Adrian Willi

Applied Statistics for Data Science

Serie 11

Aufgabe 11.1

In dieser Aufgabe verwenden wir den Datensatz **Auto**, der in der Bibliothek **ISLR** enthalten ist.

```
library(ISLR)
```

Falls eine Fehlermeldung kommt, muss die Bibliothek zuerst installiert werden (dies muss nur ein einziges Mal gemacht werden:

```
install.packages("ISLR")
```

a) Untersuchen Sie den Datensatz mit `head(Auto)` und `?Auto`.

b) Stellen Sie das Modell für eine einfache lineare Regression mit `mpg` als Zielvariable und `horsepower` als Prädiktor auf. $mpg = \beta_0 + \beta_1 \cdot horsepower$

c) Verwenden Sie den `lm()`-Befehl um diese Regression durchzuführen.
lm(Auto$mpg ~ Auto$horsepower)
Verwenden Sie den `summary()`-Befehl um die Resultate auszudrucken. Kommentieren Sie diesen:

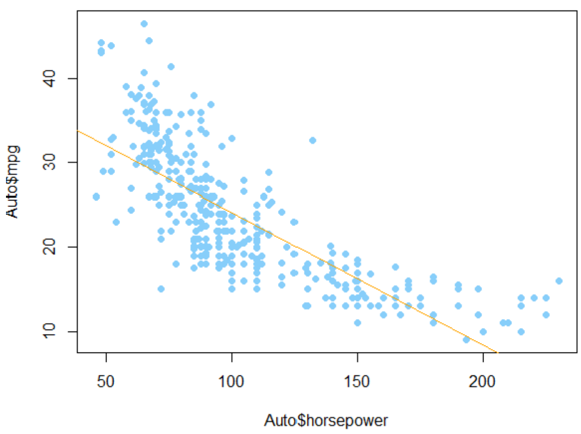
   i) Gibt es einen Zusammenhang zwischen der Zielgrösse und dem Prädiktor?
   p-Wert ist fast 0. H_0 wird verworfen ($H_0: \beta_1 = 0$)

   ii) Wie interpretieren Sie die Koeffizienten für `(intercept)` und `horsepower`? Ist der Zusammenhang positiv oder negativ? 39.93 Benzinverbrauch für 0 PS
   -0.15 pro PS sinkt Reichweite

   iii) Bestimmen Sie die Vertrauensintervalle (mit `confint()`) und interpretieren Sie diese? Die wahren Werte liegen im Vertrauensintervall

   iv) Interpretieren Sie den R^2-Wert. Gemäss R^2 Wert sind 60% der Varianz durch das Modell erklärbar. Mässig gut

d) Plotten Sie die Zielvariable und den Prädiktor mit der Regressionsgeraden (`abline`). Wie interpretieren Sie diesen Plot im Vergleich zum `summary()`-Output.



Passt gut zu den Beobachtungen aus dem summary()-Output

Aufgabe 11.2

Die **MASS**-Bibliothek enthält den **Boston**-Datensatz, der **medv** (median house value) für 506 Stadtviertel um Boston herum erfasst. Wir werden versuchen, **medv** mit 13 Prädiktoren wie **rm** (durchschnittliche Anzahl von Zimmern pro Haus) , **age** (Durchschnittsalter der Häuser) und **lstat** (Prozent der Haushalte mit niedrigem sozioökonomischen Status) vorherzusagen .

a) Um mehr über den Datensatz zu erfahren, können wir **?Boston** oder **help(Boston)** eingeben. Laden Sie zuerst die **MASS**-Bibliothek.

b) Welche Spaltennamen sind verfügbar? colnames(Boston)

c) Mit dem **attach(...)**-Befehl können wir **R** die Spaltennamen des Datensatzes **Boston** erkennen lassen. attach(Boston)

d) Wir werden damit beginnen, die **lm()**-Funktion zu verwenden, um ein einfaches lineares Regressionsmodell mit **medv** als Antwort und **lstat** als Prädiktor anzupassen.

