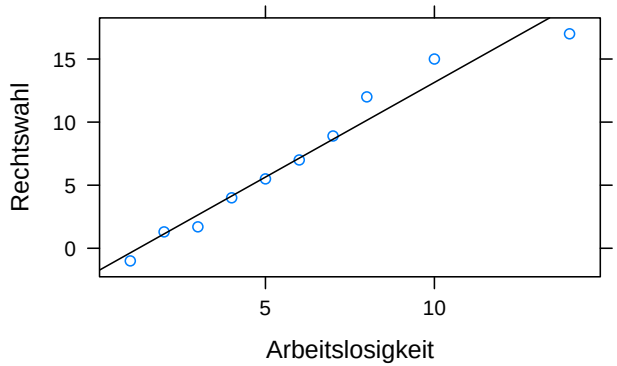
Regression I

Statistik I

Sommersemester 2009



Zum Nachlesen

- Agresti: 9.1-9.4
- Gehring/Weins: 8
- Schumann: 8.1-8.2

Was ist ein Zusammenhang?

- Gemeinsame Verteilung zweier Variablen ...
- ... weicht von Randverteilungen ab

Was ist ein Zusammenhang?

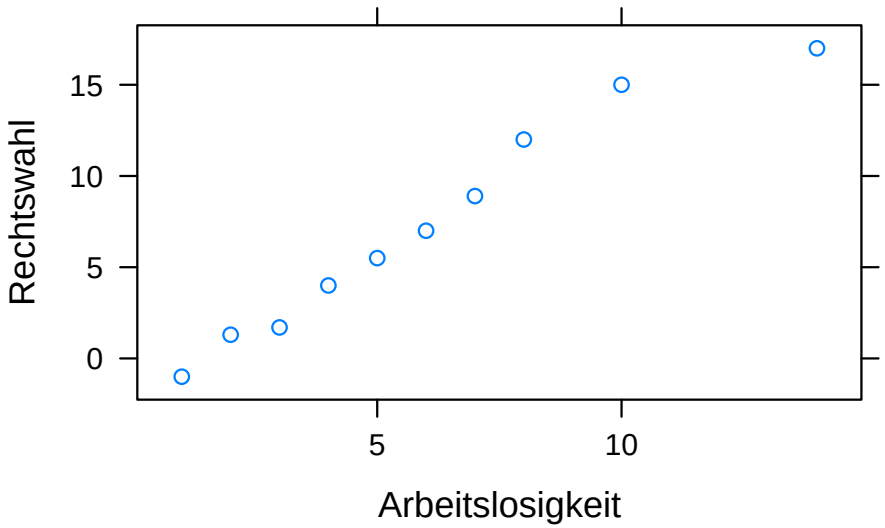
- Gemeinsame Verteilung zweier Variablen ...
- ... weicht von Randverteilungen ab
- „Muster“ – gemeinsame Verteilung $\neq$ zufällige Verteilung

Zusammenhangsmaß für intervallskalierte Daten?

- x überdurchschnittlich, y überdurchschnittlich (und umgekehrt) $\longrightarrow$ positiver Zusammenhang
- x überdurchschnittlich, y unterdurchschnittlich (und umgekehrt) $\longrightarrow$ negativer Zusammenhang
- Maß für *gemeinsame* Abweichung von zwei Mittelwerten: Abweichungsprodukt
- SAP: Welches Muster überwiegt?
- Kovarianz: SAP/n
- Teilen durch Produkt Standardabweichungen $\longrightarrow$ r, standardisiert Kovarianz auf Wertebereich [-1;1]

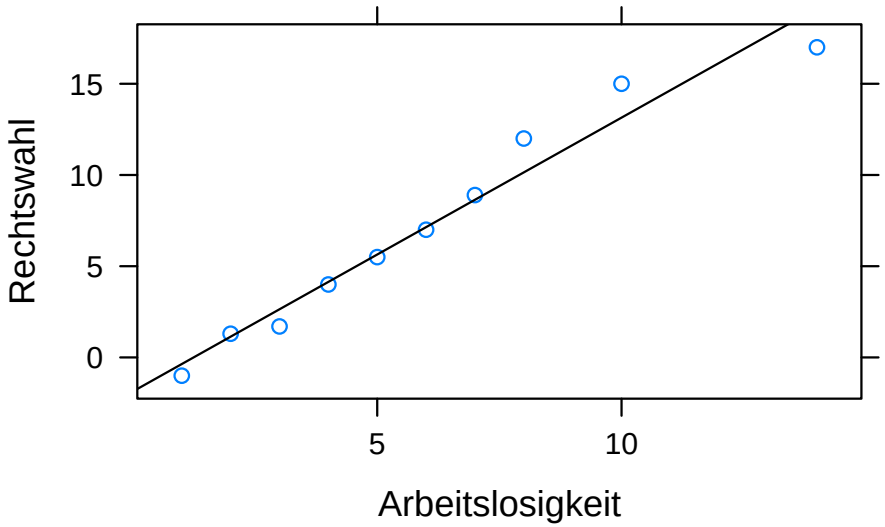
Was ist ein linearer Zusammenhang?

- *r* mißt *lineare* Zusammenhänge
- Zusammenhang zwischen zwei Variablen entspricht näherungsweise einer geraden Linie
- Formaler: Wert einer Variable könnte durch Addition/Multiplikation in Wert anderer Variable (plus Fehler) überführt werden



Was ist ein linearer Zusammenhang?

- *r* mißt *lineare* Zusammenhänge
- Zusammenhang zwischen zwei Variablen entspricht näherungsweise einer geraden Linie
- Formaler: Wert einer Variable könnte durch Addition/Multiplikation in Wert anderer Variable (plus Fehler) überführt werden



Was ist ein statistisches Modell?

- Mathematische Formalisierung von Theorie/Hypothesen
- Set von *Gleichung(en)*
  - Eine oder mehrere *abhängige* Variablen als *Funktion* von einer/mehrerer unabhängiger Variablen
  - Funktionen: Abbildungsvorschrift; input/output
- Modelle sind nicht wahr/maßstäblich
- Test von Hypothesen/extrem vereinfachten Beschreibungen

Was ist „lineare Einfachregression“?

- „linear“: nur lineare Beziehungen zwischen Variablen
- „einfach“: eine abhängige, eine unabhängige Variable
- „Regression“: „Rückführung“; konditionaler Mittelwert der abhängigen Variablen abhängig vom Niveau der unabhängigen Variablen

Was ist Regression?

