FFHS
Fernfachhochschule Schweiz
Mitglied der SUPSI

Multivariate, beschreibende Statistik und Lin. Regression

ESDS, PVA 2

Dr. Ivan Moser, 09.03.24

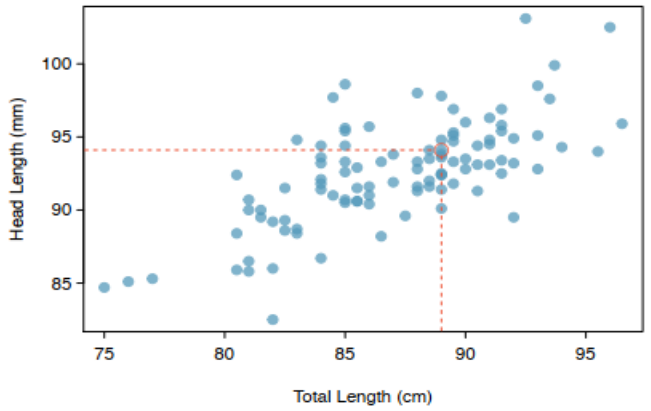
Inhalt

- (Repetition univariate, explorative Datenanalyse)
- Bivariate, explorative Datenanalyse
  - Metrische Variablen
    - Streudiagramme
  - Kategoriale Variablen
    - Kontingenztabelle
    - Säulendiagramme
    - Mosaik-Plot
- Korrelation
- Lineare Regression (deskriptiv)

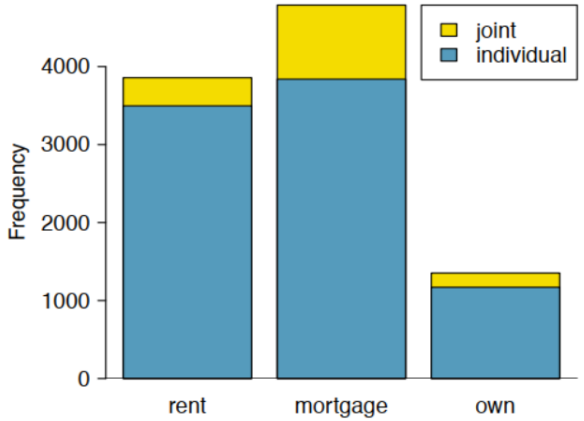
Fragen und Wünsche für die PVA?

Übersicht

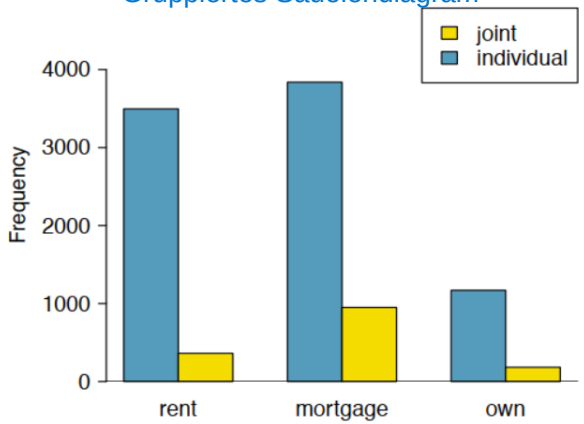
Scatterplot



Gestapeltes Saulendiagramm



Gruppiertes Sauelendiagram

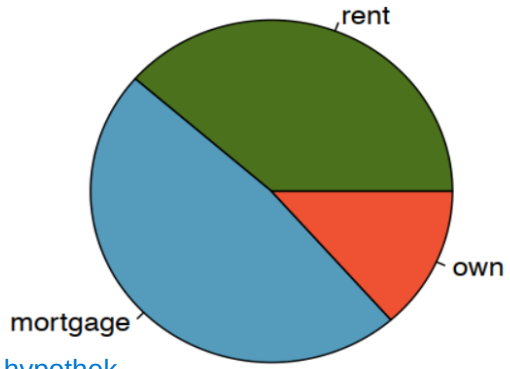


Mosaikdiagram

(b)

Standardisiertes Sauelendiagram

Bei einer Variablen



hypothek

Aufgabe bivariate, explorative Datenanalyse 1

- Welche Darstellungsformen eignen sich zur Visualisierung des Zusammenhanges zwischen zwei kategorialen Variablen?

  - Streudiagramm (scatterplot)
yes - Gestapeltes Säulendiagramm (stacked barplot)
  - Einfaches Säulendiagramm (barplot)
  - Gruppiertes Kreisdiagramm (grouped pie chart)
yes - Gruppiertes Säulendiagramm (grouped barplot)
yes - Mosaikdiagramm (mosaic plot)
  - Kreisdiagramm

Aufgabe Kontingenztafel

- Sie möchten zeigen, dass der Ehestatus einen Einfluss auf die Lebenseinstellung hat. Welche Variante einer Kontingenztafel ist am sinnvollsten und weshalb?

	dull	routine	exciting	Total
married	21	241	251	513
widowed	17	54	40	111
divorced	10	74	65	149
separated	6	11	8	25
never married	11	79	108	198
Total	65	459	472	996

a) Rohwerte

	dull	routine	exciting	Total
married	0.04	0.47	0.49	1
widowed	0.15	0.49	0.36	1
divorced	0.07	0.5	0.44	1
separated	0.24	0.44	0.32	1
never married	0.06	0.4	0.55	1

b) Relative Häufigkeiten zeilenweise

	dull	routine	exciting
married	0.32	0.53	0.53
widowed	0.26	0.12	0.08
divorced	0.15	0.16	0.14
separated	0.09	0.02	0.02
never married	0.17	0.17	0.23
Total	1	1	1

