

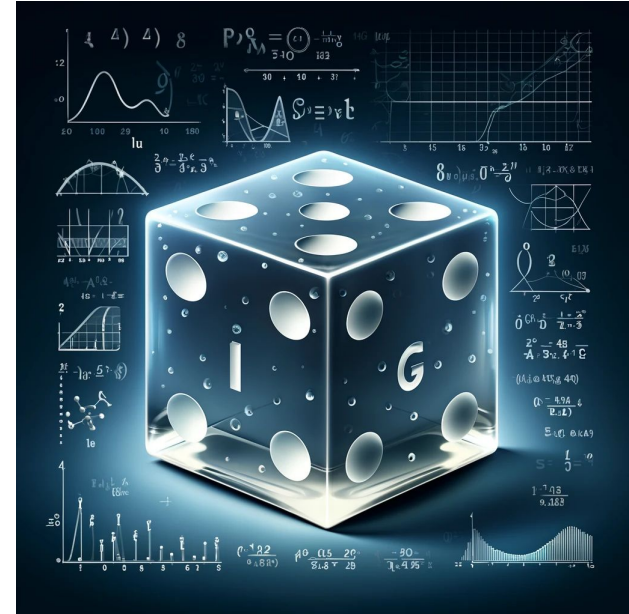


Einführung in die Statistik

Praktische Übung – Jürgen Hermes – IDH – SoSe 2025

Statistik (Stochastik) – Die Grundlagen

- Deskriptive Statistik
 - Mittelwerte
 - Streuungsmaße
 - Korrelation
 - **Regression**
- Inferenzstatistik



Hausaufgabe

- Berechnen Sie die restlichen Pearson-Korrelationen zwischen den Variablen in der Tabelle auf der Folie “Zusammenhänge”. Die entsprechenden Werte finden Sie im `openintro::cia_factbook`
- Visualisieren Sie mindestens die Korrelation zwischen Geburtenrate und Bevölkerungswachstum. In diesem Plot sehen Sie zwei extreme Ausreißer. Ermitteln Sie die zugehörigen Länder dazu. Worauf ist das Verhalten dieser beiden Ausreißer wahrscheinlich zurückzuführen? Können Sie das anhand von anderen im `cia_factbook` vorliegenden Daten bestätigen?
- Wie hoch ist die Spearman-Korrelation zwischen Hubraum (`eng_size`) und PS (horse power, `horsepwr`) im Datensatz `openintro::cars054`? Bilden Sie vor der Berechnung eine Hypothese zur Richtung der Korrelation heraus.

Zusammenhänge

	Geburten- rate	Sterberate	Kinder- sterblichkeit	Lebens- erwartung	Bevölkerungs- wachstum
Geburtenrate	1	0.15	0.87	-0.84	0.56
Sterberate		1	0.36	-0.48	-0.17
Kinder- sterblichkeit			1	-0.89	0.44
Lebens- erwartung				1	-0.35
Bevölkerungs- wachstum					1

Begriffe

Korrelationskoeffizient

Pearson

Spearman

Einseitige Steuerung

Gegenseitige Steuerung

Drittseitige Steuerung

Komplexe Steuerung

Kovariaten

Korrelation & Kausalität

Koinzidenz

bidirektional

Effekt & Rauschen



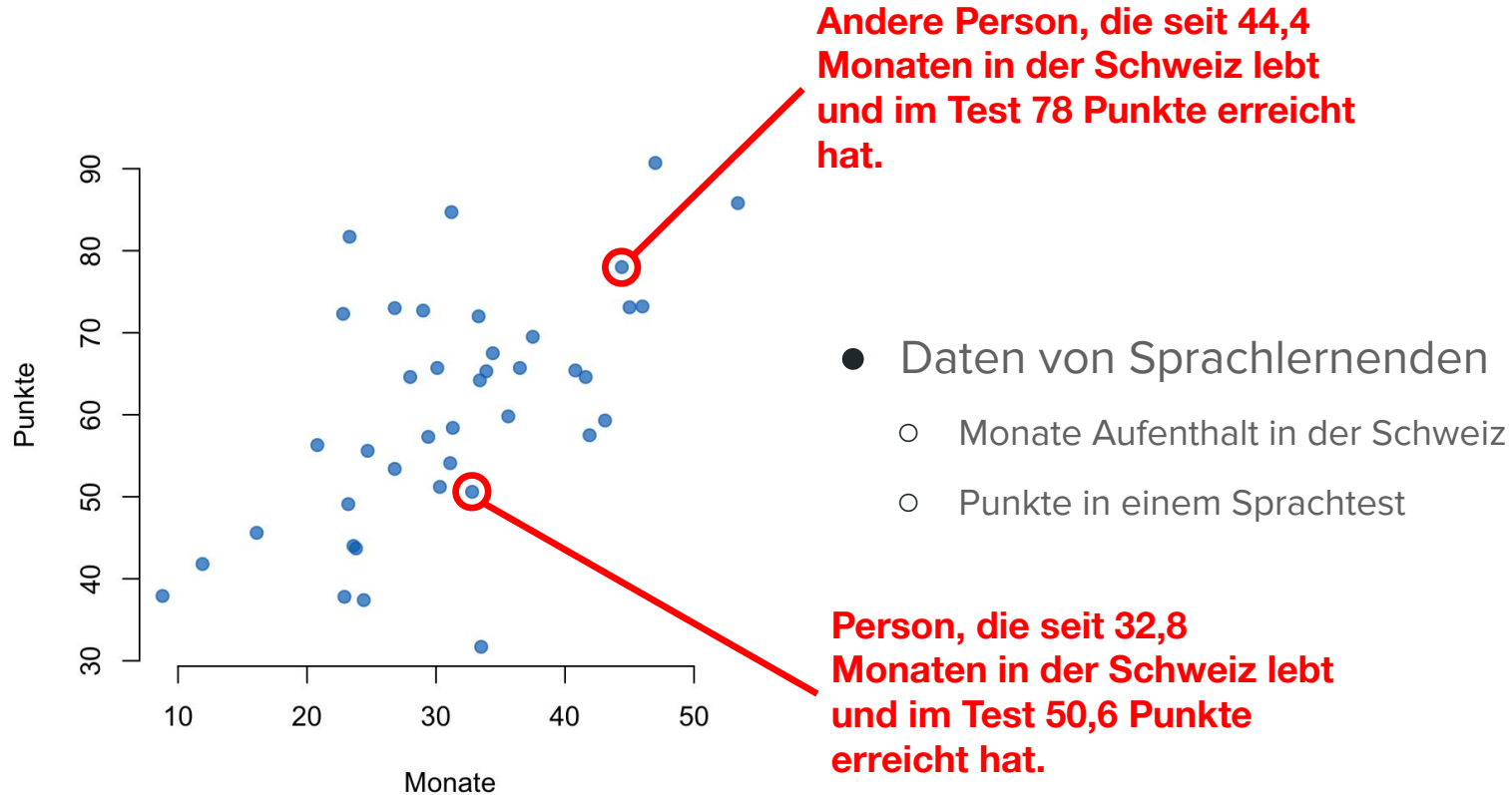


Diese und die folgenden Folien sind erstellt worden von Sascha Wolfer für seinen Kurs "Statistik mit R" an der Uni Basel. Ich nutze sie mit seiner freundlichen Genehmigung. DOI für die Materialien ist

