Multiple Lineare Regression an einem Beispiel

Anja Mühlemann

13. Mai 2024

Durchführen der multiplen linearen Regression

Der Datensatz, welcher wir in diesem Beispiel untersuchen, stammt vom *Australian Institute of Sport* und enthält Daten von 102 Athleten und 100 Athletinnen.

```
ds <- read.table("ais.txt", header = T)
str(ds)
```

```
## 'data.frame':    202 obs. of  13 variables:
##  $ Sex  : chr  "female" "female" "female" "female" ...
##  $ Sport: chr  "BBall" "BBall" "BBall" "BBall" ...
##  $ RCC  : num  3.96 4.41 4.14 4.11 4.45 4.1 4.31 4.42 4.3 4.51 ...
##  $ WCC  : num  7.5 8.3 5 5.3 6.8 4.4 5.3 5.7 8.9 4.4 ...
##  $ Hc   : num  37.5 38.2 36.4 37.3 41.5 37.4 39.6 39.9 41.1 41.6 ...
##  $ Hg   : num  12.3 12.7 11.6 12.6 14 12.5 12.8 13.2 13.5 12.7 ...
##  $ Ferr : int  60 68 21 69 29 42 73 44 41 44 ...
##  $ BMI  : num  20.6 20.7 21.9 21.9 19 ...
##  $ SSF  : num  109.1 102.8 104.6 126.4 80.3 ...
##  $ Bfat : num  19.8 21.3 19.9 23.7 17.6 ...
##  $ LBM  : num  63.3 58.5 55.4 57.2 53.2 ...
##  $ Ht   : num  196 190 178 185 185 ...
##  $ Wt   : num  78.9 74.4 69.1 74.9 64.6 63.7 75.2 62.3 66.5 62.9 ...
```

Es gilt:

- *Sport*: Sportart
- *Sex*: männlich oder weiblich
- *Ht*: Körpergrösse in cm
- *Wt*: Gewicht in kg
- *LBM*: Lean body mass
- *RCC*: Anzahl rote Blutkörperchen
- *WCC*: Anzahl weisse Blutkörperchen
- *Hc*: Hämatokrit (Volumenanteil der zellulären Elemente im Blut in %)
- *Hg*: Hämoglobin (Protein, das 90% der roten Blutkörperchen ausmacht)
- *Ferr*: Eisenkonzentration
- *BMI*: Body mass index (Gewicht/Körpergrösse^2)
- *SSF*: Summe der Hautfaltenmessungen
- *Bfat*: Körperfettanteil in %

Als erstes müssen wir sicherstellen, dass die Variablen *Sport* und *Sex* als kategoriell verstanden werden.

```
ds$Sport <- as.factor(ds$Sport)
ds$Sex <- as.factor(ds$Sex)
str(ds)
```

```
## 'data.frame':    202 obs. of  13 variables:
```

```
##  $ Sex  : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Sport: Factor w/ 10 levels "BBall","Field",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ RCC  : num  3.96 4.41 4.14 4.11 4.45 4.1 4.31 4.42 4.3 4.51 ...
##  $ WCC  : num  7.5 8.3 5 5.3 6.8 4.4 5.3 5.7 8.9 4.4 ...
##  $ Hc   : num  37.5 38.2 36.4 37.3 41.5 37.4 39.6 39.9 41.1 41.6 ...
##  $ Hg   : num  12.3 12.7 11.6 12.6 14 12.5 12.8 13.2 13.5 12.7 ...
##  $ Ferr : int  60 68 21 69 29 42 73 44 41 44 ...
##  $ BMI  : num  20.6 20.7 21.9 21.9 19 ...
##  $ SSF  : num  109.1 102.8 104.6 126.4 80.3 ...
##  $ Bfat : num  19.8 21.3 19.9 23.7 17.6 ...
##  $ LBM  : num  63.3 58.5 55.4 57.2 53.2 ...
##  $ Ht   : num  196 190 178 185 185 ...
##  $ Wt   : num  78.9 74.4 69.1 74.9 64.6 63.7 75.2 62.3 66.5 62.9 ...
```

Mit diesem Datensatz könnten wir verschiedene interessante Modelle rechnen. In diesem Beispiel versuchen wir die Lean Body Mass *LBM* durch die Sportart *Sport*, das Geschlecht *Sex*, das Gewicht {Wt} und die Körpergrösse *Ht* zu modellieren. Im folgenden betrachten wir p-Werte kleiner als 0.05 als signifikant.

```
fit <- lm(LBM ~ Sport + Sex + Wt + Ht, data = ds)
summary(fit)
```

```
##
## Call:
## lm(formula = LBM ~ Sport + Sex + Wt + Ht, data = ds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0270  -1.5501   0.1373   1.5401   8.2466
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -19.72251    5.90817  -3.338 0.001016 **
## SportField     1.89754    0.91453   2.075 0.039351 *
## SportGym       3.46566    1.61318   2.148 0.032960 *
## SportNetball  -1.62550    0.77380  -2.101 0.036996 *
## SportRow       0.82272    0.67028   1.227 0.221189
## SportSwim      2.68892    0.77528   3.468 0.000648 ***
## SportT400m     1.71299    0.81785   2.094 0.037549 *
## SportTennis    1.25376    0.97810   1.282 0.201472
## SportTSprnt    3.02548    0.92798   3.260 0.001320 **
## SportWPolo    -0.64348    0.83283  -0.773 0.440701
## Sexmale        7.53275    0.54167  13.907  < 2e-16 ***
## Wt             0.64444    0.02652  24.296  < 2e-16 ***
## Ht             0.17470    0.03733   4.679 5.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.466 on 189 degrees of freedom
## Multiple R-squared:  0.9665, Adjusted R-squared:  0.9644
## F-statistic: 454.7 on 12 and 189 DF,  p-value: < 2.2e-16
```

