Einführung in die Statistik

Praktische Übung – Jürgen Hermes – IDH – SoSe 2024

Statistik (Stochastik) – Die Grundlagen

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

Multiple Regression

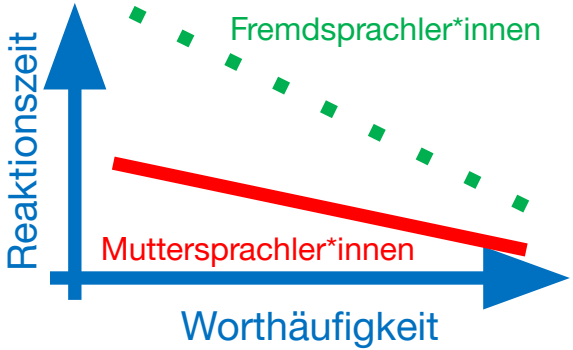
Prädiktoren werden auch **unabhängige Variablen** genannt. Die Kriteriumsvariable wird auch **abhängige Variable** genannt.

- 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
  - ...

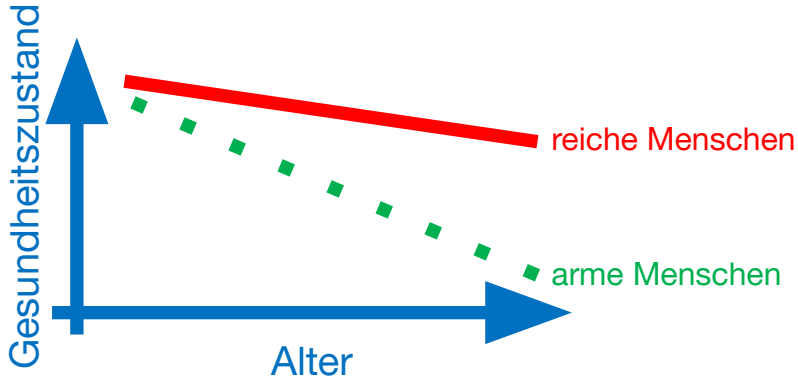
Multiple Regression

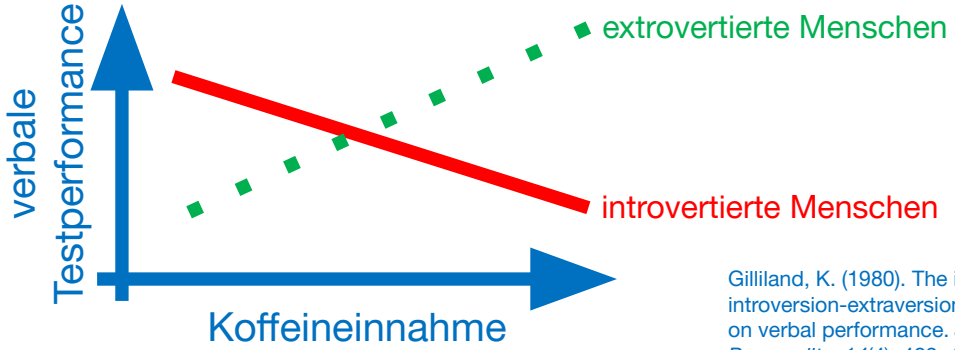
- Bei der multiplen Regression bekommt jeder Prädiktor seine eigene Steigung.
  - Auch: **(*β*-)Gewicht**, **Koeffizient**, ***coefficient***, ***estimate***
- Neben den **Einzeleffekten** (*single/main effects*) sind auch **Interaktionen** möglich.
  - Interaktion: Das Zusammenwirken von zwei oder mehr 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

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.

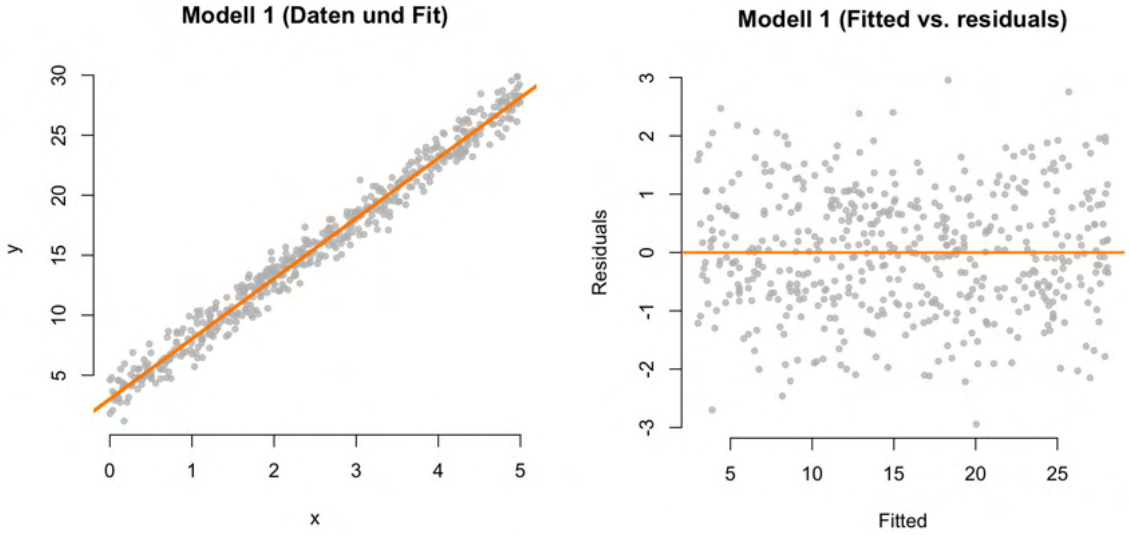
Normalverteilung
der Residuen

Die Residuen sind normalverteilt.