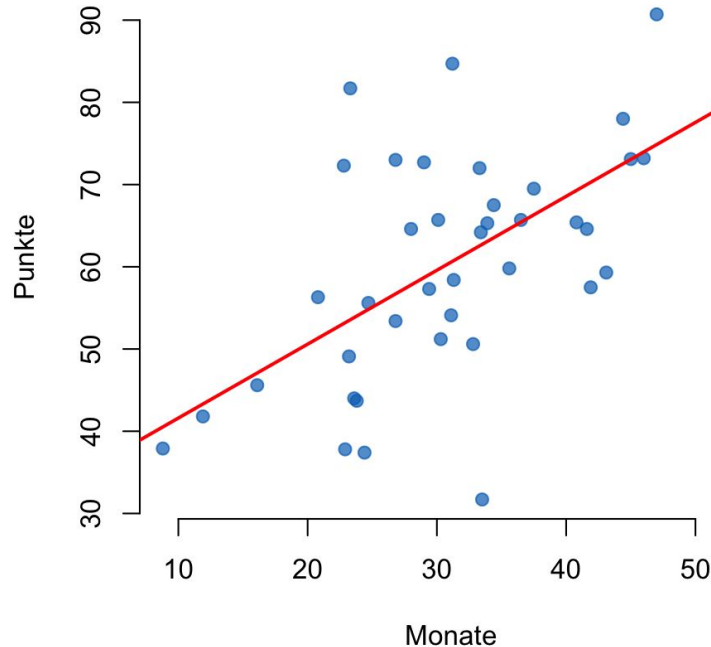
[10.5281/zenodo.7431504](https://doi.org/10.5281/zenodo.7431504)

Regression

Daten



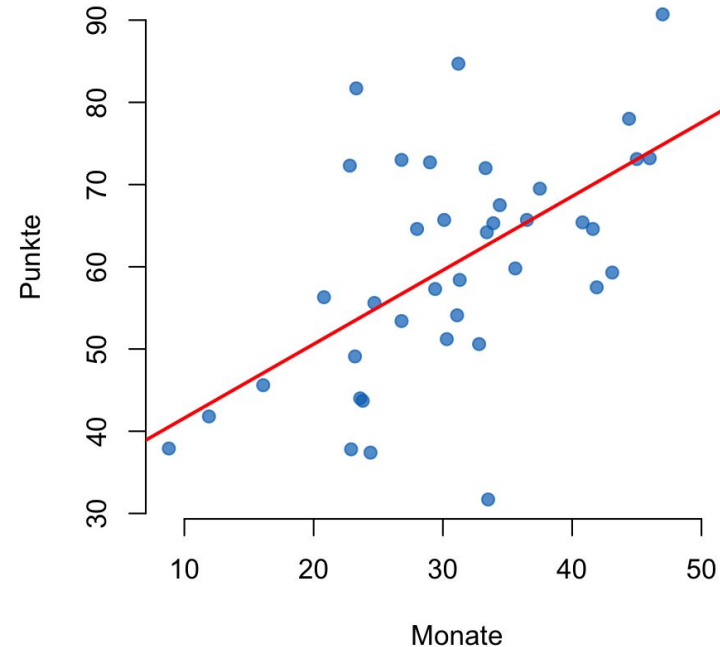
Lineare Regression



- Optimale Beschreibung einer Punktwolke durch eine Gerade
 - Modell wird angepasst oder *gefittet*.
- Zusammenhang zur Korrelation:
 - Positive Korrelation → Gerade steigt ("positive Steigung")
 - Negative Korrelation?
 - Korrelation gleich 0?

Lineare Regression

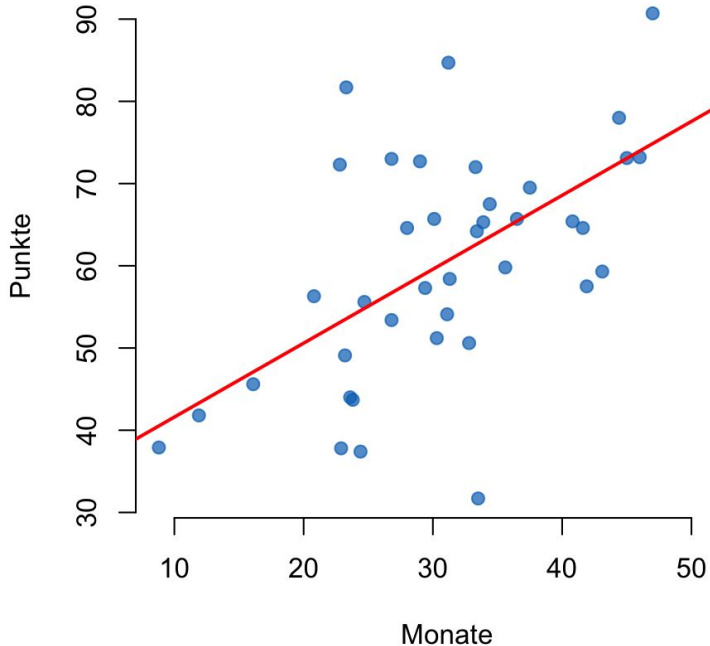
- Regressionsgeraden sind definiert durch zwei Parameter:
 - y -Achsenabschnitt / *Intercept* a
 - Steigung / *Slope* b
- y -Wert = Intercept + Slope • x -Wert
- $y = a + b \cdot x$



$$\text{Punkte} = 32,6 + 0,9 \cdot \text{Monate}$$

Regressionsparameter

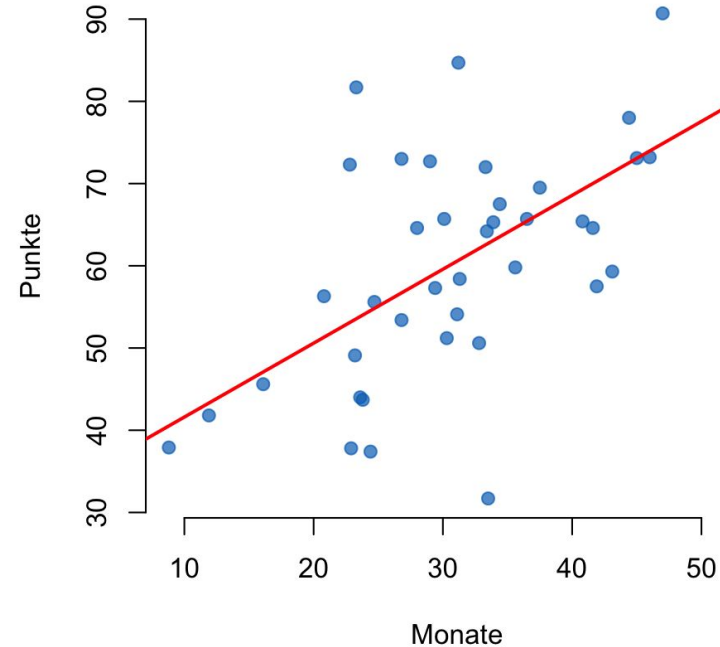
- Parameter sind interpretierbar.
 - Slope b : Um wie viel Punkte steigt das Testergebnis mit jedem Monat Aufenthalt in der Schweiz?
 - Intercept a : Wie viel Punkte erzielt man, wenn man noch nicht in der Schweiz war (Monate = 0)?
- Hier: Pro Monat ca. 1 Punkt mehr, 32,6 Punkte bei 0 Monaten.



$$\text{Punkte} = 32,6 + 0,9 \cdot \text{Monate}$$

Regression als Vorhersage

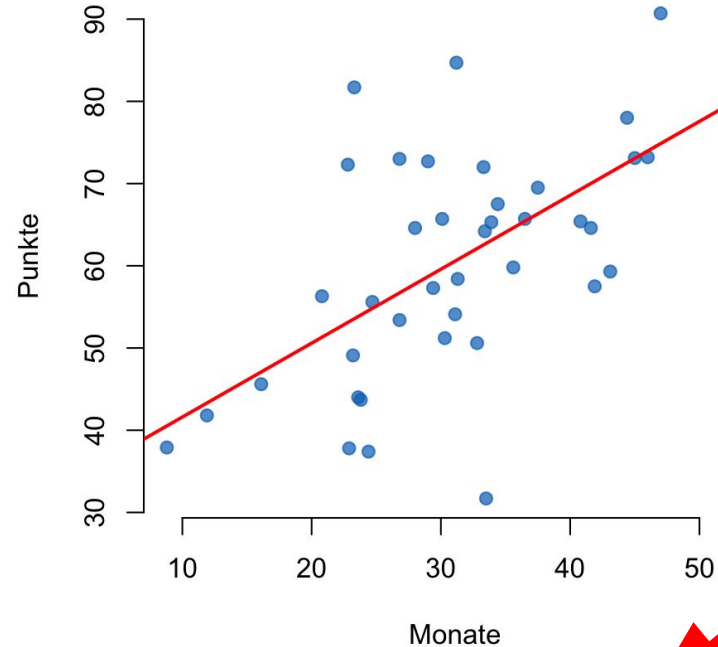
- Aufgrund der Gerade können wir bei neuen Werten von x vorhersagen, welchen y -Wert wir erwarten würden.



