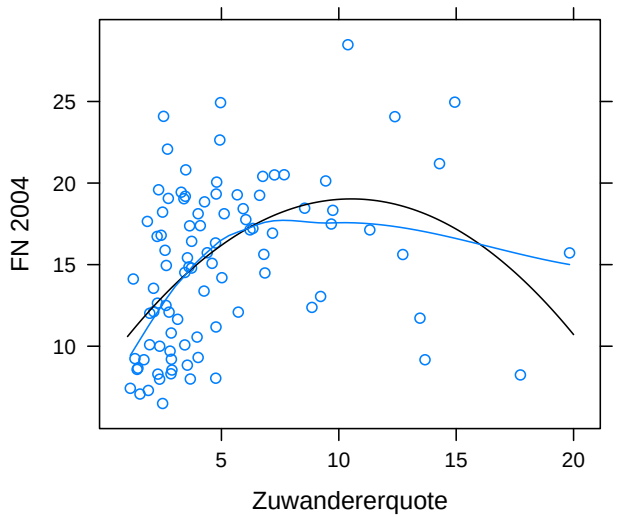
Regression III

Statistik I

Sommersemester 2009



Zum Nachlesen

- Agresti/Finlay Kapitel 10.2, 10.3, 11.1

Informationen zur Klausur

- „Hilfsmittel“
  - Einfacher Taschenrechner, d. h. keine Programmier- oder Grafikfunktionen
  - Formelsammlung (wird zur Verfügung gestellt)
- Bei Rechenaufgaben bitte runden (Zahl der geforderten Kommastellen wird angegeben)
- e-Klausur: Anmeldung notwendig
  - Einloggen bei Ilias mit ZDV-Account: `https://www.e-learning.uni-mainz.de/ilias3/login.php`
  - „Beitreten“ zum Kurs „Einführungsmodul“ in der Kategorie „Powi Einführungsmodul“
  - Paßwort zum Beitreten: `Schreibtisch`

Ilias


JOHANNES GUTENBERG UNIVERSITÄT MAINZ
eLearning - Zentrum für Datenverarbeitung
Angemeldet als Kai Arzheimer
Abmelden
Persönlicher Schreibtisch
Magazin
Suche
Mail (1 Neu)
Zuletzt besucht
Magazin > FB 02 - Sozialwissenschaften, Medien und Sport > Institut für Politikwissenschaft > POWI Einführungsmodul > Einführungsmodul
Einführungsmodul
Inhalt
Info
Mitglieder
Lernfortschritt
Kursinhalt
Probeklausur Magister (Copy)
Info
Auf den Schreibtisch
powered by ILIAS (v3.9.9 2009-04-19)

Fit

- R^2:
  - Wieviel Prozent der Gesamtvarianz von y werden vom Modell erklärt
  - Relatives Maß für die Anpassung des Modells an die Daten
- Root Mean Squared Error
  - Mittelwert der Fehler $= 0$
  - Standardabweichung der Fehler = Root Mean Squared Error = Standard Error of the Estimate
  - In der ursprünglichen Einheit – absolutes Maß für Qualität der Prognose

Residuen und einflußreiche Fälle

- Wie stark weicht Prognose von realem Wert ab $\rightarrow$ Residuum des Falles (ggf. studentisieren)
- Einfluß auf Modellschätzungen
  - Wie weit ist der Fall vom Zentrum der x-Werte entfernt (Hebel)?
  - Wie weit weicht der Fall von der Regressionslinie ab (Residuum, Berechnung ohne diesen Fall)
  - Produkt: Einfluß auf Regressionslinie

Warum mehr als eine unabhängige Variable?

- Modell – radikale Vereinfachung
- 1:1 Beziehung x – y zu starke Vereinfachung?
- (Fast) alle sozialen Prozesse multikausal

Warum mehr als eine unabhängige Variable?

- Modell – radikale Vereinfachung
- 1:1 Beziehung x – y zu starke Vereinfachung?
- (Fast) alle sozialen Prozesse multikausal
- Zweite mögliche Erklärung für Mordquote: ethnische Homogenität
  - Spannungen zwischen ethnischen Gruppen
  - Höhere Kriminalität innerhalb von Minderheiten
  - Höhere Wahrscheinlichkeit für schwarze Verdächtige verurteilt zu werden

Wie sieht ein Modell mit zwei unabhängigen Variablen aus?

Multivariate Regression

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$$

- y als Funktion von x_1 (z. B. Armut) und x_2 (z. B. Homogenität)
- Beide Variablen haben unabhängig voneinander Einfluß
- Additives Zusammenwirken
- Nicht gegenseitige Verstärkung von Armut und ethnischen Konflikten

Was sind die Ergebnisse?

Multivariates Modell

$$\text{„Morde“} = 36.2 + 0.8 \times \text{„Armut“} - 0.46 \times \text{„Prozent Weiße“}$$

Was sind die Ergebnisse?

Multivariates Modell

$$\text{„Morde“} = 36.2 + 0.8 \times \text{„Armut“} - 0.46 \times \text{„Prozent Weiße“}$$

- Für einen Bundesstaat ohne weiße Bevölkerung und ohne Armut werden 36 Morde / 100 000 Einwohner erwartet
- („Prozent Weiße“ ein guter Indikator für Homogenität?)
- Für jeden Prozentpunkt mehr Armut werden 0.8 Morde mehr erwartet
- Für jeden Prozentpunkt mehr weiße Bevölkerung werden 0.46 Morde *weniger* erwartet

Was sind die Ergebnisse?

Multivariates Modell

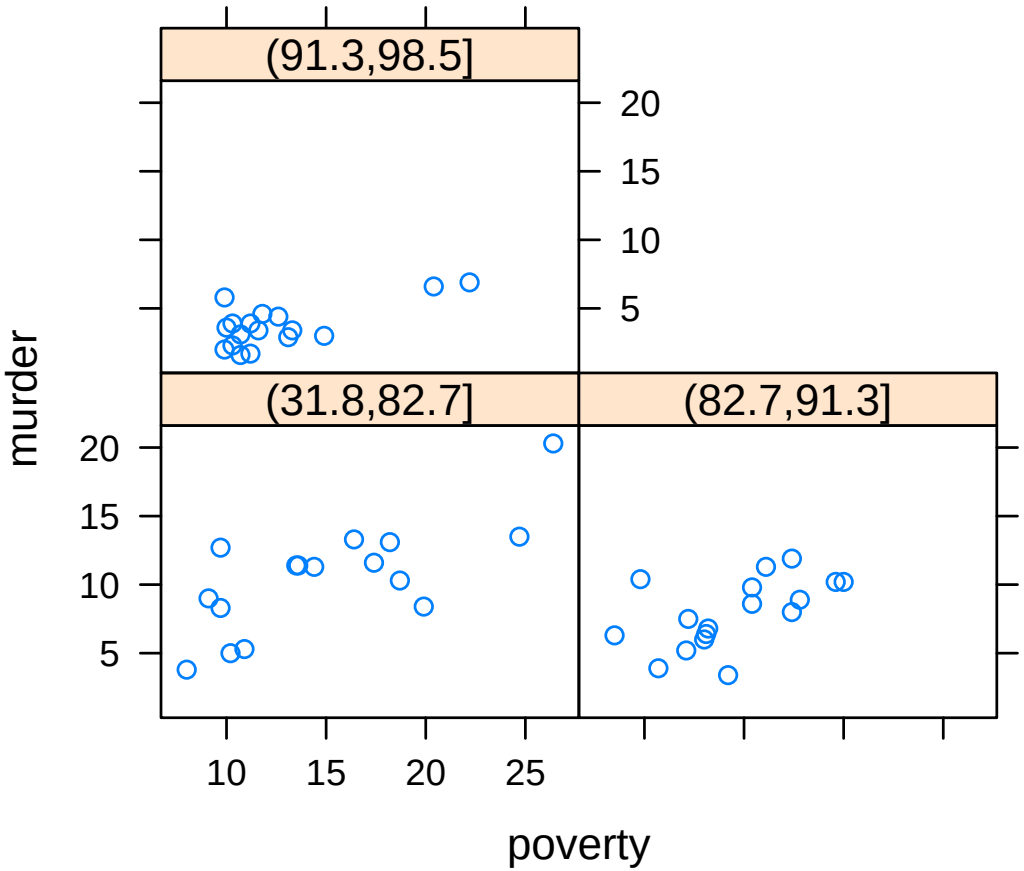
$$\text{„Morde“} = 36.2 + 0.8 \times \text{„Armut“} - 0.46 \times \text{„Prozent Weiße“}$$

