Regression zweier metrischer Variablen – Bierdaten

Daniel Stepanovic

Ausgangssituation

Der vorliegende Datensatz enthält Informationen zu 34 verschiedenen Biersorten, die in Österreich im Jahr 2016 über die Brau Union vertrieben wurden (Quelle: Open Data Portal Österreich).

Für jede Biersorte wurden die folgenden Merkmale erhoben:

Variablen	Beschreibung
Bier_name	Bezeichnung der Biersorte (nur zur Identifikation, nicht Bestandteil der Analyse)
Stammwuerze	Gehalt an Extraktstoffen vor der Gärung, gemessen in Grad Plato (°P)
Vol_alkohol	Alkoholgehalt in Volumenprozent (%)
Kalorien_je_100ml	Kaloriengehalt des Bieres je 100 ml in Kilokalorien (kcal)

Ziel der Analyse ist es, den Kaloriengehalt eines Bieres mithilfe der Prädiktoren **Stammwürze** und **Alkoholgehalt** zu modellieren. Es soll überprüft werden, wie stark diese beiden Merkmale den Energiegehalt beeinflussen. Zusätzlich wird für eine bestimmte Biersorte ("Gösser Märzen") eine Vorhersage getroffen und im Kontext der übrigen Daten interpretiert.

Die Daten stammen aus dem Datensatz `wi21b043.txt` und wurden zur Bearbeitung dieser Übung von `wi23b095` verwendet.

Datenmanagement

Die Daten wurden mit **read.table()** eingelesen:

```r
options(scipen = 9999)
bierdaten = read.table("wi21b043.txt", sep = ";", header = TRUE)
head(bierdaten)
```

```
                      Bier_name Stammwuerze Vol_alkohol Kalorien_je_100ml
1    Bierspezialitäten Winterbock        16.2         7.1                45
2 Puntigamer Das \\bierige\\ Bier        11.5         5.1                40
3             Schladminger Märzen        11.7         5.1                43
4          Zipfer Limetten Radler         9.6         2.0                38
5           Gösser Dunkles Zwickl        13.2         5.7                50
6       Starobrno Altbrünner Gold        11.5         5.0                43
```

Wir überprüfen die Stichprobe auf fehlende Werte:

```r
nrow(bierdaten)
```

```
[1] 34
```

```r
summary(bierdaten)
```

```
  Bier_name          Stammwuerze     Vol_alkohol    Kalorien_je_100ml
 Length:34          Min.   : 6.50   Min.   :0.400   Min.   :16.00
 Class :character   1st Qu.:11.50   1st Qu.:4.800   1st Qu.:40.00
 Mode  :character   Median :12.05   Median :5.100   Median :43.00
                    Mean   :11.98   Mean   :4.748   Mean   :42.06
                    3rd Qu.:12.72   3rd Qu.:5.500   3rd Qu.:45.00
                    Max.   :16.20   Max.   :7.100   Max.   :60.00
                    NA's   :2       NA's   :3       NA's   :1
```

```r
bierdaten_clean <- subset(bierdaten, complete.cases(bierdaten))
```

```r
nrow(bierdaten)          # Ursprüngliche Stichprobe
```

```
[1] 34
```

```r
nrow(bierdaten_clean)    # Bereinigte Stichprobe
```

```
[1] 28
```

```r
nrow(bierdaten) - nrow(bierdaten_clean)  # Anzahl der ausgeschlossenen Beobachtungen
```

```
[1] 6
```

Die ursprüngliche Stichprobe umfasst 34 Biersorten. Es liegen jedoch fehlende Werte in den Variablen Stammwuerze (2 Fälle), Vol_alkohol (3 Fälle) und Kalorien_je_100ml (1 Fall) vor. Für die weitere Analyse wurden nur vollständige Beobachtungen verwendet, sodass 28 Biersorten in die Auswertung eingehen.

