

Statistik zur Datenanalyse

Dr. Meike Wocken

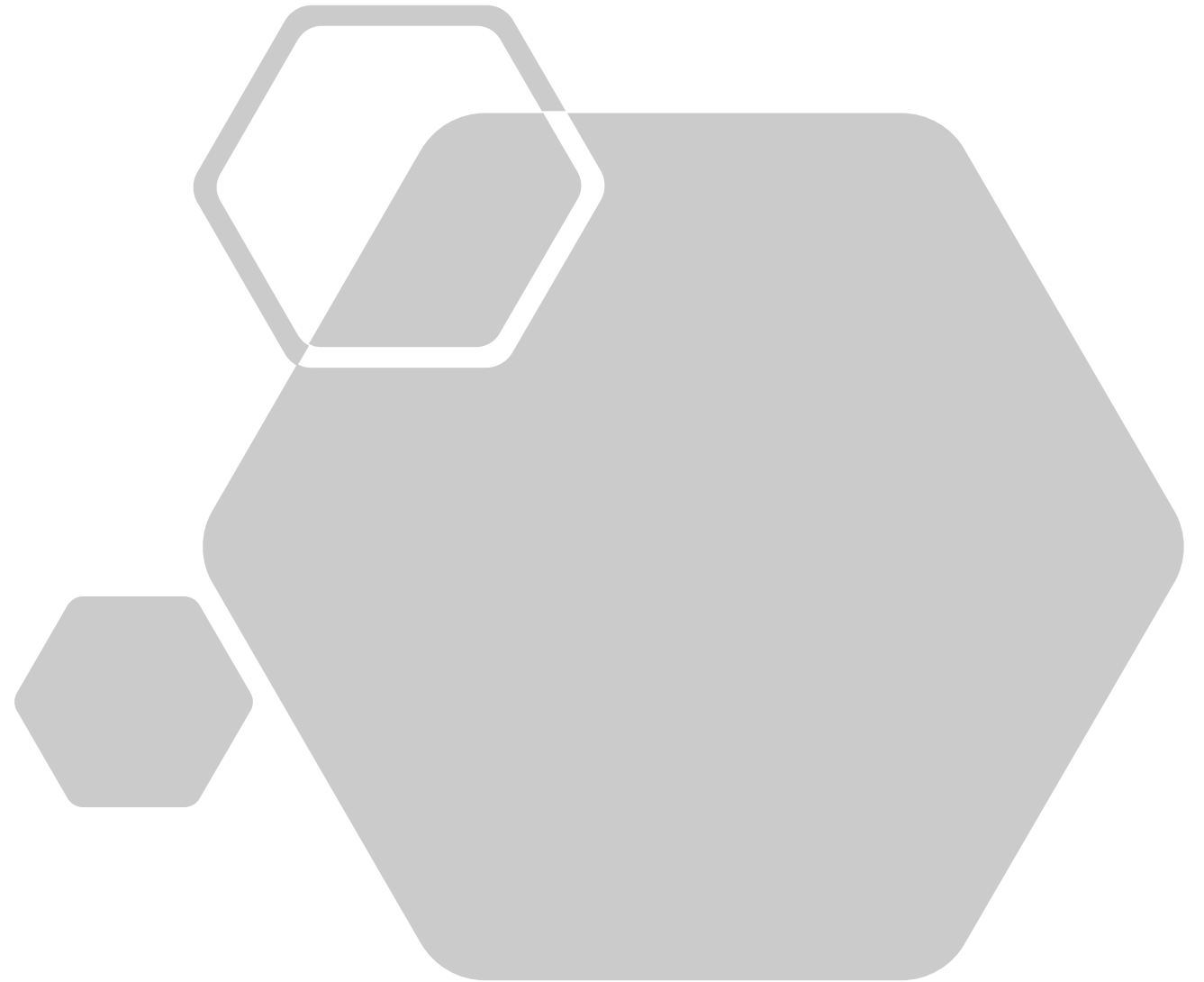
Vorlesung 2

HS Bielefeld

Digitale Technologien (M.Sc.)

WiSe 2023/24

Meike.Wocken@codecentric.de



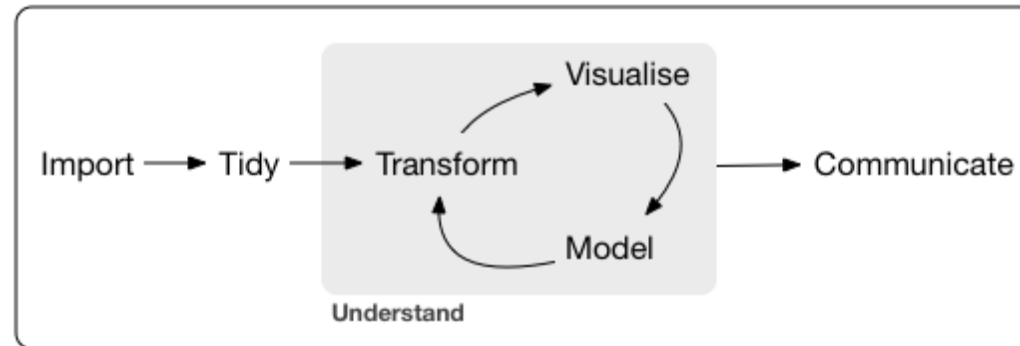
Warum Wahrscheinlichkeiten? Sprachmodelle!

S = Where are we going

- „Where are we“ ...previous Words, Context
- „Going“ ...word being predicted

$P(S) = P(\text{Where}) P(\text{are} | \text{Where}) P(\text{we} | \text{Where are}) P(\text{going} | \text{Where are we})$

Tidyverse



Am Anfang:

Data Cleaning: Bereinigung, Fehlwerte (NA), Ausreißer,

Fehlerhafte Messungen erkennen (z.B. „Trends“ in den Sensoren durch Alterung oder Verschmutzung)

GRAFISCHE EXPLORATION

Codierung, Dummy-Variablen

Open Data - <http://casas.wsu.edu/datasets/>

Drei Bedingungen müssen Daten erfüllen, damit wir von Open Data sprechen:

- **Verfügbarkeit und Zugriffsmöglichkeit:**

Die Daten stehen frei zur Verfügung, bevorzugt als Download aus dem Internet. Es darf maximal eine gut zu begründende Bearbeitungsgebühr erhoben werden. Die Daten müssen in einem zweckmäßigen und bearbeitbaren Format vorliegen.