- Regression ist der Oberbegriff für Verfahren, ...
- die die *konditionale* Verteilung einer Variablen y...
- in Abhängigkeit von einer oder mehreren anderen Variablen $x_1, x_2 \ldots x_k$ beschreiben

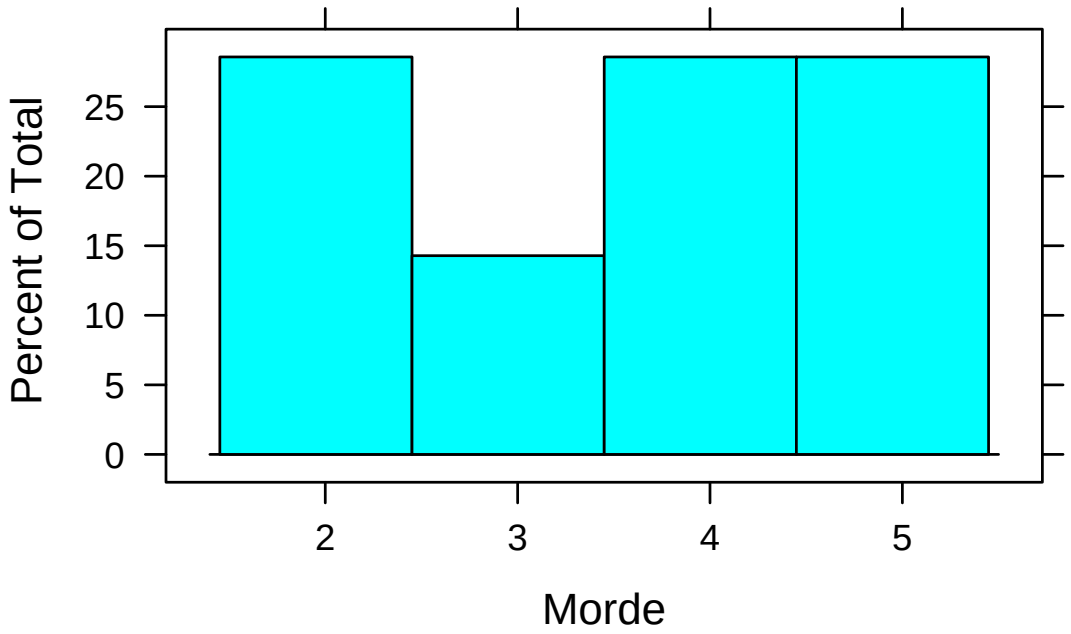
Was ist eine „konditionale Verteilung“?

- Verteilung von y (Mittelwert, Streuung etc.) ...
- innerhalb von Subgruppen, die durch $x_1, x_2 \ldots x_k$ definiert sind

Armut und Gewaltverbrechen in 51 Bundesstaaten

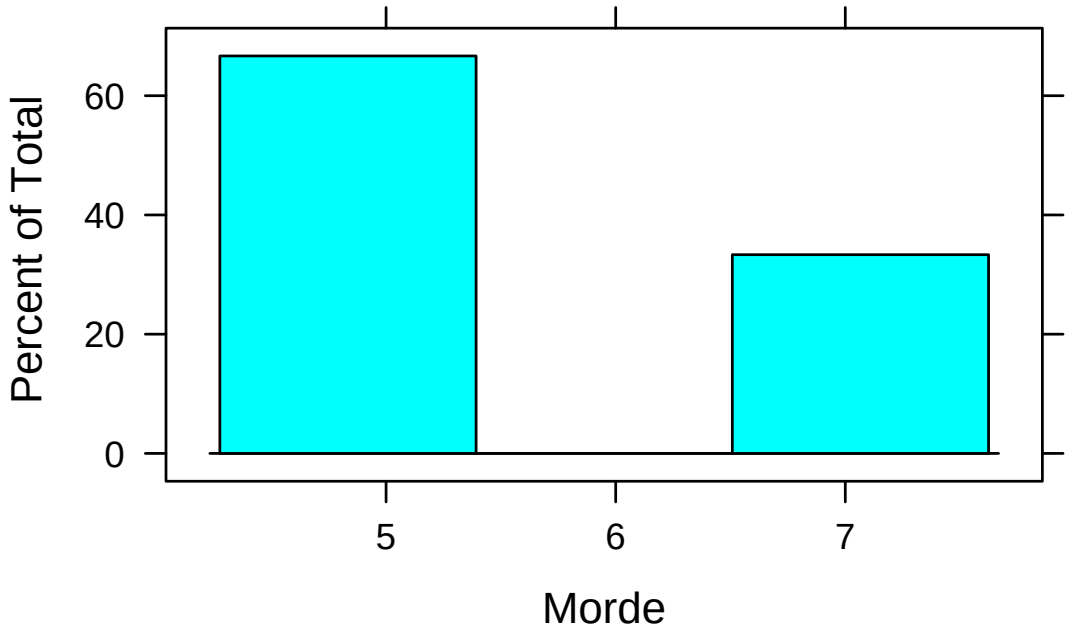
- Statistiken auf Ebene der 50 Bundesstaaten plus Washington D.C.
- Zusammenhang zwischen Anteil der Armen (in Prozent) an Gesamtbevölkerung und
- Morde pro 100 000 Einwohner?
- Konditionale Verteilung von y (Morde)
- Für verschiedene Werte von x_1 (Armutsquote)

