LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Querschnittsdatenanalyse:

## Angewandte Regressionsanalyse mit STATA

## (vorläufige Version)

Prof. Dr. Josef Brüderl

LMU München

Wintersemester 2018/19

# Inhalt

# Lernziele

- Kenntnis verschiedener Querschnitts-Regressionsmodelle
  - Keine Herleitungen, keine Berechnungen per Hand
  - Sondern ein anwendungsorientierter Überblick
- Darstellung und Interpretation der Regressionsergebnisse
  - Insbesondere die graphische Darstellung der Regressionsergebnisse wird betont („Das Zeitalter der Regressionstabelle ist vorbei“)
- Interpretation von Interaktionen
  - Selbst im linearen Modell wird hier viel falsch gemacht
  - Bei nicht-linearen Modellen ist es noch komplizierter
- Praktische Umsetzung der Regressionsmodelle mit STATA
  - Die grundlegenden STATA-Befehle sind in den Folien enthalten
  - Zusätzlich kann man anhand der begleitenden STATA Do-Files die Berechnungen nachvollziehen

# ALLBUS 2002

- Bevölkerungsumfrage alle 2 Jahre seit 1980 (N ~ 3.000)
  - Von GESIS als Service für die Sozialforschung
  - Trenddaten
- ALLBUS 2002 (N=2.820)
  - Einwohnermelderegisterstichprobe
  - Ostdeutsche überrepräsentiert
  - GG: alle deutschsprachigen Personen über 18, wohnhaft in D in Privathaushalten
  - Ausschöpfung: 47%
  - Mündliches Interview (CAPI)
  - Infos: http://www.gesis.org/allbus

# ALLBUS 2002: Datenaufbereitung

- Für den Kurs wurden einige Variablen aufbereitet
  - Abgespeichert im Datensatz: **`AllbReg.dta`**

Daten: ALLBUS 2002
Do-File: 0 Datenaufbereitung.do

- Abhängige Variablen
  - `eink`: monatliches Nettoeinkommen in Euro
  - `rechts`: Links-Rechts Selbsteinstufung (Skala 1-10)
  - `arblos`: Arbeitslosigkeit in den letzten 10 Jahren (0=nein, 1=ja)
  - `partei`: Wahlabsicht (CDU, SPD, FDP, Grüne, PDS)
  - `oecdeink`: Nettoäquivalenzeinkommen in Euro
- Unabhängige Variablen
  - `bild`: schulische und berufliche Bildung in Jahren
  - `alter`: Alter in Jahren
  - `exp`: Berufserfahrung (Berechnung: alter - bild - 6)
  - `prestv`: Berufsprestige des Vaters (Magnitude-Skala)
  - `frau`: Dummy für Frau
  - `ost`: Dummy für Ostdeutscher
  - `beruf`: berufliche Stellung (Arbeiter, Angest., Beamter, Selbst.)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

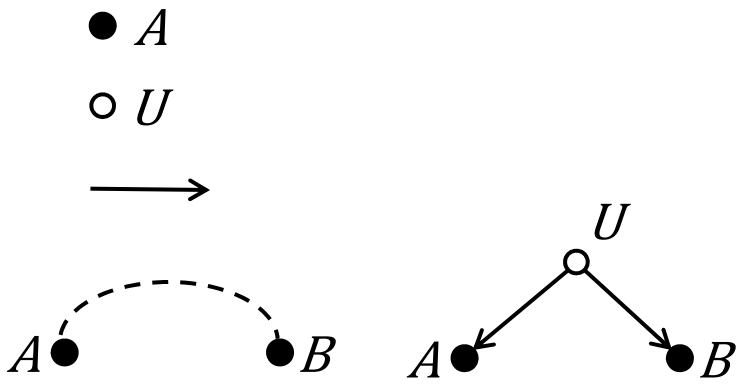
# Kapitel 1:
# Kausalität in den Sozialwissenschaften

# Ziele empirischer Sozialforschung

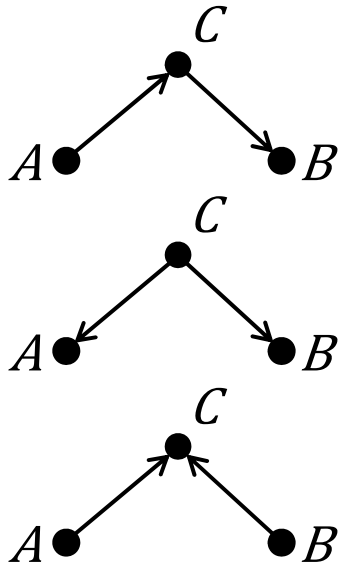
- Beschreibung (Deskription)
  - Die genaue Beschreibung der sozialen Welt kommt zuerst
    - „It is a capital mistake to theorize in advance of the facts“ (Warnung von Sherlock Holmes an Dr. Watson)
- Erklärung
  - Hat man die Fakten geklärt, kann man nach den Ursachen fragen
    - Dies ist die Suche nach Kausaleffekten (Kausalanalyse)
- Politikberatung
  - Hat man Fakten und kausale Zusammenhänge, so kann man politische Maßnahmen empfehlen
    - Dazu benötigt man aber Vorstellungen über den Soll-Zustand (politisches Ziel). Da dies normativ ist, nicht Teil der Wissenschaft.
  - Bsp.: Anstieg der Einkommensungleichheit
    - In welchen Einkommensgruppen verändert sich was?
    - Dann kann man nach den Ursachen fragen
    - Dann kann man politische Maßnahmen empfehlen

# Kausaldiagramme (DAGs)

- Knoten: beobachtete Variable
- Offener Knoten: latente Variable
- Gerichtete Kante: **Kausalbeziehung**
- Bidirektionale Kante: **Assoziation** durch gemeinsame latente Ursache („confounder“)




- Die drei grundlegenden Muster der Kausalbeziehung dreier Variablen
  - Intervenierende Variable
  - Konfundierende Variable
  - Collider Variable



# Was ist Wissenschaft?

- Ziel: Wissenschaft will Wissen schaffen
  - Nicht nur moralisieren und/oder politisieren
- Methode I:

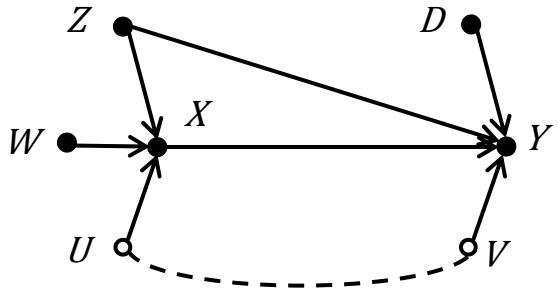
**Lee Smolin, Professor für theoretische Physik**

"[W]ir müssen darauf bestehen, sie [die Ideen] testen – oder falsifizieren – zu können. Darauf beruht der Fortschritt der Wissenschaft in den vergangenen 400 Jahren. Wenn man eine theoretische Struktur hat, die nichts erklärt und nichts vorhersagt, hört man auf, Wissenschaft zu betreiben. Dann haben wir es mit der Gefahr von nicht überprüfbaren Theorien [...] zu tun.

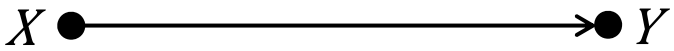
- Methode II: Konsequentes Anzweifeln aller Ergebnisse (auch der eigenen!)
  (aus: Richtlinien der LMU München zur Selbstkontrolle in der Wissenschaft)

# Theorie

- Kausalanalysen basieren auf Theorien
  - Eine “Theorie” ist eine Menge von miteinander verknüpften und logisch konsistenten Aussagen, von denen eine nichtleere Teilmenge empirisch prüfbare Aussagen (Hypothesen) sind




- Das Geschäft der (kausalanalytischen) Sozialforschung ist die empirische Überprüfung der Gültigkeit von (aus Theorien abgeleiteten) Hypothesen
  - Z.B.:



# Exkurs: Mängel soziologischer Theorien

- Viele soziologische Theorien sind allerdings wissenschaftlich unfruchtbar
  - Oft ist unklar, welche theoretischen Konzepte eigentlich eine wesentliche Rolle spielen
  - Viele Theorien sind zu ungenau, um falsch zu sein
  - Viele Theorien lassen keine Herleitung von empirisch überprüfbaren Hypothesen zu, weil sie u.a. Konstrukte verwenden,
    - welche keine Entsprechung in der Realität haben
    - und/oder kaum operationalisier- und messbar sind

# Empirische Überprüfung I

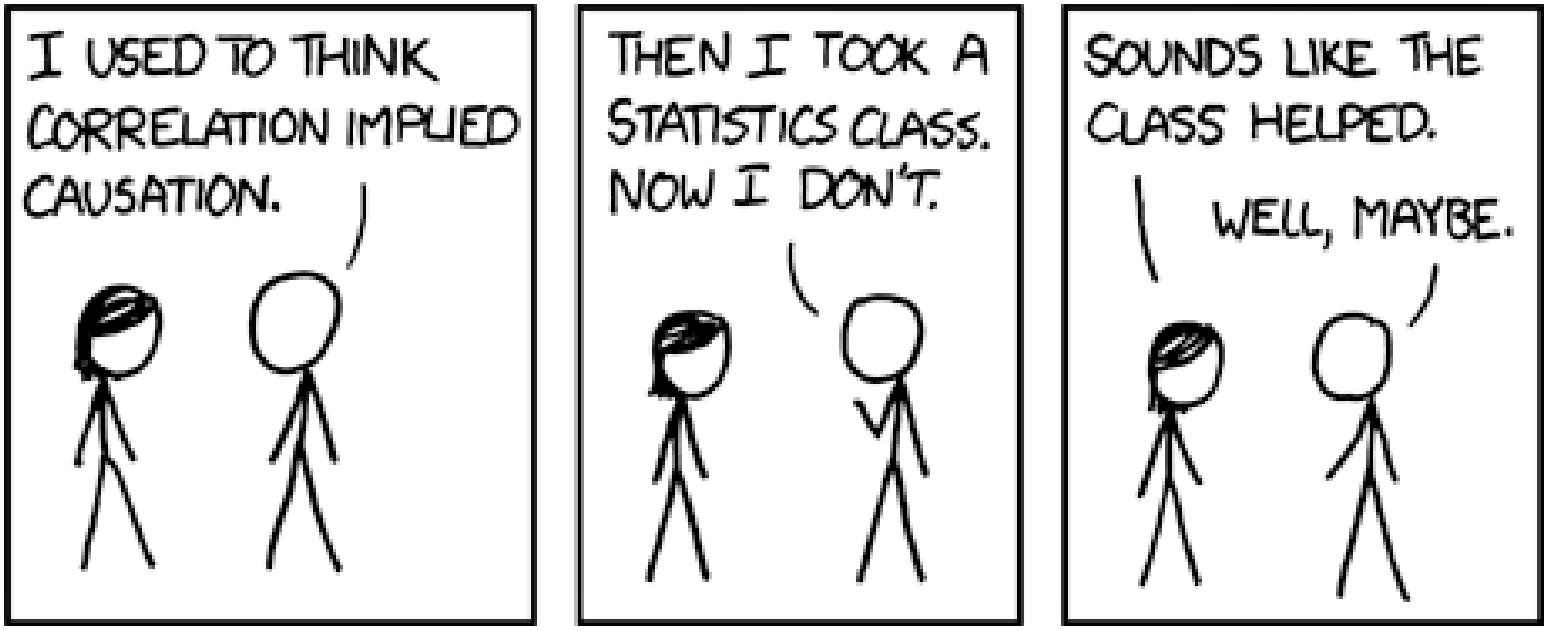
- DIE wissenschaftliche Methode: das Experiment
  - Versuchs- und Kontrollgruppe
  - Randomisierung
    - Dadurch unterscheiden sich Versuchs- und Kontrollgruppe nicht
    - Keine unbeobachtete Heterogenität
  - Kontrollierte Stimulussetzung durch Forscher
    - Damit ist sichergestellt, dass die uV der aV zeitlich vorgeht
    - Keine Endogenität
- Ein sauber durchgeführtes Experiment erlaubt einen sicheren Kausalschluss
- Experimente sind aber in den Sozialwissenschaften oft nicht praktikabel

# Empirische Überprüfung II

- Deshalb erhebt man oft Daten über X und Y ex-post-facto und berechnet deren Korrelation
- Korrelation ist aber nicht gleich Kausalität
  - Es könnte auch eine „Scheinkorrelation“ vorliegen (s.u.)