- Für einen Staat mit mittlerer Armut (13%) und mittlerem Anteil von Weißen (87%) werden etwas mehr als 6 Morde erwartet
- Für einen extrem armen (26.4%) Staat mit einer extrem gemischten (32%) Bevölkerung wie Washington D. C. werden ca. 43 Morde erwartet
- (Immer noch knapp die Hälfte weniger als beobachtet)

Wie kann man sich das vorstellen?

- Unabhängig vom Niveau von „Homogenität“ (x_2) hat x_1 stets den gleichen positiven Effekt
- Unabhängig vom Niveau von „Armut“ (x_1) hat x_2 stets den gleichen negativen Effekt

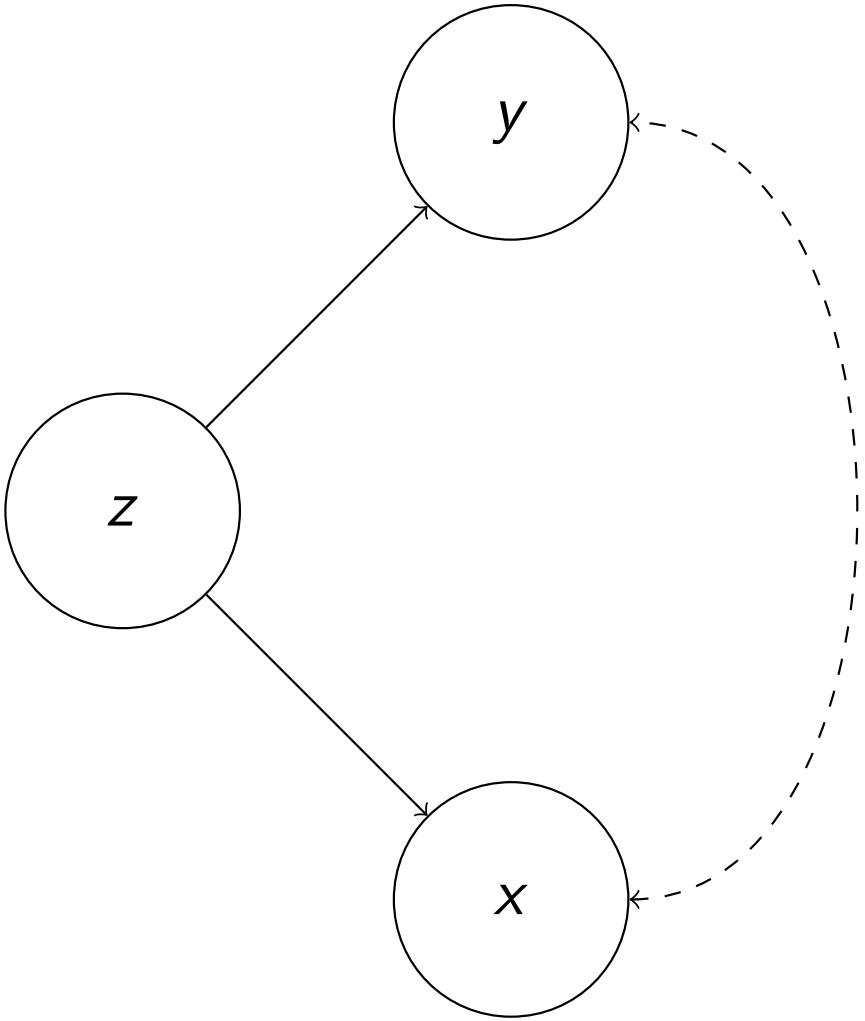
Graphisch



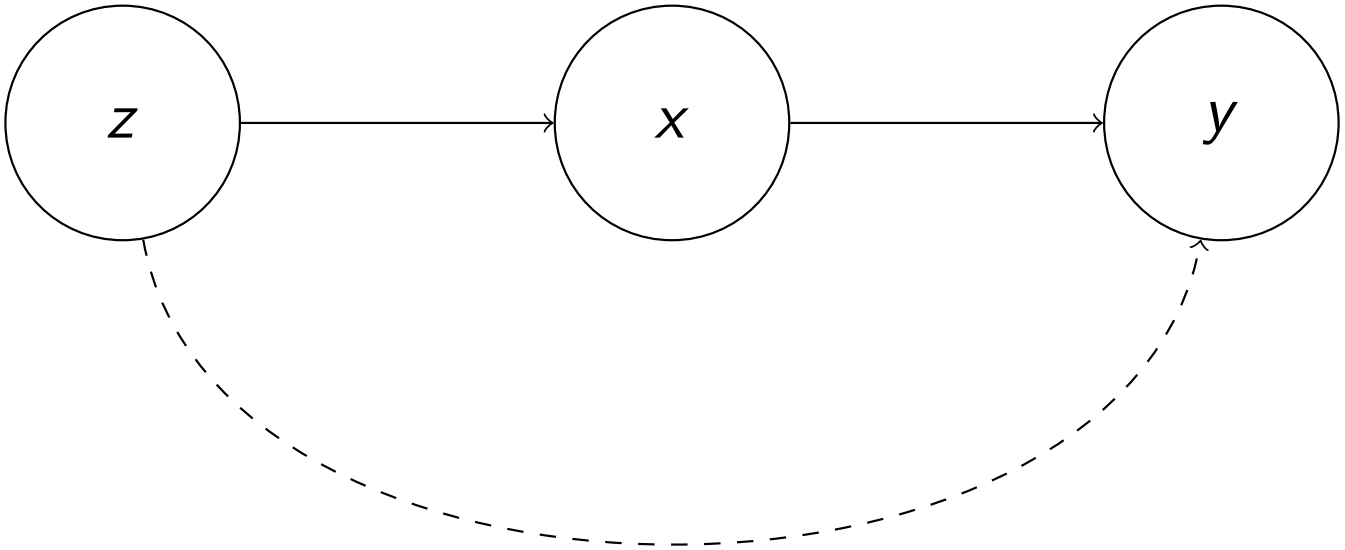
Welche Beziehungen können zwischen drei Variablen bestehen?