- **Verarbeitung und Weitergabe der Daten:**

Open Data werden unter Bedingungen veröffentlicht, die eine Verarbeitung, Weitergabe und Zusammenführung mit anderen Datensätzen erlauben. Die Daten müssen maschinenlesbar sein.

- **Universelle Partizipation:**

Open Data stützt sich auf die Teilnahme aller: jede und jeder muss die Daten Nutzen, Verarbeiten und Weitergeben dürfen. Es darf keine Diskriminierung geben gegenüber relevanten Anwendungsbereichen, Personen oder Gruppen. Es darf keine Einschränkungen der Verwendung der Daten geben (z.B. eingeschränkt auf nur nicht-kommerzielle Anwendungen).

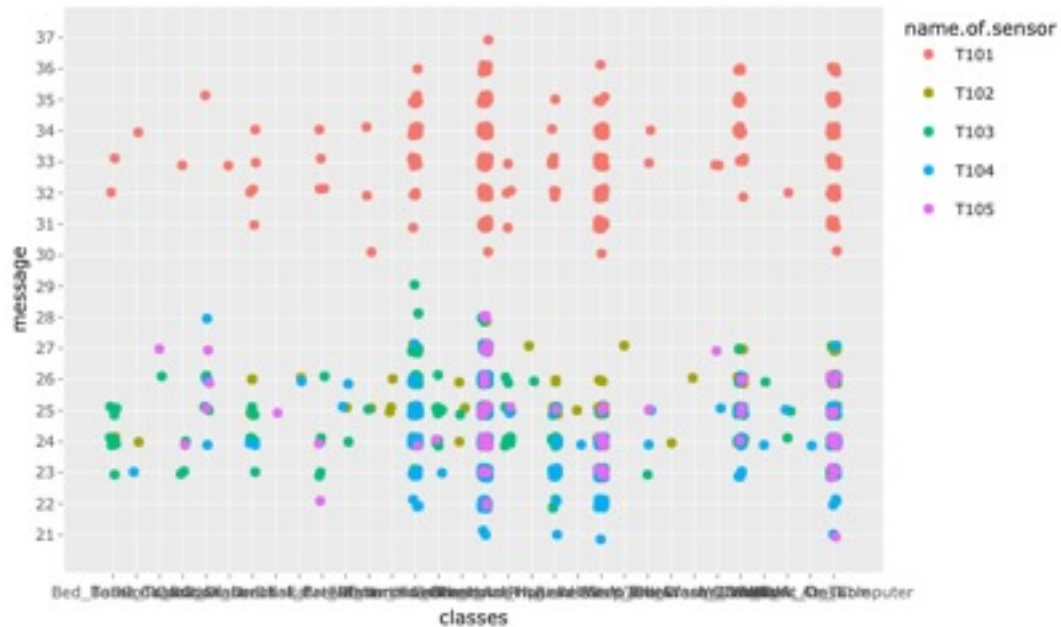
Blogbeitrag: <https://foundersfoundation.de/content-library/potenziale-open-data/>

Dataset 33

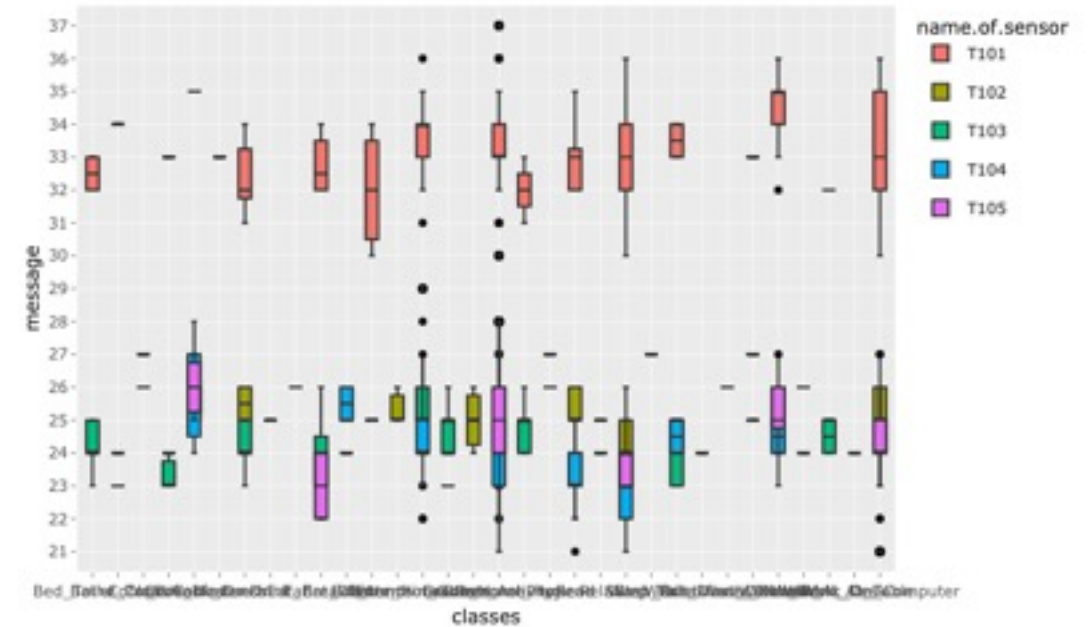
<http://casas.wsu.edu/datasets/hh111.zip>

- Datenset runterladen, Readme lesen
- Einlesen der annotierten Daten hh111.ann.txt
- Welche Sensor-Typen liegen vor?
- Wie viele Beobachtungen haben Sie von welchen Einzel-Sensoren?
- Gibt es Fehlwerte in dem Datensatz?
- Grafische Betrachtung der Temperatursensoren. Was fällt auf?

Streudiagramme, Boxplots, ...

