Lineare Modelle in R:
Zweiweg-Varianzanalyse und Kovarianzanalyse

Achim Zeileis

2009-02-20

1 Datenaufbereitung

Wie schon im Tutorium *LiMo2.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)
R> dim(gsa)

[1] 1100    5

R> attach(gsa)
```

Zuletzt wird dieser Datensatz für die weitere Analyse attached.

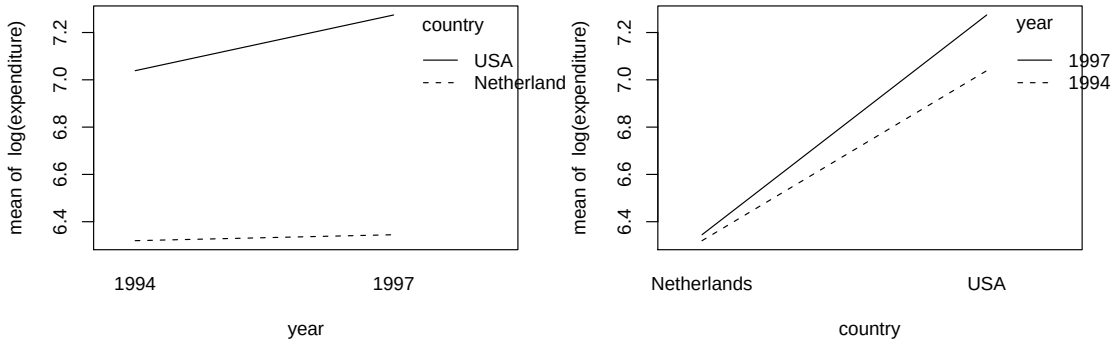
2 Zweiweg-Varianzanalyse

Als Beispiel für die Zweiweg-Analyse haben wir uns in der Lehrveranstaltung die Abhängigkeit der log-Ausgaben `log(expenditure)` von den Erklärungsvariablen `country` und `year` angeschaut. Die Mittelwerte in den vier Gruppen können durch `tapply` berechnet werden:

```
R> tapply(log(expenditure), list(year, country), mean)

     Netherlands      USA
1994    6.319761 7.038604
1997    6.344921 7.274130
```

Diese vier Mittelwerte können auch durch einen Interaktionsplot visualisiert werden:

```
R> interaction.plot(year, country, log(expenditure))
R> interaction.plot(country, year, log(expenditure))
```



In welcher Reihenfolge man die Variablen spezifiziert ist inhaltlich egal, aber oft ist eine der beiden Varianten leichter zu interpretieren. Die Visualisierung legt nahe, daß man eine Interaktion der beiden Variablen im linearen Modell berücksichtigen sollte, da offenbar 1. US-Amerikaner mehr ausgeben als Niederländer und 2. sich dieser Effekt von 1994 nach 1997 noch verstärkt.

Das zugehörige Modell wird in R angepaßt durch

```
R> fmCxY <- lm(log(expenditure) ~ country * year)
R> summary(fmCxY)

Call:
lm(formula = log(expenditure) ~ country * year)

Residuals:
      Min        1Q    Median        3Q       Max
-1.592087 -0.364250  0.007218  0.350061  2.111721

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.31976    0.02821 224.039  < 2e-16 ***
countryUSA           0.71884    0.04556  15.777  < 2e-16 ***
year1997             0.02516    0.04246   0.593  0.55361
countryUSA:year1997  0.21037    0.06597   3.189  0.00147 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5367 on 1096 degrees of freedom
Multiple R-squared: 0.3733,        Adjusted R-squared: 0.3716
F-statistic: 217.6 on 3 and 1096 DF,  p-value: < 2.2e-16
```

Aus der `summary` ist ablesbar, daß der Interaktionsterm klar signifikant ist ($p < 0.0015$). **Wichtig:** Daß hier der Koeffizient von `year` nicht signifikant ist, ist irrelevant. Solange der Interaktionsterm im Modell bleibt, verbleiben auch immer alle zugehörigen Haupteffekte im Modell. Denselben Test könnte man natürlich auch wieder per `anova` durchführen.

Die prognostizierten Werte für alle Kombinationen von `Netherlands`/`USA` und `1994`/`1997` kann man natürlich per `predict` ausrechnen:

```
R> predict(fmCxY, newdata = data.frame(country = factor(c("Netherlands",
+      "Netherlands", "USA", "USA")), year = factor(c("1994", "1997",
+      "1994", "1997"))))
```

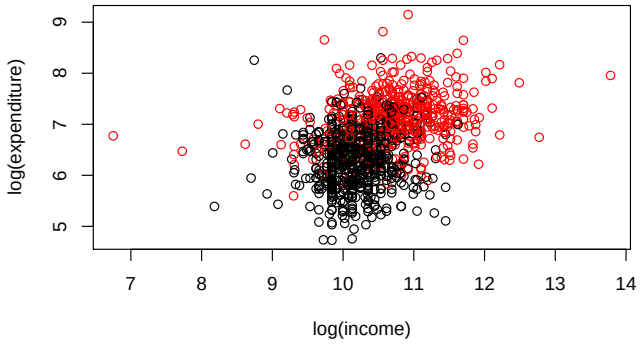
```
       1        2        3        4
6.319761 6.344921 7.038604 7.274130
```

Diese ergeben genau die vier empirischen Stichprobenmittel, die bereits oben durch `tapply` ausgerechnet wurden. Man kann diese vier Werte aber auch leicht 'zu Fuß' ausrechnen. Der `(Intercept)` Koeffizient 6.32 ist die Schätzung für die Referenzgruppe `Netherlands/1994`. Wenn man nur den USA-Effekt dazuaddiert $6.32 + 0.719 = 7.039$ erhält man die Schätzung für die Gruppe `USA/1994`. Wenn man nur den 1997-Effekt dazuaddiert $6.32 + 0.025 = 6.345$ erhält man die Schätzung für die Gruppe `Netherlands/1997`. Um die Schätzung für die letzte Gruppe `USA/1997` zu erhlaten, muß man den USA-Effek, den 1997-Effekt und deren Interaktion hinzuaddieren. Es ist also die Summe aller Koeffizienten: $6.32 + 0.719 + 0.025 + 0.21 = 7.274$.

3 Kovarianzanalyse

Als Beispiel für die Kovarianzanalyse versuchen wir die log-Ausgaben `log(expenditure)` durch das log-Einkommen `log(income)` und das Herkunftsland `country` zu erklären. Als Visualisierung dieses Problems wählen wir ein Streudiagramm, das nach dem Herkunftsland verschieden eingefärbt wird.

```
R> plot(log(expenditure) ~ log(income), col = as.numeric(country))
```



Hier ist für beide Länder ein gewisser Anstieg der Ausgaben mit dem Einkommen abzulesen. Außerdem ist sichtbar, daß die US-Amerikaner mehr ausgeben als die Niederländer. Um ein geeignetes Modell für die Daten zu finden, führen wir eine vollständige Suche durch und vergleiche alle Kandidaten vom trivialen Modell bis zum Interaktionsmodell:

```
R> fm1 <- lm(log(expenditure) ~ 1)
R> fmC <- lm(log(expenditure) ~ country)
R> fmI <- lm(log(expenditure) ~ log(income))
R> fmCI <- lm(log(expenditure) ~ country + log(income))
R> fmCxI <- lm(log(expenditure) ~ country * log(income))
```

Bisher hatten wir als Modellwahlkriterium immer den F-Test verwendet: entweder wenn `log(income)` als erster Haupteffekt hinzugefügt wird

```
R> anova(fm1, fmI, fmCI, fmCxI)
```

```
Analysis of Variance Table

Model 1: log(expenditure) ~ 1
Model 2: log(expenditure) ~ log(income)
Model 3: log(expenditure) ~ country + log(income)
Model 4: log(expenditure) ~ country * log(income)
  Res.Df    RSS   Df Sum of Sq         F Pr(>F)
1   1099 503.75
2   1098 432.72    1     71.02    247.45 <2e-16 ***
3   1097 314.59    1    118.14    411.58 <2e-16 ***
4   1096 314.59    1 5.747e-06 2.002e-05 0.9964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

oder wenn `country` als erster Haupteffekt hinzugefügt wird.

```
R> anova(fm1, fmC, fmCI, fmCxI)

Analysis of Variance Table

Model 1: log(expenditure) ~ 1
Model 2: log(expenditure) ~ country
Model 3: log(expenditure) ~ country + log(income)
Model 4: log(expenditure) ~ country * log(income)
  Res.Df    RSS   Df Sum of Sq         F    Pr(>F)
1   1099 503.75
2   1098 322.07    1    181.68   632.956 < 2.2e-16 ***
3   1097 314.59    1      7.48    26.067 3.886e-07 ***
4   1096 314.59    1 5.747e-06 2.002e-05    0.9964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Beide Varianten kommen zur selben Entscheidung: alle Variablen sind signifikant mit Ausnahme des Interaktionsterms. Auf Basis der F-Tests würde also das Haupteffektmodell `fmCI` gewählt.

Alternativ könnte man auch ein Informationskriterium zur Modellwahl verwenden. Diese werden in R von der Funktion `AIC` berechnet:

```
R> AIC(fm1, fmI, fmC, fmCI, fmCxI)

      df      AIC
fm1    2 2266.578
fmI    3 2101.402
fmC    3 1776.537
fmCI   4 1752.681
fmCxI  5 1754.681
```

Die erste Spalte zeigt dabei an, wie viele Parameter geschätzt wurden (= Anzahl Regressionskoeffizienten + Fehlervarianz), und die zweite Spalte das zugehörige AIC. Da es gilt, das Informationskriterium zu minimieren, würde sich also das AIC ebenfalls für das Modell `fmCI` entscheiden.

Auch das BIC kann mit der Funktion `AIC` berechnet werden, dabei muß nur der Koeffizient des Strafterms angegeben werden. Dieser ist beim AIC 2 und beim BIC $\log(n)$, hier also `log(1100)`. Wenn man dies als Argument `k` mit übergibt

```
R> AIC(fm1, fmI, fmC, fmCI, fmCxI, k = log(1100))
```

```
      df      AIC
fm1    2 2276.584
fmI    3 2116.411
fmC    3 1791.546
fmCI   4 1772.693
fmCxI  5 1779.696
```

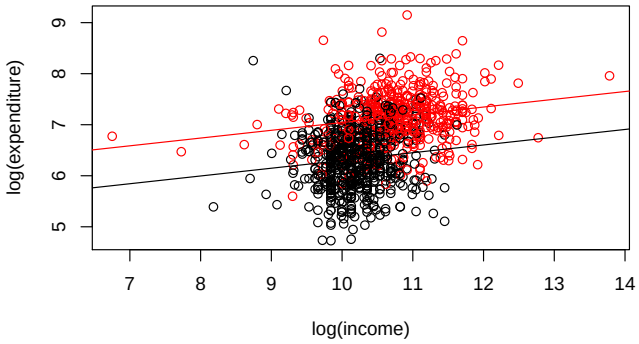
erhält man die BIC-Werte für die verschiedenen Modelle. Auch das BIC würde hier also das Modell `fmCI` wählen. **Bemerkung:** Hier sind sich alle drei Modellwahlmethoden einig, aber das muß nicht immer so sein. Grundsätzlich könnten diese auch zu unterschiedlichen Modellen gelangen.

Da sie hier alle dasselbe Modell wählen, wollen wir dies noch einmal näher anschauen:

```
R> summary(fmCI)

Call:
lm(formula = log(expenditure) ~ country + log(income))

Residuals:
      Min        1Q    Median        3Q       Max
-1.562643 -0.356784  0.003915  0.347889  2.147357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.78004    0.30434  15.706  < 2e-16 ***
countryUSA   0.74311    0.03661  20.297  < 2e-16 ***
log(income)  0.15193    0.02974   5.108 3.84e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5355 on 1097 degrees of freedom
Multiple R-squared: 0.3755,        Adjusted R-squared: 0.3744
F-statistic: 329.8 on 2 and 1097 DF,  p-value: < 2.2e-16
```

Dieses Modell spezifiziert also zwei parallele Regressionsgerade, eine für die USA, eine für die Niederlande. Die Geraden haben beide dieselbe Steigung (= Koeffizient von `log(income)`), nämlich 0.152. Nur die Achsenabschnitte unterscheiden sich: 4.78 für die Niederlande und $4.78 + 0.743 = 5.523$ für die USA. Die zugehörigen Regressionsgeraden können mit Hilfe von `abline`, dem man einen Achsenabschnitt und eine Steigung als Argumente übergibt, auch leicht visualisiert werden.

```
R> plot(log(expenditure) ~ log(income), col = as.numeric(country))
R> abline(coef(fmCI)[1], coef(fmCI)[3])
R> abline(sum(coef(fmCI)[1:2]), coef(fmCI)[3], col = 2)
```



Zum Abschluß sollen auch wieder die angelegten Objekte

```
R> detach(gsa)
R> objects()
```

```
[1] "GSA"   "fm1"   "fmC"   "fmCI"  "fmCxI" "fmCxY" "fmI"   "gsa"
```

aufgeräumt werden:

```
R> remove(GSA, gsa, fm1, fmC, fmI, fmCI, fmCxI, fmCxY)
```