Applied Statistics for Data Science

Serie 12

Aufgabe 12.1

Wir untersuchen den Datensatz `Boston` aus dem letzten Übungsblatt weiter.

Um ein multiples lineares Regressionsmodell unter Verwendung der kleinsten Quadrate anzupassen, verwenden wir wieder die Funktion `lm()`. Die Syntax `lm(y ~ x1 + x2 + x3)` wird verwendet, um eine Modell mit drei Prädiktoren, `x1` , `x2` und `x3` . Die Funktion `summary()` jetzt gibt die Regressionskoeffizienten für alle Prädiktoren aus.

a) Passen Sie ein multiples lineares Regressionsmodell mit der Zielvariable `medv` und den Prädiktoren `lstat` und `age` an.

   Definieren Sie das Modell und interpretieren Sie alle Werte in der Ausgabe `summary()`, die wir besprochen haben (Koeffizienten, seine p-Werte, R^2-Wert, p-Wert der F-Statistik).

b) Der `Boston`-Datensatz enthält 13 Variablen, und es wäre also umständlich all dies eingeben zu müssen, um eine Regression mit allen Prädiktoren. Stattdessen können wir die folgende Kurzhand `lm(medv ~., Daten = Boston)` verwenden.

   Interpretieren Sie in der `summary()` Ausgabe den Koeffizienten von `age` und den entsprechenden p-Wert, vergleichen Sie diesen mit der Ausgabe in a) und erklären Sie den Unterschied.

c) Der Wert von R^2 ist größer als der in a) berechnete Wert. Erläutern Sie.

d) Mit Hilfe der Funktion `lm()` ist es einfach, Interaktionsterme in ein lineares Modell aufzunehmen. Die Syntax `lstat:black` weist `R` an, einen Interaktionsterm zwischen `lstat` und `black`.

   Die Syntax `lstat * age` beinhaltet gleichzeitig `lstat`, `age`, und der Interaktions-Begriff `lstat × age` als Prädiktoren; es ist eine Abkürzung für `lstat + age + lstat:age`.

Diskutieren Sie nochmals alle Werte in der `summary()` von `lstat*age` wie in a).

Aufgabe 12.2

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

a) Produzieren Sie mit `pairs` Streudiagramme, die alle Variablen des Datensatzes enthält.

b) Berechnen Sie die Korrelationsmatrix zwischen den Variablen mit `cor()`. Dazu müssen wir zuerst die Variable `name` entfernen, da diese qualitativ ist und vor allem kaum einen Einfluss auf den Verbrauch hat.

```
library(ISLR)

head(Auto)

##   mpg cylinders displacement horsepower weight acceleration year
## 1  18         8          307        130   3504         12.0   70
## 2  15         8          350        165   3693         11.5   70
## 3  18         8          318        150   3436         11.0   70
## 4  16         8          304        150   3433         12.0   70
## 5  17         8          302        140   3449         10.5   70
## 6  15         8          429        198   4341         10.0   70
##   origin                      name
## 1      1 chevrolet chevelle malibu
## 2      1         buick skylark 320
## 3      1        plymouth satellite
## 4      1             amc rebel sst
## 5      1               ford torino
## 6      1          ford galaxie 500

Auto.1 <- within(Auto, rm(name))

head(Auto.1)

##   mpg cylinders displacement horsepower weight acceleration year
## 1  18         8          307        130   3504         12.0   70
## 2  15         8          350        165   3693         11.5   70
## 3  18         8          318        150   3436         11.0   70
## 4  16         8          304        150   3433         12.0   70
## 5  17         8          302        140   3449         10.5   70
## 6  15         8          429        198   4341         10.0   70
##   origin
## 1      1
## 2      1
## 3      1
```

```
## 4        1
## 5        1
## 6        1
```

Interpretieren Sie die Werte für `horsepower` und `displacement` mit den Streudiagrammen oben.

c) Wir verwenden `lm()` um eine multiple Regression mit der Zielgrösse `mpg` und allen anderen Variablen (ausser `name`) als Prädiktoren durchzuführen. Verwenden Sie wieder Output des `summary()`-Befehls zu interpretieren.

   i) Gibt es einen Zusammenhang zwischen den Prädiktoren und der Zielvariable? Begründen Sie dies mit dem p-Wert zum F-Wert.

   ii) Welche Prädiktoren scheinen statistisch signifikant einen Einfluss auf die Zielvariable zu haben?

   iii) Was deutet der Koeffizient für `year` an?

d) Untersuchen das Modell aus c) noch auf Interaktionseffekte.