In den Resultaten erkennen wir folgende Dinge:

1. *Schätzwerte:*

- Der Intercept beschreibt die Lean Body Mass einer Basketballspielerin mit Körpergrösse 0cm und Gewicht 0kg.
- Netball- und Polospielerinnen haben eine kleinere LBM als Basketballspielerinnen derselben Körpergrösse und desselben Gewichts. Field- und Gymsportlerinnen, Rudererinnen, Schwimmerinnern, 400m-Läuferinnen, Tennisspielerinnen und Sprinterinnen haben eine grössere LBM als Basketballspielerinnen derselben Körpergrösse und desselben Gewichts.
- Basketballspieler derselben Statur (Grösse & Gewicht) haben eine um ca. 7.5kg höhere LBM als Basketballspielerinnen.
- Pro zusätzlichem kg Gewicht nimmt die LBM im Schnitt um 0.64kg zu, wenn die anderen Variablen fixiert bleiben.
- Pro zusätzlichem cm Körpergrösse nimmt die LBM im Schnitt um 0.17kg zu, wenn die anderen Variablen fixiert bleiben.

2. *Test auf Zusammenhang auf 95%-Niveau:* Alle Kovariablen haben einen signifikanten Einfluss auf die LBM. Jedoch unterscheiden sich die Rudererinnen und die Polospielerinnen nicht signifikant von den Basketballspielerinnen.

3. *Adjustiertes R^2:* Etwa 96% der Varianz in der LBM wird durch die von uns gewählten Kovariablen erklärt.

4. *F-Test:* Der F-Test ist signifikant. Dies ist nicht weiter erstaunlich, dann dies bedeutet, dass mit einer Sicherheit von 95% mindestens eine der Kovariablen einen signifikanten Einfluss auf die LBM hat.

Übung:

- Schätzen Sie die LBM einer 173cm grossen und 61kg schweren Schwimmerin.
- Um wie viel kg unterscheidet sich die LBM von Sprinterinnen und 400m-Läufer gleicher Statur (Grösse und Gewicht)?
- Wie gross wäre ein 75kg schwerer Sprinter mit LBM 63kg?

Lösung:

- Die LBM einer 173cm grossen und 61kg schweren Schwimmerin ist

$$\begin{aligned}\hat{LBM} &= -19.72 + 1.90 \cdot X_{Field} + 3.47 \cdot X_{Gym} - 1.63 \cdot X_{Netball} + 0.82 \cdot X_{Row} + 2.69 \cdot X_{Swim} \\ &\quad + 1.71 \cdot X_{400m} + 1.25 \cdot X_{Tennis} + 3.03 \cdot X_{Sprint} - 0.64 \cdot X_{WPolo} + 7.53 \cdot X_{male} \\ &\quad + 0.64 \cdot Wt + 0.17 \cdot Ht \\ &= -19.72 + 1.90 \cdot 0 + 3.47 \cdot 0 - 1.63 \cdot 0 + 0.82 \cdot 0 + 2.69 \cdot 1 \\ &\quad + 1.71 \cdot 0 + 1.25 \cdot 0 + 3.03 \cdot 0 - 0.64 \cdot 0 + 7.53 \cdot 0 \\ &\quad + 0.64 \cdot 61 + 0.17 \cdot 173 \\ &= -19.72 + 2.69 \cdot 1 + 0.64 \cdot 61 + 0.17 \cdot 173 \\ &= \underline{\underline{51.42kg}}\end{aligned}$$

- Der Gewichtsunterschied lautet: Um wie viel kg unterscheidet sich die LBM von Sprinterinnen und 400m-Läufer gleicher Statur (Grösse und Gewicht)?

$$\begin{aligned}\hat{LBM}_{Sprinterin} &= -19.72 + 3.03 \cdot 1 + 0.64 \cdot Wt + 0.17 \cdot Ht \\ \hat{LBM}_{400m-Läufer} &= -19.72 + 1.71 \cdot 1 + 7.53 \cdot 1 + 0.64 \cdot Wt + 0.17 \cdot Ht\end{aligned}$$