$$\text{Punkte} = 32,6 + 0,9 \cdot \text{Monate}$$

Regression als Vorhersage

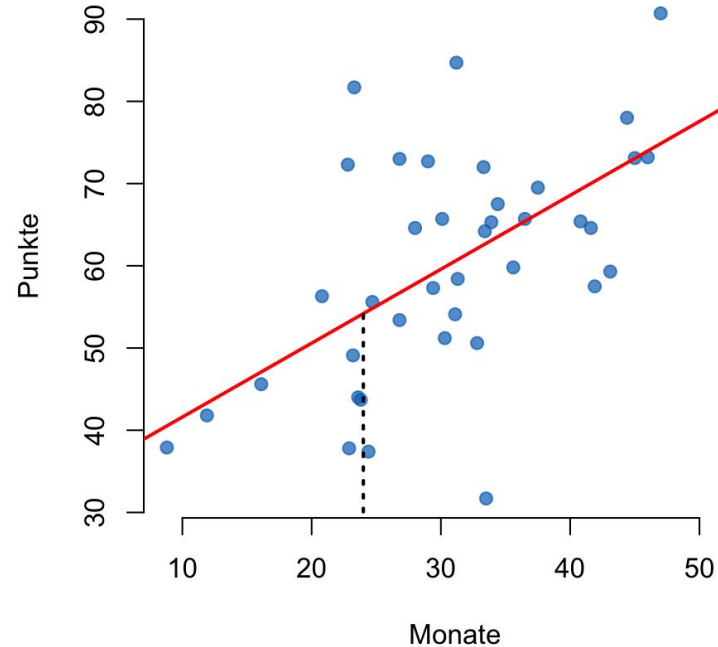
- Uns wird eine Sprachschülerin vorgestellt, die seit zwei Jahren in der Schweiz lebt.
- Wie viel Punkte wird sie wohl in dem Test erzielen?



$$\text{Punkte} = 32,6 + 0,9 \cdot 24$$

Regression als Vorhersage

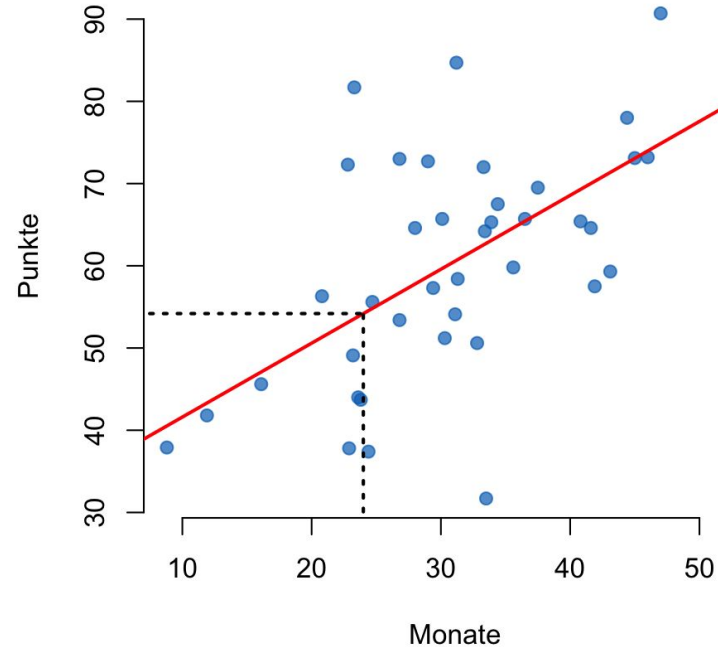
- Uns wird eine Sprachschülerin vorgestellt, die seit zwei Jahren in der Schweiz lebt.
- Wie viel Punkte wird sie wohl in dem Test erzielen?



$$\text{Punkte} = 32,6 + 0,9 \cdot 24$$

Regression als Vorhersage

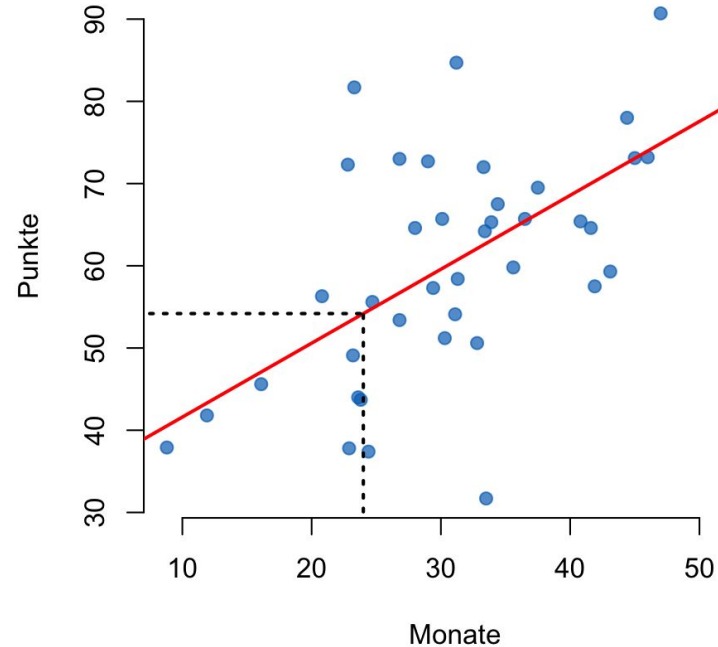
- Uns wird eine Sprachschülerin vorgestellt, die seit zwei Jahren in der Schweiz lebt.
- Wie viel Punkte wird sie wohl in dem Test erzielen?



$$54,2 = 32,6 + 0,9 \cdot 24$$

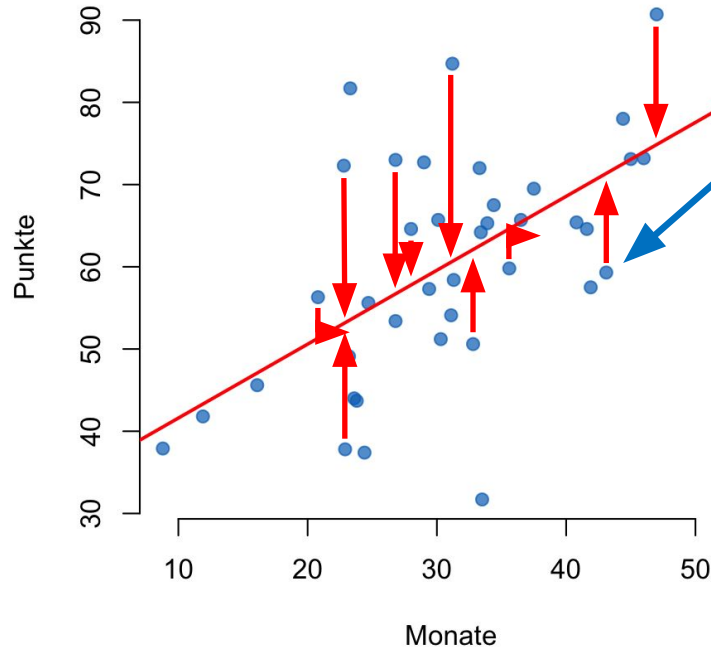
Regression als Vorhersage

- Natürlich können wir uns mit dieser Vorhersage irren.
- Sie ist aber unser "best guess" gegeben die Daten, die wir bisher gesammelt haben.



$$54,2 = 32,6 + 0,9 \cdot 24$$

Residuen (= Vorhersagefehler)



Hat diese Person "zu wenige" Punkte erreicht?

