

Beispiel in R: Modellwahl im einfachen linearen Regressionsmodell

Regina Tüchler

2006-10-09

Wir betrachten die Variablen `FBAG` und `MBAG` aus dem Datensatz `statlab`:

```
> load("statlab.rda")  
> attach(statlab)
```

Streudiagramm:

Im Streudiagramm Fig. 1 sehen wir die positive Koppelung zwischen den beiden Variablen.

```
> plot(MBAG, FBAG, pch = 19, cex = 0.4)
```

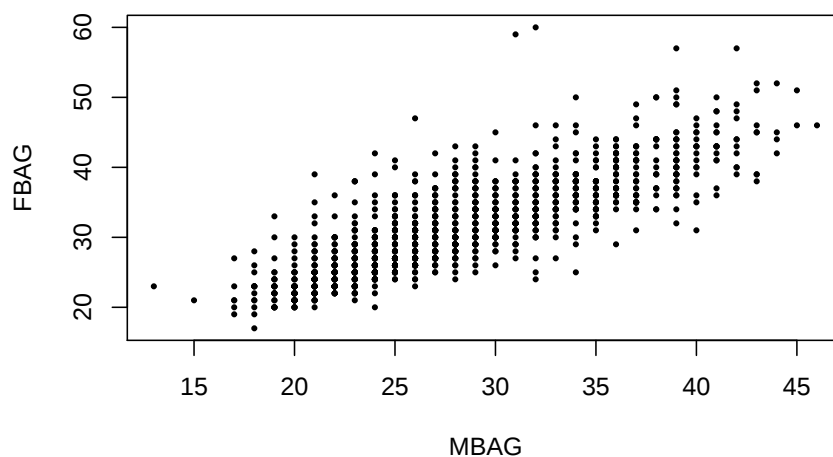


Figure 1: Streudiagramm von MBAG gegen FBAG.

Einfaches lineares Regressionsmodell:

$$\hat{FBAG}_i = \hat{\beta}_1 + \hat{\beta}_2 MBAG_i$$

Wir speichern das Ergebnis unter dem Namen `fm` ("fitted model") ab:

```
> fm <- lm(FBAG ~ MBAG)
```

Call:

```
lm(formula = FBAG ~ MBAG)
```

Coefficients:

```
(Intercept)      MBAG
      5.5638      0.9176
```

Die Regressionsgerade lautet daher:

$$\hat{FBAG}_i = 5.56 + 0.92 \cdot MBAG_i$$

In Fig. 2 wurde in das schon bestehende Streudiagramm die Regressionsgerade als rote Linie einzeichnen.

```
> plot(MBAG, FBAG, pch = 19, cex = 0.4)
> abline(fm, col = "red")
```

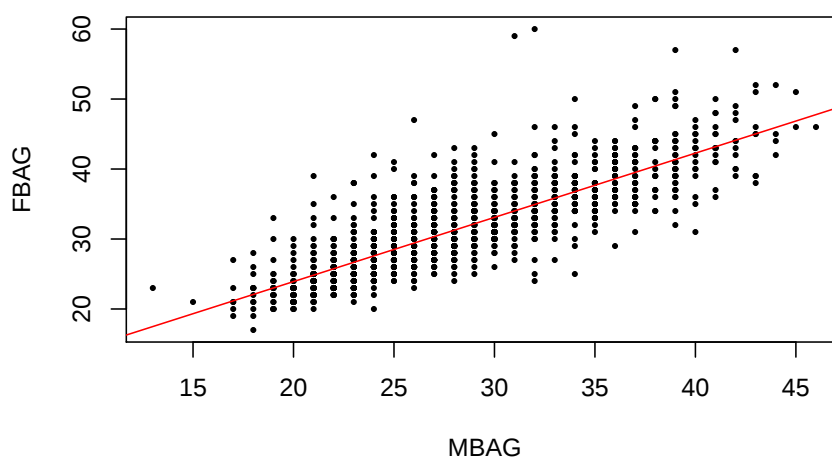


Figure 2: Regressionsgerade.

