Lineare Modelle in R: Einweg-Varianzanalyse

Achim Zeileis

2009-02-20

1 Datenaufbereitung

Wie schon in der Vorlesung wollen wir hier zur Illustration der Einweg-Analyse die logarithmierten Ausgaben der Touristen im `GSA` Datensatz erklären durch ihr Herkunftsland, wobei wir uns auf die Länder USA und Niederlande beschränken.

Um die Analyse der Vorlesung zu reproduzieren, müssen wir zunächst die zugehörigen Daten entsprechend aufbereiten. Dafür laden wir zunächst den `GSA` Datensatz:

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

Nun selektieren wir die Spalten/Variablen und die Zeilen/Beobachtungen, die wir benötigen und eliminieren dann noch die fehlenden Beobachtungen. Zuerst wählen wir die Spalten 3, 5, 6, 8 und 9 aus,

```
R> gsa <- GSA[, c(3, 5, 6, 8, 9)]
```

die zugehörigen Variablennamen sind:[1]

```
R> names(GSA)[c(3, 5, 6, 8, 9)]
```

```
[1] "income"       "country"      "expenditure"  "accomodation" "year"
```

Wir wollen uns dann nicht alle Beobachtungen anschauen, sondern nur die, die zu den USA oder den Niederlanden gehören, und rufen dafür `subset` auf:

```
R> gsa <- subset(gsa, country == "USA" | country == "Netherlands")
```

Und zuletzt lassen wir alle fehlenden Beobachtungen weg.

```
R> gsa <- na.omit(gsa)
```

Um einen kurzen Überblick über die Daten zu bekommen, erfragen wir die Dimension des Datensatzes sowie eine numerische Zusammenfassung.

```
R> dim(gsa)
```

```
[1] 1100    5
```

[1] Dieser Datensatz enthält mehr Variablen als wir für die folgende Analyse benötigen, aber da dieses Beispiel in der nächsten Einheit fortgeführt wird, werden hier schon alle auch dafür verwendeten Variablen ausgewählt.

```
R> summary(gsa)
```

```
     income                     country     expenditure         accomodation
 Min.   :   854.1   Netherlands     :648   Min.   : 113.0   Hotel     :553
 1st Qu.: 22100.0   USA             :452   1st Qu.: 485.9   Camping   :253
 Median : 31350.0   Austria (Vienna):  0   Median : 792.6   B&B       :103
 Mean   : 41567.5   Austria (other) :  0   Mean   : 991.8   Appartment:122
 3rd Qu.: 48320.0   Belgium         :  0   3rd Qu.:1284.9   Room      : 57
 Max.   :966400.0   Denmark         :  0   Max.   :9417.5   Farm      : 12
                    (Other)         :  0
   year
 1994:587
 1997:513
```

Dabei fällt auf, dass der Faktor `country` immer noch alle Levels hat, obwohl aufgrund unserer Auswahl nur noch zwei der Ausprägungen vorkommen. Um diesen Faktor neu zu kodieren, gibt es verschiedene Wege: Besonders einfach ist es, einmal die Funktion `factor` darauf anzuwenden und das Resultat wieder der entsprechenden Spalte im Datensatz zuzuweisen.

```
R> gsa$country <- factor(gsa$country)
R> summary(gsa)
```

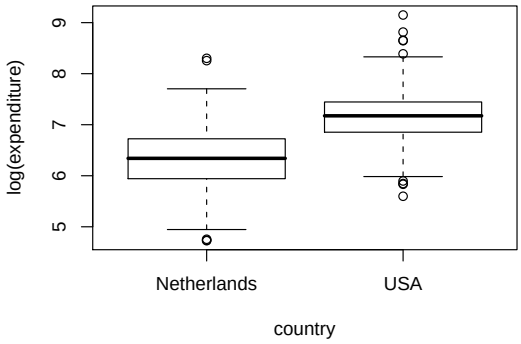
```
     income               country     expenditure         accomodation
 Min.   :   854.1   Netherlands:648   Min.   : 113.0   Hotel     :553
 1st Qu.: 22100.0   USA        :452   1st Qu.: 485.9   Camping   :253
 Median : 31350.0                     Median : 792.6   B&B       :103
 Mean   : 41567.5                     Mean   : 991.8   Appartment:122
 3rd Qu.: 48320.0                     3rd Qu.:1284.9   Room      : 57
 Max.   :966400.0                     Max.   :9417.5   Farm      : 12
   year
 1994:587
 1997:513
```

Nun hat `country` nur noch die gewünschten zwei Ausprägungen und wir können mit der eigentlichen Analyse beginnen.