1. „Scheinkorrelation“/„scheinbare Non-Korrelation“
2. Mediatorvariable
3. Multiple Verursachung
4. Interaktion
5. ...

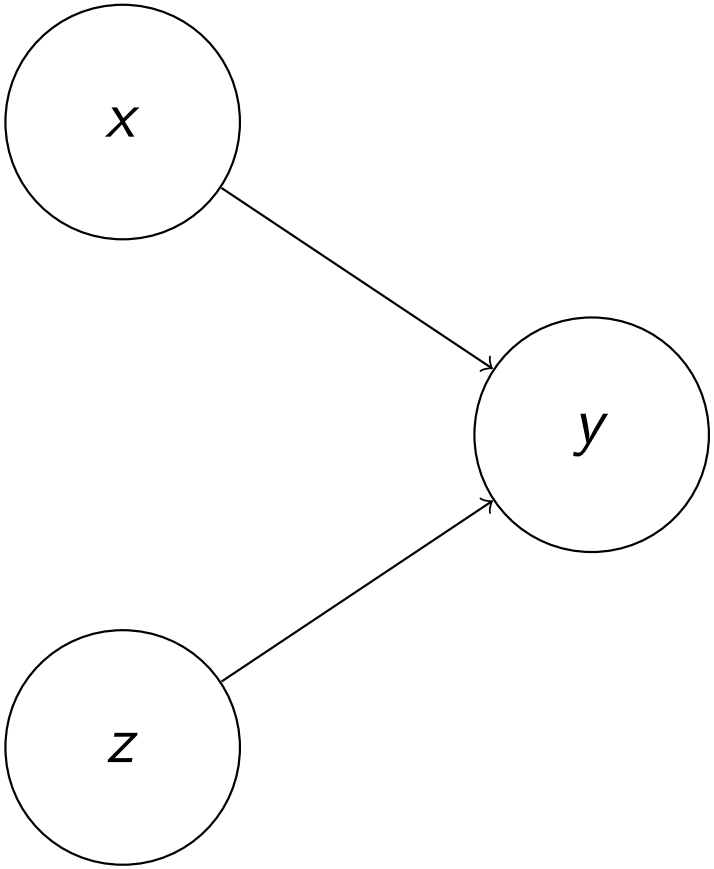
1. „Scheinkorrelation“



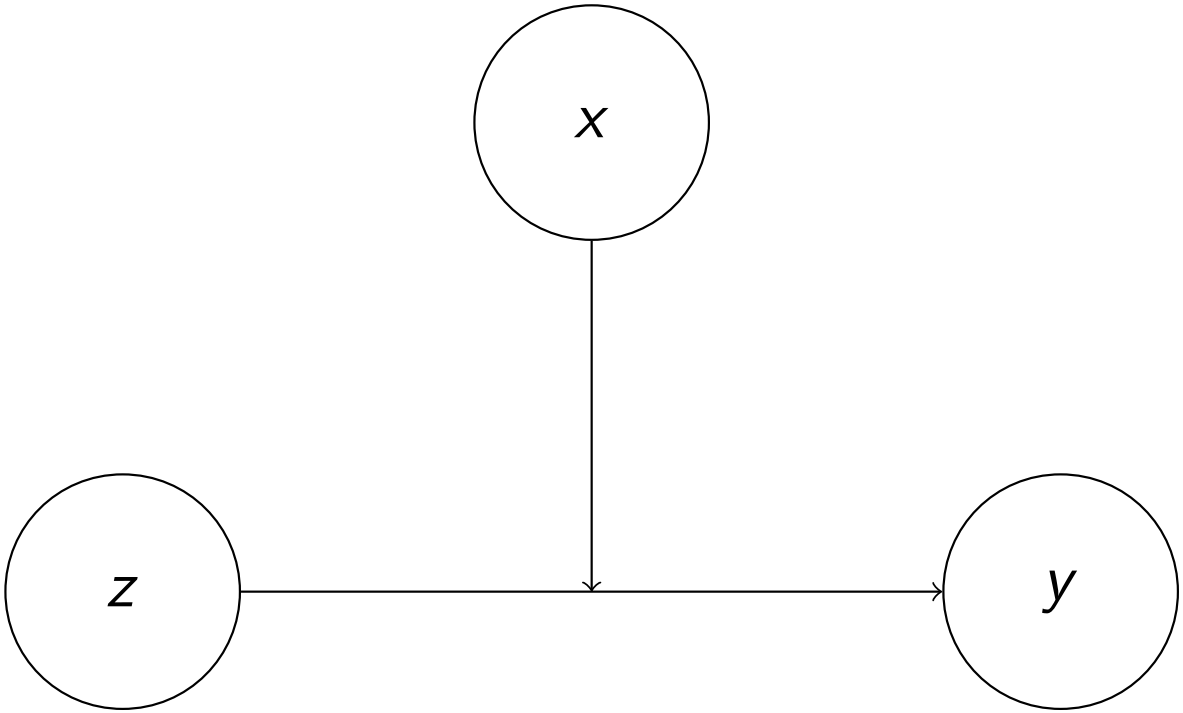
2. „Mediatorvariable“



3. „Multiple Verursachung“



4. „Interaktion“



Was tun mit „Scheinkorrelationen“?

- Hintergrundvariable z beeinflußt x und $y \longrightarrow$ „scheinbarer“ Zusammenhang zwischen x und y
- Kein Problem im Experiment
  - x von uns gesetzt; zufällige Zuweisung zu $x = 0; x = 1$
  - Korrelation zwischen z und x wird aufgebrochen
  - Verbleibende Effekte von x (wenn es sie gibt) kausal zu interpretieren
  - „Kontrolle“ von z durch zufällige Zuweisung: Mittelwerte von z für $x = 0; x = 1$ gleich
- Multivariate Modelle ermöglichen sogenannte „Drittvariablenkontrolle“ $\longrightarrow$ statistische Kontrolle von multipler Verursachung, Mediation, Scheinkorrelation

Was bedeutet „statistische Kontrolle“?

- *Wenn z* bekannt und gemessen, *ex post* Kontrolle durch multivariates Modell möglich
- Ist es wirklich Armut, die Einfluß auf Verbrechen hat?
- Oder werden Armut und Verbrechen vom Zerfall der Familien beeinflußt?

Mord, Armut, single moms

$$\text{Mord} = -10.1 + 1.32 \times \text{poverty}; R^2 = 0.31$$

Was bedeutet „statistische Kontrolle“?

- *Wenn z* bekannt und gemessen, *ex post* Kontrolle durch multivariates Modell möglich
- Ist es wirklich Armut, die Einfluß auf Verbrechen hat?
- Oder werden Armut und Verbrechen vom Zerfall der Familien beeinflußt?

