Uebung zur Einheit 06

Moritz Schrom

13. Dezember 2022

Abgabe

Lösen Sie die folgenden Aufgabe mit Hilfe der passenden R-Funktionen und laden Sie die Lösungen im Moodle-Upload als PDF-Dokument hoch. Namenskonvention: UE6_nachname.pdf

Hinweis: Schreiben Sie einen Bericht für Fachkolleg:innen!

Import

Im Datenfile Nachname.txt sind folgende Variablen enthalten:

- gewinn: Gewinnveränderung (in %)
- branche: Branche (1...Metallindustrie, 2...Textilindustrie, 3...Handel)
- groesse: Betriebsgröße (1...klein, 2...groß)

Aufgabe Konjunktur

In einem Land kommt es zu einem Konjunkturaufschwung. Es soll geprüft werden, ob drei bestimmte Branchen vom Aufschwung in gleicher Weise profitieren. Weiters soll untersucht werden, ob der Aufschwung kleine Betriebe anders betrifft als große Betriebe. Dazu wurden aus den drei Branchen jeweils X kleine und X große Betriebe zufällig ausgewählt und die im Datensatz gegebenen prozentualen Gewinnveränderungen erhoben.

Einlesen der Daten mit Hilfe von read.table

```
konjunktur <- read.table("Schrom.txt", header = TRUE, sep = ",")
```

Hinweise

Hinweis 1

Sind die Daten vollständig?

```
summary(konjunktur)
```

```
##      gewinn          branche     groesse
##  Min.   :-6.100   Min.   :1   Min.   :1.000
##  1st Qu.: 1.400   1st Qu.:1   1st Qu.:1.000
##  Median : 5.750   Median :2   Median :1.000
##  Mean   : 5.554   Mean   :2   Mean   :1.477
##  3rd Qu.: 8.575   3rd Qu.:3   3rd Qu.:2.000
##  Max.   :17.000   Max.   :3   Max.   :2.000
##                   NA's   :2   NA's   :4
```

Ein Summary der Daten zeigt. Es gibt 2 Nullwerte bei branche und 4 Nullwerte bei groesse. Diese entfernen wir:

```
konjunktur <- konjunktur[complete.cases(konjunktur), ]
summary(konjunktur)
```

```
##      gewinn          branche         groesse
##  Min.   :-6.100   Min.   :1.000   Min.   :1.000
##  1st Qu.: 1.525   1st Qu.:1.000   1st Qu.:1.000
##  Median : 5.750   Median :2.000   Median :1.000
##  Mean   : 5.455   Mean   :2.048   Mean   :1.476
##  3rd Qu.: 8.175   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :17.000   Max.   :3.000   Max.   :2.000
```

Hinweis 2

Sie müssen die Variablen branche und groesse zuerst als Faktoren definieren, damit R diese nicht aufgrund ihrer numerischen Kodierung als metrisch betrachtet und falsche Ergebnisse liefert. Geben Sie den Kategorien auch aussagekräftige Namen/Labels, z. B. so:

```
#daten$groesse <- factor(daten$groesse, labels = c("klein", "groß"))
konjunktur$branche <- factor(konjunktur$branche, labels = c("Metallindustrie", "Textilindustrie", "Hand
konjunktur$groesse <- factor(konjunktur$groesse, labels = c("klein", "groß"))
summary(konjunktur)
```

```
##      gewinn                  branche          groesse
##  Min.   :-6.100   Metallindustrie:13   klein      :22
##  1st Qu.: 1.525   Textilindustrie:14   gro<U+00DF>:20
##  Median : 5.750   Handel         :15
##  Mean   : 5.455
##  3rd Qu.: 8.175
##  Max.   :17.000
```

Hinweis 3

Die zweifachen Varianzanalysen sind in jedem Fall durchzuführen, auch wenn gewisse Voraussetzungen (Ausreißer, Varianzhomogenität, . . .) verletzt sind. Weisen Sie in diesem Fall einfach im Fließtext auf die Probleme hin.

Struktur der Analyse

Welche Branche scheint in der Stichprobe am meisten vom Konjunkturaufschwung zu profitieren?

- Hängen die Gewinnänderungen in der Stichprobe von der Betriebsgröße ab?
- Zu welchem Schluss kommt man, wenn man beide Prädiktoren als Haupteffekte ins Modell aufnimmt?
- Ist eine Wechselwirkung vorhanden?
- Für welches der möglichen fünf Modelle (konstantes Modell/einfache Varianzanalysen/zweifache Varianzanalyse ohne/mit Wechselwirkung) würden Sie sich schlussendlich entscheiden und warum

Hängen die Gewinnänderungen in der Stichrpobe von der Betriebsgröße ab?

