Statistik II

Einheit 6: Multiple Regression

Wintersemester 2025 | Prof. Dr. Stephan Goerigk

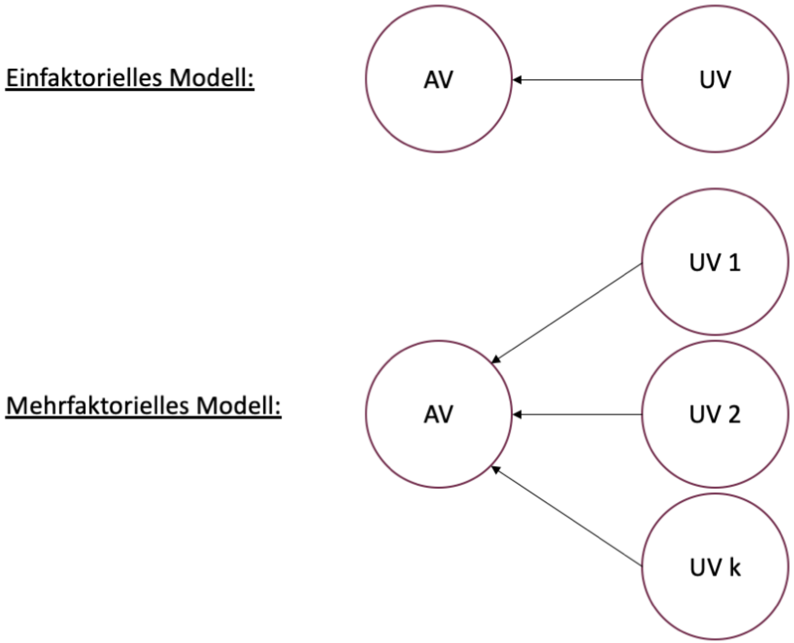
Multiple Regression

Vorbemerkungen

- multiple Regression: das Regressionsmodell enthält mehr als eine UV (Prädiktor)
- Ziel: Durch Hinzunahme weiterer Prädiktoren Vorhersagen bezogen auf die AV zu verbessern

Abgrenzung zur mehrfaktoriellen ANOVA:

- Bei der ANOVA sind UVs immer kategorial (Mittelwertesvergleiche zw. Gruppen/Kategorien)
- Im Regressionsmodell können kategoriale und stetige UVs verwendet und auch kombiniert werden

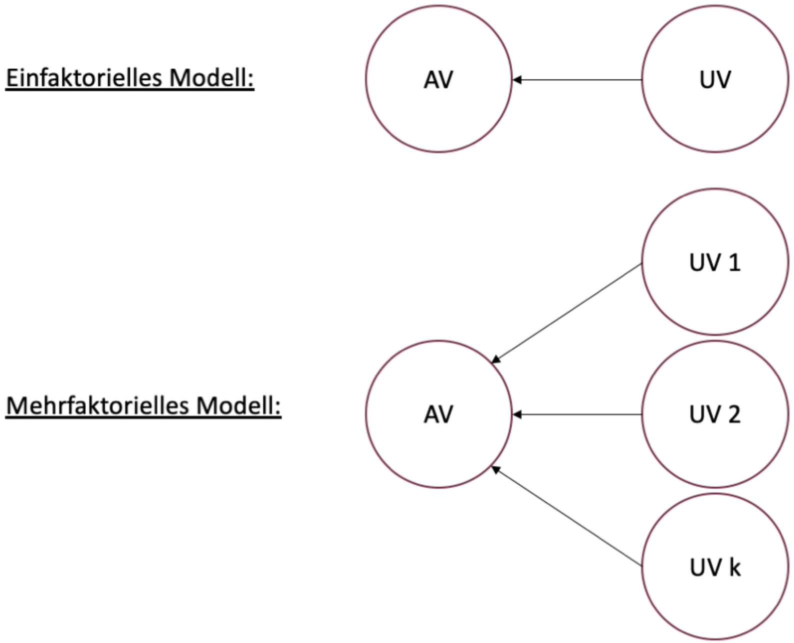


Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Weitere relevante Fragen

- Wie viel % der Gesamtvarianz der AV können die Prädiktoren **gemeinsam** erklären?
- Welcher Prädiktor hat den **größten** Vorhersagebeitrag?
- **Verändert** sich die Stärke, Richtung (und Interpretation) des Effekts eines Prädiktors, wenn weitere Prädiktoren berücksichtigt werden? (z.B. Überdeckungseffekte)



Multiple Regression

Szenario in der Vorlesung

Wir beschränken uns heute zunächst auf die einfachste Form der multiplen Regression:

- Die Beschreibung des AV-Werts Y_i durch 2 stetige Prädiktoren und die Fehlervariable.
- Hat man den Fall mit zwei Prädiktoren verstanden, ist die Generalisierung auf weitere Prädiktoren einfach.

Dies lässt sich durch die folgende **Erweiterung der Regressionsgleichung** darstellen:

$$Y_i = a + b_1 \cdot X_{i1} + b_2 \cdot X_{i2} + \epsilon_i$$

wobei:

$$\epsilon_i \; N(0, \sigma^2)$$

(Fehler normalverteilt mit Erwartungswert 0)