Mord, Armut, single moms

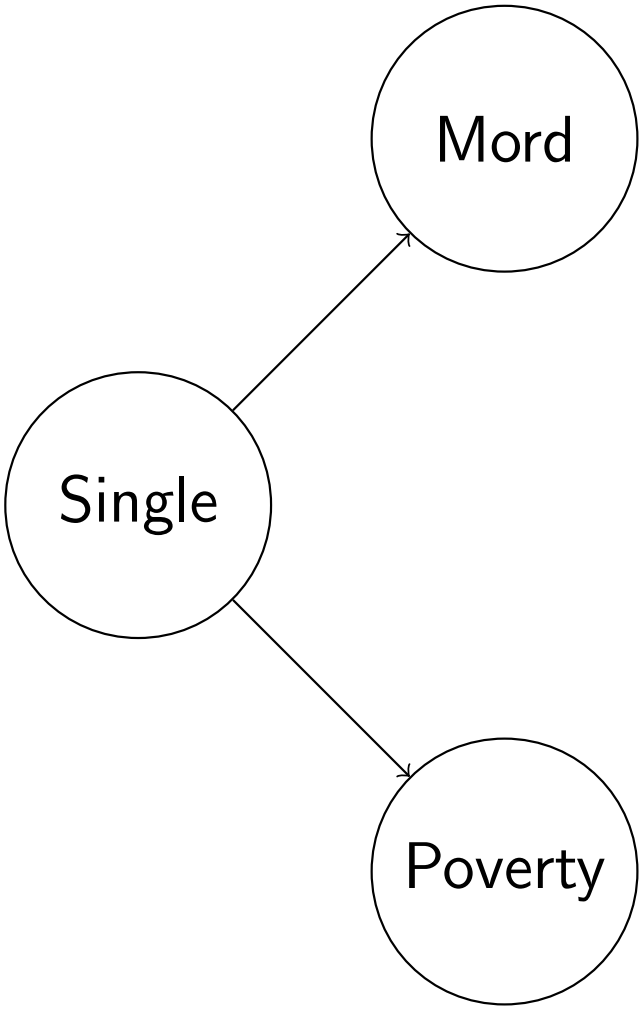
$$\text{Mord} = -10.1 + 1.32 \times \text{poverty}; R^2 = 0.31$$

$$\text{Mord} = -40.7 + 0.3 \times \text{poverty} + 4.0 \times \text{lone parent}; R^2 = 0.74$$

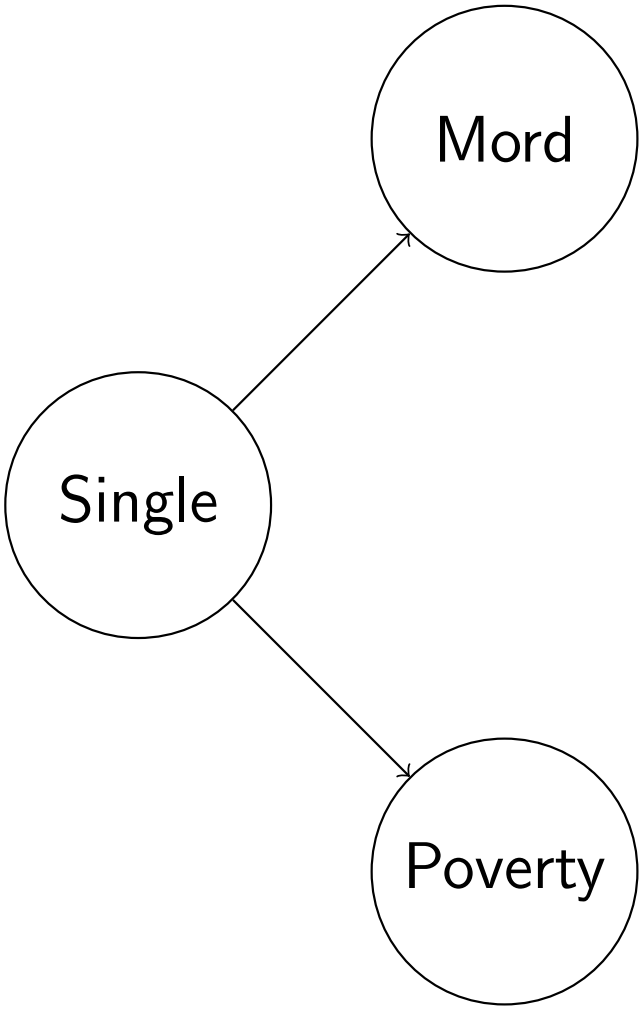
Wie kommt diesese Ergebnis zustande?

- „Lone Parents“: starker Effekt auf „Mord“; moderater Effekt auf „Armut“
- Weil „Lone Parents“ beide anderen Variablen beeinflußt: moderater Zusammenhang

