Applied Statistics for Data Science

Serie 13

Aufgabe 13.1

In der Bibliothek **ISLR** hat es den Datensatz **Carseats**. Wir möchten **Sales** (Anzahl Kinderautositze) aufgrund von verschiedenen Prädiktoren in 400 verschiedenen Standorten vorhersagen.

Der Datensatz enthält qualitative Prädiktoren, wie **ShelveLoc** als Indikator der Lage im Gestell, das heisst der Platz in einem Geschäft, wo der Autositz ausgestellt ist. Der Prädiktor nimmt die drei Werte **Bad**, **Medium** und **Good** an. Für qualitative Variablen generiert **R** Dummy-Variablen automatisch.

a) Untersuchen Sie den Datensatz mit **head(Carseat)** und **?Carseat**.

b) Finden Sie mit **lm()** ein multiples Regressionsmodell um **Sales** aus **Price**, **Urban** und **US** vorherzusagen.

c) Interpretieren Sie die Koeffizienten in diesem Modell. Achten Sie darauf, dass einige Variablen qualitativ sind.

d) Schreiben Sie das Modell in Gleichungsform. Achten Sie darauf, dass Sie die qualitativen Variablen richtig behandeln.

e) Für welche Prädiktoren kann die Nullhypothese $H_0 : \beta_j = 0$ verworfen werden?

f) Auf der Basis der vorhergehenden Frage, finden Sie ein kleineres Modell, das nur Prädiktoren verwendet für die es Hinweise auf einen Zusammenhang mit der Zielvariablen gibt.

g) Wie genau passen die Modelle in a) und e) die Daten an?

Aufgabe 13.2

Wir führen noch eine multiple lineare Regression für **Auto** aus der letzten Übung durch.

a) Bestimmen Sie mit `regsubset` mit Option `forward`, das multiple Regressionsmodell mit zwei Prädiktoren und bestimmen Sie die Koeffizienten diese Modelles. Interpretieren Sie die Koeffizienten.

b) Hätte es einen Unterschied gegeben, wenn wir die Option `backward` gewählt hätten?

Applied Statistics for Data Science

Musterlösungen zu Serie 13

Lösung 13.1

a) Datensatz:

```
library(ISLR)
head(Carseats)

##   Sales CompPrice Income Advertising Population Price ShelveLoc
## 1  9.50       138     73          11        276   120       Bad
## 2 11.22       111     48          16        260    83      Good
## 3 10.06       113     35          10        269    80    Medium
## 4  7.40       117    100           4        466    97    Medium
## 5  4.15       141     64           3        340   128       Bad
## 6 10.81       124    113          13        501    72       Bad
##   Age Education Urban  US
## 1  42        17   Yes Yes
## 2  65        10   Yes Yes
## 3  59        12   Yes Yes
## 4  55        14   Yes Yes
## 5  38        13   Yes  No
## 6  78        16    No Yes
```

b) Output:

```
fit <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit)

##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393,Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

c) Interpretation der Koeffizienten:

- Der Koeffizient 13.04 ist ein bisschen schwierig zu interpretieren. Gemäss dem Modell unter d) sind dies die mittleren Verkaufszahlen in Geschäften, die in ländlichen Gegenden ausserhalb der USA erreicht werden, wobei der Preis der Kindersitze noch \$0 ist (nicht sehr realsistisch).
- Der Koeffizient -0.05 besagt, dass für eine Zunahme von einem Dollar durchschnittlich 0.05 Einheiten Kindersitze weniger verkauft werden.
- Der Koeffizient -0.021 besagt, dass verglichen zu ländlichen Gegenden durchschnittlich 0.021 Einheiten weniger verkauft werden. Der p-Wert ist allerdings sehr hoch, so dass dies eher eine zufällige Abweichung ist.
- Der Koeffizient 1.2 besagt, dass verglichen zu Geschäften ausserhalb der USA, 1.2 Einheiten mehr verkauft werden. Vielleicht sind in den USA Kindersitze pflicht.

d) Modell: Für **`Urban`** wählen wir die Dummy-Variable:

$$x_{2i} = \begin{cases} 1 & \text{falls } i\text{-te Person lebt in der Stadt} \\ 0 & \text{falls } i\text{-te Person lebt auf dem Land} \end{cases}$$

Für **`US`** wählen wir die Dummy-Variable

$$x_{3i} = \begin{cases} 1 & \text{falls } i\text{-te Person lebt in den USA} \\ 0 & \text{falls } i\text{-te Person lebt nicht in den USA} \end{cases}$$

Das Modell lautet dann

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot \texttt{Price} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \\ &= \beta_0 + \beta_1 \cdot \texttt{Price} + \begin{cases} \beta_2 + \beta_3 + \varepsilon_i & \text{falls } i\text{-te Person urban in den USA lebt} \\ \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person urban nicht in den USA lebt} \\ \beta_3 + \varepsilon_i & \text{falls } i\text{-te Person ländlich in den USA lebt} \\ \varepsilon_i & \text{falls } i\text{-te Person ländlich nicht in den USA lebt} \end{cases} \end{aligned}$$

e) Für alle ausser **`Urban`**

f) Output:

```
fit <- lm(Sales ~ Price + US, data = Carseats)
summary(fit)

##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393,Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