Übersicht

```r
summary(bierdaten_clean)
```

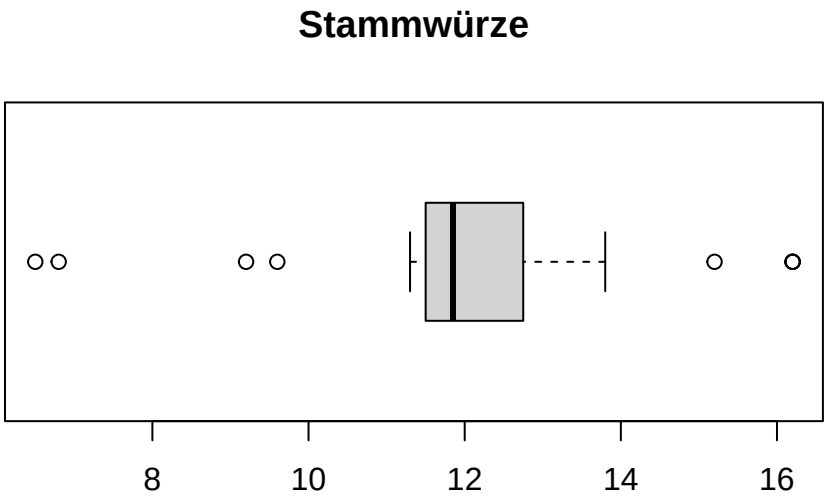
```
  Bier_name          Stammwuerze     Vol_alkohol    Kalorien_je_100ml
 Length:28          Min.   : 6.50   Min.   :0.400   Min.   :16.00
 Class :character   1st Qu.:11.50   1st Qu.:4.775   1st Qu.:40.00
 Mode  :character   Median :11.85   Median :5.100   Median :43.00
                    Mean   :12.07   Mean   :4.693   Mean   :42.54
                    3rd Qu.:12.72   3rd Qu.:5.425   3rd Qu.:45.00
                    Max.   :16.20   Max.   :7.100   Max.   :60.00
```

Die Stammwürze reicht von etwa 6,5 bis 16,2°P. Der Alkoholgehalt schwankt zwischen 0,4% bis 7,1%. Der Kaloriengehalt liegt zwischen 16 und 60 kcal je 100ml.

Visualisierung

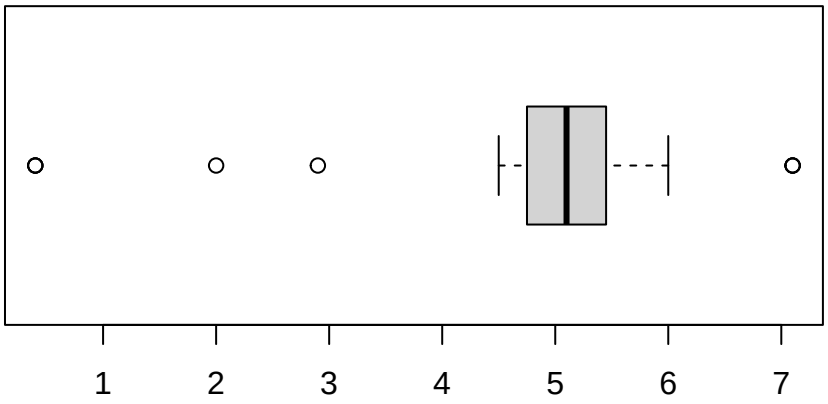
Boxplots der Variablen

```r
boxplot(bierdaten_clean$Stammwuerze, horizontal = TRUE, main ="Stammwürze")
```



