# Statistik – Regression

# Regression

- In der Bivariaten Deskriptiven Statistik haben wir die Korrelation kennengelernt

- $r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{2{,}5900}{0{,}1685 * 17{,}4929} = 0{,}8787$

- Es liegt eine starke positive Korrelation vor, d.h. steigt die Größe, steigt auch das Gewicht

- Wir können etwas über den Zusammenhang zweier Größen sagen

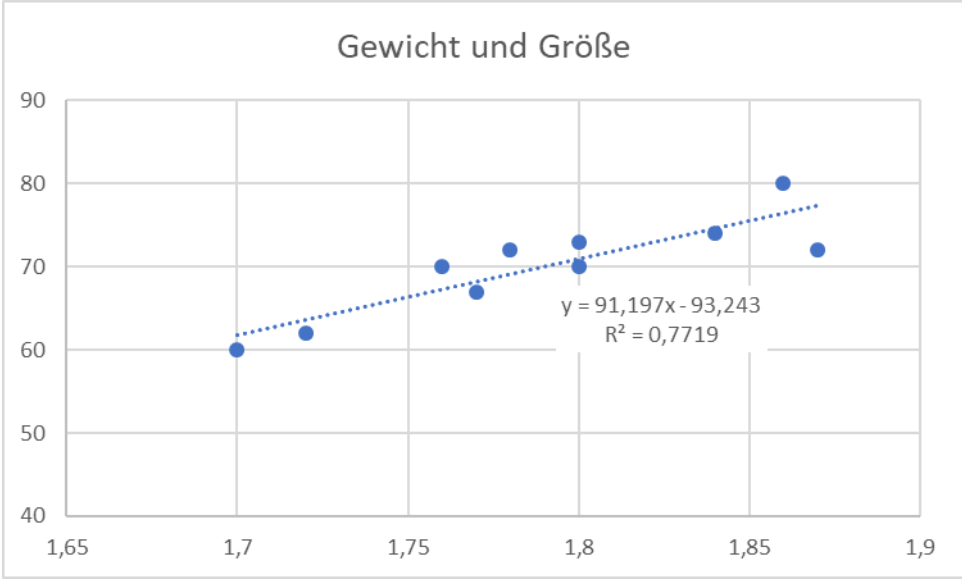
| Befragter | Größe X [m] | Gewicht Y [kg] |
|---|---|---|
| 1 | 1,87 | 72 |
| 2 | 1,70 | 60 |
| 3 | 1,80 | 73 |
| 4 | 1,84 | 74 |
| 5 | 1,78 | 72 |
| 6 | 1,80 | 70 |
| 7 | 1,72 | 62 |
| 8 | 1,76 | 70 |
| 9 | 1,86 | 80 |
| 10 | 1,77 | 67 |

# Regression

- Bisher können wir aber nichts über funktionale Zusammenhänge sagen, wir sind nicht in der Lage vorhersagen über das Gewicht zu machen, wenn wir die Größe einer Person kennen
- Diese Möglichkeit eröffnet uns die Regression
- Sie stellt uns eine Gleichung zur Verfügung
- Wir wissen, dass diese Gleichung auf Basis von Stichproben entwickelt wurde, und dass damit Fehlaussagen möglich sind

# Regression

- Wir haben möglicherweise die Regression schon in Excel kennengelernt
- Über eine Trendlinie können wir uns eine Gleichung und ein Bestimmtheitsmaß $R^2$ darstellen lassen