Applied Statistics for Data Science

Musterlösungen zu Serie 12

Lösung 12.1

a) Modell:

$$\texttt{medv} = \beta_0 + \beta_1 \cdot \texttt{lstat} + \beta_2 \cdot \texttt{age}$$

```
library(MASS)
fit <- lm(medv ~ lstat + age, data = Boston)

summary(fit)

##
## Call:
## lm(formula = medv ~ lstat + age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.981  -3.978  -1.283   1.968  23.158
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
## lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
## age          0.03454    0.01223   2.826  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.173 on 503 degrees of freedom
## Multiple R-squared:  0.5513,Adjusted R-squared:  0.5495
## F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

Die Schätzungen sind

$$\widehat{\beta}_0 = 33.22; \qquad \widehat{\beta}_1 = -1.03; \qquad \widehat{\beta}_2 = 0 - 03$$

Wir bekommen für das Modell

$$\texttt{medv} = 33.22 - 1.03 \cdot \texttt{lstat} + 0.03 \cdot \texttt{age}$$

Interpretation der Schätzungen:

a) $\widehat{\beta}_0 = 33.22$

In Vierteln, in denen es keine Bevölkerung mit niedrigerem Status und keine vor 1940 gebauten Einheiten gibt, liegt der mittlere Wert der Häuser bei \$ 33 220.

b) $\widehat{\beta}_1 = -1.03$

Für jedes zusätzliche Prozent der Bevölkerung mit niedrigerem Status sinkt der mittlere Wert um \$ 1030.

c) $\widehat{\beta}_2 = 0.03$

Für jedes zusätzliche Prozent der Einheiten, die vor 1949 gebaut wurden, erhöht sich der mittlere Wert um \$ 30.

d) Alle p-Werte sind signifikant (unterhalb des Signifikanzniveaus von 5 %), so dass alle Schätzungen einzeln signifikant zum Modell beitragen.

e) Der R^2-Wert beträgt 0.5513, daher werden etwa 55 % der Variation durch das Modell erklärt.

f) Der p-Wert des F-Wertes liegt unterhalb des Signifikanzniveaus und ist daher signifikant. Die Nullhypothese H_0

$$\beta_1 = \beta_2 = 0$$

wird abgelehnt. Einer der β's unterscheidet sich signifikant von 0. Mindestens eine Variable trägt signifikant zum Modell bei.

b)

```
fit <- lm(medv ~ ., data = Boston)

summary(fit)

##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406,Adjusted R-squared:  0.7338
```

```
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Der Wert von p ist fast 1, also überhaupt nicht signifikant. Aber in a) beträgt der p-Wert $0,005$, was signifikant ist. Das bedeutet, dass die Variable **age** stark mit anderen Variablen korrelieren muss (siehe d)).

c) Je mehr Variablen Sie haben, desto größer ist der R^2-Wert. Das bedeutet, dass die R^2 kein guter Indikator ist, um verschiedene Modelle zu vergleichen.

d) Modell:

$$\texttt{medv} = \beta_0 + \beta_1 \cdot \texttt{lstat} + \beta_2 \cdot \texttt{age} + \beta_{12} \cdot \texttt{lstat*age}$$