  Daher lauter der Unterschied

$$\hat{LBM}_{400m-Läufer} - \hat{LBM}_{Sprinterin} = 1.71 \cdot 1 + 7.53 \cdot 1 - 3.03 \cdot 1 = \quad \underline{\underline{6.21kg}}$$

- Die Grösse eines 75kg schwerer Sprinters mit LBM 63kg ist

$$\begin{aligned}63kg &= -19.72 + 3.03 \cdot 1 + 7.53 \cdot 1 + 0.64 \cdot 75 + 0.17 \cdot Ht \\ \Leftrightarrow Ht &= \underline{\underline{142.12cm}}\end{aligned}$$

Mit R könnte man die 1. Frage folgendermassen beantworten:

```
predict(fit,newdata = data.frame(Sex="female", Sport="Swim", Ht=173, Wt=61),
        interval = "prediction",level = 0.95)
```

```
##        fit      lwr      upr
## 1 52.50017 47.49765 57.50269
```

Konfidenzintervalle

Natürlich können wir uns auch noch die 95%-Konfidenzintervalle für die Schätzer anschauen

```
confint(fit)
```

```
##                     2.5 %      97.5 %
## (Intercept)  -31.37692948 -8.06808545
## SportField     0.09355167  3.70152739
## SportGym       0.28350452  6.64781156
## SportNetball  -3.15188799 -0.09910582
## SportRow      -0.49947533  2.14492257
## SportSwim      1.15961619  4.21822422
## SportT400m     0.09969856  3.32627895
## SportTennis   -0.67563962  3.18316605
## SportTSprnt    1.19493829  4.85601187
## SportWPolo    -2.28632734  0.99936484
## Sexmale        6.46426195  8.60124281
## Wt             0.59211996  0.69676588
## Ht             0.10105539  0.24834043
```

Mit den Konfidenzintervallen kommen wir zum selben Entscheid, wie bei den t-Tests oben. Zudem können wir Aussagen vom folgenden Typ machen: "Mit einer Sicherheit von 95% nimmt die LBM bei Sportler:innen der gleichen Sportart, des gleichen Geschlechts und gleicher Körpergrösse pro zusätzlichem kg Körpergewicht zwischen 0.59kg und 0.70kg zu."

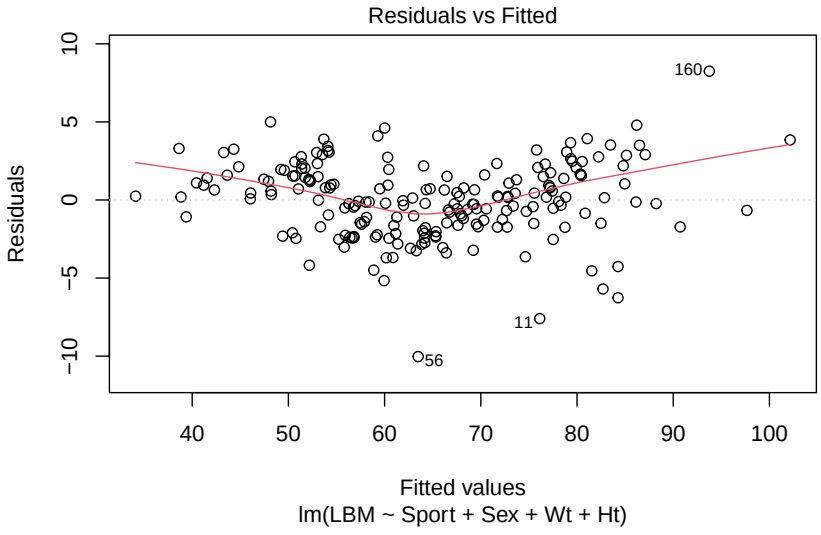
Überprüfung der Annahmen

bevor wir vorhersagen machen oder die Resultate oben als sicher behandeln, sollten wir noch die Annahmen überprüfen.

1. Linearität

Wir betrachten den Residualplot.

```
plot(fit, which=1)
```

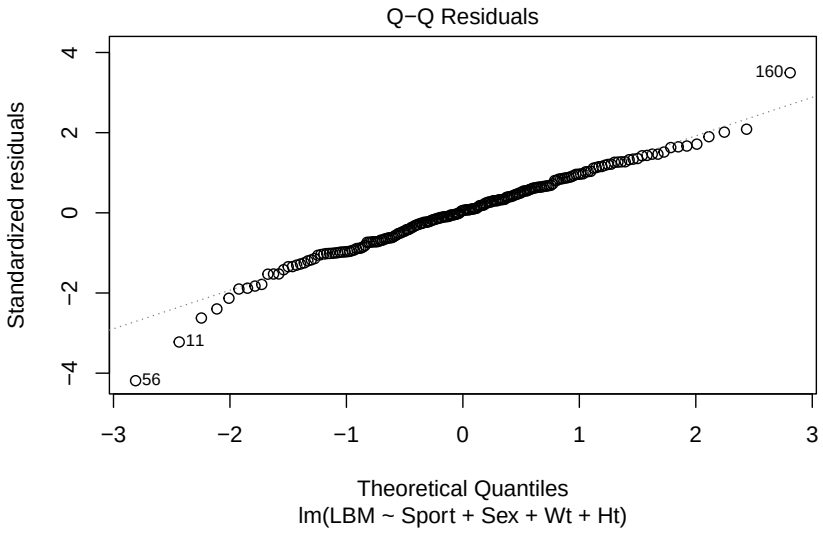


Der Residualplot deutet darauf hin, dass es noch nicht-lineare Effekte gibt, welche durch unser lineares Modell nicht abgedeckt werden.