Voraussetzungen: Linearität

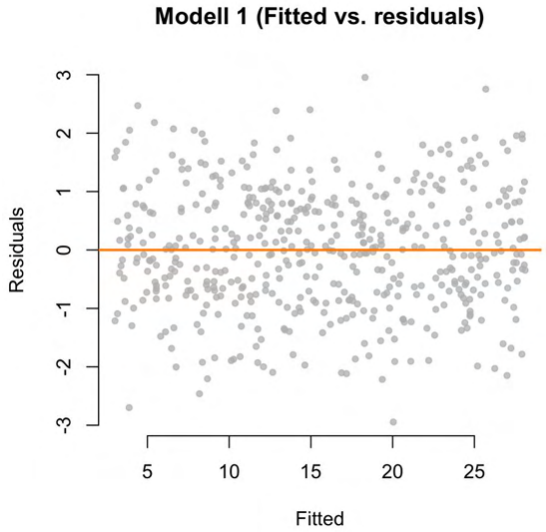
Fitted sind die geschätzten Werte, also die Werte auf der Regressionsgeraden.

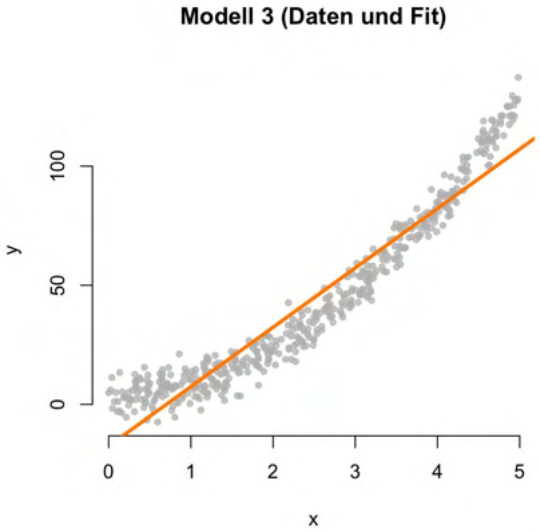
- Keine große Überraschung: Lineare Regressionen können nur lineare Zusammenhänge erfassen.
- Nützlicher Diagnostik-Plot: *Fitted vs. residuals*

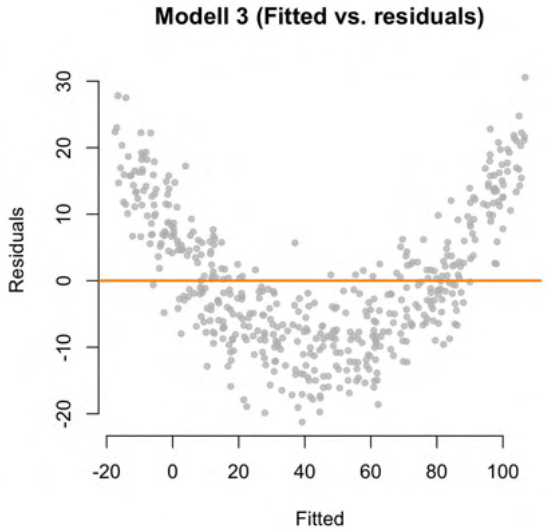


Linearität ist gegeben, wenn die Residuen sich gleichmäßig um den Fit verteilen und keine eindeutige Abweichung erkennbar ist.