2 Einweg-Varianzanalyse: 2-Stichprobenproblem

Der Faktor `country` hat zwei Ausprägungen, weshalb man hier auch von einem 2-Stichprobenproblem spricht. Zur Visualisierung erzeugen wir zunächst mal einen Boxplot der beiden Stichproben:

```
R> plot(log(expenditure) ~ country, data = gsa)
```



Daraus ist also ersichtlich, daß als die US-Amerikaner im Mittel mehr auszugeben scheinen als die Niederländer. Die Mittelwerte der beiden Gruppen sind

```
R> tapply(log(gsa$expenditure), gsa$country, mean)

Netherlands         USA
   6.330866    7.156888
```

Um durch eine Varianzanalyse nachzuweisen, daß diese Mittelwerte auch signifikant unterschiedlich sind, passen wir das entsprechende lineare Modell an, das `log(expenditure)` durch `country` erklärt.

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

Call:
lm(formula = log(expenditure) ~ country, data = gsa)

Coefficients:
(Intercept)   countryUSA
      6.331        0.826
```

In dem so angepaßten Modell wurden also durch R automatisch Treatment-Kontraste verwendet (dies ist die Voreinstellung in R). Der erste Koeffizient `(Intercept)` entspricht also dem Mittelwert in der Stichprobe der Niederländer. Der zweite Koeffizient `countryUSA` entspricht der Differenz der Mittelwerte zwischen den beiden Stichproben. Die log-Ausgaben der US-Amerikaner sind also im Mittel um 0.826 höher als die der Niederländer. Um den Mittelwert für die Stichprobe der US-Amerikaner zu berechnen, müßte man also beide Koeffizienten aufsummieren.

```
R> sum(coef(fmC))

[1] 7.156888
```

Um nun die Varianzanalyse durchzuführen gibt es in R verschiedene äquivalente Wege. In

```
R> summary(fmC)
```

```
Call:
lm(formula = log(expenditure) ~ country, data = gsa)

Residuals:
     Min       1Q   Median       3Q      Max
-1.60319 -0.37121  0.01674  0.33912  1.99344

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.33087    0.02128  297.56   <2e-16 ***
countryUSA   0.82602    0.03319   24.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5416 on 1098 degrees of freedom
Multiple R-squared: 0.3607,        Adjusted R-squared: 0.3601
F-statistic: 619.4 on 1 and 1098 DF,  p-value: < 2.2e-16
```

ist zunächst die t-Statistik für die Differenz der Mittelwerte `countryUSA` angegeben. Zu dieser gehört auch ein hochsignifikanter p-Wert. Denselben p-Wert erhalten wir auch aus der F-Statistik, die dieses Modell gegen das triviale Modell tests.

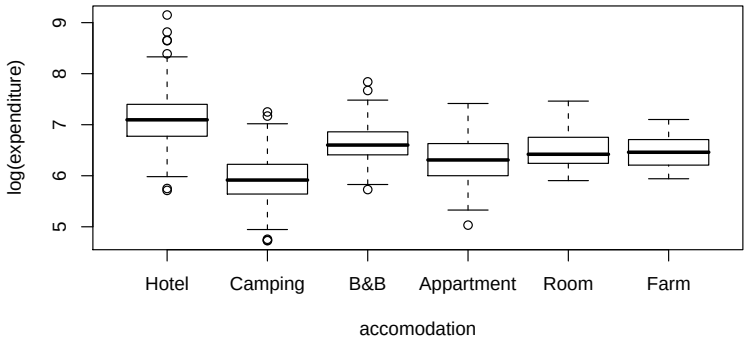
Man könnte natuerlich auch das triviale Modell explizit berechnen und dann die Varianzanalyse mit `anova` durchführen. Dies führt selbstverständlich auch zu demselben p-Wert:

```
R> fm1 <- lm(log(expenditure) ~ 1, data = gsa)
R> anova(fmC, fm1)

Analysis of Variance Table

Model 1: log(expenditure) ~ country
Model 2: log(expenditure) ~ 1
  Res.Df     RSS   Df Sum of Sq      F    Pr(>F)
1   1098  322.07
2   1099  503.75   -1   -181.68 619.38 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3 Einweg-Varianzanalyse: k-Stichprobenproblem