Oder ist gar unsere Regressionsgerade fehlerhaft?

- Die Person hat im Vergleich zu allen anderen und gegeben ihre Aufenthaltsdauer in der Schweiz zu wenige Punkte erreicht.
- Unsere Regressionsgerade beschreibt die Daten, die wir haben, optimal.

Das heisst: Die Gerade minimiert die Abweichung aller Punkte zur Geraden (= Summe der Residuen).

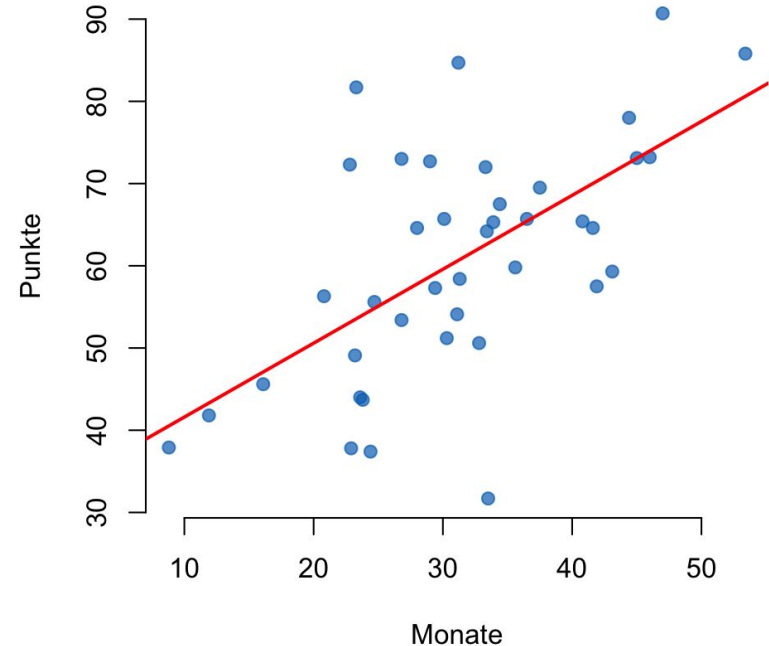
Residuen sind interpretierbar

- Residuen sind das, was durch die Vorhersage von x auf y nicht erklärt werden kann.
 - Unaufgeklärte Varianz
- Was könnte das in unserem Beispiel sein?
 - Allgemeine Sprachfähigkeit
 - Kenntnis zusätzlicher Zweitsprachen
 - Kenntnis anderer germanischer Sprachen
 - Kontakthäufigkeit mit deutschsprachigen Personen in der Schweiz

→ **Kovariaten**

Korrelation und Regression

- Korrelation hier:
 - Pearson: $r = 0,60$
 - Spearman: $r = 0,56$
- Aber: Korrelationen sind **bidirektional**.
- Regressionen sind **gerichtet**: Vorhersage von y aus x .
 - y -Abweichungen (= Residuen) werden minimiert.
 - (Bei Vorhersage von x aus y werden x -Abweichungen minimiert, dabei verändert sich die Regressionsgerade.)



Lineare Regression in R

- Vorhersage: Kontinuierliche (metrisch skalierte) Variable
- Prädiktoren: Kontinuierliche oder diskrete Variable(n)
- Funktion: `lm()` – *linear model*
- Syntax: `lm(<Formel>)`

`<Kriterium> ~ <Prädiktorstruktur>`

"predicted by"

Lineare Regression in R

Aufruf

```
Call:  
lm(formula = testat2024$`Punkte T1` ~ testat2024$`HA Ges`)
```

```
Residuals:  
      Min       1Q   Median       3Q      Max  
-30.738 -14.268  -1.208   10.012   47.262
```

Verteilung der Residuen

Coefficients:

Effektschätzer & stat. Prüfgröße

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	14.268	7.253	1.967	0.063185 .
testat2024\$`HA Ges`	8.235	2.063	3.992	0.000717 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.47 on 20 degrees of freedom

Multiple R-squared: 0.4435. Adjusted R-squared: 0.4156

F-statistic: 15.94 on 1 and 20 DF, p-value: 0.0007166

Prädiktoren

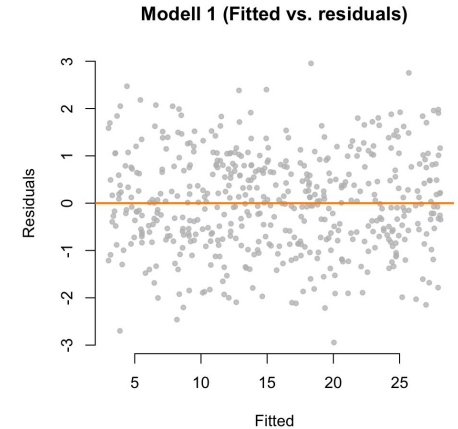
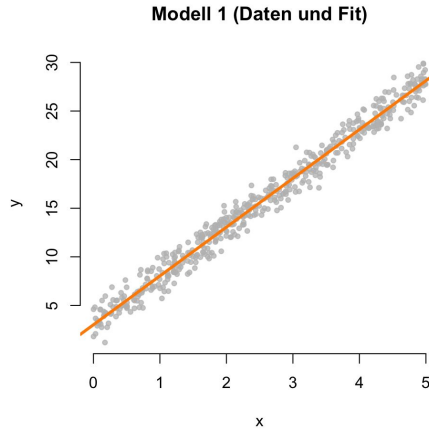
- `mod <- lm(...)`
- `summary(mod)`

Regressionsdiagnostik

```
mod <- lm(y ~ x)
```

```
plot(df$x, df$y)  
ablines(mod, col="red")
```

```
plot(fitted(mod), resid(mod))  
abline(h = 0)
```



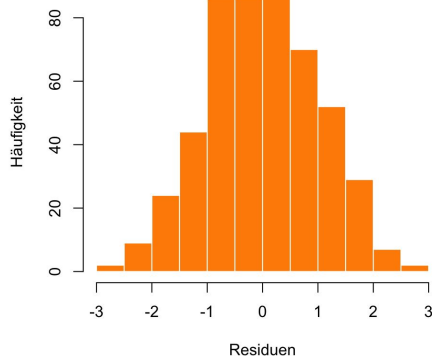
Regressionsdiagnostik

Die meisten Voraussetzungen können auch numerisch getestet werden. Hierzu z.B. <https://book.stat420.org/model-diagnostics.html>

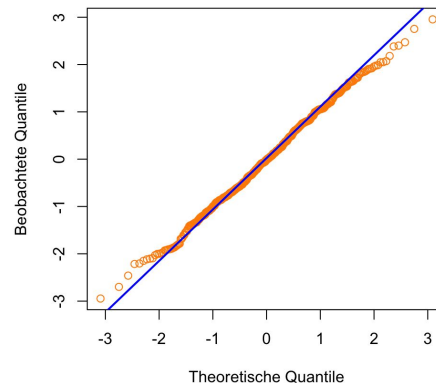
```
mod <- lm(y ~ x, data = df)
```

```
hist(resid(mod))
```

