WDDA FS 2025: Leitfaden für Aufgabenserie 6

2025-06-04

1 Einleitung

Dieser Leitfaden bietet detaillierte Erklärungen für die Übungen in WDDA FS 2025 Aufgabenserie/Übungsblatt 6. Diese Serie konzentriert sich auf **Multiple Regression** (MRM), bei der wir mehrere erklärende Variablen verwenden, um eine Zielvariable zu modellieren.

2 Aufgabe 1: Gold Chains (Multiple Regression)

Aufgabenstellung: Analysieren Sie den Datensatz **Gold Chains** mit Preis als Zielvariable und Länge sowie Breite als erklärende Variablen.

2.1 Schritt 1: Daten einlesen und erkunden

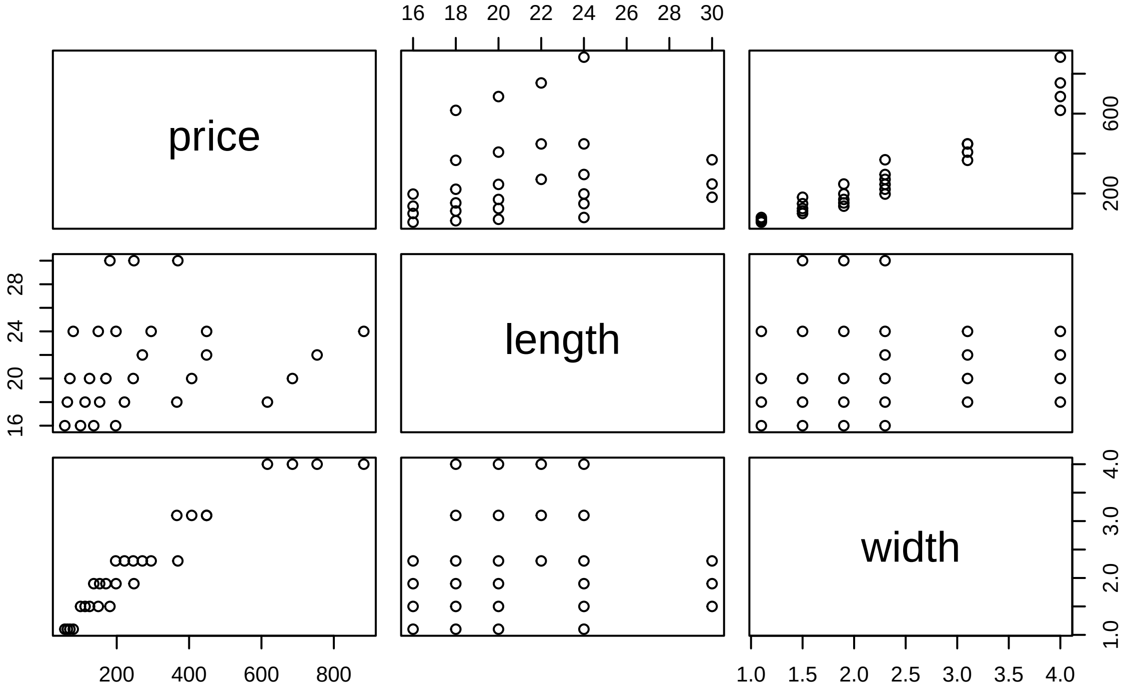
```r
gold <- read_excel("../data/WDDA_06.xlsx", sheet = "Gold Chains")
names(gold) <- c("price", "length", "width")
head(gold)
```

```
## # A tibble: 6 x 3
##   price length width
##   <dbl>  <dbl> <dbl>
## 1  56.6     16   1.1
## 2 100       16   1.5
## 3 137.      16   1.9
## 4 197.      16   2.3
## 5  63.6     18   1.1
## 6 112.      18   1.5
```

2.2 Schritt 2: Streudiagramme untersuchen (a)

```r
pairs(gold, main = "Streudiagramme Gold Chains")
```

Streudiagramme Gold Chains



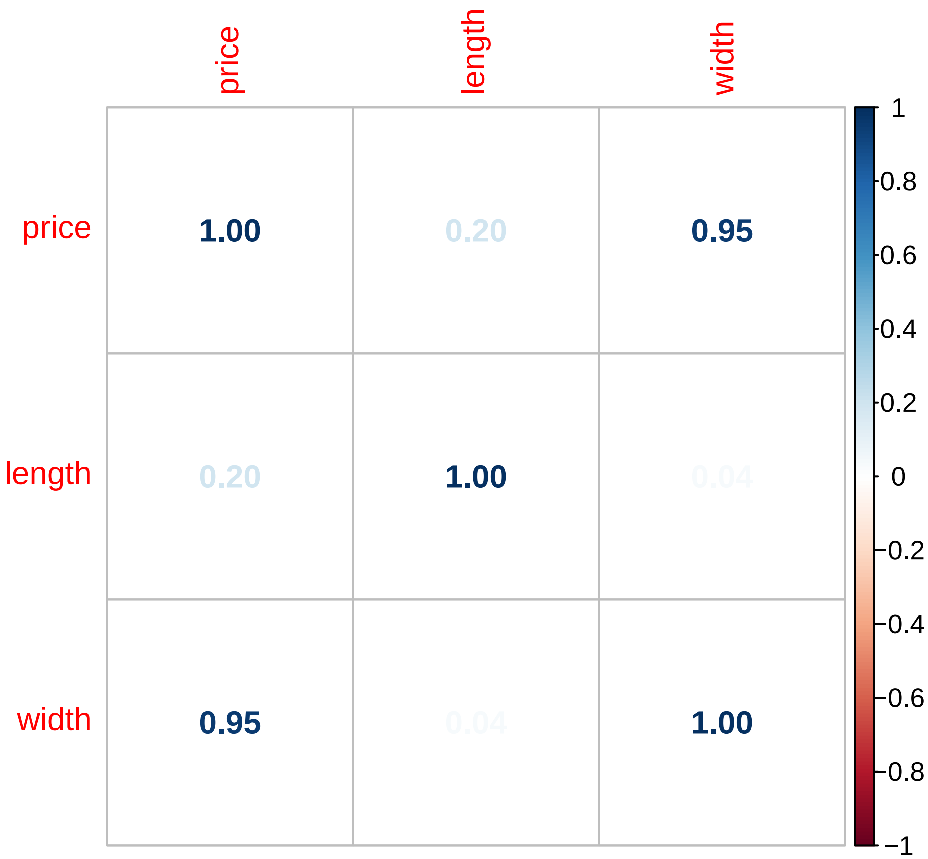
Bewertung: Die Diagramme zeigen lineare Beziehungen ohne starke Krümmung. Dies ist gut für eine multiple Regression geeignet.