Multiple Regression

Elemente der multiplen Regressionsgleichung

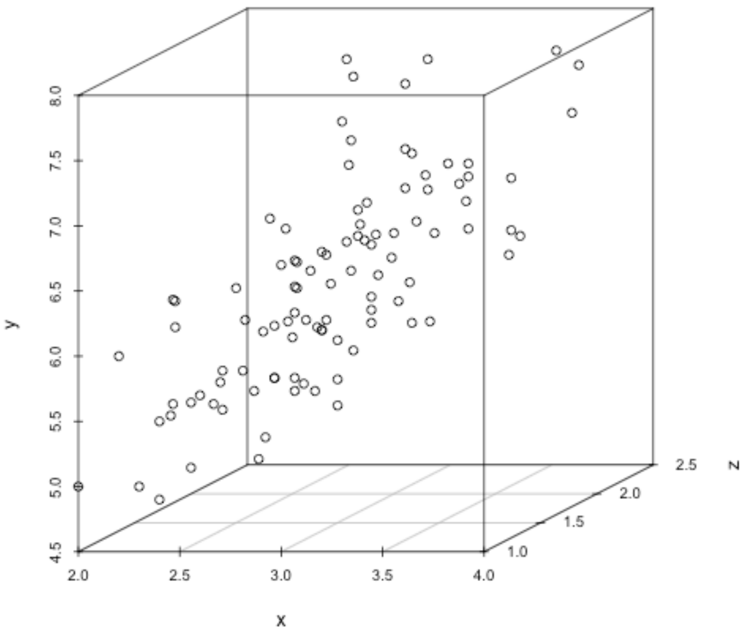
$$Y_i = a + b_1 \cdot X_{i1} + b_2 \cdot X_{i2} + \epsilon_i$$

- X_1 und X_2 sind Zufallsvariablen. Ihre Realisationen sind jeweils die Werte der zufällig gezogenen Person i bezüglich der UV_1 und der UV_2
- a, b_1, b_2 und σ^2 sind die zu schätzenden Modellparameter
    - a = Y-Achsenabschnitt
    - b_1 = Steigungsparameter der UV_1
    - b_2 = Steigungsparameter der UV_2
    - σ^2 = Varianz des Fehlerterms (für Hypothesen meist inhaltlich nicht relevant)

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Graphische Darstellung



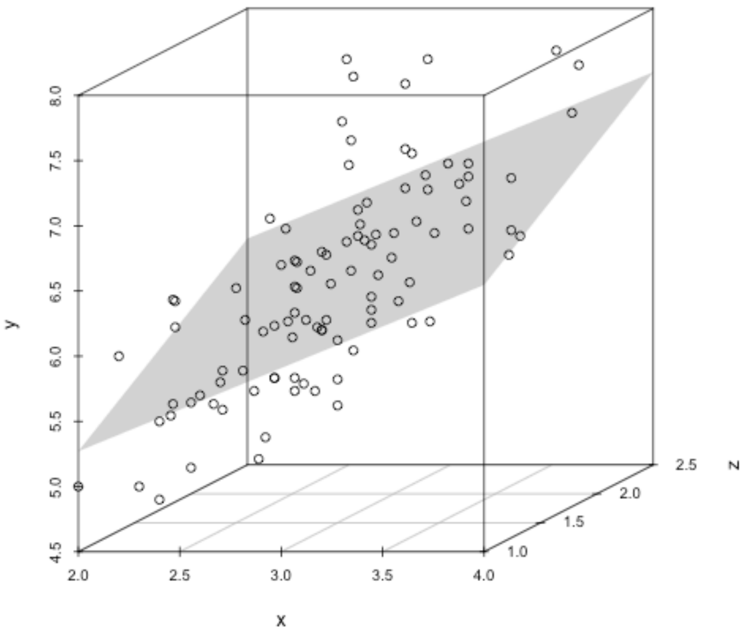
- Einfache lineare Regression: 2-dimensionales Koordinatensystem mit X-Achse und Y-Achse
- Mutiple Regression (2 UVs): 3-dimensionales Koordinatensystem mit X-Achse, Y-Achse und Z-Achse

$\rightarrow$ Es wird ein 3D-Streudiagramm dargestellt

- Punkt = Beobachtungswert einer Person
- Kombination aus AV (Y-Achse), UV_1 (X-Achse) und UV_2 (Z-Achse) Wert

Multiple Regression

Graphische Darstellung



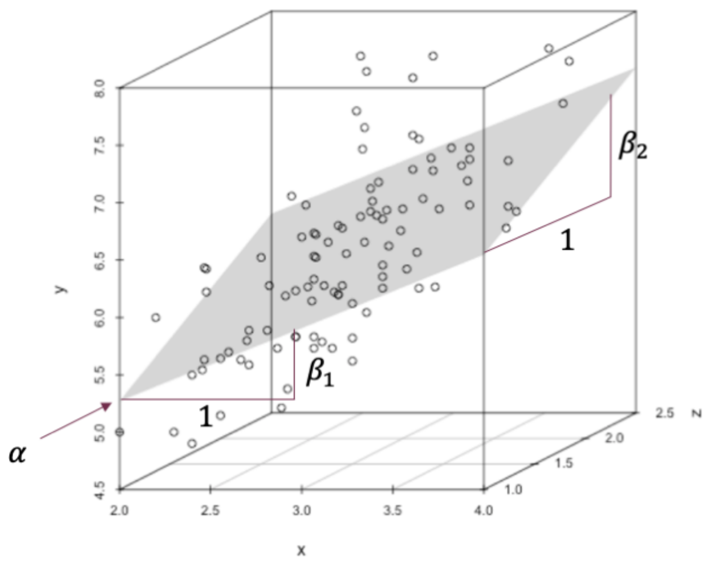
- Einfache lineare Regression: Modellfunktion dargestellt durch Regressionsgerade
    - Gerade definiert durch 1 Y-Achsenabschnitt + 1 Steigungsparameter
- Mutiple Regression (2 UVs): Modellfunktion dargestellt durch Regressionsebene
    - Ebene definiert durch 1 Y-Achsenabschnitt + 2 Steigungsparameter
- Auf der Ebene liegen alle durch das Modell erwarteten Werte

