JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Statistik II, SoSe 23

3.7.23, Lineare Regressionsanalyse, Teil 2

Simone Abendschön

Inhalte heute

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

- Hinweise Klausur
- Lineare Regression, Teil 2

Hinweise E-Klausur

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

- Wann: 18.7.23 von 9 bis 10:30 Uhr (bitte um 8.45 Uhr vor Ort sein)
- Wo: Medizinisches Lehrzentrum (MLZ), Klinikstr. 29, Hörsaal 1 (R 036)
- Bitte mitbringen: Studiausweis, Personalausweis/Reisepass, ausgedruckte Formelsammlung (ohne Notizen), Taschenrechner, Stift

→Konzeptpapier bekommen Sie von uns

- Diese und weitere Infos diese Woche über stud.ip

Lernziele Regression heute

- Sie verstehen, wie ein (bivariates) Regressionsmodell geschätzt wird
- Sie verstehen die lineare Regressionsgleichung und können Sie anwenden (heute)

Regressionsanalyse: Einführung

Was ist:

- die Richtung
- die Stärke
- die statistische Signifikanz

… des Einflusses von X auf Y?

Regressionsanalyse: Einführung

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

- kann also Fragen beantworten wie:
  - Welche Einflussgrößen tragen zur Erklärung eines Merkmals bei (bspw. Höhe des Einkommens/ Wahlverhalten einer Person/rechtsextreme Einstellungen/Lebenszufriedenheit … einer Person)?
  - Wie stark sind diese jeweiligen Einflüsse und sind sie statistisch signifikant?
  - Wie gut können wir mit diesen Einflussfaktoren gemeinsam die aV Höhe des Einkommens (usw.) bestimmen (und damit auch vorhersagen)?

Regressionsanalyse: Einführung

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Wichtig: **Theoretische Modellspezifizierung (**abgeleitet aus der konzeptionellen Forschungsarbeit/theoretische Plausibilität vorab), Beispiel:

- aV: Einkommen
- uV: Berufserfahrung, Geschlecht, Alter, Branche, Umfang Erwerbstätigkeit,…

➢ über die **Richtung** (und manchmal auch **Stärke**) des Einflusses der uVs treffen wir in **Hypothesen** bestimmte Vermutungen, die wir **empirisch überprüfen**

- z.B. H1: Eine langjährige Berufserfahrung hat einen positiven Einfluss auf die Höhe des Einkommens

Bivariates lineares Regressionsmodell

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

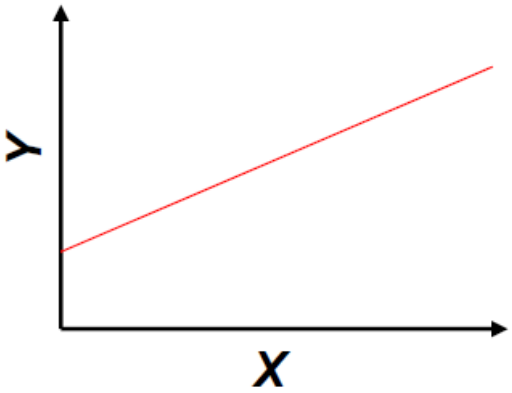
Grundprinzip

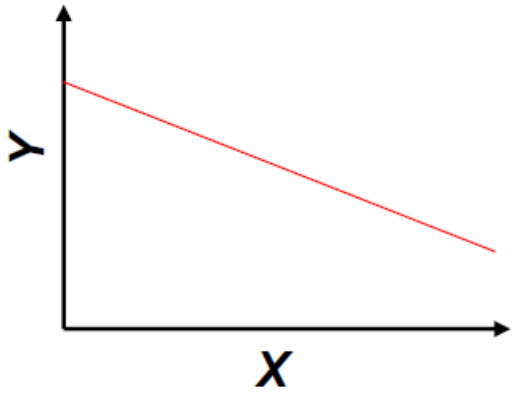
- Annahme einer **linearen Beziehung** zwischen X und Y: d.h. Stärke und Richtung des Zusammenhangs ist in jedem beliebigen Werteintervall auf der Variablen *X* gleich
- Voraussetzung: aV (pseudo-)metrisch, uV (pseudo-)metrisch bzw. dichotom

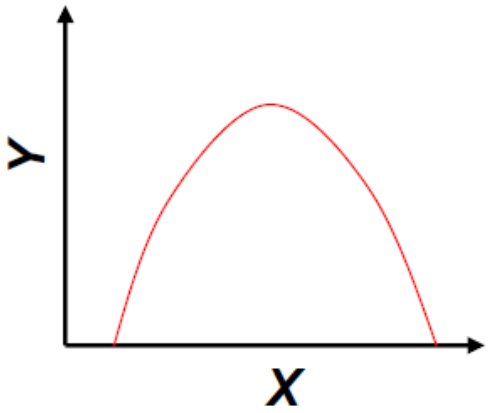
Kurze Wdh.

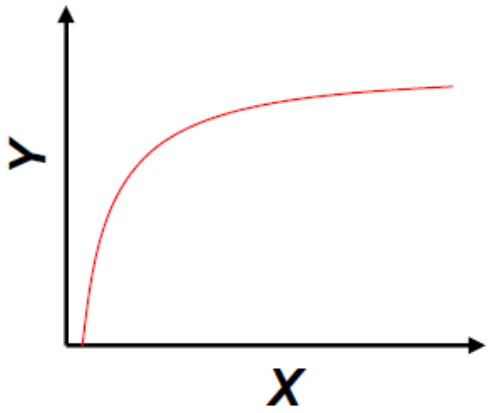
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

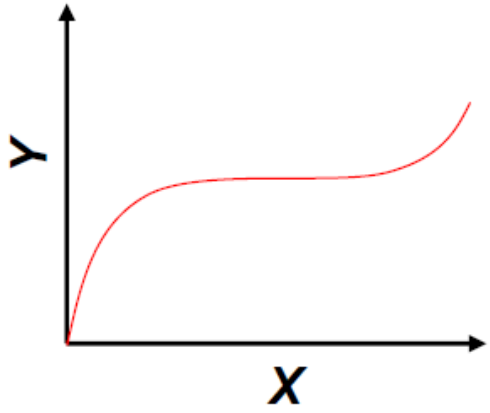
Lineare bzw. nicht-lineare Zusammenhänge