2.3 Schritt 3: Korrelationen berechnen (b)

```
cor_matrix <- cor(gold)
print(cor_matrix)
```

```
##            price     length      width
## price  1.0000000 0.19984805 0.95438935
## length 0.1998481 1.00000000 0.03547783
## width  0.9543894 0.03547783 1.00000000
```

```
corrplot(cor_matrix, method = "number")
```



Grösste Korrelation: Preis und Breite (r   0.95)

Erklärung: Breitere Ketten benötigen mehr Gold, was zu höheren Preisen führt.

2.4 Schritt 4: Marginale Steigung der Breite (c)

Die **marginale Steigung** ist der Koeffizient in einer einfachen Regression:

```r
mod_width_simple <- lm(price ~ width, data = gold)
summary(mod_width_simple)
```

```
##
## Call:
## lm(formula = price ~ width, data = gold)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -106.906  -49.106   -4.375   39.524  208.359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -220.95      33.68  -6.561 5.88e-07 ***
## width         223.87      13.73  16.299 3.63e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 67.09 on 26 degrees of freedom
## Multiple R-squared:  0.9109, Adjusted R-squared:  0.9074
## F-statistic: 265.7 on 1 and 26 DF,  p-value: 3.631e-15
```

Interpretation: Pro mm Breitenzunahme steigt der Preis um ca. 224$.

2.5 Schritt 5: Erwartung für partielle Steigung (d)

Da Länge und Breite **unkorreliert** sind (r 0.04), erwarten wir, dass die partielle Steigung der Breite ähnlich der marginalen Steigung ist.

2.6 Schritt 6: Multiple Regression anpassen (e)

```r
mod_gold <- lm(price ~ length + width, data = gold)
summary(mod_gold)
```

```
##
## Call:
## lm(formula = price ~ length + width, data = gold)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -77.84 -33.24 -23.41  25.25 185.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -405.635     62.119  -6.530  7.7e-07 ***
## length         8.884      2.654   3.347  0.00258 **
## width        222.489     11.647  19.103  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.85 on 25 degrees of freedom
## Multiple R-squared:  0.9384, Adjusted R-squared:  0.9335
## F-statistic: 190.6 on 2 and 25 DF,  p-value: 7.34e-16
```

Partielle Steigung der Breite: Pro mm Breitenzunahme (bei konstanter Länge) steigt der Preis um ca. 222$.

2.7 Schritt 7: Intercept, R^2 und Standardfehler interpretieren (f)

```r
r2_gold <- summary(mod_gold)$r.squared
rse_gold <- summary(mod_gold)$sigma
cat("R² =", round(r2_gold, 4), "\n")
```

```
## R² = 0.9384
```

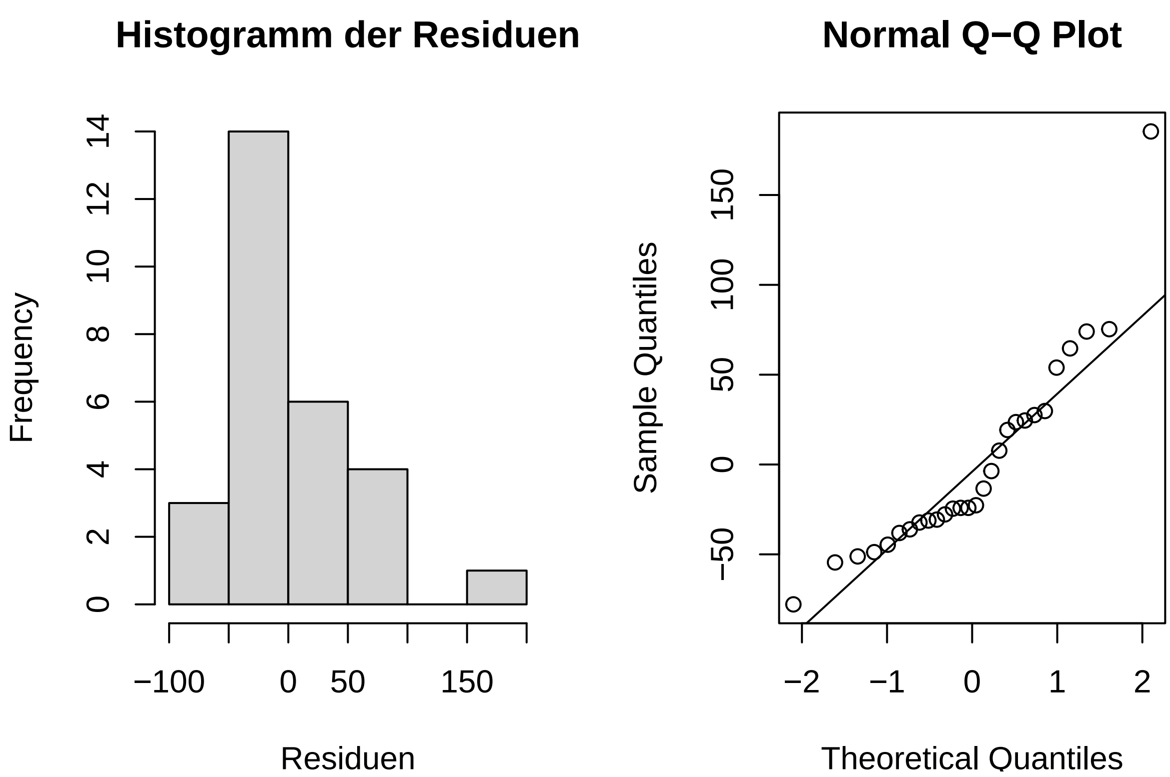
```r
cat("RSE =", round(rse_gold, 2), "$\n")
```

```
## RSE = 56.85 $
```

- **Intercept:** Nicht sinnvoll interpretierbar (Preis bei Länge=0, Breite=0)
- **R²:** Das Modell erklärt 94% der Preisvariation
- **RSE:** Typische Abweichung von ±57$ vom geschätzten Preis

2.8 Schritt 8: Residuen analysieren (g)

```r
resid_gold <- resid(mod_gold)
par(mfrow = c(1,2))
hist(resid_gold, main = "Histogramm der Residuen", xlab = "Residuen")
qqnorm(resid_gold)
qqline(resid_gold)
```




```r
par(mfrow = c(1,1))

mean_resid_gold <- mean(resid_gold)
sd_resid_gold <- sd(resid_gold)
cat("Residuen-Mittelwert:", round(mean_resid_gold, 2), "\n")
```

```
## Residuen-Mittelwert: 0
```