   i) Definieren Sie das einfache Regressionsmodell unter Verwendung der beiden obigen Variablen. $medv = \beta_0 + \beta_1 \cdot lstat$

   ii) Die grundlegende Syntax lautet **lm(y~x,data)** Dabei ist **y** die Antwort, **x** der Prädiktor und Daten der Datensatz, in dem diese beiden Variablen enthalten sind.

```
lm.fit <- lm(...)
summary(lm.fit)
```

e) Wir können die Funktion **names(...)** verwenden, um herauszufinden, welche anderen Informationen in **lm.fit** gespeichert sind.

f) Obwohl wir diese Grössen über den Namen zugreifen können (beispielsweise **lm.fit$coefficients**), ist es sicherer, Funktionen wie **coef(...)** zu verwenden, um auf sie zuzugreifen. coef(lm.fit)

   Interpretieren Sie diese Werte und die entsprechenden p–Werte in der obigen Zusammenfassung. $medv = 34.55 - 0.95 \cdot lstat$ Insider

   Preis ohne Untermenschen Abnahme Preis für jedes Prozent Untermensch

g) Um ein Vertrauensintervall für die Koeffizientenschätzungen zu erhalten, können wir den Befehl **confint(...)** verwenden.

   Geben Sie eine Interpretation dieser Werte an.

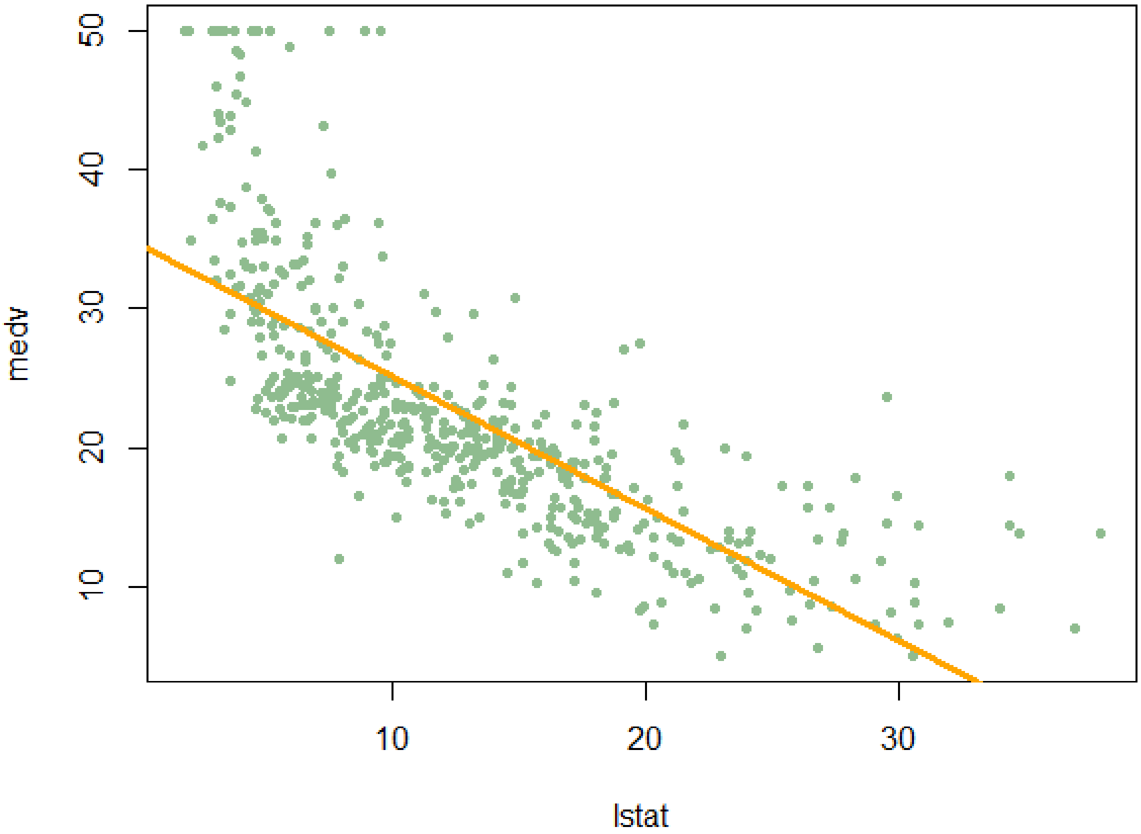
   Beide wahren Werte passen zum Vertrauensintervall

h) Wir werden nun **medv** und **lstat** zusammen mit der Regression der kleinsten Quadrate auftragen Zeile mit den Funktionen **plot(...)** und **abline()** (siehe Übung 3).

