Beispiel in R: Modellwahl im multiplen Regressionsmodell

Regina Tüchler

2006-10-09

Wir betrachten aus dem Datensatz `statlab` die Responsevariable `CTWGT`, die mit Hilfe eines multiplen Regressionsmodells durch Prädiktoren `MBAG` und `FBAG` erklärt werden soll.

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

Es soll nun die Variablenwahl aus diesen Prädiktoren durchgeführt werden: Wir verwenden dabei die Notation zur Modellbezeichnung aus den Folien mit $x_{i2} = FBAG$, $x_{i3} = MBAG$.

1. Schritt: Modellvergleich M_1 mit M_{23}:

Wir berechnen die folgenden 2 Modelle

$$M_1 : C\hat{TWGT}_i = \hat{\beta}_1$$

$$M_{23} : C\hat{TWGT}_i = \hat{\beta}_1 + \hat{\beta}_2 FBAG_i + \hat{\beta}_3 MBAG_i$$

und speichern die Ergebnisse unter dem Namen `fm.1` bzw `fm.23` ab:

```
> fm.1 <- lm(CTWGT ~ 1)

Call:
lm(formula = CTWGT ~ 1)

Coefficients:
(Intercept)
      70.94

> fm.23 <- lm(CTWGT ~ FBAG + MBAG)

Call:
lm(formula = CTWGT ~ FBAG + MBAG)

Coefficients:
(Intercept)         FBAG         MBAG
   61.97228      0.07701      0.23115
```

Die R-Funktion `anova` vergleicht all jene Modelle miteinander, die im Argument angegeben werden.

```
> an <- anova(fm.1, fm.23)

Analysis of Variance Table

Model 1: CTWGT ~ 1
Model 2: CTWGT ~ FBAG + MBAG
```

```
  Res.Df    RSS   Df Sum of Sq      F   Pr(>F)
1   1295 252087
2   1293 247729    2      4358 11.373 1.27e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wir bekommen eine Tabelle mit:
In der ersten Spalte wird angegeben, welches Modell (hier 1 und 2) gerechnet wurde;
für diese Modelle wird in der Spalte `RSS` die Abweichungsquadratsumme SSE_M berechnet;
in der Spalte `Sum of Sq` wird die Differenz der Abweichungsquadratsummen der beiden Modelle berechnet: Für das Modell M_{23} stehen Prädiktoren zur Verfügung. Daher wird in diesem Modell sicher mehr erklärt und daher ist die $SSE_{M23} < SSE_{M_1}$. Die Frage, ob die Differenz $SSE_{M_1} - SSE_{M23}$ signifikant ist, und daher im statistischen Test bewiesen wird, dass das M_{23} zu wählen ist, wird mit Hilfe der F-Größe und dem dazugehörigen p-Wert beantwortet.

Im Beispiel lautet die F-Größe 11.37 und beim p-Wert zeigen die 3 Steren an, dass ein hoch signifikantes Ergebnis vorliegt. Daher wird die Nullhypothese: M_1 verworfen und das Modell M_{23} gewählt.

Man weiss jetzt, dass nicht beide Parameter β_2 und β_3 0 sind. Daher untersuchen wir nun Teilmodelle, um festzustellen, ob einer der Parameter 0 gesetzt werden soll:

2. Schritt: Modellvergleich M_1, M_2, M_{23}:

```
> fm.2 <- lm(CTWGT ~ FBAG)

Call:
lm(formula = CTWGT ~ FBAG)

Coefficients:
(Intercept)         FBAG
    63.0811       0.2493
```

Die ANOVA zum Vergleich dieser 3 Modelle lautet:

```
> an2 <- anova(fm.1, fm.2, fm.23)

Analysis of Variance Table

Model 1: CTWGT ~ 1
Model 2: CTWGT ~ FBAG
Model 3: CTWGT ~ FBAG + MBAG
  Res.Df    RSS   Df Sum of Sq       F    Pr(>F)
1   1295 252087
2   1294 248517    1      3569 18.6307 1.707e-05 ***
3   1293 247729    1       788  4.1151   0.04271 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Die Tabelle enthält zeilenweise die 3 Modelle.
in den Spalten `Sum of Sq` und `F` (mit p-Wert daneben) werden immer 2 Modelle miteinander verglichen:
Man beginnt mit der letzten Zeile: Vergleich von M_2 und M_{23}: Die F-Grösse hilft beim Testen von

Nullhypothese: M_2
Alternativhypothese: M_{23}

Da der dazugehörige p-Wert signifikant (auf dem 95 %-Niveau) ist, wird M_2 zugunsten des

Modells M_{23} verworfen. Der zu $MBAG$ gehörige Parameter β_3 ist also ungleich 0.

Es stellt sich jetzt nur mehr die Frage, ob vielleicht β_2 0 gesetzt werden soll:

3. Schritt: Modellvergleich M_1, M_3, M_{23}:

```
> fm.3 <- lm(CTWGT ~ MBAG)

Call:
lm(formula = CTWGT ~ MBAG)

Coefficients:
(Intercept)         MBAG
    62.4007       0.3018
```

Die ANOVA zum Vergleich dieser 3 Modelle lautet:

```
> an3 <- anova(fm.1, fm.3, fm.23)

Analysis of Variance Table

Model 1: CTWGT ~ 1
Model 2: CTWGT ~ MBAG
Model 3: CTWGT ~ FBAG + MBAG
  Res.Df    RSS   Df Sum of Sq       F    Pr(>F)
1   1295 252087
2   1294 247836    1      4250 22.1833 2.745e-06 ***
3   1293 247729    1       108  0.5624    0.4534
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In der letzten Zeile werden M_3 und M_{23} verglichen: Die F-Grösse hilft beim Testen von

Nullhypothese:	M_3
Alternativhypothese:	M_{23}

Da der dazugehörige p-Wert nicht signifikant ist, wird M_3 angenommen. β_2 wird also 0 gesetzt und das gewählte Modell ist das einfache lineare Regressionsmodell mit Prädiktor `MBAG`.

(Die F-Größe in der 2. Zeile testet Nullhypothese M_1 und Gegenhypothese M_3 und ist hoch signifikant.)