Bildung auf Einkommen für diese Personengruppe anders ist, als für die Gesamtpopulation.

Wenn dem so ist, dann messen wir nicht den allgemeinen Effekt der Bildung auf das Einkommen (den Average Treatment Effect (ATE)), sondern nur den Effekt der Bildung für diese Gruppe (den Local Average Treatment Effect (LATE)).

# Schwache Instrumente ( *weak instruments* )

Was passiert, wenn ein Instrument nicht (oder nur marginal) relevant ist?

- + Relevanz unserer Instrumente testen
- + Mittels `ivreg` direkt möglich. (`diagnostics = TRUE`)

Wir führen ein zweites Instrument `nearc2` ein (in der Nähe ist ein 2-Jahres College)

- + F-Statistik (gemeinsame Signifikanz der Instrumente) in der *first-stage* ist mit 8,8 unter 10
- + Besser das schwächste Instrument wieder zu entfernen

```
weak_iv <- ivreg(lwage ~ educ + smsa + exper + expersq + south |
                 nearc4 + nearc2 + smsa + exper + expersq +
                 data = card)

summary(weak_iv, diagnostics = T)
```

```
Call:
ivreg(formula = lwage ~ educ + smsa + exper + expersq + south |
       nearc4 + nearc2 + smsa + exper + expersq + south, data = card)

Residuals:
    Min       1Q   Median       3Q      Max
-1.95566 -0.24783  0.02474  0.26737  1.46939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.4675334   0.8199110   4.229 2.42e-05 ***
educ         0.1494138   0.0485931   3.075 0.002125 **
smsa         0.1182924   0.0286581   4.128 3.76e-05 ***
exper        0.1127343   0.0218558   5.158 2.66e-07 ***
expersq      -0.0022691   0.0003468  -6.543 7.07e-11 ***
south       -0.1320361   0.0345091  -3.826 0.000133 ***

Diagnostic tests:
              df1  df2  statistic  p-value
Weak instruments  2 3003    8.811 0.000153 ***
Wu-Hausman       1 3003    2.206 0.137556
Sargan           1   NA    1.820 0.177328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4044 on 3004 degrees of freedom
Multiple R-Squared: 0.171, Adjusted R-squared: 0.1696
Wald test: 95.98 on 5 and 3004 DF, p-value: < 2.2e-16
```

# Vorgehensweise bei der Instrumentalvariablen-schätzung

Sie sollten sich folgende Fragen stellen:

- + Ist das Instrument relevant?
  - + Instrument mit der Intervention korreliert; *first-stage* F-Statistik (bzw. "weak instruments test") > 10
- + Erfüllt das Instrument das Ausschließlichkeitskriterium
  - + Das Instrument beeinflusst das Ergebnis **ausschließlich** durch die Intervention
- + Ist das Instrument exogen?
  - + Keine Pfeile zum Instrument im DAG
- + Anwenden von 2-stage least squares (2SLS)
  - + Nutzen des R-Pakets `ivreg()`