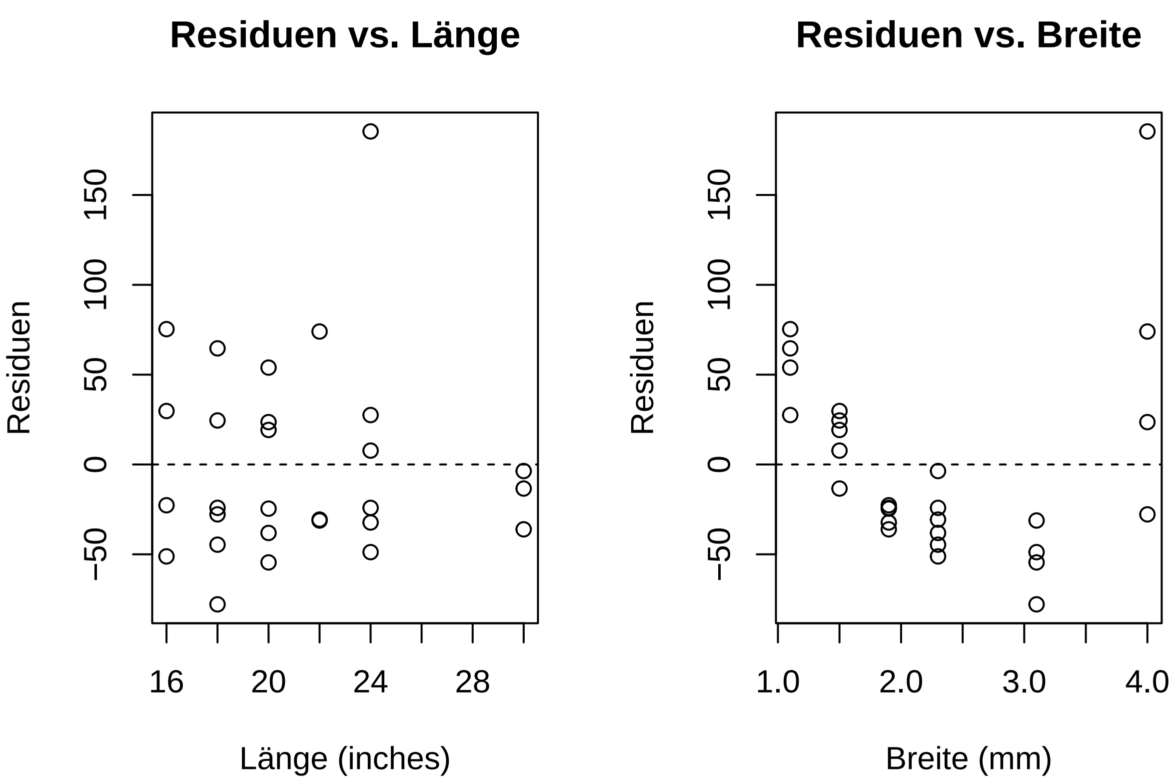
```r
cat("Residuen-SD:", round(sd_resid_gold, 2), "$\n")
```

```
## Residuen-SD: 54.71 $
```

Bewertung: Residuen sind ungefähr glockenförmig mit Mittelwert 0.

2.9 Schritt 9: Residuen vs. erklärende Variablen (h)

```r
par(mfrow = c(1,2))
plot(gold$length, resid_gold, main = "Residuen vs. Länge",
     xlab = "Länge (inches)", ylab = "Residuen")
abline(h = 0, lty = 2)
plot(gold$width, resid_gold, main = "Residuen vs. Breite",
     xlab = "Breite (mm)", ylab = "Residuen")
abline(h = 0, lty = 2)
```




```r
par(mfrow = c(1,1))
```

Bewertung: - Länge: OK (konstante Streuung) - Breite: Problematisch (U-förmiges Muster)

2.10 Schritt 10: MRM-Bedingungen erfüllt? (i)

1. **Linearität:** OK aus Streudiagrammen
2. **Konstante Varianz:** Problematisch bei Breite

3. **Normalität:** Ungefähr erfüllt
4. **Unabhängigkeit:** Angenommen

Fazit: Nicht alle Bedingungen erfüllt wegen Heteroskedastizität.

2.11 Schritt 11: Länge und Breite kombinieren (j)

```
gold$volume <- gold$length * gold$width
mod_gold_vol <- lm(price ~ length + width + volume, data = gold)
summary(mod_gold_vol)
```

```
##
## Call:
## lm(formula = price ~ length + width + volume, data = gold)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -91.17 -32.04 -22.90  38.82 106.91
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  103.002    150.208   0.686  0.49945
## length       -15.720      7.173  -2.192  0.03835 *
## width        -37.485     72.830  -0.515  0.61147
## volume        12.502      3.472   3.601  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.75 on 24 degrees of freedom
## Multiple R-squared:   0.96,  Adjusted R-squared:  0.955
## F-statistic: 192.2 on 3 and 24 DF,  p-value: < 2.2e-16
```

Das Volumen (Länge $\times$ Breite) könnte eine wichtige Variable sein, da es das Goldgewicht approximiert.

2.12 Schritt 12: Weitere Analysen (k-p)

```
# (k) Residuum der 1. Beobachtung
pred_1 <- predict(mod_gold)[1]
resid_1 <- gold$price[1] - pred_1
cat("Residuum 1. Beobachtung:", round(resid_1, 2), "$\n")
```

```
## Residuum 1. Beobachtung: 75.33 $
```