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Graphische Darstellung

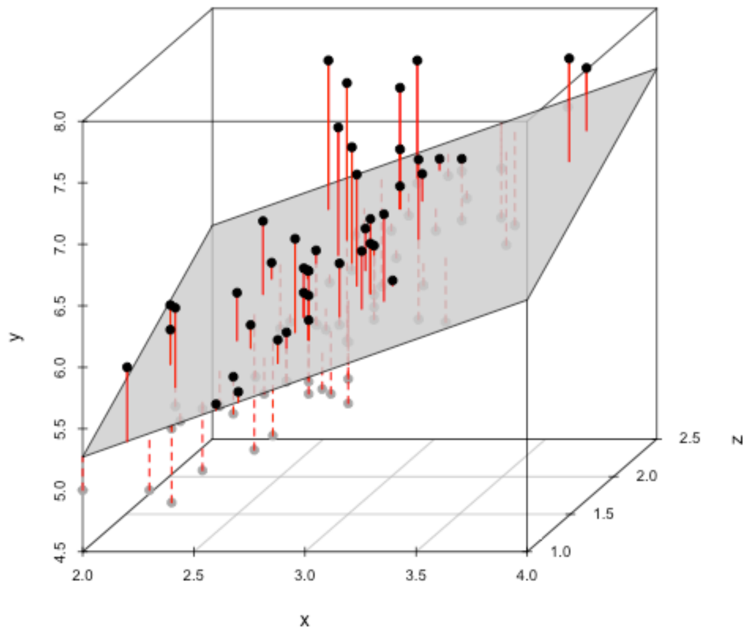
$$Y_i = a + b_1 \cdot X_{i1} + b_2 \cdot X_{i2} + \epsilon_i$$



- a gibt den Y-Achsenabschnitt an
  - a ist der Wert der AV, wenn UV_1 und UV_2 gleich 0 sind
  - $a = a + b_1 \cdot 0 + b_2 \cdot 0$
  - Ob a sinnvoll interpretiert werden kann, hängt davon ab, ob $UV_1 = 0$ und $UV_2 = 0$ inhaltlich sinnvolle Werte darstellen
- b_1 gibt an, wie stark die Regressionsebene auf der xy-Gerade steigt bzw. fällt, wenn UV_1 um 1 Einheit zunimmt.
- b_2 gibt an, wie stark die Regressionsebene auf der zy-Gerade steigt bzw. fällt, wenn UV_2 um 1 Einheit zunimmt.

Multiple Regression

Parameterschätzung



- Die unbekannten Modellparameter a, b_1 und b_2 können mit der **Methode der kleinsten Quadrate** bestimmt werden (wie bei einfacher Regression)
- Die Ebene wird so definiert, dass die Residuen minimiert werden
- Die Formeln sind aufwendig, weswegen wir uns hier auf die Berechnung in R beschränken

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Standardfehler der Modellparameter

- Während wir die Schätzung der Modellparameter R überlassen, schauen wir uns einmal die Berechnung der Standardfehler für b_1 und b_2 an.
- Diese brauchen wir, um Hypothesentests/Konfidenzintervalle für diese Parameter zu berechnen

$$SE(B_1) = \sqrt{Var(B_1)} = \sqrt{\frac{1}{1 - r^2_{x1x2}} \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}}$$

$$SE(B_2) = \sqrt{Var(B_2)} = \sqrt{\frac{1}{1 - r^2_{x1x2}} \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}}$$

- r^2_{x1x2} stellt die quadrierte Korrelation zwischen den beiden Prädiktoren dar
- σ^2 wird durch die Stichprobenvarianz s^2 geschätzt.

Multiple Regression

Konfidenzintervalle der Modellparameter

Die Konfidenzintervalle für b_1 und b_2 lassen sich wie folgt berechnen:

$$b \pm t_{1-\frac{\alpha}{2}} \cdot SE(B_j)$$

- Die Freiheitsgerade für den t-Wert errechnen sich als $df = n - \mathbf{3}$

Multiple Regression

Beispiel: Risikofaktoren für Aggression bei Kindern

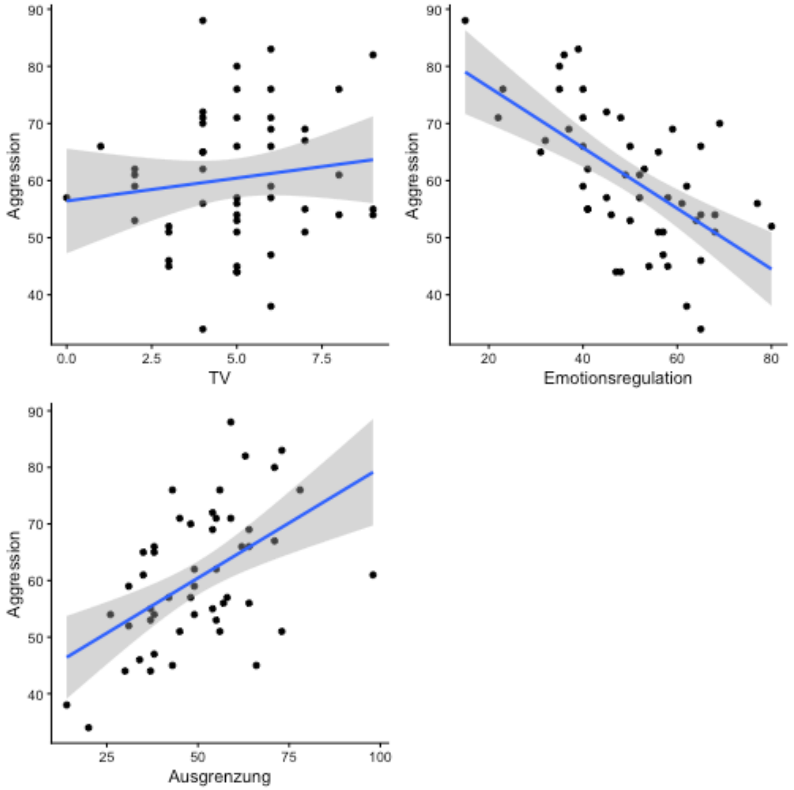
- Wissenschaftler:innen haben Daten erhoben $(N = 50)$, um Risikofaktoren für Aggression bei Kindern zu identifizieren.
- Folgende Variablen wurden gemessen
  - Aggression (AV, 1-100 Punkte)
  - TV (UV, in Stunden/Tag)
  - Emotionsregulation (UV, 1-100 Punkte)
  - Ausgrenzungserfahrung (UV, 1-100 Punkte)
- Die ersten 15 Fälle sind in der Tabelle rechts dargestellt.