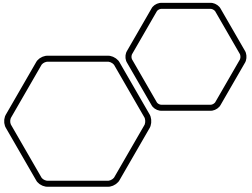


Boxplot: Min, Lower Fence, Q1, Median, Q3, Upper Fence, Max
Erstes Quartil = $Q1 = 25\%$ Quantil = 25% der Beobachtungen
sind kleiner gleich $Q1$
Zweites Quartil = $Q2 = \text{Median} = 50\%$ Quantil
Drittes Quartil = $Q3 = 75\%$ Quantil



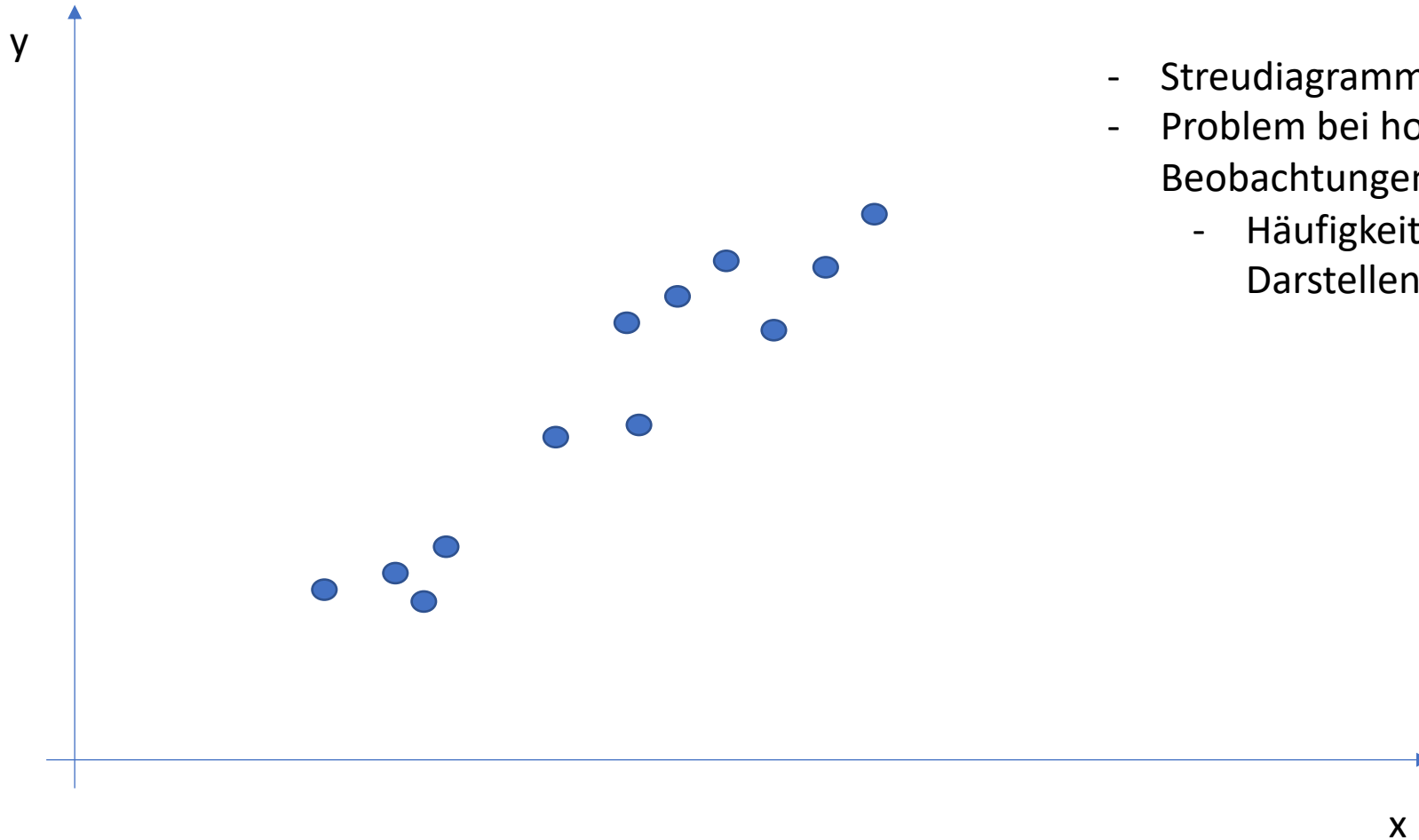
Interquartilsabstand = $Q3 - Q1$ hier liegen 50% der Beobachtungen

Lower/Upper Fence: $\text{Median} \pm 1,5 \cdot \text{Interquartilsabstand}$, alles darunter/darüber potentielle Ausreißer