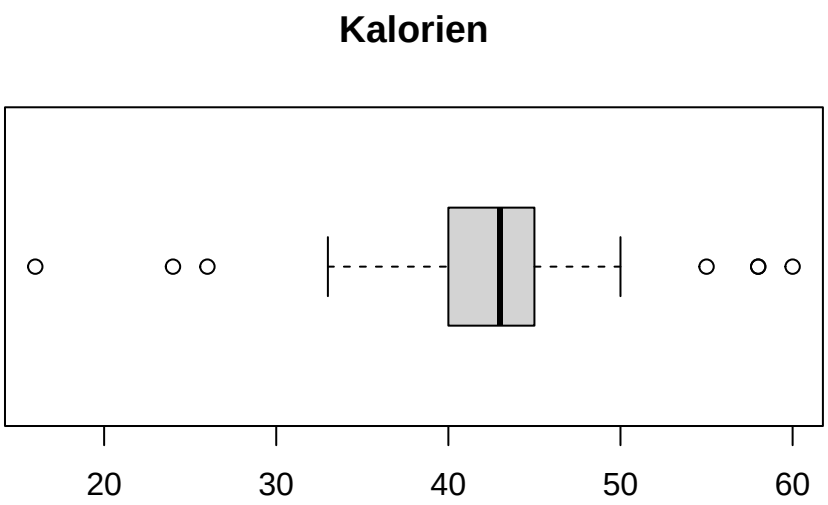
Die Verteilung der **Stammwürze** ist annähernd symmetrisch mit mehreren Ausreißern an beiden Enden. Der Mittelwert liegt bei 12,07 Grad Plato, der Median bei 11,85.

```
boxplot(bierdaten_clean$Vol_alkohol, horizontal = TRUE, main ="Alkoholgehalt")
```



Der **Alkoholgehalt** zeigt eine leicht linksschiefe Verteilung mit mehreren Ausreißern im Bereich unter 2%. Der Mittelwert liegt bei rund 4,69 Vol. %, der Median bei 5,1 Vol. %.

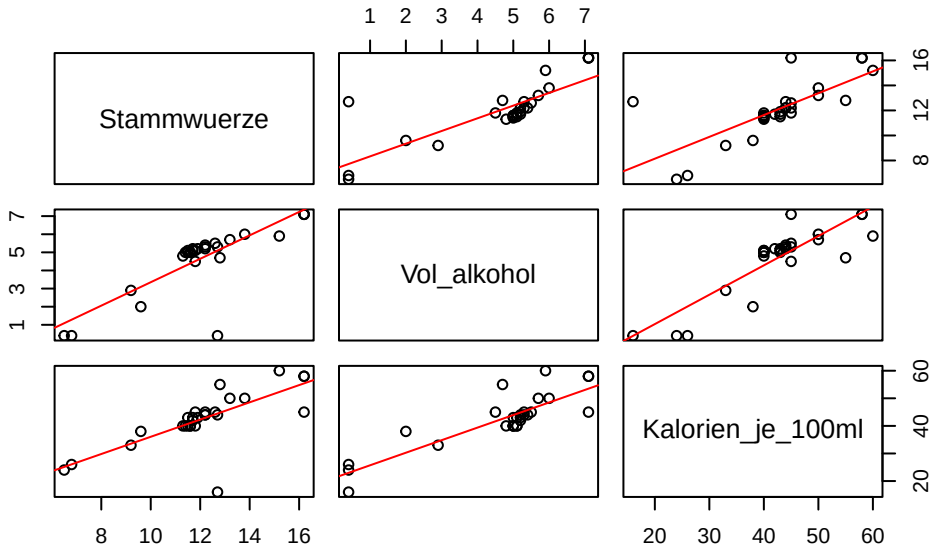
```
boxplot(bierdaten_clean$Kalorien_je_100ml, horizontal = TRUE, main ="Kalorien")
```



Die **Kalorienverteilung** ist leicht linksschief. Einige Biere weisen einen besonders hohen Kaloriengehalt (über 50 kcal) auf. Der Mittelwert liegt bei 42,54 kcal, der Median bei 43 kcal je 100 ml.