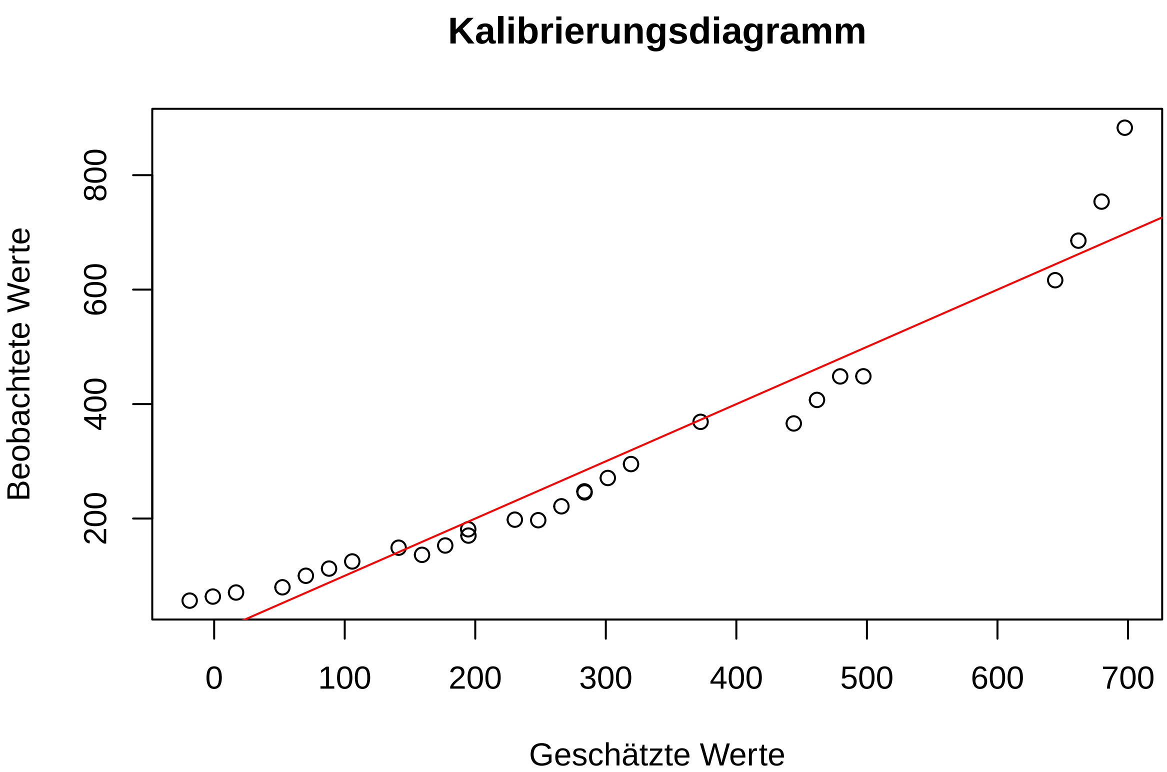
```
# (l) 25. Beobachtung extrem hoch?
pred_25 <- predict(mod_gold)[25]
ci_25 <- predict(mod_gold, interval = "prediction")[25,]
cat("25. Beobachtung - Preis:", gold$price[25], "$\n")
```

```
## 25. Beobachtung - Preis: 882.9 $
```

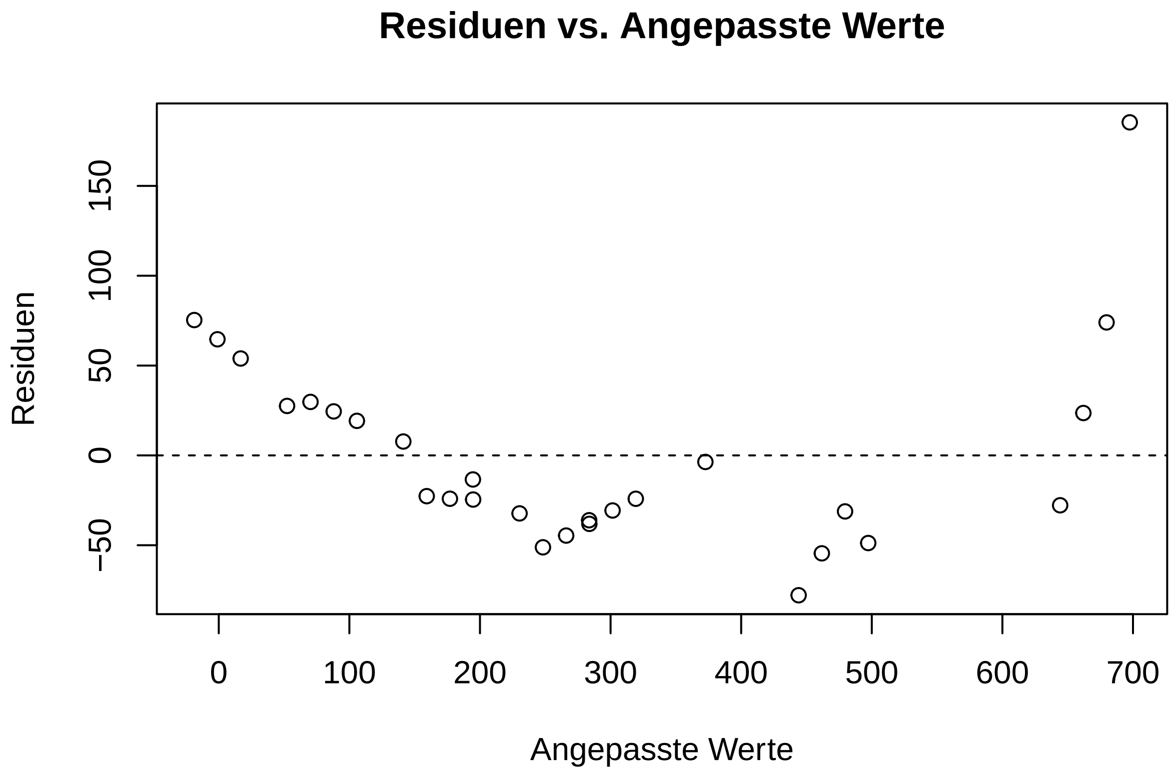
```r
cat("95% Prognose-Intervall: [", round(ci_25["lwr"], 2), ",", round(ci_25["upr"], 2), "] $\n")
```

```
## 95% Prognose-Intervall: [ 570.58 , 824.48 ] $
```

```r
# (m) Kalibrierungsdiagramm
plot(fitted(mod_gold), gold$price,
     main = "Kalibrierungsdiagramm",
     xlab = "Geschätzte Werte", ylab = "Beobachtete Werte")
abline(0, 1, col = "red")
```




```r
# (n) Residuen vs. angepasste Werte
plot(fitted(mod_gold), resid_gold,
     main = "Residuen vs. Angepasste Werte",
     xlab = "Angepasste Werte", ylab = "Residuen")
abline(h = 0, lty = 2)
```



3 Aufgabe 2: HR Regression (Pfaddiagramm-Interpretation)

Aufgabenstellung: Interpretieren Sie das gegebene Pfaddiagramm für Gehaltsvorhersage basierend auf Alter und Testscore.

3.1 Pfaddiagramm-Analyse

Aus dem Pfaddiagramm lesen wir ab: - **Age** → **Salary:** 5 \$000/year - **Test Score** → **Salary:** 2 \$000/point - **Age**   **Test Score:** 5 points/year (Korrelation)

3.2 Schritt 1: Gleichungen notieren (a)

```r
cat("MRM: Salary = b0 + b1*Age + b2*TestScore\n")
```

```
## MRM: Salary = b0 + b1*Age + b2*TestScore
```

```r
cat("Angepasstes MRM: Salary = b0 + 5*Age + 2*TestScore\n")
```

```
## Angepasstes MRM: Salary = b0 + 5*Age + 2*TestScore
```

Interpretation: - Pro Jahr Alter: +5000\$ (bei konstantem Testscore) - Pro Testpunkt: +2000\$ (bei konstantem Alter)

3.3 Schritt 2: Nötige Informationen für Schätzung? (b)

Nein! Das **Intercept (b)** fehlt im Pfaddiagramm. Ohne diesen können wir keine konkreten Gehaltsschätzungen machen.