Aggression	TV	Emotionsregulation	Ausgrenzung
53	5	50	55
57	5	58	58
80	5	35	71
61	8	49	35
62	4	41	49
82	9	36	63
66	1	40	62
44	5	47	30
51	5	56	56
54	5	68	38
76	6	35	78
65	4	56	38
65	4	31	35
61	2	52	98
53	2	64	37

Multiple Regression

Beispiel: Risikofaktoren für Aggression bei Kindern

- Um einen 1. Eindruck zu gewinnen, lohnt es sich, die Daten zu visualisieren
- Wir schauen uns dafür die bivariaten Streudiagramme an:



Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Modellschätzung in R

```
model = lm(Aggression ~ TV + Emotionsregulation, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4040  -6.2847   0.7681   7.6023  19.7061
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        86.24294    6.70484  12.863 < 0.0000000000000002 ***
## TV                  0.11254    0.68819   0.164                0.871
## Emotionsregulation -0.52752    0.09961  -5.296           0.00000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.604 on 47 degrees of freedom
## Multiple R-squared:  0.3852,    Adjusted R-squared:  0.3591
## F-statistic: 14.73 on 2 and 47 DF,  p-value: 0.00001083
```

Die Schätzwerte für a, b_1 und b_2 können in der Spalte `Estimate` abgelesen werden

Interpretation:

- Der durchschnittliche Aggressionswert eines Kindes, das 0h TV sieht und einen Emotionsregulationsscore von 0 hat ist $a = 86.24$
- Mit jeder zusätzlichen Stunde TV nimmt der Aggressionswert um $b_1 = 0.11$ Punkte zu.
- Mit jedem zusätzlichen Punkt auf der Emotionsregulationsskale nimmt der Aggressionswert um $b_2 = -0.53$ Punkte ab.

Multiple Regression

Schätzung der unbekannten Fehlervarianz σ^2

Die Schätzfunktion für die unbekannte Fehlervarianz lässt sich darstellen als

$$\hat{\sigma}^2 = \frac{1}{n-3}\sum_{i=1}^{n}(Y_i - (A + B_1 \cdot X_{i2}))^2$$

- Nach Umstellen und ziehen der Wurzel erhält man den Standardschätzfehler (wie in der einfachen Regression):

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-3}}$$

$\rightarrow$ Dies entspricht der Wurzel aus der Summe der quadrierten Residuen geteilt durch $n - 3$

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Schätzung der unbekannten Fehlervarianz σ^2

```r
model = lm(Aggression ~ TV + Emotionsregulation, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4040  -6.2847   0.7681   7.6023  19.7061
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        86.24294    6.70484  12.863 < 0.0000000000000002 ***
## TV                  0.11254    0.68819   0.164                0.871
## Emotionsregulation -0.52752    0.09961  -5.296           0.00000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.604 on 47 degrees of freedom
## Multiple R-squared:  0.3852,    Adjusted R-squared:  0.3591
## F-statistic: 14.73 on 2 and 47 DF,  p-value: 0.00001083
```

Dieser Wert findet sich im unteren Bereich des R Outputs:

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-3}} = 9.60$$

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Aufstellen der Modellgleichung

```
model = lm(Aggression ~ TV + Emotionsregulation, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4040  -6.2847   0.7681   7.6023  19.7061
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        86.24294    6.70484  12.863 < 0.0000000000000002 ***
## TV                  0.11254    0.68819   0.164                0.871
## Emotionsregulation -0.52752    0.09961  -5.296           0.00000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.604 on 47 degrees of freedom
## Multiple R-squared:  0.3852,    Adjusted R-squared:  0.3591
## F-statistic: 14.73 on 2 and 47 DF,  p-value: 0.00001083
```

Allgemeine Form:

$$Y_i = a + b_1 \cdot X_{i1} + b_2 \cdot X_{i2} + \epsilon_i$$

mit ϵ_i $N(0, \sigma^2)$

In unserem Fall ergibt sich die Modellgleichung:

$$Y_i = 86.24 + 0.11 \cdot X_{i1} + -0.53 \cdot X_{i2} + \epsilon_i$$

mit ϵ_i $N(0, 9.60^2)$

$\rightarrow$ Damit ließe sich ein konkreter Wert Y_i der AV schätzen.

Multiple Regression

Hypothesentests

Je nach konkreter Fragestellung muss entschieden werden, welche Parameter geschätzt werden sollen bzw. welche Hypothesen getestet werden sollen.