Beispiel: Konditionale Verteilung Mordquote für Armutsquote $\approx 10.5\%$



- AM $= 3.6$

Beispiel: Konditionale Verteilung Mordquote für Armutsquote $\approx 12.5\%$

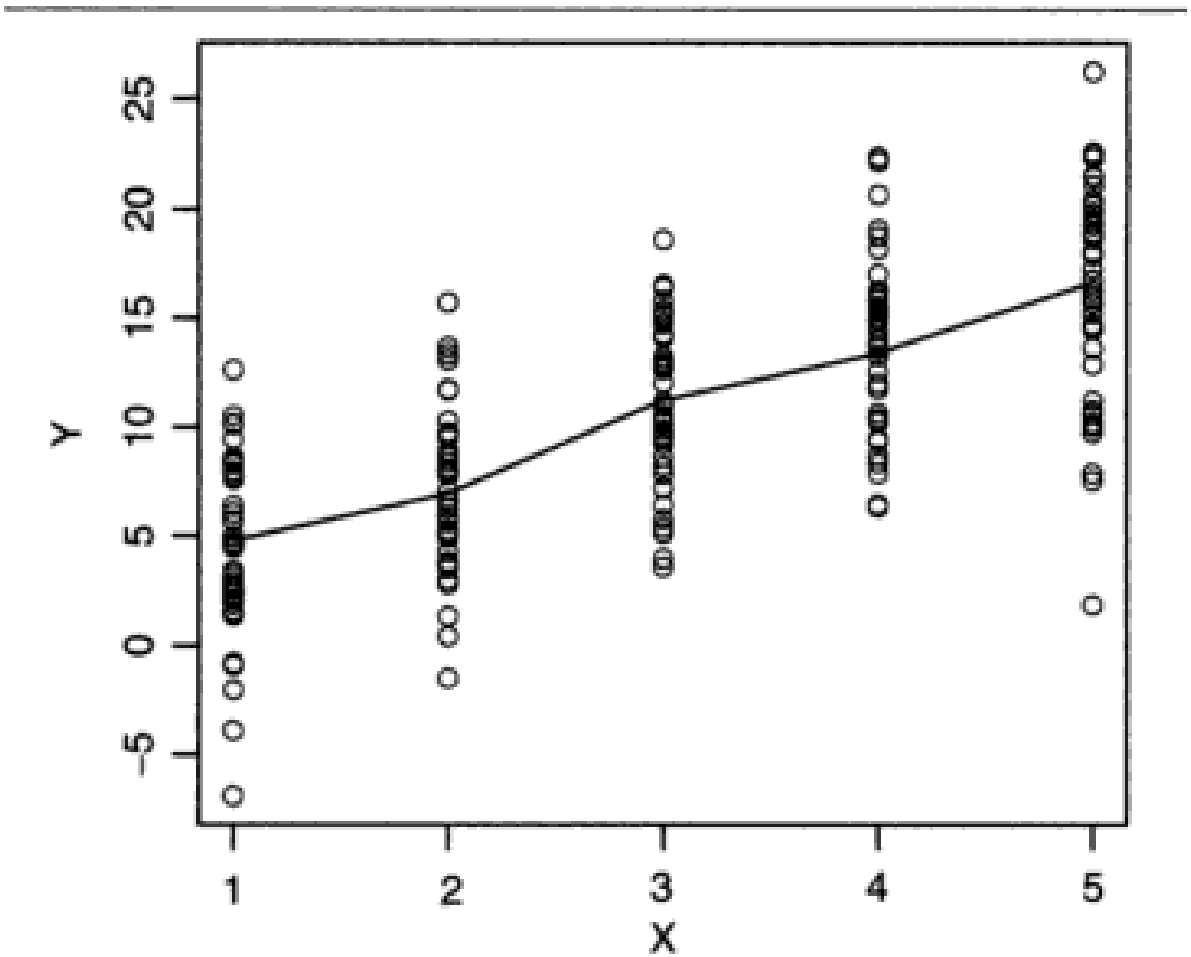


- AM $= 5.7$

Was ist lineare Regression? II

- Konditionale Mittelwerte können durch gerade Linie verbunden werden
- Konditionale Variation um konditionalen Mittelwert

Wie war das mit dem konditionalen Mittelwert?



Was ist lineare Regression? II

- Konditionale Mittelwerte können durch gerade Linie verbunden werden
- Konditionale Variation um konditionalen Mittelwert
- Übergang zum Modell: Annahmen über die Eigenschaften der Linie kommen von außen
- „Abhängige“ / „unabhängige“ Variable kommen ebenfalls von außen
- Was ist damit gewonnen?

Was ist lineare Regression? II

- Konditionale Mittelwerte können durch gerade Linie verbunden werden
- Konditionale Variation um konditionalen Mittelwert
- Übergang zum Modell: Annahmen über die Eigenschaften der Linie kommen von außen
- „Abhängige“ / „unabhängige“ Variable kommen ebenfalls von außen
- Was ist damit gewonnen?
  - Wenn Annahme über „Linie“ (Modell) gültig
  - Allgemeiner Mechanismus $\longrightarrow$ einfach

Was ist lineare Regression? II

- Konditionale Mittelwerte können durch gerade Linie verbunden werden
- Konditionale Variation um konditionalen Mittelwert
- Übergang zum Modell: Annahmen über die Eigenschaften der Linie kommen von außen
- „Abhängige“ / „unabhängige“ Variable kommen ebenfalls von außen
- Was ist damit gewonnen?
  - Wenn Annahme über „Linie“ (Modell) gültig
  - Allgemeiner Mechanismus $\longrightarrow$ einfach
  - Lineare Regression setzt Kausalität nicht voraus
  - Kann Kausalität nicht bestätigen

Was macht man mit Modellen?

- Wenn Modell korrekt (angemessen), y-Werte für nicht beobachtete x
- Allgemeiner, einfacher, eleganter Mechanismus

1. Kompakte Beschreibung
2. Besseres Verständnis/Hypothesentest
3. Prognose für zusätzliche/hypothetische/zukünftige x

Kausalität und Korrelation

- Was ist Kausalität?

Kausalität und Korrelation

- Was ist Kausalität?
  - x verursacht y. Hypothetisch: ohne x kein y; anderes x – anderes y
  - Voraussetzungen: theoretischer Zusammenhang, zeitliche Reihenfolge, keine anderen Variablen ($\rightarrow$ Theorie und Forschungsdesign) + empirischer Zusammenhang (Korrelation)
  - Näherungsweise Prüfung von Kausalität in experimentellen Designs
  - In Ex-post-facto Designs nur hypothetische Prüfung; statistische Kontrolle von Drittvariablen

Kausalität und Korrelation

- Was ist Kausalität?
  - x verursacht y. Hypothetisch: ohne x kein y; anderes x – anderes y
  - Voraussetzungen: theoretischer Zusammenhang, zeitliche Reihenfolge, keine anderen Variablen ($\rightarrow$ Theorie und Forschungsdesign) + empirischer Zusammenhang (Korrelation)
  - Näherungsweise Prüfung von Kausalität in experimentellen Designs
  - In Ex-post-facto Designs nur hypothetische Prüfung; statistische Kontrolle von Drittvariablen
- Korrelation notwendig, aber nicht hinreichend für Kausalität
- Korrelation kann durch gemeinsame Hintergrundvariablen zustande kommen
- Korrelation kann durch andere Variablen unterdrückt werden, obwohl kausaler Zusammenhang besteht

Welche Form hat das Modell der linearen Einfachregression?

$$y = a + b_1 x_1 + e \quad \text{bzw.}$$
$$y = b_0 + b_1 x_1 + e \quad \text{bzw.}$$

$$y = \alpha + \beta_1 x_1 + \epsilon \quad \text{bzw.}$$
$$y = \beta_0 + \beta_1 x_1 + \epsilon \quad \text{bzw.}$$

Welche Form hat das Modell der linearen Einfachregression?

$$y = a + b_1 x_1 + e \quad \text{bzw.}$$
$$y = b_0 + b_1 x_1 + e \quad \text{bzw.}$$

$$y = \alpha + \beta_1 x_1 + \epsilon \quad \text{bzw.}$$
$$y = \beta_0 + \beta_1 x_1 + \epsilon \quad \text{bzw.}$$

- Deskription vs. Inferenz – lateinische vs. griechische Buchstaben

Welche Form hat das Modell der linearen Einfachregression?