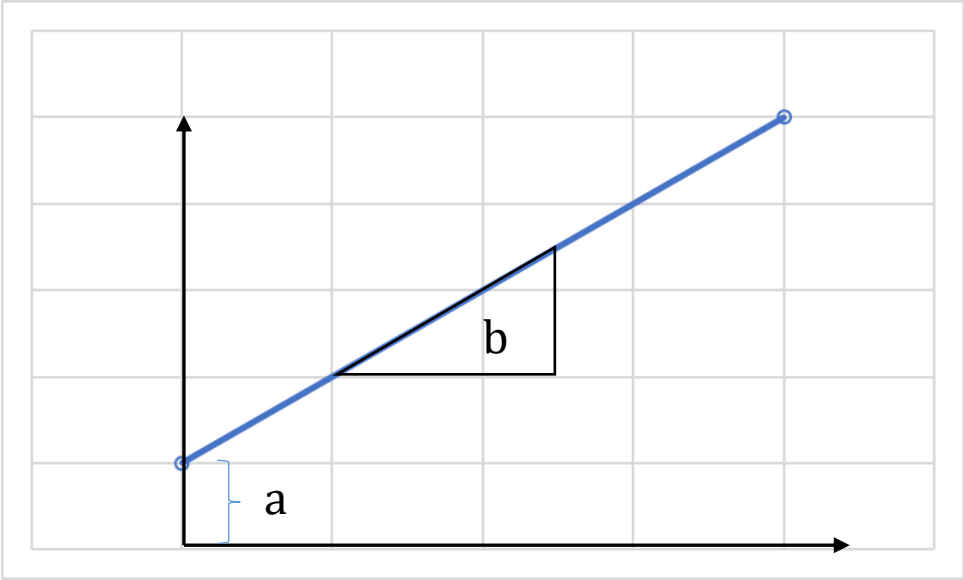
# Regression

## Entwicklung einer Regressionsgleichung

- Beschränkung auf den einfachsten Fall der Linearen Regression mit einer unabhängigen Variablen
- Mehr Komplexität ist aber möglich:
  - Multiple Lineare Regression: Mehrere unabhängige Variablen
  - Nicht-lineare Regression: Die abhängige Variable folgt einer anderen Gleichung als der Geradengleichung

# Regression

- Lineare Regression: Die zugrunde liegende Funktion folgt einer Geradengleichung
- $y = f(x) = a + b * x$
- $y$ *abhängige Variable*
- $x$ *unabhängige Variable*
- $a$ Schnittpunkt mit der y-Achse
- $b$ Steigung der Geraden



# Regression

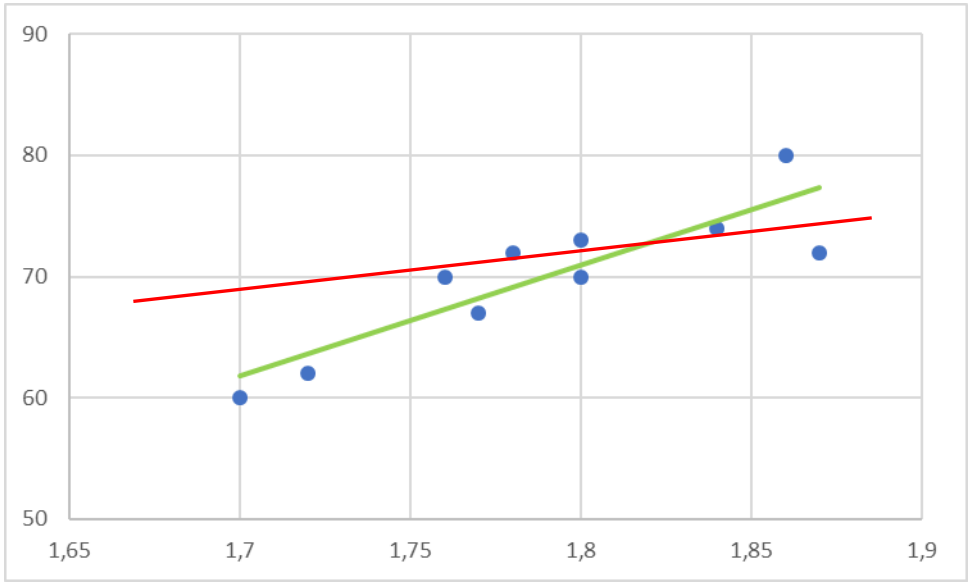
## Regressionsgleichung

- $\hat{y}_i = a + b * x_i + \varepsilon$
- In der Realität wird die einfache Geradengleichung nicht exakt erfüllt, wir werden immer einen gewissen Fehler $\varepsilon$ berücksichtigen
- Die Regressionsrechnung sucht nun eine Gleichung, die diesen Fehler minimiert
- In die endgültige Gleichung werden wir Werte $x_i$ einsetzen können und eine Vorhersage für $y_i$ erhalten