Voraussetzungen: Linearität

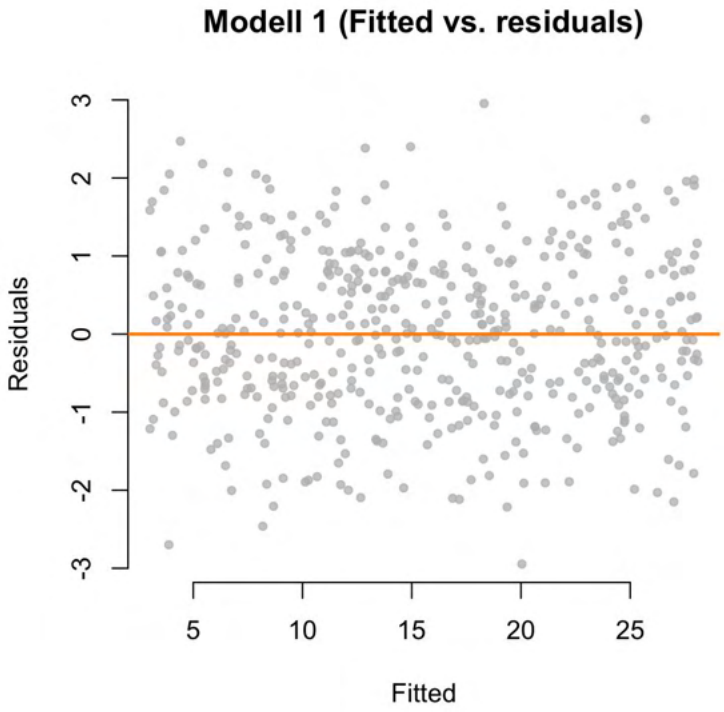




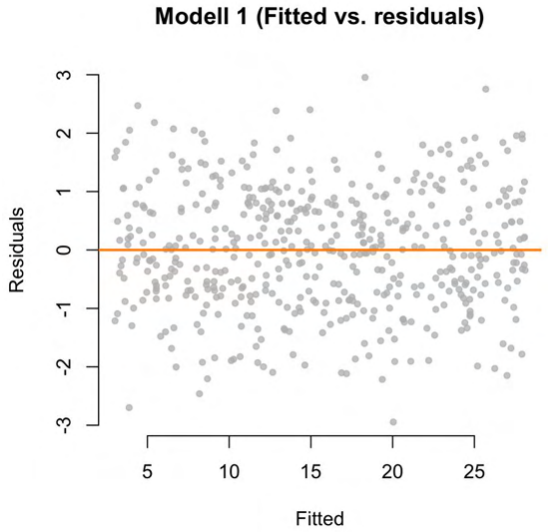


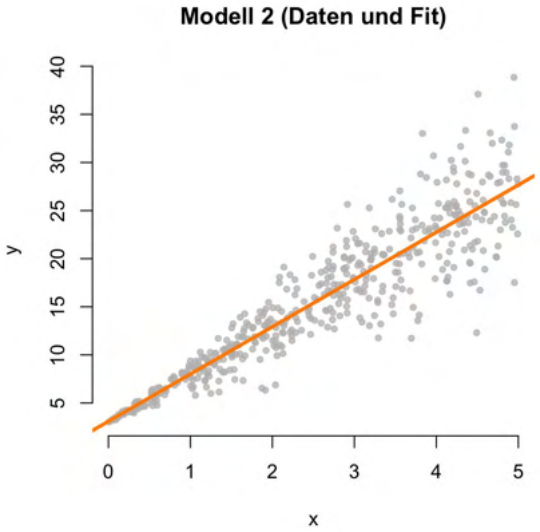
Voraussetzungen: Varianzhomogenität

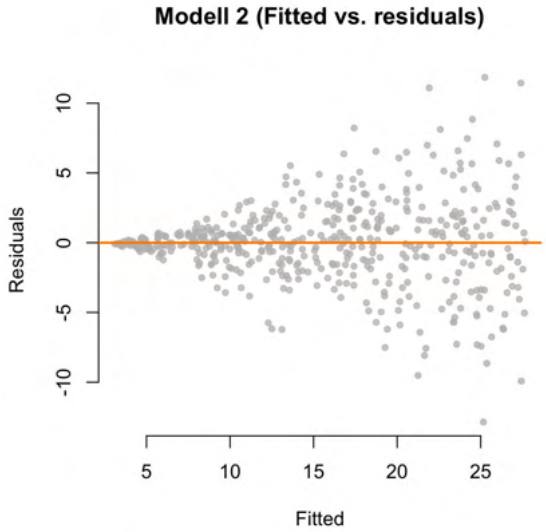
- Die Varianz der Residuen muss für alle Abschnitte des Prädiktors (der Prädiktoren) ungefähr gleich sein.
- **Heteroskedastizität** ist die Verletzung dieses Prinzips.