Wir besprechen (zunächst) zwei Arten von Hypothesentests für die multiple Regression:

1. Hypothesentests für einzelne Modellparameter (z.B. eine Steigung)
    - $H_0 : b_j = 0$
    - Geeignet für Zusammenhangshypothesen
    - keine Steigung = kein Zusammenhang (UV kann AV nicht systematisch vorhersagen)

2. Omnibus Tests
    - basieren auf Vergleich der Varianzaufklärung (wie ANOVA)
    - prüfen Signifikanz des Gesamtmodells (H_0 : alle Steigungen sind 0)
    - erlauben Vergleich von Teilmodellen (z.B. Modell mit weiterem Prädiktor vs. Modell ohne weiteren Prädiktor)

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Hypothesentests für einzelne Modellparameter

```
model = lm(Aggression ~ TV + Emotionsregulation, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4040  -6.2847   0.7681   7.6023  19.7061
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        86.24294    6.70484  12.863 < 0.0000000000000002 ***
## TV                  0.11254    0.68819   0.164                0.871
## Emotionsregulation -0.52752    0.09961  -5.296           0.00000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.604 on 47 degrees of freedom
## Multiple R-squared:  0.3852,    Adjusted R-squared:  0.3591
## F-statistic: 14.73 on 2 and 47 DF,  p-value: 0.00001083
```

1. Hypothesentest Y-Achsenabschnitt:
   - $H_0 : a = 0$
   - $H_1 : a \neq 0$
2. Hypothesentest Steigung (TV):
   - $H_0 : b_1 = 0$
   - $H_1 : b_1 \neq 0$
3. Hypothesentest Steigung (Emotionsregulation):
   - $H_0 : b_2 = 0$
   - $H_1 : b_2 \neq 0$

Unser Beispiel: Es soll überprüft werden, ob TV-Sehen bzw. Emotionsregulation linear mit Aggression zusammenhängt, wenn der jeweils andere Prädiktor konstant gehalten wird.

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Hypothesentests für einzelne Modellparameter

```
model = lm(Aggression ~ TV + Emotionsregulation, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4040  -6.2847   0.7681   7.6023  19.7061
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        86.24294    6.70484  12.863 < 0.0000000000000002 ***
## TV                  0.11254    0.68819   0.164                0.871
## Emotionsregulation -0.52752    0.09961  -5.296           0.00000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.604 on 47 degrees of freedom
## Multiple R-squared:  0.3852,    Adjusted R-squared:  0.3591
## F-statistic: 14.73 on 2 and 47 DF,  p-value: 0.00001083
```

$$t_a = \frac{a}{SE(a)} = \frac{86.24}{6.70} = 12.86$$

$$t_{b_1} = \frac{b_1}{SE(b_1)} = \frac{0.11}{0.69} = 0.16$$

$$t_{b_2} = \frac{b_2}{SE(b_2)} = \frac{-0.53}{0.10} = -5.30$$

- Unter der Geltung der H_0 folgen diese Teststatistiken jeweils einer t-Verteilung mit $df = n - 3$
- Der kritische Bereich ist jeweils beidseitig.
- p-Werte $< .05$ zeigen signifikantes Ergebnis an (Koeffizient $\neq 0$)

Multiple Regression

Hypothesentests für einzelne Modellparameter - Konfidenzintervalle

```
model = lm(Aggression ~ TV + Emotionsregulation, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4040  -6.2847   0.7681   7.6023  19.7061
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        86.24294    6.70484  12.863 < 0.0000000000000002 ***
## TV                  0.11254    0.68819   0.164                0.871
## Emotionsregulation -0.52752    0.09961  -5.296           0.00000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.604 on 47 degrees of freedom
## Multiple R-squared:  0.3852,    Adjusted R-squared:  0.3591
## F-statistic: 14.73 on 2 and 47 DF,  p-value: 0.00001083
```

```
confint(model)
```

```
##                         2.5 %     97.5 %
## (Intercept)        72.7545432 99.7313311
## TV                 -1.2719190  1.4970047
## Emotionsregulation -0.7279233 -0.3271248
```

- KI zeigt Bereich an, in welchem der Parameter mit 95% Sicherheit liegt.
- Umschließt KI die 0 nicht (Koeffizient ungleich 0), kommt dies einem signifikanten Testergebnis gleich

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Omnibus-Test

Omnibustest des Gesamtmodells kann folgende Hypothese prüfen:

- $H_0 : b_1 = b_2 = 0$
- $H_1 : b_j \neq 0$

$\rightarrow$ Mithilfe des Omnibus-Tests kann überprüft werden, ob bei zumindest einer der UVs der lineare Zusammenhang mit der AV ungleich 0 ist (bei Konstanthaltung der jeweils anderen UV).

Anders gesagt:

- Prüfung, ob Modell mit Prädiktoren signifikant mehr Varianz der AV erklärt als ohne.
- Es werden Varianzen verwendet $\rightarrow$ Teststatistik ist wieder der von der ANOVA bekannte F-Wert

Multiple Regression

Omnibus-Test

Die Teststatistik des Omnibus-Tests ist wie folgt definiert:

$$F = \frac{\frac{1}{2}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\frac{1}{n-3}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} = \frac{\frac{1}{2}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{s^2}$$

- Unter der Geltung der Nullhypothese folgt diese Teststatistik einer F-Verteilung.
- Der kritische Bereich liegt auf der rechten Seite.

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Omnibus-Test

```
model = lm(Aggression ~ TV + Emotionsregulation, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4040  -6.2847   0.7681   7.6023  19.7061
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        86.24294    6.70484  12.863 < 0.0000000000000002 ***
## TV                  0.11254    0.68819   0.164                0.871
## Emotionsregulation -0.52752    0.09961  -5.296           0.00000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.604 on 47 degrees of freedom
## Multiple R-squared:  0.3852,    Adjusted R-squared:  0.3591
## F-statistic: 14.73 on 2 and 47 DF,  p-value: 0.00001083
```

Dieser Wert findet sich im unteren Bereich des R Outputs:

$$F = \frac{\frac{1}{2}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{s^2} = 14.73$$

mit $df_{Zähler} = 2$ und $df_{Nenner} = 47$

- $p = 0.00001083 < .05$
- Das Gesamtmodell mit den Prädiktoren erklärt signifikant mehr Varianz, als das Modell ohne Prädiktoren.