Bivariates lineares Regressionsmodell

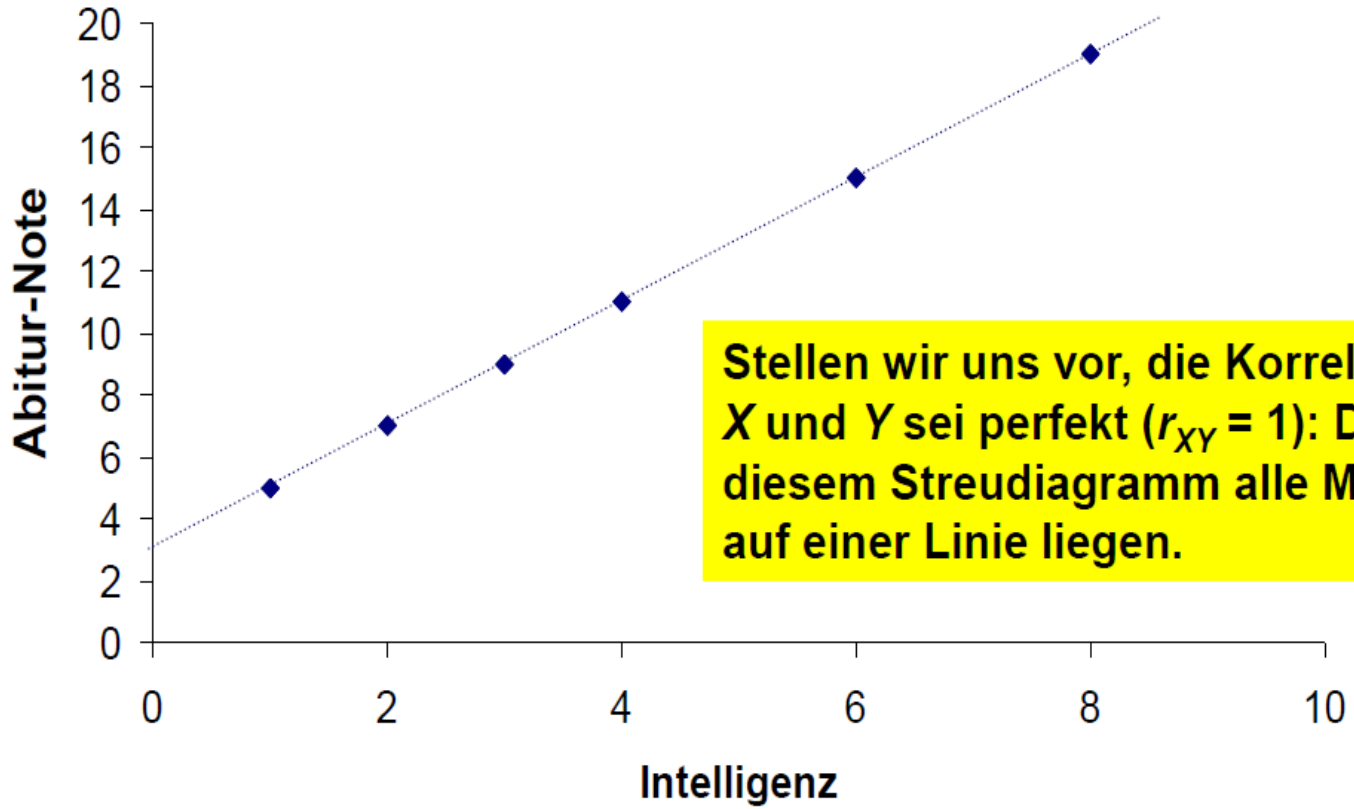
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Beispiel: Vorhersage der Abitur-Note einer Person (y) auf der Basis ihrer Intelligenz (x)

- Möglich, wenn Intelligenz und Abitur-Note miteinander korrelieren
- Allgemein gilt: Je höher die Korrelation zwischen X und Y, desto zuverlässiger gelingt die Vorhersage von Y durch X.
- Ohne Korrelation zwischen X und Y: Vorhersage des y-Wertes auf Basis des x-Wertes genau so gut/schlecht wie ohne Kenntnis von x → geringster Vorhersagefehler, wenn man für diese Person den Mittelwert von Y schätzt

Bivariates lineares Regressionsmodell

Beispiel: Einfluss der Intelligenz auf die Abinote: **Abinote= f(Intelligenz)**



Stellen wir uns vor, die Korrelation zwischen *X* und *Y* sei perfekt ($r_{XY} = 1$): Dann würden in diesem Streudiagramm alle Messwertpaare auf einer Linie liegen.