- Um von einer Korrelation auf Kausalität schließen zu können, müssen folgende Bedingungen gelten:
  - X und Y sind korreliert
  - X geht Y zeitlich voran (keine Endogenität)
  - Die Korrelation von X und Y bleibt erhalten, auch wenn man für dritte Variablen kontrolliert (keine unbeobachtete Heterogenität)

# Korrelation ist nicht gleich Kausalität



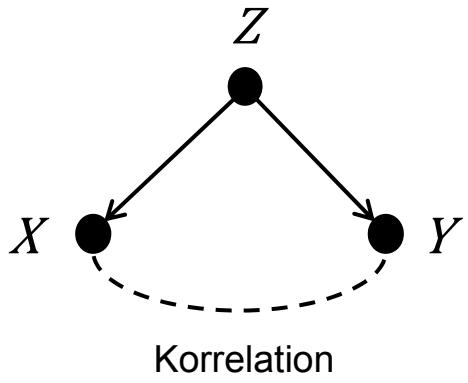

https://xkcd.com/552/

# Das Problem: Selbstselektion

- Bei einem Experiment werden die Vpn vom Forscher den beiden Gruppen per Randomisierung zugewiesen
- Bei ex-post facto Designs überlässt man es den Personen selbst, in welche Gruppe sie gehen (Selbstselektion)
- Das führt leicht zu unbeobachteter Heterogenität
- Selbstselektion ist **das** allgegenwärtige methodische Problem in der Sozialforschung!
  - Beispiele, bei denen dieses Problem nicht bedacht wurde und die so (oder ähnlich) in Studien berichtet wurden und werden(!)
    - Ehemänner leben länger
    - Ehemänner verdienen mehr
    - Ärmere Menschen leben kürzer
    - Bewohner von Betonblöcken sind häufiger krank
    - Häufiges Fernsehen auf Privatsendern macht dumm
    - Zähneputzen senkt das Herzinfarktrisiko

# Folge: Scheinkorrelation/Konfundierung

- X und Y korrelieren zwar, aber Grund hierfür ist eine dritte Variable Z, die sowohl X als auch Y kausal verursacht
  - Die Korrelation ist „echt“, aber die Kausalität ist „scheinbar“ (Scheinkausalität)
- Schematisch anhand eines DAG




- Z ist eine „antezedierende“ Variable (Drittvariable, Confounder)
- Durch die beiden Kausaleffekte entsteht eine Korrelation von X und Y
- Es wäre ein Fehler diese Korrelation als kausal zu interpretieren

- Durch Kontrolle von Z (Drittvariablenkontrolle) kann man das Problem beheben

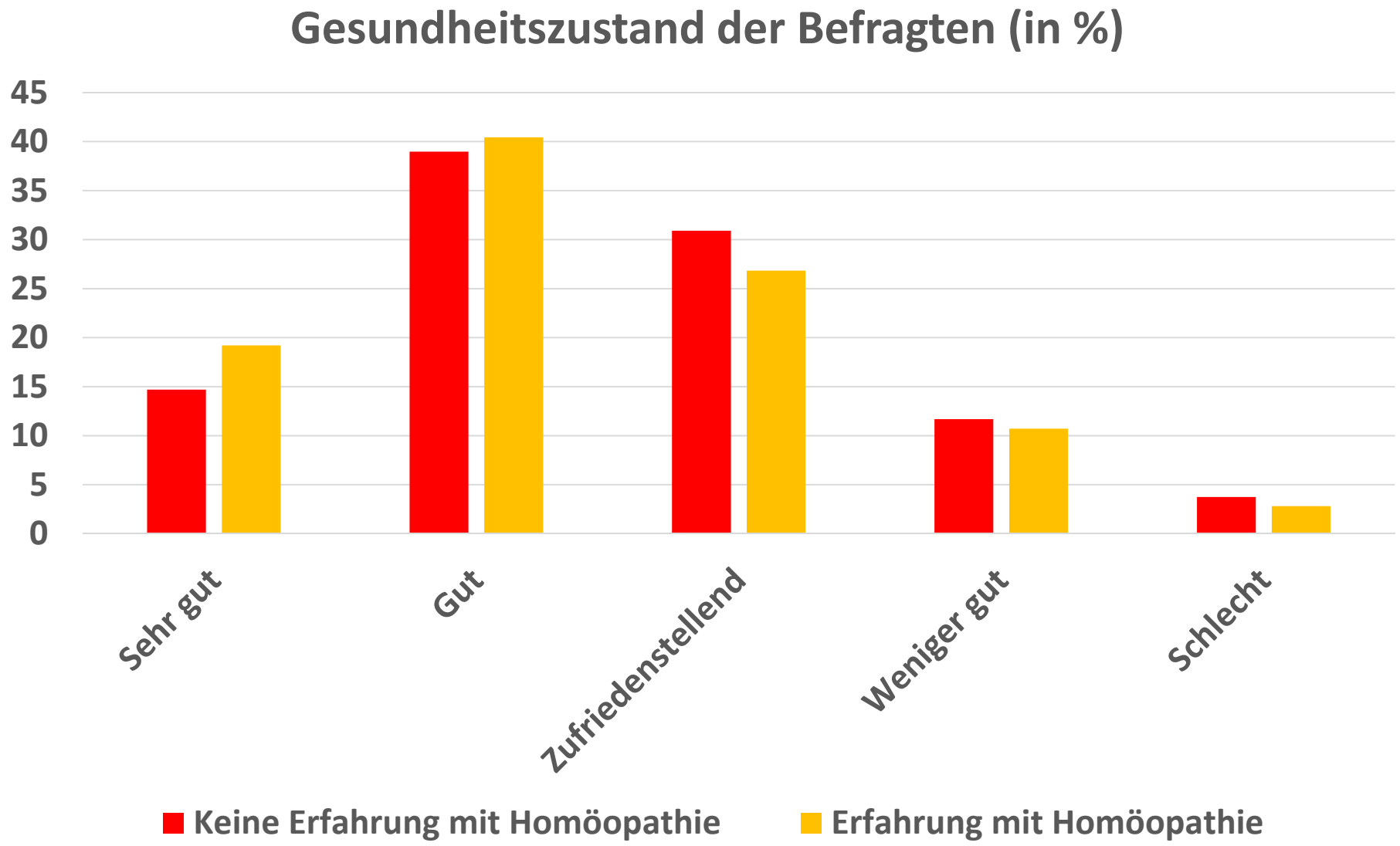
# Drittvariablenkontrolle

- Z.B. mittels konditionaler Kreuztabellen (Partialtabellen)
  - Z wird konstant gehalten: Für jede Ausprägung von Z wird eine eigene Kreuztabelle (X × Y) erstellt (dreidimensionale Kreuztabelle)
  - Damit erhält man für jede Ausprägung von Z einen eigenen, konditionalen Korrelationskoeffizienten:

$$r_{XY \bullet Z_1}, r_{XY \bullet Z_2}, \text{usw.}$$

  - Die messen die Korrelation von X und Y unter Kontrolle von Z
    - Damit ist Z jeweils konstant und kann nicht mehr die Ursache für eine eventuelle Korrelation von X und Y sein
    - Sind die konditionalen Korrelationskoeffizienten ungleich Null, so können wir von Kausalität ausgehen
- Problem: es kann mehrere Drittvariablen geben
  - Lösung: multivariate Analyseverfahren (Regression)
- Regression hat in der Sozialforschung eine zentrale Rolle: Sie ist **der** Ersatz für das Experiment

# Bsp.: Macht Homöopathie gesund?




Daten: Allbus 2012
Quelle: Stefanie Heyne

# Bsp.: Macht Homöopathie gesund?

| | Modell 1 | Modell 2 |
|---|---|---|
| Erfahrung mit Homöopathie | **0.129*** ** | **-0.002** |
| | (3.68) | (-0.05) |
| Bildung (Ref. Niedrige Bildung) | | |
| Mittlere Bildung | | 0.390*** |
| | | (9.73) |
| Hohe Bildung | | 0.673*** |
| | | (15.80) |
| Konstante | 3.491*** | 3.194*** |
| N | 3419 | 3419 |

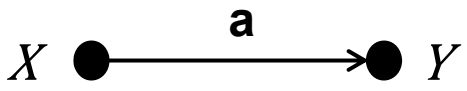
Abhängige Variable: Gesundheitseinschätzung 1 - 5
Quelle: Allbus 2012, Stefanie Heyne

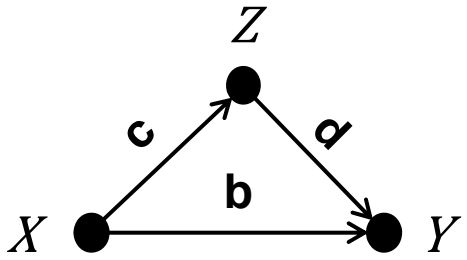
# Methoden der Kausalanalyse

- „Selection on observables“: Confounder gemessen
  - Man kann für sie kontrollieren
    - Regression **(1. Sem.: „Querschnittsdatenanalyse“)**
    - Matching **(3. Sem.: „Kausalanalyse“)**
- „Selection on unobservables“: Confounder nicht gemessen
  - Regressions- und Matching-Schätzer sind verzerrt
    - Unbeobachtete Heterogenität, „omitted variable bias“
  - Unverzerrte Schätzer kann man erhalten mit
    - Instrumentalvariablen Ansatz **(3. Sem.: „Kausalanalyse“)**
      - Exogene Variation identifiziert den Kausaleffekt
    - Regression Discontinuity Ansatz **(3. Sem.: „Kausalanalyse“)**
      - Homogenität von Personen an einer Schwelle
    - Within-Panelanalyse **(2. Sem.: „Längsschnittdatenanalyse“)**
      - Homogenität einer Person über die Zeit

# Intervention

- Bei Konfundierung ist Z „antezedierend“ (zeitlich vor X und Y)
- Ist Z intervenierend (zeitlich zwischen X und Y), so liegt eine Intervention vor
- Eine Intervention ist im Unterschied zur Konfundierung kein Problem der Kausalanalyse, sondern der 2. Schritt
  - Z ist ein „kausaler Mechanismus“
  - Man weiß nun, wie der Kausaleffekt von X auf Y zustande kommt




- (Totaler) Kausaleffekt: a
- Direkter Kausaleffekt: b
- Indirekter Kausaleffekt: c · d
- Es gilt: a = b + c · d
- Kausalanalyse
  - 1. Schritt: schätze den totalen KE
  - 2. Schritt: kontrolliere für Z
    - b < a : Z ist ein Mechanismus

# Beispiel: Kirchgangshäufigkeit

- Mit dem ALLBUS 1994 untersuchen wir, wie sich der Wohnort (West/Ost) auf den Kirchgang auswirkt

Kirchgang nach Wohnort

| | West | Ost |
|---|---|---|
| Selten/nie | 55% | 84% |
| Öfter | 45% | 16% |
| N | 2339 | 1104 |

