Einführung in die Statistik

Praktische Übung – Jürgen Hermes – IDH – SoSe 2023

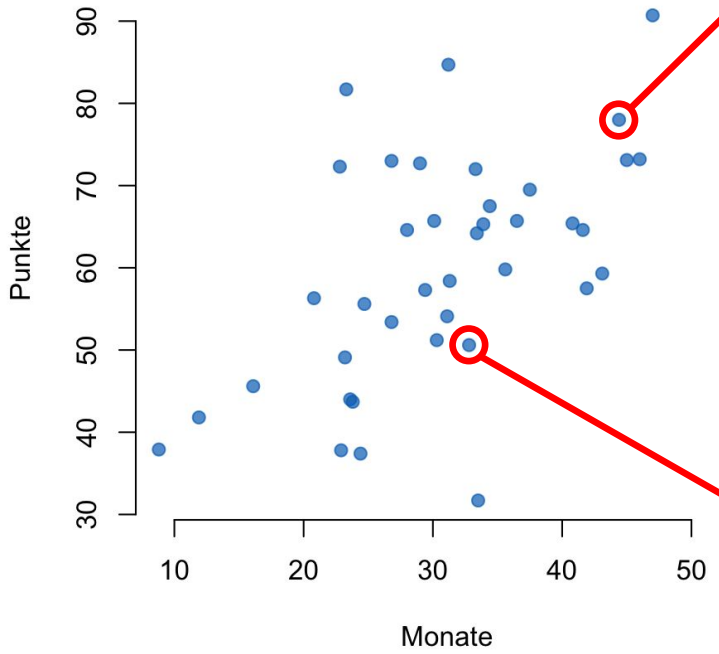
Programm heute

- Korrelation: Wiederholung
- Lineare Regression
- Multiple Regression (eher nicht Testat-Relevant)
- Binär-Logistische Regression (eher nicht Testat-Relevant)
- Zusammenfassung / Hausaufgabe

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

Regression

Daten

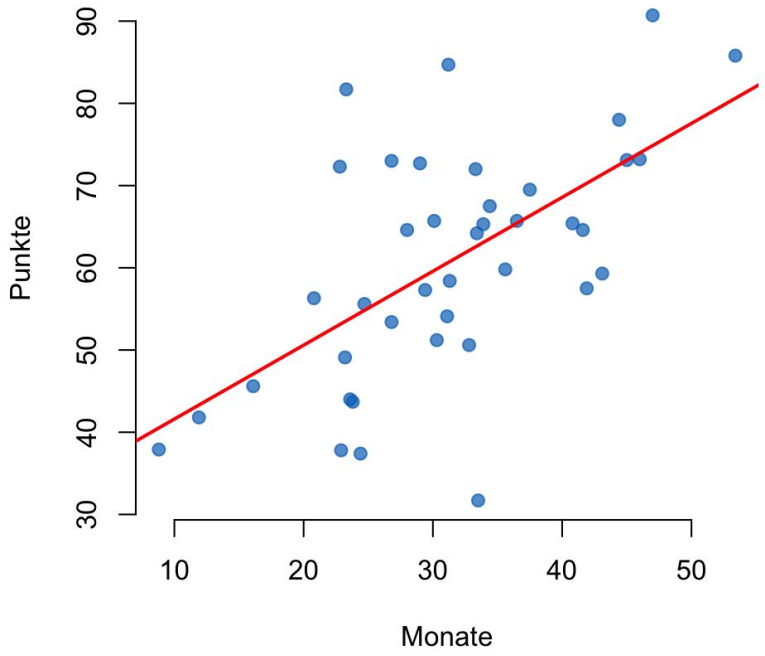


Andere Person, die seit 44,4 Monaten in der Schweiz lebt und im Test 78 Punkte erreicht hat.

- Daten von Sprachlernenden
  - Monate Aufenthalt in der Schweiz
  - Punkte in einem Sprachtest

Person, die seit 32,8 Monaten in der Schweiz lebt und im Test 50,6 Punkte erreicht hat.

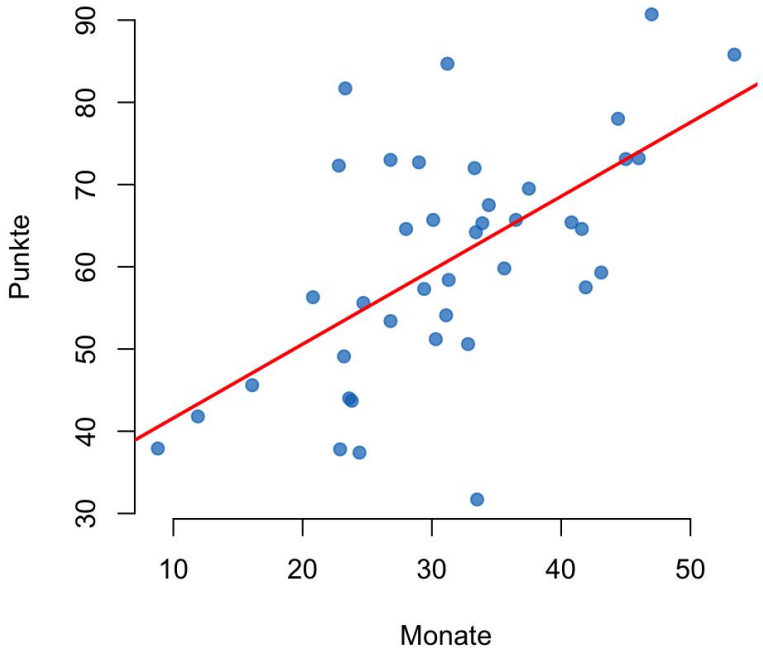
Lineare Regression



- 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

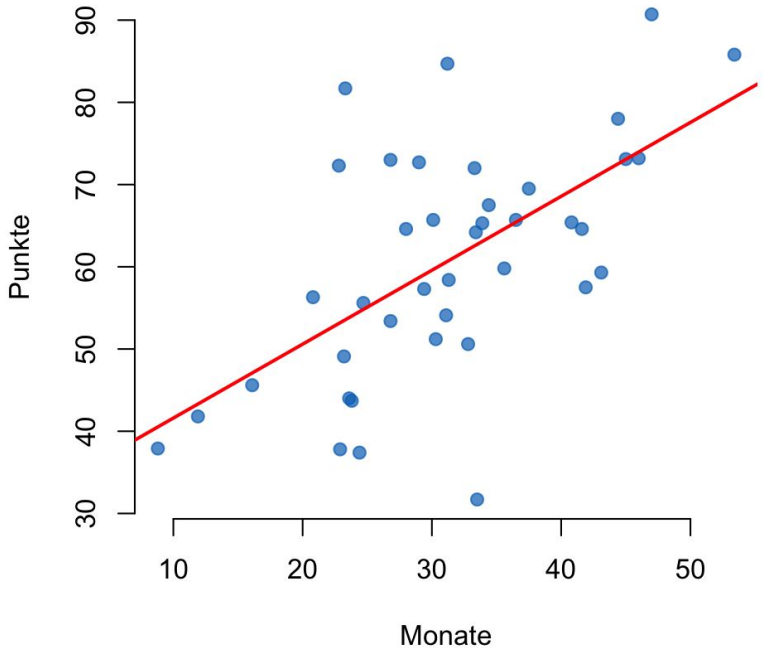
- Regressionsgeraden sind definiert durch zwei Parameter:
  - *y*-Achsenabschnitt / *Intercept* ***a***
  - Steigung / *Slope* ***b***
- *y*-Wert = Intercept + Slope • *x*-Wert
- $y = a + b \bullet x$