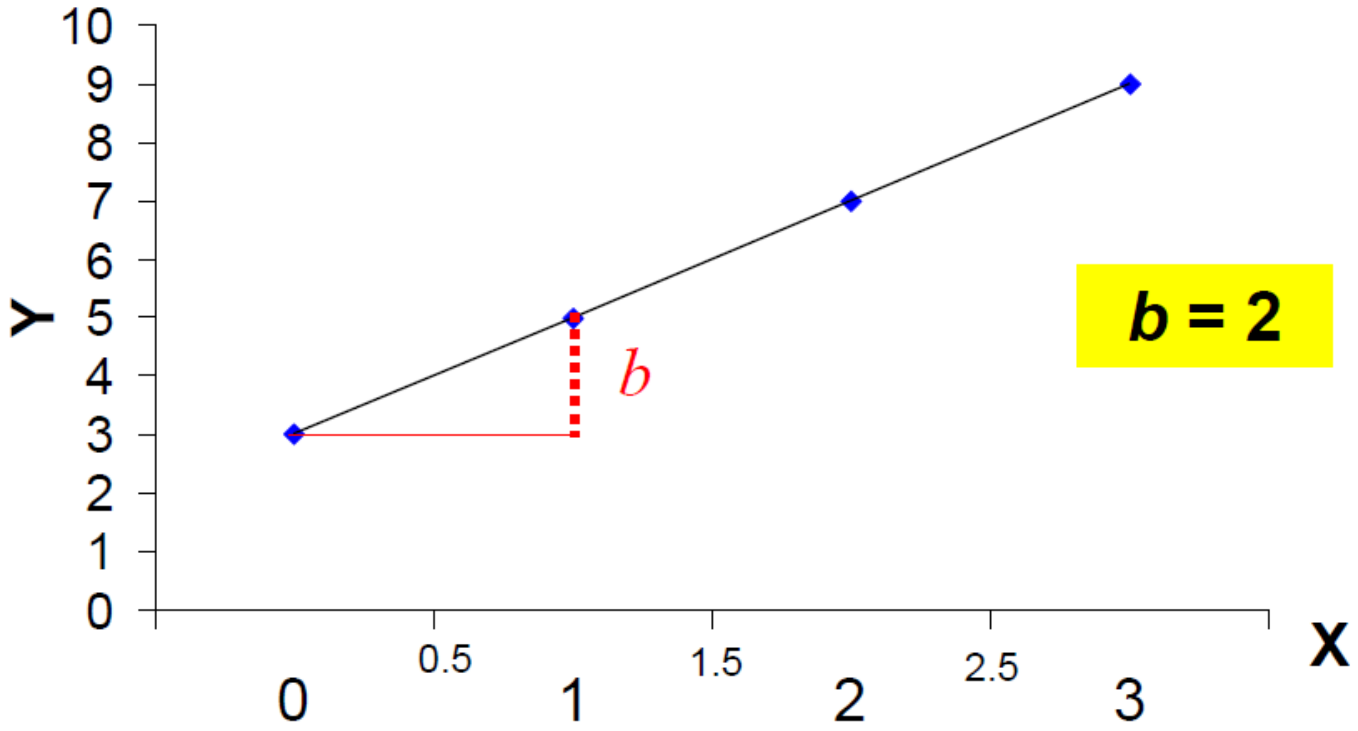
Bivariates lineares Regressionsmodell

Geraden sind lineare Funktionen der allgemeinen Form

$$y = a + b \cdot x$$

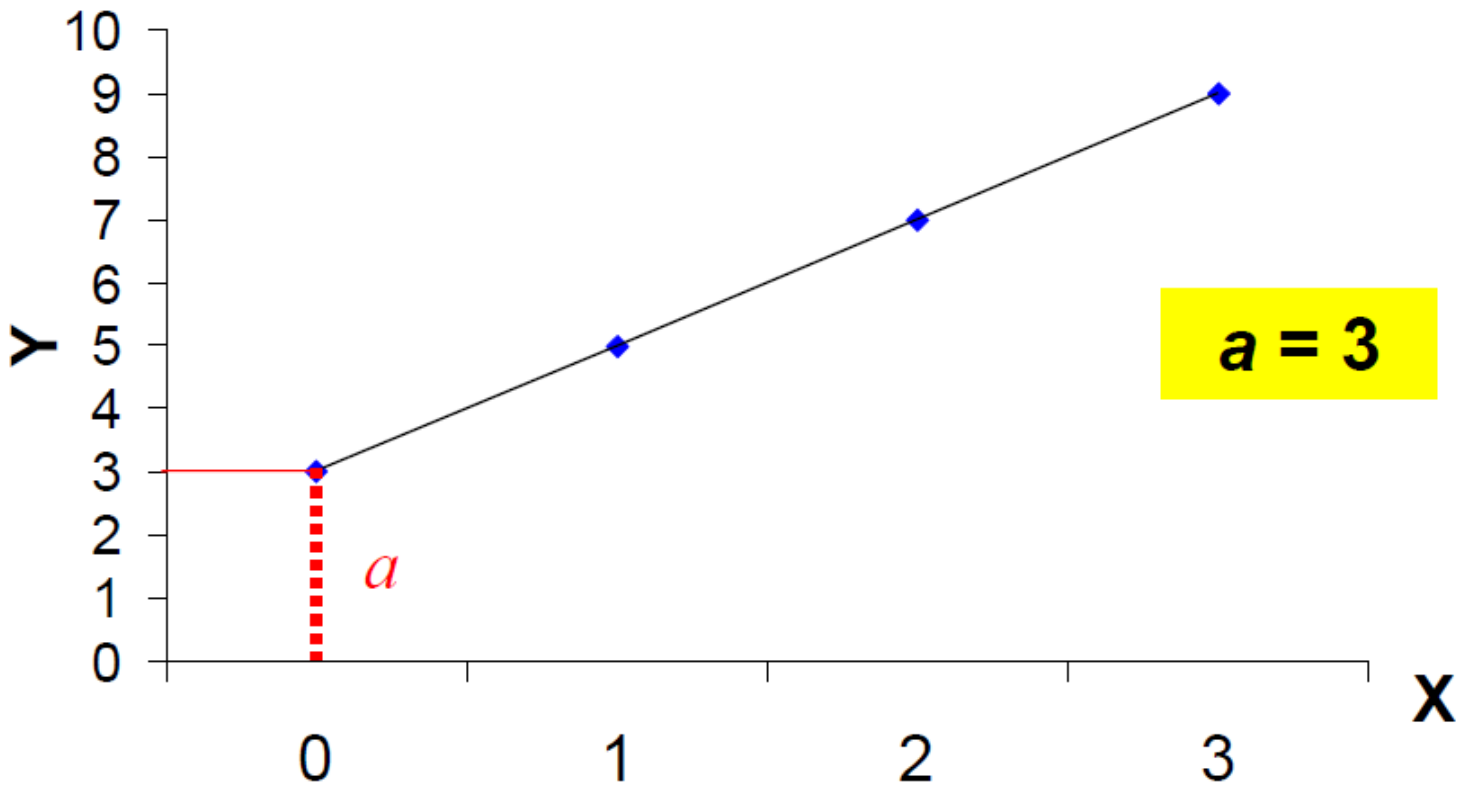
- b: Steigungs-/Regressionskoeffizient (engl. *slope*)
- a: Konstante (engl. *intercept*); beschreibt Höhenlage der Geraden bei x = 0 bzw. den Schnittpunkt mit der y-Achse

Grafische Darstellung b



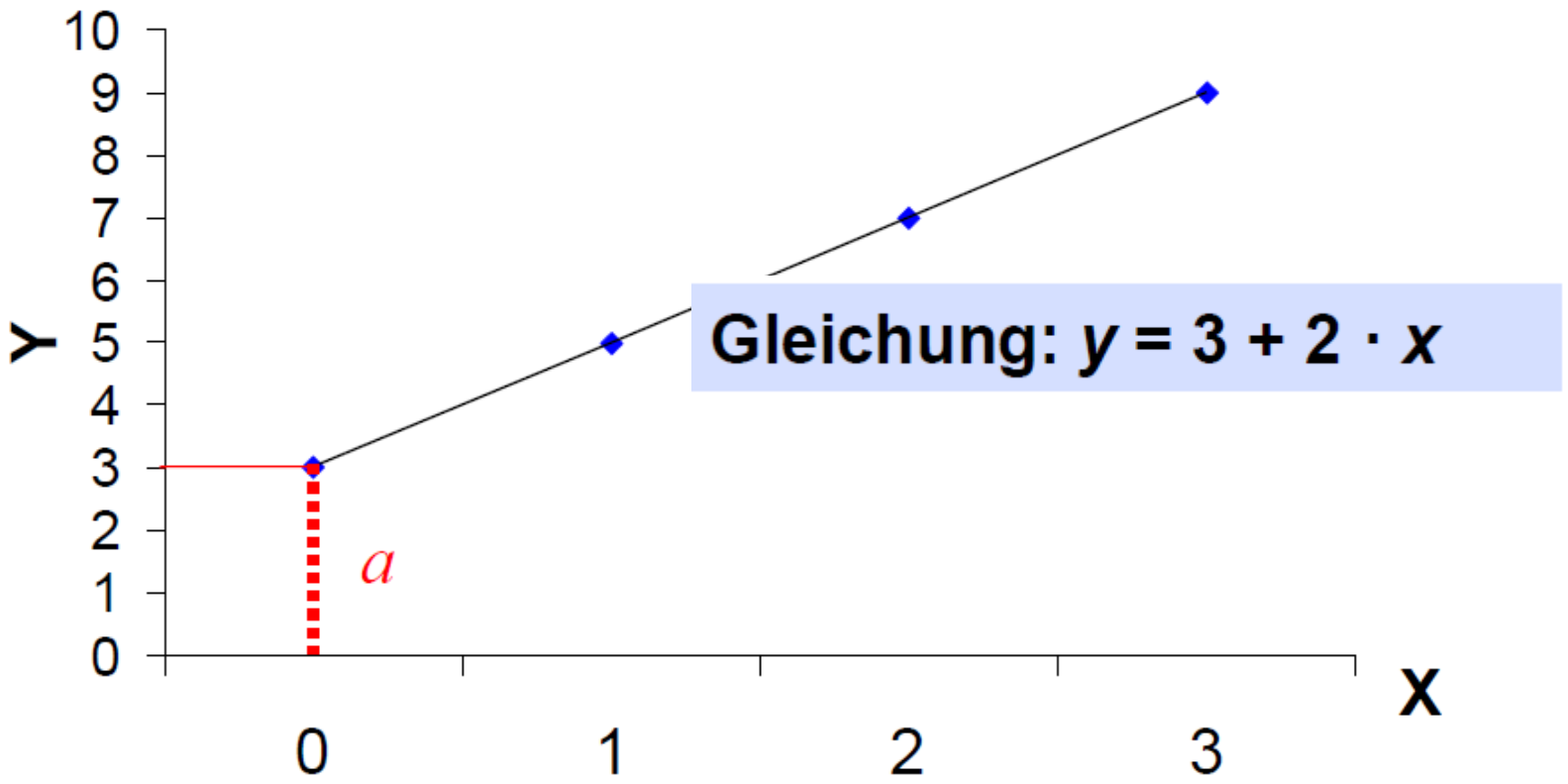
Um wie viele Einheiten ändert sich Y wenn sich X um eine Einheit nach rechts bewegt?

Grafische Darstellung a



Welchen Wert hat Y wenn x=0?

Grafische Darstellung



Regressionsgleichung Teil 1 formal

- $y_i=f(x_i) = \alpha + \beta \cdot x_i$

Oder (Formelsammlung): $y_i = \beta_0 + \beta_1 * x_i$

  - β/β_1 (od. b): Regressions- oder Steigungskoeffizient (*slope*), Regressionsgewicht
  - α / β_0 (od. a): Konstante (*intercept*); beschreibt die Höhenlage der Geraden bei x = 0 bzw. den Schnittpunkt mit der y-Achse an diesem Punkt
- Abinote= f(Intelligenz)
- Abinote= α + β(Intelligenz)

Bivariates lineares Regressionsmodell

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

- Aber: In der Regel gibt es in der Sozialforschung **keinen** perfekten Zusammenhang zwischen zwei Variablen
  - zu viele Störvariablen im „wirklichen Leben“
  - Verschiedene Faktoren beeinflussen einen Sachverhalt
  - Messfehler bei der Datenerfassung