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 \cdots + e \quad \text{bzw.}$$
$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \cdots + e \quad \text{bzw.}$$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots + \epsilon \quad \text{bzw.}$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots + \epsilon \quad \text{bzw.}$$

- Deskription vs. Inferenz – lateinische vs. griechische Buchstaben
- Erweiterung zum multivariaten Modell möglich

Welche Bestandteile hat dieses Modell?

1. y: abhängige Variable, soll „erklärt“ werden
2. a oder b_0 oder α oder β_0: „Achsenabschnitt“: Wert von y wenn $x_{(1)} = 0$
3. b_1 oder β_1: „Steigung“; erwarteter Effekt auf y wenn x_1 um eine Einheit zunimmt
4. e oder ϵ: Abweichung zwischen erwartetem und beobachtetem y; zufällige Einflüsse

- 2 und 3 bilden die systematische/deterministische Komponente
- 4 ist die stochastische/zufällige Komponente

Warum ϵ?

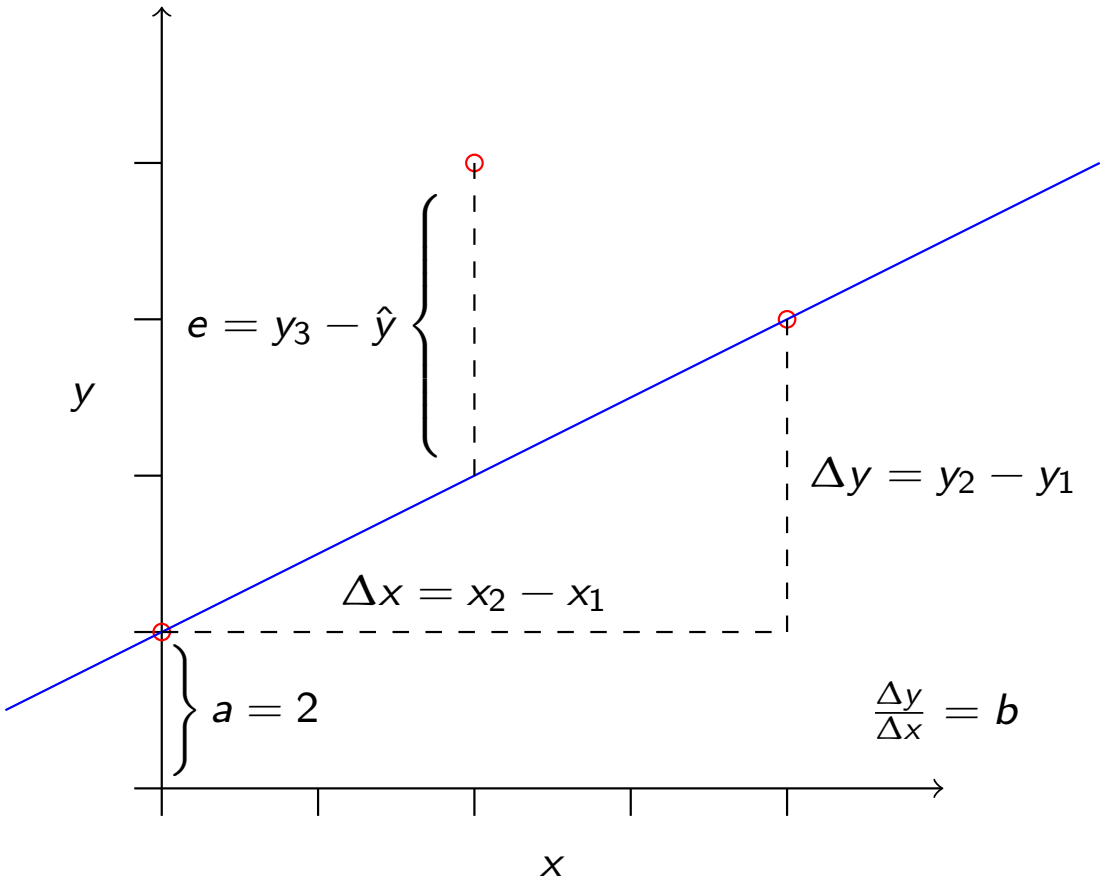
„To err is human, to forgive divine, but to include errors into your design is statistical“ (Leslie Kish)

Warum ϵ?

„To err is human, to forgive divine, but to include errors into your design is statistical“ (Leslie Kish)

- Kein Modell paßt perfekt
- Zahl der unabhängigen Variablen beschränkt
- ϵ umfaßt
  - Nichtgemessene Einflüsse, die vernachlässigt werden können
  - Genuin zufällige Einflüsse
- Modell $\neq$ Wirklichkeit $\longrightarrow \epsilon$

Wie sieht das praktisch aus?



$$\hat{y} = a + bx + e = 2 + 0.5 \times x + e$$

Wie interpretiert man das Modell der linearen Einfachregression?

- b: Stärke/Richtung des Effektes von x auf y
- Bei Interpretation Wertebereiche von x und y beachten
- $\hat{y}$ ist der für einen x-Wert vorhergesagte Wert von y
  - Konditionaler Mittelwert von y
  - *Unter Berücksichtigung der Modellannahmen*
- Vergleich Modell/Empirie: $\hat{y}_i$ vs. y_i
  - Warum weicht y_i ab? Untypisch?
- Hypothetische Werte: Wieviel Morde würden wir in einem Bundesstaat ohne Armut erwarten (wenn Modell gilt)
- Prognose: Wie wirkt sich ein Anstieg der Armut aus? Wie stark läßt sich Verbrechen durch Armutsbekämpfung reduzieren?