Pairs-Plot mit Regressionsgeraden

```
pairs(bierdaten_clean[, c("Stammwuerze", "Vol_alkohol", "Kalorien_je_100ml")],
        panel = function(x, y) {
        points(x, y)
        abline(lm(y ~ x), col = "red")
      })
```



Die Grafiken deuten auf lineare Zusammenhänge zwischen allen drei Variablen hin:

- Zwischen **Stammwürze** und **Kaloriengehalt** zeigt sich ein klar positiver Zusammenhang – höhere Stammwürze führt tendenziell zu mehr Kalorien.
- Auch der **Alkoholgehalt** hängt positiv mit dem **Kaloriengehalt** zusammen, wenn auch etwas schwächer.
- Zwischen **Stammwürze** und **Alkoholgehalt** besteht eine sehr starke lineare Korrelation, was auf Multikollinearität hinweisen könnte – das wird bei der Modellwahl berücksichtigt.

```
cor(bierdaten_clean$Stammwuerze, bierdaten_clean$Vol_alkohol)
```

```
[1] 0.809924
```

Ein Korrelationskoeffizient von r 0,81 bestätigt eine starke Multikollinearität zwischen Stammwürze und Alkoholgehalt.

Modellschätzung

```r
modell <- lm(Kalorien_je_100ml ~ Stammwuerze + Vol_alkohol, data = bierdaten_clean)
summary(modell)
```

```
Call:
lm(formula = Kalorien_je_100ml ~ Stammwuerze + Vol_alkohol, data = bierdaten_clean)

Residuals:
   Min     1Q Median     3Q    Max
-9.574 -2.779 -0.902  2.964 12.059

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3692     5.9877   2.901 0.007653 **
Stammwuerze   0.5192     0.7308   0.710 0.483972
Vol_alkohol   4.0268     0.9153   4.399 0.000177 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.119 on 25 degrees of freedom
Multiple R-squared:  0.7434,    Adjusted R-squared:  0.7229
F-statistic: 36.22 on 2 and 25 DF,  p-value: 0.00000004121
```

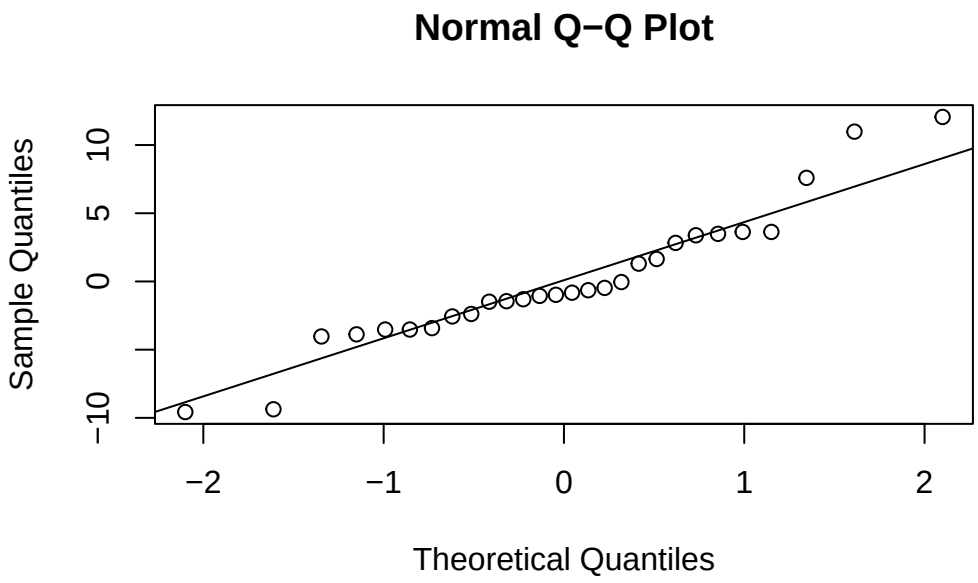
Sowohl der Achsenabschnitt als auch der Koeffizient von Vol_alkohol sind signifikant auf dem 0,05-Niveau, während der Einfluss der Stammwuerze statistisch nicht signifikant ist. Das Bestimmtheitsmaß beträgt 0,74, das Modell erklärt somit rund 74% der Varianz der Kaloriemenge pro 100ml Bier. Die Regressionsgleichung für den erwarteten Kaloriengehalt, gegeben Stammwürze und Alkoholgehalt, lautet daher:

$$\text{E (Kalorien\_je\_100ml} \quad \text{Stammwuerze, Vol\_alkohol)} = 17{,}37 + 0{,}52\,\text{Stammwuerze} + 4{,}03\,\text{Vol\_alkohol}$$

Der Kaloriengehalt steigt demnach im Mittel um etwa 4 kcal pro zusätzliches Volumenprozent Alkohol, während der Einfluss der Stammwürze vernachlässigbar ist. Insgesamt liefert das Modell eine gute Vorhersagequalität, wobei insbesondere der Alkoholgehalt eine zentrale Rolle spielt.