Nun können qualitative Erklärungsvariablen nicht nur 2 Ausprägungen haben, sondern im Allgemeinen k verschiedene Ausprägungen. Um zu illustrieren, welche Konsequenzen dies auf die entsprechenden linearen Modelle hat, wollen wir hier die log-Ausgaben `log(expenditure)` durch die Art der Unterbringung `accomodation` erklären. Die Frage ist also, ob Touristen, die im Hotel wohnen mehr Geld ausgeben als beispielsweise Touristen in einem Bed & Breakfast oder einem Appartment usw. Zur Visualisierung verwenden wir wieder parallele Boxplots

```
R> plot(log(expenditure) ~ accomodation, data = gsa)
```



aus denen schon ablesbar ist, dass der Hotelurlaub der teuerste zu sein scheint, während Camping besonders günstig ist und die anderen Unterbringungsmöglichkeiten ungefähr ähnlich sind. Die Mittelwerte der log-Ausgaben in den $k = 6$ Stichproben betragen hier

```
R> tapply(log(gsa$expenditure), gsa$accomodation, mean)
```

```
     Hotel    Camping        B&B Appartment       Room       Farm
  7.110292   5.937905   6.644212   6.296507   6.540692   6.473815
```

Das zugehörige lineare Modell passen wir dann mit `lm` an:

```
R> fmA <- lm(log(expenditure) ~ accomodation, data = gsa)
R> fmA

Call:
lm(formula = log(expenditure) ~ accomodation, data = gsa)

Coefficients:
           (Intercept)     accomodationCamping         accomodationB&B
                7.1103                -1.1724                -0.4661
accomodationAppartment        accomodationRoom        accomodationFarm
               -0.8138                -0.5696                -0.6365
```

Hier ist also zu sehen, daß der `(Intercept)` Koeffizient der Schätzung für die Hotel-Stichprobe entspricht. Dies ist also die Referenz-Kategorie. Alle übrigen Koeffizienten geben die Differenz der Mittelwerte zur Referenzkategorie an. Das heißt also, dass der Mittelwert bei Camping um 1.172 niedriger ist, bei Bed & Breakfast um 0.466 niedriger usw. Die entsprechenden Prognosen kann man also wieder durch Aufsummieren der entsprechenden Koeffizienten erhalten. Für Camping beispielsweise als

```
R> coef(fmA)[2] + coef(fmA)[1]

accomodationCamping
           5.937905
```

Um zu prüfen, ob `accomodation` nun einen signifikanten Einfluß auf `log(expenditure)` hat, verwenden wir wieder die zugehörigen `summary`.

```
R> summary(fmA)
```

```
Call:
lm(formula = log(expenditure) ~ accomodation, data = gsa)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39688 -0.30997 -0.02192  0.28815  2.04003

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             7.11029    0.02002 355.183  < 2e-16 ***
accomodationCamping    -1.17239    0.03573 -32.812  < 2e-16 ***
accomodationB&B        -0.46608    0.05052  -9.226  < 2e-16 ***
accomodationAppartment -0.81379    0.04709 -17.282  < 2e-16 ***
accomodationRoom       -0.56960    0.06549  -8.698  < 2e-16 ***
accomodationFarm       -0.63648    0.13736  -4.634 4.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4708 on 1094 degrees of freedom
Multiple R-squared: 0.5187,        Adjusted R-squared: 0.5165
F-statistic: 235.8 on 5 and 1094 DF,  p-value: < 2.2e-16
```

Hierbei ist nun zu beachten, daß die t-Statistiken und die F-Statistik unterschiedliche Interpretationen haben. Der zu `accomodationCamping` gehörende t-Test überprüft ausschließlich, ob sich der Mittelwert zwischen Hotel und Camping unterscheidet. Analog für die übrigen t-Statistiken. Um zu überprüfen, ob der Faktor insgesamt eine zusätzliche Erklärung gegenüber dem trivialen Modell bringt, kann hingegen die F-Statistik verwendet werden, die alle $k - 1 = 5$ Koeffizienten von `accomodation` gleichzeitig testet. Diese ist natürlich hochsignifikant, woraus geschloßen werden kann, daß dieser Faktor einen signifikanten Einfluß hat. Dieselbe F-Statistik mit demselben p-Wert erhält man alternativ auch wieder per `anova` und dem expliziten Vergleich mit dem trivialen Modell.

```
R> anova(fm1, fmA)

Analysis of Variance Table

Model 1: log(expenditure) ~ 1
Model 2: log(expenditure) ~ accomodation
  Res.Df    RSS   Df Sum of Sq      F    Pr(>F)
1   1099 503.75
2   1094 242.44    5    261.30 235.82 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wie eine Mehrweg-Varianzanalyse sowie Kovarianzanalyse mit Hilfe derselben Funktionen durchgeführt werden kann, wird der Inhalt des nächsten Tutoriums sein.

Zum Abschluß sollen auch wieder die angelegten Objekte

```
R> objects()

[1] "GSA" "fm1" "fmA" "fmC" "gsa"
```

aufgeräumt werden:

```
R> remove(GSA, gsa, fm1, fmA, fmC)
```