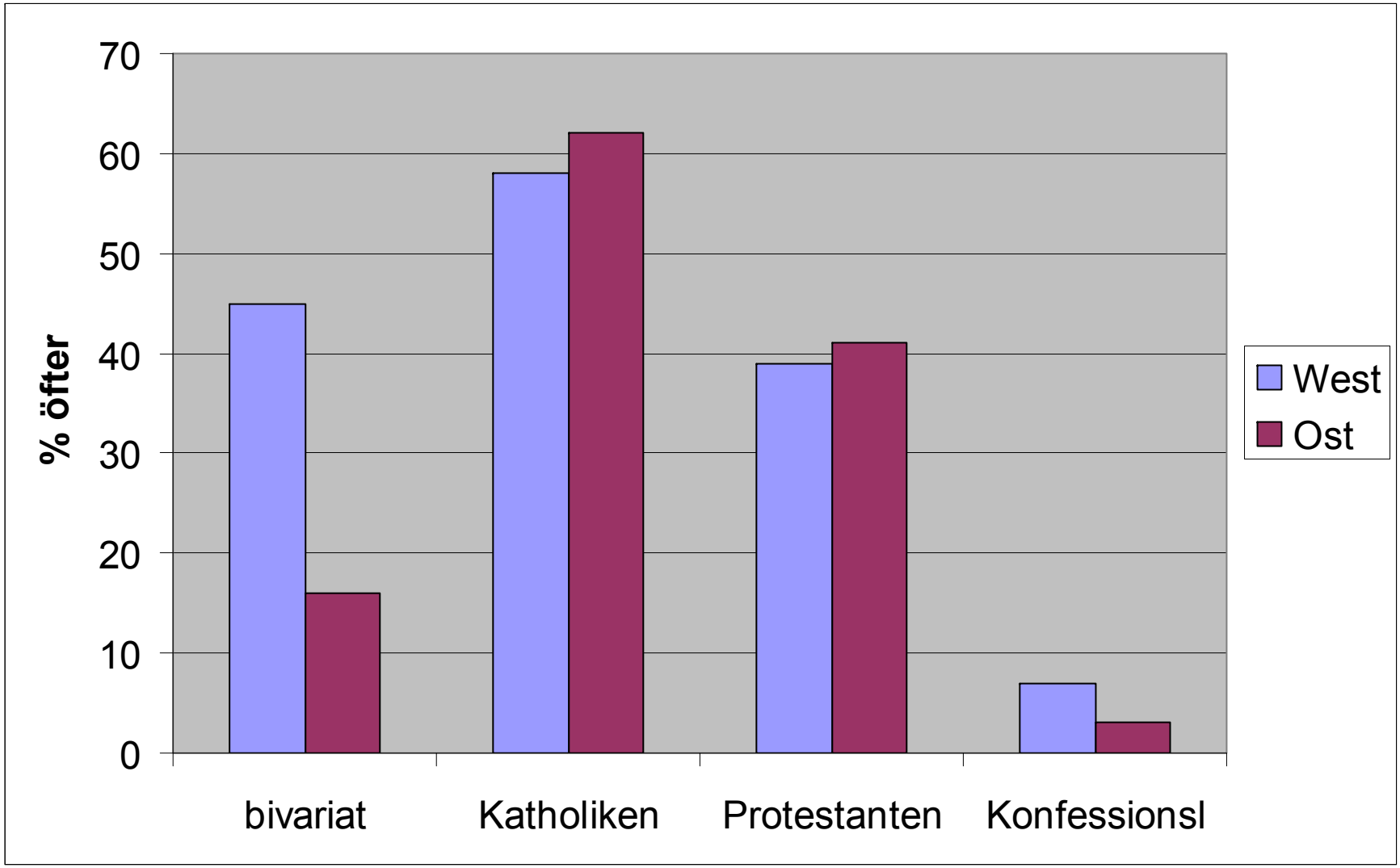
V = 0,28

- Könnte „Konfession“ ein intervenierender Mechanismus sein? Deshalb kontrollieren wir für Konfession (2 × 2 × 3-Tabelle)

Kirchgang nach Konfession und Wohnort

| | Katholiken | | Protestanten | | Konfessionslose | |
|---|---|---|---|---|---|---|
| | West | Ost | West | Ost | West | Ost |
| Selten/nie | 42% | 38% | 61% | 59% | 93% | 97% |
| Öfter | 58% | 62% | 39% | 41% | 7% | 3% |
| | V = 0,01 | | V = 0,01 | | V = 0,07 | |

# Beispiel: Kirchgangshäufigkeit



# Beispiel: Kirchgangshäufigkeit

- Um den Kausalmechanismus ganz zu verstehen, erstellen wir noch die Kreuztabellen X × Z und Z × Y

Konfession nach Wohnort

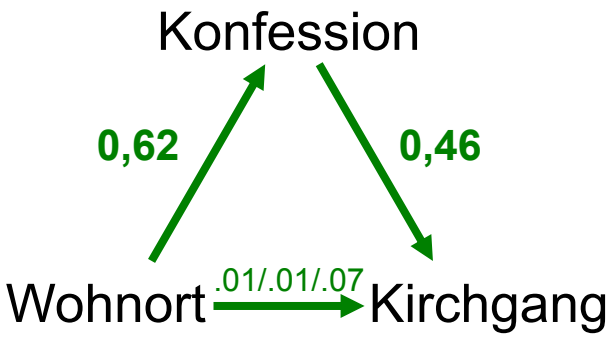
| | West | Ost |
|---|---|---|
| Katholik | 47% | 3% |
| Protestant | 41% | 26% |
| Konfessionsl. | 12% | 71% |

V = 0,62

Kirchgang nach Konfession

| | selten | öfter |
|---|---|---|
| Katholik | 41% | 59% |
| Protestant | 60% | 40% |
| Konfessionsl. | 96% | 4% |

V = 0,46




Das gesamte Kausalmodell präsentieren wir übersichtlich in einem „Pfaddiagramm“. An die Pfeile schreiben wir die bivariaten Korrelationskoeffizienten. Die konditionalen Korrelationskoeffizienten von „Wohnort“ auf „Kirchgang“ sind fast null und machen deutlich, dass hier praktisch kein direkter Kausaleffekt vorliegt. „Konfession“ ist der kausale Mechanismus, der den KE von „Wohnort“ auf „Kirchgang“ vollständig erklärt.

# Extrem: Suppression

- Falls der indirekte Effekt ein anderes Vorzeichen hat, als der direkte Effekt, fällt der totale Effekt kleiner aus

$$a = b + c \cdot d$$

  - In extremen Fällen ist der totale Effekt gar nahe Null (verdeckte Korrelation, Suppression)
  - Oder hat gar das entgegengesetzte Vorzeichen

- Bsp.: Diskriminierung von Frauen bei der Studienzulassung?
  - Aus: Krämer, Walter (1995) Denkste! Campus-Verlag
  - Fiktives Beispiel

| | M | F | Σ |
|---|---|---|---|
| nicht zugel. | 400 | 450 | 850 |
| zugelassen | 100 (20%) | 50 (10%) | 150 |
| Σ | 500 | 500 | 1000 |

Φ = (-) 0.14

# Beispiel: Studienzulassung

Kontrolliert man für die intervenierende Variable „Studienfach“ verschwindet der Effekt: es gibt nur einen schwachen direkten Effekt.

Mathe

| | M | F | Σ |
|---|---|---|---|
| nicht zug. | 100 | 10 | 110 |
| zug. | 80 (44%) | 10 (50%) | 90 |
| Σ | 180 | 20 | 200 |

Φ = (+) 0.03

SoWi

| | M | F | Σ |
|---|---|---|---|
| nicht zug. | 300 | 440 | 740 |
| zug. | 20 (6%) | 40 (8%) | 60 |
| Σ | 320 | 480 | 800 |

Φ = (+) 0.04

# Beispiel: Studienzulassung

Frauen bewerben sich häufiger für Sowi

| | M | F | Σ |
|---|---|---|---|
| Mathe | 180 | 20 | 200 |
| Sowi | 320 | 480 | 800 |
| Σ | 500 | 500 | 1000 |

Φ = (+) 0.40

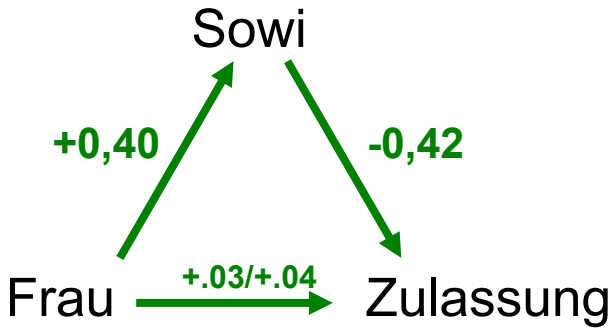
Sowi hat niedrigere Zulassungsquoten als Mathe

| | Mathe | Sowi | Σ |
|---|---|---|---|
| nicht zug. | 110 | 740 | 850 |
| zug. | 90 | 60 | 150 |
| Σ | 200 | 800 | 1000 |

Φ = (-) 0.42

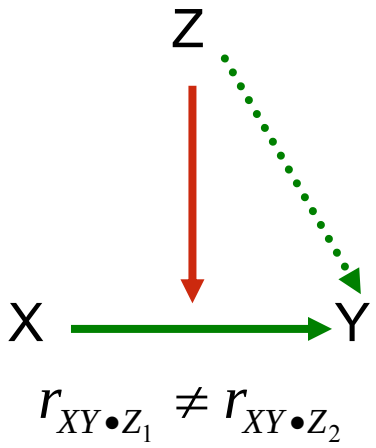
- Der direkte Effekt ist Null
- Der indirekte Effekt ist negativ
- Deshalb ist der totale Effekt negativ

- Der totale Effekt wird vollständig durch die Studienfachwahl erklärt
  - Es gibt keine Diskriminierung



# Interaktion

Die Beziehung von X und Y fällt unterschiedlich aus, je nachdem welchen Wert Z annimmt (Z heißt auch „Moderator“)




$$r_{XY \bullet Z_1} \neq r_{XY \bullet Z_2}$$

Beispiele:

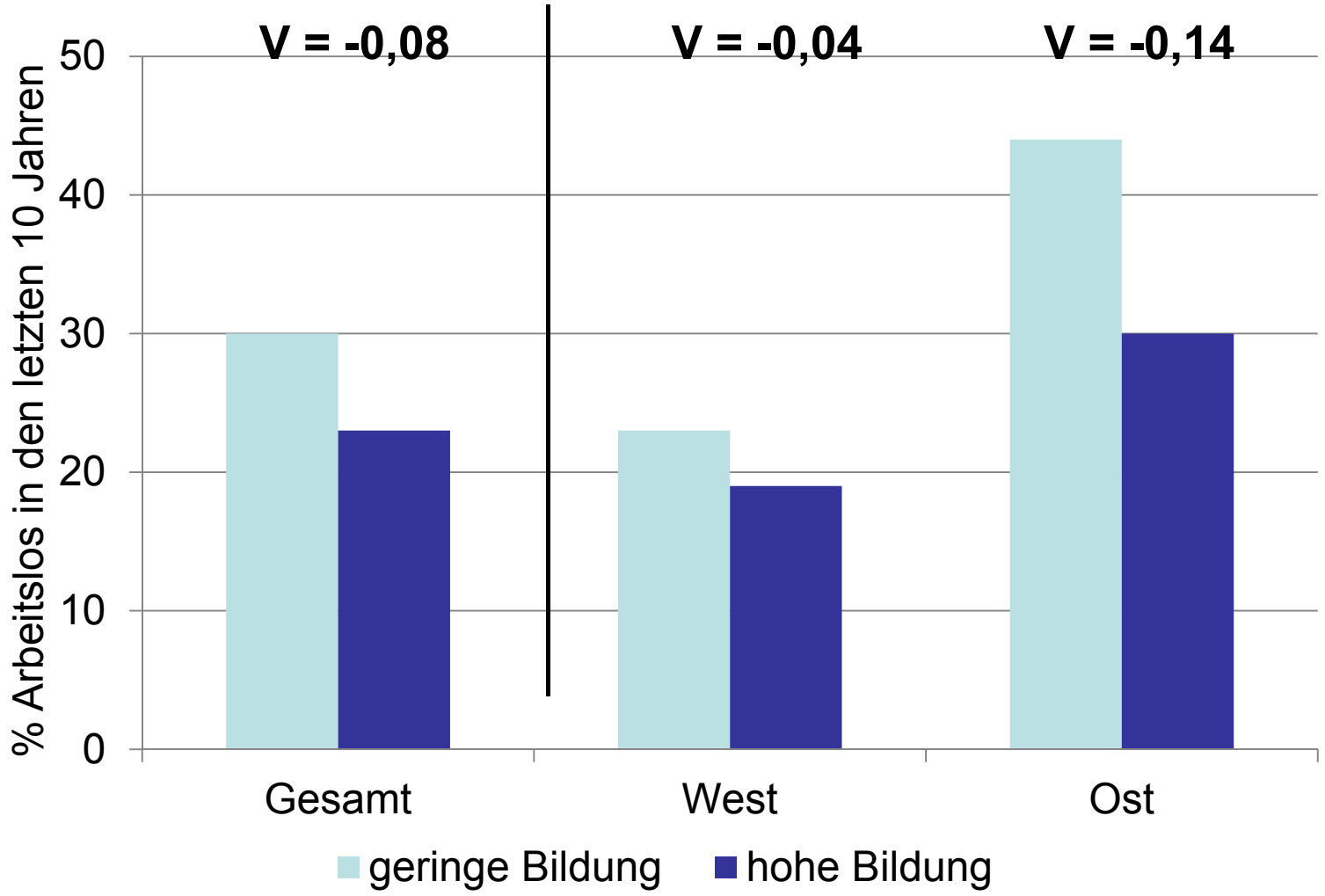
- Sport (X), Gesundheit (Y), Erkältung (Z)
- Einsatz (X), Erfolg (Y), Motivation (Z)