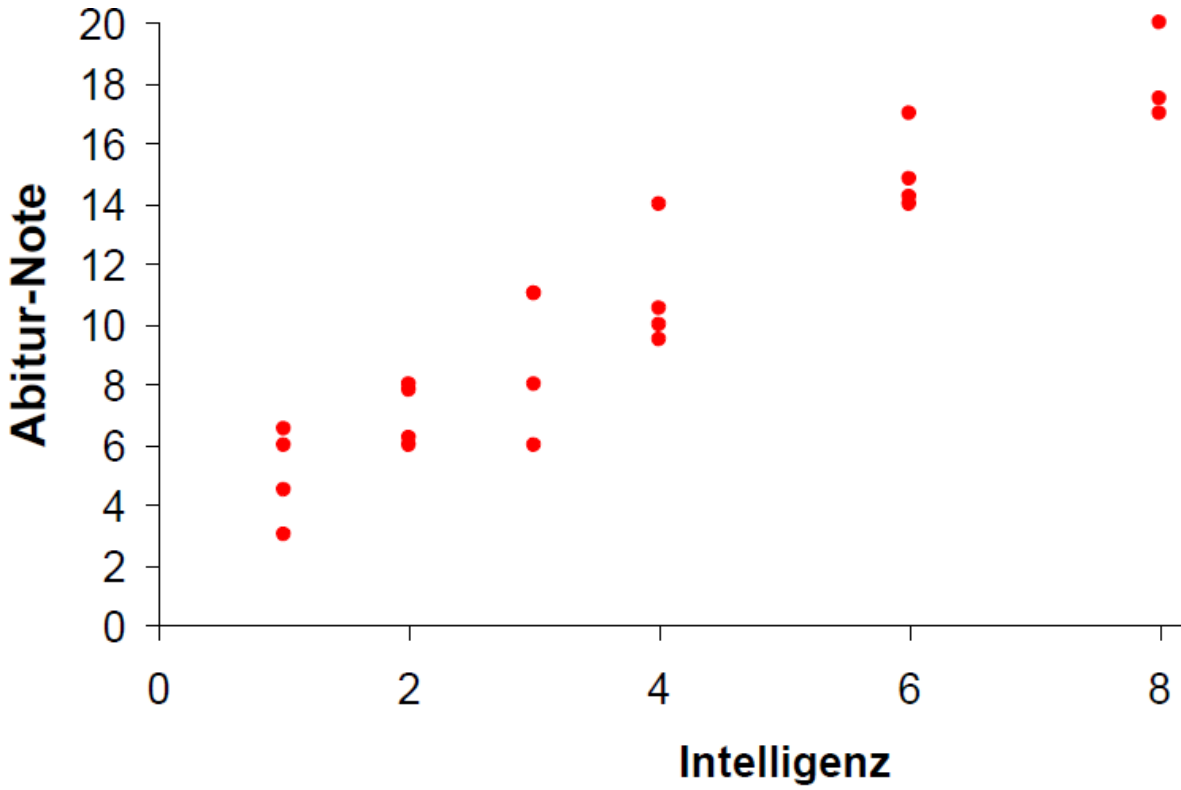
  →D.h. in der Regel haben wir es NICHT mit perfekten Zusammenhängen zwischen aV und uV zu tun!

  →D.h. auch: unsere Vorhersagen sind mit Unsicherheit behaftet

Bivariates lineares Regressionsmodell

Daten bilden keine Gerade, sondern Punktwolke!

Grundlagen bivariate lineare Regression



Regressere = Zurückführen auf

Ziel (bivariate) Regression: „Zurückführen von y auf x“

Allgemein: Punktwolke durch das beste lineare *Modell* erklären

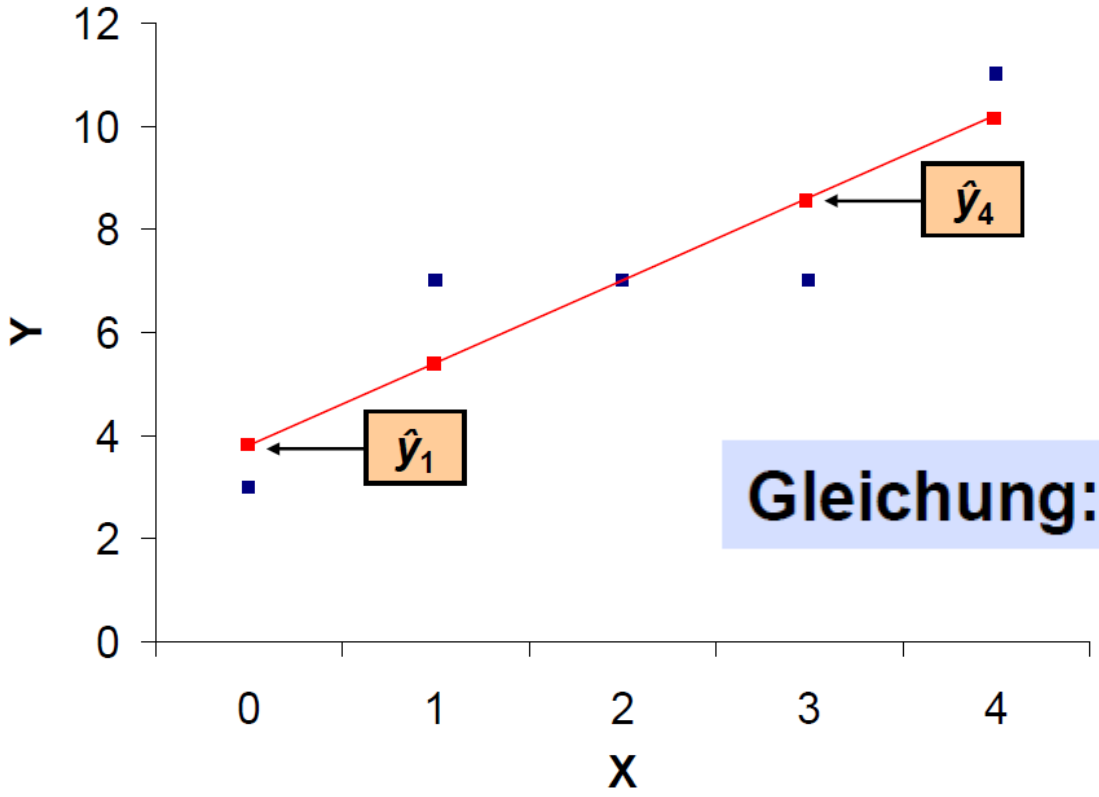
Bivariates lineares Regressionsmodell

- Eine Vorhersage ist trotz Störfaktoren möglich, aber: wird mit abnehmender Stärke des Zusammenhangs zwischen X und Y ungenauer
- Annahme eines linearen Zusammenhangs zwischen X und Y bedeutet, dass man eine Gerade durch das Streudiagramm legen kann
- Wie kann man die Genauigkeit maximieren? Beliebig viele Geraden möglich – wir wollen die „beste“ finden
- Vorgehen: **Man schätzt die vorhergesagten Werte so, dass der Vorhersagefehler *über alle Werte hinweg* so gering wie möglich ist**

