Lineare Modelle in R: Schrittweise Modellselektion

Achim Zeileis

2009-02-20

Wie schon im Tutorium *LiMo3.pdf* laden wir den `GSA` Datensatz

```
R> load("GSA.rda")
```

und wählen die Variablen aus, die wir verwenden wollen, lassen die fehlenden Werte weg und rekodieren `country`, das jetzt nur noch zwei Levels hat.

```
R> gsa <- GSA[, c(3, 5, 6, 8, 9)]
R> gsa <- subset(gsa, country == "USA" | country == "Netherlands")
R> gsa <- na.omit(gsa)
R> gsa$country <- factor(gsa$country)
```

Dann passen wir einen ersten Versuch eines Regressionsmodells an:

```
R> fm1 <- lm(log(expenditure) ~ country + log(income), data = gsa)
```

Dies nutzt von den vorhandenen potentiellen Erklärungsvariablen `country`, `income`, `accomodation` und `year` nur die ersten zwei. Es ist ein Regressionsmodell mit einem Steigungsparameter bzgl. `log(income)` und unterschiedlichen Achsenabschnitten nach `country`.

Um zu überprüfen, ob das Hinzufügen von zusätzlichen Erklärungsvariablen (oder das Weglassen verwendeter Variablen) das Modell verbessert, wird eine AIC-basierte schrittweise Modellwahl mit Hilfe von `step` durchgeführt. Dabei kann mit dem Argument `scope` das einfachste und kompliziertesteste der zu durchsuchenden Modelle angegeben werden. Die Listenkomponente `lower` gibt das kleinste denkbare Modell an, hier im Beispiel das triviale Modell. Die Listenkomponente `upper` gibt das größte denkbare Modell an, hier das Modell in dem alle Variablen interagieren, d.h. es wird von jede Kombination von `country`, `year` und `accomodation` ein eigener Achsenabschnitt und `log(income)`-Steigungsparameter geschätzt. Das Resultat von `step` ist wieder ein angepaßtes `"lm"`-Objekt und wird dem Objekt `fm2` zugewiesen.

```
R> fm2 <- step(fm1, scope = list(lower = log(expenditure) ~ 1, upper = log(expenditure) ~
+      country * year * accomodation * log(income)))

Start:  AIC=-1370.98
log(expenditure) ~ country + log(income)

                     Df Sum of Sq      RSS      AIC
+ accomodation        5    108.92   205.66 -1828.51
+ year                1      2.02   312.57 -1376.07
<none>                             314.59 -1370.98
+ country:log(income)  1 5.747e-06   314.59 -1368.98
```

```
- log(income)          1      7.48   322.07 -1347.13
- country              1    118.14   432.72 -1022.26

Step:  AIC=-1828.51
log(expenditure) ~ country + log(income) + accomodation

                          Df Sum of Sq      RSS      AIC
+ year                     1      0.44   205.22 -1828.87
<none>                                   205.66 -1828.51
+ country:log(income)      1      0.02   205.65 -1826.59
+ country:accomodation     5      0.46   205.21 -1820.95
+ accomodation:log(income) 5      0.22   205.45 -1819.67
- log(income)              1      4.59   210.26 -1806.22
- country                  1     20.83   226.50 -1724.37
- accomodation             5    108.92   314.59 -1370.98

Step:  AIC=-1828.87
log(expenditure) ~ country + log(income) + accomodation + year

                          Df Sum of Sq      RSS      AIC
+ country:year             1      3.10   202.13 -1843.59
+ year:accomodation        5      2.41   202.82 -1831.84
+ year:log(income)         1      0.50   204.72 -1829.56
<none>                                   205.22 -1828.87
- year                     1      0.44   205.66 -1828.51
+ country:log(income)      1      0.02   205.20 -1826.96
+ country:accomodation     5      0.49   204.73 -1821.51
+ accomodation:log(income) 5      0.22   205.01 -1820.02
- log(income)              1      4.06   209.28 -1809.33
- country                  1     21.11   226.33 -1723.17
- accomodation             5    107.35   312.57 -1376.07

Step:  AIC=-1843.59
log(expenditure) ~ country + log(income) + accomodation + year +
    country:year

                          Df Sum of Sq      RSS      AIC
<none>                                   202.13 -1843.59
+ country:log(income)      1      0.02   202.10 -1841.73
+ year:log(income)         1      0.01   202.12 -1841.63
+ country:accomodation     5      0.64   201.48 -1837.09
+ year:accomodation        5      0.60   201.53 -1836.86
+ accomodation:log(income) 5      0.26   201.87 -1835.01
- country:year             1      3.10   205.22 -1828.87
- log(income)              1      3.58   205.71 -1826.28
- accomodation             5    108.01   310.13 -1382.68
```