Anmerkung: $*$ im **lstat*age** bedeutet *nicht* Multiplikation, sondern nur Interaktion.

```
fit <- lm(medv ~ lstat * age, data = Boston)

summary(fit)

##
## Call:
## lm(formula = medv ~ lstat * age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.806  -4.045  -1.333   2.085  27.552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
## lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
## age         -0.0007209  0.0198792  -0.036   0.9711
## lstat:age    0.0041560  0.0018518   2.244   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.149 on 502 degrees of freedom
## Multiple R-squared:  0.5557,Adjusted R-squared:  0.5531
## F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

Die Schätzungen sind

$$\widehat{\beta}_0 = 36,10; \qquad \widehat{\beta}_1 = -1.39; \qquad \widehat{\beta}_2 = -0.0007; \qquad \widehat{\beta}_{12} = 0.004$$

Wir bekommen für das Modell

$$\texttt{medv} = 36.10 - 1.39 \cdot \texttt{lstat} - 0.00072 \cdot \texttt{age} + 0.0041 \cdot \texttt{lstat*age}$$

Interpretation der Schätzungen:

a) $\widehat{\beta}_0 = 36.10$

In Vierteln, in denen es keine Bevölkerung mit niedrigerem Status und keine vor 1940 gebauten Einheiten gibt, liegt der mittlere Wert der Häuser bei \$ 36 100.

b) $\widehat{\beta}_1 = -1.39$

Für jedes zusätzliche Prozent der Bevölkerung mit niedrigerem Status sinkt der mittlere Wert um \$ 1930.

c) $\widehat{\beta}_2 = -0.00072$

Für jedes zusätzliche Prozent der Einheiten, die vor 1949 gebaut wurden, sinkt der mittlere Wert um \$ 0.27.

Wie Sie sich vorstellen können, ist dieser Wert nicht signifikant, wie Sie aus der Ausgabe ersehen können.

d) $\widehat{\beta}_{12} = 0.004$

Dieser Koeffizient ist etwas schwierig zu interpretieren, und wir haben es im Unterricht nicht gemacht.

e) Nicht mehr alle p-Werte sind signifikant (unterhalb des Signifikanzniveaus von 5 %).

Der p-Wert für **age** ist $0,97$, also nicht mehr signifikant, während er es ohne Interaktion war. Was ist der Grund dafür?

Der p-Wert des Interaktionstermins beträgt $0,0252$ und liegt damit unter dem Signifikanzniveau von 5 %. Die Nullhypothese H_0, dass es keine Interaktion gibt, wird zurückgewiesen. Es besteht eine statistisch signifikante Interaktion.

Werfen wir nun einen Blick auf den Korrelationskoeffizienten der beiden erklärenden Variablen **lstat** und **age**.

```
cor(Boston["lstat"], Boston["age"])

##              age
## lstat 0.6023385
```

Dieser Wert ist recht hoch. Eine Erklärung könnte sein, dass die Menschen in den ärmeren Vierteln nicht das Geld hatten, um neue Häuser zu bauen, so dass es mehr Häuser gibt, die vor 1940 gebaut wurden.

f) Der Wert von R^2 beträgt $0,56$, daher wird etwa 56 % der Variation durch das Modell erklärt.

g) Der p-Wert des F-Wertes liegt unterhalb des Signifikanzniveaus und ist daher signifikant. Die Nullhypothese H_0

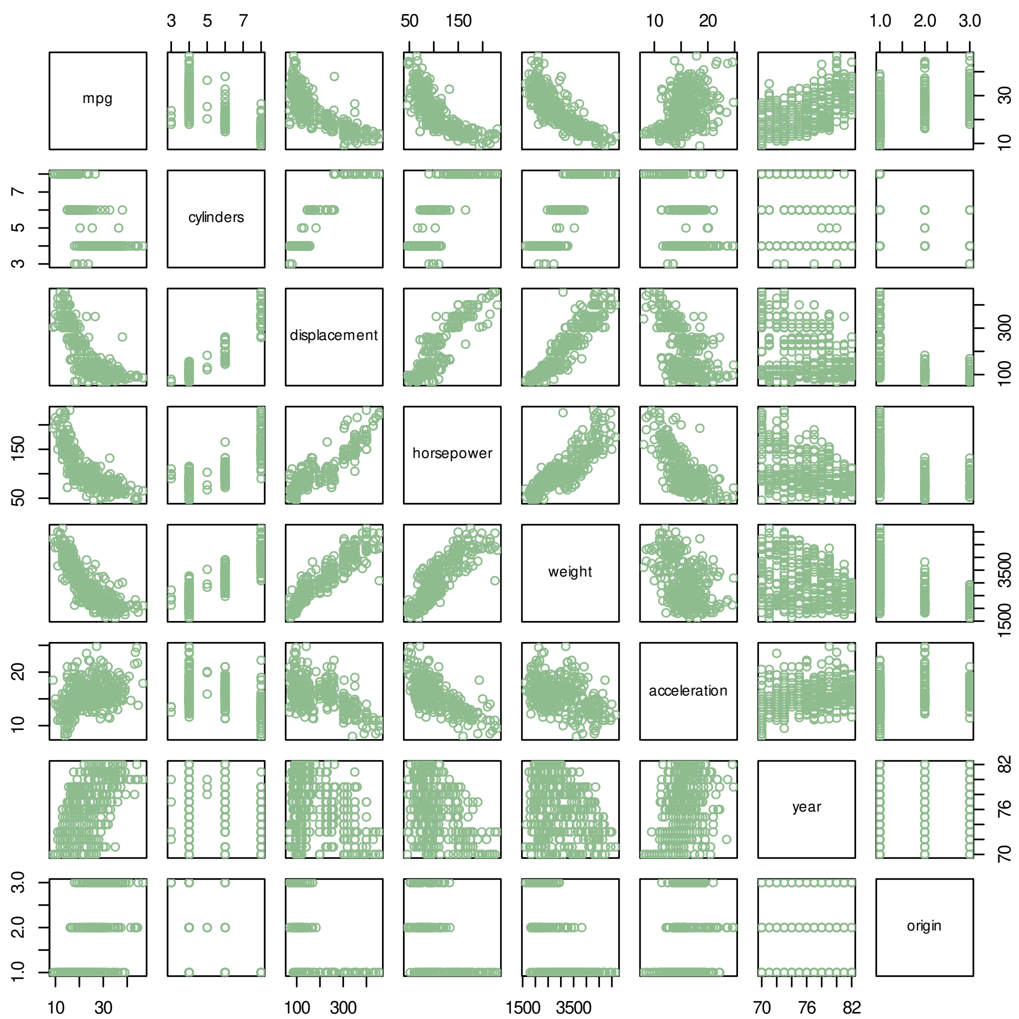
$$\beta_1 = \beta_2 = \{\beta_{12} = 0$$

wird abgelehnt. Einer der β's unterscheidet sich signifikant von 0. Mindestens eine Variable trägt signifikant zum Modell bei.

Lösung 12.2

a) Streudiagramm:

```
pairs(Auto.1, col = "darkseagreen")
```



b) Korrelationsmatrix

```
round(cor(Auto.1), 2)
```

```
##                 mpg cylinders displacement horsepower weight
## mpg            1.00     -0.78        -0.81      -0.78  -0.83
## cylinders     -0.78      1.00         0.95       0.84   0.90
## displacement -0.81      0.95         1.00       0.90   0.93
## horsepower    -0.78      0.84         0.90       1.00   0.86
## weight        -0.83      0.90         0.93       0.86   1.00
## acceleration  0.42     -0.50        -0.54      -0.69  -0.42
## year           0.58     -0.35        -0.37      -0.42  -0.31
## origin         0.57     -0.57        -0.61      -0.46  -0.59
##              acceleration  year origin
## mpg                  0.42  0.58   0.57
## cylinders           -0.50 -0.35  -0.57
## displacement        -0.54 -0.37  -0.61
## horsepower          -0.69 -0.42  -0.46
## weight              -0.42 -0.31  -0.59
## acceleration         1.00  0.29   0.21
## year                 0.29  1.00   0.18
## origin               0.21  0.18   1.00
```

Der Korrelationskoeffizient ist 0.9. Das heisst, je grösser **horsepower** ist, umso grösser ist **displacement**. Die beiden Variablen korrelieren also. Dies ist auch aus a) ersichtlich. Das Streudiagramm zeigt deutlich einen positiven linearen Zusammenhang.

c) Output

```
fit <- lm(mpg ~ ., data = Auto.1)
summary(fit)

##
## Call:
## lm(formula = mpg ~ ., data = Auto.1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215,Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i) Der p-Wert zum zugehörigen F-Wert ist praktisch 0 und somit besteht ein statistisch signifikanter Zusammenhang zwischen Zielvariable und den Prädiktoren.

ii) Dies sind die Koeffizienten mit $**$ oder $***$ (`displacement`, `weight`, `year` und `origin`).

iii) Der Koeffizient für `year` ist positiv. Das heisst, man mit jüngeren Autos weiter pro Gallone Benzin kommt. Die neueren Autos sind als im Allgemeinen sparsamer.

d) Output:

```
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
```

Der p-Wert des Interaktionsterm ist von der Grössenordnung 10^{-14}, also sehr nahe bei 0. Die Nullhypothese, dass keine Interaktion vorliegt, wird also verworfen.

Dies lässt sich damit erklären, dass das Gewicht mit den jüngeren Autos immer kleiner geworden ist.