2. Normalverteilung

Wir betrachten den Q-Q-Plot.

```
plot(fit, which=2)
```

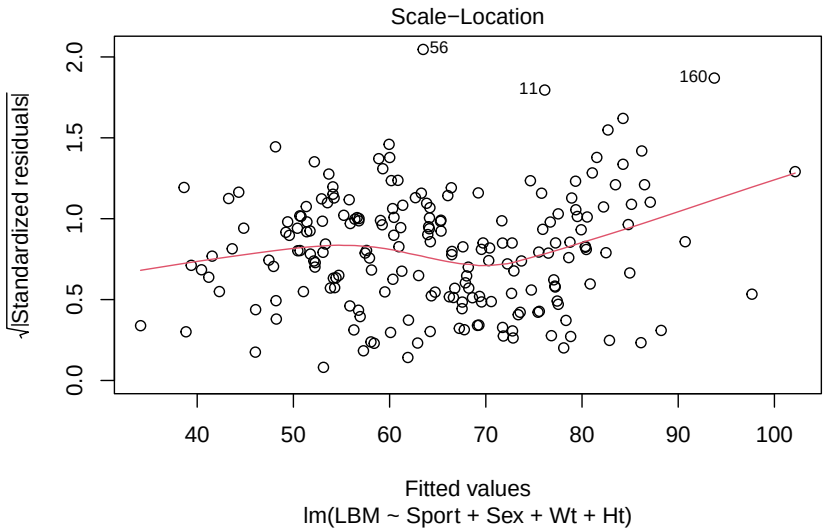


In den Enden weichen die beobachteten Quantile von denjenigen ab, welche wir unter einer Normalverteilung erwarten würden. Das passiert jedoch häufig und ist in diesem Fall noch in einem annehmbaren Bereich.

3. Homoskedastizität

Der Residualplot zeigte keinen Hinweis auf Heteroskedastizität. Wir betrachten aber zur Sicherheit noch den Scale-Location-Plot.

```
plot(fit, which=3)
```

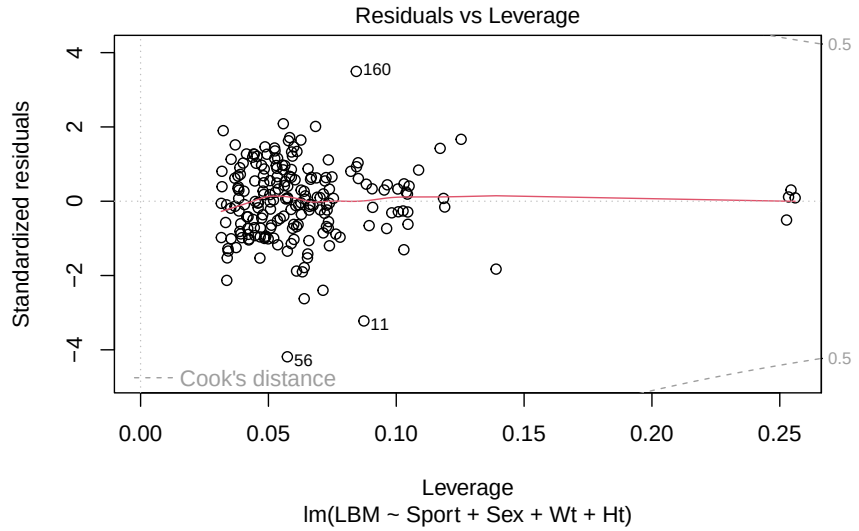


Das Resultat des Scale-Location-Plot ist ähnlich wie beim Residualplot. Es sieht nicht aus, als wäre die Homoskedastizität verletzt. Aber es gibt evtl. einen nicht-linearen Effekt, welcher nicht durch unser Modell erklärt wird.

4. Hebelwirkung

Wir betrachten den, "Leverage vs Residuals"-Plot.

```
plot(fit, which=5)
```



Es scheint keine Beobachtung mir übermässiger Hebelwirkung zu geben.

5. Multikollinearität

Wir betrachten nun wie stark die nummerischen Merkmale miteinander korrelieren. In unserem Modell hatten wir nur zwei nummerische Merkmale, nämlich `Wt` und `Ht`.

```r
round(cor(ds[,c("Wt",  "Ht")]),2)
```

```
##      Wt   Ht
## Wt 1.00 0.78
## Ht 0.78 1.00
```

Wie zu erwarten gibt es eine moderate bis starke Korrelation zwischen Gewicht und Körpergrösse. Die Korrelation ist aber nicht so stark, dass wir unseren Schlussfolgerungen nicht mehr vertrauen könnten.