# Regression

## Regressionsgleichung

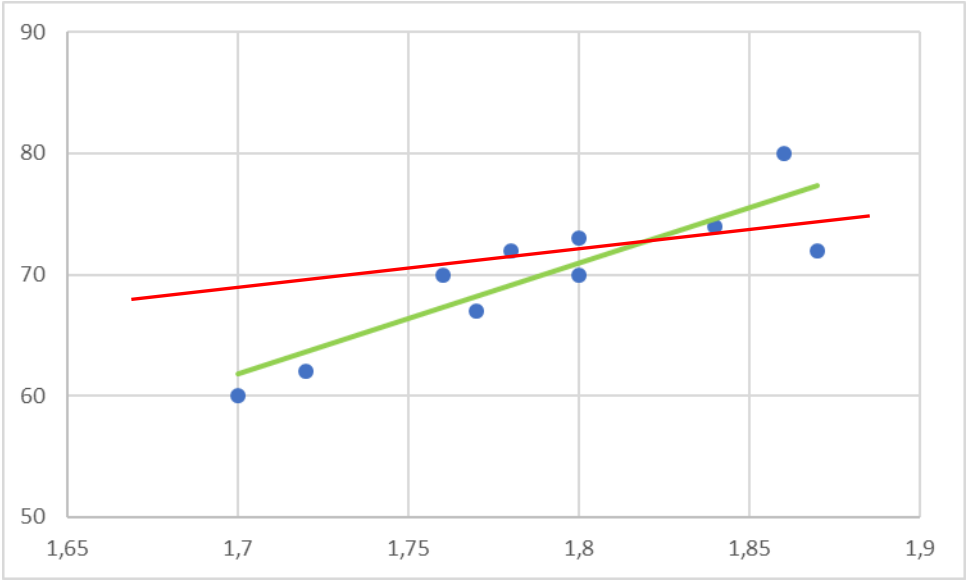
- Aus den vorliegenden Daten erhalten wir eine Punktewolke, in die wir eine Gerade so platzieren werden, dass der entstehende Fehler (Abweichung der einzelnen Punkte von der Geraden) klein wird



# Regression

## Regressionsgleichung

- Residuum: Abstand der einzelnen Punkte von der Regressionsgeraden



# Regression

**Methode der kleinsten Fehlerquadrate**

- Suche nach einer Geraden, die die Residuen aller Datenpunkte in Summe möglichst klein macht
- Quadrieren der einzelnen Fehler (Residuen) verhindert, dass sich positive und negative Werte gegenseitig auslöschen

# Regression

## Methode der kleinsten Fehlerquadrate

- Die Minimierungsaufgabe

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = min$$

$$\sum_{i=1}^{n}\left(y_i - (a + b * x_i)\right)^2 = min$$

# Regression

## Methode der kleinsten Fehlerquadrate

- Die Minimierungsaufgabe lauft nach einigem rechnen auf folgende Bestimmungsgleichungen für die Komponenten der Regressionsgleichung hinaus

$$b = \frac{cov_{x,y}}{s_x^2}$$

$$cov_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

$$a = \bar{y} - b * \bar{x}$$

# Regression

## Methode der kleinsten Fehlerquadrate

$$cov_{x,y} = 0{,}2878$$

$$b = 91{,}1972$$

$$a = -93{,}2429$$

# Regression

## Regressionsgleichung

- Die Regressionsgleichung

$$\hat{y}_i = -93{,}2429 + 91{,}1972 * x_i$$

- Mit dieser Gleichung kann jetzt eine Vorhersage gemacht werden, wenn die Größe bekannt ist (! Und man zu der Grundgesamtheit gehört, aus der diese Stichprobe gezogen wurde!)