„Scheinkorrelation“

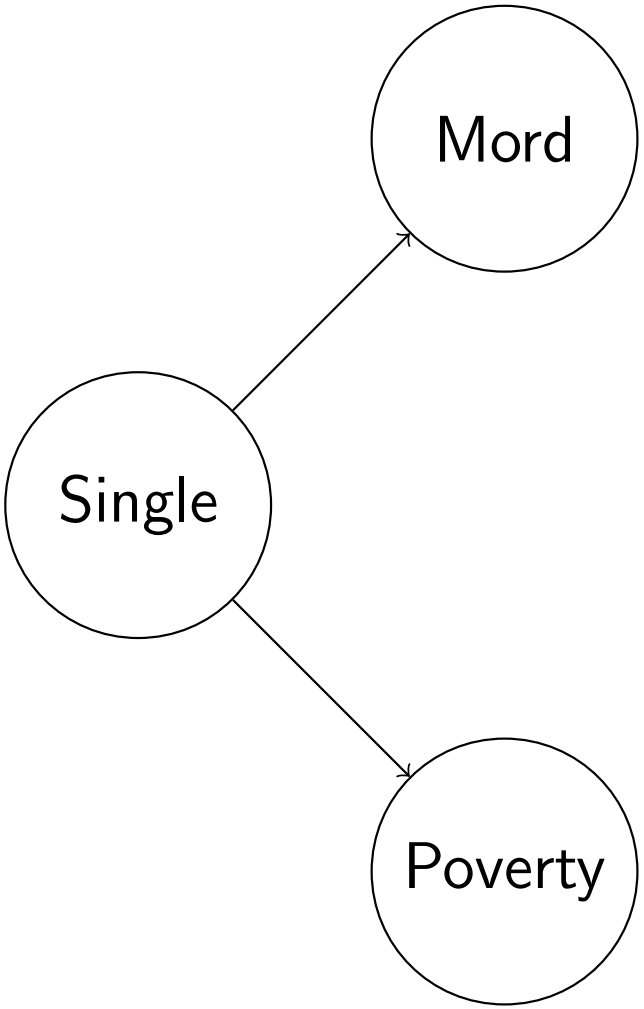


„Scheinkorrelation“



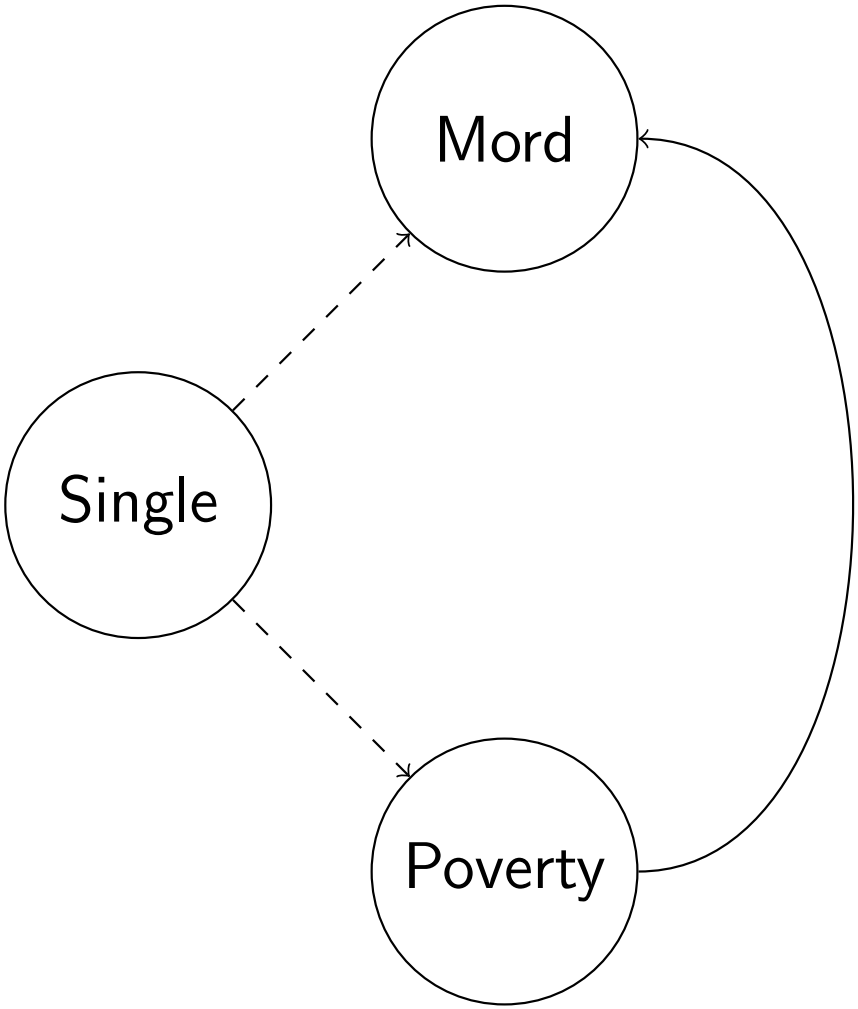
Mord = -40.4 + 4.3 × lone parent;
$R^2 = 0.74$

„Scheinkorrelation“



Mord = -40.4 + 4.3 × lone parent;
$R^2 = 0.74$
Poverty = 0.8 + 1.2 × lone parent;
$R^2 = 0.30$

„Scheinkorrelation“

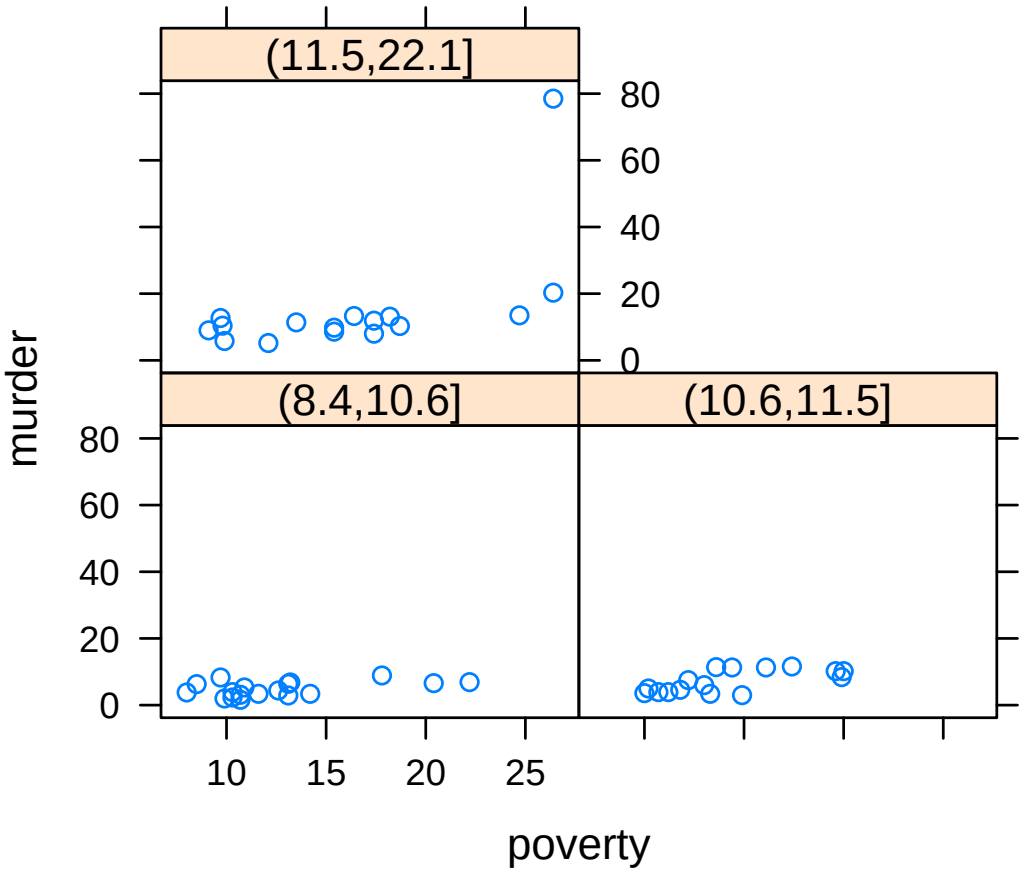


$\text{Mord} = \text{-10.1} + 1.3 \times \text{poverty};$
$R^2 = 0.32$

Wie kommt diesese Ergebnis zustande?

- „Lone Parents“: starker Effekt auf „Mord“; moderater Effekt auf „Armut“
- Weil „Lone Parents“ beide anderen Variablen beeinflußt: moderater Zusammenhang
- Für verschiedene Niveaus von „Lone Parents“ kaum Zusammenhang zwischen „Armut“ und „Mord“

Familien, Armut, Mord: Statistische Kontrolle



Wie kommt diesese Ergebnis zustande?

- „Lone Parents“: starker Effekt auf „Mord“; moderater Effekt auf „Armut“
- Weil „Lone Parents“ beide anderen Variablen beeinflußt: moderater Zusammenhang
- Für verschiedene Niveaus von „Lone Parents“ kaum Zusammenhang zwischen „Armut“ und „Mord“
- Multivariate Regression: Schätzt Effekt von „single“ für alle denkbaren Niveaus von „poverty“ (und umgekehrt)
- „Konstant halten“ = „statistische Kontrolle“
- Trotzdem Vorsicht mit kausalen Interpretationen

Warum multivariate Regression?

- Scheinkorrelation, multiple Verursachung, Mediatorvariable: bivariate Ergebnisse führen in die Irre
- Effekte zu stark, zu schwach, umgekehrte Vorzeichen: statistische Kontrolle als mächtiges Werkzeug
- Aber: additives Zusammenwirken vs. Interaktion

Was ist Interaktion?

- Zwei (oder mehr) unabhängige Variablen wirken *nicht* additiv zusammen
- Wirkung von x_1 hängt ab vom Niveau von x_2 (und umgekehrt)
- Beispiel: Wirkungen verstärken sich gegenseitig
- Bzw. x_2 hat nur einen Effekt, wenn Minimum von x_1 vorhanden

Beispiel: Arbeitslosigkeit, Einwanderer, Wahl des FN 2004

- Aggregatdaten: Französische Regionalwahlen 2004
- Einheiten: 94 Départements auf dem französischen Festland

- Zusammenhang zwischen FN-Ergebnis, Arbeitslosigkeit, Zuwanderung, Interaktion???