Voraussetzungen: Varianzhomogenität



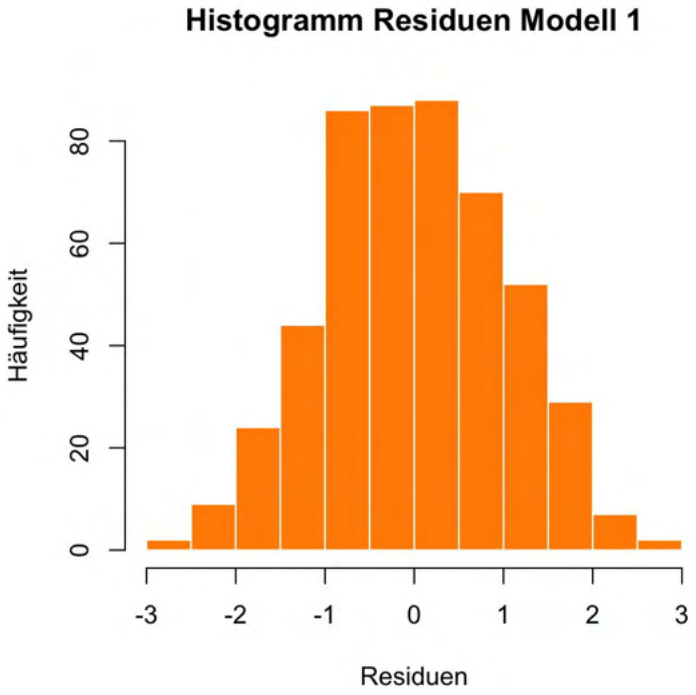


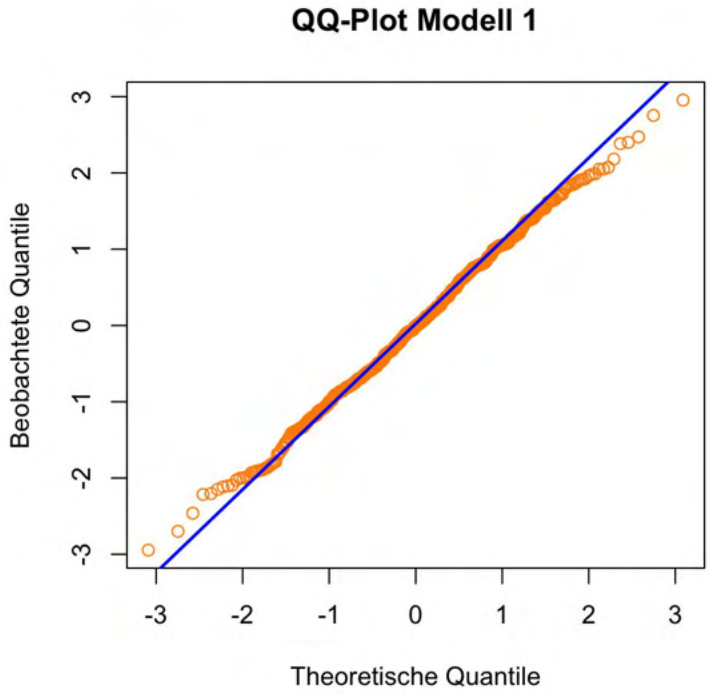


Voraussetzungen: Normalverteilung

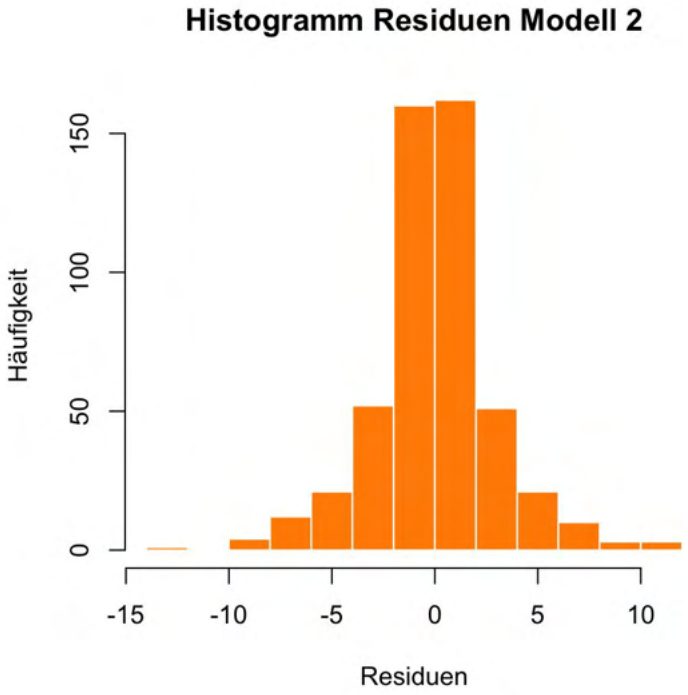
- Residuen müssen normalverteilt sein.
- Geeignete Diagnostik-Plots: Histogramm und QQ-Plot für die Residuen aus dem Regressionsmodell
- Histogramm zeigt die Anzahl an Datenpunkten in bestimmten Abschnitten (= Verteilung).
- QQ-Plot plottet theoretische Quantile (laut Normalverteilung) gegen beobachtete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