Schauen wir uns zunächst die Homogenität der Varianzen unseres Modells mit dem Levene-Test an:

```
library(car)
```

```
## Loading required package: carData
```

```
model1 <- lm(gewinn ~ groesse, data = konjunktur)
leveneTest(model1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  4.8199  0.034 *
##       40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Das Ergebnis: 0.034, ist auf dem 0.05 Niveau signifikant. Was heißt das? Wir müssten die Nullhypothese "Gleiche Varianzen" eigentlich verwerfen, und könnten mit der ANOVA nicht fortfahren. Fahren wir trotzdem fort:

```
summary(model1)
```

```
##
## Call:
## lm(formula = gewinn ~ groesse, data = konjunktur)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3550  -3.9660   0.2136   2.5590  11.7450
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.6364     1.1593   4.862 1.84e-05 ***
## groessegro<U+00DF>  -0.3814     1.6800  -0.227    0.822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.438 on 40 degrees of freedom
## Multiple R-squared:  0.001287,   Adjusted R-squared:  -0.02368
## F-statistic: 0.05153 on 1 and 40 DF,  p-value: 0.8216
```

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: gewinn
##           Df  Sum Sq Mean Sq F value Pr(>F)
## groesse    1    1.52  1.5236  0.0515 0.8216
## Residuals 40 1182.68 29.5670
```

Was heißt das für uns im konkreten Fall? Wir können unsere Nullhypothese für den Faktor "groesse": "Alle Gewinnentwicklungen sind gleich" nicht verwerfen: Die Gewinnentwicklungen sind also gleich.

Auch das Summary des Modells zeigt keine Signifikanz der Unternehmensgröße (das Modell würde nur 0,1% der Varianz erklären, das ist extrem wenig...)

Zu welchem Schluss kommt man, wenn man beide Prädiktoren als Haupteffekte ins Modell aufnimmt?

Dafür erstellen wir ein neues Modell, mit beiden Prädikatoren: Außerdem schauen wir uns wieder mit dem Levene-Test die Homogenität der Varianzen an:

```
leveneTest(gewinn ~ groesse * branche, data = konjunktur)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
```

```
## group  5  1.1061 0.3743
##       36
```

Auch hier können wir die Nullhypothese der Gleichverteilung nicht verwerfen... Fahren wir trotzdem mit unserem Modell fort, und wenden die zweifache Anova an, sowie lassen uns eine Summary des Modells ausgeben.

```
model2 <- lm(gewinn ~ groesse + branche, data = konjunktur)
Anova(model2)
```

```
## Anova Table (Type II tests)
##
## Response: gewinn
##            Sum Sq Df F value Pr(>F)
## groesse      1.14  1  0.0388 0.8449
## branche     61.59  2  1.0437 0.3620
## Residuals 1121.09 38
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = gewinn ~ groesse + branche, data = konjunktur)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1704  -3.3609  -0.5323   2.7091  10.1364
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.0911     1.6941   3.005  0.00468 **
## groessegro<U+00DF>      -0.3307     1.6791  -0.197  0.84490
## brancheTextilindustrie  -0.6900     2.0931  -0.330  0.74346
## brancheHandel            2.1032     2.0582   1.022  0.31331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.432 on 38 degrees of freedom
## Multiple R-squared:  0.05329,    Adjusted R-squared:  -0.02145
## F-statistic: 0.713 on 3 and 38 DF,  p-value: 0.5503
```

Beide Faktoren sind nicht signifikant.

Ist eine Wechselwirkung vorhanden?

```
model3 <- lm(gewinn ~ groesse * branche, data = konjunktur)
Anova(model3)
```

```
## Anova Table (Type II tests)
##
## Response: gewinn
##                 Sum Sq Df F value    Pr(>F)
## groesse           1.14  1  0.0755    0.7851
## branche          61.59  2  2.0303    0.1461
## groesse:branche 575.10  2 18.9592 2.376e-06 ***
## Residuals       546.00 36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Und hier ist zu sehen: Der Interaktionsterm ist hochsignifikant (auf dem 0.001 Niveau). Es besteht also ein gemeinsamer Einfluss von Größe und Branche auf die Gewinnentwicklung.

Es gibt somit eine Wechselwirkung!

Lassen wir uns wieder ein Summary des Modells ausgeben, um den Erklärungswert zu ermitteln