Wir wir gesehen haben, sind die Annahmen der multiplen linearen Regression nur teilweise erfüllt. Daher modifizieren wir unser Modell, indem wir die Interaktionen mit einbeziehen und schauen, ob unser Modell so besser wird.

Multiple lineare Regression mit Interaktionen

```r
fit2 <- lm(LBM ~ Sport + (Sex + Wt + Ht)^2, data = ds)
summary(fit2)
```

```
##
## Call:
## lm(formula = LBM ~ Sport + (Sex + Wt + Ht)^2, data = ds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0963 -0.9707  0.0456  1.2283  6.4395
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.770774  17.762201  -1.958 0.051776 .
## SportField     1.465670   0.813051   1.803 0.073058 .
## SportGym       0.583369   1.567396   0.372 0.710175
## SportNetball  -1.592528   0.679320  -2.344 0.020118 *
## SportRow       1.013799   0.582311   1.741 0.083338 .
## SportSwim      2.333746   0.678709   3.439 0.000722 ***
## SportT400m     1.831432   0.702318   2.608 0.009855 **
## SportTennis    0.379654   0.846586   0.448 0.654349
## SportTSprnt    3.113377   0.805097   3.867 0.000152 ***
## SportWPolo    -1.303456   0.727425  -1.792 0.074778 .
## Sexmale      -35.434627   9.767822  -3.628 0.000369 ***
## Wt             1.036188   0.261589   3.961 0.000106 ***
## Ht             0.306487   0.103759   2.954 0.003544 **
## Sexmale:Wt     0.208851   0.039071   5.345 2.62e-07 ***
## Sexmale:Ht     0.154312   0.058381   2.643 0.008913 **
## Wt:Ht         -0.002891   0.001476  -1.958 0.051716 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.11 on 186 degrees of freedom
## Multiple R-squared:  0.9759, Adjusted R-squared:  0.9739
## F-statistic:   502 on 15 and 186 DF,  p-value: < 2.2e-16
```

Wir ziehen nun noch die Zweifachinteraktionen von Geschlecht, Gewicht und Grösse mit in unser Modell ein. Die Zweifachinteraktionen mit der Sportart lassen wir vorerst weg, weil wir ansonsten sehr viele Parameter schätzen müssten.

1. *Schätzwerte:* Der Schätzer für Geschlecht=männlich hat in diesem neuen Modell neu ein negatives

Vorzeichen. Das mag auf den ersten Blick erstaunen, wird jedoch durch die Interaktion von Geschlecht und Gewicht bzw. Geschlecht und Grösse wieder kompensiert. Bei der Interpretation des Einflusses eines Merkmals müssen nun auch die Interaktionen beachtet werden.

2. *Test auf Zusammenhang auf 95%-Niveau:* Nicht mehr alle Sportarten, welche im 1. Modell signifikant waren, sind in diesem Modell erneut signifikant. Dafür ist jetzt noch die Interaktion zwischen Geschlecht und Gewicht bzw. Grösse signifikant.
3. *Adjustiertes R^2:* Das neue Modell erklärt die Varianz zu etwa 97% - also mehr als das vorherige Modell.
4. *F-Test:* Der F-Test ist signifikant. Dies ist nicht weiter erstaunlich, da dies bedeutet, dass mit einer Sicherheit von 95% mindestens eine der Kovariablen einen signifikanten Einfluss auf die LBM hat.

Übung: Beantworten Sie die Fragen von vorhin erneut in Bezug auf das neue Modell.

- Schätzen Sie die LBM einer 173cm grossen und 61kg schweren Schwimmerin.
- Um wie viel kg unterscheidet sich die LBM von Sprinterinnen und 400m-Läufer gleicher Statur (Grösse und Gewicht)?
- Wie gross wäre ein 75kg schwerer Sprinter mit LBM 63kg?

Lösung:

- Die LBM einer 173cm grossen und 61kg schweren Schwimmerin ist

$$\begin{aligned}
\hat{LBM} &= -34.77 + 1.47 \cdot X_{Field} + 0.58 \cdot X_{Gym} - 1.59 \cdot X_{Netball} + 1.01 \cdot X_{Row} + 2.33 \cdot X_{Swim} \\
&\quad + 1.83 \cdot X_{400m} + 0.38 \cdot X_{Tennis} + 3.11 \cdot X_{Sprint} - 1.30 \cdot X_{WPolo} - 35.43 \cdot X_{male} \\
&\quad + 1.04 \cdot Wt + 0.31 \cdot Ht + 0.21 \cdot X_{male} \cdot Wt + 0.15 \cdot X_{male} \cdot Ht - 0.003 \cdot Wt \cdot Ht \\
&= -34.77 + 1.47 \cdot 0 + 0.58 \cdot 0 - 1.59 \cdot 0 + 1.01 \cdot 0 + 2.33 \cdot 1 \\
&\quad + 1.83 \cdot 0 + 0.38 \cdot 0 + 3.11 \cdot 0 - 1.30 \cdot 0 - 35.43 \cdot 0 \\
&\quad + 1.04 \cdot 61 + 0.31 \cdot 173 + 0.21 \cdot 0 \cdot 61 + 0.15 \cdot 0 \cdot 173 - 0.003 \cdot 61 \cdot 173 \\
&= -34.77 + 2.33 \cdot 1 + 1.04 \cdot 61 + 0.31 \cdot 173 - 0.003 \cdot 61 \cdot 173 \\
&= \underline{\underline{52.971}}
\end{aligned}$$