Daten: Armut und Morde in 51 Bundesstaaten

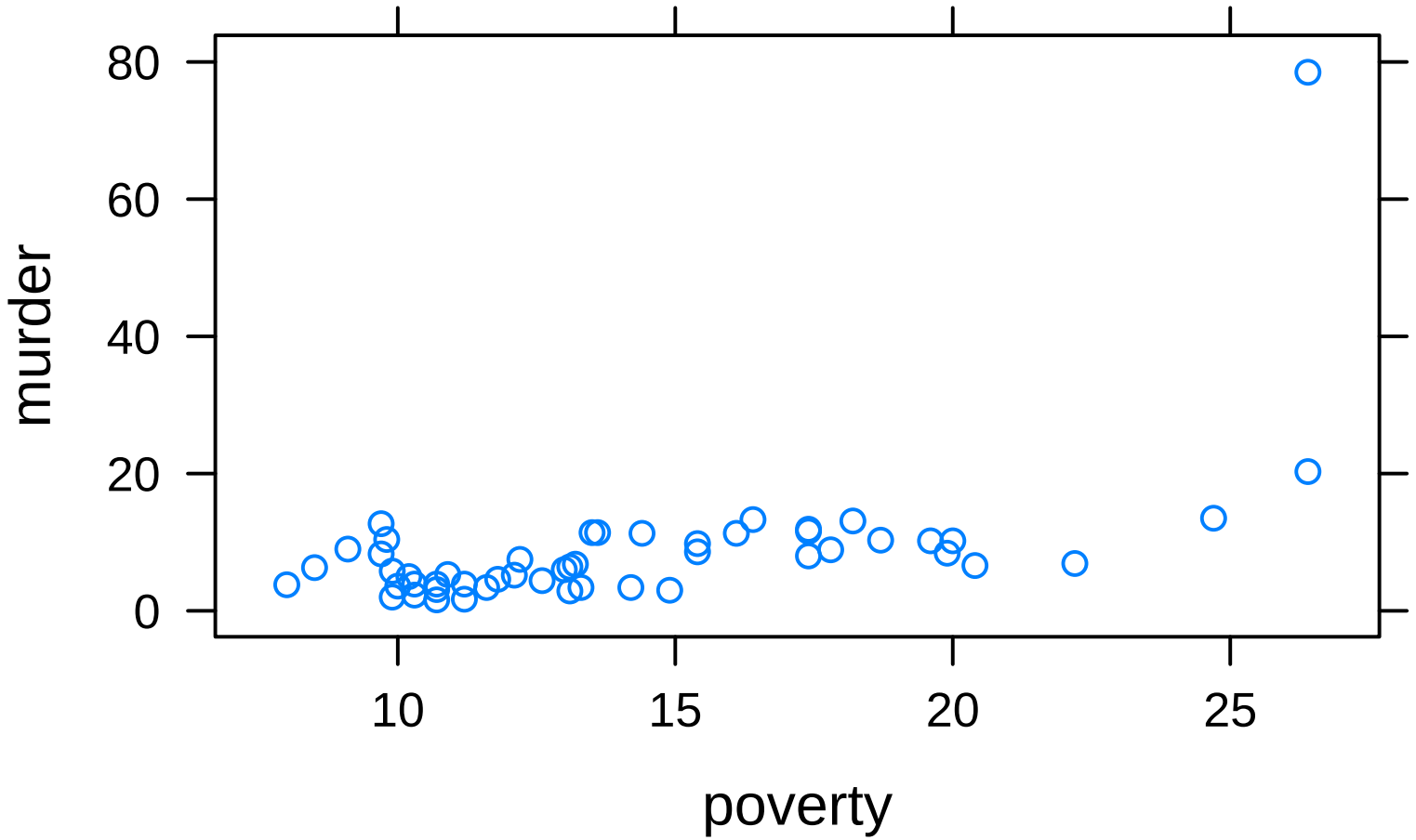
- Strukturdaten für 51 US-Staaten (Morde, Armut, % Weiße, % urbane Bevölkerung etc.)
- Unterschiede in x und y
  - Hawaii: 8 Prozent Arme, Louisiana 26.4 Prozent
  - Maine: 1.6 Morde pro 100 000 Einwohner, Washington D.C. 78.5
- Sehr hohe Aggregationsebene
- Sehr große Unterschiede in Größe der Einheiten
  - Wyoming: 522,830 Einwohner
  - California: 36,553,215 Einwohner

Was erwarten wir? Warum?

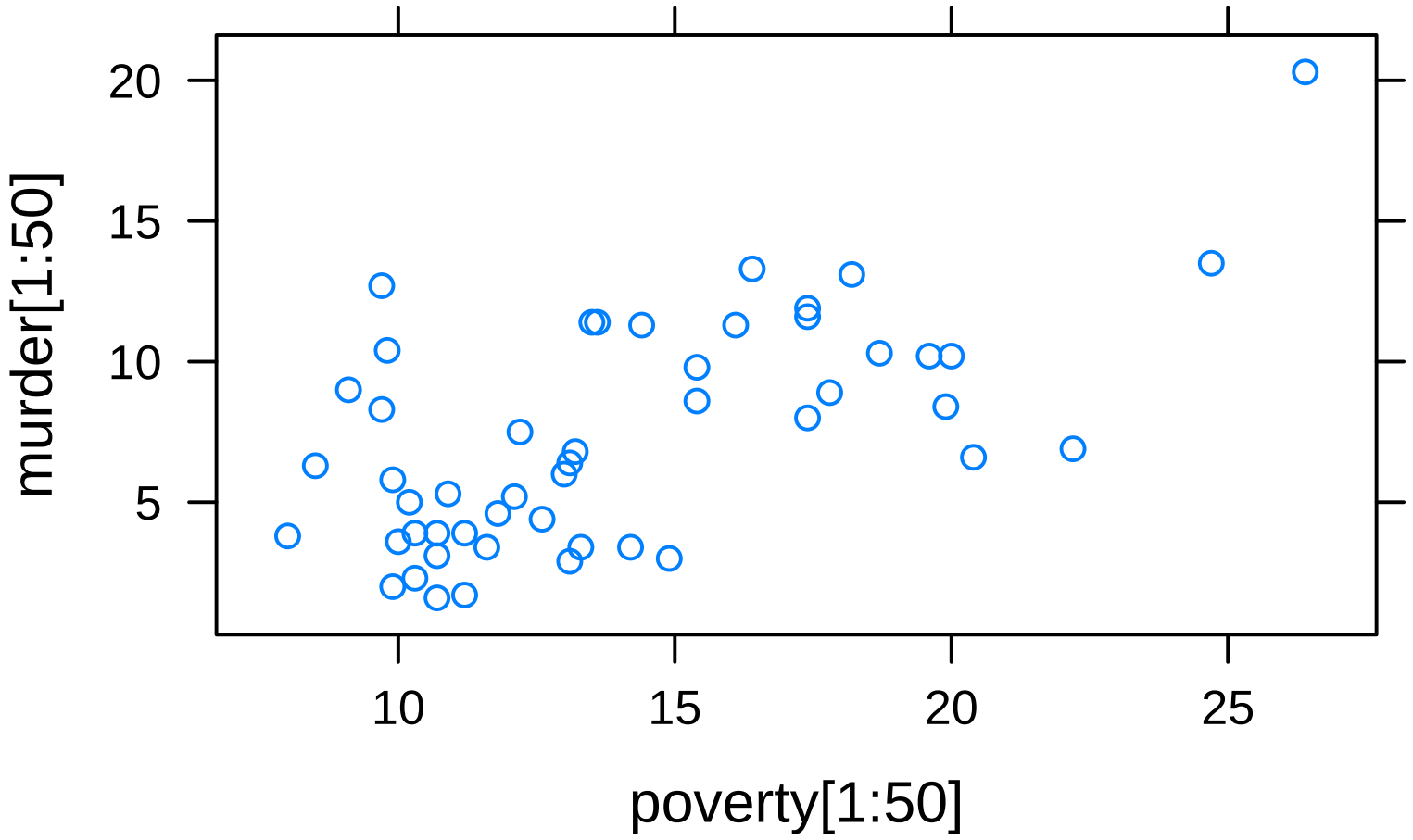
Was erwarten wir? Warum?