Bivariate Regressionsanalyse

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

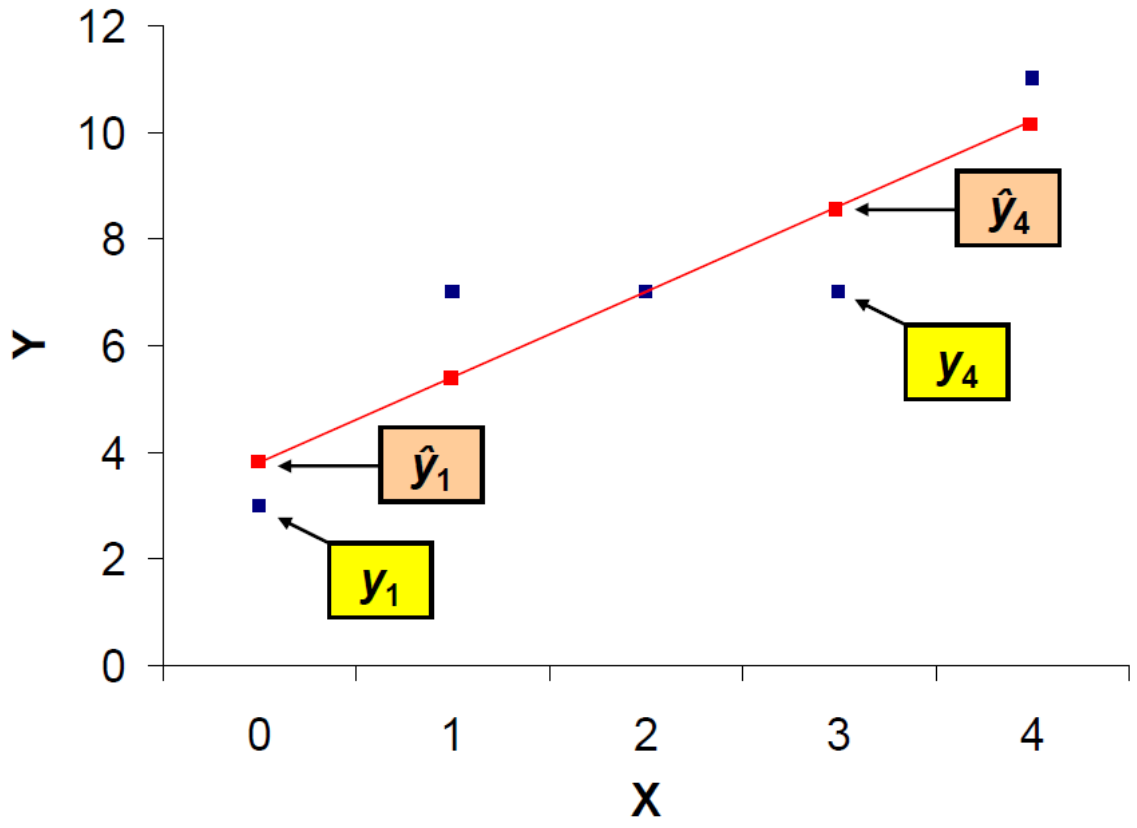
Punktwolke aV und uV mit Gerade



Gleichung: $y = 3 + 2 \cdot x$

Bivariate Regressionsanalyse

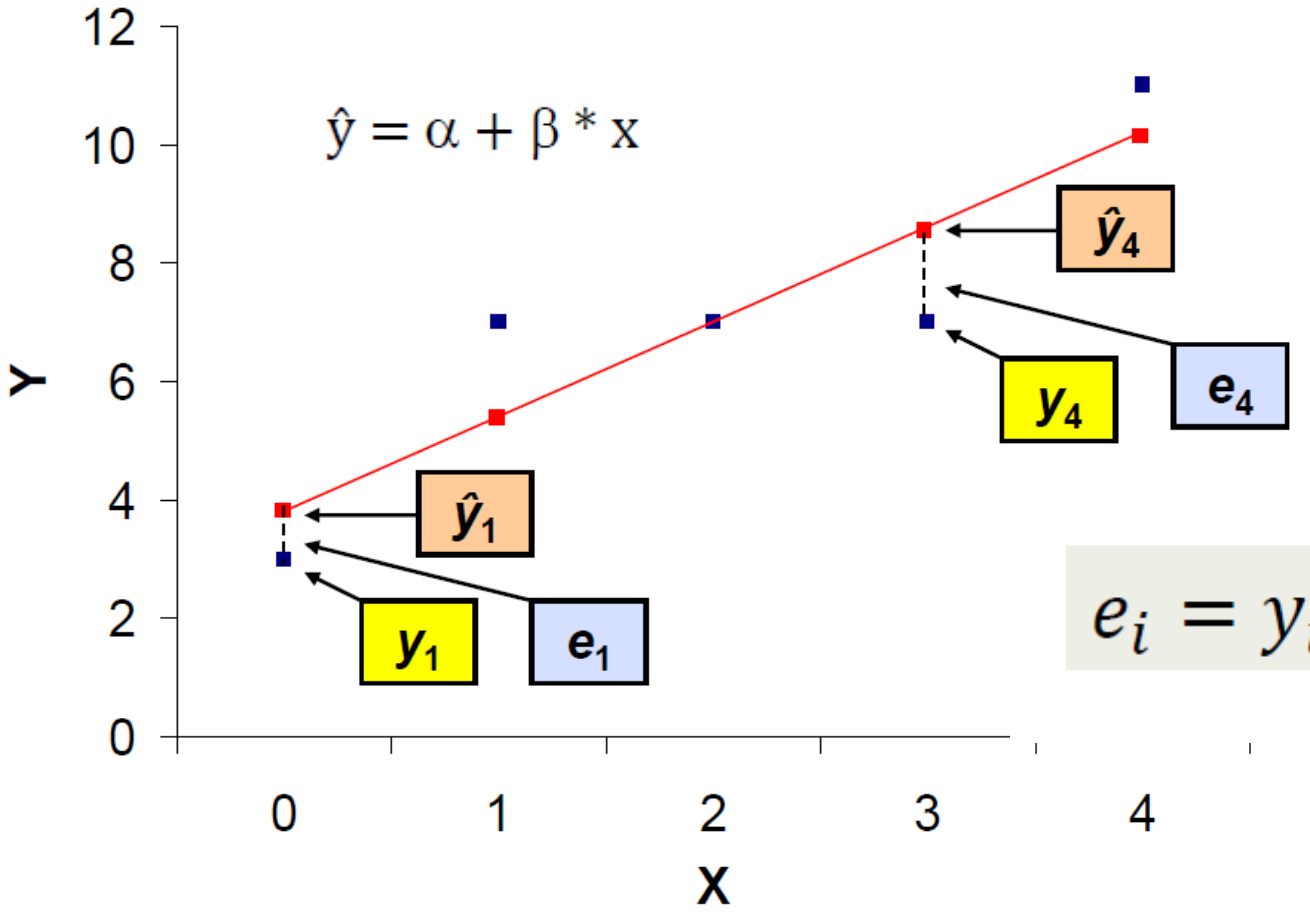
Punktwolke aV und uV mit Gerade und tatsächlichen sowie „geschätzten“ y-Werten



Bivariate Regressionsanalyse

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Punktwolke aV und uV mit Gerade, tatsächlichen und geschätzten y-Werten sowie Fehlerterm (Residuen e)



$$e_i = y_i - \hat{y}_i$$

Lineare Regression - OLS

- **Wir suchen die Gerade, bei der der *Abstand* aller Punkte zur Gerade minimal ist**
- **Wir nutzen diese Gerade, um die y-Werte bestmöglich vorherzusagen, zu „schätzen“**

$$\hat{y} = \alpha + \beta * x \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$$

- **Gleichung (Teil 2) Regressionsfunktion unter Berücksichtigung der Residuen:**

$$y = \alpha + \beta * x + e$$

Lineare Regression – OLS-Methode

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

- Regressionsgerade soll so durch die Punktewolke gelegt werden, dass **die Summe der quadrierten Regressionsresiduen minimal ist**

→ Ordinary Least Squares-Verfahren (OLS), Kriterium der kleinsten Quadrate

→ Formal:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \alpha + \beta * x_i)^2 = \text{Minimum}$$

Lineare Regression - OLS

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

OLS = Ordinary Least Square = Kleinste Quadrate Schätzer (Methode der kleinsten Quadrate)

- mathematisches Verfahren, das eine Konstante a und eine Steigung b schätzt, und das die lineare Beziehung zwischen X und Y am Besten beschreibt
- minimiert die Quadrate der geschätzten Fehler
- ist **BLUE** (*best unbiased linear estimator*) nach Gauss-Markov

Lineare Regression - OLS

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

- **Minimierungsvorschrift ist erfüllt, wenn** die Regressionsparameter wie folgt bestimmt werden:

- (Geschätzter) Regressionskoeffizient β_1 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- (geschätzte) Regressionskonstante $\propto$ oder β_0 :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \bar{x}$$

Beispiel: lineare Regression

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Bivariates Regressionsmodell zur Erklärung der Lebenszufriedenheit durch Einkommen

Die Variable Lebenszufriedenheit wird im ALLBUS mit einer *elfstufigen Skala* erfasst.

Frageformulierung: *„Und jetzt noch eine allgemeine Frage. Wie zufrieden sind Sie gegenwärtig – alles in allem – mit ihrem Leben?“ Die Befragten können dabei Werte von 0 bis 10 angeben, wobei der Wert 0 „ganz und gar unzufrieden“ und der Wert 10 „ganz und gar zufrieden“ bedeutet*.