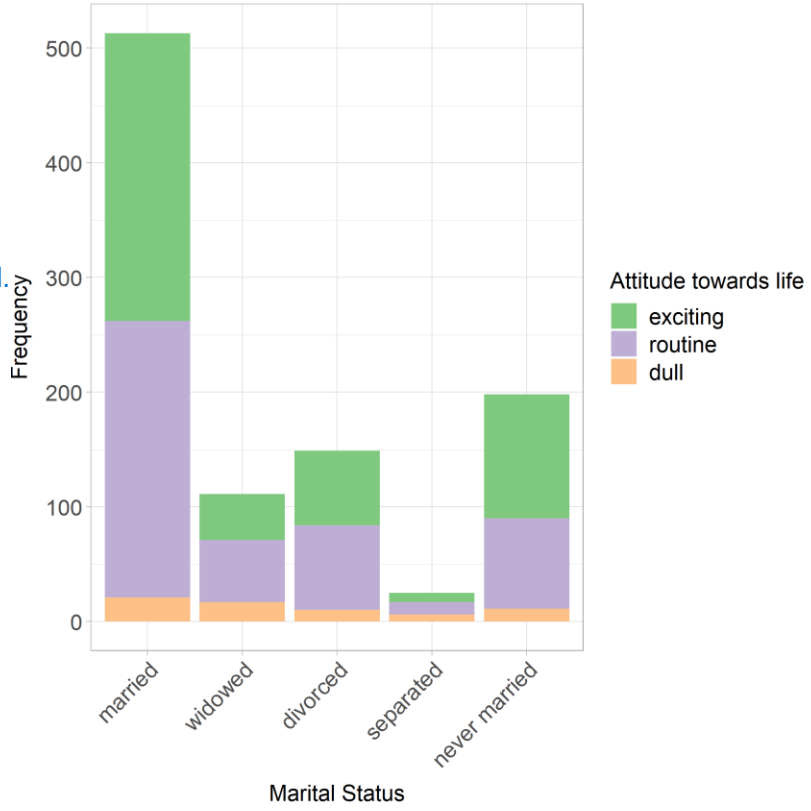
c) Relative Häufigkeiten spaltenweise

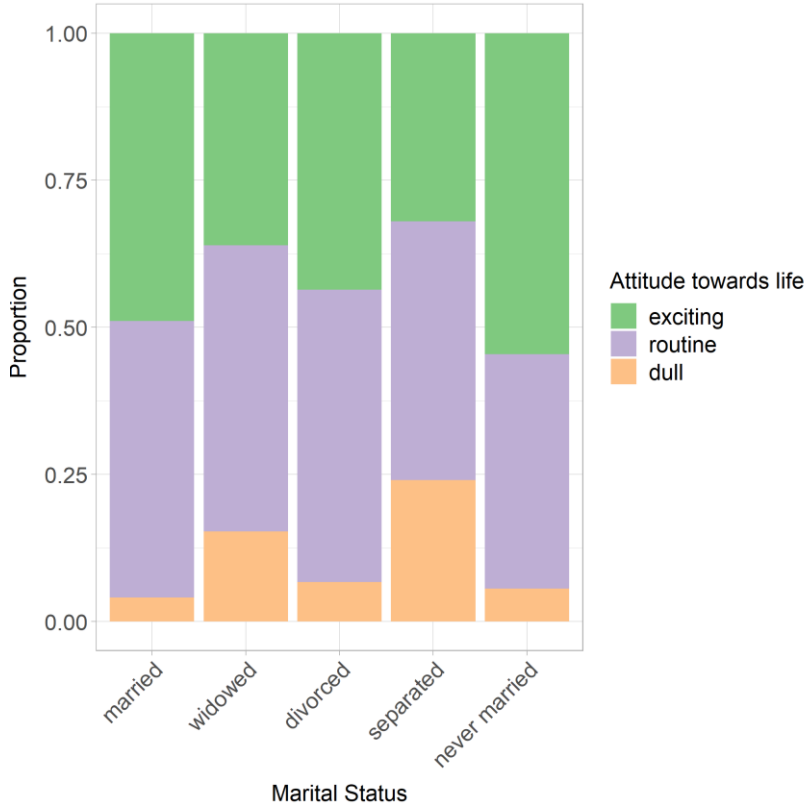
Aufgabe bivariate, explorative Datenanalyse 2

- Was sind die Vor- und Nachteile von gestapelten Säulendiagrammen und standardisierten, gestapelten Säulendiagrammen?

Links: Nicht standardisiert
Man sieht das nur wenige Personen angegeben haben, dass sie getrennt sind.

Diese Information sieht man rechts nicht.





Aufgabe Mosaikplot

Es ist rude vor allem fuer die Personen welche es selbst nie machen.

Quelle: Vincenzo Pinto / Getty Images

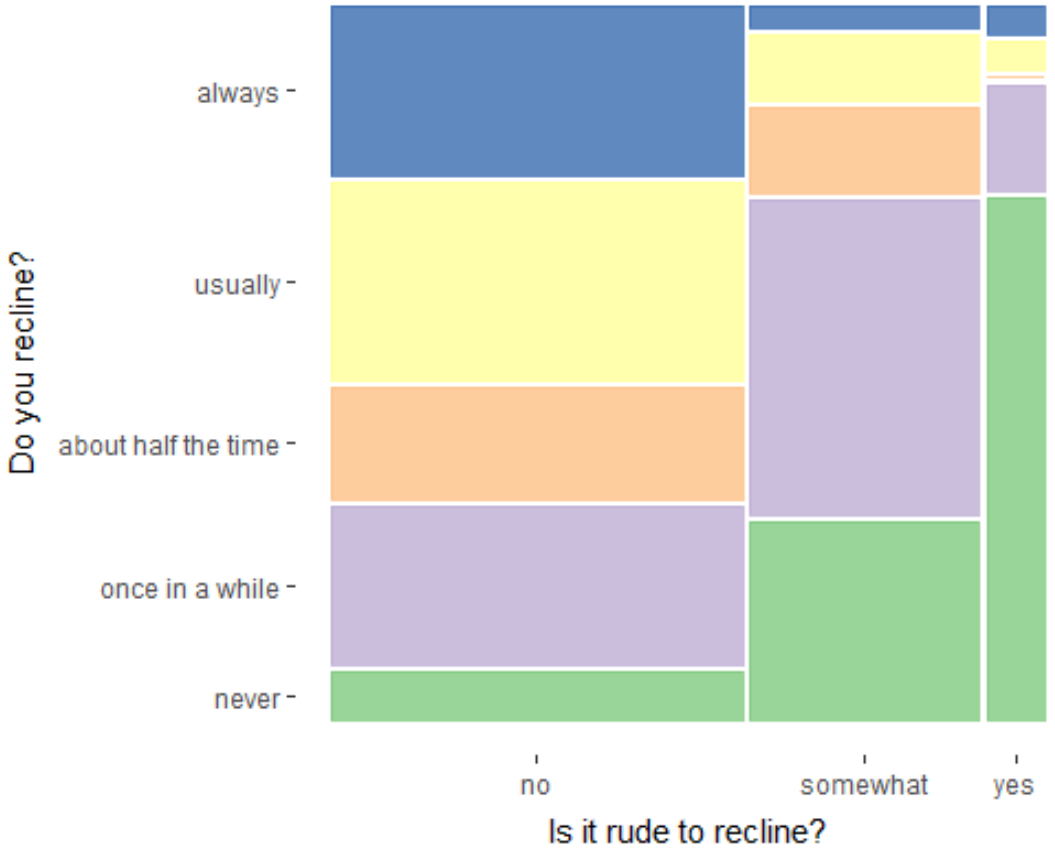
In einer Studie wurden Flugpassagiere dazu befragt, ob sie es unhöflich finden, wenn ein Flugpassagier vor ihnen den Sitz nach hinten klappt ("Is it rude to recline") und wie häufig sie selbst den Sitz nach hinten klappen ("Do you recline?"). Die Resultate wurde mittels Mosaikplot dargestellt.
Welche Aussage zum Mosaikplot ist korrekt?

Falsch
- In der Kategorie der Personen , die den Sitz nie nach hinten klappen (never) und es unhöflich finden, den Sitz nach hinten zu klappen (yes) befinden sich absolut gesehen die meisten Personen.

Richtig
- Der Anteil der Personen die den Sitz nie nach hinten klappen ist unter denjenigen Personen am höchsten, die es unhöflich finden, wen jemand den Sitz nach hinten klappt.

Richtig
- Es gibt keinen erkennbaren Zusammenhang zwischen der Einstellung gegenüber dem Runterklappen und dem eigenen Verhalten.

- Die vorliegende Wahl der Achsen für die beiden Variablen eignet sich vor allem dann, wenn das eigene Verhalten als Erklärvariable für die Einstellung gegenüber dem Runterklappen gesehen wird.