- Sind Arme gewalttätiger?
- Macht Armut gewalttätig?
- Gibt es gemeinsame Ursachen?
- Sind weniger gewalttätige Staaten wirtschaftlich erfolgreicher?
- ...
- Kein Rückschluß auf individuelles Verhalten (ökologischer Fehlschluß)

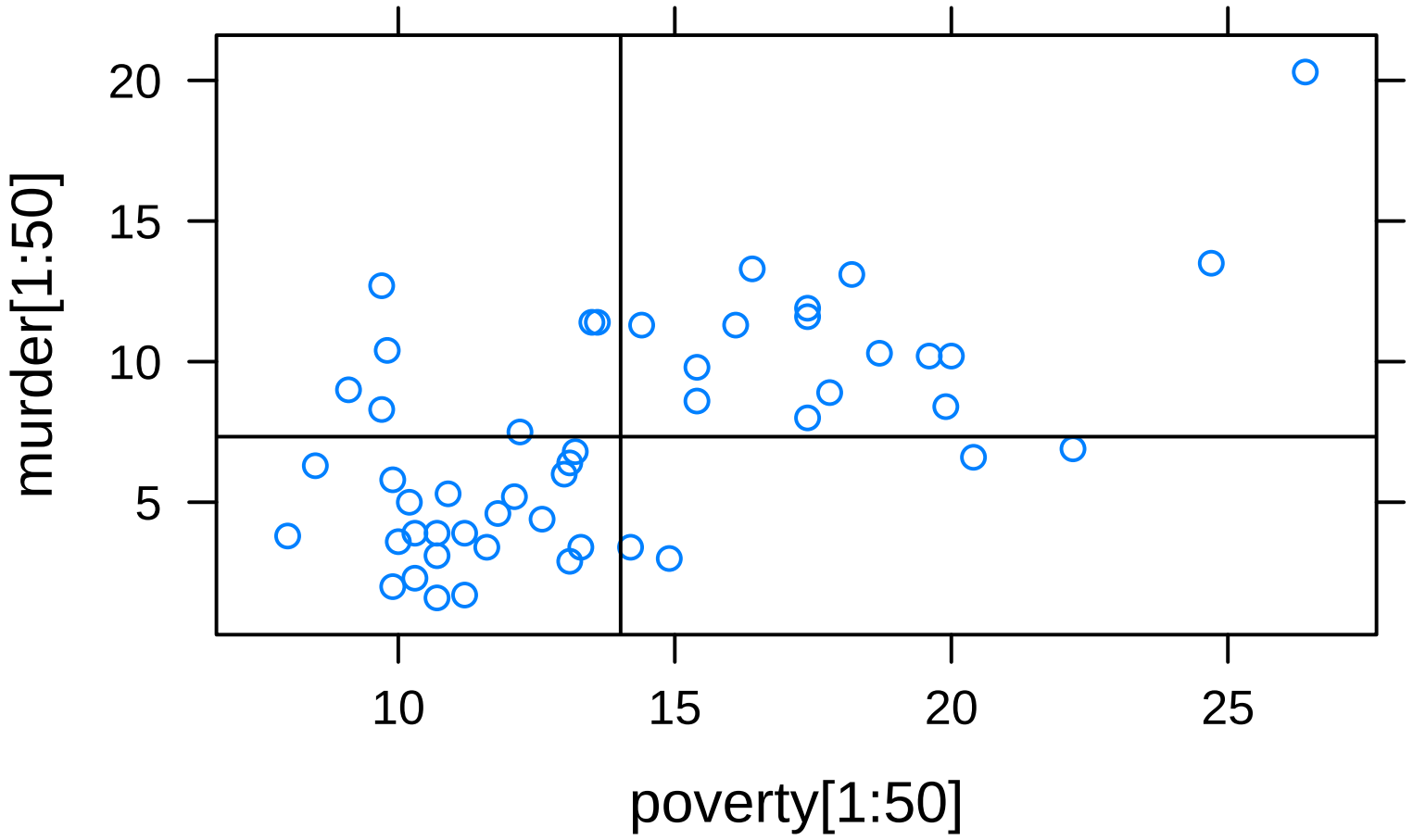
Wie sieht die bivariate Verteilung aus?



Wie sieht die bivariate Verteilung aus?



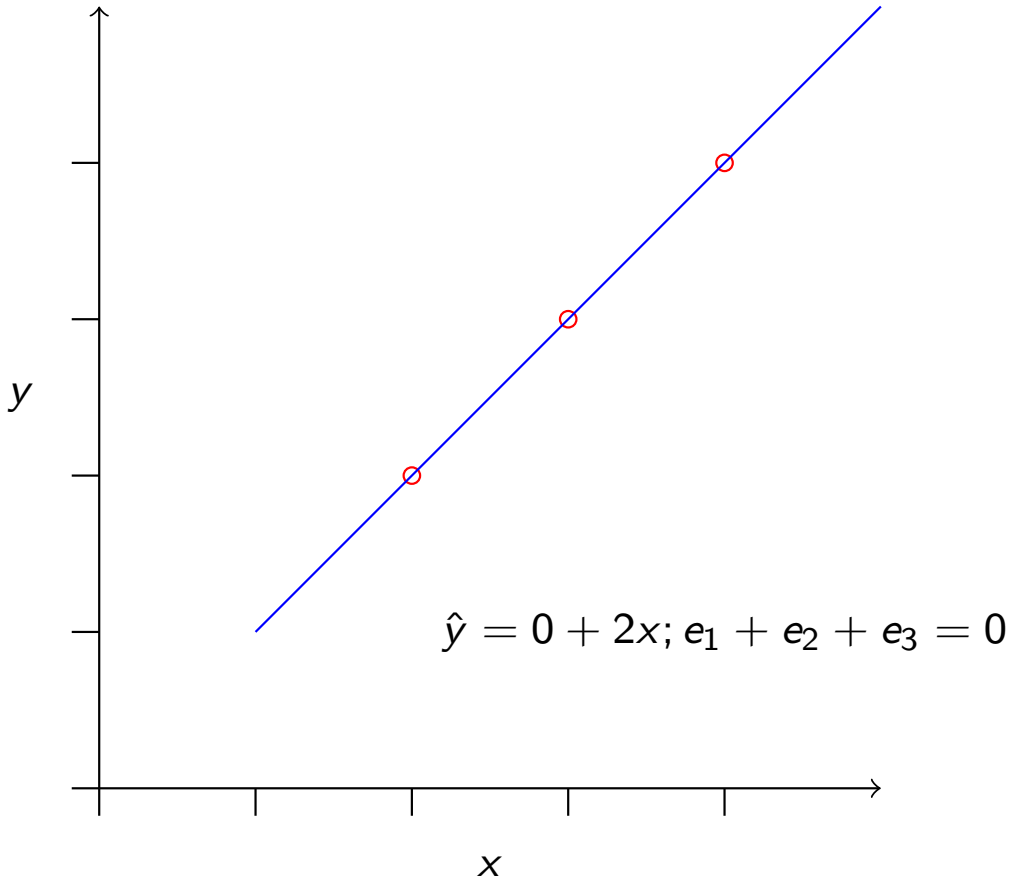
Wie sieht die bivariate Verteilung aus?