Bsp. Lineare Regression - OLS

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

ID	Lebenszufriedenheit (Wert auf Skala von 0 bis 10)	Nettoeinkommen im Monat in Euro
1	7	2000
2	10	4550
3	2	1003
4	9	3200
5	7	2900
6	6	2850
7	4	1900
8	6	3700

Quelle: Eigene Darstellung

$\bar{y}$=6,38

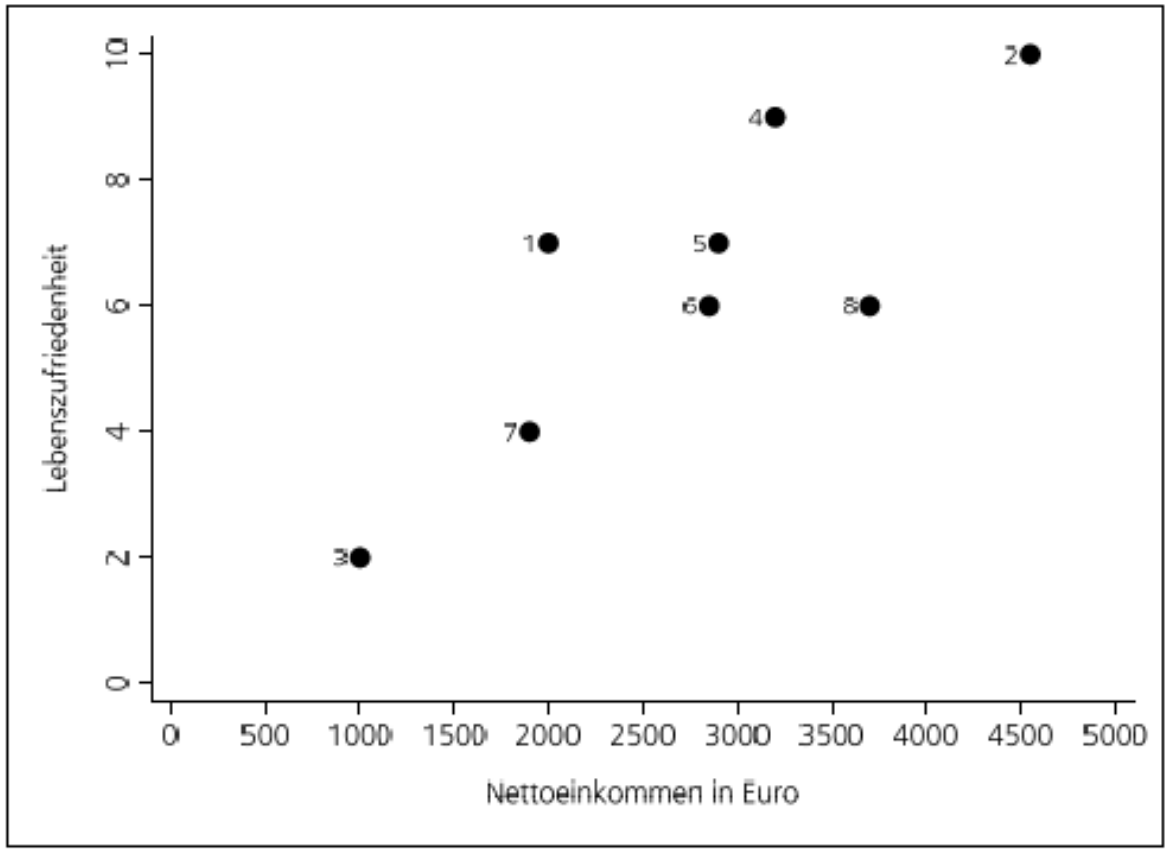
s^2=6,55

$\bar{x}$=2762,88

s^2=1244740,41

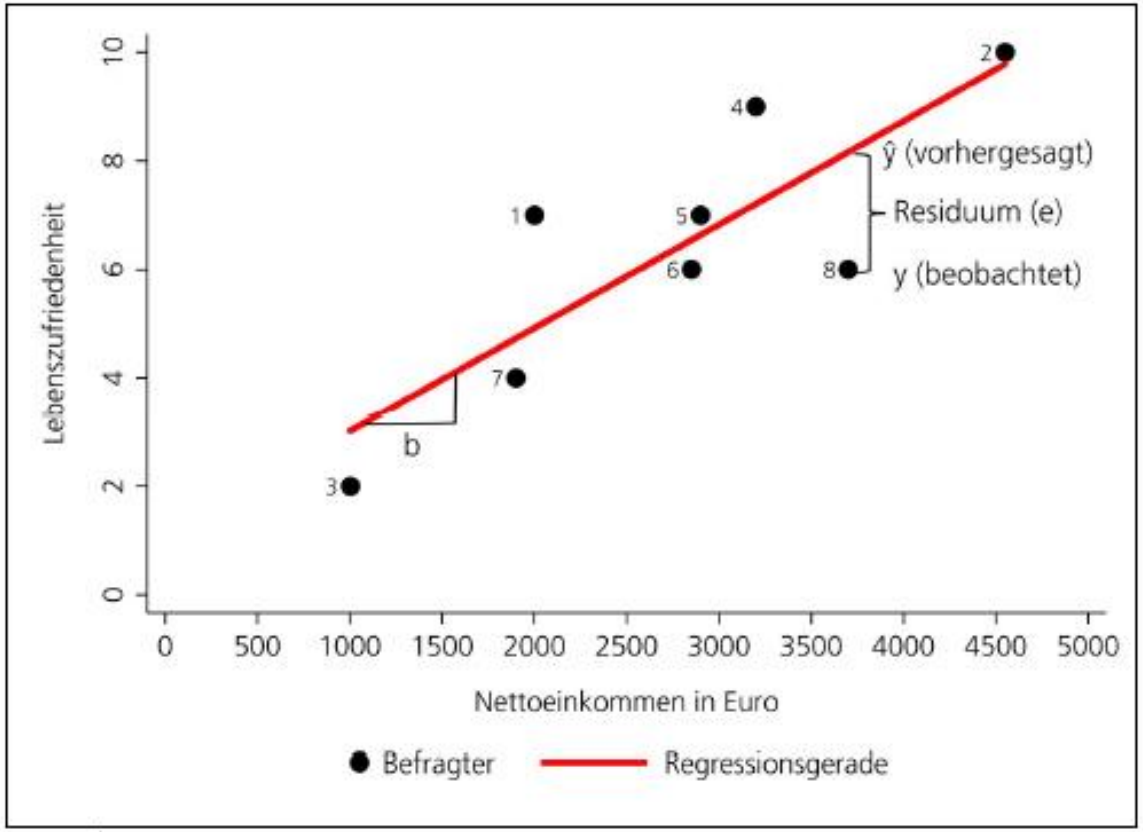
Cov_{xy}=2371,34

Bsp. Lineare Regression - OLS



Quelle: Eigene Darstellung

Bsp. Lineare Regression - OLS



Quelle: Eigene Darstellung

Bsp. Lineare Regression

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

ID	Lebenszufriedenheit (Wert auf Skala von 0 bis 10)	Nettoeinkommen im Monat in Euro
1	7	2000
2	10	4550
3	2	1003
4	9	3200
5	7	2900
6	6	2850
7	4	1900
8	6	3700