Beispiel: M. Halbwachs (1930) Les Causes du Suicide

Halbwachs stellte fest, dass es einen Zusammenhang zwischen Konfession und Selbstmordrate gibt: Katholiken 19,9 Selbstmorde (pro 100.000), Protestanten 39,6 Selbstmorde (pro 100.000). Kontrolliert man den Wohnort, so verschwindet der Zusammenhang für Städte, auf dem Land wird er stärker.

| Wohnort | Katholik | Protestant |
|---|---|---|
| Stadt | 39,9 | 37,8 |
| Land | 8,8 | 41,4 |
| Alle | 19,9 | 39,6 |

# Beispiel: Arbeitslosigkeit in Abhängigkeit von Bildung




Daten: ALLBUS 2002
Do-File: 6 LinReg Interaktion.do

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Kapitel 2:
# Explorative Datenanalyse

# Anscombes Quartett

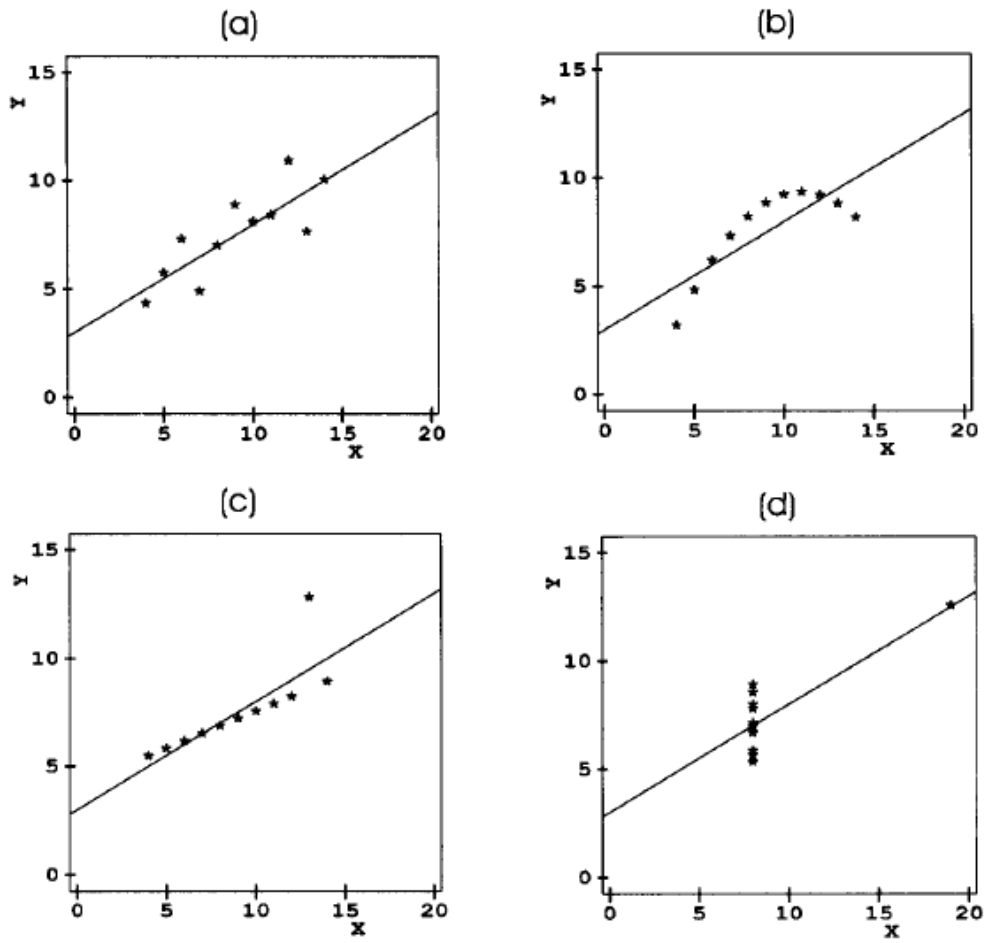
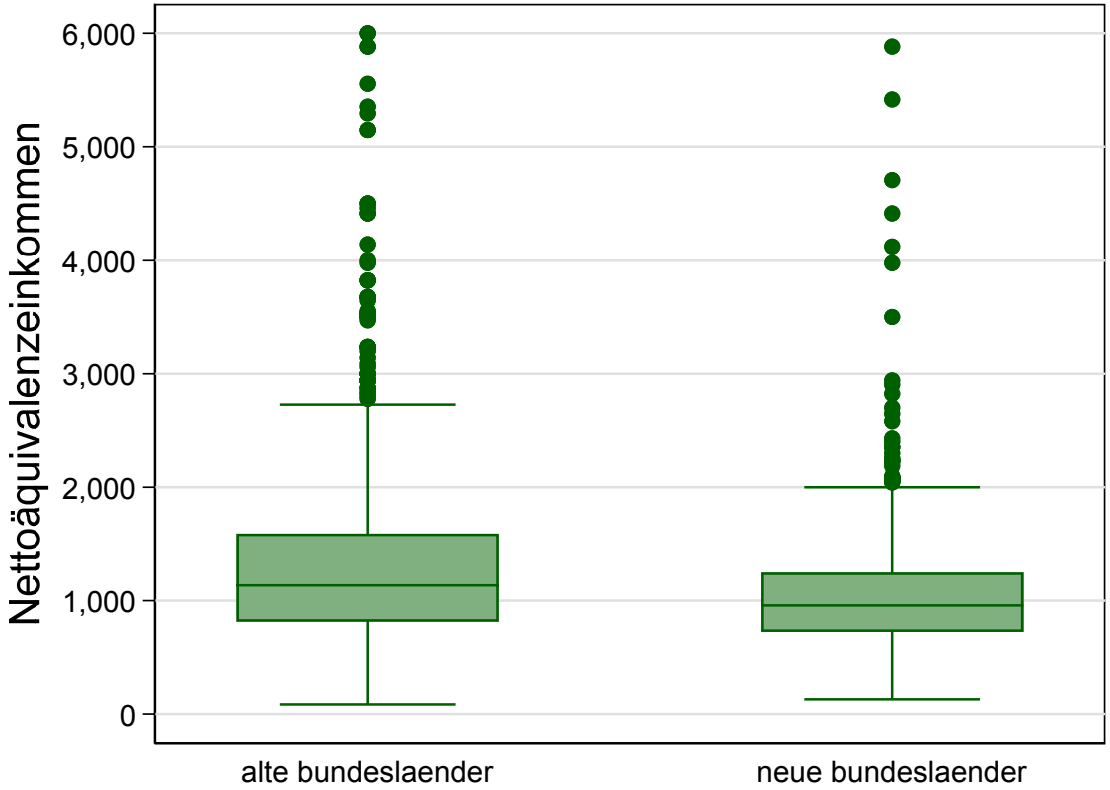



**Figure 3.1.** Four datasets, due to Anscombe (1973), with identical least-squares regressions. In (*a*), the linear regression is an accurate summary; in (*b*), the linear regression distorts the curvilinear relationship between $Y$ and $X$; in (*c*), the linear regression is drawn toward an outlier; in (*d*), the linear regression "chases" the influential observation at the right.

# Univariate Verteilungen
# Bsp.: Armut in Deutschland

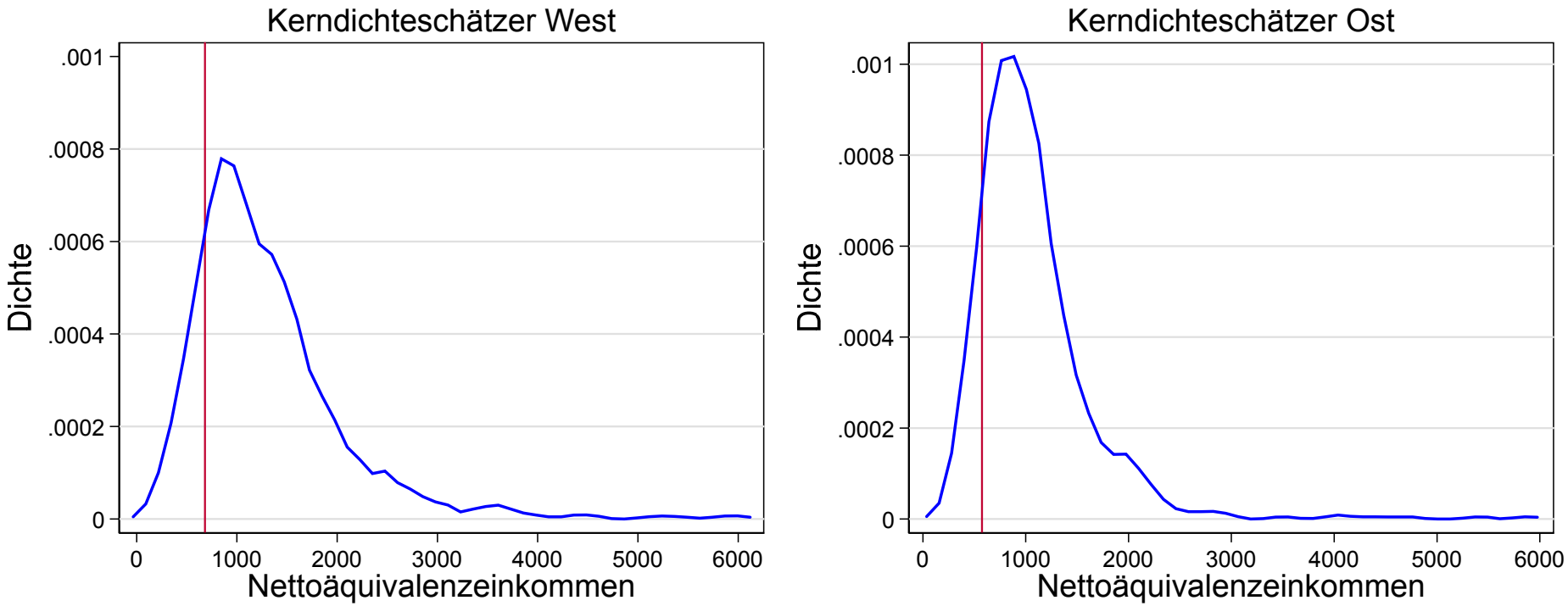
```
* Gruppiertes Boxplot West/Ost
graph box oecdeink, over(v3)
```




| | West | Ost |
|---|---|---|
| Armutsgrenze (60% Median) | 682 | 575 |
| Armutsquote | 15,9% | 12,1% |