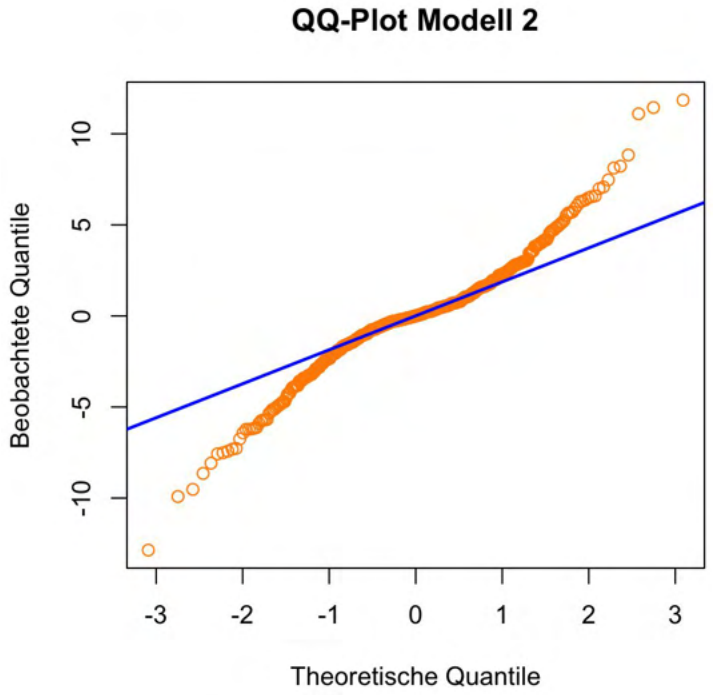
Voraussetzungen: Normalverteilung



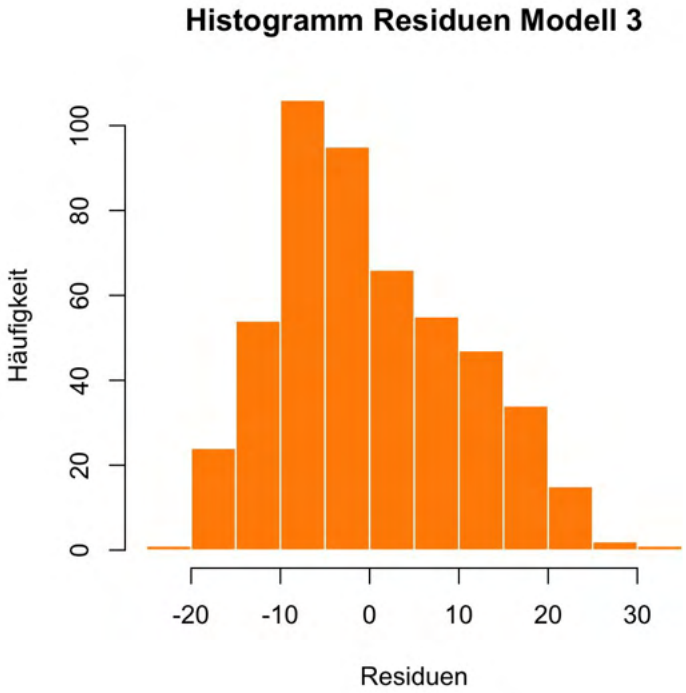


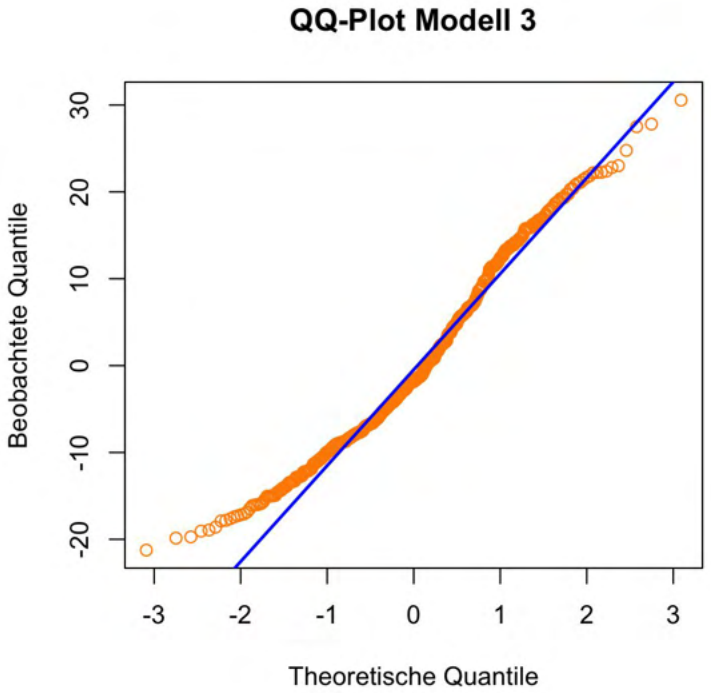
Voraussetzungen: Normalverteilung





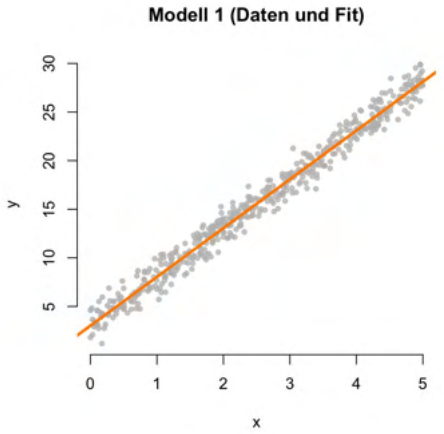
Voraussetzungen: Normalverteilung

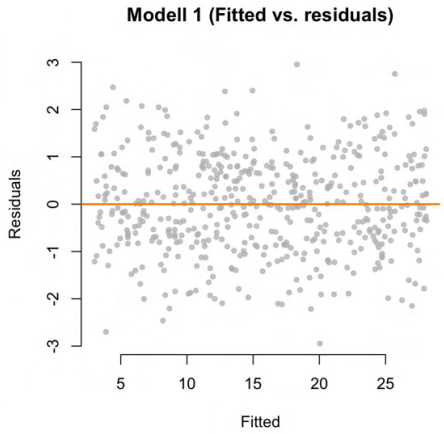


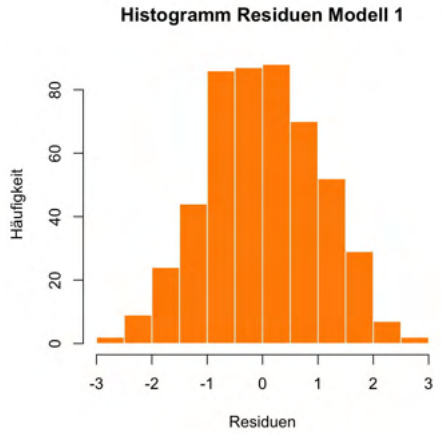


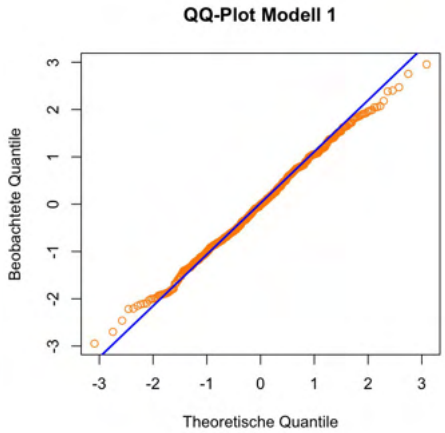
Voraussetzungen

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







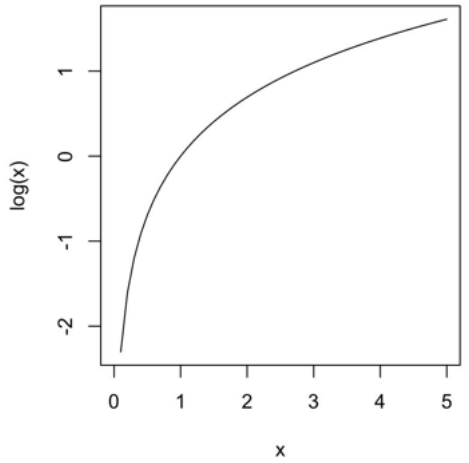


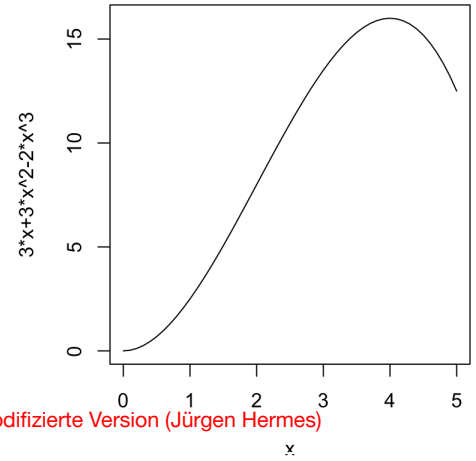
Exkurs: Nicht-lineare Regression

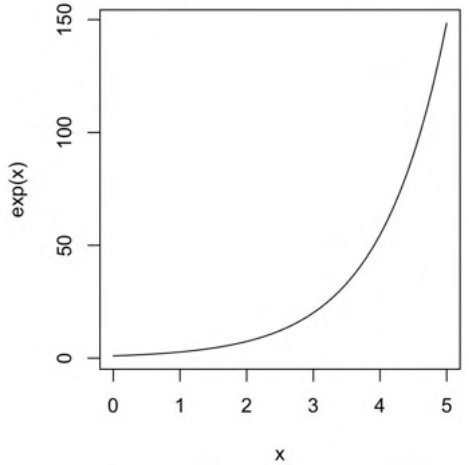
- Fitten eines nicht-linearen Zusammenhangs zwischen Prädiktoren und Kriterium
  - Logarithmische Funktionen `log()`
  - Exponentialfunktion `exp()`
  - Polynome `poly()`
  - ...