Wie macht man das?

- Interaktion = Zusammenwirken von zwei Variablen
- Bildung einer zusätzlichen Variable durch Multiplikation der Ausgangsvariablen
- $Z = x_1 \times x_2$
  - $Z = 0$ wenn $x_1 = 0$ und $x_2 = 0$
  - $Z > 0$ wenn $x_1 > 0$ und $x_2 > 0$ oder wenn $x_1 < 0$ und $x_2 < 0$
  - $Z < 0$ wenn $x_1 > 0$ und $x_2 < 0$ oder wenn $x_1 < 0$ und $x_2 > 0$
- „Künstliche“ Variable Z bildet Zusammenwirken von x_1 und x_2 ab
- Eventuell x_1 und x_2 vorab zentrieren (Interpretation?)

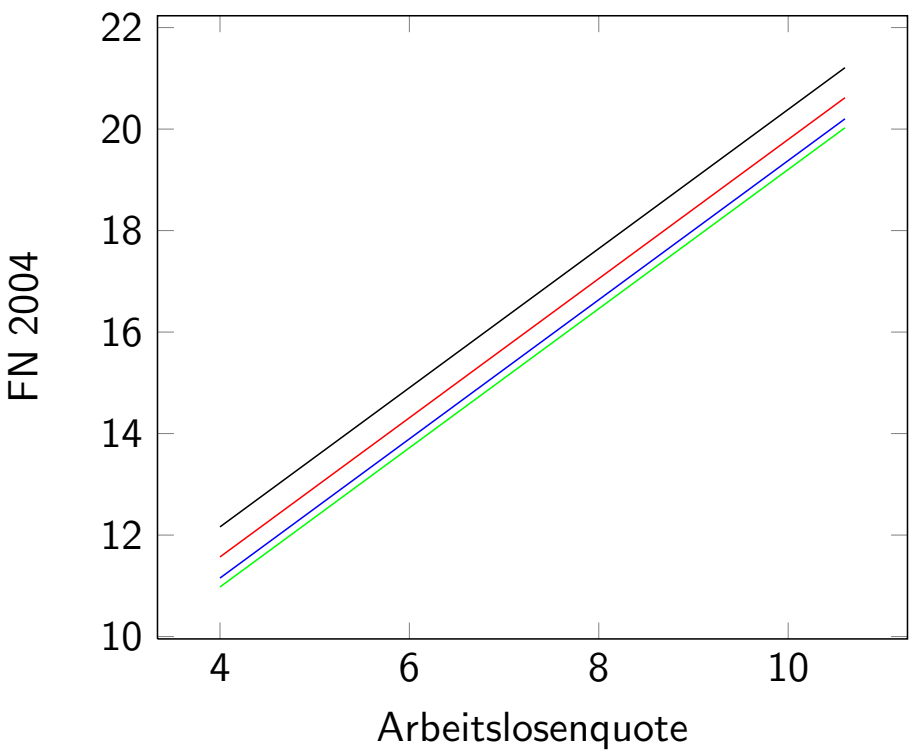
Im Beispiel

- Arbeitslosigkeit und Zuwandereranteil stets $> 0 \rightarrow$ Interaktionsvariable stets größer 0
- Koeffizient für $Z = 0 \rightarrow$ keine Interaktion, additives Zusammenwirken
- Koeffizient für $Z > 0 \rightarrow$ positive Interaktion
  - Gegenseitige Verstärkung; großer Effekt von Arbeitslosigkeit wenn viel Zuwanderung und umgekehrt
  - Schwacher Effekt von Arbeitslosigkeit wenn wenig Zuwanderung
- Koeffizient für $Z < 0 \rightarrow$ negative Interaktion
  - Schwacher Effekt von Arbeitslosigkeit wenn viel Zuwanderung und umgekehrt
  - Starker Effekt von Arbeitslosigkeit wenn wenig Zuwanderung
- Nicht kompliziert, nur etwas komplex

Was kommt raus?

Modell ohne Interaktion

$$\text{FN2004} = 5.1 + 1.4 \times \text{Arbeitslosenquote} + 0.2 \times \text{Zuwandererquote}$$



Zuwandererquoten: 2.6% (grün), 3.7%, 6.3%, 10% (schwarz); $R^2 = 0.19$

Was kommt raus? II

Modell mit Interaktion

FN 2004 = -3.2 + 2.5 × Arbeitslosenquote + 1.7 × Zuwandererquote -0.2 × Arbeitslosenquote × Zuwandererquote

Was kommt raus? II

Modell mit Interaktion

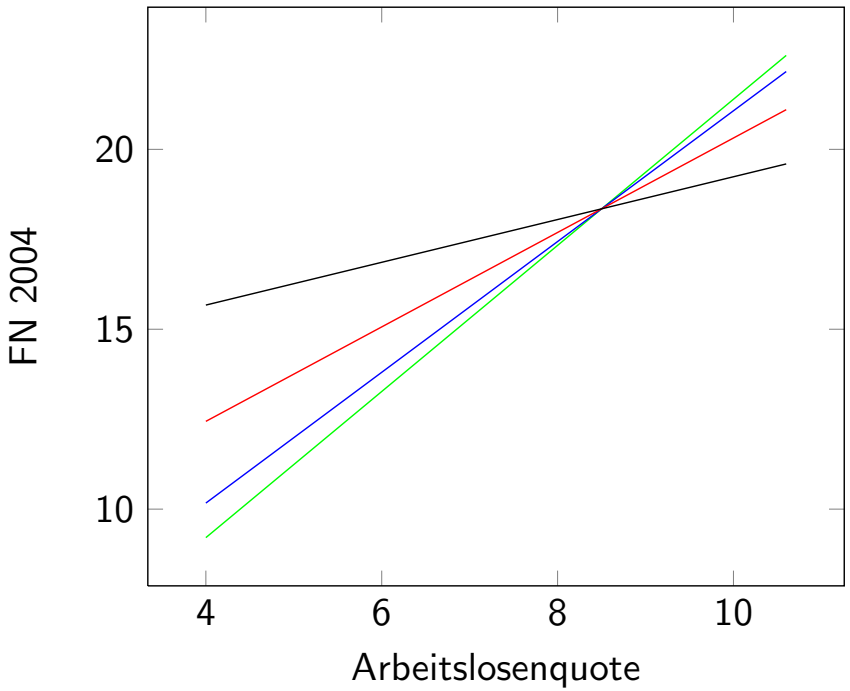
FN 2004 = -3.2 + 2.5 × Arbeitslosenquote + 1.7 × Zuwandererquote -0.2 × Arbeitslosenquote × Zuwandererquote

- 2.5: Effekt der Arbeitslosenquote wenn Zuwandererquote = 0
- 1.7: Effekt der Zuwandererquote wenn Arbeitslosenquote = 0
- Für alle anderen Konstellationen: einsetzen und ausrechnen

Was kommt raus? II

Modell mit Interaktion

FN 2004 = -3.2 + 2.5 × Arbeitslosenquote + 1.7 × Zuwandererquote -0.2 × Arbeitslosenquote × Zuwandererquote