Daten: ALLBUS 2002
Do-File: 0 Armut.do

# Beispiel: Armut in Deutschland



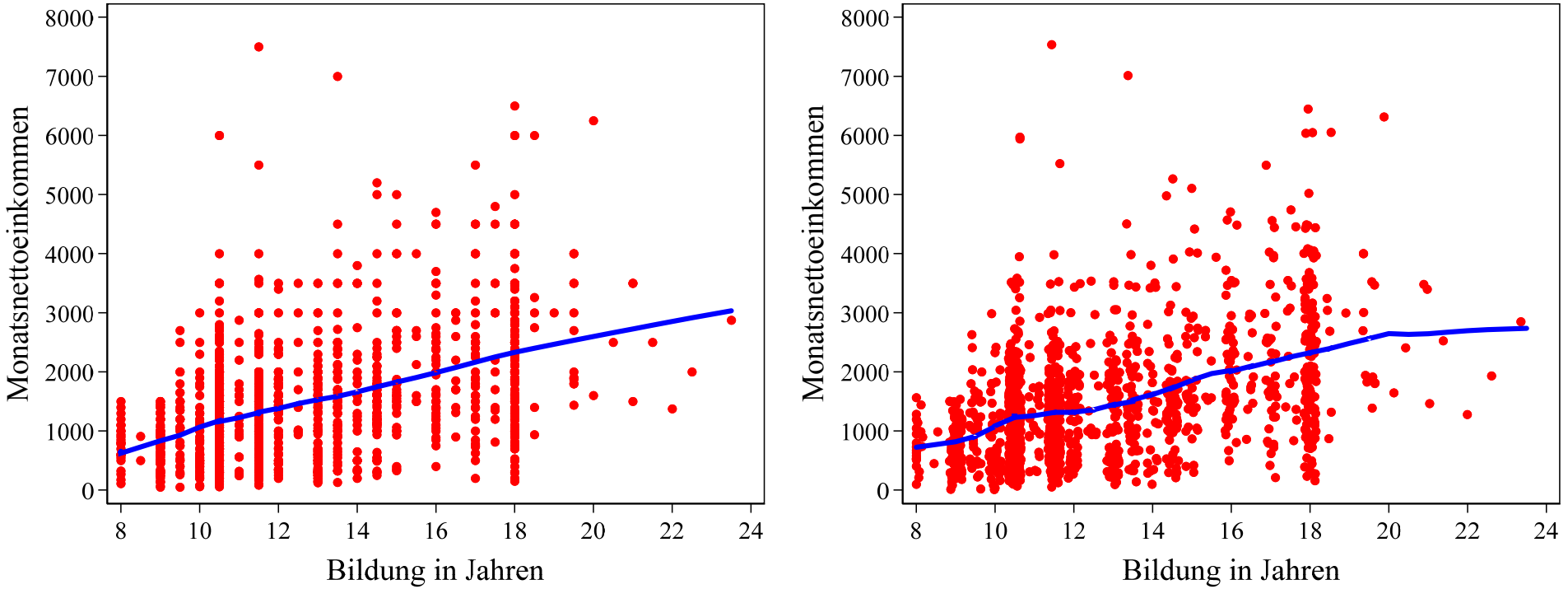

Eingezeichnet sind die Armutsgrenzen

Daten: ALLBUS 2002
Do-File: 0 Armut.do

# Bivariate Verteilungen: Scatterplot

- Bivariate Zusammenhänge veranschaulicht man am besten mit einem Streudiagramm
  - Überdecken sich die Daten stark, so, „jittered“ man am besten
  - Einen Eindruck von der Art des Zusammenhangs bekommt man mittels einer nicht-parametrischen Regression. Bewährt hat sich hierfür der Lowess-Smoother (locally weighted scatterplot smoother).
    - An der Stelle $x_i$ wird eine lineare Regression berechnet, in die die Daten in der Umgebung gewichtet eingehen. Die Breite der Umgebung ist steuerbar durch ,,bandwidth“ (z.B. bwidth=0.8). Es wird trikubisch gewichtet. Anhand der Regressionsparameter wird dann $\widehat{y_i}$ berechnet. Dies wird für alle X-Werte gemacht. Die Verbindung der $(x_i, \widehat{y_i})$ ergibt die Lowess-Kurve. Je kleiner die Umgebung, desto näher an den Daten ist die Kurve.

# Bivariate Verteilungen: Scatterplot

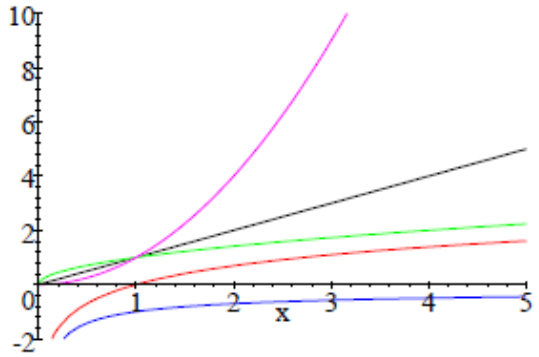



Daten: ALLBUS 2002
Do-File: 1 Regression.do

- Beispiel: Einkommen in Abhängigkeit von Bildung
  - Nur Westdeutschland, max. 8 Tsd. Euro (N=1.468)
  - Links ist nicht gejittered, es kommt zu starker Überdeckung. Rechts ist gejittered („jitter(2)“: 2% der Zeichenfläche)
  - Die blaue Kurve ist der Lowess-Smoother
    - Links werden zur Berechnung jeweils 80% der Fälle in der Umgebung verwendet, rechts nur 30%. Die rechte Kurve folgt deshalb wesentlich genauer den Daten, ist dafür aber unregelmäßiger.
  - In beiden Fällen erkennt man kaum Nicht-Linearitäten

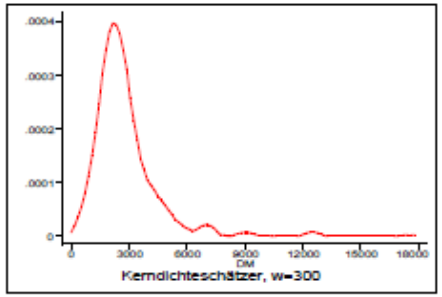
# Exkurs: Datentransformationen

- Schiefe und Ausreißer sind für Regressionen ein Problem
- Durch Potenz-Transformationen kann man Schiefe reduzieren
- Tukeys “ladder of powers”

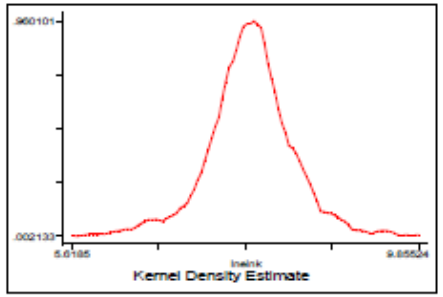



| | | | |
|---|---|---|---|
| $x^3$ | $q = 3$ | | produziert |
| $x^{1.5}$ | $q = 1.5$ | cyan | Rechtsschiefe |
| $x$ | $q = 1$ | schwarz | |
| $x^{.5}$ | $q = .5$ | grün | produziert |
| $\ln x$ | $q = 0$ | rot | Linksschiefe |
| $-x^{-.5}$ | $q = -.5$ | blau | |

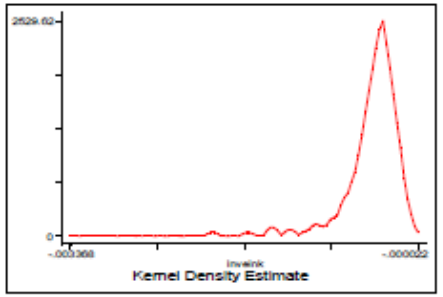
**Beispiel: Einkommensverteilung**




q=1




q=0




q=-1

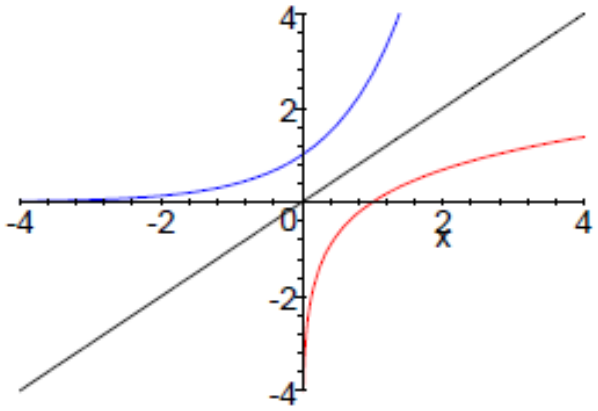
# Exkurs: Potenzfunktionen, ln und e

$$x^{0.5} = x^{\frac{1}{2}} = \sqrt[2]{x}, \qquad x^{-0.5} = \frac{1}{x^{0.5}} = \frac{1}{\sqrt[2]{x}}, \qquad x^0 = 1$$

Mit $\ln$ notieren wir den (natürlichen) Logarithmus zur Basis $e = 2,71828\ldots$:

$$y = \ln x \Leftrightarrow e^y = x$$

Daraus folgt $\ln(e^y) = e^{\ln y} = y$.

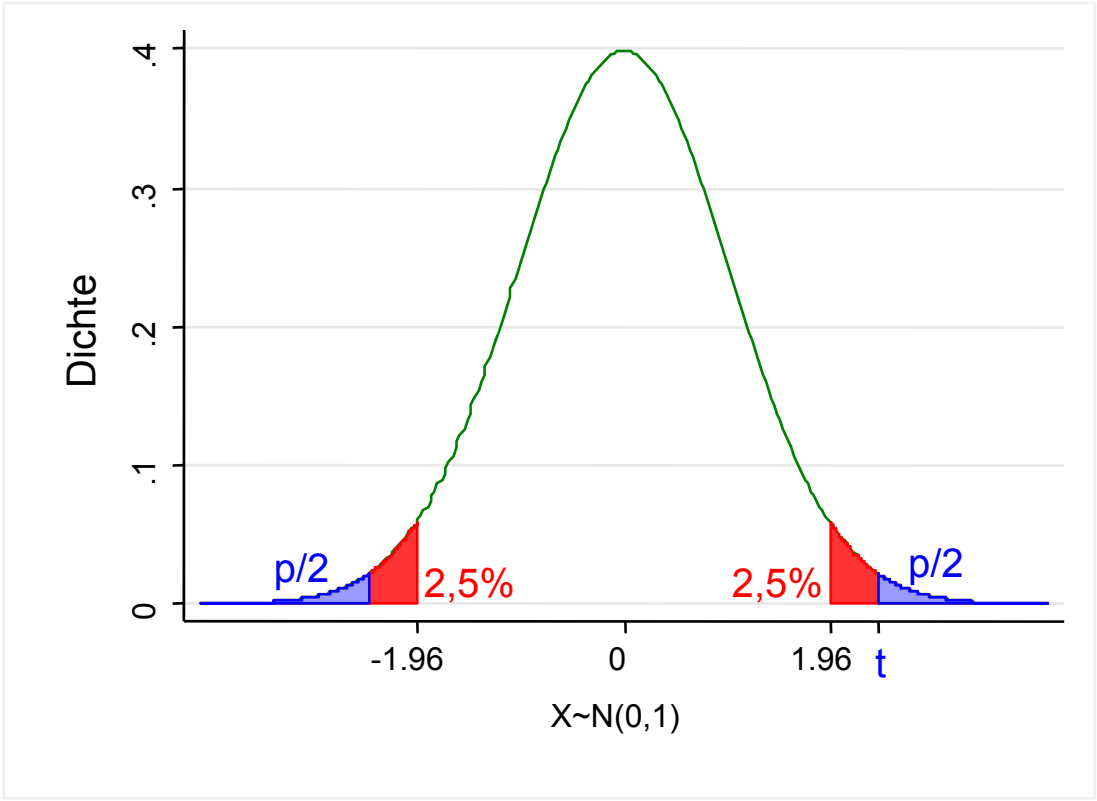



Rechenregeln

| | |
|---|---|
| $e^x e^y = e^{x+y}$ | $\ln(xy) = \ln x + \ln y$ |
| $e^x / e^y = e^{x-y}$ | $\ln(x/y) = \ln x - \ln y$ |
| $(e^x)^y = e^{xy}$ | $\ln x^y = y \ln x$ |

# Exkurs: p-Wert

- Der p-Wert gibt bei Gültigkeit der $H_0$ die Wahrscheinlichkeit an, dass die Teststatistik den berechneten Wert oder einen, der noch weiter in Richtung der Alternativhypothese liegt, annimmt
- Die Nullhypothese wird dann verworfen, wenn **$p \leq \alpha$**




$H_0$ : diff = 0

Prüfverteilung: Bei n>30 ist die t-Verteilung eine Standardnormalverteilung

Die kritischen Werte sind auf dem 5%-Niveau -1,96 und 1,96

Liegt die Teststatistik **t** z.B. bei 2,4, so kann die $H_0$ abgelehnt werden

Rechts von **t** liegt p/2 (die Whs., dass noch was Extremeres rauskommt) (hier ist p=0,016)

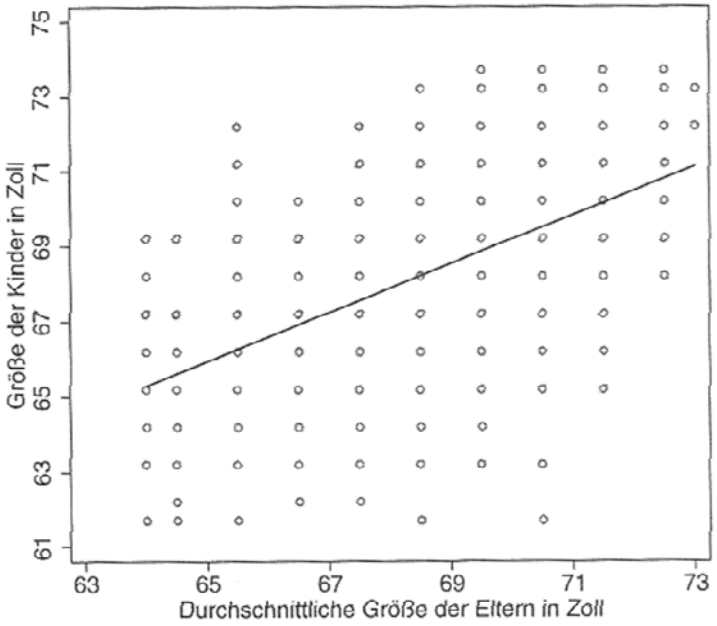
Da offensichtlich $p < 0,05$, kann die $H_0$ abgelehnt werden

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

# Kapitel 3:
# Einführung in die Regression

# Zum Begriff „Regression“

- Regression toward the mean:
  „The stature of the adult offspring must on the whole, be more *mediocre* than the stature of their parents“ (Sir Francis Galton, 1889)
- Galton fittete (visuell!) eine Gerade
  - Sein Ergebnis: Steigung von 0,67
- Später wurde dies mit OLS gemacht und auf das Fitten von Geraden mit OLS der Begriff „Regression“ übertragen
  - OLS Ergebnis: Steigung von 0,64
- Die erste inhaltliche Anwendung prägte also den Begriff für ein statistisches Verfahren!