- Vorhersagen sind übrigens nur im Bereich $[x_{min}, x_{max}]$ zulässig

# Regression

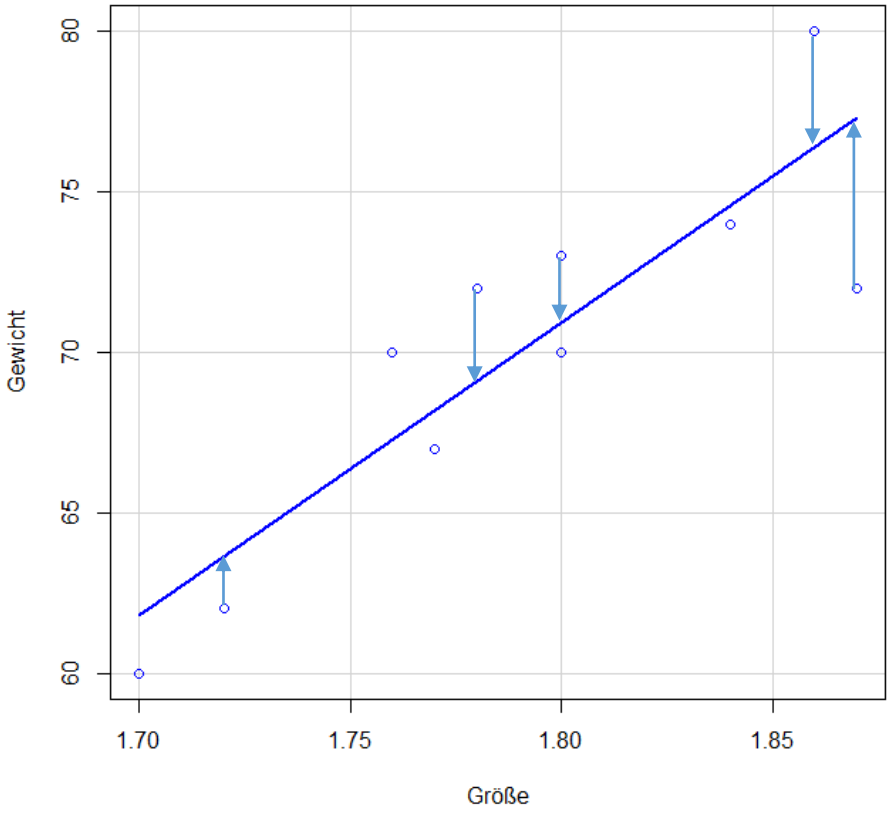
## Vorhersagegüte

- Die Regressionsgleichung ist entwickelt
- Offen ist aber die Frage, wie gut die Gleichung unsere Daten wiederspiegelt
- Auch die Vorhersagegüte hängt von der Qualität der Gleichung ab
- Ein erstes Signal liefert uns ein Streudiagramm mit der Regressionsgeraden

# Regression

## Vorhersagegüte

- Die Vorhersagegüte unserer Regressions-gleichung wird bestimmt durch die Abstände der Datenpunkte von der Regressionsgeraden



# Regression

## Vorhersagegüte

- Zur Bestimmung der Vorhersagegüte nutzen wir Varianzen

$s_y^2$ Gesamtvarianz: Quadrierte Abweichung aller Werte vom Mittelwert

$s_{\hat{y}}^2$ Regressionsvarianz: Quadrierte Abweichung aller vorhergesagten Werte vom Mittelwert

$s_{y^*}^2$ Fehlervarianz: Quadrierte Abweichung aller Werte vom vorhergesagten Wert

# Regression

## Vorhersagegüte

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$s_{\hat{y}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$s_{y^*}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - y_i^*)^2$$

# Regression

## Vorhersagegüte

$s_y^2 = 34{,}0000$

$s_{\hat{y}}^2 = 26{,}2445$

$s_{y^*}^2 = 7{,}7555$

# Regression

## Vorhersagegüte

Teststatistik: Quotient aus Regressionsvarianz und Fehlervarianz

$$F = \frac{s_{\hat{y}}^2}{s_{y^*}^2}$$

Für ein gutes Regressionsmodell sollte der Wert über $F > 1$ gelten

Der Wert ist aber nur mit weiteren Modellen vergleichbar

# Regression

## Vorhersagegüte

$$F = \frac{26{,}2445}{7{,}7555} = 3{,}39$$

Für ein gutes Regressionsmodell sollte der Wert über $F > 1$ gelten

Der Wert ist aber nur mit weiteren Modellen vergleichbar

# Regression

## Vorhersagegüte

- Besser interpretierbar ist der Determinationskoeffizient $R^2$
- Quotient aus Regressionsvarianz und Gesamtvarianz
- Entspricht dem quadrierten Korrelationskoeffizient
- $R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = r^2 = \frac{26{,}2445}{34{,}0000} = 0{,}7718$

# Regression

**Voraussetzungen für die lineare Regression**

- Mindestens intervallskalierte unabhängige Variable
- Mindestens intervallskalierte abhängige Variable
- Linearer Zusammenhang muss gegeben sein
- Wenige Ausreißer

# Regression

## Hypothesen für Faktoren, Wechselwirkungen und Konstanten

$H_0$ Das untersuchte Element ist keine wichtige Größe in der Regressionsgleichung

$H_1$ Das untersuchte Element ist signifikant wichtig in der Regressionsgleichung

Beispiel:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -93.24      31.39  -2.971 0.017854 *
Größe          91.20      17.53   5.203 0.000819 ***
```

p-Werte < $\alpha$: Konstante (Intercept) und Faktor (Größe) sind signifikant wichtig

# Regression

## Hypothesen für die Modellgüte

$H_0$ Das Modell beschreibt nicht die vorliegenden Daten, die Regressionsgleichung ist keine gute Beschreibung der Daten

$H_1$ Das Modell beschreibt die vorliegenden Daten signifikant, die Regressionsgleichung ist eine gute Beschreibung der Daten

Beispiel:

```
F-statistic: 27.07 on 1 and 8 DF,  p-value: 0.0008193
```

p-Wert < $\alpha$: Das Modell beschreibt die Daten signifikant

# Regression