FFHS
Fernfachhochschule Schweiz
Mitglied der SUPSI

Korrelation

Grundidee Korrelationskoeffizient

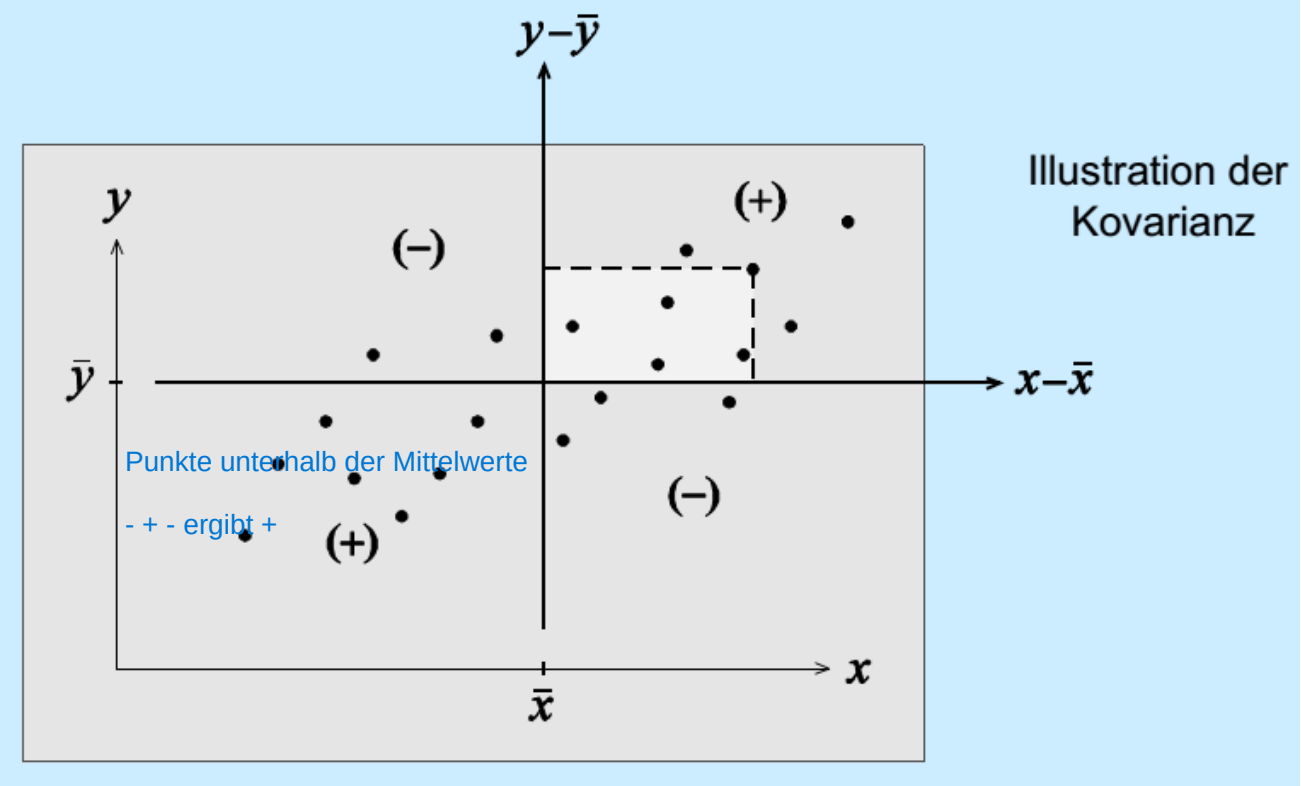
- **Kovarianz** (cov(x, y)): Das „miteinander variieren“
  - je weiter der Wert von Null entfernt ist, desto enger der lineare Zusammenhang zwischen den zwei Variablen
    - positiver Wert: gleichsinniger linearer Zusammenhang
    - negativer Wert: gegensinniger linearer Zusammenhang

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Grundidee Korrelationskoeffizient

- Illustration Kovarianz

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$



Summen werden immer aufsummiert

Grundidee Korrelationskoeffizient

- **Korrelationskoeffizient** (r): standardisiertes Zusammenhangsmass
  - gibt Auskunft über die **Stärke des linearen Zusammenhangs**/des Kovariierens
  - möglicher Wertebereich: -1 bis 1, weil hier...
  - ...die Kovarianz in einen einheitlichen Wertebereich transformiert wird (Division durch das Produkt der Standardabweichungen der beiden Variablen)

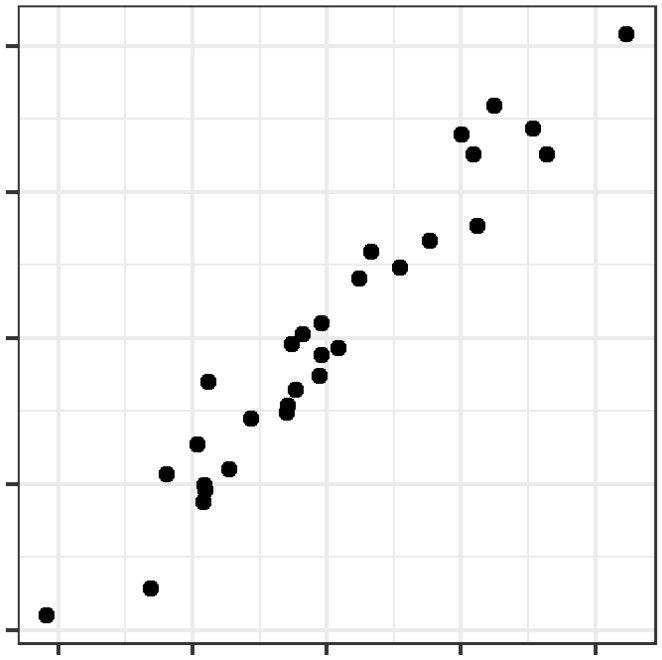
Pearson-Bravais Korrelationskoeffizient:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

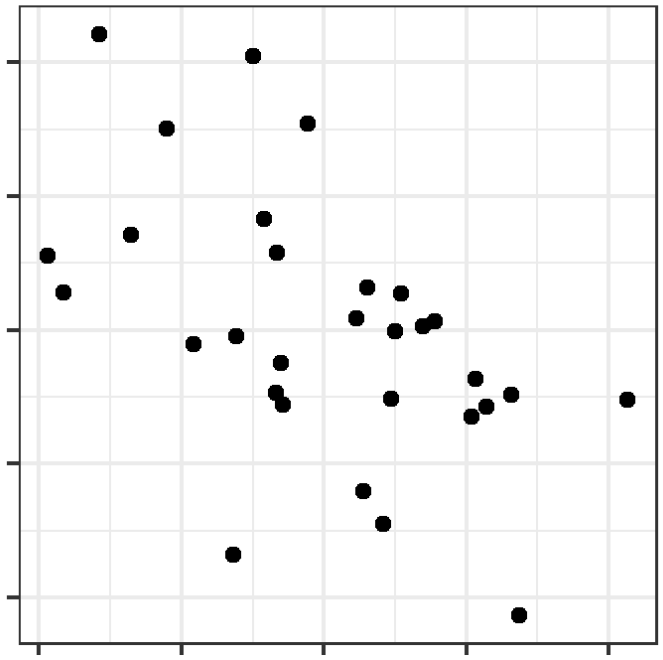
Höhe und Richtung des Korrelationskoeffizienten

ab 0.5 ist es eine hohe Korrelation

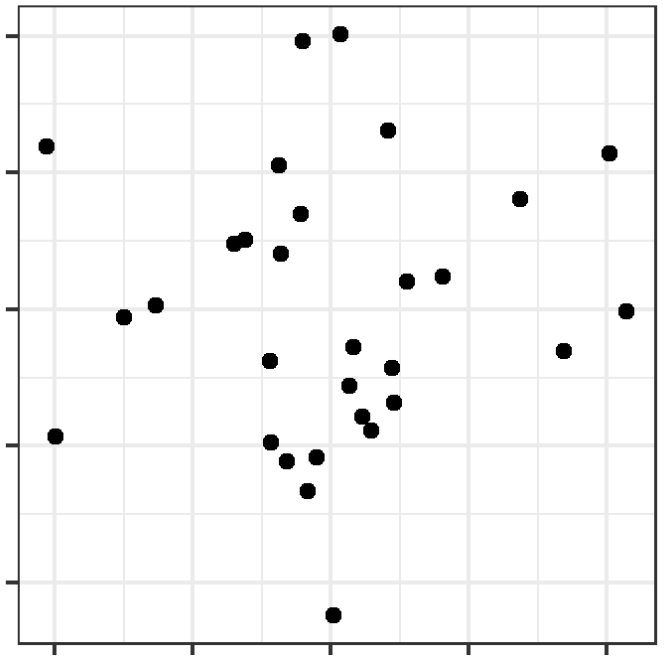
- Schätzt den Korrelationskoeffizienten für folgende «Punktewolken»