Punkte = 32,6 + 0,9 • Monate

Regressionsparameter

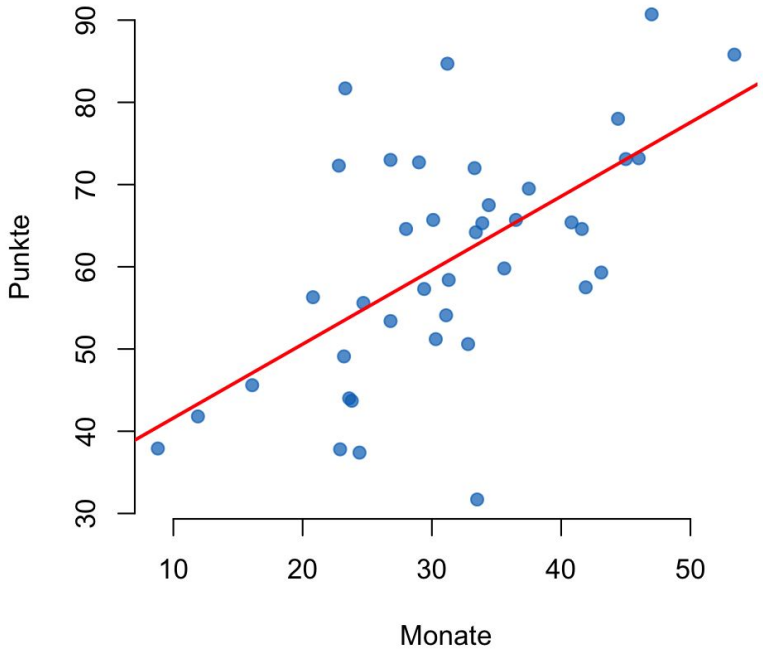
- Parameter sind interpretierbar.
  - Slope *b*: Um wie viel Punkte steigt das Testergebnis mit jedem Monat Aufenthalt in der Schweiz?
  - Intercept *a*: Wie viel Punkte erzielt man, wenn man noch nicht in der Schweiz war (Monate = 0)?
- Hier: Pro Monat ca. 1 Punkt mehr, 32,6 Punkte bei 0 Monaten.



Punkte = 32,6 + 0,9 • Monate

Regression als Vorhersage

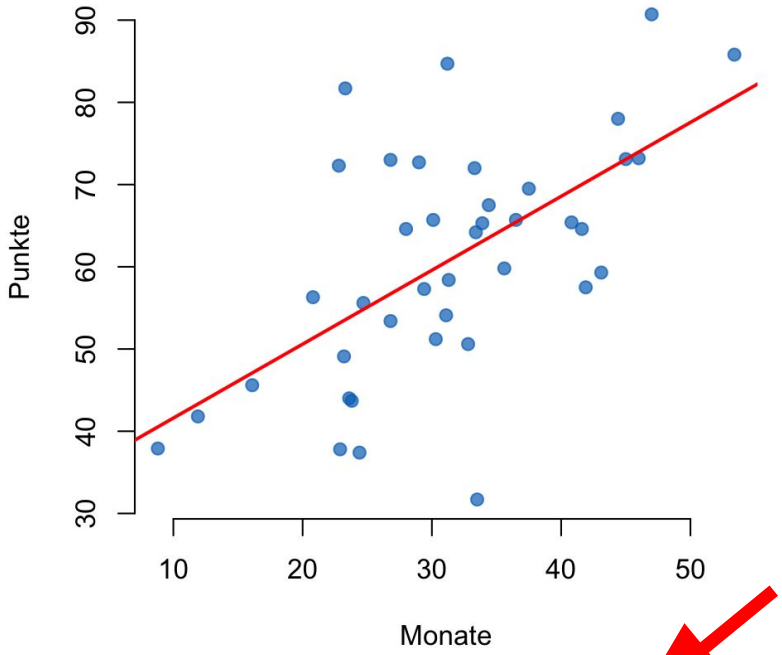
- Aufgrund der Gerade können wir bei neuen Werten von *x* vorhersagen, welchen *y*-Wert wir erwarten würden.



Punkte = 32,6 + 0,9 • Monate

Regression als Vorhersage

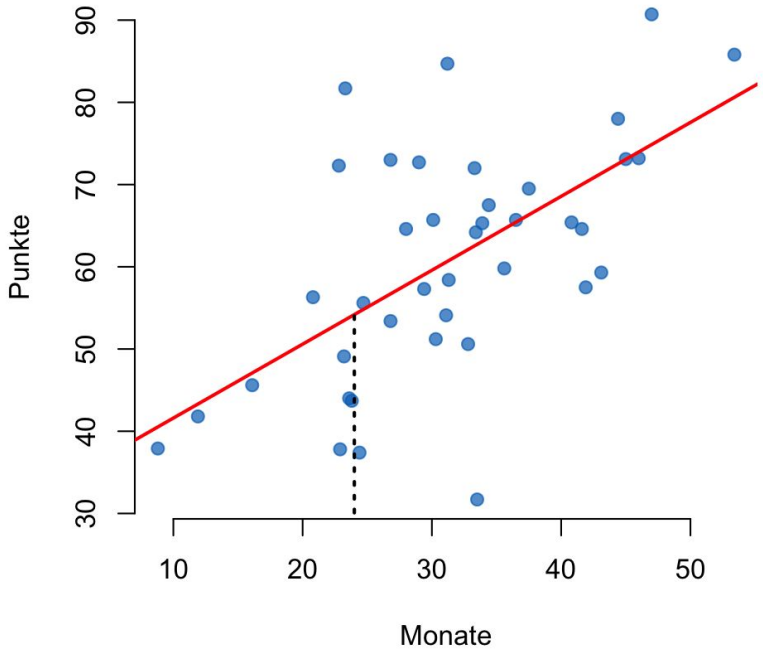
- Uns wird eine Sprachschülerin vorgestellt, die seit zwei Jahren in der Schweiz lebt.
- Wie viel Punkte wird sie wohl in dem Test erzielen?



Punkte = 32,6 + 0,9 • 24

Regression als Vorhersage

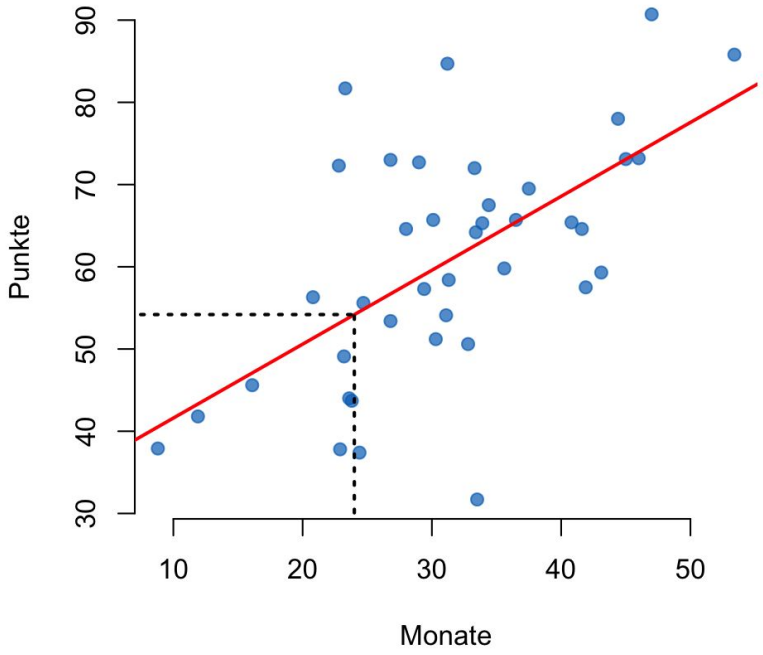
- Uns wird eine Sprachschülerin vorgestellt, die seit zwei Jahren in der Schweiz lebt.
- Wie viel Punkte wird sie wohl in dem Test erzielen?