3.4 Schritt 3: Direkter vs. indirekter Effekt (c)

```r
direct_effect <- 2  # $000/point
indirect_effect <- 5 * 2  # 5 points/year * 2 $000/point
cat("Direkter Effekt:", direct_effect, "$000/point\n")
```

```
## Direkter Effekt: 2 $000/point
```

```r
cat("Indirekter Effekt:", indirect_effect, "$000/year\n")
```

```
## Indirekter Effekt: 10 $000/year
```

Indirekter Effekt ist grösser ($10 > 2$).

3.5 Schritt 4: Marginale Steigung (d)

Die **marginale Steigung** berücksichtigt sowohl direkte als auch indirekte Effekte:

$$\text{Marginale Steigung} = \text{Direkter Effekt} + \text{Indirekter Effekt}$$

3.6 Schritt 5: Kurs-Investition bewerten (e)

```r
cat("Nutzen: 5 Punkte × 2000$/Punkt = 10.000 USD\n")
```

```
## Nutzen: 5 Punkte × 2000$/Punkt = 10.000 USD
```

```r
cat("Kosten: 25.000 USD\n")
```

```
## Kosten: 25.000 USD
```

```r
cat("Partielle Steigung ist relevant (2000$/Punkt)\n")
```

```
## Partielle Steigung ist relevant (2000$/Punkt)
```

Fazit: Der Kurs lohnt sich nur, wenn man länger als 2.5 Jahre im Unternehmen bleibt.

4 Aufgabe 3: Download (Netzwerk-Performance)

Aufgabenstellung: Erweitern Sie die Download-Analyse um die Variable “Stunden nach 8AM”.

4.1 Schritt 1: Daten einlesen und Korrelationen (a)

```r
download <- read_excel("../data/WDDA_06.xlsx", sheet = "Download")
names(download) <- c("time_sec", "size_mb", "hours_after_8", "vendor")

cor_download <- cor(download$time_sec, download$size_mb)
print(cor_download)
```

```
## [1] 0.790286
```

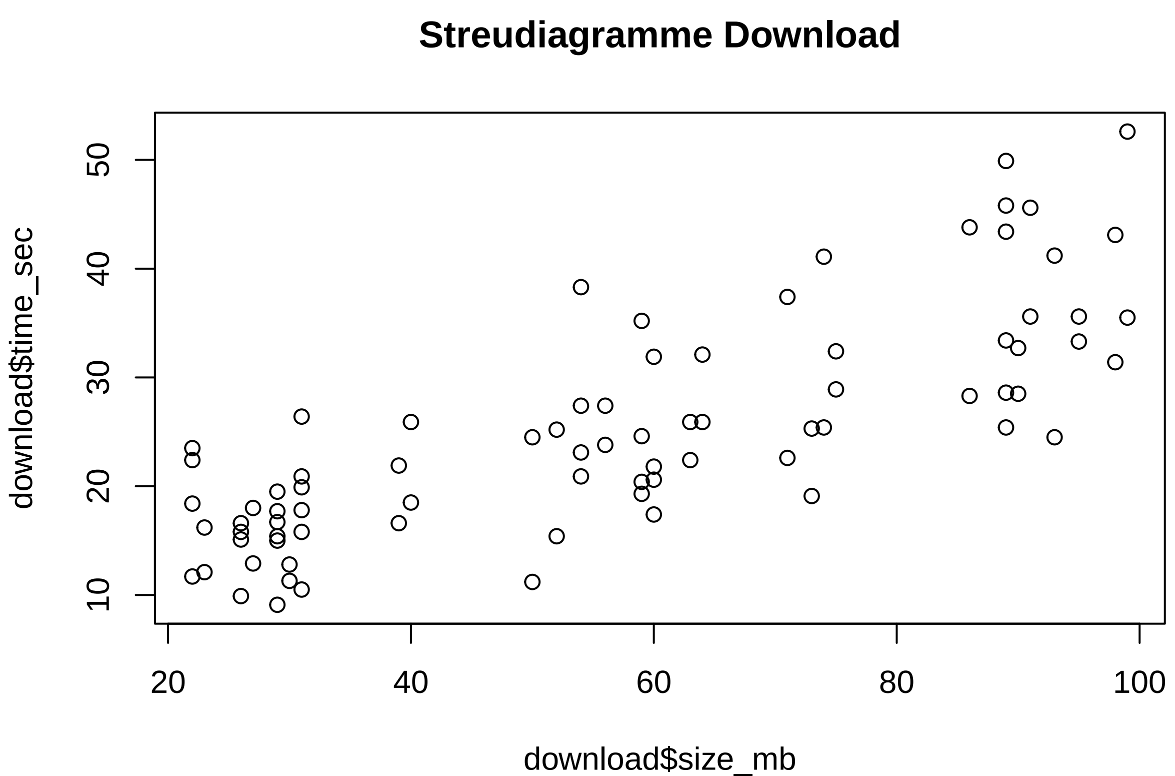
```r
cor_download2 <- cor(download$hours_after_8, download$size_mb)
print(cor_download2)
```

```
## [1] 0.988079
```

Wichtige Beobachtung: Dateigrösse und Stunden sind sehr stark korreliert (r 0.99)!

4.2 Schritt 2: Streudiagramme (b)

```r
plot(download$size_mb, download$time_sec, main = "Streudiagramme Download")
```



Bewertung: Lineare Beziehungen, aber starke Korrelation zwischen den erklärenden Variablen.