Multivariate Deskription und Exploration

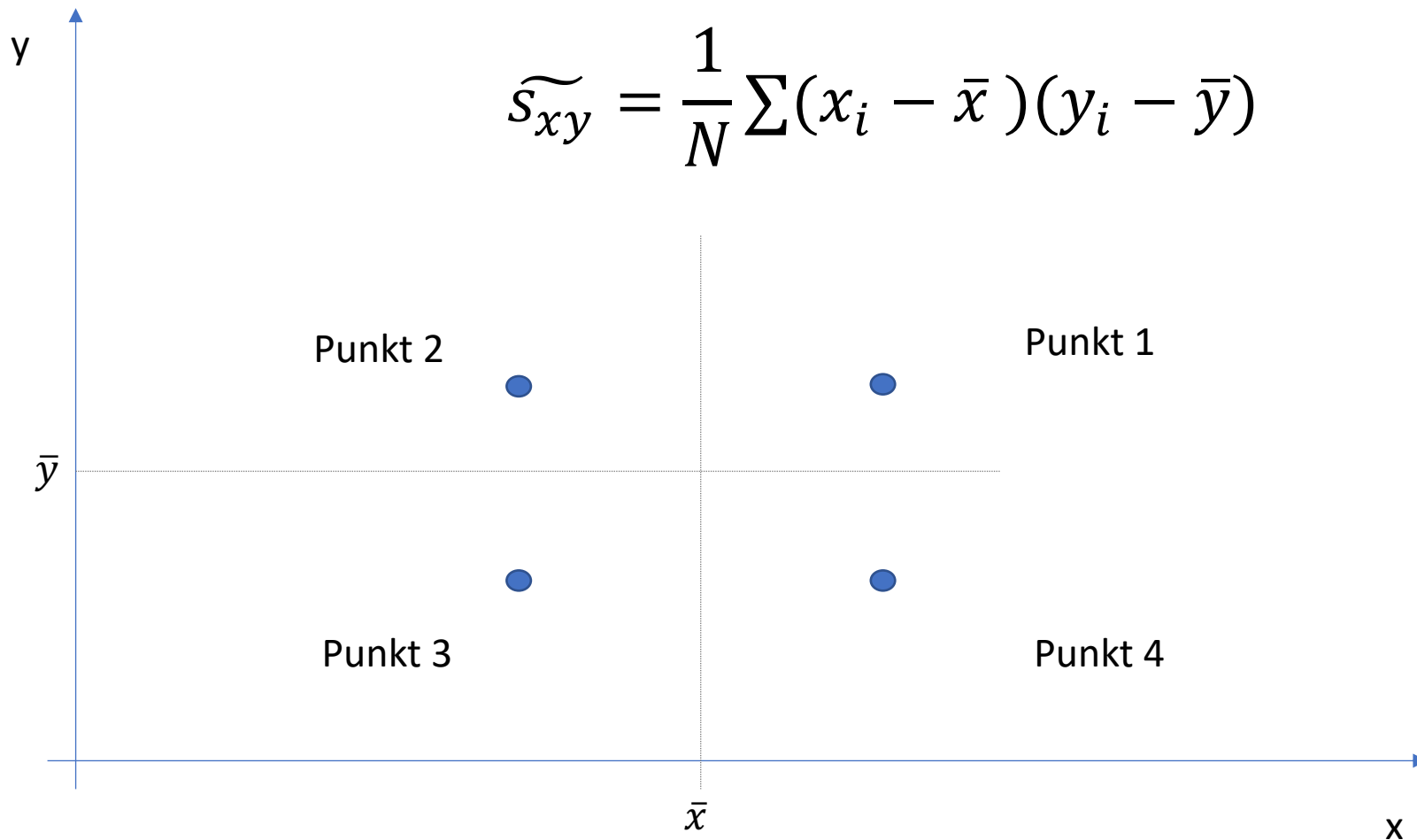
Darstellung von Beziehungen



- Streudiagramm, Matrix an Streudiagrammen
- Problem bei hoher Anzahl an Beobachtungen? Überlappungen!!
 - Häufigkeiten über Farbe oder Größe Darstellen, jitter, geom_hex,

Quantifizierung von Beziehungen

Empirische Kovarianz = Kennzahl **GEMEINSAME Streuung**



Quantifizierung von Beziehungen

Punktwolken im Bereich Punkt 1 und Punkt 3 haben positiven Zusammenhang:

$$\widetilde{s}_{xy} > 0$$

Punktwolken im Bereich Punkt 2 und Punkt 4 haben negativen Zusammenhang:

$$\widetilde{s}_{xy} < 0$$

Quantifizierung von Beziehungen

Korrelationskoeffizient

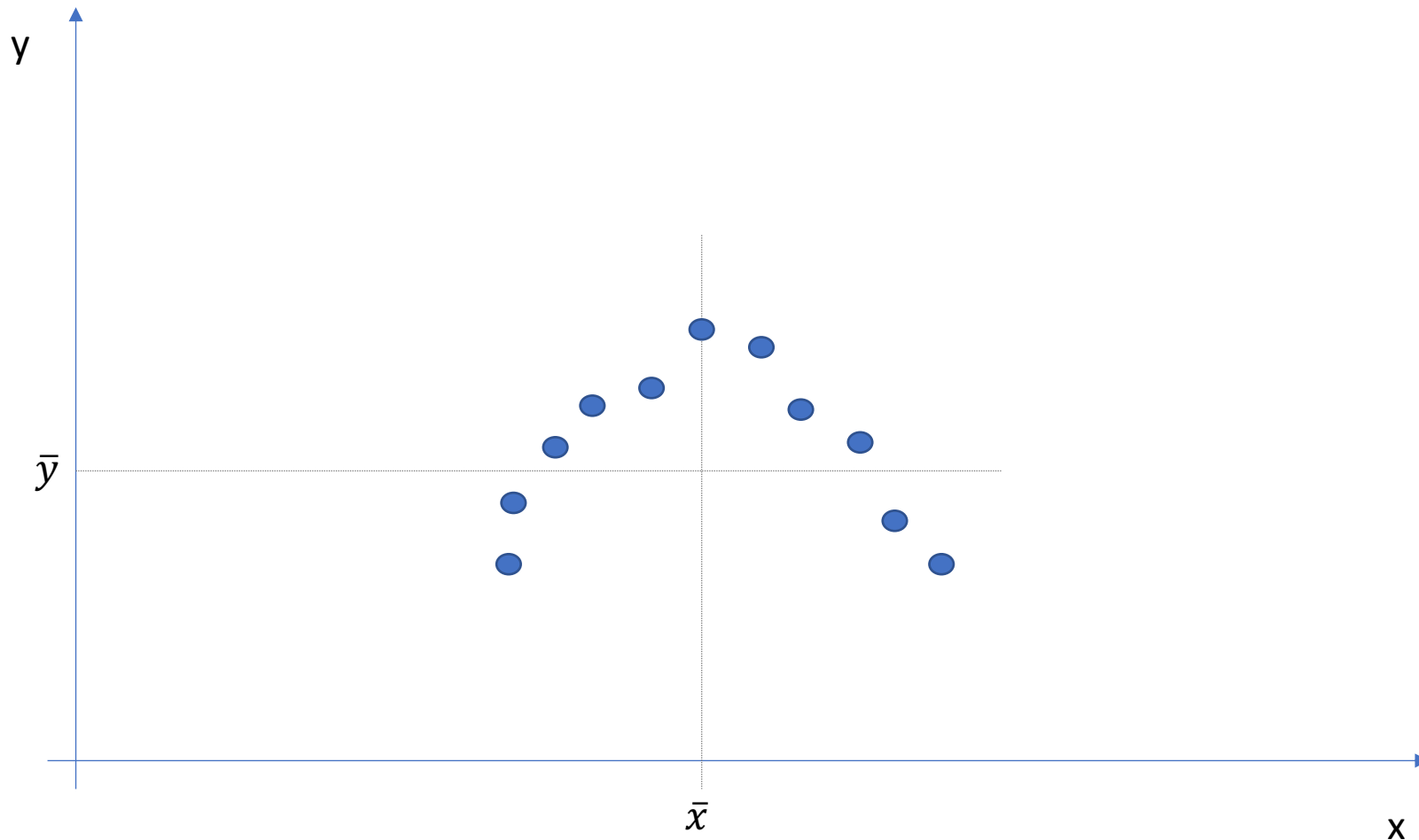
$$r = \frac{\widetilde{s}_{xy}}{\widetilde{s}_x \widetilde{s}_y}, \quad -1 \leq r \leq 1$$