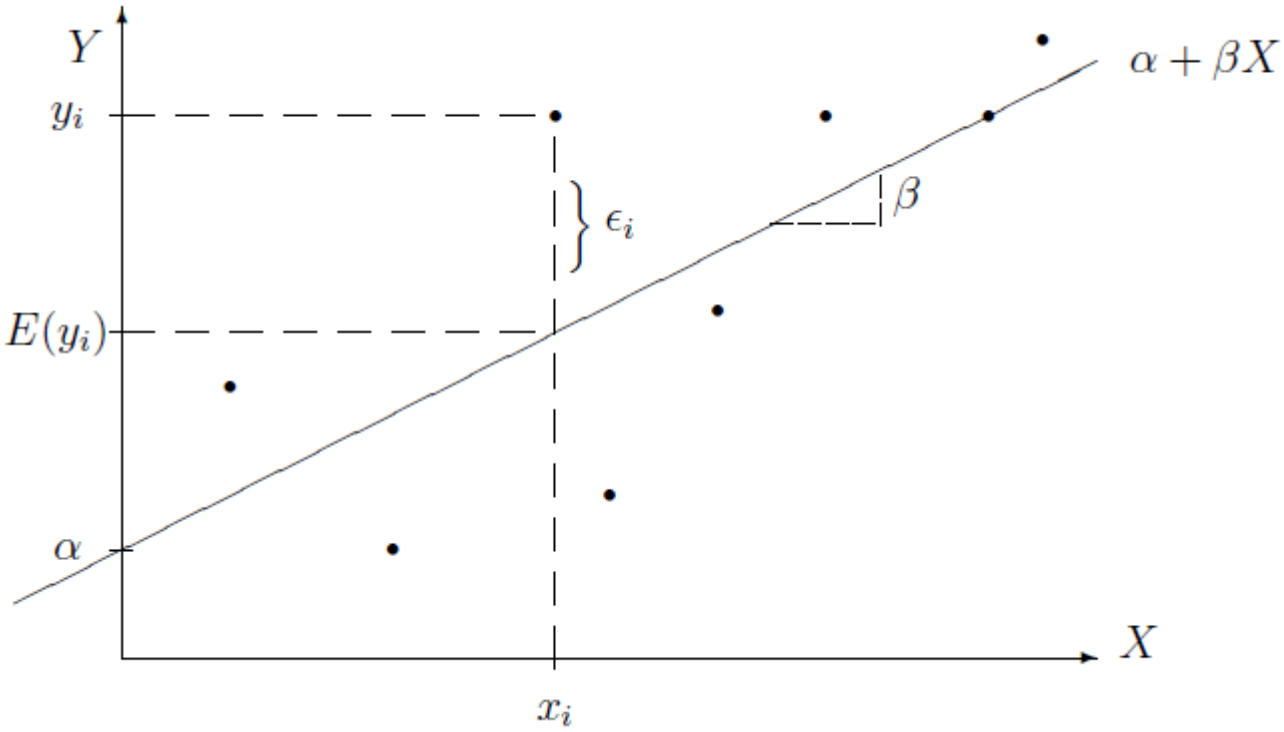

Galtons Daten mit Regressionsgerade
Quelle: Fahrmeir et al. (2007)

# Das einfache Regressionsmodell

- Man formuliert folgendes lineare Modell des Zusammenhangs:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- α und β sind die Regressionskoeffizienten
  - α: Achsenabschnitt, β: Steigung
  - β: um wie viel Einheiten ändert sich Y, wenn X um eine Einheit steigt
- $\varepsilon_i$ ist der Fehlerterm (Abweichung der Daten von der Modellgerade)



# OLS-Schätzer

- Man schätzt die Regressionskoeffizienten, indem man die Fehlerquadratsumme minimiert (ordinary least squares, OLS)

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \varepsilon_i^{\,2} = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Ableiten dieses Ausdrucks, Nullsetzen und Auflösen der beiden daraus resultierenden Gleichungen, liefert die OLS-Schätzer:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

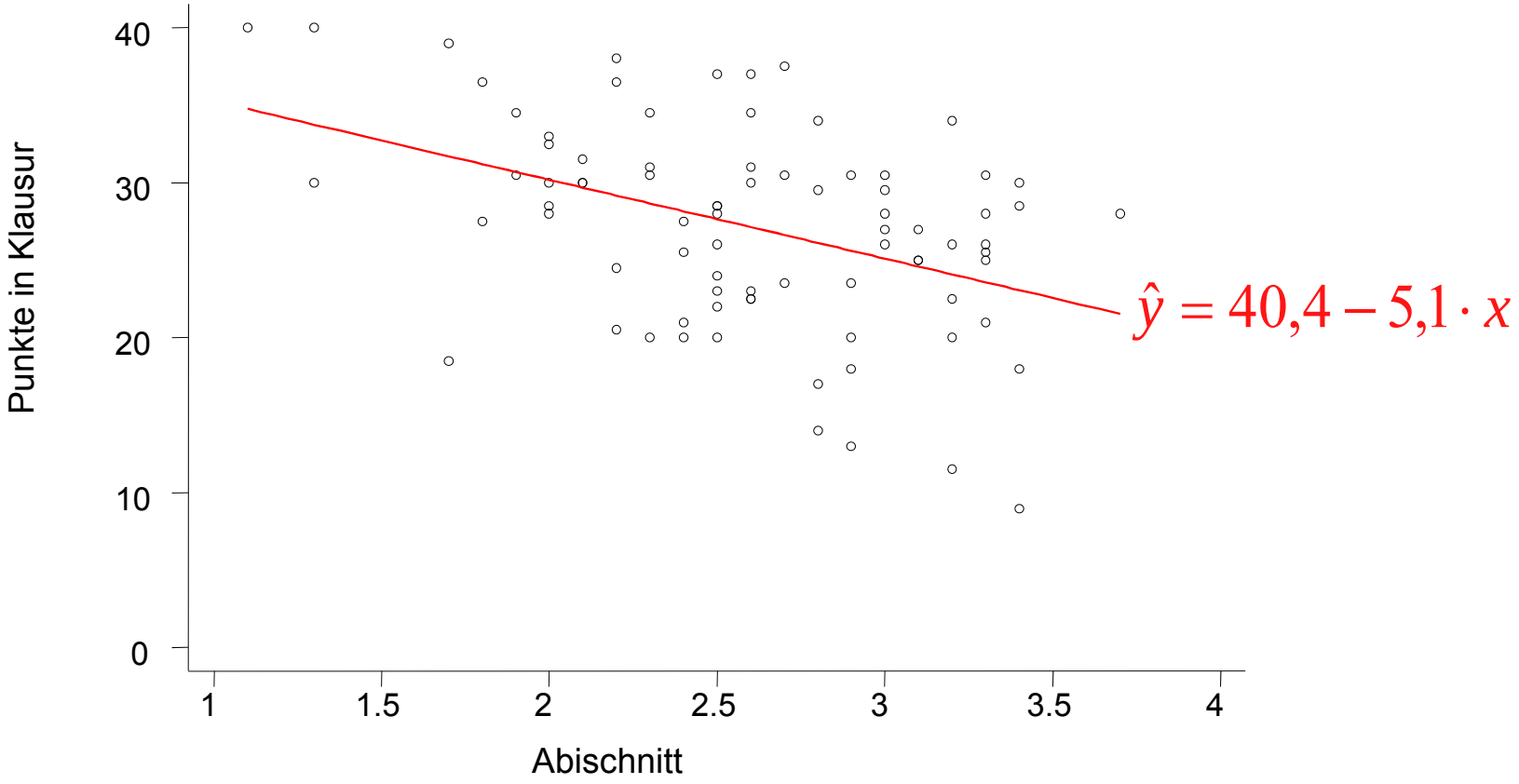
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{XY}}{s_X^{\,2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2}$$

- Die vom Regressionsmodell vorhergesagten Werte sind $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

- Die geschätzten Fehler (Residuen) sind damit $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$

# Beispiel

**Abinote und Klausurerfolg**



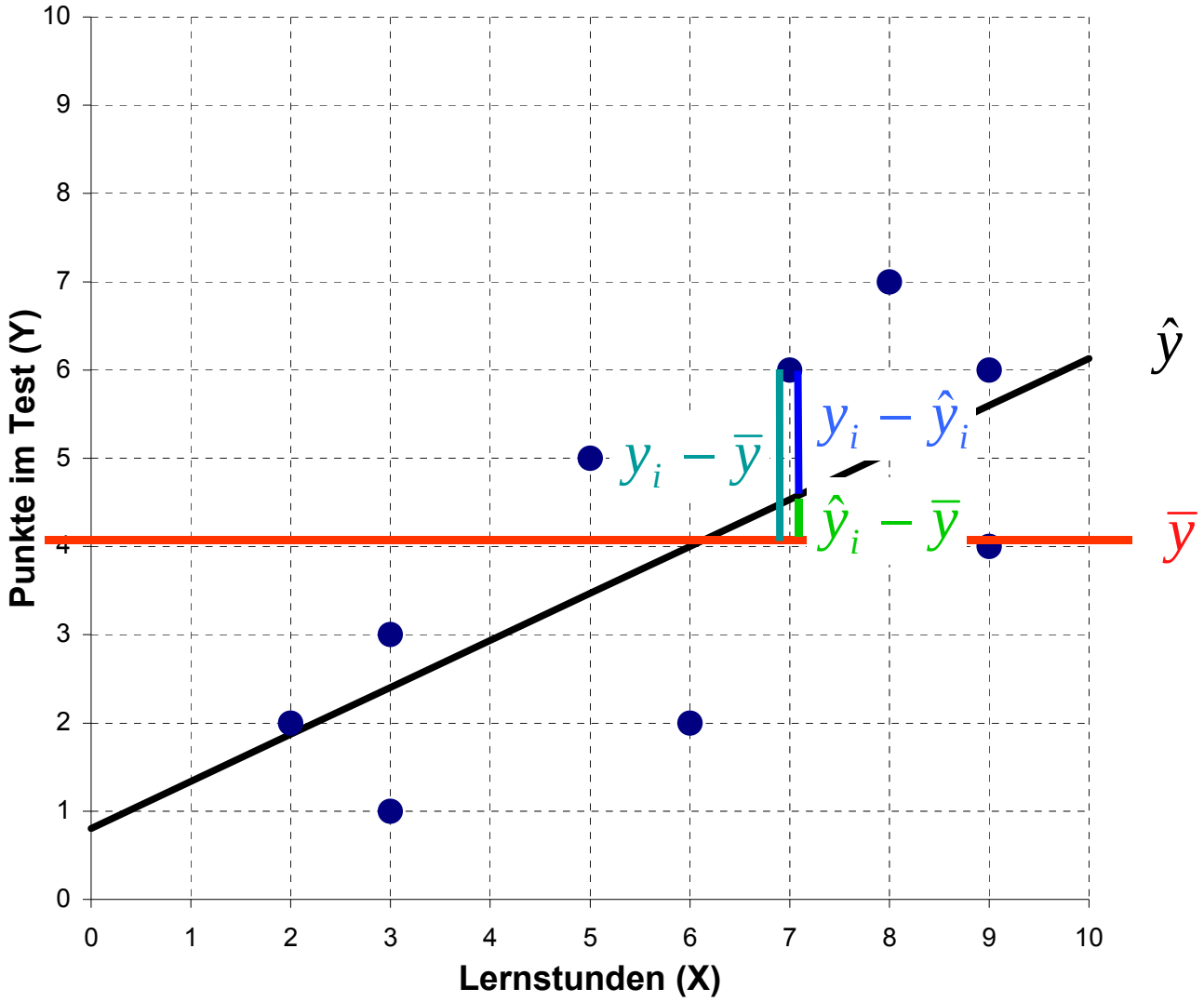
# Das Bestimmtheitsmaß $R^2$

- Wie gut passt das Regressionsmodell auf die Daten?
- Die Grundidee ist: Welcher Anteil der Streuung von Y wird durch das Regressionsmodell „erklärt“?
- Streuungszerlegung
  - Total sum of squares (TSS): $$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
  - Model sum of squares (MSS): $$MSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
  - Residual sum of squares (RSS): $$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
  - Die gesamte Streuung kann damit in zwei Teile zerlegt werden