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

Anzahl der Terme, Vorsicht vor Overfitting!







leicht modifizierte Version (Jürgen Hermes)

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.**

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

Inferenzstatistik

Inferenzstatistik

Man nennt dieses Vorgehen auch *null-hypothesis significance testing (NHST).*

- Mit inferenzstatistischen Verfahren wollen wir von unserer **Stichprobe** auf die zugrundeliegende **Grundgesamtheit** (**Population**) verallgemeinern.
- Wir testen anhand unserer Stichprobe eine **Hypothese**, von der wir wissen möchten, ob sie in der Population gilt.
- Die Hypothese, die getestet wird, ist immer die **Nullhypothese**.
  - H_0: Es gibt *keinen* Zusammenhang / Unterschied (= Effekt).
- Unsere Forschungshypothese ist stets die **Alternativhypothese**.
  - H_1: Es gibt einen Zusammenhang / Unterschied.
- Wir wollen herausfinden, ob wir die Nullhypothese mit ausreichend großer Wahrscheinlichkeit **ablehnen** können. Dafür benötigen wir einen Test.

Inferenzstatistik – Beispiel

- Unsere **Stichprobe** sind Studierende, die letztes Jahr ein Statistik-Testat geschrieben haben. Unsere **Grundgesamtheit** (**Population**) sind alle Menschen, die dieses Statistik-Testat hätten schreiben können.
- Die **Hypothese** ist, das die Abgabe von Hausaufgaben in der Übung das Ergebnis des finalen Testats positiv beeinflusst.
- **Nullhypothese**: Die Abgabe von Hausaufgaben hat keinen Effekt auf das Testat-Ergebnis.
- **Alternativhypothese**: Die Abgabe von Hausaufgaben hat einen positiven Effekt auf das Testat-Ergebnis.
- Wir wollen herausfinden, ob wir die Nullhypothese mit ausreichend großer Wahrscheinlichkeit **ablehnen** können.

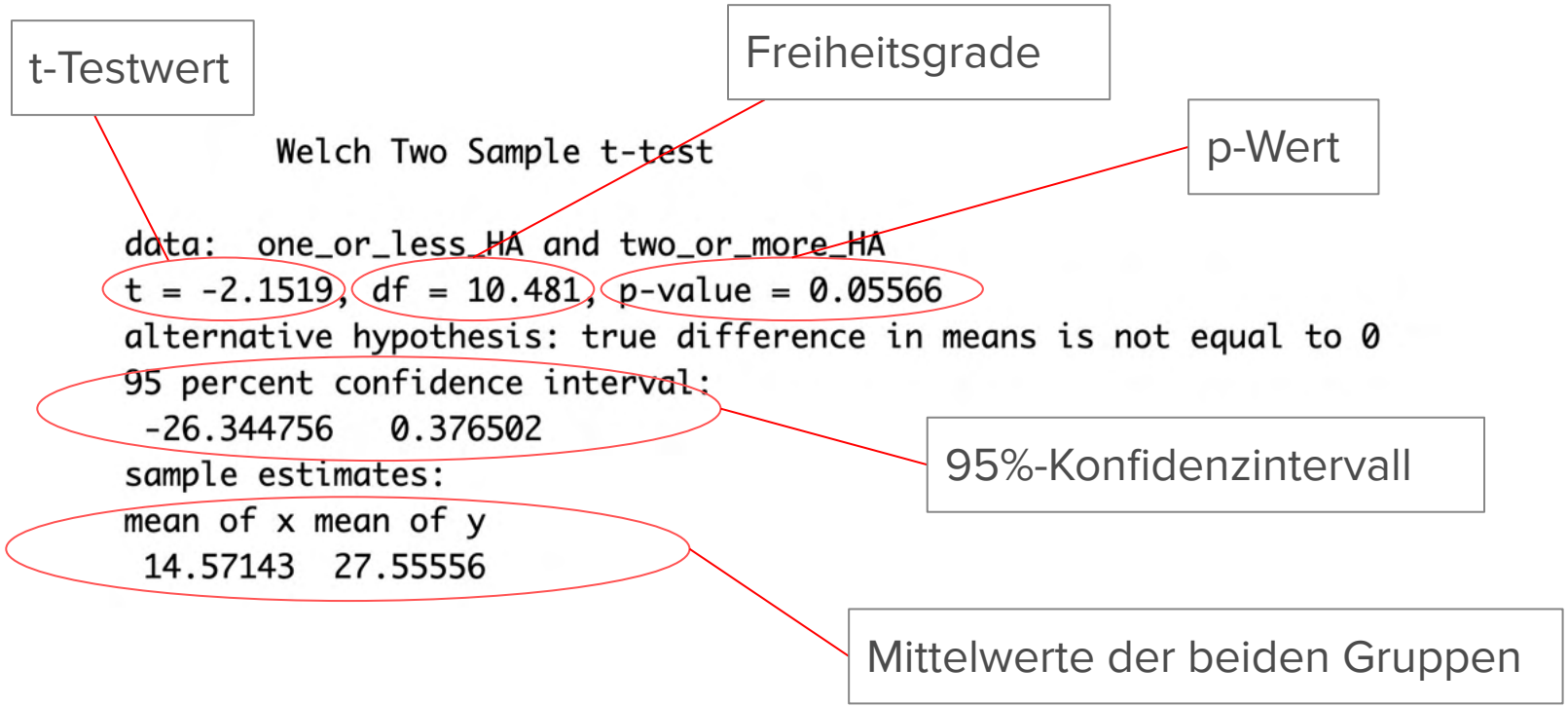
 völlig neue Folie (Jürgen Hermes)

Beispiel für ein statistisches Testverfahren: t-test

- **Vergleich von Mittelwerten**: Ein t-Test wird verwendet, um zu überprüfen, ob die Mittelwerte von zwei unabhängigen Gruppen signifikant unterschiedlich sind.
- **Hypothesentest**: Ein t-Test testet die Nullhypothese (H0), dass die Mittelwerte der beiden Gruppen gleich sind, gegen die Alternativhypothese (H1), dass die Mittelwerte unterschiedlich sind.
- **Statistische Signifikanz**: Der t-Test berechnet eine Teststatistik und einen p-Wert, um zu bestimmen, ob die beobachteten Unterschiede durch Zufall erklärt werden können oder statistisch signifikant sind (d.h., nicht durch Zufall bedingt).