Korrelatonswert zwischen 0,8 und 0,9



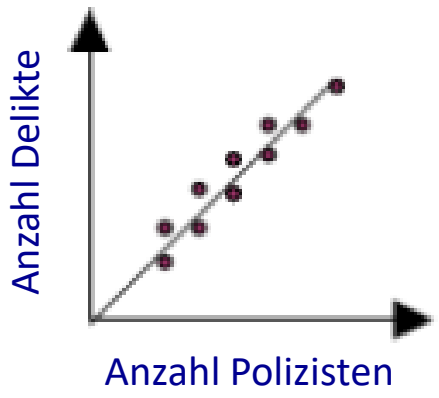
negativ weil oben links nach unten rechts
-0.52

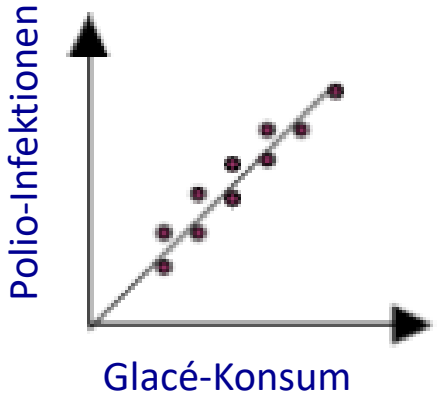


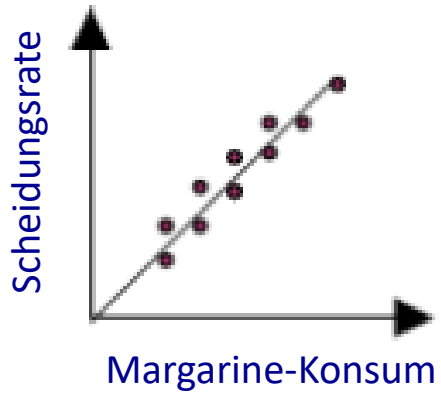
Korrelation um 0

Vorsicht bei der Interpretation

- **Korrelation ≠ Kausalität (Ursache – Wirkung)!**
  - X führt zu Y
  - Y führt zu X
  - Wechselseitige Beziehung
  - eine/mehrere Drittvariable(n) führen zur Korrelation zwischen A und B
  - zufällige Korrelation (kein sachlicher Zusammenhang)



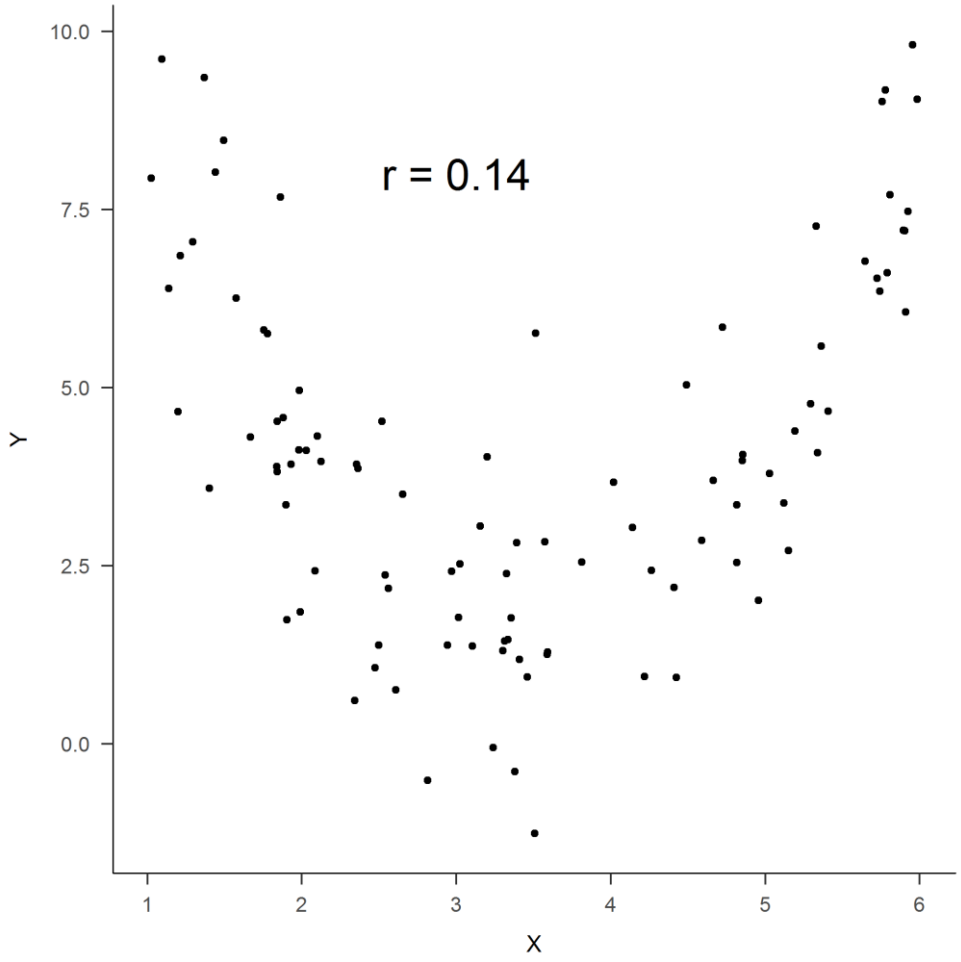




Vorsicht bei der Interpretation

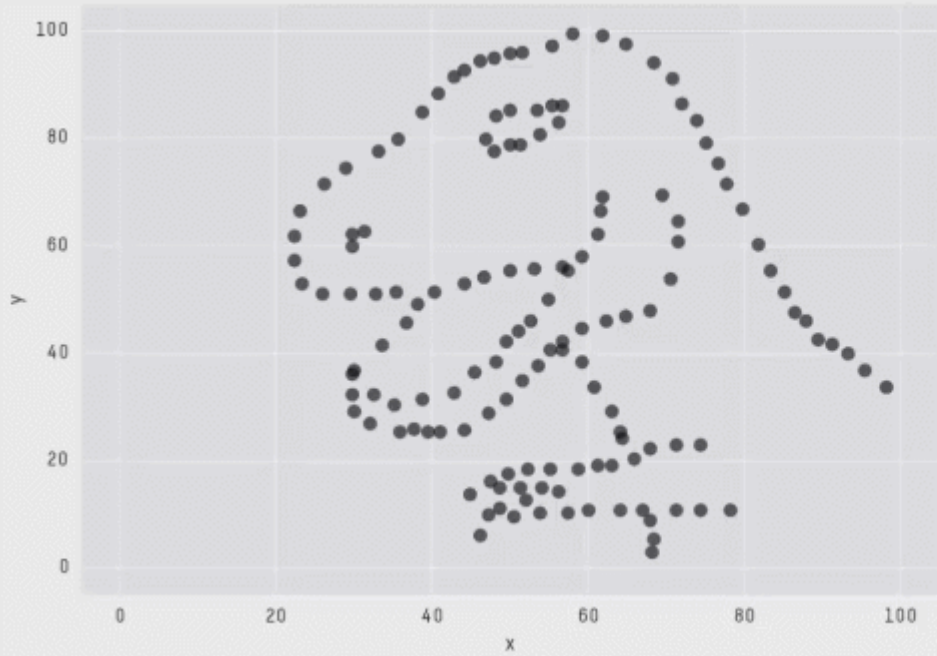
- Nehmen wir an, wir erstellen ein Streudiagramm für zwei Variablen X und Y (siehe rechts). Ausserdem berechnen wir die Korrelation, sie beträgt r = 0.14. Bedeutet dies, dass es keinen Zusammenhang zwischen A und B gibt?
  Weshalb müssen wir hier vorsichtig sein?

Zusammenhang ja aber nicht linear



Vorsicht bei der Interpretation

Daten immer plotten



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Aufgabe Korrelation

Kovarianz ist abhaengig von der Einheit.

Welche der folgenden Aussagen ist korrekt?

- Eine Korrelation $r_{XY} = 0.54$ bedeutet, dass der Zusammenhang von zwei Variablen X und Y am besten als ein linearer Zusammenhang beschrieben werden kann.
- Die Korrelation eignet sich besser als die Kovarianz, um die Zusammenhänge zwischen verschiedenen Variablen zu vergleichen.