Punkte = 32,6 + 0,9 • 24

Regression als Vorhersage

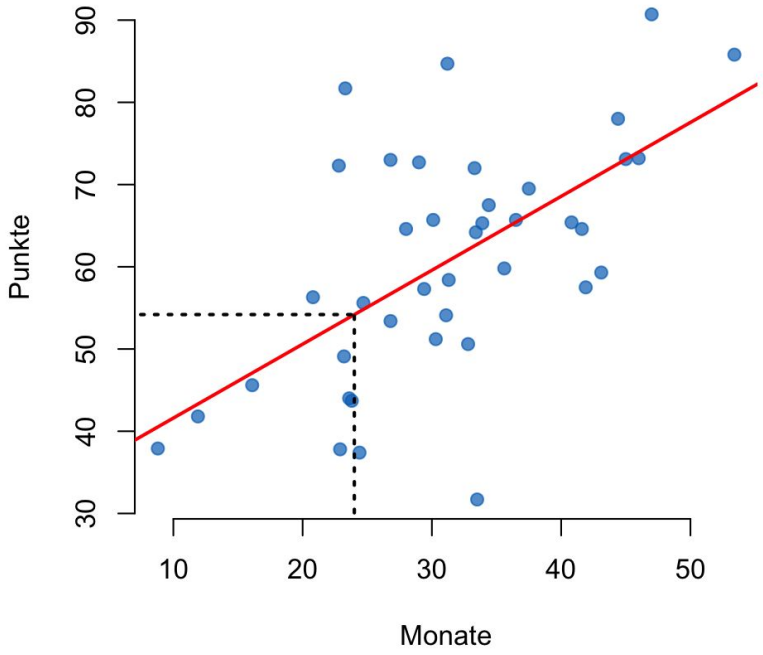
- Uns wird eine Sprachschülerin vorgestellt, die seit zwei Jahren in der Schweiz lebt.
- Wie viel Punkte wird sie wohl in dem Test erzielen?



54,2 = 32,6 + 0,9 • 24

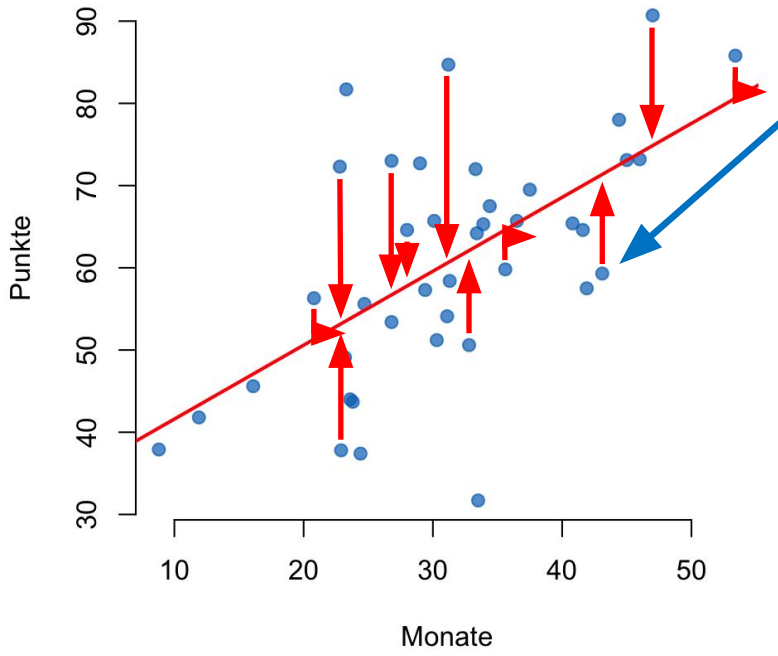
Regression als Vorhersage

- 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 • 24

Residuen (= Vorhersagefehler)



Hat diese Person "zu wenige" Punkte erreicht?

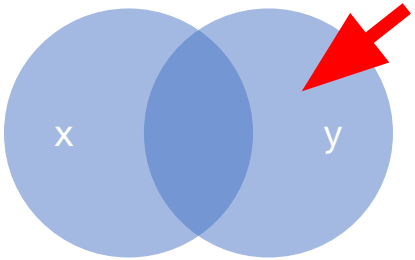
Oder ist gar unsere Regressionsgerade fehlerhaft?

- 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

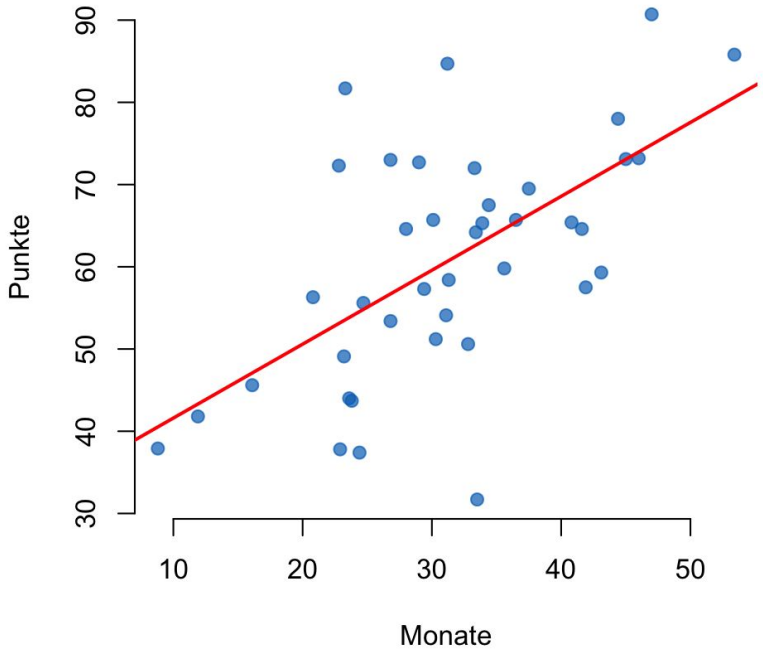
- Residuen sind das, was durch die Vorhersage von *x* auf *y* nicht erklärt werden kann.
  - Unaufgeklärte Varianz
- Was könnte das in unserem Beispiel sein?
  - Allgemeine Sprachfähigkeit
  - Kenntnis zusätzlicher Zweitsprachen
  - Kenntnis anderer germanischer Sprachen
  - Kontakthäufigkeit mit deutschsprachigen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

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

Aufruf

```
Call:
lm(formula = RT ~ NativeLanguage * Frequency, data = lexdec)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67232 -0.14728 -0.03079  0.11713  1.06986 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               6.466060   0.027796 232.626  < 2e-16 ***
NativeLanguageOther       0.286343   0.042459   6.744 2.12e-11 ***
Frequency                -0.031098   0.005651  -5.504 4.31e-08 ***
NativeLanguageOther:Frequency -0.027472   0.008631  -3.183  0.00149 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2218 on 1655 degrees of freedom
Multiple R-squared:  0.1584,	Adjusted R-squared:  0.1569 
F-statistic: 103.8 on 3 and 1655 DF,  p-value: < 2.2e-16
```