```
Call:lm(formula = Gewicht ~ Größe, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2958 -1.5062 -0.7359  2.5739  3.6162

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -93.24      31.39  -2.971 0.017854 *
Größe          91.20      17.53   5.203 0.000819 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.954 on 8 degrees of freedom
Multiple R-squared:  0.7719,  Adjusted R-squared:  0.7434
F-statistic: 27.07 on 1 and 8 DF,  p-value: 0.0008193
```

Wir finden unsere Werte wieder!

# Regression

```
F-statistic: 27.07 on 1 and 8 DF,  p-value: 0.0008193
```

- p-Wert < 0,05 Das Modell liefert einen signifikanten Erklärungsbeitrag

# Regression

```
Multiple R-squared:  0.7719,  Adjusted R-squared:  0.7434
```

- Determinationsquotient $R^2$: wieviel Prozent der Varianz der abhängigen Variable (hier: Gewicht) wird erklärt
- Je höher desto besser
- Der korrigierte $\mathrm{R}^2$ (Adjusted R-squared) spielt in der einfachen linearen Regression keine Rolle (wichtig in multiplen linearen Regression)

# Regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -93.24      31.39  -2.971 0.017854 *
Größe          91.20      17.53   5.203 0.000819 ***
```

- Größe und Signifikanz der Regressionskoeffizienten
- Intercept: Die Konstante a in der Gleichung
- Größe: Die unabhängige Variable
- Für beide wird ein p-Wert angegeben
- $p<\alpha$: Wert ist signifikant (verbleibt im Modell)

# Multiple Lineare Regression

Erweiterung der linearen Regression auf die multiple lineare Regression

- Bisher gehen wir von einer unabhängigen Variablen ($x$) aus, die unser Problem beschreibt
- Was passiert, wenn mehrere unabhängige Variablen ($x_i$) das Problem beschreiben?
- Beispiel: Beispiel_Regression.xlsx
- Neben der abhängigen Variablen Gewicht gibt es zwei unabhängige Variablen Größe und Schuhgröße

# Multiple Lineare Regression

Frage: Können wir mit den beiden unabhängigen Variablen Größe bzw. Schuhgröße jeweils das Gewicht darstellen?

- Zwei Berechnungen der linearen Regression

# Multiple Lineare Regression

- **Größe als unabhängige Variable**

```
Residuals:
    Min      1Q  Median      3Q     Max
-18.884  -5.200  -1.730   7.996  26.593

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -113.9931    39.6218  -2.877    0.0083 **
Größe          1.0870     0.2229   4.877 0.0000568 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.82 on 24 degrees of freedom
Multiple R-squared:  0.4977,      Adjusted R-squared:  0.4768
F-statistic: 23.78 on 1 and 24 DF,  p-value: 0.00005685
```

Signifikantes Modell ($p < \alpha = 5\%$), dass aber nicht viel Streuung erklärt ($R^2 < 50\%$)

# Multiple Lineare Regression

- **Schuhgröße als unabhängige Variable**

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.1401 -6.2864 -0.6401  4.6099 16.0500