Histogramm Residuen Modell 1



QQ-Plot Modell 1



```
qqnorm(resid(mod))  
qqline(resid(mod))
```

Multiple Regression

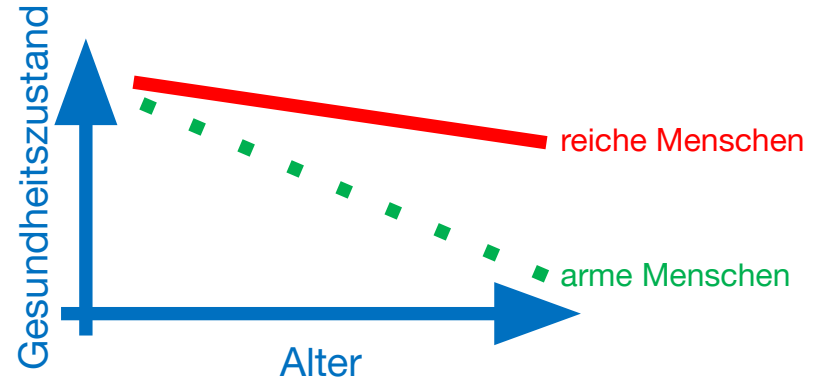
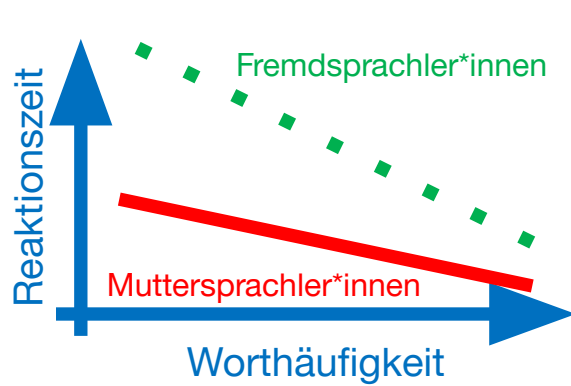
Prädiktoren werden auch **unabhängige Variablen** genannt. Die Kriteriumsvariable wird auch **abhängige Variable** genannt.

- Bisher haben wir eine y -Variable aus einer x -Variable vorhergesagt.
- Typischerweise benutzen wir mehrere **Prädiktoren**, um die **Kriteriumsvariable** vorherzusagen.
- Beispiele:
 - Korpusfrequenz + Wortart + Einbettungstiefe → Lesezeit
 - Ticketpreis + Wetter + Beliebtheit der Band → Anzahl Konzertbesucher
 - Fahrgastaufkommen + Wetter + Streckenzustand → Verspätungen
 - Hausaufgaben + Anwesenheit + Folienstudium → Testaterfolg
 - ...

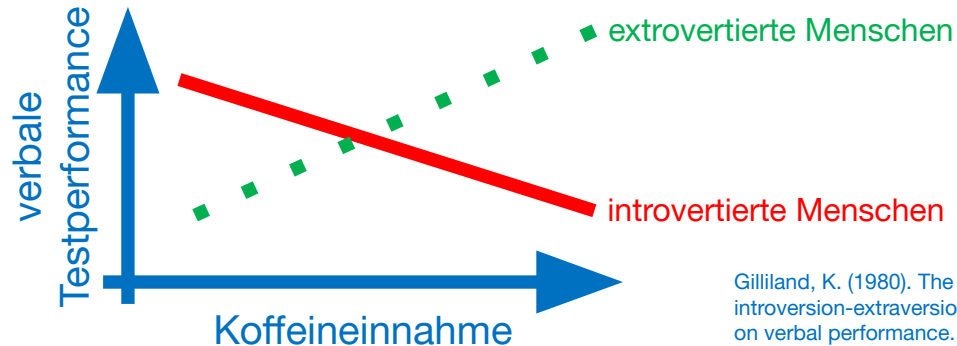
Multiple Regression

- Bei der multiplen Regression bekommt jeder Prädiktor seine eigene Steigung.
 - Auch: **(β -)Gewicht, Koeffizient, *coefficient*, *estimate***
- Neben den **Einzeleffekten** (*single/main effects*) sind auch **Interaktionen** möglich.
 - Interaktion: Das Zusammenwirken von zwei oder mehr Prädiktoren auf die Kriteriumsvariable.
 - Beispiele:
 - Das Wetter hat nur bei unbeliebteren Bands einen Einfluss auf die Anzahl der Gäste.
 - Je höher ein Wort eingebettet ist, desto mehr Einfluss hat die Worthäufigkeit auf die Lesezeit.

Interaktionen: Beispiele



Quelle: Datensatz lexdec aus {languageR}



Gilliland, K. (1980). The interactive effect of introversion-extraversion with caffeine induced arousal on verbal performance. *Journal of Research in Personality*, 14(4), 482–492.
[https://doi.org/10.1016/0092-6566\(80\)90006-9](https://doi.org/10.1016/0092-6566(80)90006-9)

Regression: Voraussetzungen

Linearität des Zusammenhangs

Das Kriterium kann als eine lineare Kombination der Prädiktoren ausgedrückt werden.

Varianzhomogenität der Residuen

Die Fehlervarianz ist überall ungefähr gleich.

Normalverteilung der Residuen

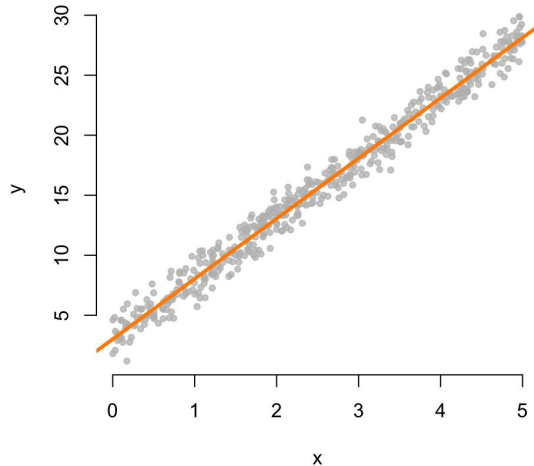
Die Residuen sind normalverteilt.

Voraussetzungen: Linearität

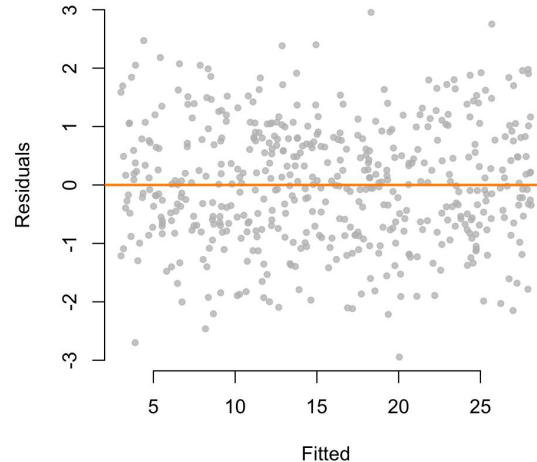
Fitted sind die geschätzten Werte, also die Werte auf der Regressionsgeraden.

- Keine große Überraschung: Lineare Regressionen können nur lineare Zusammenhänge erfassen.
- Nützlicher Diagnostik-Plot: *Fitted vs. residuals*

Modell 1 (Daten und Fit)



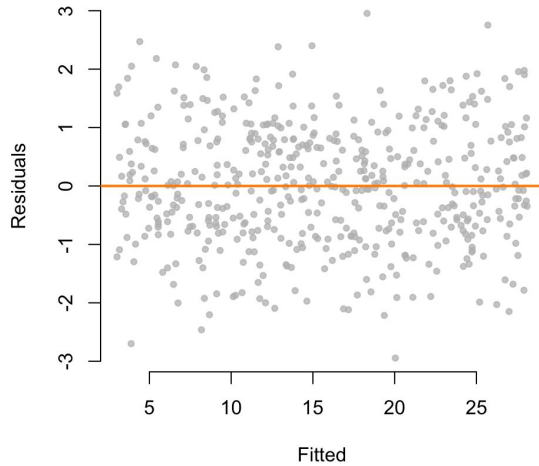
Modell 1 (Fitted vs. residuals)



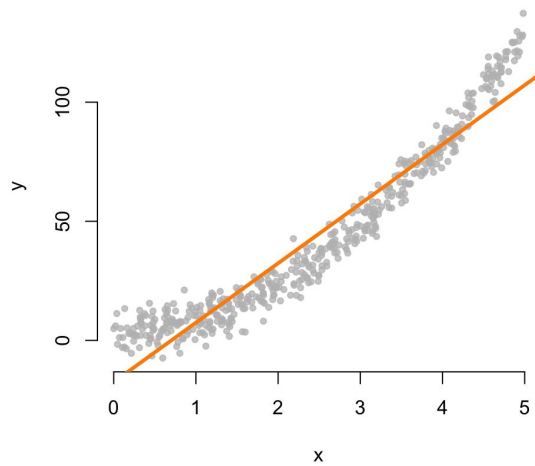
Linearität ist gegeben, wenn die Residuen sich gleichmäßig um den Fit verteilen und keine eindeutige Abweichung erkennbar ist.