- Der Gewichtsunterschied lautet: Um wie viel kg unterscheidet sich die LBM von Sprinterinnen und 400m-Läufer gleicher Statur (Grösse und Gewicht)?

$$\begin{aligned}
\hat{LBM}_{Sprinterin} &= -34.77 + 3.11 \cdot 1 + 1.04 \cdot Wt + 0.31 \cdot Ht - 0.003 \cdot Wt \cdot Ht \\
\hat{LBM}_{400m-Läufer} &= -34.77 + 1.83 \cdot 1 - 35.43 \cdot 1 + 1.04 \cdot Wt + 0.31 \cdot Ht \\
&\quad + 0.21 \cdot 1 \cdot Wt + 0.15 \cdot 1 \cdot Ht - 0.003 \cdot Wt \cdot Ht
\end{aligned}$$

  Daher lauter der Unterschied

$$\hat{LBM}_{400m-Läufer} - \hat{LBM}_{Sprinterin} = 1.83 \cdot 1 - 35.43 \cdot 1 + 0.21 \cdot 1 \cdot Wt + 0.15 \cdot 1 \cdot Ht - 3.11 \cdot 1$$

- Die Grösse eines 75kg schwerer Sprinters mit LBM 63kg ist

$$\begin{aligned}
&63kg = -34.77 + 3.11 \cdot 1 - 35.43 \cdot 1 + 1.04 \cdot 75 + 0.31 \cdot Ht + 0.21 \cdot 1 \cdot 75 + 0.15 \cdot 1 \cdot Ht - 0.003 \cdot 75 \cdot Ht \\
&\Leftrightarrow Ht = \underline{154.638cm}
\end{aligned}$$

Mit R könnte man die 1. Frage folgendermassen beantworten:

```
predict(fit2,newdata = data.frame(Sex="female", Sport="Swim", Ht=173, Wt=61),
        interval = "prediction",level = 0.95)
```

```
##        fit      lwr      upr
## 1 53.28737 49.00216 57.57258
```

Wir können natürlich auch für dieses Modell die 95%-Konfidenzintervalle berechnen.

```
confint(fit2)
```

```
##                      2.5 %        97.5 %
## (Intercept)  -69.812045851  2.704987e-01
## SportField    -0.138317530  3.069657e+00
## SportGym      -2.508790313  3.675528e+00
## SportNetball  -2.932691365 -2.523642e-01
## SportRow      -0.134983985  2.162582e+00
## SportSwim      0.994789697  3.672702e+00
## SportT400m     0.445899120  3.216965e+00
## SportTennis   -1.290491631  2.049799e+00
## SportTSprnt    1.525082294  4.701672e+00
## SportWPolo    -2.738519315  1.316076e-01
## Sexmale      -54.704586558 -1.616467e+01
## Wt             0.520125212  1.552250e+00
## Ht             0.101791662  5.111828e-01
## Sexmale:Wt     0.131772217  2.859292e-01
## Sexmale:Ht     0.039138701  2.694862e-01
## Wt:Ht         -0.005803092  2.173408e-05
```

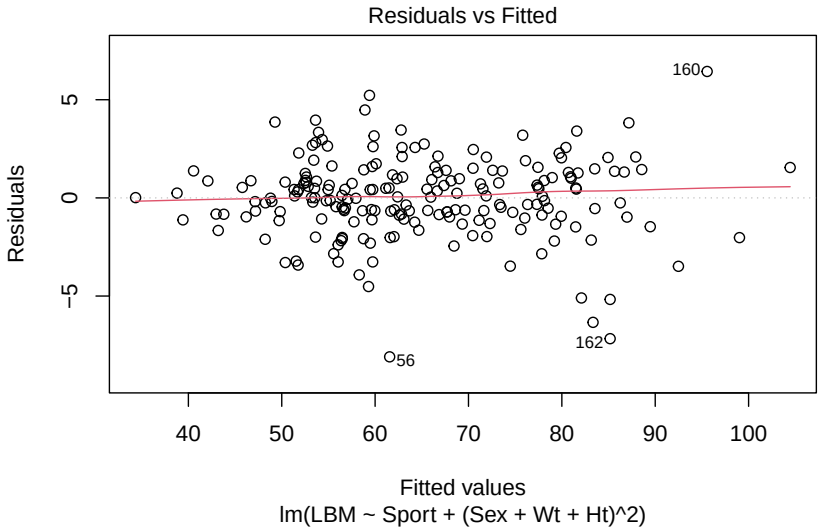
Überprüfung der Annahmen

Auch für dieses Modell sollten wir die Annahmen überprüfen.