Q-Q-Plot

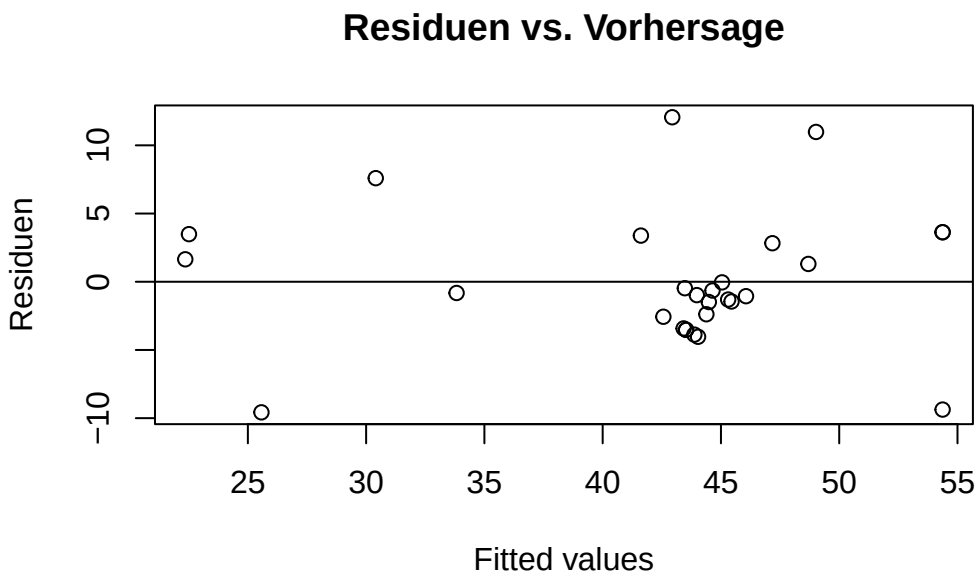
```r
qqnorm(residuals(modell))
qqline(residuals(modell))
```



Der Q-Q-Plot zeigt eine weitgehende Übereinstimmung mit der Normalverteilung. Leichte Abweichungen an den Rändern sind erkennbar, aber tolerierbar.

Residuen vs. Fitted Values

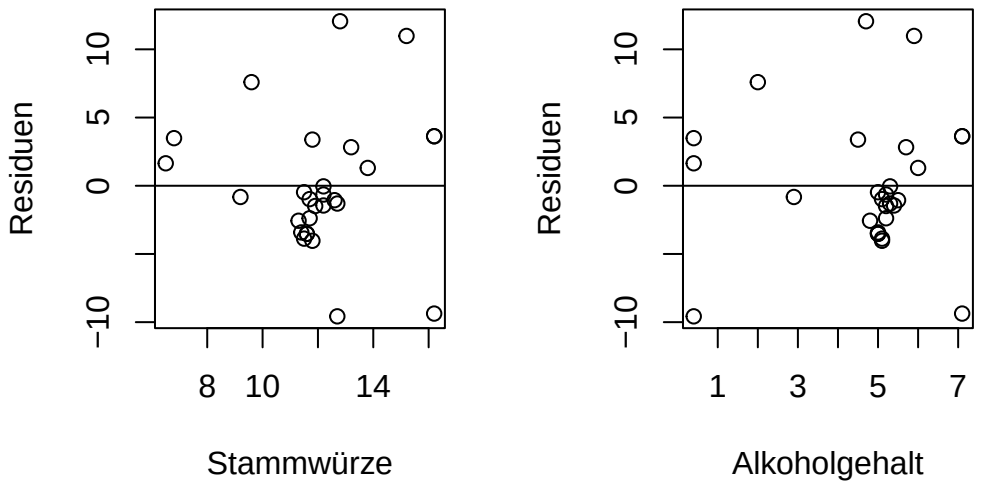
```r
plot(residuals(modell) ~ fitted(modell),
     main = "Residuen vs. Vorhersage",
     ylab = "Residuen", xlab = "Fitted values")
abline(h = 0)
```