Coefficients:
             Estimate Std. Error t value      Pr(>|t|)
(Intercept) -127.0577    22.1583  -5.734 0.00000657471 ***
Schuhgröße     4.9049     0.5265   9.316 0.00000000193 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.103 on 24 degrees of freedom
Multiple R-squared:  0.7834,      Adjusted R-squared:  0.7743
F-statistic: 86.78 on 1 and 24 DF,  p-value: 0.000000001927
```

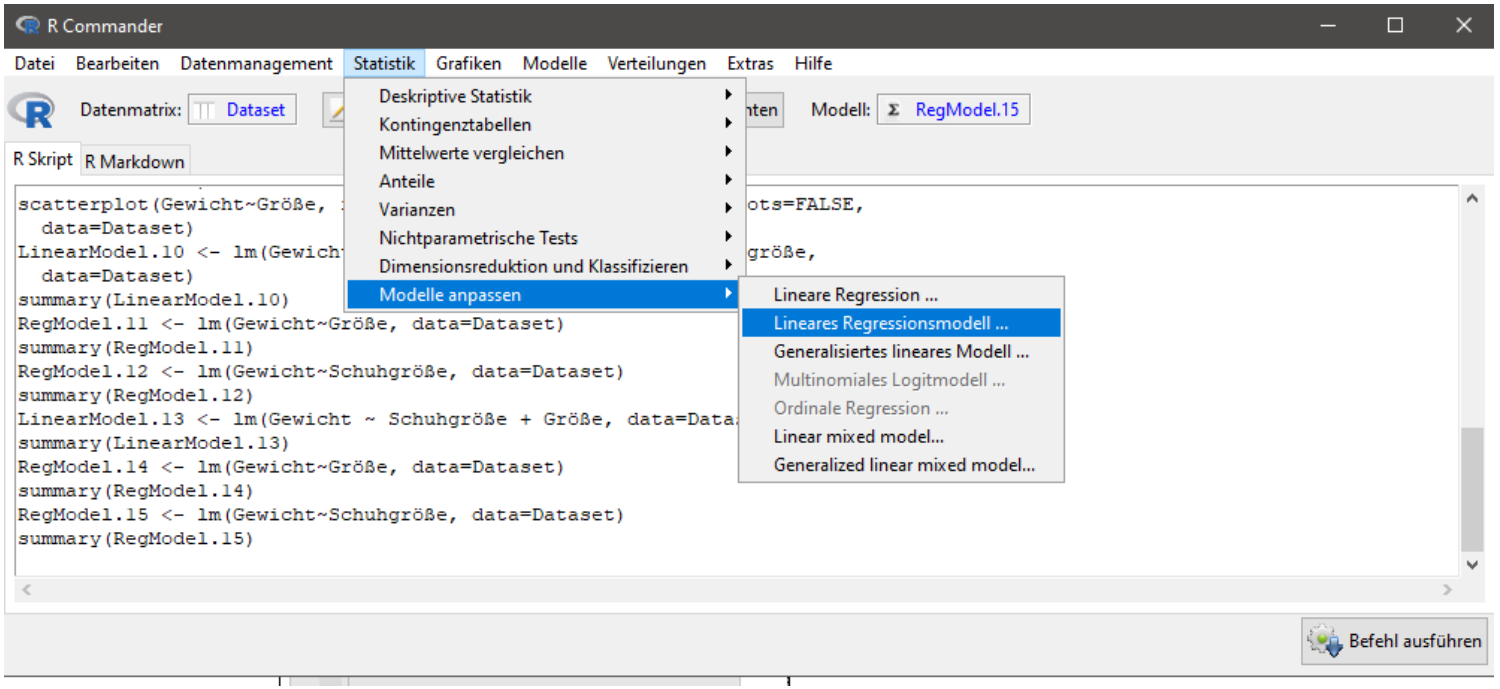
Signifikantes Modell (p < a = 5%), deutlich besseres Modell ($R^2 > 50\%; F_{Schuhgröße} > F_{Größe}$)

# Multiple Lineare Regression

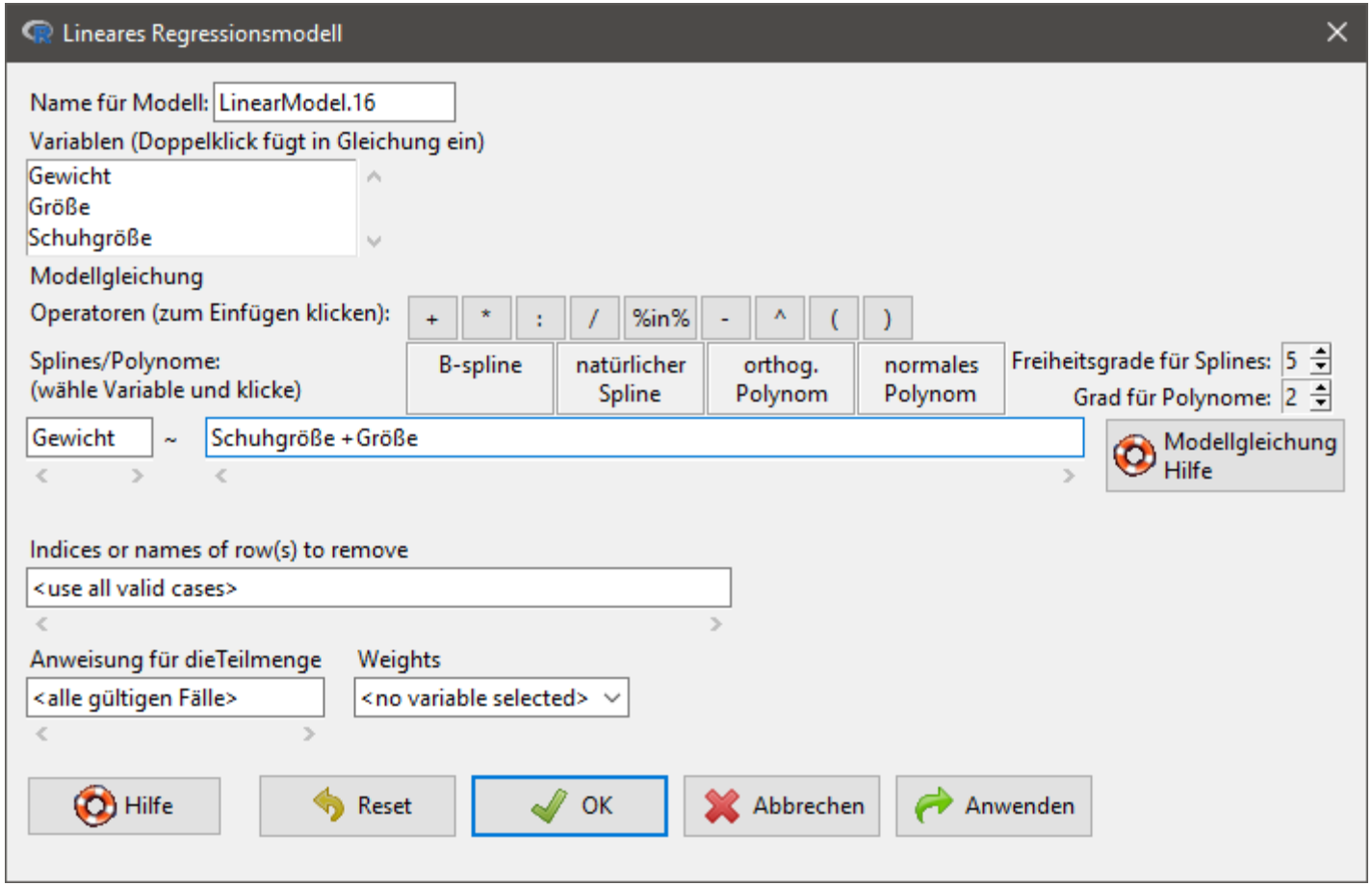
- Was passiert, wenn wir beide unabhängigen Variablen gleichzeitig in der Regressionsrechnung verwenden?
- Wir benötigen eine neue Regressionsgleichung:
- $\hat{y}_i = a + b_1 * x_{i,1} + b_2 * x_{i,2} + \cdots + b_p * x_{i,p} + \varepsilon$
- Auch diese Gleichung lässt sich mit der Methode der kleinsten Fehlerquadrate lösen, auf eine Herleitung wird hier aber verzichtet

# Multiple Lineare Regression

- **Schuhgröße <u>und</u> Größe als unabhängige Variablen**
- **Wir wechseln von *Lineare Regression…* zu *Lineares Regressionsmodell…***



# Multiple Lineare Regression



# Multiple Lineare Regression

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.1694 -4.4693 -0.9884  3.5260 19.4014

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) -104.1928    25.0535  -4.159   0.000379 ***
Schuhgröße     6.5396     1.0711   6.105 0.00000315 ***
Größe         -0.5156     0.2978  -1.731   0.096787 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.825 on 23 degrees of freedom
Multiple R-squared:  0.8083,  Adjusted R-squared:  0.7917
F-statistic:  48.5 on 2 and 23 DF,  p-value: 0.000000005615
```