1. Linearität

Wir betrachten den Residualplot.

```
plot(fit2, which=1)
```

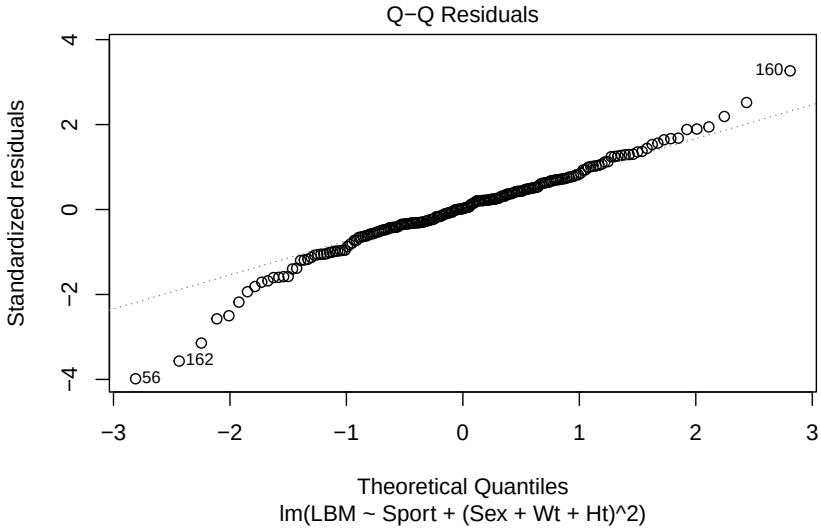


Der Residualplot für dieses Modell sieht sehr viel besser aus als vorhin. Es gibt keine starken Hinweise auf nicht-abgedeckte Nichtlinearitäten.

2. Normalverteilung

Wir betrachten den Q-Q-Plot.

```
plot(fit2, which=2)
```

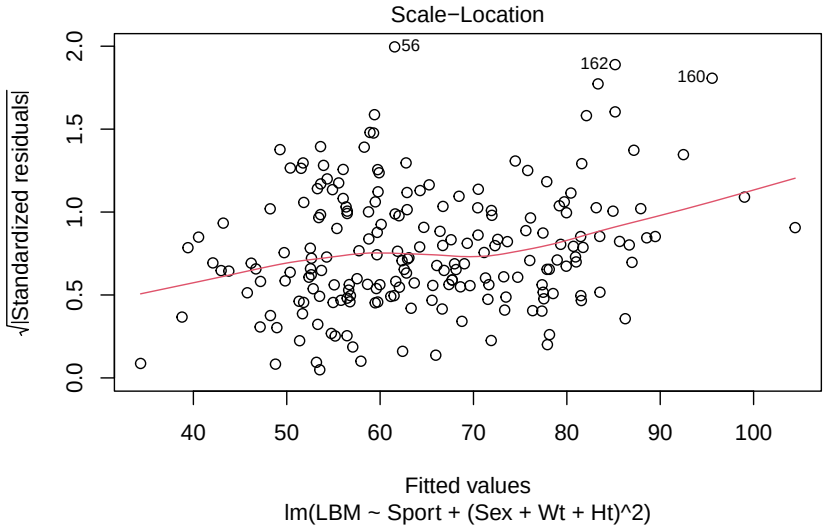


Der Q-Q-Plot ein bisschen weniger gut aus als vorhin. Gerade für sehr kleine Residuen ist die Normalverteilungsannahme erneut ein bisschen kritisch.

3. Homoskedastizität

Der Residualplot zeigte keinen Hinweis auf Heteroskedastizität. Wir betrachten aber zur Sicherheit noch den Scale-Location-Plot.

```
plot(fit2, which=3)
```



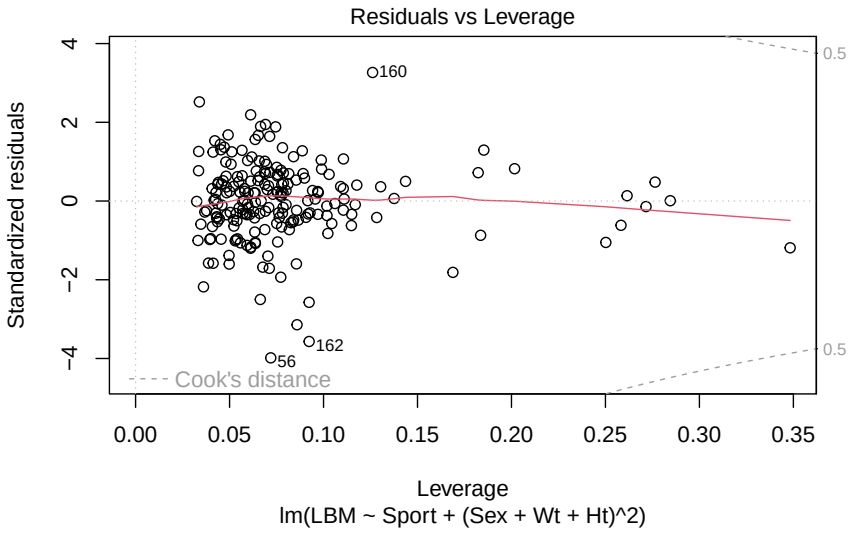
Die rote Linie ist zwar nicht exakt konstant, aber wir erkennen kein spezielles Muster. Da es auch beim Resid-

ualplot keinen speziellen Hinweis auf Heteroskedastizität gab, gehen wir davon aus dass die Homoskedastizität nicht verletzt ist.