Mit empirischer Standardabweichung

$$\widetilde{s}_x = \sqrt{\frac{1}{N} \sum (x_i - \bar{x})^2} \quad \widetilde{s}_y = \sqrt{\frac{1}{N} \sum (y_i - \bar{y})^2}$$

WICHTIG: es wird hier nur der lineare Zusammenhang gemessen!!

Quadratische Beziehung: $r \approx 0$
trotzdem liegt eine Beziehung in den Daten vor



(Pearsons) Korrelationskoeffizient

- Metrische Werte
- Was passiert bei ORDINALER Skala? Spearmans Korrelationskoeffizient! Es wird der Rang der Beobachtungen verwendet zur Berechnung des Korrelationskoeffizienten
- Beispiel: Bewertung auf 5-stufiger Skala und Alter der Kundinnen
- Spearmans Korrelationskoeffizient misst damit die Stärke eines MONOTONEN Zusammenhangs (nicht mehr zwangsweise linear!)

Korrelation und Kausalität: Theoretische Überlegungen notwendig!!

WICHTIG:

wie ist der Zusammenhang, in welche Richtung geht die Beeinflussung:
von x zu y oder y zu x?

Probleme:

- Scheinkorrelation
- Verdeckte Korrelation

Zusammenhang in den Daten

Bis jetzt: Quantifizierung des LINEAREN Zusammenhangs mittels Korrelationskoeffizienten

Theoretische Überlegung führt zur Richtung der Beeinflussung:

$$Y = f(X) + \epsilon$$

f **funktionaler Zusammenhang** ist unbekannt, f ist die systematische Information, die uns X über Y liefert. ϵ ist **Störgröße** in dem Zusammenhang.

Wozu brauchen wir f ?

- Prognose (exakte Prognosen)
- Inferenz (exakte Funktionsform)

Wie erhalten wir f ?

1. Modellgetrieben

1. Annahmen zur Funktionsform
2. Schätzung der Funktionsparameter
3. Funktionsform kann sich vom wahren f unterscheiden

Reale Zusammenhänge sind i.d.R. nicht global, sondern oft nur lokal linear:

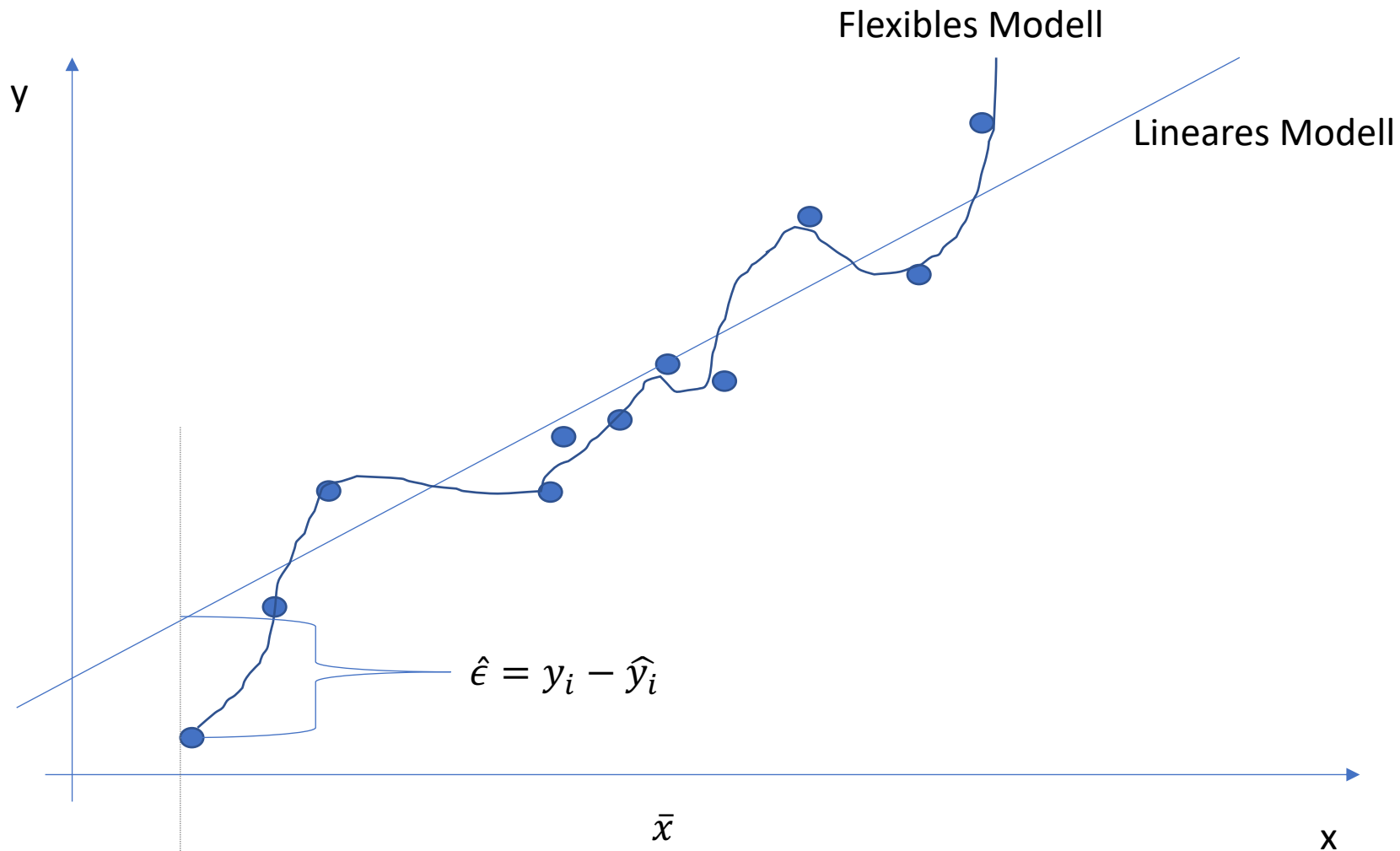
Lineare Modelle liefern oft keine exakten Prognose, sind aber gut zu verstehen und zu interpretieren.