4.3 Schritt 3: Marginale vs. partielle Steigung (c-e)

```r
# Marginale Steigung
mod_size_simple <- lm(time_sec ~ size_mb, data = download)
marginal_slope <- coef(mod_size_simple)["size_mb"]

# Multiple Regression
mod_download <- lm(time_sec ~ size_mb + hours_after_8, data = download)
partial_slope <- coef(mod_download)["size_mb"]

cat("Marginale Steigung:", round(marginal_slope, 3), "s/MB\n")
```

```
## Marginale Steigung: 0.313 s/MB
```

```r
cat("Partielle Steigung:", round(partial_slope, 3), "s/MB\n")
```

```
## Partielle Steigung: 0.324 s/MB
```

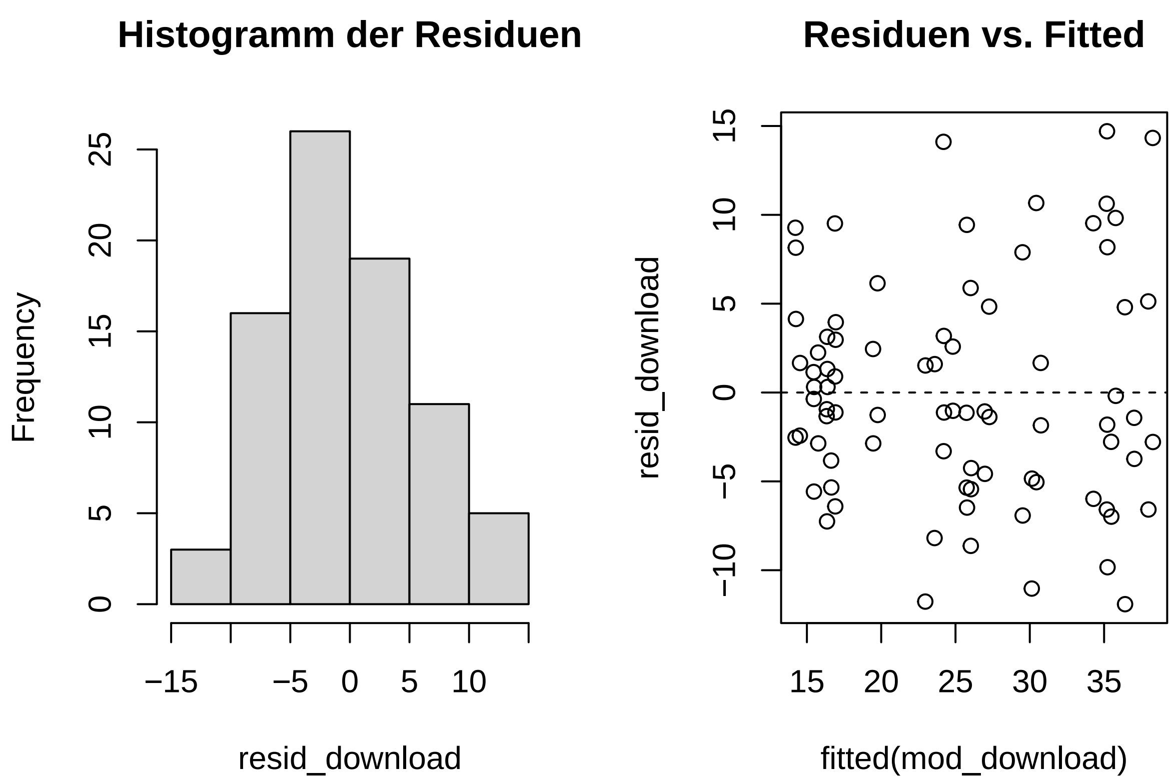
Erwartung: Wegen der starken Korrelation (r = 0.99) erwarten wir deutliche Unterschiede zwischen marginaler und partieller Steigung.

4.4 Schritt 4: Modell-Diagnostik (f-i)

```r
# Modell-Zusammenfassung
summary(mod_download)
```

```
##
## Call:
## lm(formula = time_sec ~ size_mb + hours_after_8, data = download)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.911  -4.644  -1.093   3.378  14.703
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.1388     2.8857   2.474   0.0156 *
## size_mb         0.3237     0.1798   1.800   0.0757 .
## hours_after_8  -0.1857     3.1619  -0.059   0.9533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.284 on 77 degrees of freedom
## Multiple R-squared:  0.6246, Adjusted R-squared:  0.6148
## F-statistic: 64.05 on 2 and 77 DF,  p-value: < 2.2e-16
```

```r
# Residuen-Analyse
resid_download <- resid(mod_download)
par(mfrow = c(1,2))
hist(resid_download, main = "Histogramm der Residuen")
plot(fitted(mod_download), resid_download, main = "Residuen vs. Fitted")
abline(h = 0, lty = 2)
```




```r
par(mfrow = c(1,1))
```

MRM-Bedingungen: - Linearität: OK - Konstante Varianz: OK - Normalität: OK - **Multikollinearität:** Problematisch (r = 0.99)

4.5 Schritt 5: Modellvergleich (n-o)

```r
# SRM vs MRM Vergleich
mod_download_srm <- lm(time_sec ~ size_mb, data = download)
r2_srm <- summary(mod_download_srm)$r.squared
r2_adj_srm <- summary(mod_download_srm)$adj.r.squared
r2_mrm <- summary(mod_download)$r.squared
r2_adj_mrm <- summary(mod_download)$adj.r.squared

cat("SRM R²:", round(r2_srm, 4), ", Adj-R²:", round(r2_adj_srm, 4), "\n")
```

```
## SRM R²: 0.6246 , Adj-R²: 0.6197
```

```
cat("MRM R²:", round(r2_mrm, 4), ", Adj-R²:", round(r2_adj_mrm, 4), "\n")
```

```
## MRM R²: 0.6246 , Adj-R²: 0.6148
```

Empfehlung: SRM bevorzugen wegen Multikollinearität.

5 Aufgabe 4: BFH (Körpergrösse-Modellierung)

Aufgabenstellung: Modellieren Sie die Körpergrösse mit verfügbaren Variablen im BFH-Datensatz.

5.1 Schritt 1: Mögliche erklärende Variablen (a-b)

```
bfh <- read_excel("../data/WDDA_06.xlsx", sheet = "BFH")
head(bfh)
```

```
## # A tibble: 6 x 23
##   class gender dob        height  foot  hair eyetext maths  cash house transport
##   <chr> <chr>  <chr>       <dbl> <dbl> <dbl> <chr>   <dbl> <dbl> <dbl> <chr>
## 1 2ab   Male   1992-08-28    178    26    20 Blau  ~   4    25.6    21 Bus
## 2 2ab   Male   1996-09-09    182    27    10 blau      4   250      14 Train
## 3 2xyz  Male   1997-02-06    174    26     3 braun     3.5  25      18 Bus
## 4 2xyz  Male   1983-09-04    181    27    10 schwarz   2    25     149 Bus
## 5 2xyz  Female 1997-07-15    164    26    43 brown     4    50      19 Other
## 6 2xyz  Male   1997-05-17    178    24    28 Braun     3.5  40      16 Train
## # i 12 more variables: costs <dbl>, distance <dbl>, postcode <dbl>, jar <dbl>,
## #   reaction1 <dbl>, reaction2 <dbl>, siblings <dbl>, present <dbl>,
## #   sleep <dbl>, handed <chr>, eye <chr>, football <dbl>
```

Mögliche Variablen: - **gender:** Geschlecht beeinflusst Körpergrösse stark - **foot:** Fussgrösse korreliert biologisch mit Körpergrösse - **dob:** Alter könnte relevant sein - **siblings:** Genetische Faktoren - **sleep:** Weniger wahrscheinlich relevant

Beste Einzelwahl: foot (Fussgrösse) wegen starker biologischer Korrelation.