Verteilung der Residuen

Effektschätzer & stat. Prüfgröße

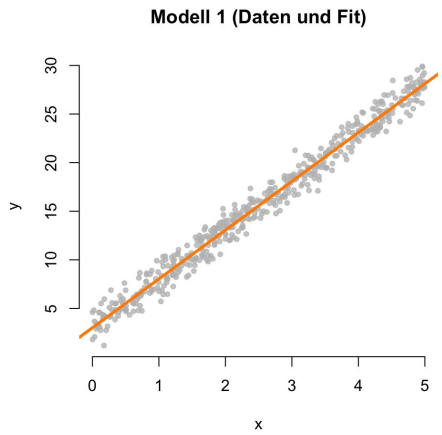
p-Werte

Prädiktoren

Varianzaufklärung

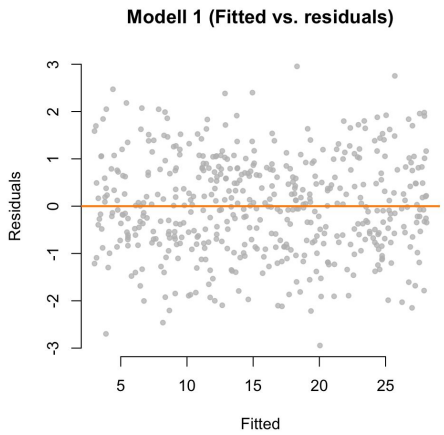
Regressionsdiagnostik

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




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

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

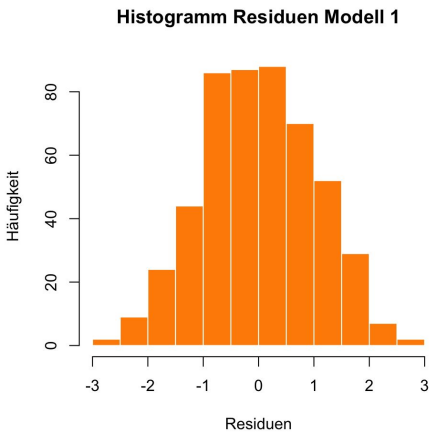


Regressionsdiagnostik

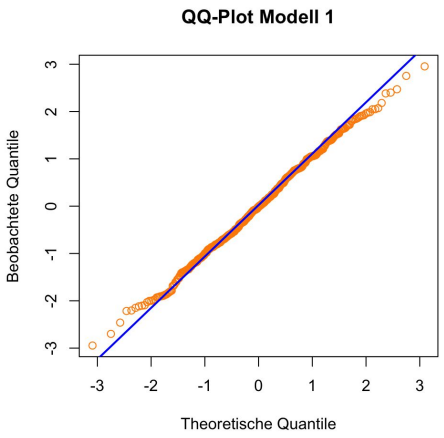
Die meisten Voraussetzungen können auch numerisch getestet werden. Hierzu z.B. https://book.stat420.org/**model-diagnostics**.html

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

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




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



Hausaufgabe

- Laden Sie aus dem Package `openintro` das Dataframe `tip` 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.

Multiple Regression

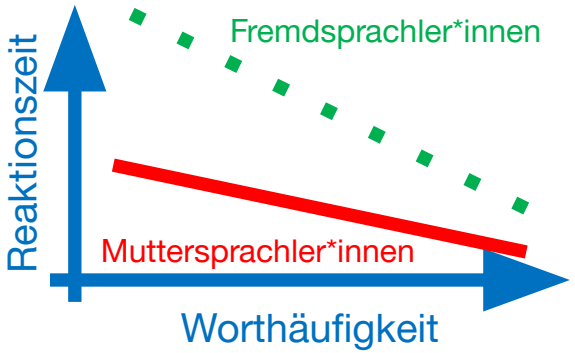
Prädiktoren werden auch **unabhängige Variablen** genannt. Die Kriteriumsvariable wird auch **abhängige Variable** genannt.

- 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
  - …

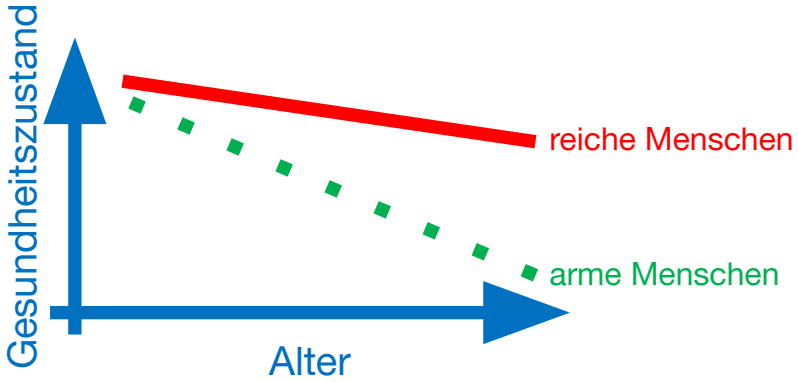
Multiple Regression

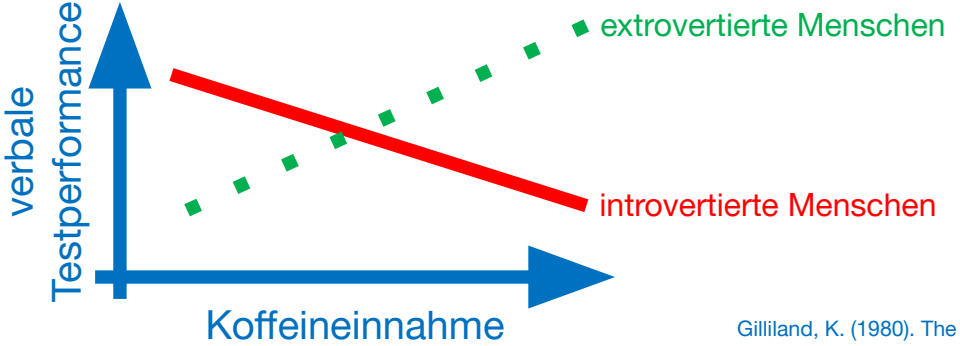
- Bei der multiplen Regression bekommt jeder Prädiktor seine eigene Steigung.
  - Auch: **(*β*-)Gewicht**, **Koeffizient**, ***coefficient***, ***estimate***
- Neben den **Einzeleffekten** (*single/main effects*) sind auch **Interaktionen** möglich.
  - Interaktion: Das Zusammenwirken von zwei oder mehr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

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.