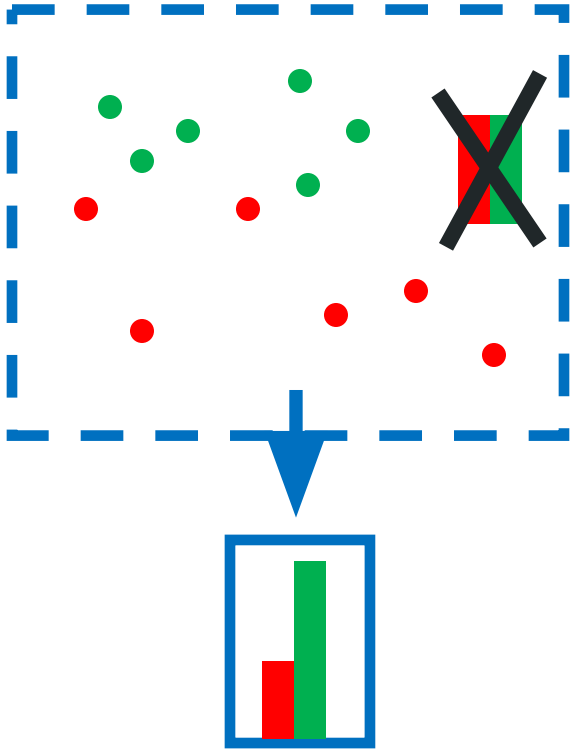
 völlig neue Folie (Jürgen Hermes)

t-test: Interpretation



Signifikanzniveau *p*

- Wir nehmen *hypothetisch* an, dass in der Grundgesamtheit die Nullhypothese gilt.
- Wir ziehen zufällig eine Stichprobe aus der Grundgesamtheit.
- *p* gibt uns an, wie wahrscheinlich die Ergebnisse in der Stichprobe sind, wenn in der Grundgesamtheit tatsächlich die Nullhypothese gelten sollte.
- Wenn *p* **klein genug** ist, lehnen wir die Nullhypothese ab und nehmen stattdessen die Alternativhypothese an.



Signifikanzniveau *p*

- "Klein genug": kleiner 5% ($p < 0{,}05$)
  - Andere Schwellenwerte: 0,01 (1%); 0,001 (0,1%)
- Unbedingte Voraussetzung: Zufällige Stichprobenziehung, ansonsten *sampling bias*
- *p* kann uns **nicht** dabei helfen, die Nullhypothese zu **bestätigen**.
  - Nicht-Effekte können also nicht interpretiert werden! Es ist somit in diesem Paradigma äußerst schwierig, Gleichheit oder Nicht-Zusammenhänge nachzuweisen.
- Alternativhypothese wird nie "bewiesen", wir können lediglich mit hoher Wahrscheinlichkeit die Nullhypothese ablehnen.

p-Werte in R & Effektgrößen

- Viele statistische Methoden geben *p*-Werte aus.
  - `summary()` von linearen Modellen
  - `cor.test()` statt `cor()`
  - `t.test()`
- Statistische Signifikanz ist aber nicht alles. Auch (sehr) kleine Effekte können signifikant sein.
  - z. B. sehr kleine Unterschiede zwischen Gruppen, schwache Zusammenhänge
- Auch sehr große Effekte sind manchmal *nicht* signifikant …
  - … z. B., weil das assoziierte Rauschen sehr groß ist.
- 5% ist eine **arbiträre Grenze**!
  - Ist $p = 0{,}049$ wirklich veröffentlichungswert und $p = 0{,}051$ nicht?

Freiheitsgrade

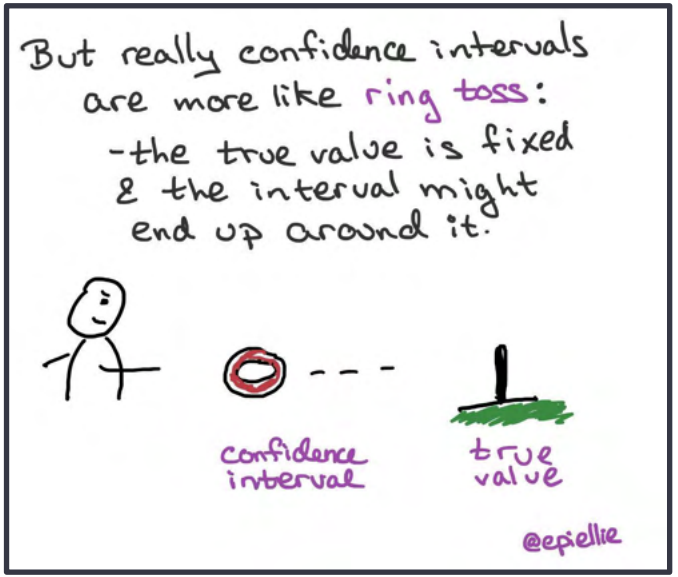
- Anzahl der unabhängigen Werte, die in einer Berechnung variieren können, ohne durch bekannte Summen oder Mittelwerte eingeschränkt zu sein.
- Höhere Freiheitsgrade führen zu einer Verteilung, die näher an der Normalverteilung liegt, was die Präzision des Tests erhöht.
- In einfachen t-test (Students t-test) werden sind Freiheitsgrade einfach die Anzahl der Elemente der Vektoren - 2. In Welch's t-Test passen sich die Freiheitsgrade an ungleiche Varianzen und Stichprobengrößen an, was zu einer robusteren Analyse führt.