5.2 Schritt 2: MRM anpassen (c)

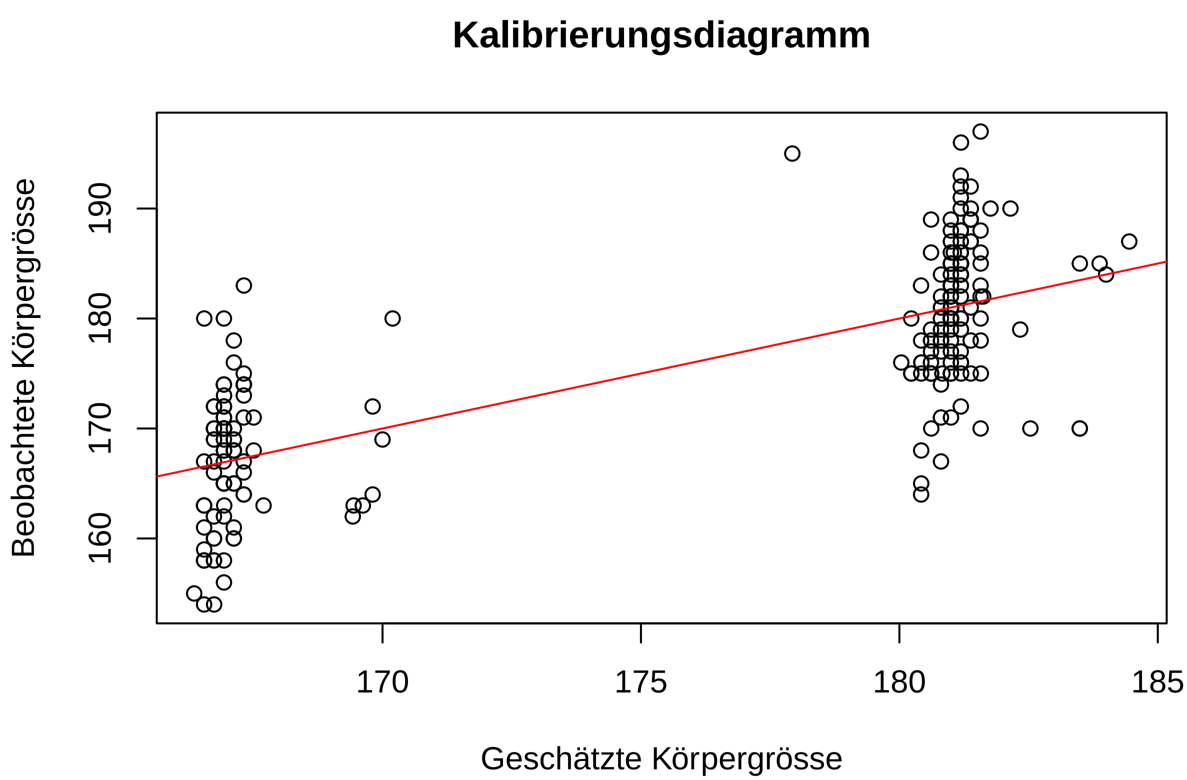
```
# Daten bereinigen
bfh_clean <- bfh[!is.na(bfh$height) & !is.na(bfh$foot) & !is.na(bfh$gender), ]
bfh_clean$age <- as.numeric(Sys.Date() - as.Date(bfh_clean$dob)) / 365.25

mod_bfh <- lm(height ~ foot + gender + age, data = bfh_clean)
summary(mod_bfh)
```

```
##
## Call:
## lm(formula = height ~ foot + gender + age, data = bfh_clean)
##
```