Standardisiertes Mass und nicht abhaengig von Werten

- Eine positive Korrelation von 0.4 zwischen sportlicher Aktivität und psychischer Gesundheit bedeutet, dass mehr sportliche Aktivität zu besserer psychischer Gesundheit führt.
- Eine Korrelation von 0 bedeutet, dass es keinen Zusammenhang zwischen zwei Variablen gibt.

FFHS
Fernfachhochschule Schweiz
Mitglied der SUPSI

Lineare Regression

Lineare Regression

- **Ziel:**
  - Beschreibung von linearen Zusammenhängen
  - Vorhersage von neuen Daten

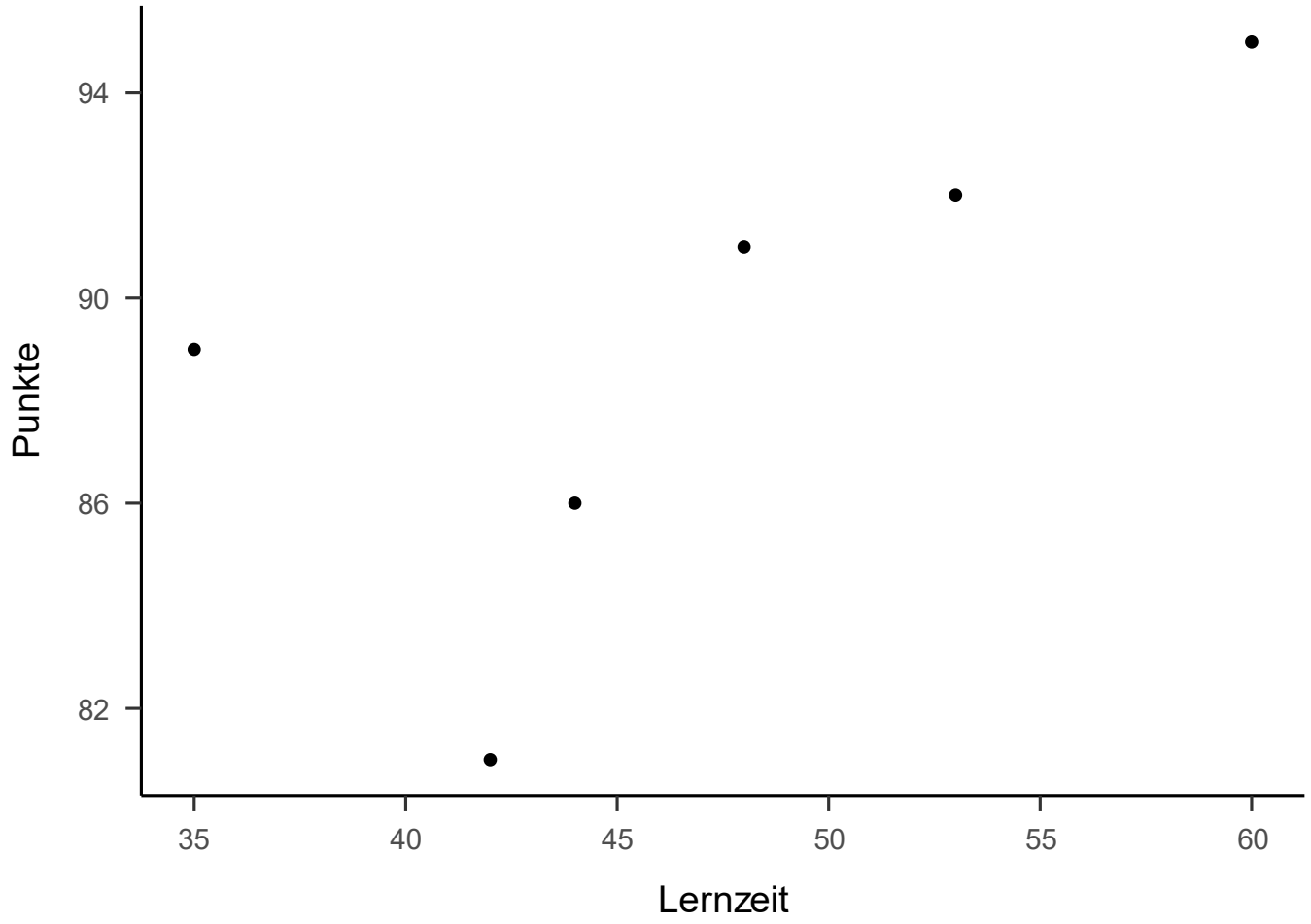
- **Regressionsgleichung:**
  - Abhängige Variable Y wird von unabhängiger Variable X vorhergesagt.

$$\hat{y}_i = a + bx_i$$

  - Beobachtete Werte: $y_i = a + bx_i + e_i$
  - Senkrechte Abweichungen (Residuen): $e_i = y_i - \hat{y}_i$
  - b: Steigung $\quad b = \frac{s_{XY}}{s_X^2} = \frac{s_{XY} \cdot s_Y}{s_X \cdot s_X \cdot s_Y} = r_{XY} \cdot \frac{s_Y}{s_X}$
  - a: y-Achsenabschnitt $\quad a = \bar{y} - b\bar{x}$

Lineare Regression

- **Beispiel:**
  - Die Punktezahl in einer Statistik-Prüfung soll aufgrund der Zeit (in Stunden) vorhergesagt werden, die die Studierenden mit Lernen verbracht haben.