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Omnibus-Test - Modellvergleiche

- Der Omnibus-Test ermöglicht uns auch den Vergleich von 2 Modellen miteinander
- Voraussetzung ist, dass das eine Modell (komplex) das andere Modell (einfach) enhält (geschachtelte Modelle; engl.: "*nested models*").
- Dies z.B. der Fall, wenn wir zu einem bestehenden Modell einen Prädiktor hinzunehmen
- Wir probieren dies in unserem Beispiel, indem wir zu unserem Modell den Prädiktor "Ausgrenzungserfahrung" hinzunehmen
    - **Szenario 1:** Ausgrenzungserfahrung ist kein relevanter Prädiktor - Modell ohne Ausgrenzungserfahrung erklärt Daten zumindest gleich gut
    - **Szenario 2:** Modell mit Ausgrenzungserfahrung hat signifikant bessere Modellpassung (kann AV besser vorhersagen)

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Omnibus-Test - Modellvergleiche

```
# Aufstellen einfaches Modell:
model1 = lm(Aggression ~ TV + Emotionsregulation, data = df)
# Aufstellen komplexes Modell:
model2 = lm(Aggression ~ TV + Emotionsregulation + Ausgrenzung, data = df)
# Das einfache Modell ist in das komplexe Modell "geschachtelt"
```

```
# Omnibus-Test zum Vergleich "geschachtelter" Modelle
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: Aggression ~ TV + Emotionsregulation
## Model 2: Aggression ~ TV + Emotionsregulation + Ausgrenzung
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     47 4335.5
## 2     46 3508.3  1     827.2 10.846 0.001909 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Das komplexe Modell (inklusive Prädiktor Ausgrenzung) kann AV signifikant besser vorhersagen als das einfache Modell $(p = 0.001909 < .05)$

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Hinzunahme weiterer Prädiktoren

```r
# Aufstellen einfaches Modell:

model1 = lm(Aggression ~ TV + Emotionsregulation, data = df)
summary(model1)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4040  -6.2847   0.7681   7.6023  19.7061
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        86.24294    6.70484  12.863 < 0.0000000000000002 ***
## TV                  0.11254    0.68819   0.164                0.871
## Emotionsregulation -0.52752    0.09961  -5.296           0.00000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.604 on 47 degrees of freedom
## Multiple R-squared:  0.3852,    Adjusted R-squared:  0.3591
## F-statistic: 14.73 on 2 and 47 DF,  p-value: 0.00001083
```

```r
# Aufstellen komplexes Modell:

model2 = lm(Aggression ~ TV + Emotionsregulation + Ausgrenzung, data = df)
summary(model2)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation + Ausgrenzung,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6668  -7.6041   0.2665   6.8444  18.4565
##
## Coefficients:
##                    Estimate Std. Error t value      Pr(>|t|)
## (Intercept)        65.59233    8.74569   7.500 0.00000000163 ***
## TV                  0.40650    0.63209   0.643       0.52335
## Emotionsregulation -0.41784    0.09651  -4.330 0.00007997688 ***
## Ausgrenzung         0.27409    0.08322   3.293       0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.733 on 46 degrees of freedom
## Multiple R-squared:  0.5025,    Adjusted R-squared:  0.4701
## F-statistic: 15.49 on 3 and 46 DF,  p-value: 0.0000004219
```

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Hinzunahme weiterer Prädiktoren

Was passiert?

- Weiterer Prädiktor wird an Modell "drangehängt"
- Er erhält ebenfalls einen Steigungsparameter, dieser erhält einen Signifikanztest
  - Mit weiterem Punkt Ausgrenzungserfahrung nimmt Aggression um $b_3 = 0.27$ Punkte zu
  - Ausgrenzungserfahrung kann Aggression signifikant vorhersagen $(p = 0.00191)$
- Y-Achsenabschnitt ist nun der Wert der AV, wenn alle 3 Prädiktoren = 0 sind.

```r
# Aufstellen komplexes Modell:

model2 = lm(Aggression ~ TV + Emotionsregulation + Ausgrenzung, data = df)
summary(model2)
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation + Ausgrenzung,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6668  -7.6041   0.2665   6.8444  18.4565
##
## Coefficients:
##                    Estimate Std. Error t value      Pr(>|t|)
## (Intercept)        65.59233    8.74569   7.500 0.00000000163 ***
## TV                  0.40650    0.63209   0.643       0.52335
## Emotionsregulation -0.41784    0.09651  -4.330 0.00007997688 ***
## Ausgrenzung         0.27409    0.08322   3.293       0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.733 on 46 degrees of freedom
## Multiple R-squared:  0.5025,    Adjusted R-squared:  0.4701
## F-statistic: 15.49 on 3 and 46 DF,  p-value: 0.0000004219
```

Multiple Regression

Modellpassung

- Auch für die multiple Regression lässt sich die Güte des Modells über R^2 schätzen
- Zur Erinnerung:
  - Verhältnis aufgeklärter zu gesamter Streuung
  - $0 \leq R^2 \leq 1$
  - Je näher R^2 an 1, desto besser passt sich Modell an Beobachtungspunkte an
- Die Hinzunahme weiterer Prädiktoren erhöht i.d.R. die Modellpassung

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Modellpassung

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4040  -6.2847   0.7681   7.6023  19.7061
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        86.24294    6.70484  12.863 < 0.0000000000000002 ***
## TV                  0.11254    0.68819   0.164                0.871
## Emotionsregulation -0.52752    0.09961  -5.296           0.00000308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.604 on 47 degrees of freedom
## Multiple R-squared:  0.3852,    Adjusted R-squared:  0.3591
## F-statistic: 14.73 on 2 and 47 DF,  p-value: 0.00001083
```

```
##
## Call:
## lm(formula = Aggression ~ TV + Emotionsregulation + Ausgrenzung,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6668  -7.6041   0.2665   6.8444  18.4565
##
## Coefficients:
##                    Estimate Std. Error t value      Pr(>|t|)
## (Intercept)        65.59233    8.74569   7.500 0.00000000163 ***
## TV                  0.40650    0.63209   0.643       0.52335
## Emotionsregulation -0.41784    0.09651  -4.330 0.00007997688 ***
## Ausgrenzung         0.27409    0.08322   3.293       0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.733 on 46 degrees of freedom
## Multiple R-squared:  0.5025,    Adjusted R-squared:  0.4701
## F-statistic: 15.49 on 3 and 46 DF,  p-value: 0.0000004219
```