Wie erhalten wir f ?

2. Datengetrieben

1. Keine Annahme über f
2. Große Anzahl an Beobachtungen notwendig
3. f wird nahe an den Daten geschätzt (komplexe Verfahren)
4. Ergebnis i.d.R. schwer zu interpretieren, z.B. neuronale Netze

Wie erhalten wir f ?



Wie erhalten wir f ?

Overfitting

Im Training ist der Fehler ≈ 0

Auf unbekannten Testdaten großer Fehler und schlechte Performance

Erkennen von Overfitting: Training- und Testdaten verwenden

Beispiel Einfache Regression

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \forall i = 1, \dots, n$$

Schätzung Parameter: Ziel ist globales Minimum der Abweichungen Y_i und \hat{Y}_i

Güte des Modells

Wir kennen das wahre f nicht, aber wie können wir die Güte trotzdem beurteilen?

Residuen als Schätzer der Abweichungen!!

$$RSE = \sqrt{\frac{1}{N - (p + 1)} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2}$$

RSE Residual Standard Error = Standardfehler der Regression

Residuum $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ ist die geschätzte Störgröße

RSE ist ein absolutes Maß (wird in Einheit von Y gemessen) der Passgenauigkeit zwischen Modell und Daten

Güte des Modells

MSE Mean Squared

Gemessen in quadratischer Einheit von Y

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

Streuungszerlegung

SQT = Summe der Quadrate Total = SST = Sum Squared Total

SQE = Summe der Quadrate Erklärt = SSE = Sum Squared Explained

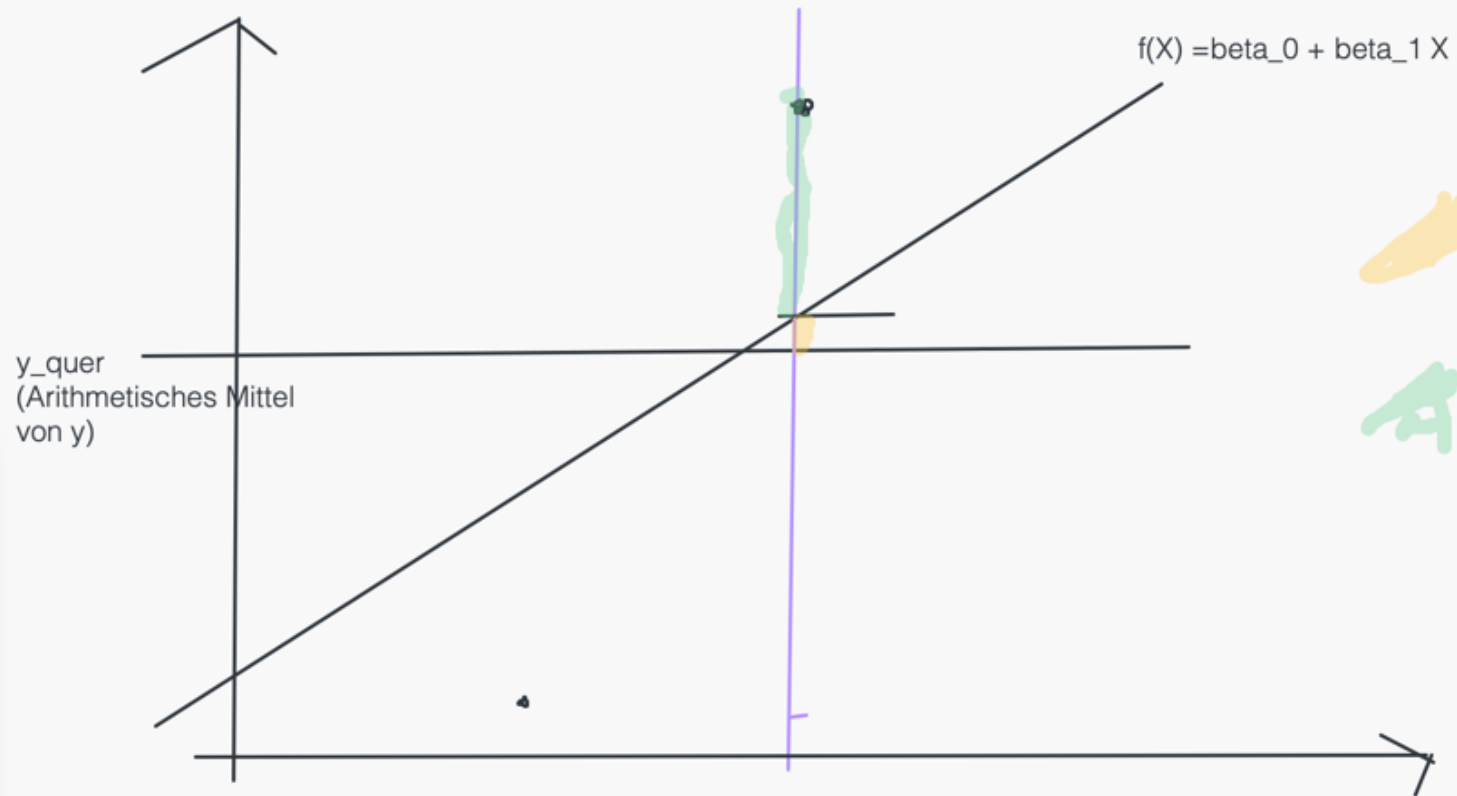
SQR = Summe der Quadrate der Residuen = SSR = Sum Squared Residuals

$$\begin{array}{ccccccc} \text{SQT} & & = & & \text{SQE} & & + & & \text{SQR} \\ \Sigma(Y_i - \bar{Y})^2 & & = & & \Sigma(\hat{Y}_i - \bar{Y})^2 & & + & & \Sigma(Y_i - \hat{Y}_i)^2 \end{array}$$