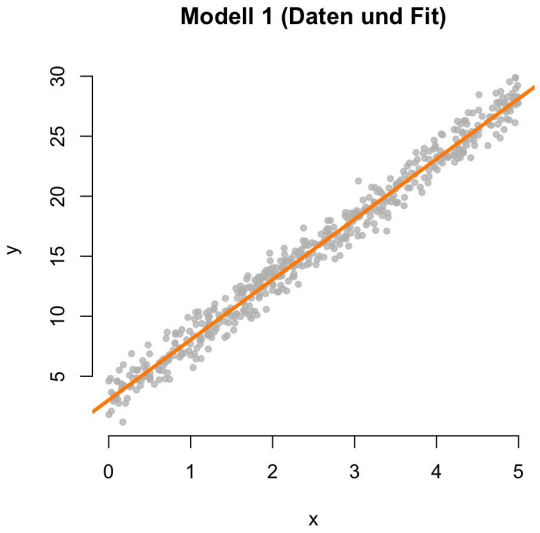
Normalverteilung
der Residuen

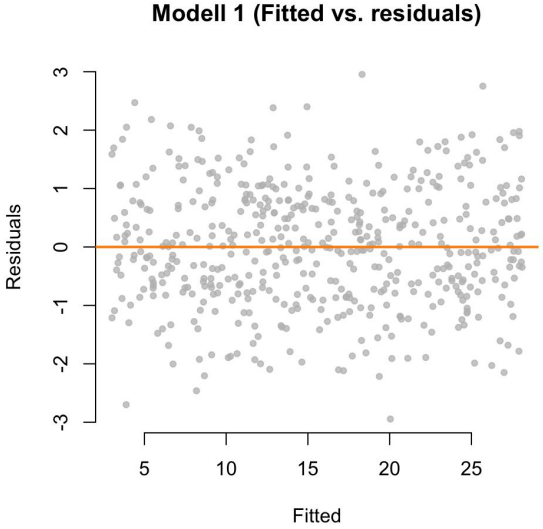
Die Residuen sind normalverteilt.

Voraussetzungen: Linearität

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

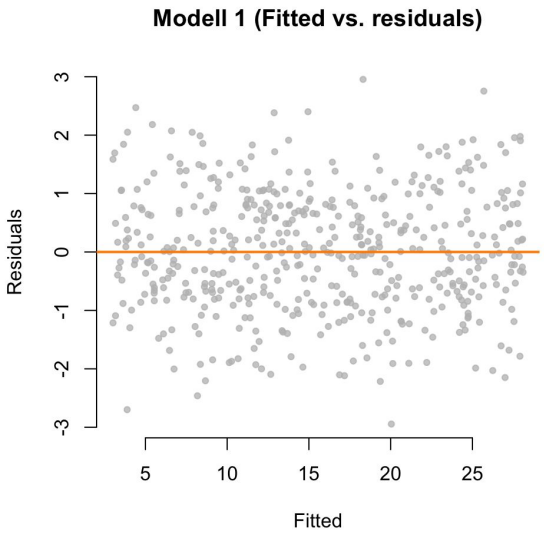
- Keine große Überraschung: Lineare Regressionen können nur lineare Zusammenhänge erfassen.
- Nützlicher Diagnostik-Plot: *Fitted vs. residuals*

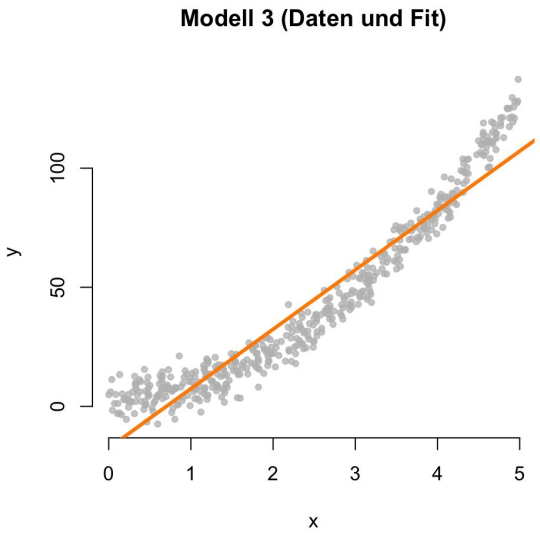


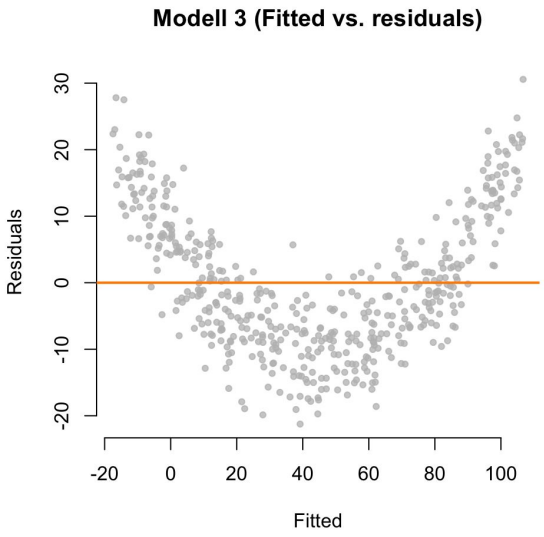


Linearität ist gegeben, wenn die Residuen sich gleichmäßig um den Fit verteilen und keine eindeutige Abweichung erkennbar ist.