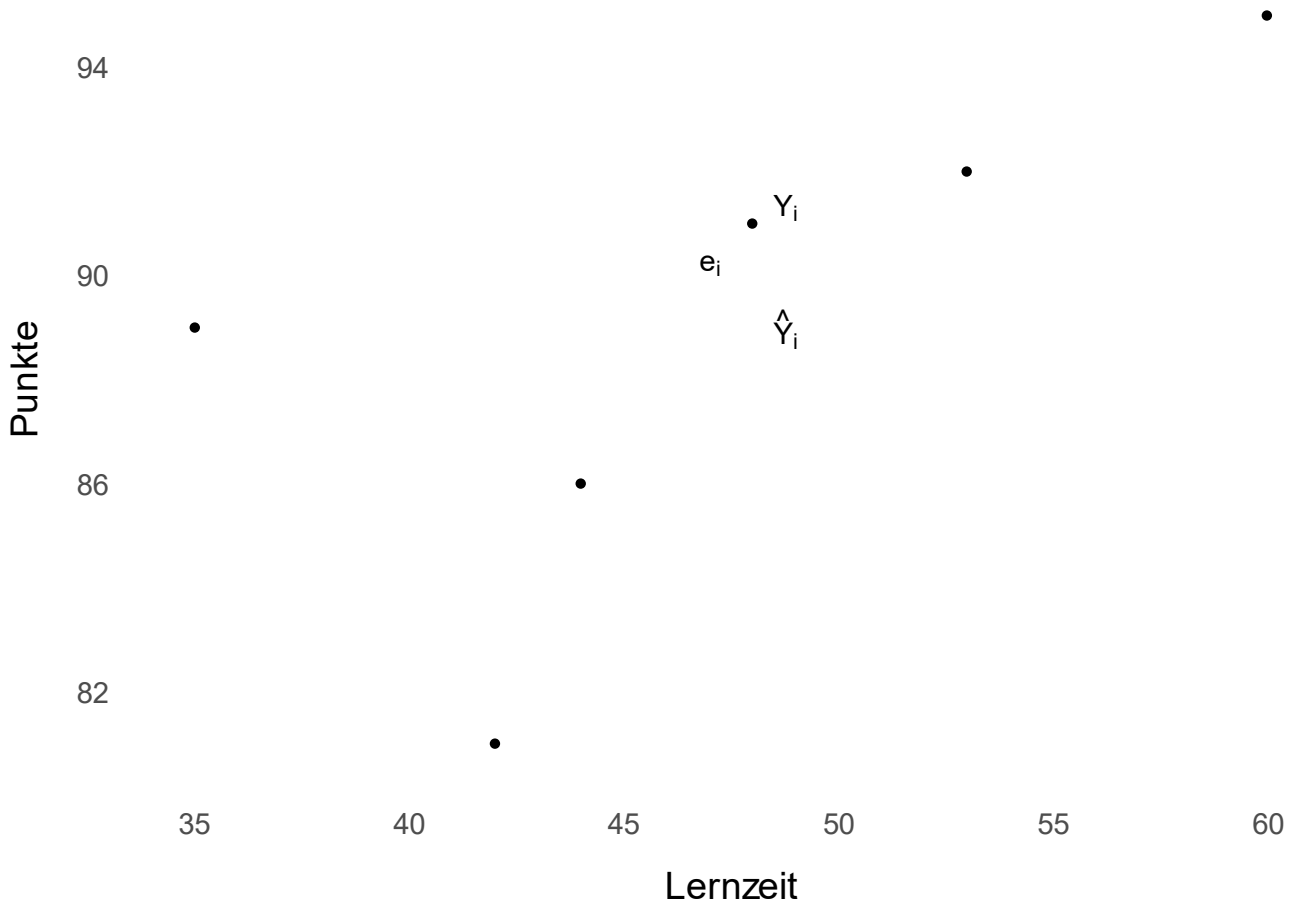


Lineare Regression

- **Beispiel:**
  - Die Punktezahl in einer Statistik-Prüfung soll aufgrund der Zeit (in Stunden) vorhergesagt werden, die die Studierenden mit Lernen verbracht haben.

- Bestimmung der Regressionsgeraden:
  - Methode der kleinsten Quadrate (OLS)
  - Minimierung der Fehlerfunktion:

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$



Lineare Regression

$$\widehat{Punkte} = 71.01 + 0.38 \cdot Lernzeit$$

- **Beispiel:**
  - Die Punktezahl in einer Statistik-Prüfung soll aufgrund der Zeit (in Stunden) vorhergesagt werden, die die Studierenden mit Lernen verbracht haben.

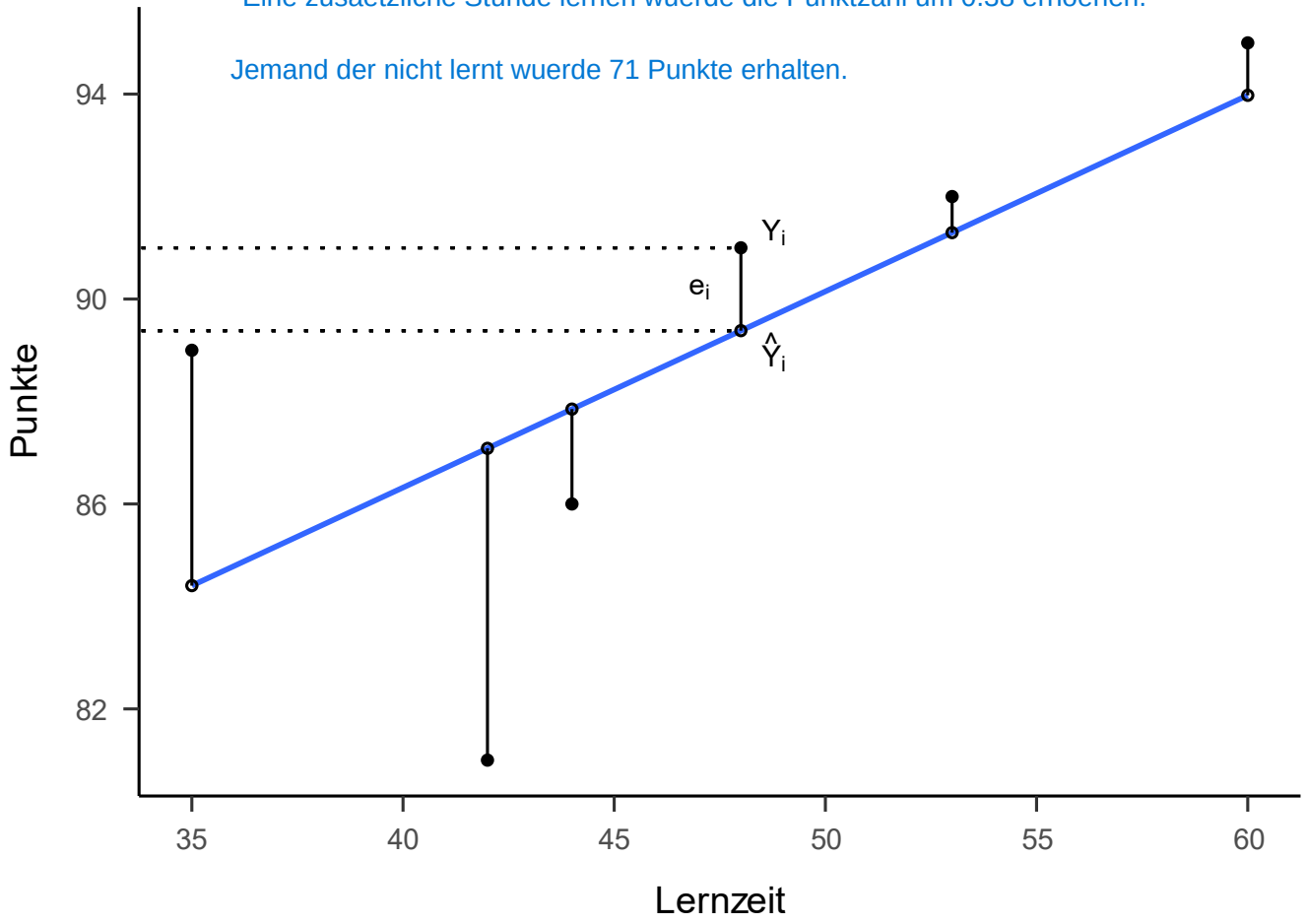
- Was bedeuten die Regressionskoeffizienten?

71.01?

0.38?

Was bedeutet was?
0.71 ist die Punkte wo es beginnt
Eine zusaetzliche Stunde lernen wuerde die Punktzahl um 0.38 erhoehen.

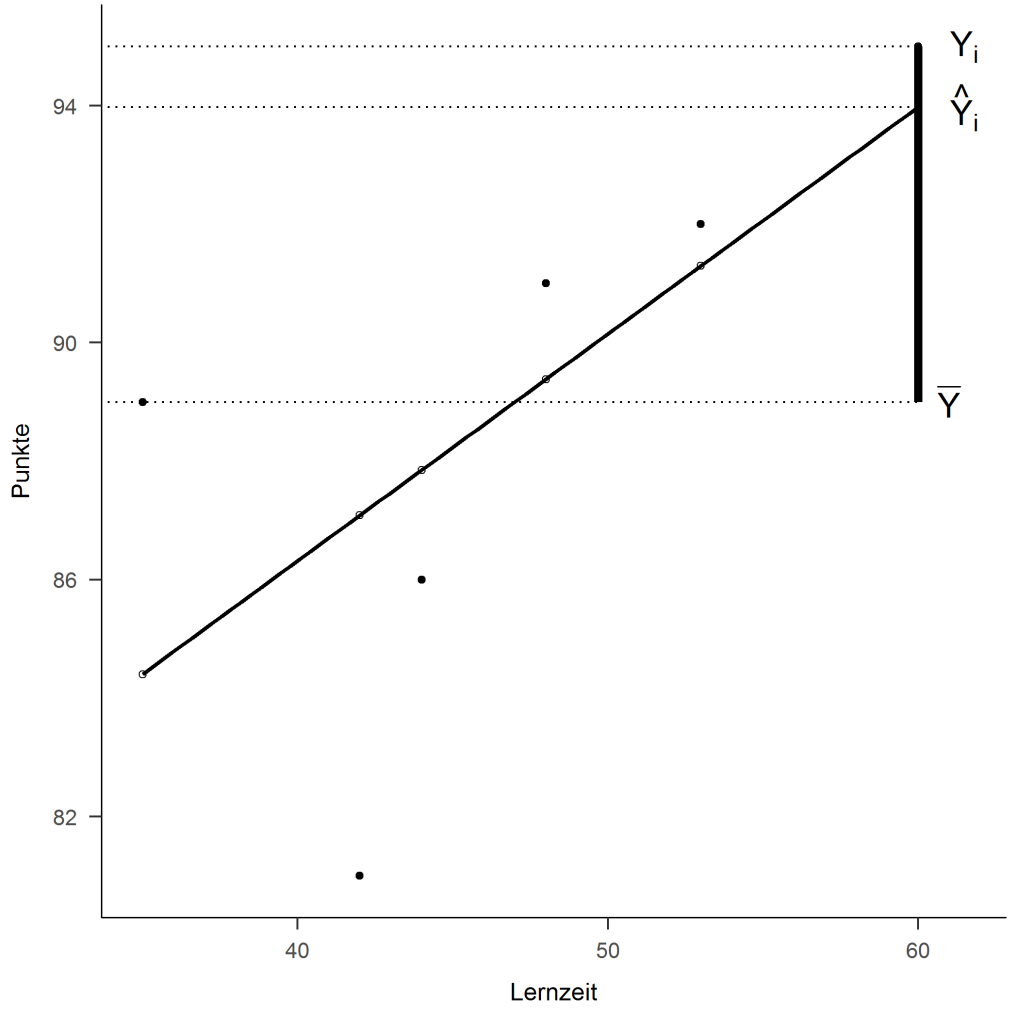
Jemand der nicht lernt wuerde 71 Punkte erhalten.