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = MSS + RSS$$

# Graphische Interpretation der Streuungszerlegung



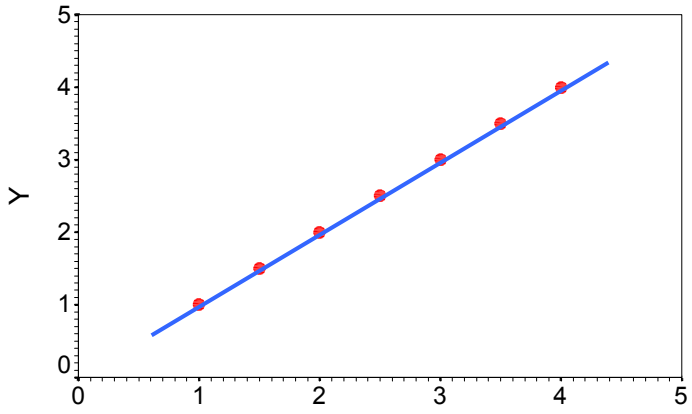
# Das Bestimmtheitsmaß $R^2$

- Das Bestimmtheitsmaß ist nun definiert als

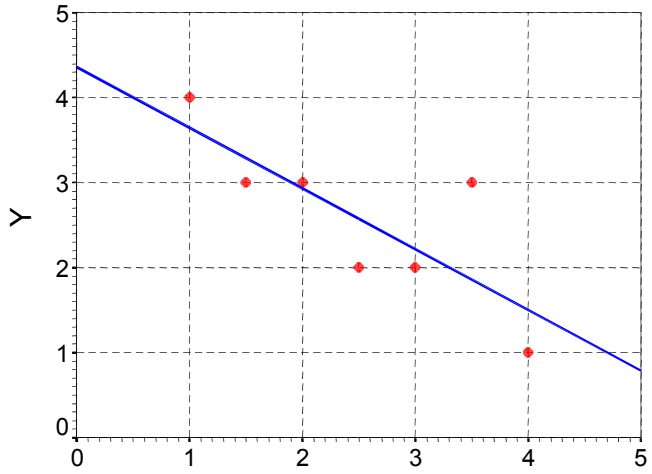
$$R^2 = \frac{\text{erklärte Streuung}}{\text{gesamte Streuung}} = \frac{MSS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Es gilt: $0 \leq R^2 \leq 1$
- $R^2$ lässt sich interpretieren als der Anteil der Varianz, der durch die Regressionsgerade (und damit durch X) erklärt wird
- Es gilt: $R^2 = r^2$

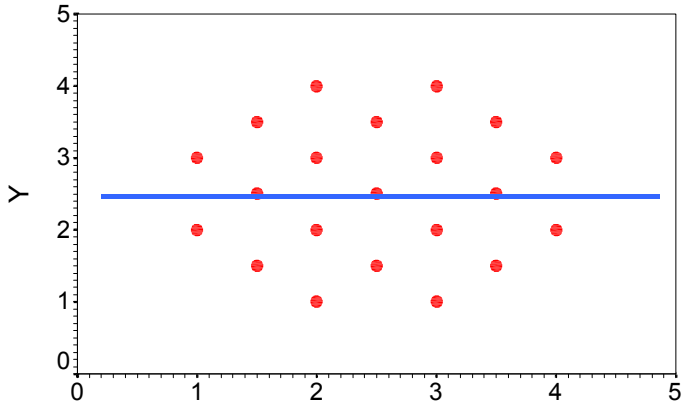
# r und $R^2$



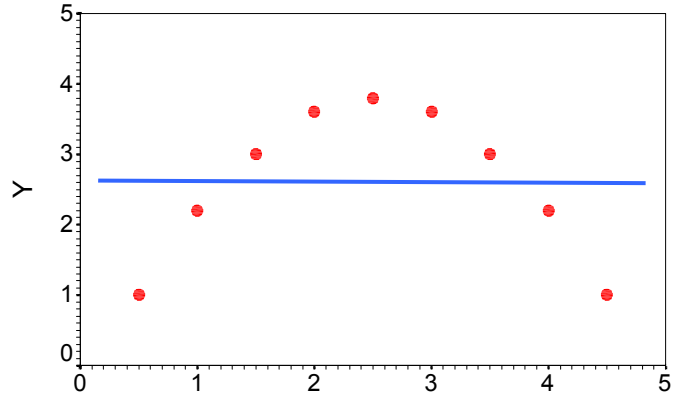

r = 1, $R^2$ = 1




r = -0.79 , $R^2$ = 0,62




r = 0 , $R^2$ = 0




r = 0 , $R^2$ = 0

# Signifikanztest für $\hat{\beta}$

- $\hat{\beta}$ ist ein Schätzer
  - Mit einer Stichprobenverteilung
  - Und einem Standardfehler $\hat{\sigma}_{\hat{\beta}}$
- Damit kann man auch ein Konfidenzintervall schätzen
- Ebenso kann man einen Signifikanztest durchführen
  - Nullhypothese: X hat keinen Einfluss auf Y (kein Zusammenhang)
    $H_0$: β = 0
  - Die Teststatistik (t-Wert) ist

$$T = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \sim t(n-2)$$

  - Die $H_0$ wird abgelehnt, falls $|T| > t_{1-\alpha/2}(n-2)$
    - Ab n>30 das $z_{1-\alpha/2}$ Quantil (Faustregel für α=5%: |T| > 2)
  - Können wir die $H_0$ verwerfen, so spricht man davon, dass X einen signifikanten Einfluss auf Y hat

# Annahmen der Regression

- A1: Linearitätsannahme (**Linearität**)

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

- A2: Im Mittel ist der „Fehler“ null

$$E(\varepsilon_i) = 0, \quad \text{für alle i}$$

- A3: Die Fehlervarianz ist konstant (**Homoskedastizität**)

$$V(\varepsilon_i) = \sigma^2, \quad \text{für alle i}$$

- A4: Die Fehlerkovarianzen sind null (**keine Autokorrelation**)

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad \text{für alle i} \neq \text{j}$$

- A5: Regressor und Fehler sind unkorreliert (**Exogenität**)

$$Cov(x_i, \varepsilon_j) = 0, \quad \text{für alle i und j}$$

- A6: Fehler normalverteilt (für Sig.tests) (**Normalverteilung**)

$$\varepsilon_i \sim N(0, \sigma^2)$$

# Eigenschaften der OLS-Schätzer

- Bei Gültigkeit von A1 bis A5 haben die OLS-Schätzer gewisse wünschenswerte Eigenschaften: Sie sind
- unverzerrt (erwartungstreu): $E(\hat{\beta}) = \beta$
- in der Klasse der linearen, unverzerrten Schätzer die mit der kleinsten Stichprobenvarianz
  - best linear unbiased estimate (BLUE)
  - Gauß-Markov Theorem
- Dabei bedeutet:
  - „linear“: die Schätzer lassen sich als lineare Funktionen der Daten berechnen
  - „unbiased“: die Schätzer sind erwartungstreu
  - „best“: die Schätzer sind effizienter als alle anderen linearen Schätzer

# Standardisierte Regressionskoeffizienten

- β hängt von der Maßeinheit von X und Y ab
- Um Vergleichbarkeit herzustellen, wählt man manchmal die Standardabweichung als Maßeinheit
  - Standardisierung von Y und X (Z-Transformation)

$$y_i^* = \frac{y_i - \bar{y}}{s_Y}, \quad x_i^* = \frac{x_i - \bar{x}}{s_X}$$

  - Die Regressionsgleichung lautet nun

$$y_i^* = \alpha^* + \beta^* x_i^* + \varepsilon_i^*$$

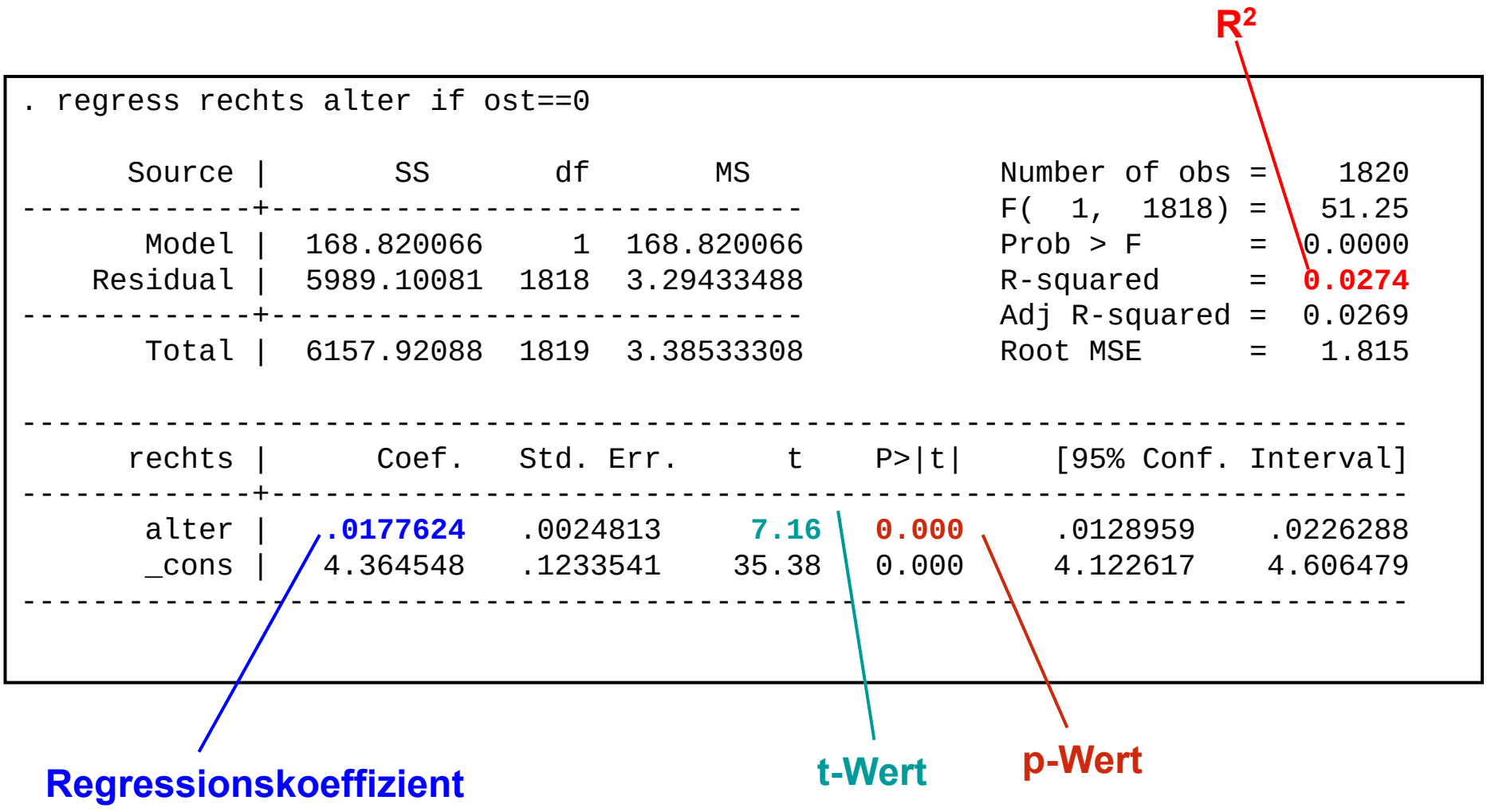
  - Für die standardisierten Regressionskoeffizienten ergibt sich