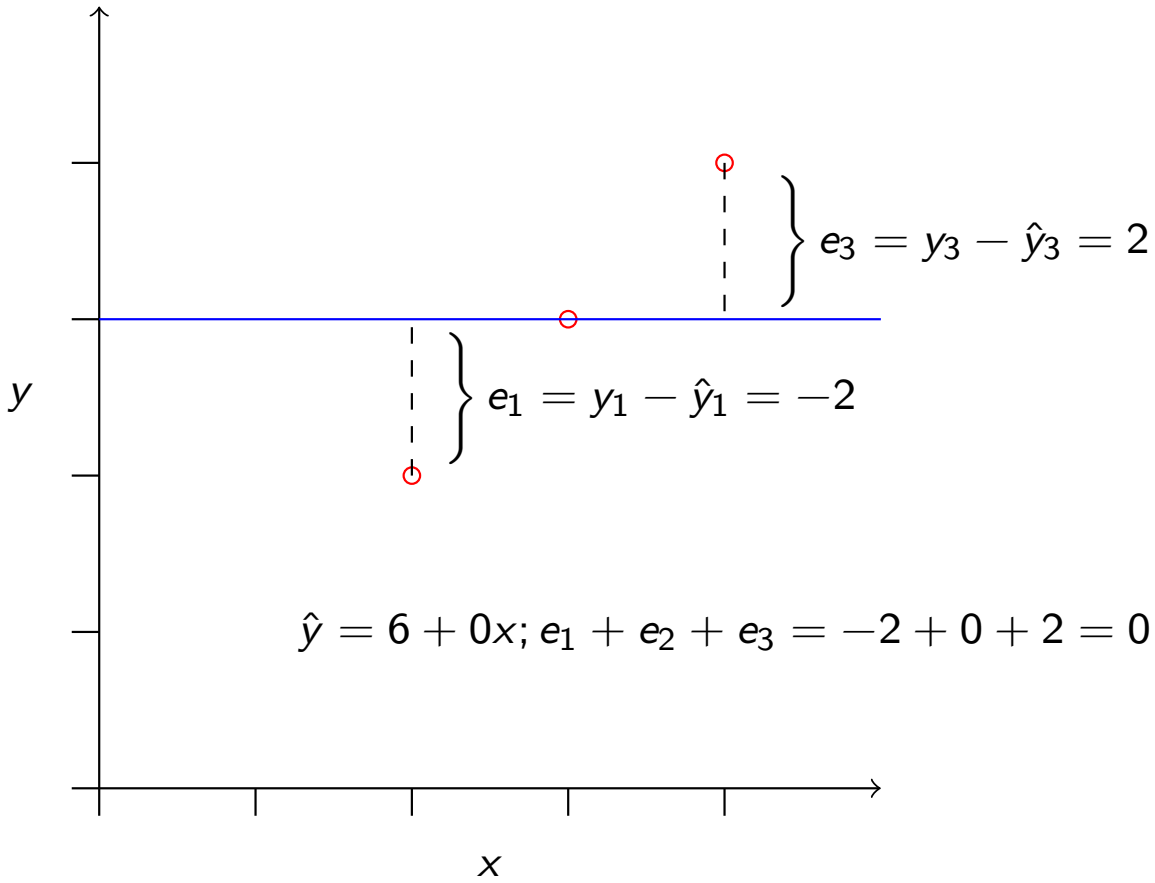
Was ist ein Kriterium für „gute“ Parameter?

- Gesucht werden Werte für a und b, die
  - *die Summe der quadrierten Abweichungen*
  - *in y-Richtung*
  - *minimieren*
- „Gute“ Trendgerade
- Warum?
  - Vorhersagefehler für y optimieren $\longrightarrow$ guter fit (innerhalb Modellannahmen)
  - Einfache Abweichungen nicht eindeutig

Warum werden nicht die einfachen Abweichungen verwendet?



Warum werden nicht die einfachen Abweichungen verwendet?



$\hat{y} = 6 + 0x; e_1 + e_2 + e_3 = -2 + 0 + 2 = 0$

Was ist ein Kriterium für „gute“ Parameter?

- Gesucht werden Werte für a und b, die
  - *die Summe der quadrierten Abweichungen*
  - *in y-Richtung*
  - *minimieren*
- „Gute“ Trendgerade
- Warum?
  - Vorhersagefehler für y optimieren $\rightarrow$ guter fit (innerhalb Modellannahmen)
  - Einfache Abweichungen nicht eindeutig
  - Quadrierte Abweichungen eindeutig + Gewichtung große Abweichungen
  - Optimales *Schätzverfahren*, wenn Bedingungen erfüllt

Wie werden die Parameter bestimmt?

- Gesucht: Werte für a und b, die SAQ_y (bezogen auf $\hat{y}$) minimal machen
- „Verlustfunktion“, abhängig von Daten und Parameterschätzungen

Verlustfunktion

$$SAQ_y = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{1i}))^2$$
$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i})^2$$

Wie finde ich das Minimum dieser Funktion?

- Analytische Lösung
- Notwendige Bedingung für einen Extremwert: 1. Ableitung gleich 0 (Tangente ist an dieser Stelle flach)
- Funktion hat zwei Variablen $\rightarrow$ zwei partielle Ableitungen (nach b_0 und b_1) betrachten

Wie finde ich die partiellen Ableitungen?

$$SAQ = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i})^2 \tag{1}$$

$$\frac{\partial SAQ}{b_0} = \sum_{i=1}^{n} -1 \times 2 \times (y_i - b_0 - b_1 x_{1i}) = 0 \tag{2}$$

$$\frac{\partial SAQ}{b_1} = \sum_{i=1}^{n} -x_{1i} \times 2 \times (y_i - b_0 - b_1 x_{1i}) = 0 \tag{3}$$

Wie finde ich die Werte für b_0 und b_1 ?

- Durch Umformen/Auflösen ergibt sich:

$$b_0 = \bar{y} - b_1 \bar{x_1} \tag{4}$$

$$b_1 = \frac{\sum (x_{1i} - \bar{x_1})(y_i - \bar{y})}{\sum (x_{1i} - \bar{x_1})^2} = \tag{5}$$

Wie finde ich die Werte für b_0 und b_1 ?