Lineare Regression

- Bestimmtheitsmass (erklärte Varianz)
  R^2

Erklaerte Varianz in der abhaengigkeit.

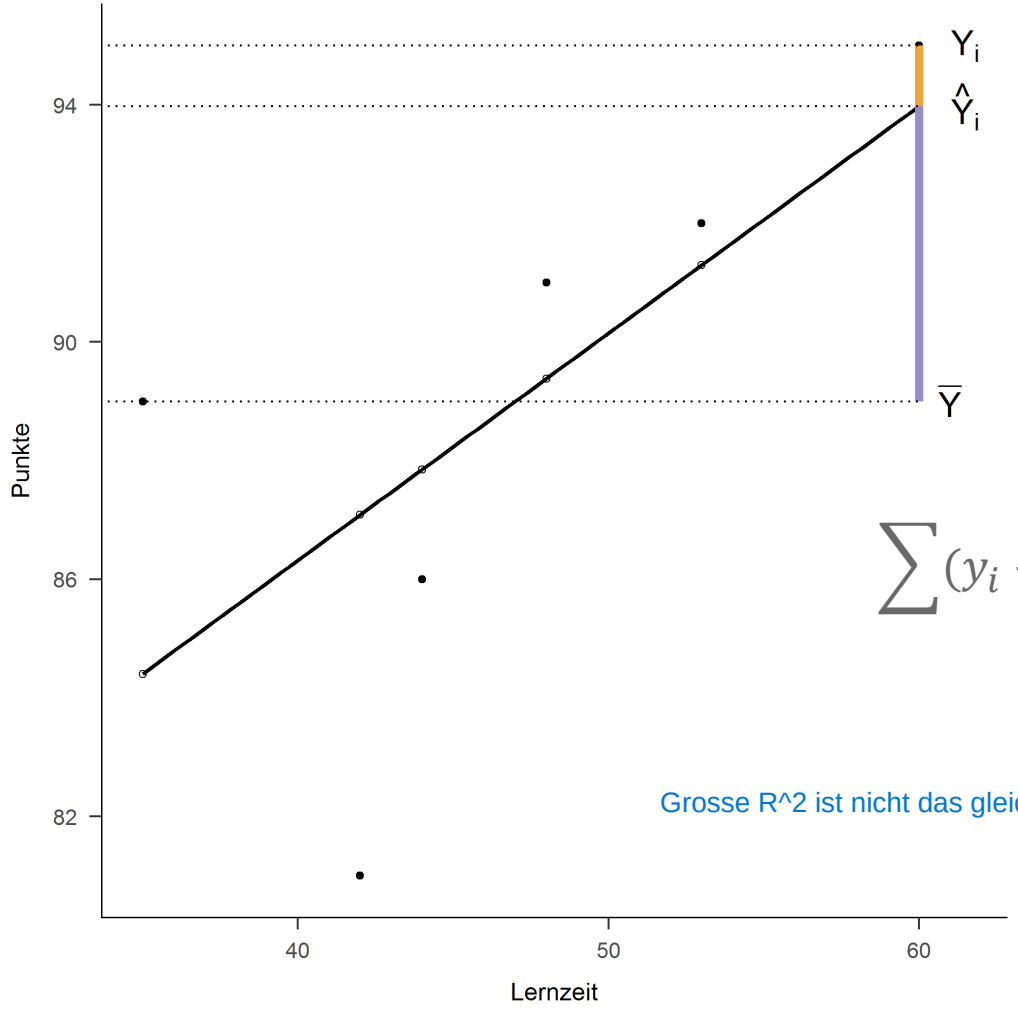


Das waere der Fehler ohne predictior.

Man kann sagen je mehr man lernt desto besser ist man.
Die Grafik sagt aus wieviel die Lernzeit an der Punkte ausmacht.

Lineare Regression

- Bestimmtheitsmass (erklärte Varianz) R^2



Orange=Tatsaechliche Abweichung vom vorhergesagten. Fehler.

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Grosse R^2 ist nicht das gleiche wie R

$$R^2 = \frac{s^2_{\hat{Y}}}{s^2_Y}$$

Lineare Regression

- **Voraussetzungen** für die Interpretation der Resultate der linearen Regression
  - Metrische Daten
  - Linearer Zusammenhang — **L**inearity
  - Unabhängige Messungen — **I**ndependency of errors
  - Normalverteilte Residuen — **N**ormality of residuals
  - Homoskedastizität — **E**quality of variance