Voraussetzungen: Linearität

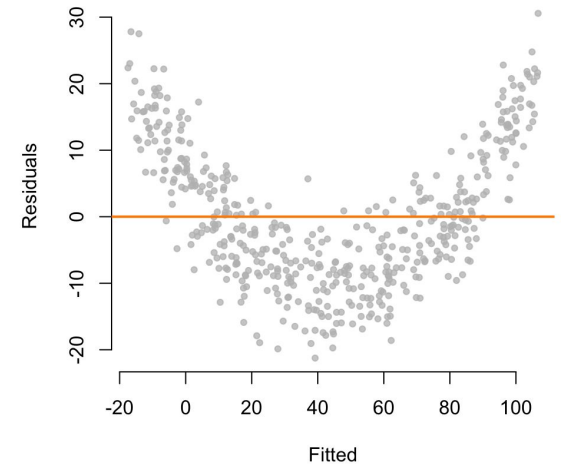
Modell 1 (Fitted vs. residuals)



Modell 3 (Daten und Fit)

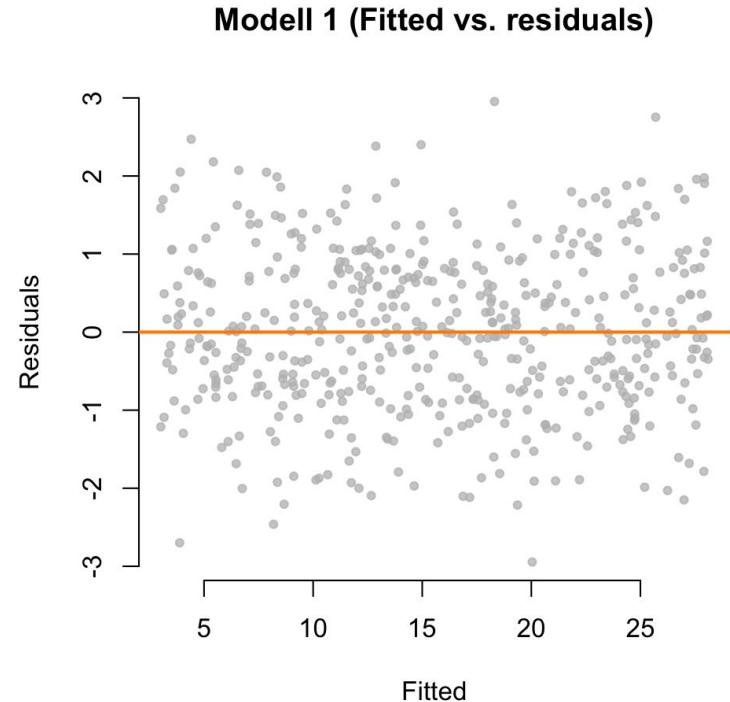


Modell 3 (Fitted vs. residuals)



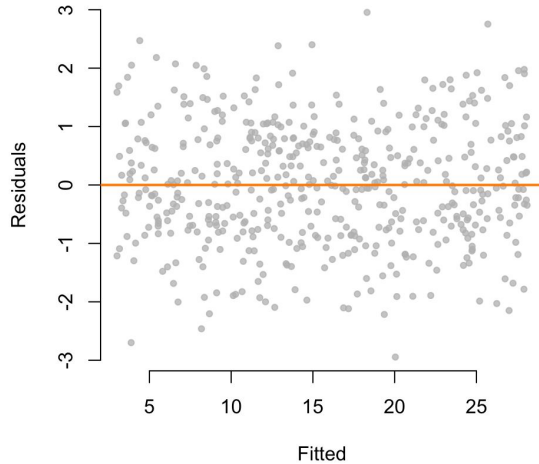
Voraussetzungen: Varianzhomogenität

- Die Varianz der Residuen muss für alle Abschnitte des Prädiktors (der Prädiktoren) ungefähr gleich sein.
- **Heteroskedastizität** ist die Verletzung dieses Prinzips.

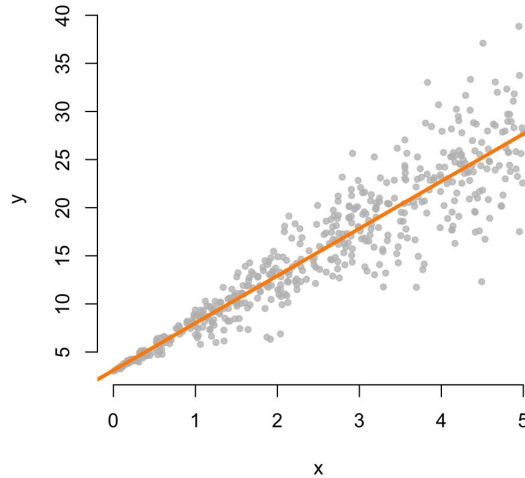


Voraussetzungen: Varianzhomogenität

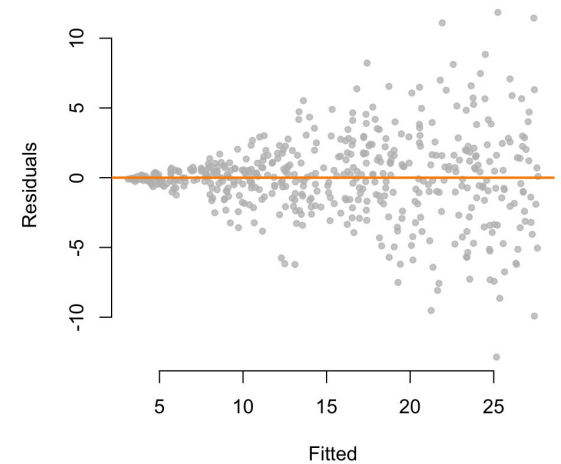
Modell 1 (Fitted vs. residuals)



Modell 2 (Daten und Fit)



Modell 2 (Fitted vs. residuals)