 völlig neue Folie (Jürgen Hermes)

https://twitter.com/EpiEllie/status/1073385427317465089

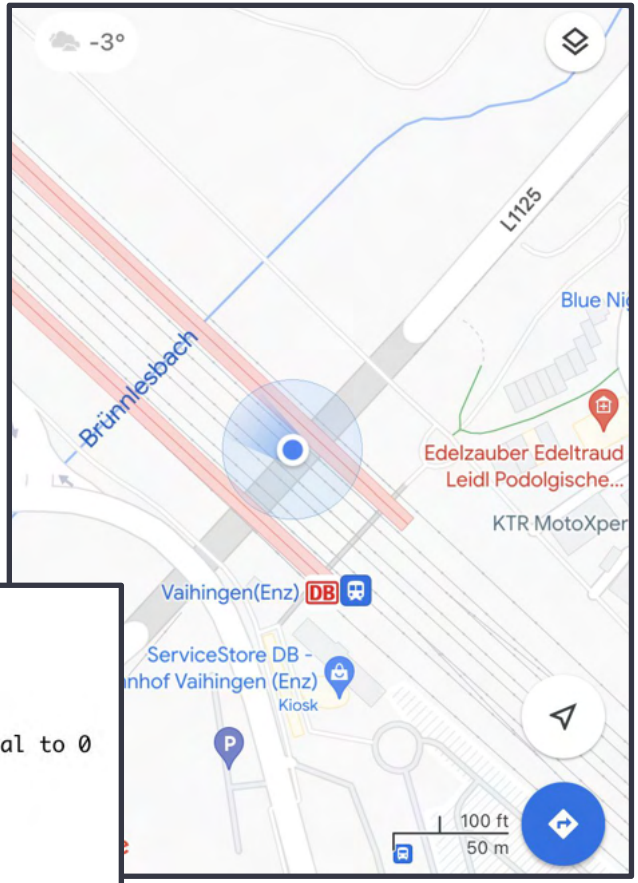
Konfidenzintervalle

- Wichtiges Konzept in der frequentistischen Inferenzstatistik
- Schätzen wir einen Parameter der Population mit einem bestimmten Verfahren, enthält ein bestimmter Anteil der Intervalle den wahren Wert.
  - 95%, 99%, 99,9%, ...
- Werfen wir wiederholt Ringe, um den echten Wert in der Population zu treffen, liegen 95%, 99%, ... der Ringe um den Wert.



Konfidenzintervalle

- Konfidenzintervalle werden in R für eine Reihe von Tests ausgegeben.
- `confint(<Modell>)` gibt Konfidenzintervalle für die Effektschätzer aus.
  - `level = .95`

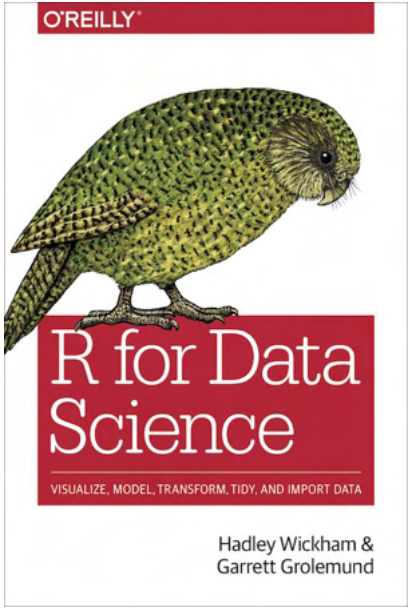


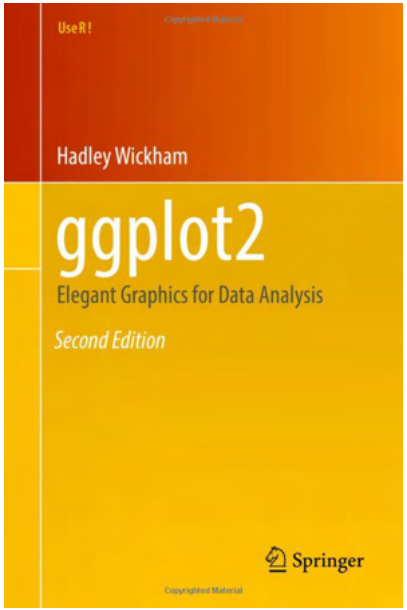

```
        Pearson's product-moment correlation

data:  lexdec$Frequency and lexdec$RT
t = -9.4587, df = 1657, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2715046 -0.1801720
sample estimates:
       cor
-0.2263358
```

Hier geht's weiter...

- Fortgeschrittene Regressionsverfahren: Gemischte Modelle
- Explorative Verfahren: Clusteranalysen, Hauptkomponentenanalyse, Multidimensionale Skalierung
- Visualisierungen mit dem Paket `{ggplot2}`
- Datenmanipulation mit dem Paket `{dplyr}`
- ...

https://r4ds.had.co.nz





https://ggplot2-book.org

Hausaufgabe

- Laden Sie aus dem Package `openintro` das Dataframe `tip` auf eine eigene Variable (dieses haben Sie auch schon in der letzten Hausaufgabe bearbeitet)
- Unterteilen Sie die Daten nach Wochentagen in zwei Gruppen (Di und Fr)
- Führen Sie einen t-test durch, um zu prüfen, ob die beiden Gruppen derselben Grundgesamtheit angehören.
- Interpretieren Sie das Ergebnis in einem Kommentar.

völlig neue Aufgabe (Jürgen Hermes)