Die Residuen sind überwiegend zufällig um die Nulllinie verteilt. Es ist keine systematische Struktur erkennbar, sodass die Annahme der Homoskedastizität (konstanten Varianz) als erfüllt gelten kann.

Residuen vs. Prädiktoren

```
par(mfrow = c(1, 2))
plot(residuals(modell) ~ bierdaten_clean$Stammwuerze, ylab = "Residuen", xlab = "Stammwürze"
abline(h = 0)
plot(residuals(modell) ~ bierdaten_clean$Vol_alkohol, ylab = "Residuen", xlab = "Alkoholgeha
abline(h = 0)
```




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

Auch gegenüber den Prädiktoren zeigen sich keine systematischen Muster – die Modellannahmen sind erfüllt.

Modellwahl und Fazit

```
modell_step <- step(modell)
```

```
Start:  AIC=94.27
Kalorien_je_100ml ~ Stammwuerze + Vol_alkohol

              Df Sum of Sq     RSS     AIC
- Stammwuerze  1     13.23  668.24  92.828
<none>                      655.01  94.268
- Vol_alkohol  1    507.06 1162.07 108.321

Step:  AIC=92.83
Kalorien_je_100ml ~ Vol_alkohol
```

```
              Df Sum of Sq     RSS     AIC
<none>                       668.24  92.828
- Vol_alkohol  1    1884.7 2552.96 128.359
```

```
summary(modell_step)
```

```
Call:
lm(formula = Kalorien_je_100ml ~ Vol_alkohol, data = bierdaten_clean)

Residuals:
   Min     1Q Median     3Q    Max
-8.497 -3.251 -1.256  2.912 12.432

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  21.1668     2.6730   7.919 0.00000002141 ***
Vol_alkohol   4.5535     0.5317   8.563 0.00000000482 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.07 on 26 degrees of freedom
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7282
F-statistic: 73.33 on 1 and 26 DF,  p-value: 0.000000004823
```

Das reduzierte Modell enthält nur noch Vol_alkohol als Prädiktor. Stammwürze wurde entfernt, da sie keine zusätzliche Erklärungskraft liefert. Die Modellgüte bleibt nahezu unverändert ($R^2 = 0{,}738$).

Fazit

Die Analyse zeigt, dass der Alkoholgehalt ein starker und signifikanter Einflussfaktor auf den Kaloriengehalt ist. Die Stammwürze korreliert stark mit dem Alkoholgehalt, liefert jedoch keine zusätzliche Erklärungskraft. Die Modellannahmen sind erfüllt, und das vereinfachte Modell mit nur einem Prädiktor ist genauso leistungsfähig wie das vollständige Modell. Es eignet sich daher besser für Prognosezwecke.

Die Biersorte “Gösser Märzen” lässt sich mit dem Modell sehr gut vorhersagen und zeigt ein typisches Kalorienprofil.

Prognose – Gösser Märzen

Für „Gösser Märzen“ (Stammwürze = 11,9, Alkohol = 5,2):

```r
gosser <- data.frame(Stammwuerze = 11.9, Vol_alkohol = 5.2)
predict(modell_step, newdata = gosser, interval = "confidence")
```

```
       fit      lwr      upr
1 44.84499 42.79912 46.89087
```

Der beobachtete Wert liegt bei 43 kcal/100 ml. Das finale Modell (nur mit Alkoholgehalt als Prädiktor) prognostiziert für Gösser Märzen etwa 44,84 kcal/100 ml. Gösser Märzen liegt somit im erwarteten Bereich und zeigt kein auffälliges Kalorienverhalten.