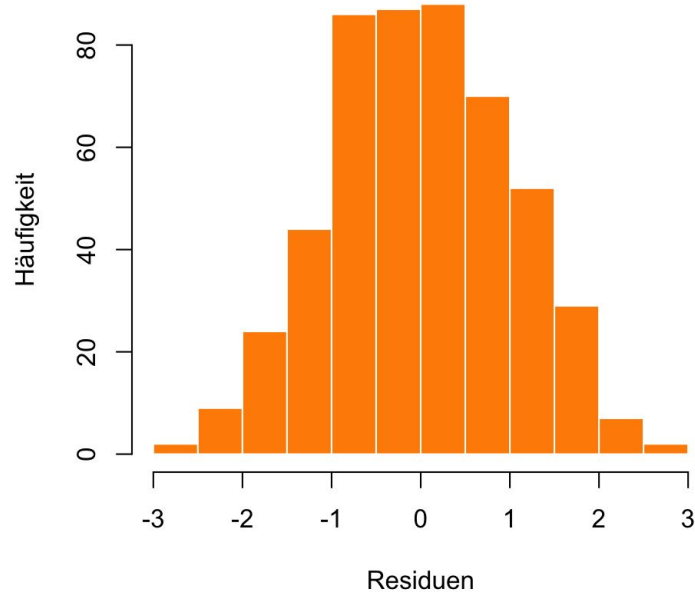


Voraussetzungen: Normalverteilung

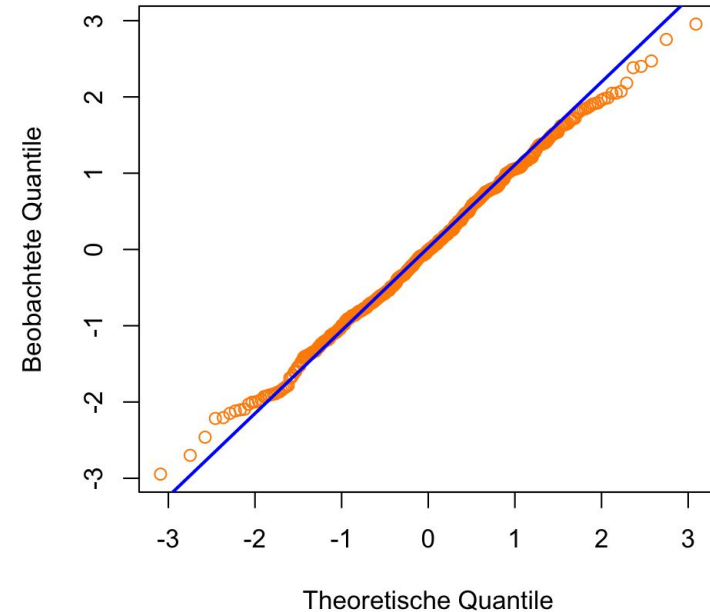
- Residuen müssen normalverteilt sein.
- Geeignete Diagnostik-Plots: Histogramm und QQ-Plot für die Residuen aus dem Regressionsmodell
- Histogramm zeigt die Anzahl an Datenpunkten in bestimmten Abschnitten (= Verteilung).
- QQ-Plot plottet theoretische Quantile (laut Normalverteilung) gegen beobachtete Quantile.