ANOVA:

Zur Durchführung der Varianzanalyse steht in R der Befehl `anova` zur Verfügung. Das Ergebnis speichern wir unter dem Namen `av` ab:

```
> av <- anova(fm)
```

Analysis of Variance Table

Response: FBAG

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MBAG	1	39285	39285	2797.6	< 2.2e-16 ***
Residuals	1294	18171	14		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Es wird die ANOVA-Tabelle ausgegeben (vgl. Folien). Hinter den p-Werten (Pr(>F)) stehen im Fall von signifikanten Testgrößen Zeichen für das Ausmaß der Signifikanz.

Die F-Größe beträgt 2797.65, beim dazugehörigen p-Wert weisen die 3 Sterne auf eine hoch signifikante Testgröße hin. Die Alternativhypothese, dass das einfache lineare Regressionsmodell signifikant mehr zur Erklärung von FBAG beiträgt, wurde also statistisch bewiesen.

Bestimmtheitsmaß, Schwankungsbreite und t-Größe:

Die Funktion `summary` gibt für das Regressionsmodell alle wichtigen Statistiken aus. Wir speichern ihr Ergebnis unter `sumy` ab:

```
> sumy <- summary(fm)

Call:
lm(formula = FBAG ~ MBAG)

Residuals:
    Min       1Q   Median       3Q      Max
-11.7614  -2.5032  -0.7505   1.8553  25.0738

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.56378     0.50148   11.10  <2e-16 ***
MBAG           0.91758     0.01735   52.89  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.747 on 1294 degrees of freedom
Multiple R-Squared:  0.6837,    Adjusted R-squared:  0.6835
F-statistic: 2798 on 1 and 1294 DF,  p-value: < 2.2e-16
```

Die aus den Folien bekannten Maßzahlen sind die folgenden: Unter **Coefficients** stehen zu den Schätzwerte der Regressionsparameter die Standardabweichung (**Std.Error**), die t-Werte und die dazugehörigen p-Werte. In der letzten Spalte stehen wieder Zeichen für etwaige signifikante Testgrößen. Dabei ist für uns laut Folien nur der Signifikanztest für β_2 von Bedeutung.

In unserem Beispiel haben wir eine Stichprobengröße von 1296. Daher können wir die kritischen Werte bei der Berechnung der Konfidenzintervalle aus der Standardnormalverteilung wählen. Wir wählen die Sicherheit 95 % mit den ungefähren kritischen Wert $c_\alpha = 2$ und bekommen die folgenden Konfidenzintervalle für die beiden Parameter:

$$\begin{aligned}\beta_1 &\in [\hat{\beta}_1 - 2 \cdot SD_1; \hat{\beta}_1 + 2 \cdot SD_1] = [5.56 - 2 \cdot 0.5; 5.56 + 2 \cdot 0.5] = [4.56; 6.57] \\ \beta_2 &\in [\hat{\beta}_2 - 2 \cdot SD_2; \hat{\beta}_2 + 2 \cdot SD_2] = [0.92 - 2 \cdot 0.02; 0.92 + 2 \cdot 0.02] = [0.88; 0.95]\end{aligned}$$

Der Test auf Signifikanz des Parameters β_2 kann mit Hilfe der t-Größe durchgeführt werden. Sie lautet hier 52.89. Der dazugehörige p-Wert ist fast 0 und durch die 3 Sterne wird angezeigt, dass eine hoch signifikante t-Größe vorliegt. Die Nullhypothese, dass der Koeffizient β_2 gleich 0 gesetzt werden kann, wird also verworfen.

Das Bestimmtheitsmaß entspricht ja dem Korrelationskoeffizienten zum Quadrat, der ebenfalls ausgegeben wird: `Multiple R-Squared = 0.68`.