Voraussetzungen: Linearität

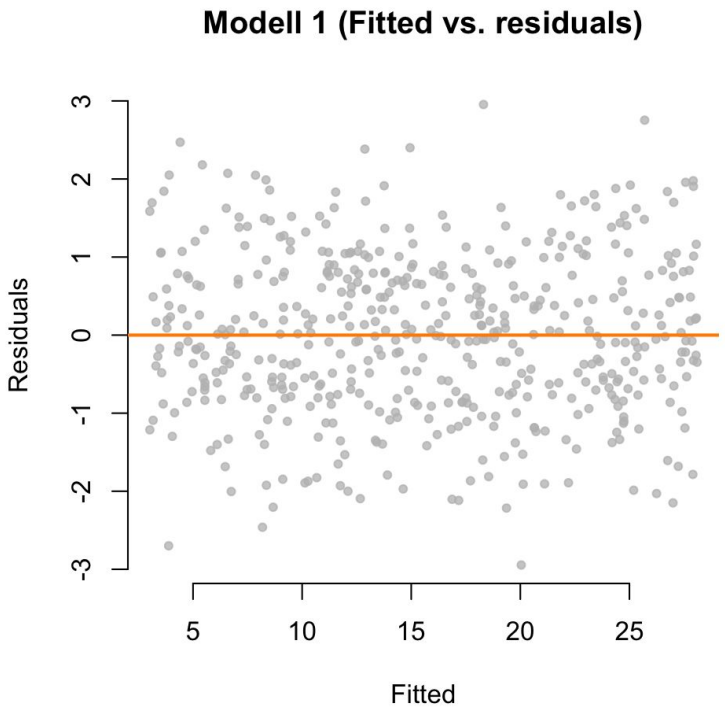




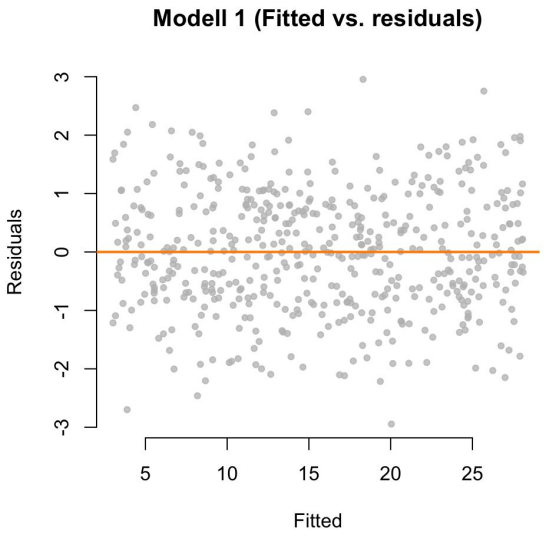


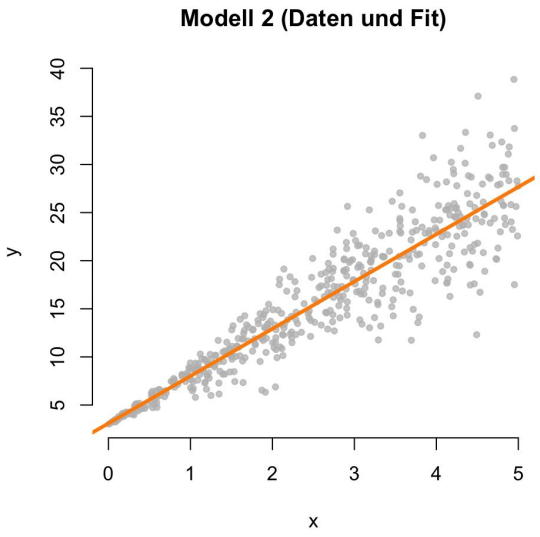
Voraussetzungen: Varianzhomogenität

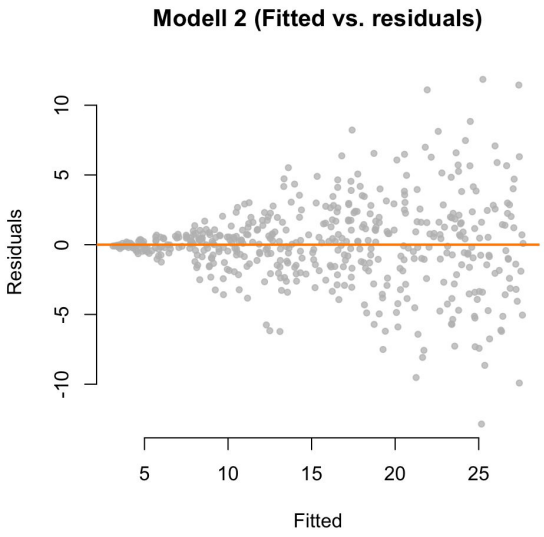
- Die Varianz der Residuen muss für alle Abschnitte des Prädiktors (der Prädiktoren) ungefähr gleich sein.
- **Heteroskedastizität** ist die Verletzung dieses Prinzips.



Voraussetzungen: Varianzhomogenität



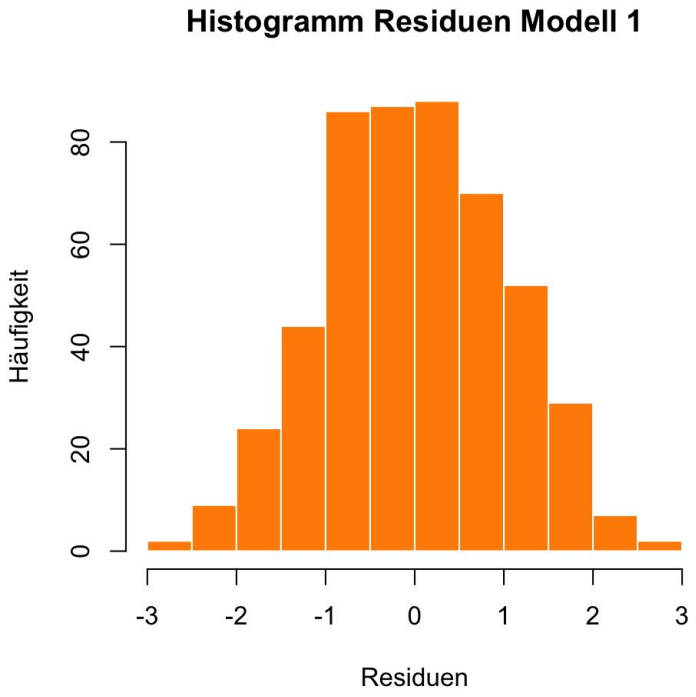


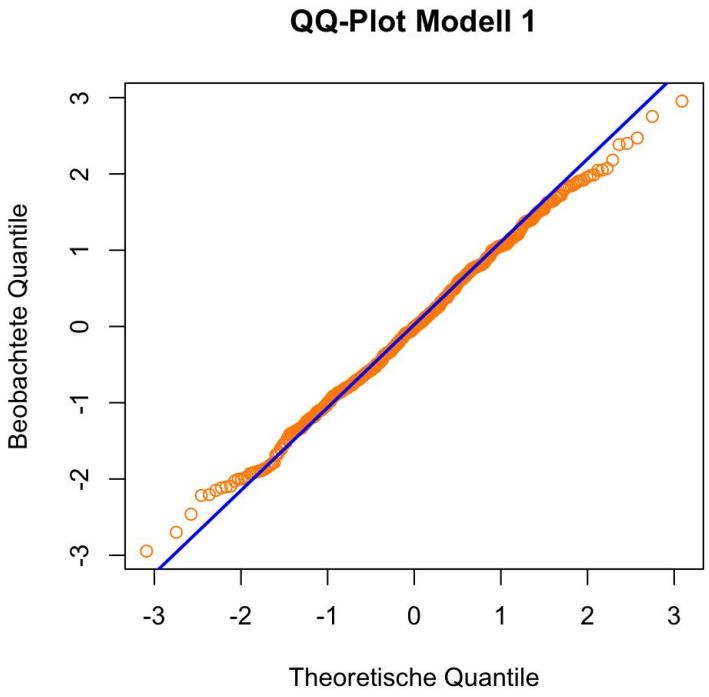


Voraussetzungen: Normalverteilung

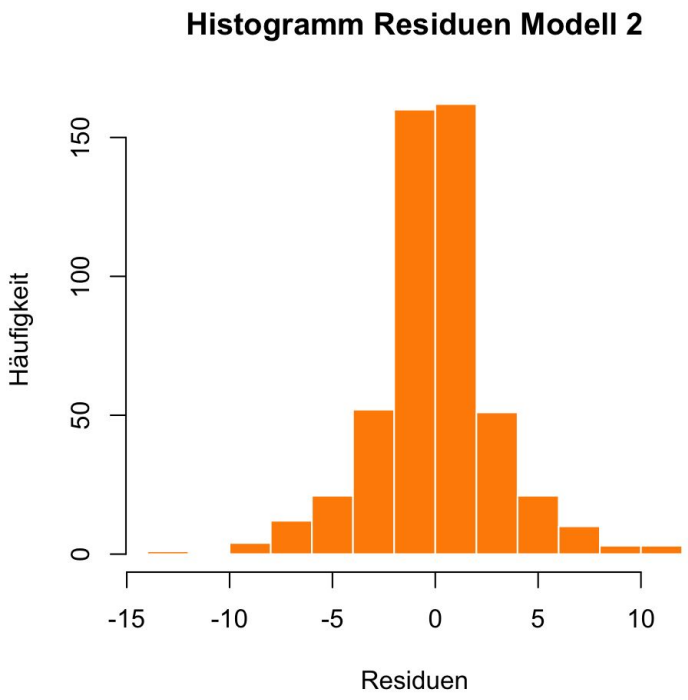
- Residuen müssen normalverteilt sein.
- Geeignete Diagnostik-Plots: Histogramm und QQ-Plot für die Residuen aus dem Regressionsmodell
- Histogramm zeigt die Anzahl an Datenpunkten in bestimmten Abschnitten (= Verteilung).
- QQ-Plot plottet theoretische Quantile (laut Normalverteilung) gegen beobachtete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