Voraussetzungen: Normalverteilung

Histogramm Residuen Modell 1

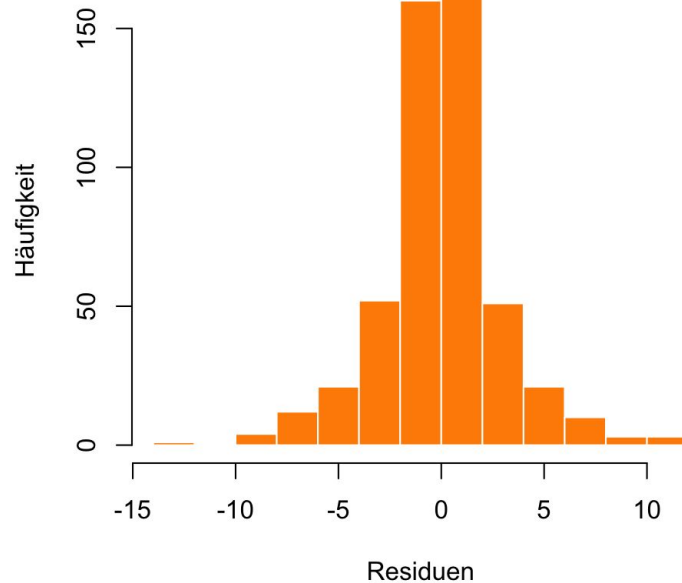


QQ-Plot Modell 1

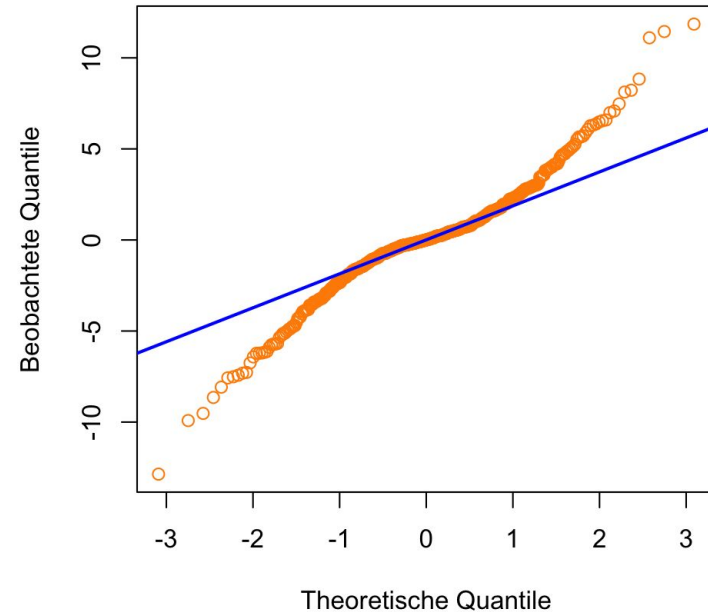


Voraussetzungen: Normalverteilung

Histogramm Residuen Modell 2

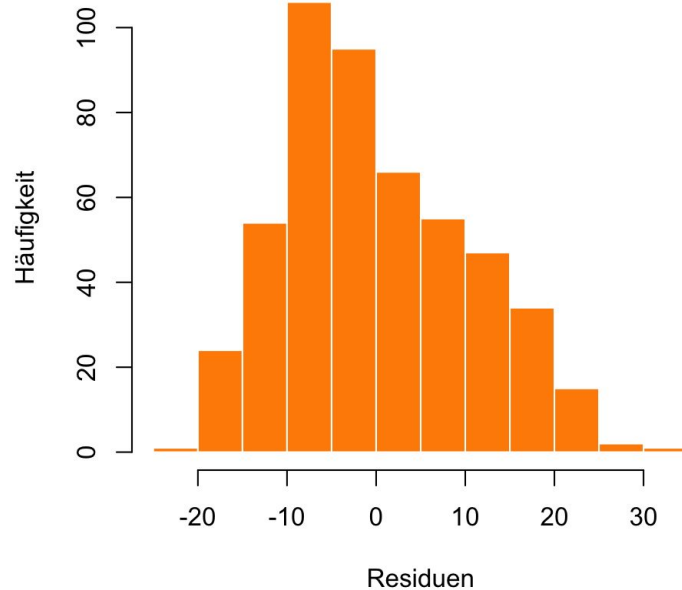


QQ-Plot Modell 2

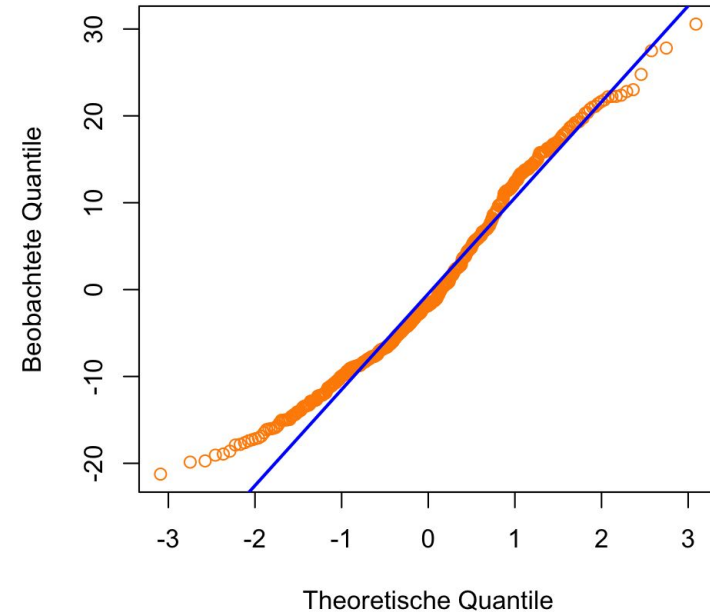


Voraussetzungen: Normalverteilung

Histogramm Residuen Modell 3



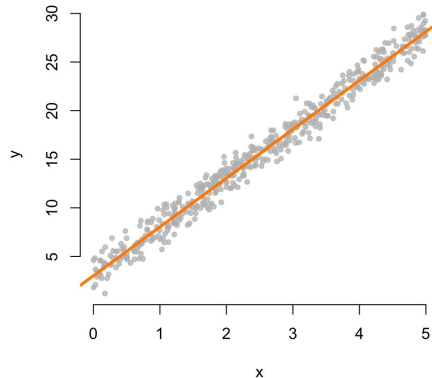
QQ-Plot Modell 3



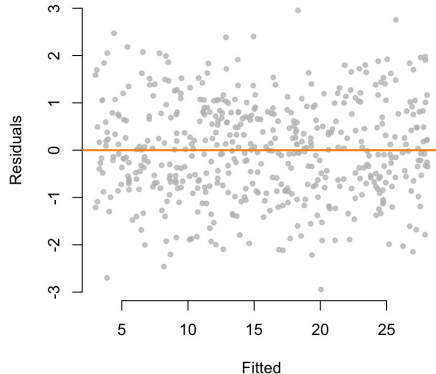
Voraussetzungen

- Von den vorherigen Modellen würde nur Modell 1 alle Voraussetzungen eindeutig erfüllen.

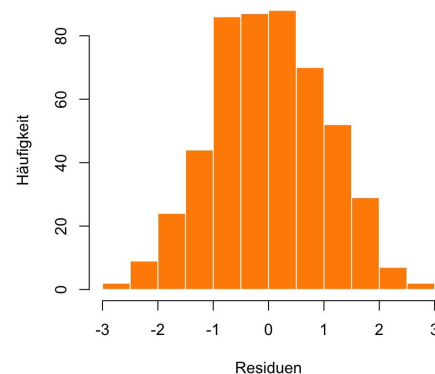
Modell 1 (Daten und Fit)



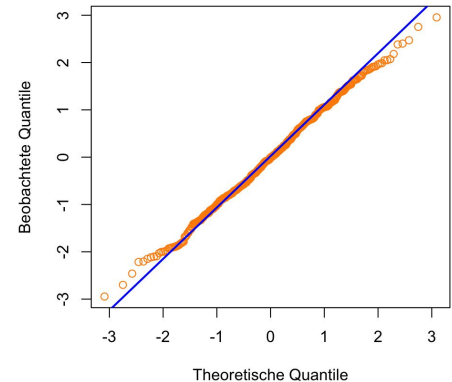
Modell 1 (Fitted vs. residuals)



Histogramm Residuen Modell 1



QQ-Plot Modell 1

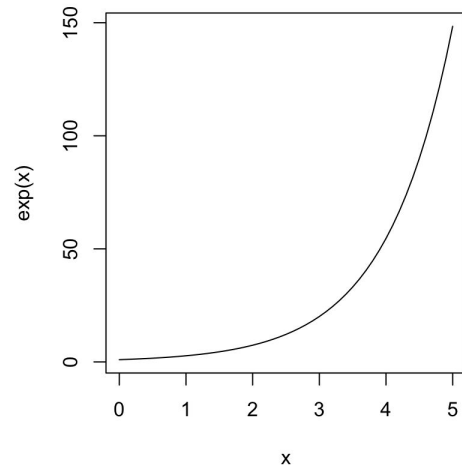
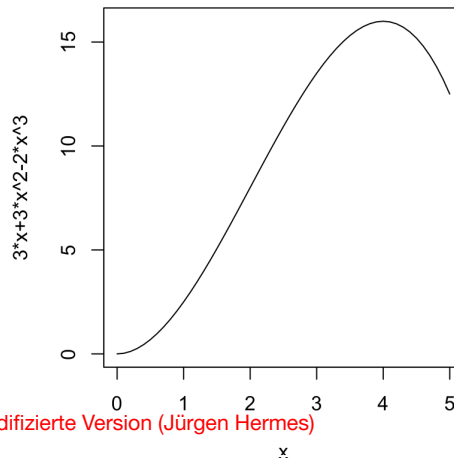
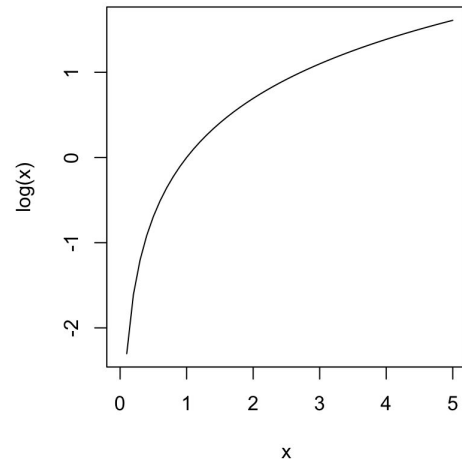


Exkurs: Nicht-lineare Regression

- Fitten eines nicht-linearen Zusammenhangs zwischen Prädiktoren und Kriterium
 - Logarithmische Funktionen `log()`
 - Exponentialfunktion `exp()`
 - Polynome `poly()`
 - ...

```
y ~ log(x)  
y ~ exp(x)  
y ~ poly(x, 3)
```

Anzahl der Terme, Vorsicht vor Overfitting!



Zusammenfassung

- Mit Regressionen beschreiben wir den (rechnerischen) Zusammenhang zwischen Variablen.
- Es wird immer **eine** Variable vorhergesagt.
 - Linear: Lineare Regression
 - Binär: Binär-logistische Regression (ausgelassen)
- Mehrere Prädiktoren möglich, ggf. auch Interaktionen
 - Zusammenwirken von Prädiktoren auf Kriterium
- Jeder Prädiktor/jede Interaktion bekommt einen Effektschätzer.
- Die Residuen sind die Vorhersagefehler.

Zusammenfassung

- Regressionen sind (im Gegensatz zu Korrelationen) **gerichtet**.
- Lineare Regressionsanalysen haben Voraussetzungen:
 - Linearität, Varianzhomogenität, Normalverteilung
- Die sog. Formel (*formula*) gibt, welche Zusammenhänge wir modellieren wollen.
 - ~ "predicted by"
- Nicht-lineare Regressionen: Modellierung nicht-linearer Zusammenhänge.

Begriffe

Intercept

Prädiktoren

Heteroskedastizität

Steigung / Slope

Kriteriumsvariable

Histogramm

Vorhersage

Interaktion

QQ-Plot

Residuen

Linearität

Binär-logistische R.

Kovariaten

Varianzhomogenität

Formel / *formula*

Multiple Regression

Normalverteilung

Nicht-lineare Regr.

Hausaufgabe

- Laden Sie aus dem Package `openintro` das Dataframe `tips` auf eine eigene Variable
- Plotten Sie (Punktdiagramm) die Werte der Spalten `bill` und `tip`.
- Berechnen Sie die Korrelation zwischen diesen beiden Variablen.
- Berechnen Sie die den Einfluss, den die Höhe der `bill` zur Vorhersage des `tip` hat (lineare Regression). Geben Sie eine Summary des linearen Modells aus.
- Plotten Sie die Regressionsgerade in rot in den vorhandenen Plot.
- Ergänzen Sie den Datensatz um eine Spalte `tipRate`, in der Sie den prozentualen Anteil des `tip` an der `bill` berechnen.
- Berechnen Sie das arithmetische Mittel und die Standardabweichung für die `tipRate` für die unterschiedlichen Wochentage (mit `tapply`).
- Zeichnen Sie einen Boxplot der `tipRate` in Abhängigkeit von den Wochentagen und einen weiteren für die unterschiedlichen Wochen.