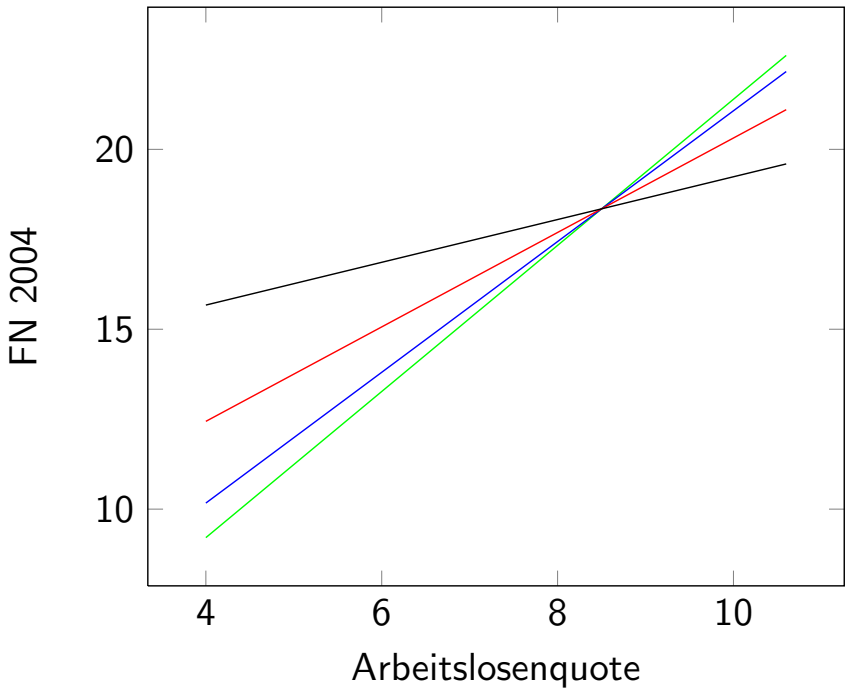
Zuwandererquoten: 2.6% (grün), 3.7%, 6.3%, 10% (schwarz); $R^2 = 0.25$

- Interpretation?
- Effekt der ALQ bei niedriger Zuwanderung am stärksten
- Effekt Zuwanderung stark und positiv bei niedriger ALQ

Was kommt raus? II

Modell mit Interaktion

FN 2004 = -3.2 + 2.5 × Arbeitslosenquote + 1.7 × Zuwandererquote -0.2 × Arbeitslosenquote × Zuwandererquote



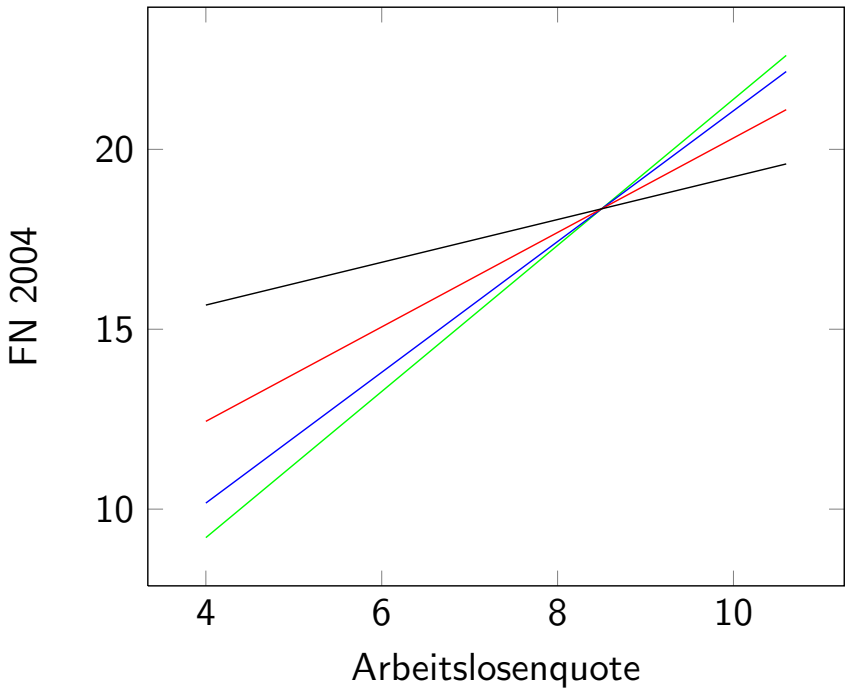
Zuwandererquoten: 2.6% (grün), 3.7%, 6.3%, 10% (schwarz); $R^2 = 0.25$

- Interpretation?
- Effekt der ALQ bei niedriger Zuwanderung am stärksten
- Effekt Zuwanderung stark und positiv bei niedriger ALQ
- Schwächt sich ab mit höherer ALQ

Was kommt raus? II

Modell mit Interaktion

FN 2004 = -3.2 + 2.5 × Arbeitslosenquote + 1.7 × Zuwandererquote -0.2 × Arbeitslosenquote × Zuwandererquote



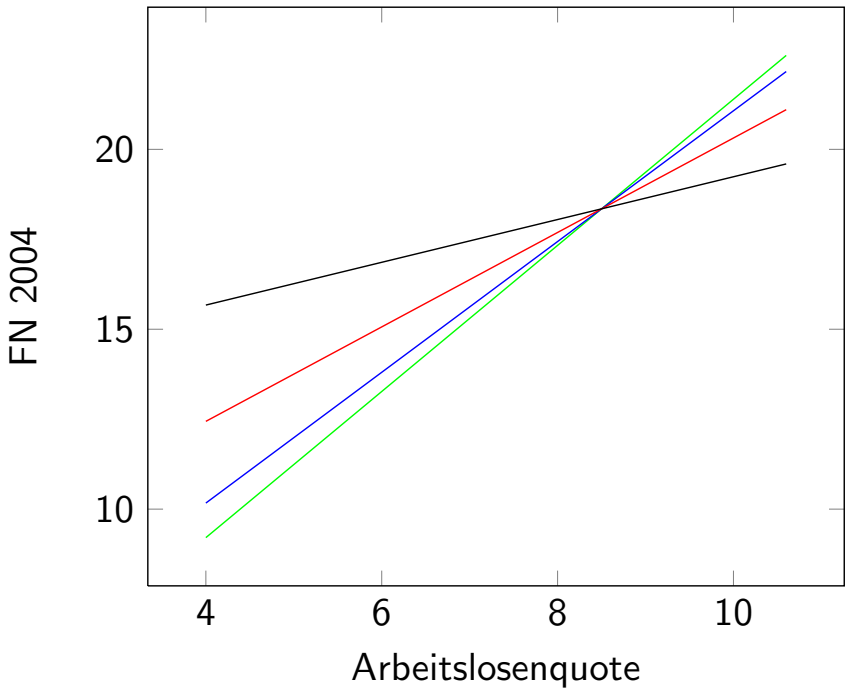
Zuwandererquoten: 2.6% (grün), 3.7%, 6.3%, 10% (schwarz); $R^2 = 0.25$

- Interpretation?
- Effekt der ALQ bei niedriger Zuwanderung am stärksten
- Effekt Zuwanderung stark und positiv bei niedriger ALQ
- Schwächt sich ab mit höherer ALQ
- *Negativ* bei ALQ > 8.5 (1.7/0.2)

Was kommt raus? II

Modell mit Interaktion

FN 2004 = -3.2 + 2.5 × Arbeitslosenquote + 1.7 × Zuwandererquote -0.2 × Arbeitslosenquote × Zuwandererquote



Zuwandererquoten: 2.6% (grün), 3.7%, 6.3%, 10% (schwarz); $R^2 = 0.25$

- Interpretation?
- Effekt der ALQ bei niedriger Zuwanderung am stärksten
- Effekt Zuwanderung stark und positiv bei niedriger ALQ
- Schwächt sich ab mit höherer ALQ
- *Negativ* bei ALQ > 8.5 (1.7/0.2)
- Warum?