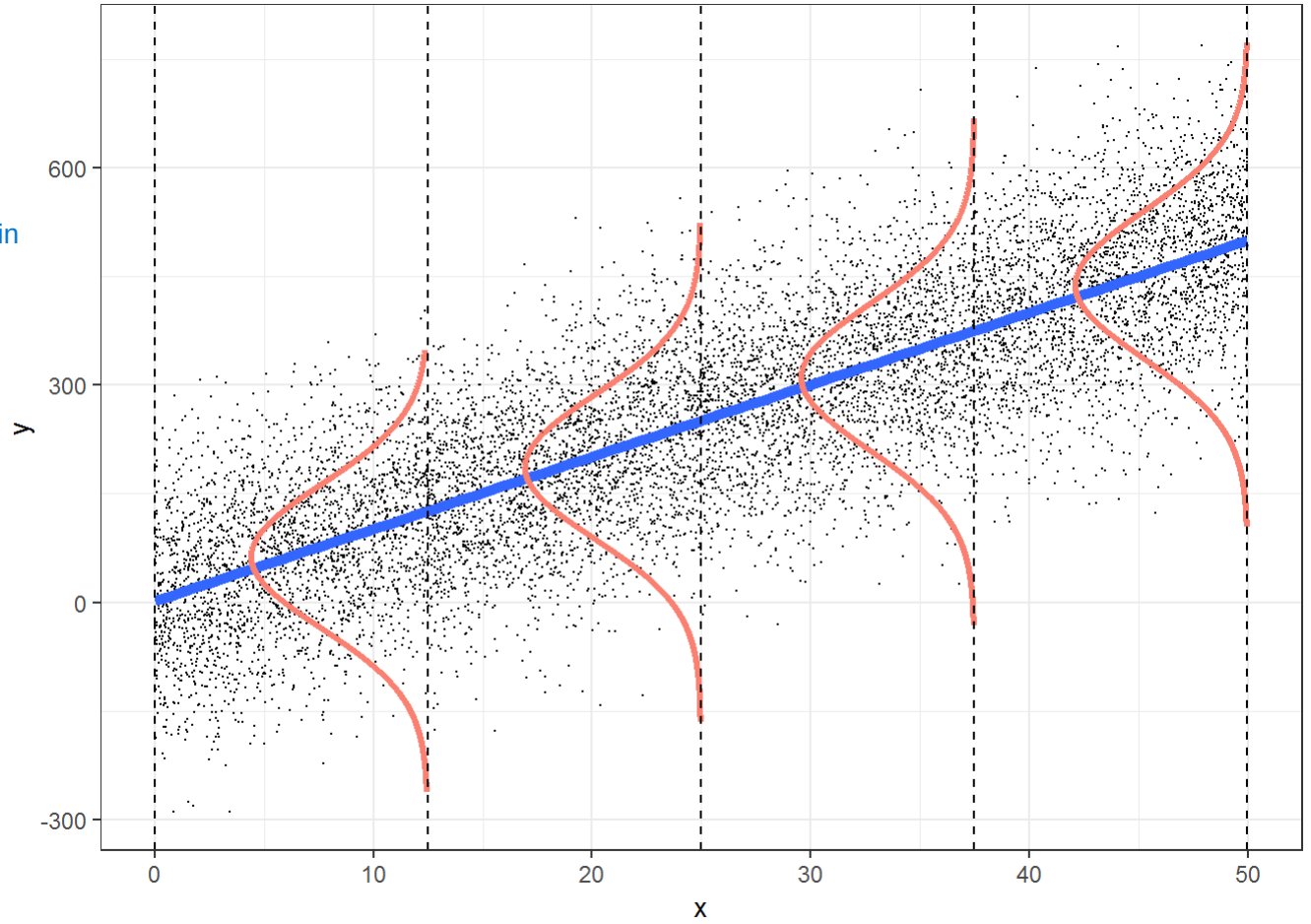
Können mit diagnostischen Plots überprüft werden (machen wir erst später in diesem Semester)

Wann sind Daten nicht mehr unabhaengig, wenn zwei messungen von der gleichen Person bei gleichen Temperaturen gemacht wurden.

Lineare Regression

Visualisierung
«Idealfall»

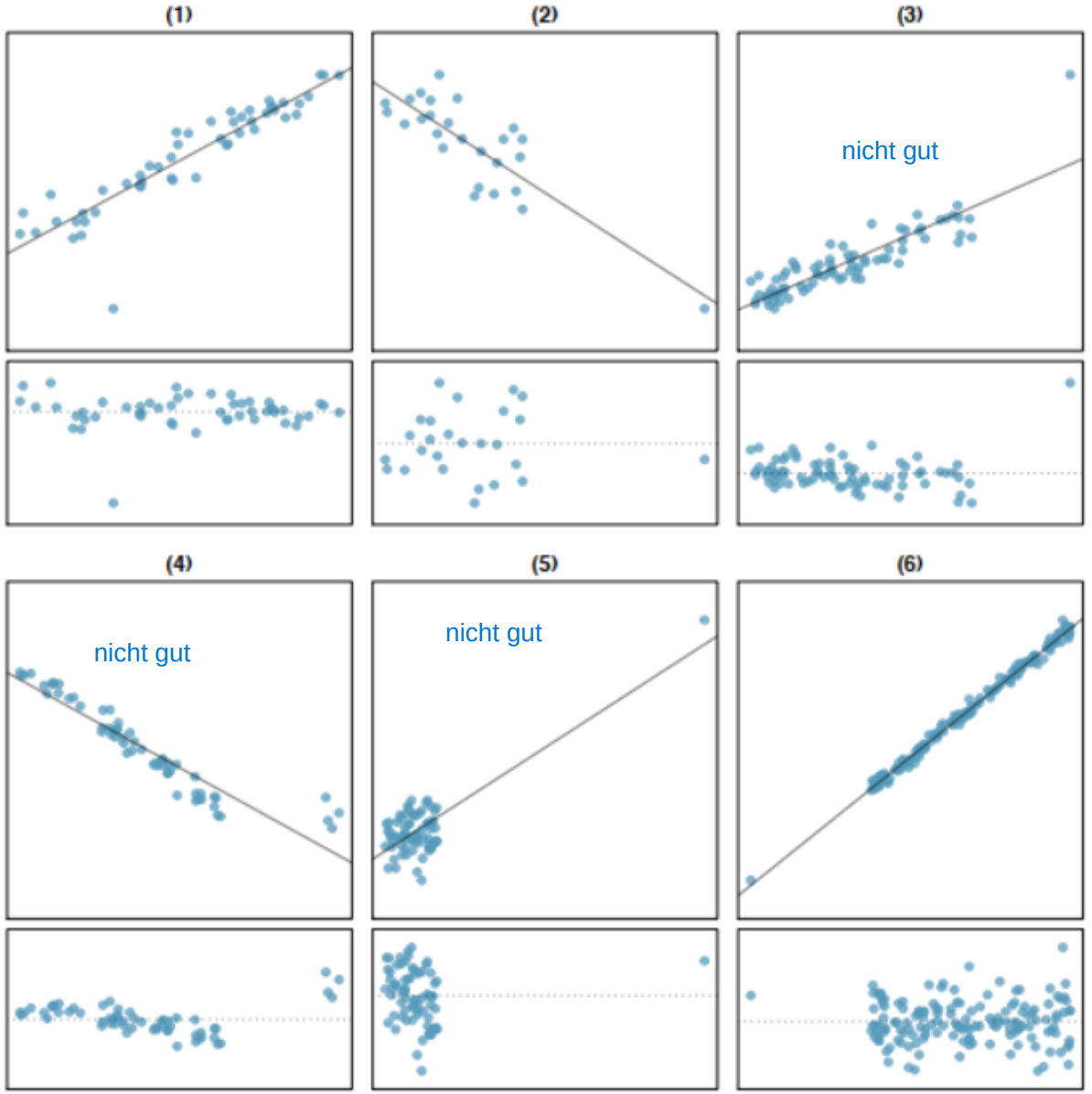
streuung sollte gleich sein



Lineare Regression

- Arten von Ausreissern (Outlier)

- Diskussion
  - Welche Ausreisser sind bedeutsam?
  - Darf man Ausreisser einfach so aus einem Datensatz ausschliessen?



OpenIntro Statistics, S. 329

FFHS
Fernfachhochschule Schweiz
Mitglied der SUPSI

Anwendung in R

Starwars

- Im Datensatz «starwars.csv» finden Sie die Daten einer Umfrage unter 1186 Personen zu ihren Erfahrungen und Einstellungen bezüglich Star Wars - Filmen. Wir möchten explorativ untersuchen, ob sich die Altersgruppen (Variable: *Age*) bezüglich dem Gefallen an Star Wars (Variable: *Do.you.consider.yourself.to.be.a.fan.of.the.Star.Wars.film.franchise*) unterscheiden.

  - Importieren Sie den Datensatz in R
  - Erstellen Sie eine geeignete Kontingenztabelle und grafische Darstellung, um ihre Fragestellung zu untersuchen.

Bildquelle: Chesnot / Getty Images

Datenquelle: https://fivethirtyeight.com/features/americas-favorite-star-wars-movies-and-least-favorite-characters/

Mario Kart

- Der Datensatz «mariokart» des openintro-Packages enthält die Daten von 141 Ebay-Auktionen des beliebten Nintendo-Spieles Mariokart.
- Untersuchen Sie mit einer explorativen Datenanalyse und linearen Regression, ob mittels des Startpreises (Variable start_pr) eine Voraussage des finalen Verkaufspreises (total_pr) möglich ist. Identifizieren Sie dabei allfällige Probleme des linearen Modelles und diskutieren Sie Lösungsvorschläge (bzw. alternative Modelle).