```
summary(model3)
```

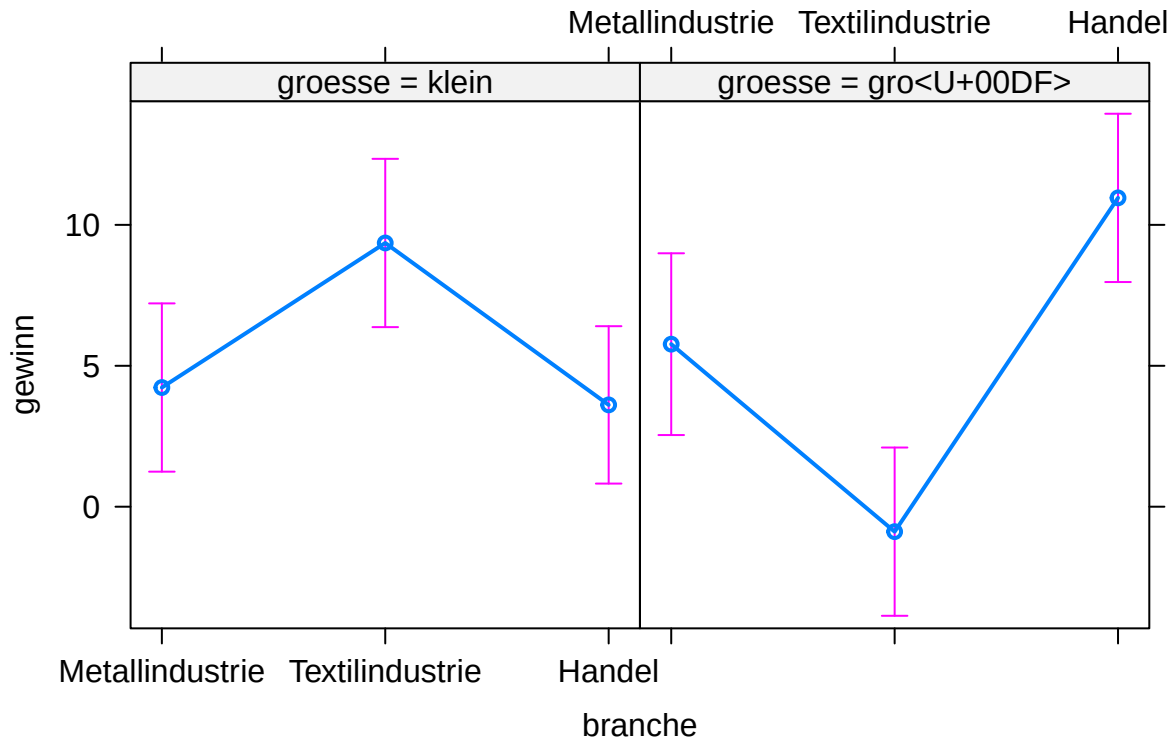
```
##
## Call:
## lm(formula = gewinn ~ groesse * branche, data = konjunktur)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7571 -3.2129  0.7643  2.3250  7.7333
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                4.2286     1.4720   2.873 0.006783
## groessegro<U+00DF>                         1.5381     2.1667   0.710 0.482345
## brancheTextilindustrie                     5.1286     2.0817   2.464 0.018663
## brancheHandel                             -0.6161     2.0156  -0.306 0.761626
## groessegro<U+00DF>:brancheTextilindustrie -11.7810     3.0046  -3.921 0.000379
## groessegro<U+00DF>:brancheHandel            5.8065     2.9592   1.962 0.057506
##
## (Intercept)                                ** 
## groessegro<U+00DF>
## brancheTextilindustrie                     *
## brancheHandel
## groessegro<U+00DF>:brancheTextilindustrie ***
## groessegro<U+00DF>:brancheHandel           .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.894 on 36 degrees of freedom
## Multiple R-squared:  0.5389, Adjusted R-squared:  0.4749
## F-statistic: 8.416 on 5 and 36 DF,  p-value: 2.382e-05
```

Unser Modell erklärt 53 Prozent der Varianz. Lassen wir uns diese Erkenntnis mit einem Effektplot darstellen: Hier wird verdeutlicht: Große Unternehmen in der Handelsbranche haben signifikant mehr vom Konjunkturaufschwung profitiert, als Unternehmen anderer Größe in anderen Branchen. Die Konfidenzintervalle überschneiden sich teilweise bis nicht. Großen Unternehmen aus der Textilindustrie haben jedoch etwas schlechter abgeschnitten. Bei den kleinen Unternehmen haben jedoch wiederum Textilunternehmen signifikant mehr profitiert als die anderen Branchen...