- Durch Umformen/Auflösen ergibt sich:

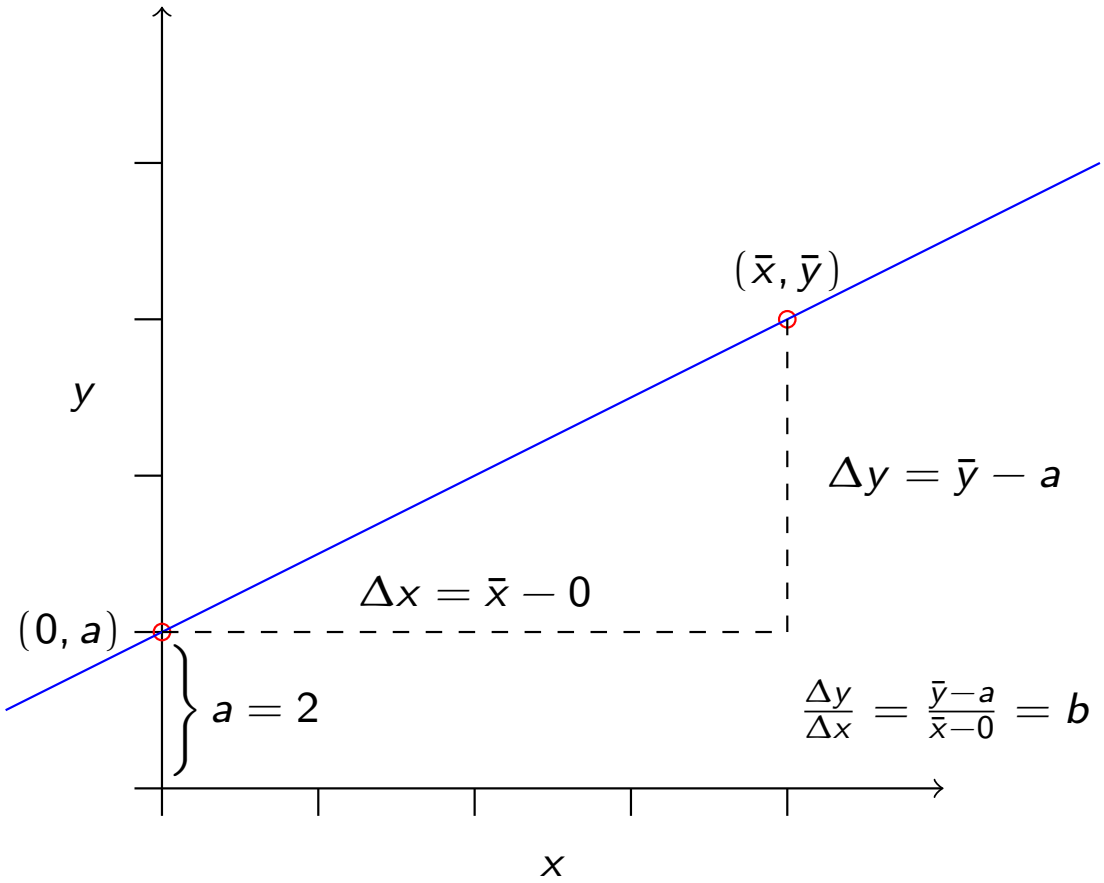
$$b_0 = \bar{y} - b_1 \bar{x_1} \tag{4}$$

$$b_1 = \frac{\sum (x_{1i} - \bar{x_1})(y_i - \bar{y})}{\sum (x_{1i} - \bar{x_1})^2} = \frac{SAP_{xy}}{SAQ_x} \tag{5}$$

Die einfache Variante

- $b_1 = \frac{SAP}{SAQ_x}$
- D. h. mit den aus der Berechnung von r bekannten SAP und SAQ b bestimmen
- Diese Gerade läuft durch Punkt $(\bar{x}, \bar{y}) \rightarrow$ Steigungsdreieck konstruieren, a berechnen

Wie sieht das praktisch aus?



Die einfache Variante

- $b_1 = \frac{SAP}{SAQ_x}$
- D. h. mit den aus der Berechnung von r bekannten SAP und SAQ b bestimmen
- Diese Gerade läuft durch Punkt $(\bar{x}, \bar{y}) \rightarrow$ Steigungsdreieck konstruieren, a berechnen

$$b_1 = \frac{\Delta y}{\Delta x} = \frac{\bar{y} - a}{\bar{x} - 0} = \frac{\bar{y} - a}{\bar{x}}$$

$$a = b_0 = \bar{y} - b_1 \times \bar{x}$$

Wie sieht das im Beispiel aus?

- $\bar{x} = 14.26$
- $\bar{y} = 8.727$
- $SAQ_x = 1050.8$
- $SAP = 1390.1$

Koeffizienten

Wie sieht das im Beispiel aus?

- $\bar{x} = 14.26$
- $\bar{y} = 8.727$
- $SAQ_x = 1050.8$
- $SAP = 1390.1$

Koeffizienten

- $b_1 = \frac{SAP}{SAQ_x} = \frac{1390.1}{1050.8} = 1.32$

Wie sieht das im Beispiel aus?

- $\bar{x} = 14.26$
- $\bar{y} = 8.727$
- $SAQ_x = 1050.8$
- $SAP = 1390.1$

Koeffizienten

- $b_1 = \frac{SAP}{SAQ_x} = \frac{1390.1}{1050.8} = 1.32$
- $a = b_0 = \bar{y} - b_1 \times \bar{x} = -10.14$

Was bedeutet das?

- Statistisch ist bei einem Anstieg der Armutsquote um einen Punkt etwas mehr als ein zusätzlicher Mord / 100 000 Einwohner zu erwarten
- Louisiana sollte $-10.136 + 1.323 \times 26.4 = 24.8$ Morde haben, hat aber nur 20.3
- Keine Armut: -10.1 Morde?
- Bei etwa 8 Prozent Armutsquote wären gar keine Morde mehr zu erwarten? $\rightarrow$ Rolle von ϵ

Zusammenfassung

- Lineare Einfachregression als einfachstes statistisches Modell
- Beschreibt Zusammenhang zwischen zwei Variablen mit Veränderungsrate und Konstante
- Muster für eine ganze Reihe von komplexeren Modellen
- Modelle
  - Ermöglichen kompakte *Beschreibung* und
  - Erleichtern *Verständnis* der Zusammenhänge
  - Ermöglichen ggf. *Schluß auf eine Grundgesamtheit*
  - Gestatten *Prognosen* für *zukünftige* und *hypothetische* Fälle
  - Wenn Modell korrekt spezifiziert und Annahmen realistisch