- Einfaches Modell (links): $R^2 = .39 \rightarrow$ Es können 39% der AV (Aggression) durch TV und Emotionsregulation erklärt werden.
- Komplexes Modell (rechts): $R^2 = .50 \rightarrow$ Es können weitere 11% der AV durch Ausgrenzung erklärt werden.

Multiple Regression

Vorhersagewerte - Anwendung des Modells

$$Y_i = 86.24 + 0.11 \cdot X_{i1} + -0.53 \cdot X_{i2} + \epsilon_i$$

- Mit der aufgestellten Modellgleichung, können wir Werte vorhersagen.

Beispiel für Person $i = 1$:

$$\hat{Y}_i = 86.24294 + 0.11254 \cdot 5 + -0.52752 \cdot 50 = 60.43$$

- Laut unserem Modell mit 2 Prädiktoren, sollte Person 1 einen Aggressionswert von 60.29 Punkten haben.

- Da der tatsächlich beobachtete Wert 53 ist, beträgt der Modellfehler $60.29 - 53 = 7.29$ Punkte.

Aggression	TV	Emotionsregulation	Ausgrenzung
53	5	50	55
57	5	58	58
80	5	35	71
61	8	49	35
62	4	41	49
82	9	36	63
66	1	40	62
44	5	47	30
51	5	56	56
54	5	68	38
76	6	35	78
65	4	56	38
65	4	31	35
61	2	52	98
53	2	64	37

Multiple Regression

Vorhersagewerte - Anwendung des Modells

Wir können mit R automatisch die Wert für unsere Modelle schätzen:

Für das Modell mit TV und Emotionsregulation:

```
df$pred_model1 = round(predict(model1, newdata = df), 2)
```

Für das Modell mit TV, Emotionsregulation und Ausgrenzung:

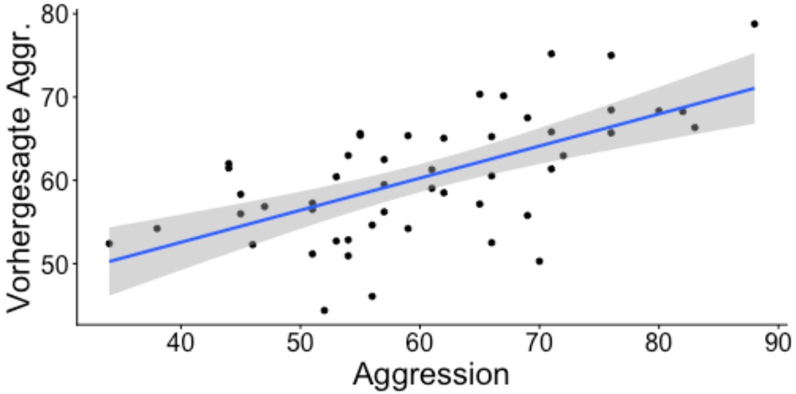
```
df$pred_model2 = round(predict(model2, newdata = df), 2)
```

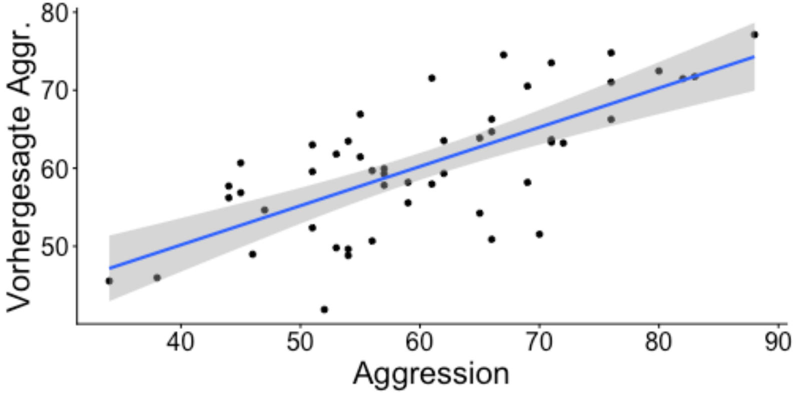
Aggression	TV	Emotionsregulation	Ausgrenzung	pred_model1	pred_model2
53	5	50	55	60.43	61.81
57	5	58	58	56.21	59.29
80	5	35	71	68.34	72.46
61	8	49	35	61.29	57.96
62	4	41	49	65.06	63.52
82	9	36	63	68.26	71.48
66	1	40	62	65.25	66.28
44	5	47	30	62.01	56.21
51	5	56	56	57.26	59.57
54	5	68	38	50.93	49.63
76	6	35	78	68.45	74.79
65	4	56	38	57.15	54.23
65	4	31	35	70.34	63.86
61	2	52	98	59.04	71.54
53	2	64	37	52.71	49.80

Multiple Regression

Vorhersagewerte - Anwendung des Modells

- Je besser das Modell passt, desto stärker der Zusammenhang zwischen beobachteten und vorhergesagten Werten:





Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Voraussetzungen der multiplen Regression

Wie bei einfacher Regression:

1) Das Kriterium (AV) muss intervallskaliert und normalverteilt sein.