Modell: Für **US** wählen wir die Dummy-Variable

$$x_{2i} = \begin{cases} 1 & \text{falls } i\text{-te Person lebt in den USA} \\ 0 & \text{falls } i\text{-te Person lebt nicht in den USA} \end{cases}$$

Das Modell lautet dann

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot \texttt{Price} + \beta_2 x_{2i} + \varepsilon_i \\ &= \beta_0 + \beta_1 \cdot \texttt{Price} + \begin{cases} \beta_2 + \varepsilon_i & \text{falls } i\text{-te Person in den USA lebt} \\ \varepsilon_i & \text{falls } i\text{-te Person nicht in den USA lebt} \end{cases} \\ &= 13.03 - 0.055 \cdot \texttt{Price} + \begin{cases} 1.2 + \varepsilon_i & \text{falls } i\text{-te Person in den USA lebt} \\ \varepsilon_i & \text{falls } i\text{-te Person nicht in den USA lebt} \end{cases} \end{aligned}$$

g) Bei beiden Modellen ist zwar der Zusammenhang belegt (p-Wert für F-Wert praktisch 0), aber wenn wir die R^2-Werte betrachten, so ist der mit 0.2393 relativ schlecht. Das heisst, obwohl der Zusammenhang gesichtert ist die Passung schlecht, da nur 23 % der Variabilität der **Sales** durch das Modell erklärt werden kann.

Lösung 13.2

a) Output:

```
library(ISLR)
Auto.1 <- within(Auto, rm(name))
library(leaps)
reg <- regsubsets(mpg ~ ., data = Auto.1, method = "forward", nvmax = 6)
summary(reg)$which

##   (Intercept) cylinders displacement horsepower weight
## 1        TRUE     FALSE        FALSE      FALSE   TRUE
## 2        TRUE     FALSE        FALSE      FALSE   TRUE
## 3        TRUE     FALSE        FALSE      FALSE   TRUE
## 4        TRUE     FALSE         TRUE      FALSE   TRUE
## 5        TRUE     FALSE         TRUE       TRUE   TRUE
## 6        TRUE      TRUE         TRUE       TRUE   TRUE
##   acceleration  year origin
## 1        FALSE FALSE  FALSE
## 2        FALSE  TRUE  FALSE
## 3        FALSE  TRUE   TRUE
## 4        FALSE  TRUE   TRUE
## 5        FALSE  TRUE   TRUE
## 6        FALSE  TRUE   TRUE
```

$$\texttt{mpg} = \beta_0 + \beta_1 \cdot \texttt{weight} + \beta_2 \cdot \texttt{year}$$

```
fit <- lm(mpg ~ weight + year, data = Auto)
summary(fit)$coef

##                 Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept) -14.347253018 4.0065185631  -3.580978  3.856624e-04
## weight       -0.006632075 0.0002145559 -30.910708 8.361624e-107
## year          0.757318281 0.0494726873  15.307806  9.772260e-42
```

$$\texttt{mpg} = -14.35 - 0.007 \cdot \texttt{weight} + 0.757 \cdot \texttt{year}$$

Bei einem Gewicht von 0 lbs (1 lbs $\approx$ 0.45 kg) und im Jahr 1900 würde diese Auto -14.35 Meilen pro Gallone machen. Das macht natürlich alles keinen Sinn: Erstens gibt es kein Auto mit 0 lbs , noch gab es 1900 sehr viele Autos.

Nimmt das Gewicht das Autos um 1 lbs zu (bei konstantem Alter), so kann das Auto 0.007 Meilen weniger pro Gallone fahren.

Nimmt das Alter des Autos um 1 lbs ab (bei konstantem Gewicht), so kann das Auto 0.757 Meilen mehr pro Gallone fahren.

```
fit <- lm(mpg ~ weight * year, data = Auto)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ weight * year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0397 -1.9956 -0.0983  1.6525 12.9896
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.105e+02  1.295e+01  -8.531 3.30e-16 ***
## weight       2.755e-02  4.413e-03   6.242 1.14e-09 ***
## year         2.040e+00  1.718e-01  11.876  < 2e-16 ***
## weight:year -4.579e-04  5.907e-05  -7.752 8.02e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.193 on 388 degrees of freedom
## Multiple R-squared:  0.8339,Adjusted R-squared:  0.8326
## F-statistic: 649.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

b) Output:

```
library(leaps)
reg <- regsubsets(mpg ~ ., data = Auto.1, method = "backward", nvmax = 6)
summary(reg)$which

##   (Intercept) cylinders displacement horsepower weight
## 1        TRUE     FALSE        FALSE      FALSE   TRUE
## 2        TRUE     FALSE        FALSE      FALSE   TRUE
## 3        TRUE     FALSE        FALSE      FALSE   TRUE
## 4        TRUE     FALSE         TRUE      FALSE   TRUE
## 5        TRUE     FALSE         TRUE       TRUE   TRUE
## 6        TRUE      TRUE         TRUE       TRUE   TRUE
##   acceleration  year origin
## 1        FALSE FALSE  FALSE
## 2        FALSE  TRUE  FALSE
## 3        FALSE  TRUE   TRUE
## 4        FALSE  TRUE   TRUE
## 5        FALSE  TRUE   TRUE
## 6        FALSE  TRUE   TRUE
```

Es kommt dasselbe Modell heraus.