Gesamte Streuung, die
ich erklären möchte

Streuung, die das
Modell erklärt

Rest (= $\hat{\epsilon}_i^2$)



$y_{\text{geschätzt}} - y_{\text{quer}} = \text{durch das Modell erklärt}$

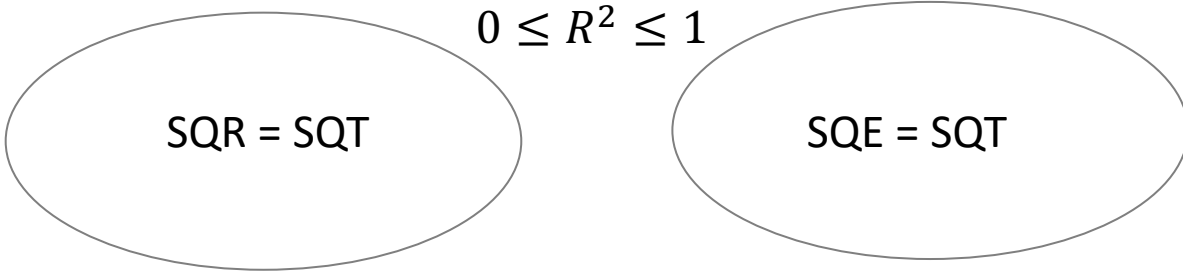
beobachtete y -
geschätztem y =
Residuum

Bestimmtheitsmaß

R^2 gibt den Anteil der Gesamtstreuung der y_i an, der durch das Modell erklärt wird

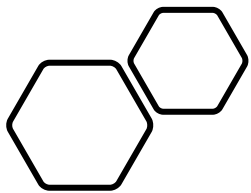
$$R^2 = \frac{SQE}{SQT} = \frac{SQT - SQR}{SQT} = 1 - \frac{SQR}{SQT}$$

$$0 \leq R^2 \leq 1$$


$$SQR = SQT$$

$$SQE = SQT$$

Nur im einfachen linearen Modell (nur eine erklärende Variable x !) gilt
 $R^2 = r_{xy}^2$ (Korrelationskoeffizient)



Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Schätzen der Parameter

$$\min_{\beta_0, \beta_1, \dots, \beta_p} \sum \hat{\epsilon}_i^2$$

Annahmen

1. Linearität & Vollständigkeit
2. Keine Multikollinearität
3. Keine Konstanten zur Erklärung
4. Erwarteter Fehler gleich Null
5. Homoskedastizität
6. Keine Autokorrelation der Störgrößen
7. Normalverteilung der Störgrößen

Aufg. 6 (Skript): Auto

6. Verwenden Sie bitte den gegebenen Datensatz **Auto** aus dem R Paket **ISLR**:

8 Regression

- 104 -

Die vier Datenpunkte gehören zu Aufgabe 5!

x_i	1	2	3	4
y_i	3	2	4	5

- a) Unterteilen Sie den Datensatz bitte in einen Test- und einen Trainingsdatensatz. Schätzen Sie bitte das Modell

$$mpg = \beta_0 + \beta_1 cylinders + \beta_2 displacement + \beta_3 horsepower + \beta_4 weight + \epsilon$$

und berechnen den RMSE jeweils auf Test- und Trainingsdaten.

- b) Stellen Sie die Residuen in den Test- und Trainingsdaten grafisch dar.
- c) Berechnen Sie die Korrelationen zwischen den Variablen und stellen ein Streudiagramm dar.
- d) Schätzen Sie ein verbessertes Modell.

$$RMSE = \sqrt{\frac{1}{n} \sum (\hat{y}_i - y_i)^2}$$

Abweichung zwischen geschätzten Werten und beobachteten Werten.

Zusatzfrage

- Was ist – ceteris paribus - der durchschnittliche Effekt einer Änderung von horsepower um 1 Einheit?

$$E(mpg) = f(\dots, horsepower, \dots)$$

$$\frac{d \widehat{mpg}}{d horsepower} = ?$$

Berechnet den Effekt sowohl für Modell 1 als auch Modell 3!

Aufg. 7 (Skript): Modellvergleich

7. In der Tabelle 8 sind die Ergebnisse von zwei einfachen linearen Regressionen dargestellt. Die Daten und die angepassten Ausgleichsgeraden sind in Abbildung 36 zu sehen. Vergleichen Sie bitte die beiden Modelle.

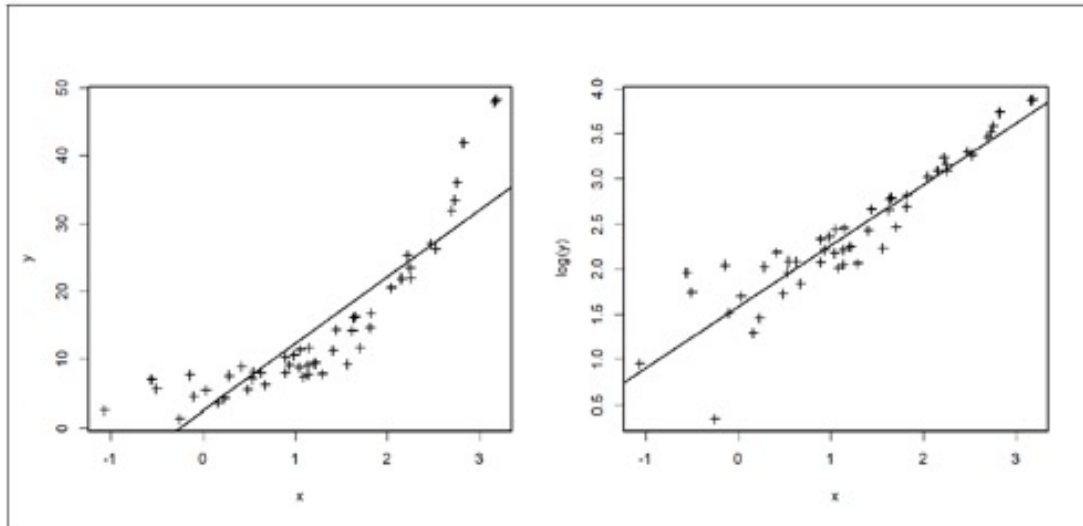


Abbildung 36: Modellvergleich Aufgabe 7.