4. Hebelwirkung

Wir betrachten den, "Leverage vs Residuals"-Plot.

```
plot(fit2, which=5)
```



Der Cook-Abstand ist erneut für alle Residuen relativ kein. Es gibt also keine einzelnen Residuen, welche die Schätzer übermässig stark beeinflusst haben.

Auch wenn die Normalverteilungsannahme ein bisschen kritisch ist, sehen wir, dass das neue Modell besser abschneidet als das vorherige..

Prognosen

Wir schauen uns nun noch an, wie Prognosen machen können. Um die Güte eines Modells zu beurteilen, wird das Modell häufig nur auf einem Teil des Datensatzes (Trainingsdatensatz) geschätzt. Anschliessend schaut man, wie gut sich das Modell auf den nicht verwendeten Daten (Testdatensatz) schlägt. Diese Vorgehensweise hilft uns auch verschiedene Modelle zu vergleichen. Idealerweise wiederholt man das Prozedere mehrmals um robustere Schlussfolgerungen zu erhalten. Dies lassen wir hier der Übersichtlichkeit halber sein.

```
set.seed(09052024) # Einen Seed wir gesetzt damit Zufallsstichproben reproduzierbar sind
index <- sample(1:202, size = 30, replace = FALSE) # wähle zufällig 30 Zeilen
test <- ds[index,]
training <- ds[-index, ]
fit1 <- lm(LBM ~ Sport + Sex + Wt + Ht, data = ds) # lineares Modell
fit2 <- lm(LBM ~ Sport + (Sex + Wt + Ht)^2, data = ds) # lineares Modell mit Interaktionen

predictions1 <- predict(fit1,newdata = test, interval = "prediction",level = 0.95)
predictions2 <- predict(fit2,newdata = test, interval = "prediction",level = 0.95)
data.frame(pred1=predictions1[,1], pred2=predictions2[,1], true=test$LBM)
```

```
##        pred1    pred2  true
## 66  48.22627 49.72691 48.57
## 200 68.61108 67.36940 68.00
```

```
## 88  49.57880 51.37918 51.48
## 75  73.36867 70.51676 72.98
## 37  49.40476 50.38355 47.09
## 180 69.54564 68.60550 68.00
## 127 79.32852 80.43319 83.00
## 4   60.86538 59.48512 57.18
## 10  52.24111 52.83239 53.42
## 122 78.92788 79.95866 82.00
## 116 76.49957 77.34869 78.00
## 115 75.42847 76.02818 75.00
## 28  60.18027 59.74189 56.48
## 63  59.28972 58.91286 63.39
## 130 72.77529 73.47934 73.00
## 39  47.45086 48.79410 48.78
## 125 80.53898 81.72998 83.00
## 148 63.01527 61.49467 62.00
## 36  48.14072 49.27806 53.14
## 144 71.73737 71.97138 70.00
## 188 66.48906 65.25567 68.00
## 123 80.35962 81.55111 82.00
## 102 69.27990 68.13110 69.00
## 41  53.54084 53.63328 56.45
## 183 67.53221 66.41741 68.00
## 137 80.34905 81.49661 82.00
## 158 73.70362 73.63070 75.00
## 85  50.66697 52.37642 53.11
## 164 69.71545 69.33390 68.00
## 103 73.57321 73.23698 74.00
```

Wir können nun für jeden Athleten / jede Athletin im Testdatensatz vergleichen, welche der beiden Vorhersagen näher an den wahren Wert kommt. Alternativ können auch die mittlere quadratische Abweichung oder die mittlere absolute Abweichung für beide Vorhersagen berechnen.

```
# Mittlere Quadratische Abweichung
mean((test$LBM - predictions1[,1])^2) # 1. Modell
```

```
## [1] 4.899416
```

```
mean((test$LBM - predictions2[,1])^2) # 2. Modell
```

```
## [1] 3.748729
```

```
# Mittlere Absolute Abweichung
mean(abs(test$LBM - predictions1[,1]))# 1. Modell
```

```
## [1] 1.818332
```

```
mean(abs(test$LBM - predictions2[,1])) # 2. Modell
```

```
## [1] 1.547891
```

Wir sehen, dass für beide Masse das neue Modell besser abschneidet. Dies stimmt mit dem überein, was wir durch Betrachtung des adjustierten Bestimmtheitsmasses vermutet hatten.