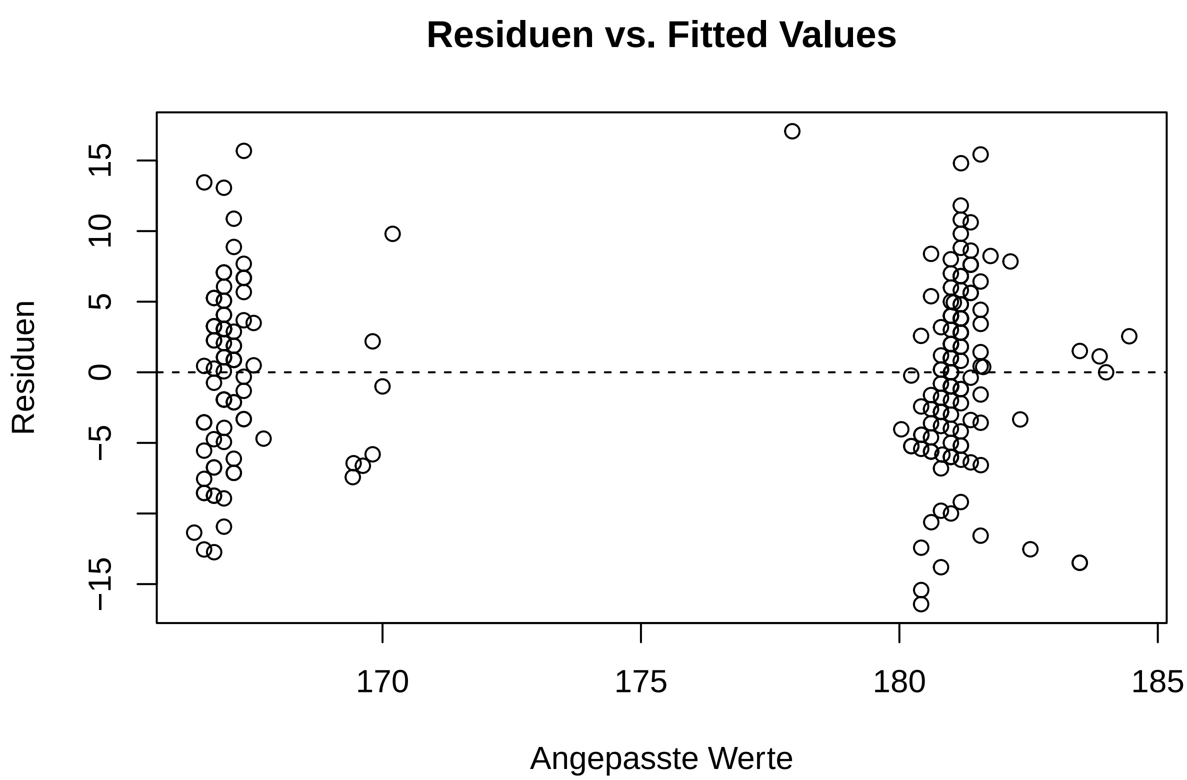
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.4205  -4.4247   0.3785   3.8134  17.0714
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.623e+02  2.859e+00  56.781   <2e-16 ***
## foot               1.918e-01  1.094e-01   1.753   0.0810 .
## genderMale         1.349e+01  9.132e-01  14.772   <2e-16 ***
## genderNon binary 1.630e+01  6.336e+00   2.573   0.0108 *
## age                5.053e-04  3.168e-03   0.160   0.8734
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.293 on 212 degrees of freedom
## Multiple R-squared:  0.5499, Adjusted R-squared:  0.5414
## F-statistic: 64.75 on 4 and 212 DF,  p-value: < 2.2e-16
```

5.3 Schritt 3: Modell-Diagnostik (d-f)

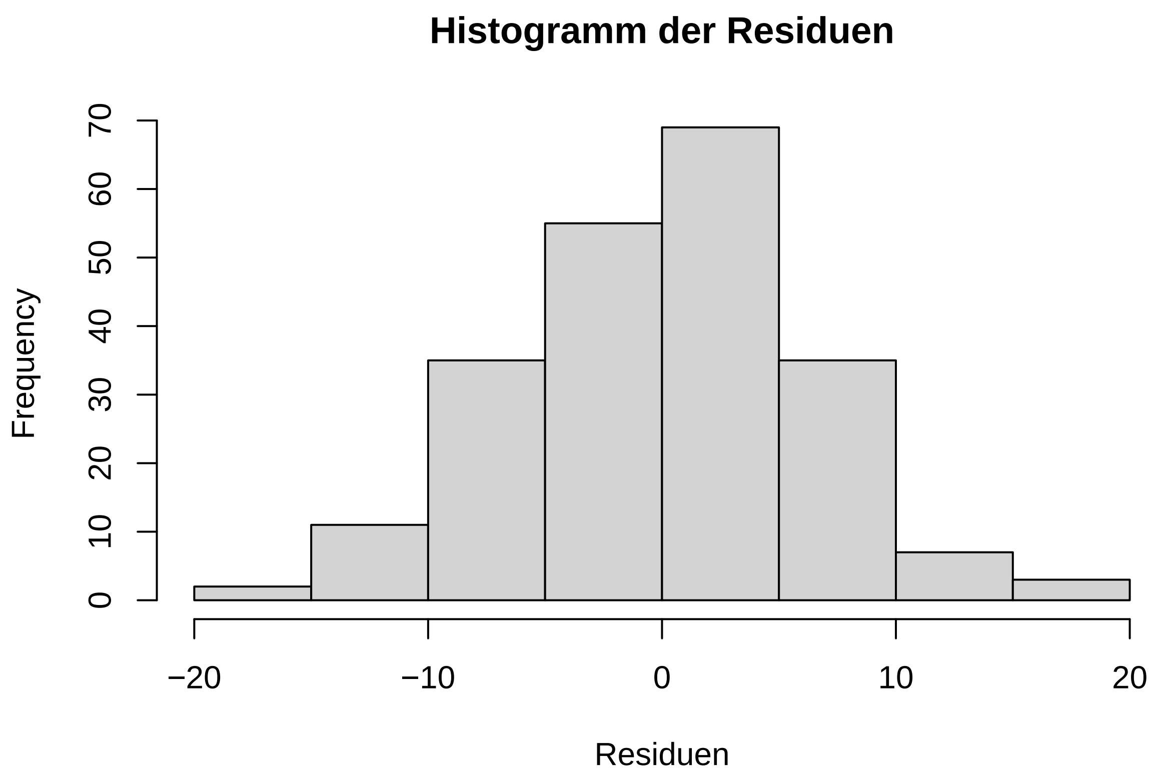
```r
# Linearität beurteilen
plot(fitted(mod_bfh), bfh_clean$height,
     main = "Kalibrierungsdiagramm",
     xlab = "Geschätzte Körpergrösse", ylab = "Beobachtete Körpergrösse")
abline(0, 1, col = "red")
```



```r
# Konstante Streuung
resid_bfh <- resid(mod_bfh)
plot(fitted(mod_bfh), resid_bfh,
     main = "Residuen vs. Fitted Values",
     xlab = "Angepasste Werte", ylab = "Residuen")
abline(h = 0, lty = 2)
```




```r
# Normalität
hist(resid_bfh, main = "Histogramm der Residuen", xlab = "Residuen")
```



5.4 Schritt 4: Modell-Optimierung (g-i)

```
# Modell-Bewertung
r2_bfh <- summary(mod_bfh)$r.squared
cat("R² =", round(r2_bfh, 4), "\n")
```

```
## R² = 0.5499
```

```
# Schrittweise Regression für optimale Variablenkombination
mod_step <- step(mod_bfh, direction = "both", trace = FALSE)
summary(mod_step)
```

```
##
## Call:
## lm(formula = height ~ foot + gender, data = bfh_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.4323  -4.4323   0.4214   3.8035  17.0514
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       162.3476     2.8439  57.085   <2e-16 ***
## foot                0.1910     0.1091   1.752   0.0813 .
## genderMale         13.4995     0.9090  14.851   <2e-16 ***
## genderNon binary   16.3030     6.3218   2.579   0.0106 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.279 on 213 degrees of freedom
## Multiple R-squared:  0.5498, Adjusted R-squared:  0.5435
## F-statistic: 86.72 on 3 and 213 DF,  p-value: < 2.2e-16
```

Interpretation der Koeffizienten: - **foot:** Pro cm Fusslänge steigt die Körpergrösse um X cm - **gender:** Geschlechtsunterschied in der Körpergrösse - **age:** Alterseffekt (falls signifikant)

6 Zusammenfassung

Diese Aufgabenserie führt in die **Multiple Regression** ein und zeigt wichtige Konzepte:

1. **Marginale vs. partielle Steigungen**
2. **Multikollinearität** und ihre Auswirkungen
3. **Modell-Diagnostik** für MRM
4. **Pfaddiagramme** zur Visualisierung komplexer Beziehungen
5. **Modellvergleich** und -optimierung

Wichtige Erkenntnisse: - Korrelationen zwischen erklärenden Variablen können Interpretationen erschweren - Residuen-Analyse ist entscheidend für Modellvalidierung - Nicht immer ist das komplexeste Modell das beste