```
library(effects)
```

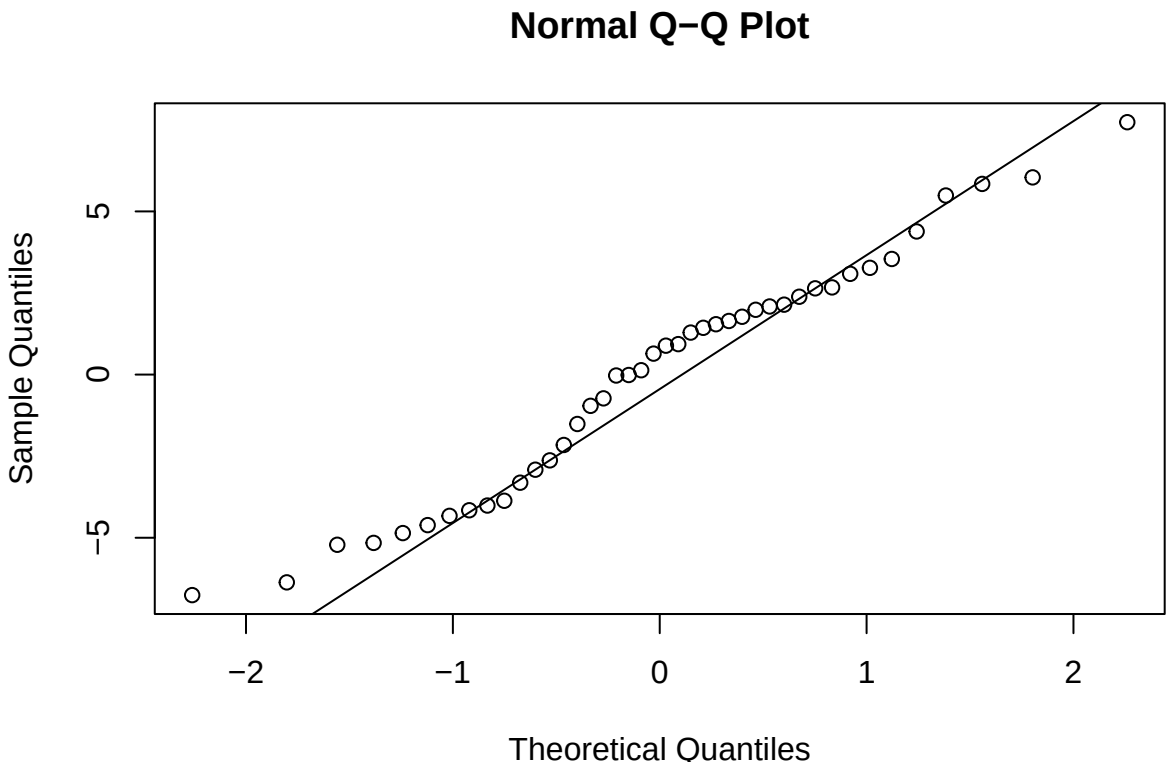
```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
plot(allEffects(model3))
```



Schauen wir uns noch die Normalverteilung der Residuen an:

```
qqnorm(residuals(model3))
qqline(residuals(model3))
```



Dieser zeigt Auffälligkeiten in Form von Ausreißern und Mustern... Daher sollten wir unser Modell nochmal hinterfragen.

Für welches der möglichen fünf Modelle (konstantes Modell/einfache Varianzanalysen/zweifache Vari- anzanalyse ohne/mit Wechselwirkung) würden Sie sich schlussendlich entscheiden und warum

Schlussendlich würde ich mich für die zweifache Varianzanalyse mit Wechselwirkung entscheiden, da diese beide Faktoren (Größe und Branche) miteinbezieht, sowie die Wechselwirkungen zwischen diesen. Es hat sich gezeigt, dass weder die Größe alleine, noch die Branche alleine signifikant für die Gewinnentwicklung war, sehr wohl jedoch die Kombination aus ebendiesen zwei Faktoren.