	Dependent variable:	
	y	log(y)
	(1)	(2)
x	9.858*** (0.769)	0.678*** (0.039)
Constant	2.460* (1.231)	1.581*** (0.062)
Observations	50	50
R ²	0.774	0.863
Adjusted R ²	0.769	0.861
Residual Std. Error (df = 48)	5.497	0.278
F Statistic (df = 1; 48)	164.523***	303.505***

Note: *p<0.1; **p<0.05; ***p<0.01

Tabelle 8: Lineare Regression

Aufgabe 7

- Nur Modelle, die die gleiche ZU ERKLÄRENDE VARIABLE besitzen, können hinsichtlich der Güte miteinander verglichen werden. Es gilt:

$$SQT_y \neq SQT_{\log(y)}$$

- ABER: log-Transformation der zu erklärenden Variablen führt zu LINEARITÄT!

Homoskedastizität und Heteroskedastizität

- Problem bei Heteroskedastizität:
 - Schätzer nicht effizient
 - Standardfehler nicht konsistenz (das ist notwendig für Hypothesentests: Aussagen von statistischen Tests bei Heteroskedastizität nicht verlässlich)
- Lösung:
 - Transformation der Daten oder
 - Robuste Kleinste-Quadrate-Schätzer Methode

Aufg. 8 (Skript): Schätzung des Gehalts

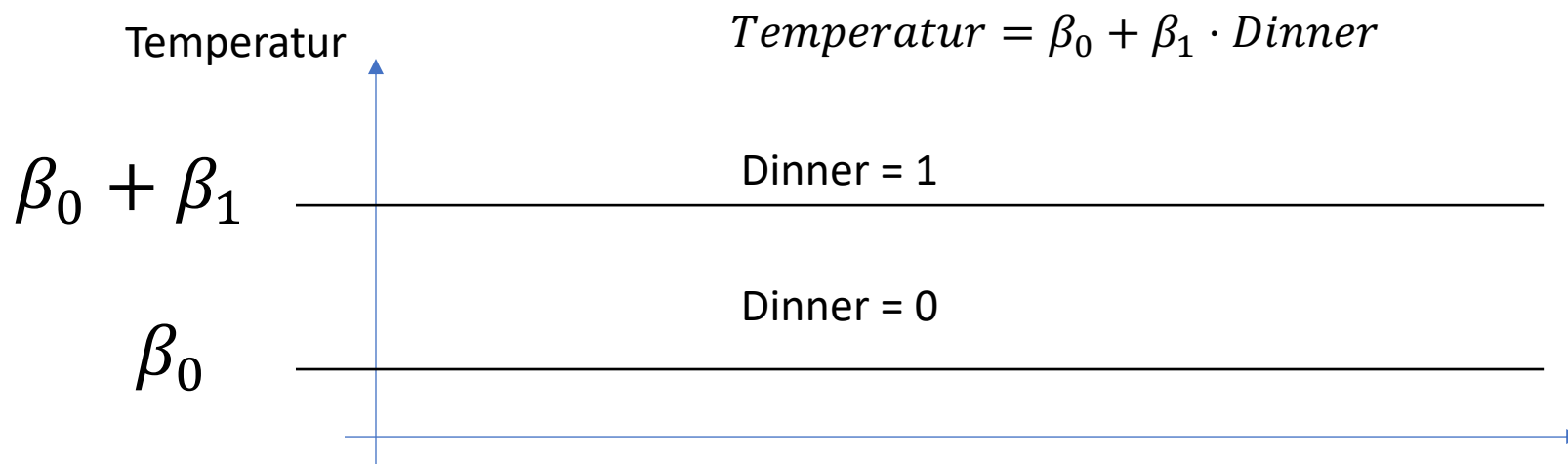
8. Datensatz **Hitter** aus dem **R** Paket **ISLR**: Mithilfe des Datensatzes soll das Gehalt eines Baseball Spielers basierend auf verschiedenen Spieler-Statistiken und der Vorjahres-Leistung geschätzt werden. Führen Sie eine Forward-Selektion zur Auswahl eines finalen Modells durch. Interpretieren Sie das Modellergebnis.

Wie funktioniert der Algorithmus der Forward-Selektion?

$$korr. R^2 = 1 - \frac{N - 1}{N - p - 1} (1 - R^2)$$

Kategoriale erklärende Variablen

- Erklärende Variablen müssen nicht metrisch sein
- Einfaches Beispiel: zwei Klassen miteinander vergleichen.
Dinner= 1 für Aktivität „Dinner“ und 0 für Aktivität „Breakfast“



Mehr als zwei Klassen

$$beste.lage = \begin{cases} 1 & \text{Wohnlage} = 1 \text{ (Beste Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

$$gute.lage = \begin{cases} 1 & \text{Wohnlage} = 2 \text{ (Gute Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

$$normale.lage = \begin{cases} 1 & \text{Wohnlage} = 3 \text{ (Normale Wohnlage)} \\ 0 & \text{sonst} \end{cases}$$

- Bodenrichtwert.csv
- 1=beste Lage, 2= gute Lage, 3= normale Lage
- Für jede Kategorie wird ein separater Effekt geschätzt, dafür wird aber immer eine Dummy Variable weniger ins Modell aufgenommen, als Kategorien insgesamt da sind (sonst perfekte Kollinearität). Eine Kategorie ist Referenzkategorie (im ersten Beispiel Breakfast). Hier also 2 Dummy Variablen im Modell (gute und normale Wohnlage im Vergleich zu beste Wohnlage).
- Schätzergebnisse im Vergleich zur Referenzkategorie zu interpretieren
- R kümmert sich automatisch um das erstellen der Dummy-Variablen