Quelle: Eigene Darstellung

$\bar{y}$=6,38

s^2=6,55

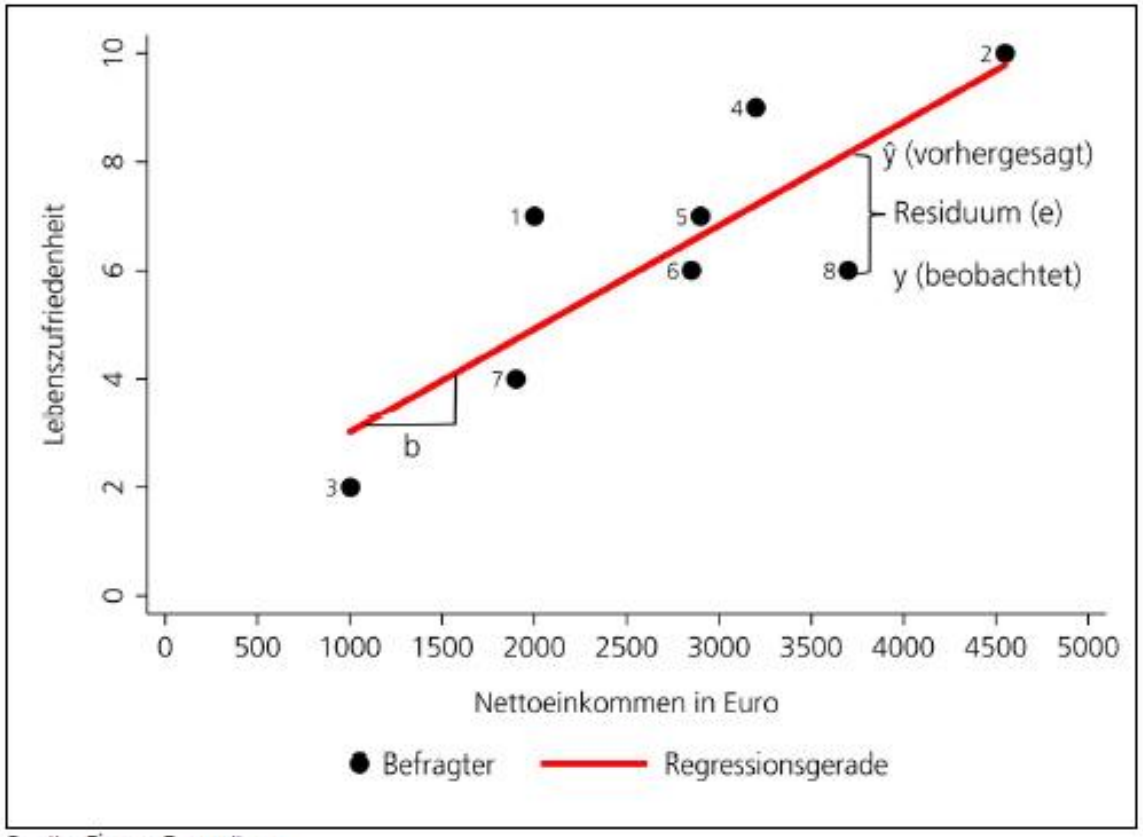
$\bar{x}$=2762,88

s^2=1244740,41

Cov_{xy}=2371,34

Bsp. Lineare Regression - OLS

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \bar{x}$$



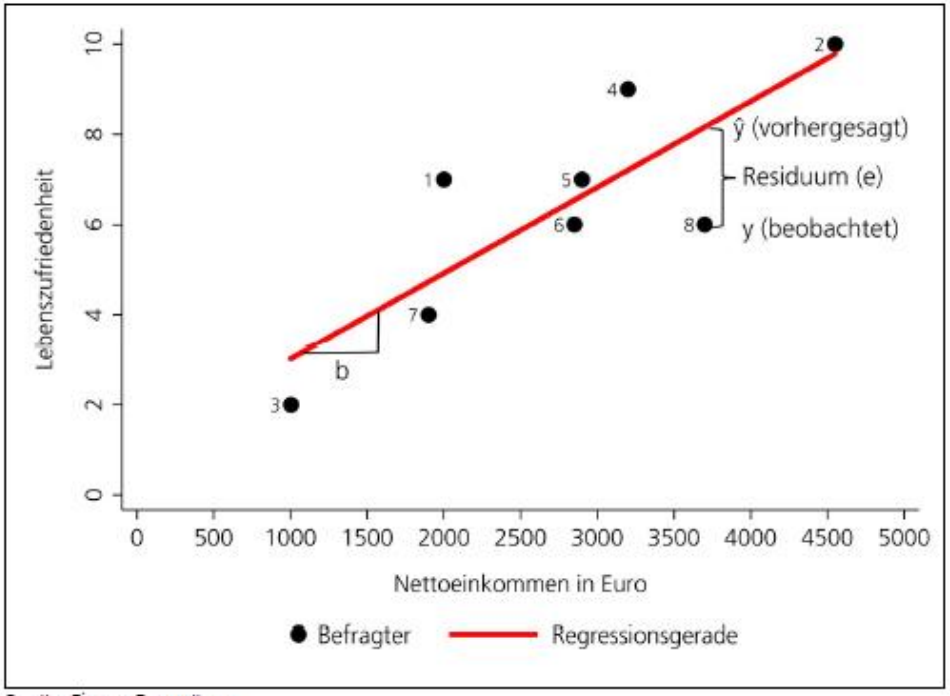
Quelle: Eigene Darstellung

$$\widehat{\beta_1} = \frac{2371{,}34}{1.244.740{,}41} = 0{,}002$$

$$\widehat{\beta_0} = 6{,}38 - \frac{2371{,}34}{1.244.740{,}41} * 2762{,}88 = 1{,}11$$

Bsp. Lineare Regression - OLS

$\hat{y}_i = 1{,}11 + 0{,}002x_i$; oder $\text{Lebenszufriedenheit}_i = 1{,}11 + 0{,}002\ \text{Einkommen}_i$

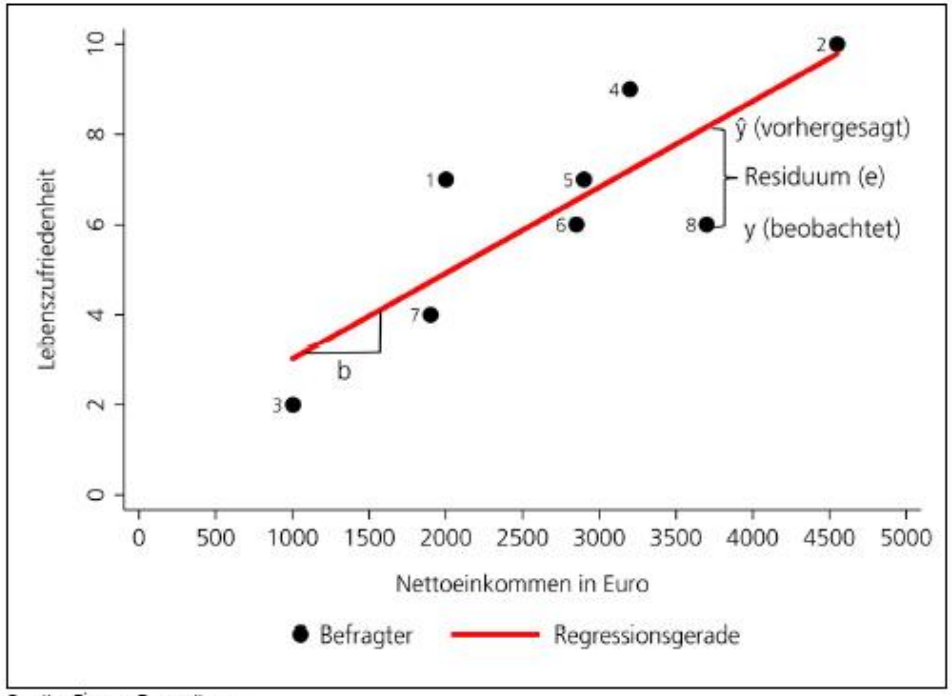


Quelle: Eigene Darstellung

Übungsfrage: Wie hoch ist die geschätzte Lebenszufriedenheit einer Person 9 mit 3000 Euro Einkommen?

Bsp. Lineare Regression - OLS

$\hat{y}_i = 1{,}11 + 0{,}002x_i$ oder Lebenszufriedenheit$_i$ = 1,11 + 0,002 Einkommen$_i$



Quelle: Eigene Darstellung

Übungsfrage: Wie hoch ist die geschätzte Lebenszufriedenheit einer Person 9 mit 3000 Euro Einkommen?

$\hat{y}_9 = 1{,}11 + 0{,}002x_i$ oder
Lebenszufriedenheit$_9$ =
1,11 + 0,002 * 3000= 7,11

Grundlagen bivariate Regression

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Was ist:

- die Richtung (Vorzeichen des Regressionskoeffizienten b!)
- die Stärke (Regressionskoeffizient b!)
- die statistische Signifikanz (nächste Einheit)

… des Einflusses von X auf Y?