Im ersten Schritt wird zunächst reportiert, daß das aktuelle Modell ein AIC von -1370.98 hat. Die Hinzunahme der Variable `accomodation` verbessert dies sehr deutlich auf -1828.51 bei Schätzung von weiteren 5 Parametern. Die Hinzunahme von `year` würde auch zu einer Verbesserung führen, jedoch nicht zu so einer deutlichen. Dagegen würde die Hinzunahme der Interaktion `country:log(income)` das Modell bzgl. des AIC auf -1368.98 verschlechtern. Ebenso führt das Weglassen von `log(income)` und `country` zu jeweils deutlichen Verschlechterungen.

Deshalb wird also `accomodation` in das Modell aufgenommen und dieselbe Prozedur noch einmal

durchgeführt. In diesem Schritt führt nur die Hinzunahmen von `year` zu einer Verbesserung des Modells, weshalb diese Variable ebenfalls in das Modell aufgenommen wird.

Im darauffolgenden Schritt führt die Interaktion `country:year` zu einer Verbesserung. (Diese stand vorher nicht zur Auswahl, weil der zugehörige Haupteffekt `year` noch nicht im Modell vertreten war.)

Im letzten Schritt kann weder die Hinzunahme weiterer Interaktionen noch das Weglassen vorhandener Variablen bzw. deren Interaktionen zu einer Verbesserung des Modells bzgl. des AIC beitragen. Also ist die Schrittweise Modellwahl abgeschlossen. Das gewählte Modell

```
R> summary(fm2)

Call:
lm(formula = log(expenditure) ~ country + log(income) + accomodation +
    year + country:year, data = gsa)

Residuals:
      Min        1Q    Median        3Q       Max
-1.256955 -0.286229 -0.009696  0.264602  1.970420

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.74906    0.25185  22.827  < 2e-16 ***
countryUSA              0.25923    0.04205   6.165 9.89e-10 ***
log(income)             0.10729    0.02442   4.394 1.22e-05 ***
accomodationCamping    -0.90674    0.03816 -23.764  < 2e-16 ***
accomodationB&B        -0.30279    0.04803  -6.304 4.20e-10 ***
accomodationAppartment -0.56677    0.04666 -12.147  < 2e-16 ***
accomodationRoom       -0.44297    0.06057  -7.314 5.03e-13 ***
accomodationFarm       -0.36915    0.12716  -2.903  0.00377 **
year1997               -0.04848    0.03446  -1.407  0.15973
countryUSA:year1997     0.21715    0.05314   4.087 4.70e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4306 on 1090 degrees of freedom
Multiple R-squared: 0.5988,        Adjusted R-squared: 0.5954
F-statistic: 180.7 on 9 and 1090 DF,  p-value: < 2.2e-16
```

hat also eine Steigung bzgl. `log(income)` und verschiedene Achsenabschnitte bzgl. der Variablen `country`, `year` und `accomodation`. Diese sind parametrisiert über alle drei Haupteffekte sowie eine Interaktion zwischen `country` und `year`.

Zum Abschluß sollen auch wieder die angelegten Objekte aufgeräumt werden:

```
R> remove(GSA, gsa, fm1, fm2)
```