2) Die Prädiktoren (UV) können entweder intervallskaliert und normalverteilt oder dichotom nominalskaliert sein.

3) Die Werte der einzelnen Versuchspersonen müssen unabhängig voneinander sein

4) Die Zusammenhänge müssen theroretisch linear sein (sonst andere Regressionsmodelle nutzen).

5) Streuungen der Wertepaare müssen über ganzen Wertebereich von X und Z homogen sein (Homoskedastizität).

Nur bei multipler Regression:

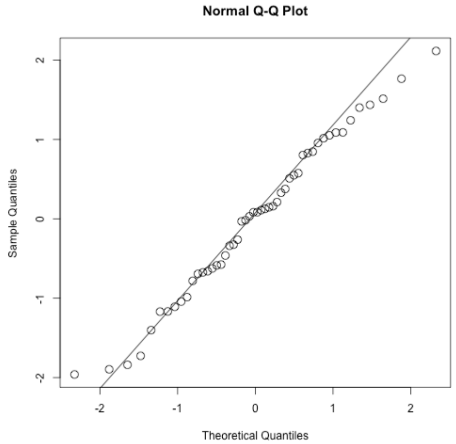
6) Multikollinearität: Prädiktoren sollten nicht zu stark miteinander korrelieren

Multiple Regression

Voraussetzungen der multiplen Regression

Normalverteilung der Residuen:

```
qqnorm(rstandard(model1), cex = 1.5)
qqline(rstandard(model1))
```




```
model1 = lm(Aggression ~ TV + Emotionsregulation, data = df)
shapiro.test(rstandard(model1))
```

```
##
##     Shapiro-Wilk normality test
##
## data:  rstandard(model1)
## W = 0.98337, p-value = 0.6997
```

Benchmarks:

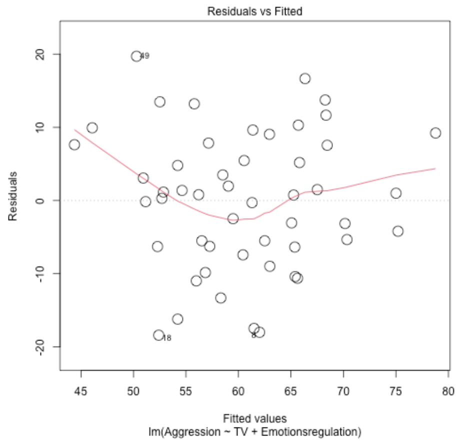
- QQ-Plot: Punkte sollten möglichst auf der 45 Grad Diagonalen liegen
- Shapiro-Wilk Test: p-Wert sollte > als $\alpha = .05$ sein

Multiple Regression

Voraussetzungen der multiplen Regression

Homoskedastizität:

```
model1 = lm(Aggression ~ TV + Emotionsregulation, data =
plot(model1, 1, cex = 2)
```



- Plot der standardisierten Residuen gegen die standardisierten vorhergesagten Werte
- Ideal ist eine Punktewolke ohne Systematik (Pattern)
- Die Linie sollte relativ horizontal verlaufen

→ dann ist Homoskedastizitätsannahme gegeben

Multiple Regression

Voraussetzungen der multiplen Regression

Multikollinearität:

Drei Methoden zur Prüfung von Multikollinearität:

- Korrelationsmatrix für hohe Korrelationen scannen
- Variance inflation factor (VIF)
- Toleranz-Statistik (1/VIF)

Benchmarks für potentielle Multikolliniaritätsproblematik:

- Korrelationen mit $r > .9$ können Probleme bereiten
- größter VIF größer als 10 (Bowerman & O'Connel, 1990)
- Durchschnittlicher VIF substanziell größer als 1
- Toleranz niedriger als 0.1 (ernstes Problem)
- Toleranz niedriger als 0.2 (potentielles Problem)

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Voraussetzungen der multiplen Regression

Multikollinearität:

Korrelationsmatrix für hohe Korrelationen scannen:

```
cor(df[, c("Aggression", "TV", "Emotionsregulation")])
```

```
##                    Aggression         TV Emotionsregulation
## Aggression          1.0000000  0.1357292         -0.6203926
## TV                  0.1357292  1.0000000         -0.1891767
## Emotionsregulation -0.6203926 -0.1891767          1.0000000
```

$\rightarrow$ Keine der bivariaten Korrelationen zwischen den Prädiktoren ist $r > .9$

Multiple Regression

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Voraussetzungen der multiplen Regression

Multikollinearität:

VIF und Toleranz berechnen:

```
library(olsrr)

ols_vif_tol(model1)
```

```
##            Variables Tolerance      VIF
## 1                 TV 0.9642122 1.037116
## 2 Emotionsregulation 0.9642122 1.037116
```

- Kein VIF größer als 10
- Durchschnittlicher VIF nicht substanziell größer als 1
- Toleranz nicht niedriger als 0.2

→ Es scheint kein Multikollinearitätsproblem vorzuliegen.

Take-aways

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

- Die Multiple Regression ermöglicht die Erweiterung des Regressionsmodells um **weitere Prädiktoren**.
- Im Gegensatz zur mehrfaktoriellen ANOVA, dürfen **auch stetige UVs** verwendet werden.
- Damit kann die **Modellpassung** und somit die **Vorhersagegüte** erhöht werden.
- Darstellung bei 2 Prädiktoren entspricht einer **Ebene** im 3D Raum.
- Es können **Hypothesentests** für die einzelnen Koeffizienten (Prädiktoren) und für das Gesamtmodell (Omnibustest) geprüft werden.
- Je mehr systematisch prädiktive UVs das Modell enthält, desto eher werden **vorhergesagte Werte** den tatsächlich beobachteten entsprechen.
- Als zusätzliche Modellvoraussetzung muss die **Multikollinearität** geprüft werden