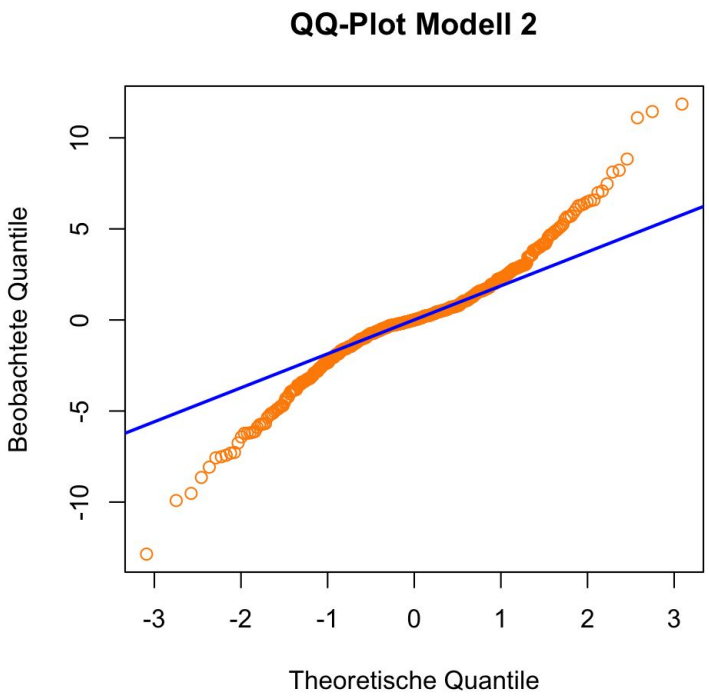
Voraussetzungen: Normalverteilung



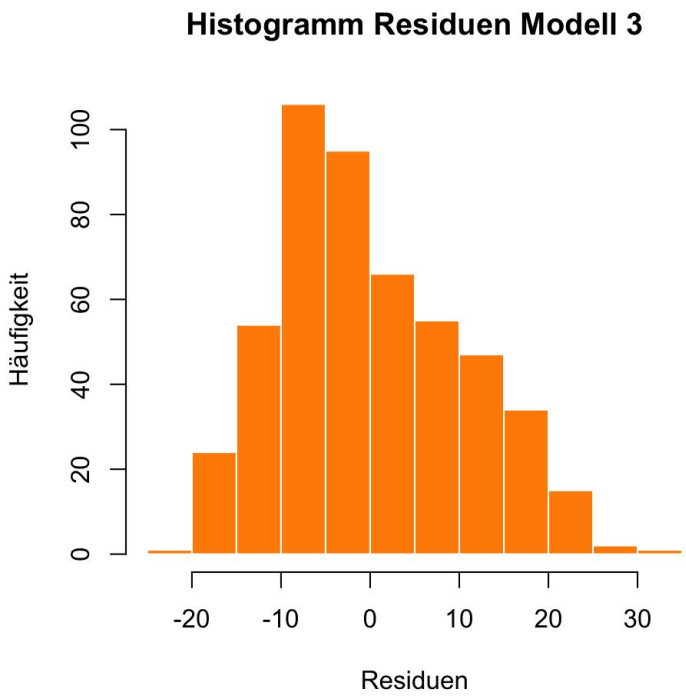


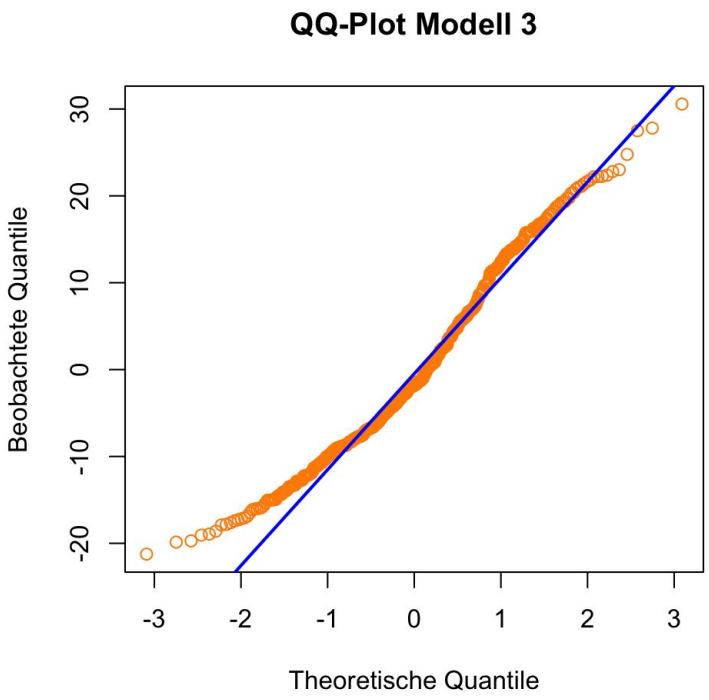
Voraussetzungen: Normalverteilung





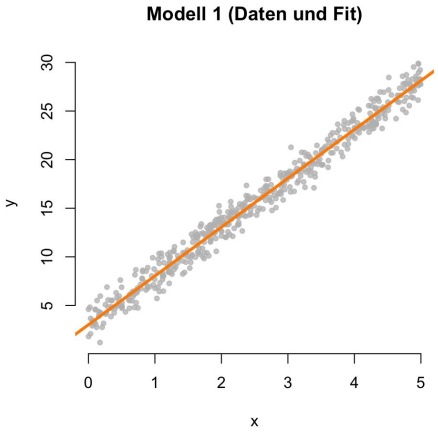
Voraussetzungen: Normalverteilung

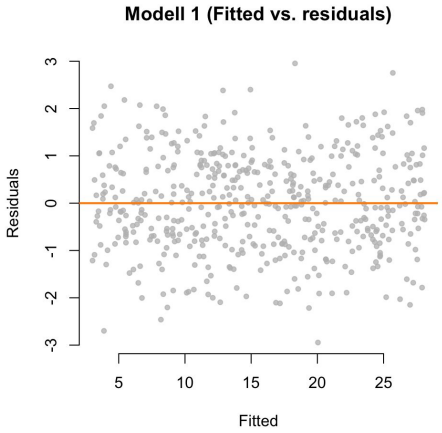


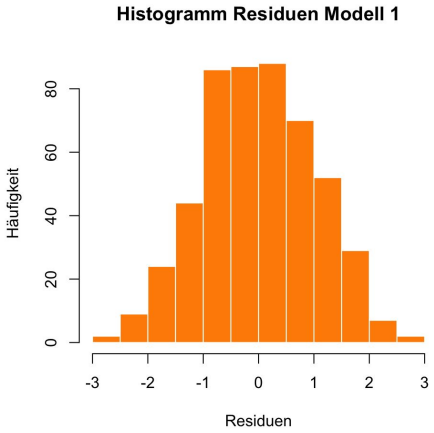


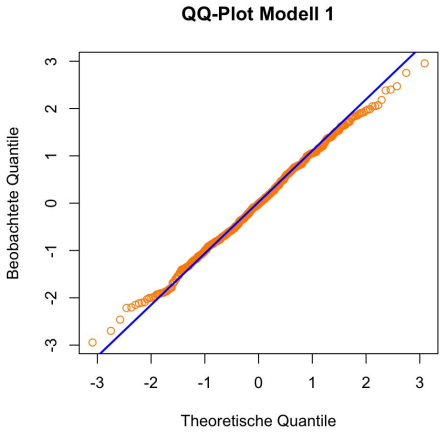
Voraussetzungen

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









Binär-logistische Regression

- Bei der logistischen Regression wird eine **binäre** Kriteriumsvariable vorhergesagt.
  - korrekt/falsch; vorhanden/nicht vorhanden; eine von zwei Realisierungen; Erfolg/Misserfolg; ja/nein
- Es darf theoretisch keine dritte Möglichkeit denkbar sein! Geschätzt wird die Wahrscheinlichkeit des Eintretens (0 bis 1).
- `glm(<Formel>, data = <Daten>, family = "binomial")`

generalized linear model

Beispiel: Binär-logistische Regression

Mary gave **[the book]**$_{\textbf{Theme}}$ to **[the man]**$_{\textbf{Recipient}}$.