Signifikantes Modell ($p < \alpha = 5\%$)

# Multiple Lineare Regression

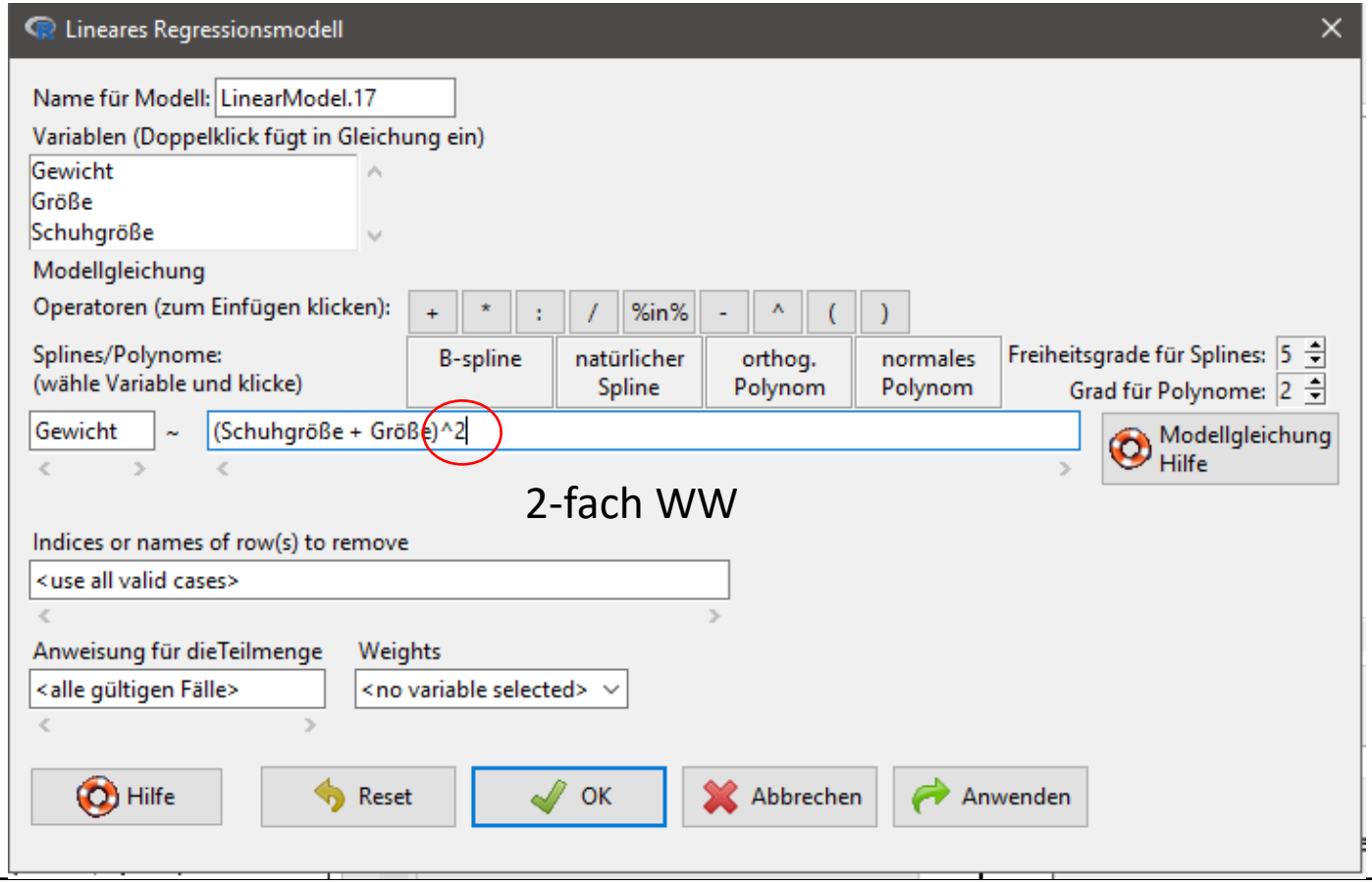
- Das neue Modell ist signifikant ($p < \alpha = 5\%$) und erklärt einen Großteil der Streuung (ca. 80%)
- Der F-Wert sinkt aber im Vergleich zum einfachen Modell (nur Schuhgröße) deutlich ab

$$F_{Schuhgröße} = 86{,}78 > F_{Schuhgröße+Größe} = 48{,}5$$

- Die unabhängige Variable Größe ist im Modell nicht signifikant! ($p < \alpha = 5\%$)

# Multiple Lineare Regression

- Zusätzlich können jetzt noch Wechselwirkungen der beteiligten unabhängigen Faktoren berücksichtigt werden



# Multiple Lineare Regression

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.7201 -4.5812  0.4169  3.8348 18.5759

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -422.30324  306.94812  -1.376   0.1827
Schuhgröße         14.23581    7.47833   1.904   0.0701 .
Größe               1.31126    1.78187   0.736   0.4696
Schuhgröße:Größe   -0.04405    0.04237  -1.040   0.3097
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.813 on 22 degrees of freedom
Multiple R-squared:  0.8173,  Adjusted R-squared:  0.7924
F-statistic: 32.81 on 3 and 22 DF,  p-value: 0.0000000267
```

# Multiple Lineare Regression

- Das neue Modell mit Wechselwirkung ist signifikant ($p < \alpha = 5\%$) und erklärt einen Großteil der Streuung (ca. 82%)
- Der F-Wert sinkt aber im Vergleich zum Modell ohne Wechselwirkung weiter ab
- Unabhängige Variablen und Wechselwirkung sind im Modell nicht mehr signifikant!
- Nicht einmal die Schuhgröße ist noch signifikant, der Wechsel zu diesem Modell kann nicht empfohlen werden