$$\hat{\alpha}^* = \bar{y}^* - \hat{\beta}^* \bar{x}^* = 0$$

$$\hat{\beta}^* = \frac{s_{X^*Y^*}}{s_{X^*}^2} = r$$

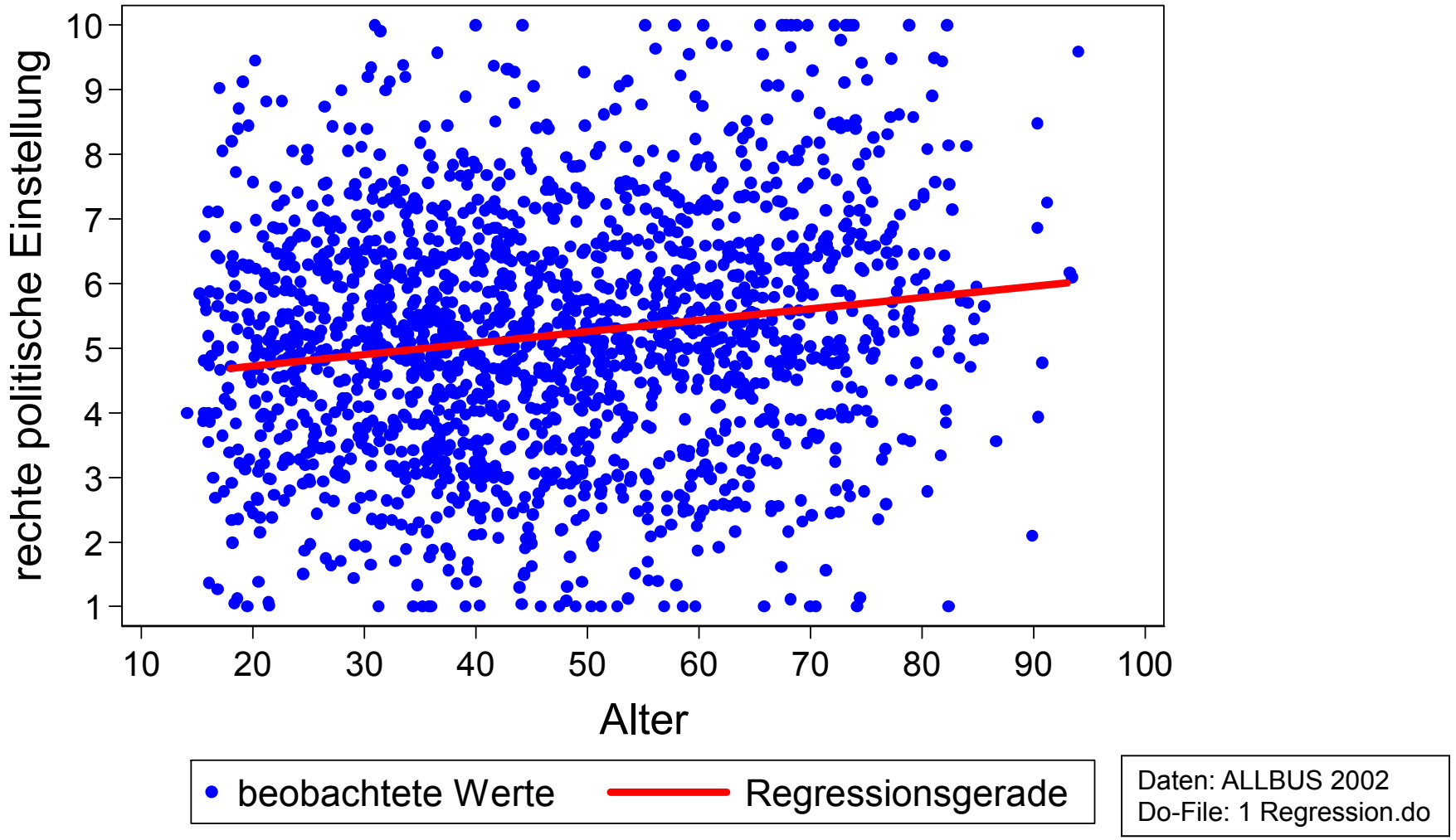
  - Beispiel „Abinote und Klausurerfolg“: r = -0,43
    - Steigt die Note um eine Standardabweichung, so verringert sich die Punktzahl um 0,43 Standardabweichungen

# Stata-Bsp.: Politische Einstellung auf Alter



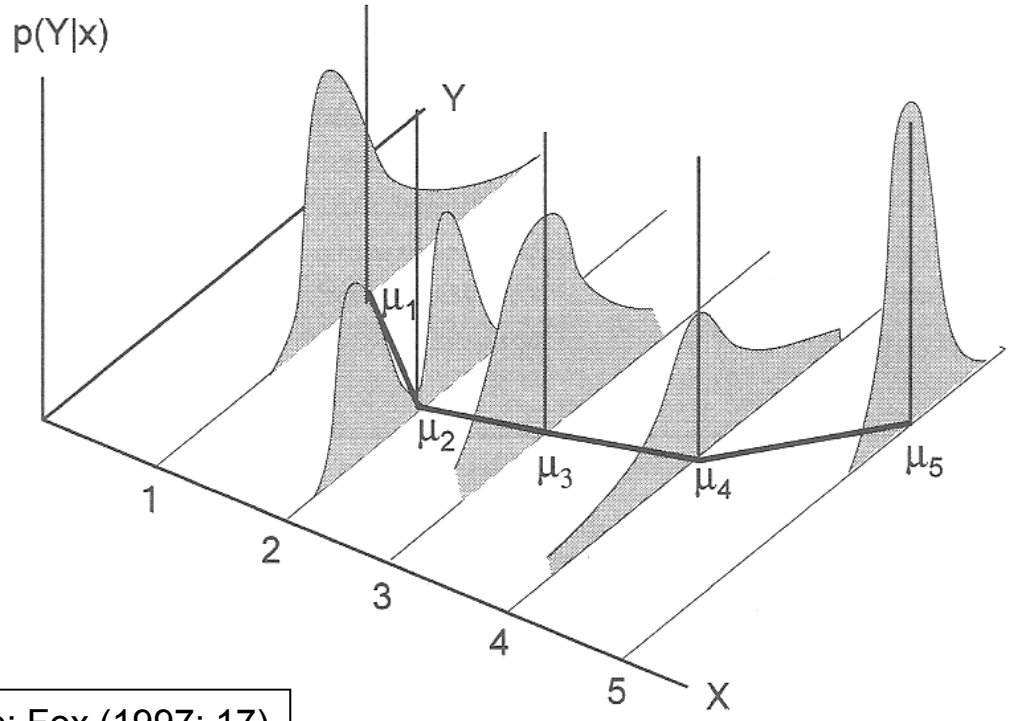

Daten: ALLBUS 2002
Do-File: 1 Regression.do

# Regression: politische Einstellung auf Alter



# Exkurs: Regression als bedingte Verteilung

- Zwei Variablen Y und X
  - mit Realisierungen $(y_i, x_i)$, für i=1,…,n
- Die Regression von Y auf X
  - ist die bedingte Verteilung: $f(Y \mid X=x)$



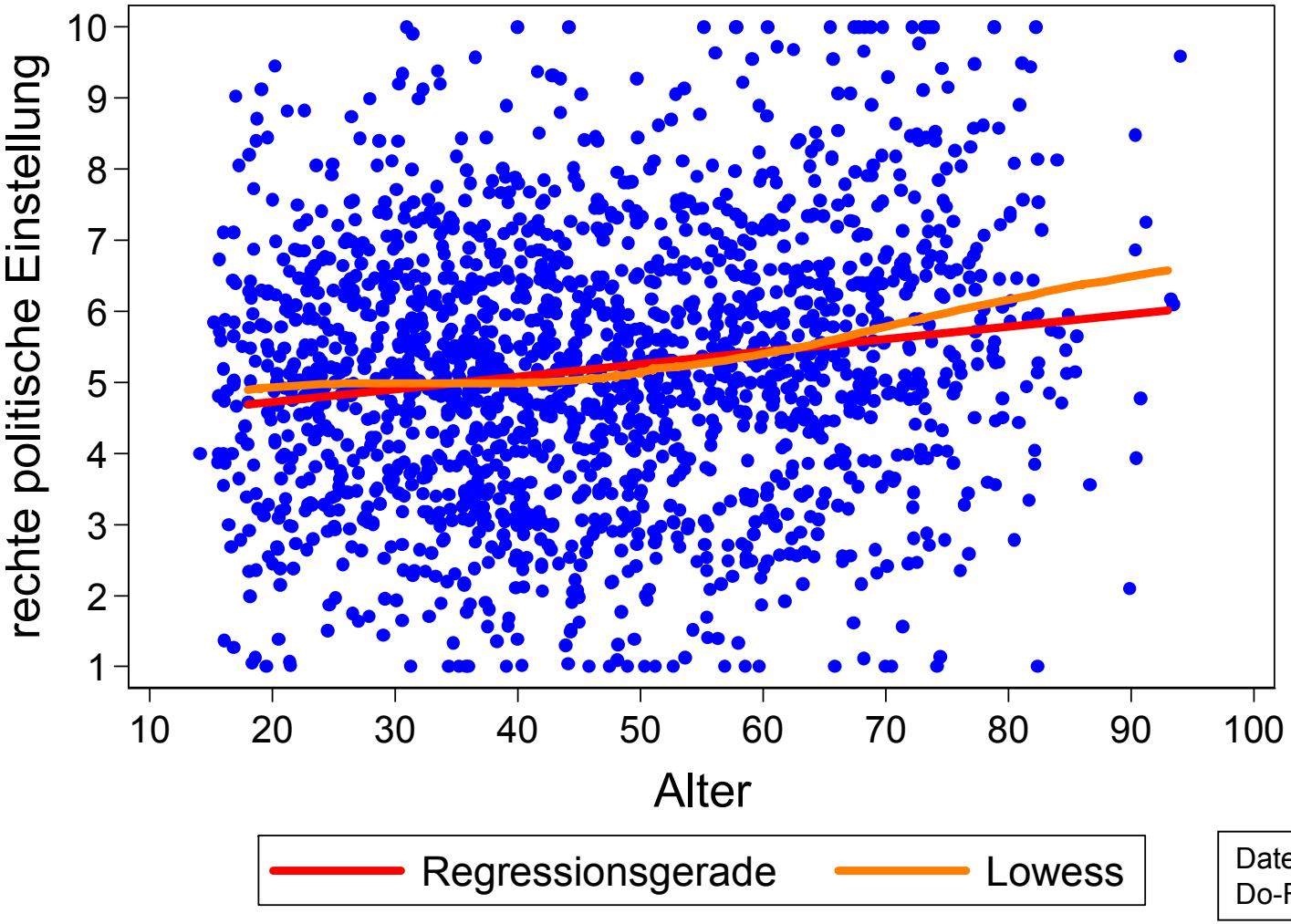

Quelle: Fox (1997: 17)

- X ist hier kategorial.
- Die bedingten Verteilungen haben sehr unterschiedliche Form.
- Eingetragen sind auch die bedingten Mittelwerte. Der Zusammenhang ist U-förmig.

# Exkurs: Regression als bedingte Verteilung

- Eine solche allgemeine Regression enthält zu viel Information
- Informationsreduktion:
  Charakterisierung der Verteilung durch geeignete Kennzahlen
  - Y metrisch: bedingtes arithmetisches Mittel
  - Y metrisch, ordinal: bedingtes Quantil
  - Y nominal: bedingte Häufigkeiten (Kreuztabelle)
- Nicht-parametrische Regression
  - Benutze die Y-Werte in einer Umgebung von x zur Berechnung der Kennzahl (local averaging)
    - Lokale mean (median) Regression
    - Lowess Smoother
- Parametrische Regression
  - Weitere Informationsreduktion: man nimmt an, dass die bedingten Kennzahlen einer Funktion folgen
    - Lineare Mittelwertsregression (OLS Regression)
    - Lineare Medianregression
    - Quantilsregression

# Nicht-parametrische und parametrische Regression




Daten: ALLBUS 2002
Do-File: 1 Regression.do