Und:

- **Wie gut „erklärt“ die Regressionsgerade (unser Modell) „die Realität“?**

Determinationskoeffizient R^2

- Auch ***Bestimmtheitsmaß*** oder ***Prozentsatz der erklärten Varianz***
- Maß für die Güte der Anpassung der Regressionsfunktion an die beobachteten Daten
- Bestimmung durch globale Prüfung der Regressionsfunktion; gibt an, wie gut die einbezogene(n) UV(s) die aV erklären
- Ist ein **PRE-Maß**
- Kann Werte zwischen 0 und 1 annehmen. Je näher R^2 an 1 ist, desto besser erklärt das spezifizierte Modell die Streuung
- Beispiel: $R^2 = 0{,}5$ => 50% der Varianz der abhängigen Variable kann durch das spezifizierte Modell erklärt werden

Varianzerklärung durch R^2

Wie gut erklärt die Schätzung die Realität?

Bei der Berechnung wird die Gesamtvarianz s_y^2 der abhängigen Variable y in zwei Teile zerlegt:

1) in die durch die (geschätzte) Regressionsfunktion erklärte Varianz $s_{\hat{y}}^2$

2) in die nicht erklärte „Restvarianz“

$$R^2 = \frac{\text{Varianz der vorhergesagten Werte}}{\text{Varianz der beobachteten Werte}}$$

R^2

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

$$R^2 = \frac{\text{Varianz der vorhergesagten Werte}}{\text{Varianz der beobachteten Werte}} \qquad R^2 = \frac{SS_{model}}{SS_{total}}$$

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

- Anteil der durch die Regressionsfunktion aufgeklärte Streuung; Güte der Anpassung der Regressionsfunktion an die empirischen Daten
- berechnet den **Anteil** der Varianz von *Y*, der durch *X* erklärt werden kann

Übung R^2

Zuhause bzw. Tutorium

- Wir wollen für das Beispiel (Lebenszufriedenheit und Einkommen) den Determinationskoeffizienten berechnen. Wie gehen Sie vor und wie viel Prozent der Varianz wird durch unser „Modell“ erklärt?

Prüfung / Interpretation Regressionsfunktion

- R^2: Wie gut erklärt das Gesamtmodell die „Realität“?
- Wie gut erklären die einzelnen Variablen das geschätzte Modell?
- Regressionskoeffizient(en):
  - geben Ausmaß der Steigerung von Y an für den Fall dass X um eine Einheit steigt
  - in welche Richtung geht die Beziehung zwischen Y und X?
- Signifikanzprüfung (nächste Woche):
  - T-Tests prüfen die einzelnen Regressionskoeffizienten auf stat. Signifikanz
  - F-Test prüft Gesamtgüte des Modells (R^2) auf stat. Signifikanz

Multiple lineare Regression

Hintergrund

- Bivariate lineare Regressionsanalyse: eine abhängige Y-Variable (aV) wird anhand einer unabhängigen X-Variablen (uV) vorhergesagt → lineare „Einfachregression“

- Beispiel: Lebenszufriedenheit (Y) wird vorhergesagt durch Einkommen (X)

Multiple lineare Regression

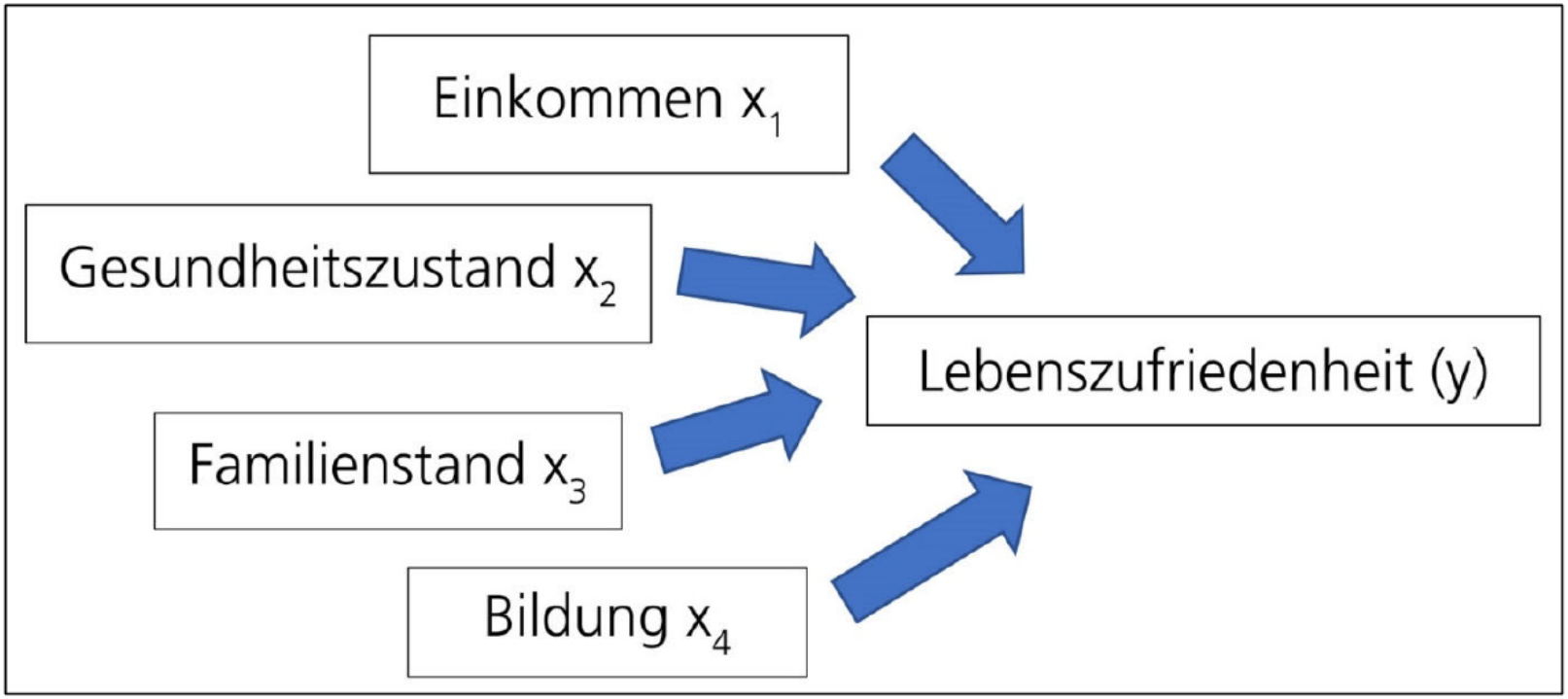
- Aber: i.d.R. gehen wir davon aus, dass mehrere Merkmale einen Sachverhalt „erklären“
- Beispiel: Studiendauer wird durch Anzahl Wochenstunden (x_1), durch allgemeine Studienbedingungen (x_2), durch Nebentätigkeit (x_3),…(x_k) beeinflusst

- Oder: Beispiel Lebenszufriedenheit

Multiple lineare Regression

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Abbildung 18: Schematische Darstellung der vermuteten multivariaten Einflussstruktur



Quelle: Eigene Darstellung

Diese und weitere Abbildungen wurden aus Kapitel 4 des Lehrbriefs entnommen

Multiple lineare Regression

- Zur Verbesserung der Vorhersage und/oder zum Test mehrerer theoretischer Annahmen bietet es sich an, mehrere uVs in das Regressionsmodell mit aufzunehmen

→ Ziel der multiplen Regression:

  → Y auf Grundlage von zwei bzw. mehr uVs (X_1, X_2,…, X_k) bestmöglichst vorherzusagen

  → Verschiedene uVs vergleichend in ihrem Einfluss prüfen

- Weitgehend parallele Logik und Interpretation wie im einfachen bivariaten Modell
- Wesentliche Veränderung: Vorstellung einer durch die Regressionsgleichung beschriebenen Regressionsgeraden lässt sich nicht mehr beibehalten (eher „Regressionsebene“)

Der Begriff der „Kontrolle“

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

- In sozialwissenschaftlichen Analysen üben in der Regel viele (unabhängigen) Merkmale einen Einfluss auf die abhängige Variable aus
- Mit Hilfe der multiplen linearen Regression können wir alle uVs in das Regressionsmodell integrieren
- Um die Auswirkung der Änderung einer Variablen zu untersuchen, werden dabei alle anderen uVs konstant gehalten.
- Dies nennt man auch für andere Variablen zu „kontrollieren“