Realisierung des
Rezipienten in **PP**

Mary gave **[the man]**$_{\textbf{Recipient}}$ **[the book]**$_{\textbf{Theme}}$.

Realisierung des
Rezipienten in **NP**

Frage: Beeinflussen die folgenden Prädiktoren die Realisierung des Rezipienten?

- Modalität (spoken vs. written)
- Länge des **Themes** (in Wörtern)
- Animiertheit des **Rezipienten**

Beispiel: Binär-logistische Regression

Frage: Beeinflussen die folgenden Prädiktoren die Realisierung des Rezipienten?

- Modalität (spoken vs. written)
- Länge des **Themes** (in Wörtern)
- Animiertheit des **Rezipienten**

```
library(languageR)


dative$PP_real <- dative$RealizationOfRecipient == "PP"

mod <- glm(PP_real ~ Modality + LengthOfTheme + AnimacyOfRec,
           data = dative, family = "binomial")
summary(mod)
```

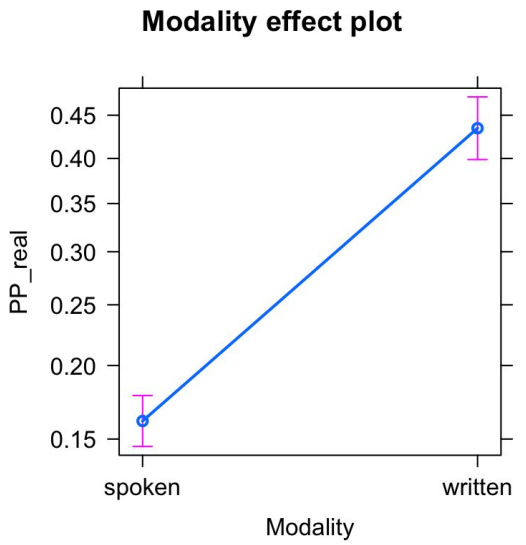
```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.72104    0.07104 -10.149  < 2e-16
***
Modalitywritten        1.38698    0.09842  14.092  < 2e-16
***
LengthOfTheme         -0.23628    0.01854 -12.745  < 2e-16
***
AnimacyOfRecinanimate  1.10800    0.14730   7.522 5.38e-14
***
```

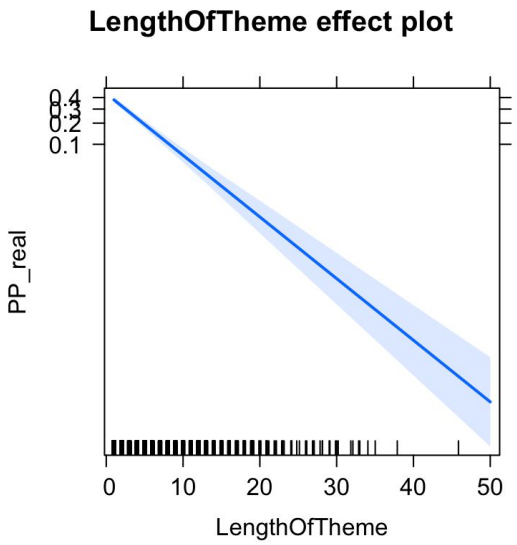
log-odds, umformbar in Wahrscheinlichkeiten

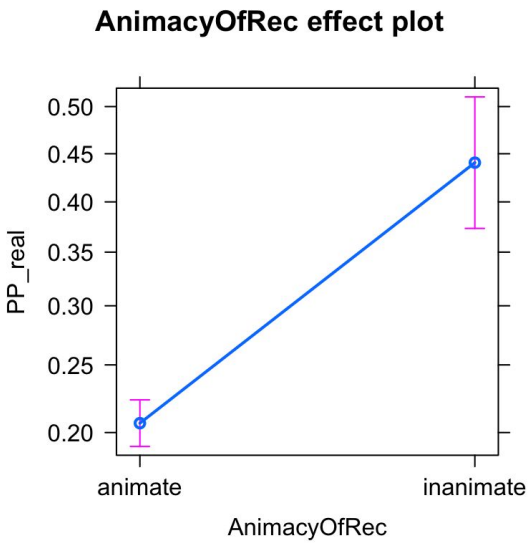
Extrahieren von Effekten: {effects}

```
mod <- glm(PP_real ~ Modality + LengthOfTheme + AnimacyOfRec,
           data = dative, family = "binomial")
library(effects)
plot(allEffects(mod))
```

```
Coefficients:
                        Estimate
(Intercept)             -0.72104
Modalitywritten          1.38698
LengthOfTheme           -0.23628
AnimacyOfRecinanimate    1.10800
```





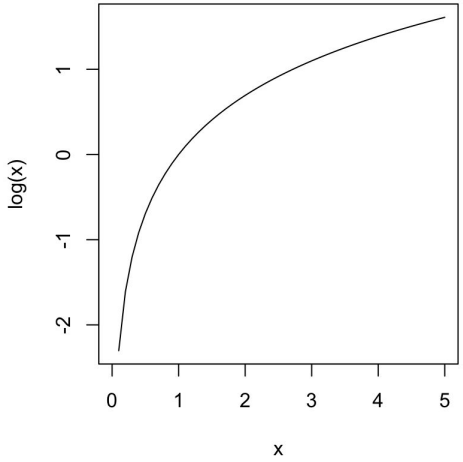


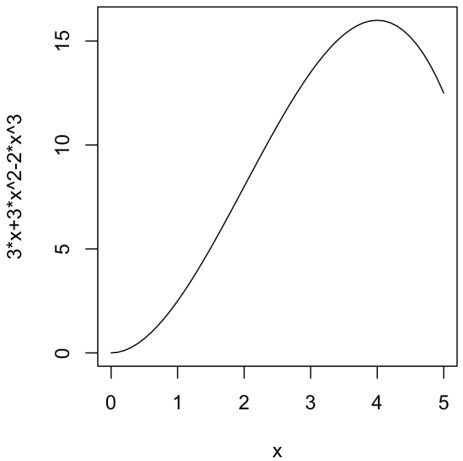
Nicht-lineare Regression

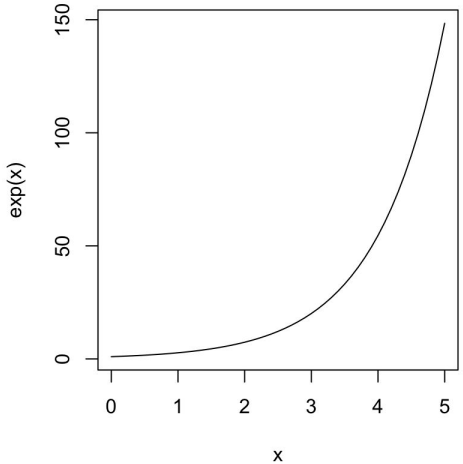
- Fitten eines nicht-linearen Zusammenhangs zwischen Prädiktoren und Kriterium
  - Logarithmische Funktionen `log()`
  - Exponentialfunktion `exp()`
  - Polynome `poly()`
  - …

```
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
- 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"; **+** *single effect*; ***** *single effect* & Interaktion
- `lm()` / `glm(..., family = "binomial")`
- Nicht-lineare Regressionen: Modellierung nicht-linearer Zshge.

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.