Zwischenfazit

- Berechnung der Parameter mechanische Prozedur, fast immer möglich
- Qualität: Fit, Prognosefähigkeit, Ausreißer, einflußreiche Werte
- Schätzungen führen in die Irre (bias), wenn relevante Drittvariablen nicht kontrolliert werden

Was sind nicht-lineare Effekte?

- Auch bei Interaktion lineare Wirkung von $x_1, x_2, x_1 \times x_2$
- Nicht immer plausibel
- Wichtige Alternative: kurvilinearer (U-förmiger) Zusammenhang
- Wie macht man das?

Was sind nicht-lineare Effekte?

- Auch bei Interaktion lineare Wirkung von $x_1, x_2, x_1 \times x_2$
- Nicht immer plausibel
- Wichtige Alternative: kurvilinearer (U-förmiger) Zusammenhang
- Wie macht man das?
- Wieder „künstliche“ Variable: x mit sich selbst multiplizieren $\longrightarrow x^2$
- x und x^2 müssen in Modell enthalten sein, sonst zu unflexibel

Beispiel: FN 2004 und Zuwandererquote, quadratischer Effekt

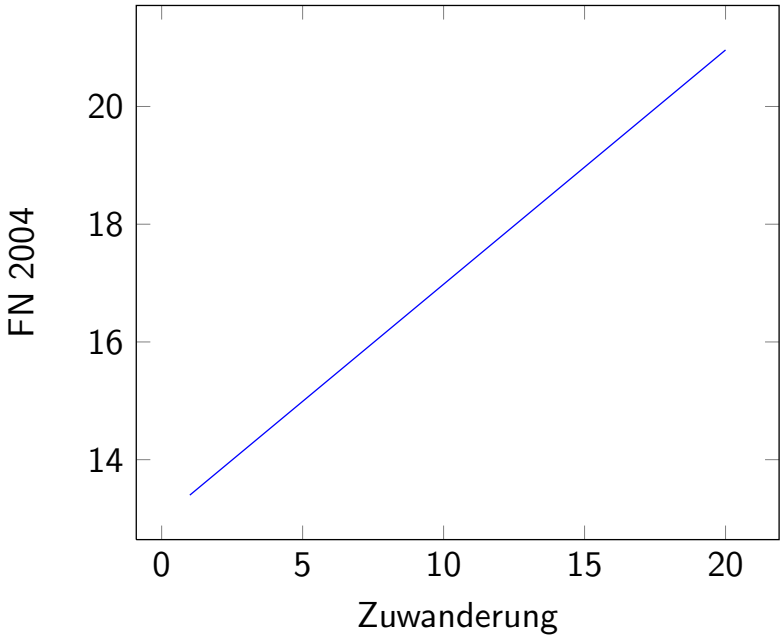
- Effekt der Zuwandererquote nicht linear – warum?

Beispiel: FN 2004 und Zuwandererquote, quadratischer Effekt

- Effekt der Zuwandererquote nicht linear – warum?

Einfaches Modell

FN 2004 $= 13 + 0.4 \times$ Zuwandererquote; $R^2 = 0.08$



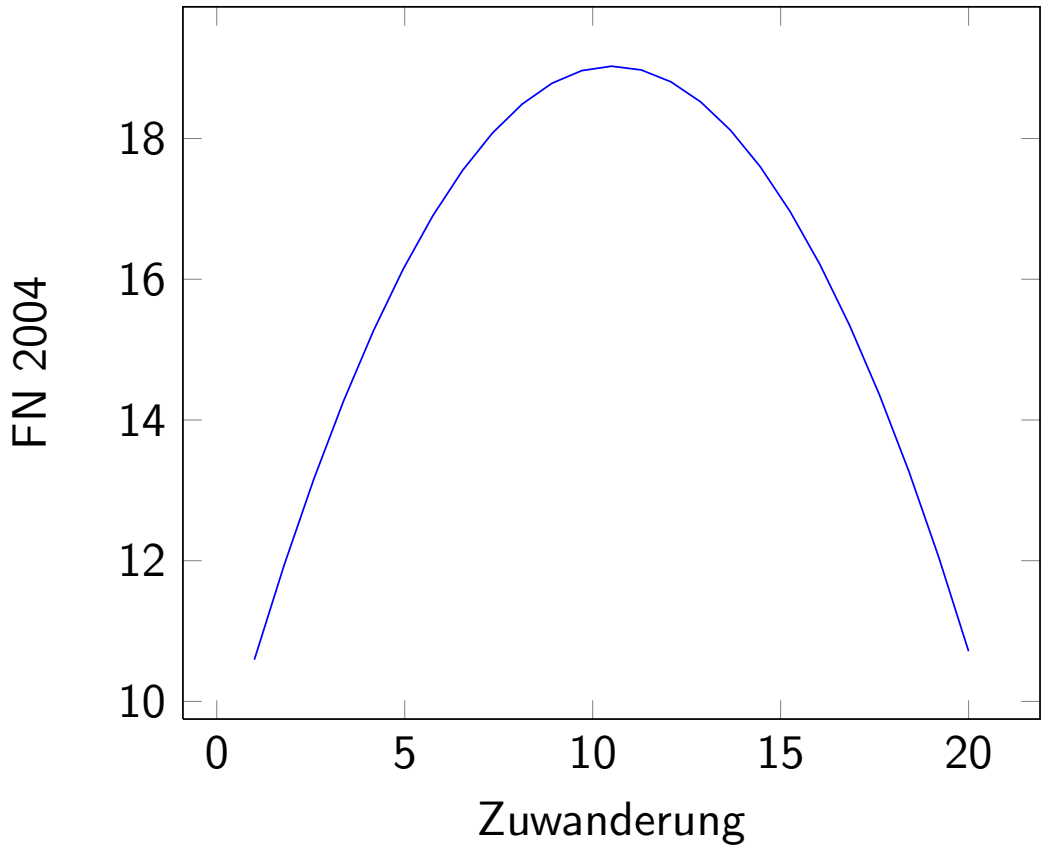
Beispiel: FN 2004 und Zuwandererquote, quadratischer Effekt

Modell mit quadratischem Effekt

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\text{FN 2004} = 8.73 + 1.96 \times \text{Zuwanderer} - 0.09 \times \text{Zuwanderer}^2$$

Beispiel: FN 2004 und Zuwandererquote, quadratischer Effekt



Zusammenfassung

- Alle sozialen Prozesse multikausal – multivariate Modelle
- Bivariate Zusammenhänge: potentieller bias
- Multivariate Modelle besser
- Nicht lineare Zusammenhänge
  - Parametrisch, z. B. mit Polynomen
  - Non-parametrisch: z. B. lokale Regression