Verwenden Sie **lty = ...**, **pch = ...** und **col = ...**, um Graphik schöner aussehen zu lassen.

i) Interpretieren Sie den Wert R^2 in der **summary**–Ausgabe oben.

Es werden etwa 54% der Varianz durch das Regressionsmodell erklärt



Applied Statistics for Data Science

Musterlösungen zu Serie 11

Lösung 11.1

a) Tabelle:

```
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
```

b) Lineare Regression:

$$\texttt{mpg} = \beta_0 + \beta_1 \cdot \texttt{horsepower}$$

c) Output:

```
fit <- lm(mpg ~ horsepower, data = Auto)
# Oder: fit <- lm(Auto$mpg ~ Auto$horsepower)

summary(fit)

##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059,Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i) Der p-Wert für `horsepower` ist fast 0 und somit wird die Nullhypothese ($\beta_1 = 0$) verworfen. Der Treibstoffverbrauch hängt von den PS ab.

ii) Der Wert 39.93 für den `intercept` gibt den Benzinverbrauch (miles pro gallon) bei 0 PS an. Dieser Wert hat hier natürlich keine praktische Bedeutung.

Interessanter ist der Wert -0.15 für `horsepower`. Dieser bedeutet, dass pro PS dass Auto 0.15 Meilen weniger weit kommt für eine Gallone ($\approx$ 3.8 l) Benzin.

Der Zusammenhang ist also negativ: je mehr PS umso weniger weit kommt pro Gallone.

iii) Vertrauensintervall:

```
confint(fit)

##                  2.5 %      97.5 %
## (Intercept) 38.525212 41.3465103
## horsepower  -0.170517 -0.1451725
```

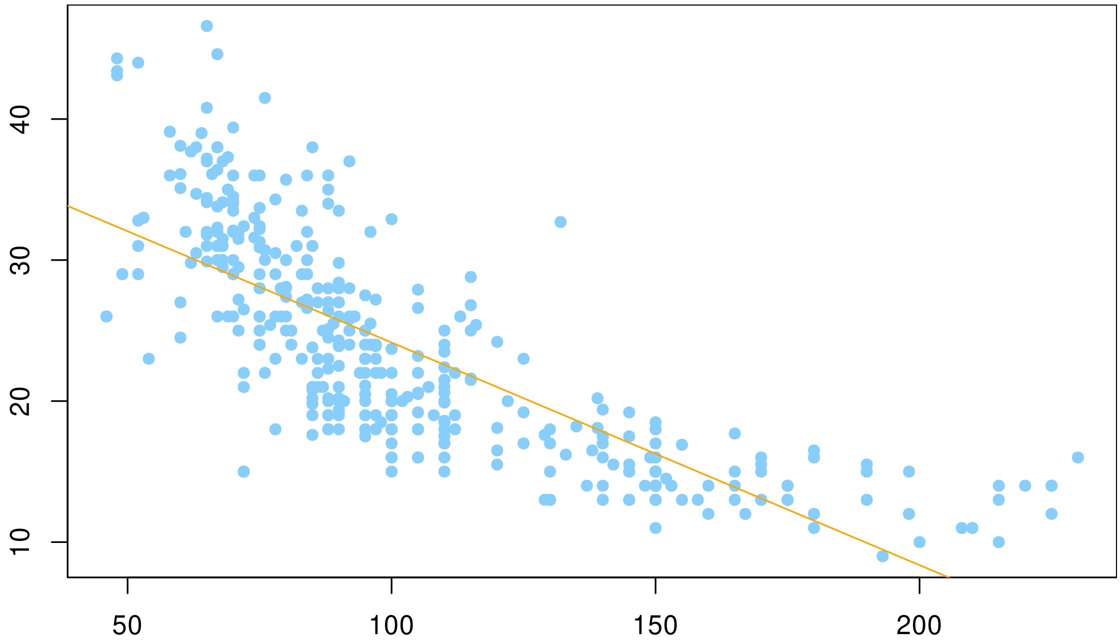
Die wahren Werte für `intercept` und `horsepower` liegen zu 95 % in den entsprechenden Intervallen. Die Intervalle sind recht schmall, so dass die Aussagekraft dieser Intervalle recht gross ist.

iv) Der R^2-Wert ist 0.606. Dieser gibt an, dass die Variabilität zu 60 % durch das Modell ist.

Das ist ok, aber nicht besonders gut, da noch andere Prädiktoren Einfluss auf den Benzinverbrauch haben.

d) Plot:

```
plot(Auto$horsepower, Auto$mpg, pch = 16, col = "lightskyblue")
abline(lm(Auto$mpg ~ Auto$horsepower), col = "orange")
```



Die sinkende Tendenz ist deutlich sichtbar, deshalb der tiefe p-Wert. Allerdings fällt die Punktwolke nicht linear (schwacher R^2-Wert).

Lösung 11.2

a) Laden von **MASS**

```
library(MASS)
help(Boston)
```

Boston {MASS}

Housing Values in Suburbs of Boston

Description

The `Boston` data frame has 506 rows and 14 columns.

Usage

`Boston`

Format

This data frame contains the following columns:

`crim`
: per capita crime rate by town.

`zn`
: proportion of residential land zoned for lots over 25,000 sq.ft.

`indus`
: proportion of non-retail business acres per town.

`chas`
: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

`nox`
: nitrogen oxides concentration (parts per 10 million).

`rm`
: average number of rooms per dwelling.

`age`
: proportion of owner-occupied units built prior to 1940.

`dis`
: weighted mean of distances to five Boston employment centres.

`rad`
: index of accessibility to radial highways.

`tax`
: full-value property-tax rate per \$10,000.

`ptratio`
: pupil-teacher ratio by town.

`black`
: *1000(Bk - 0.63)^2* where *Bk* is the proportion of blacks by town.

`lstat`
: lower status of the population (percent).

`medv`
: median value of owner-occupied homes in \$1000s.

Source

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. *J. Environ. Economics and Management* **5**, 81–102.

Belsley D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity.* New York: Wiley.

b) Spaltennamen

```
colnames(Boston)

##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"
```

```
## [7] "age"     "dis"     "rad"     "tax"     "ptratio" "black"
## [13] "lstat"  "medv"
```

c) Anhängen

```
attach(Boston)
```

d) i) Das Modell ist wie folgt definiert:

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

ii) Ausgabe

```
lm.fit <- lm(medv ~ lstat)
summary(lm.fit)

##
## Call:
## lm(formula = medv ~ lstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441,Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

e)

```
names(lm.fit)

##  [1] "coefficients"  "residuals"     "effects"
##  [4] "rank"          "fitted.values" "assign"
##  [7] "qr"            "df.residual"   "xlevels"
## [10] "call"          "terms"         "model"
```

f)

```
coef(lm.fit)

## (Intercept)       lstat
##  34.5538409  -0.9500494
```

Ersetzen Sie diese Werte in dem einfachen linearen Regressionsmodell oben

$$\texttt{medv} = 34.554 - 0.95 \cdot \texttt{lstat}$$

Die Werte $34,55$ ist der Intercept, das ist der Wert für `lstat = 0` (null Prozent des unteren Status der Bevölkerung). Der mittlere Hauswert ist \$34 554 in Nachbarschaften mit 0 Prozent Bevölkerung mit niedrigerem Status.

Der Wert $-0,95$ ist die Steigung der Regressionsgeraden. Wir können diesen Wert wie folgt interpretieren: Für jedes zusätzliche Prozent in der Bevölkerung mit niedrigerem Status sinkt der mittlere Hauswert um \$950.

g)
```
confint(lm.fit)

##                  2.5 %     97.5 %
## (Intercept) 33.448457 35.6592247
## lstat       -1.026148 -0.8739505
```

Der wahre Wert des Abschnitts liegt mit 95 %iger Wahrscheinlichkeit im Intervall

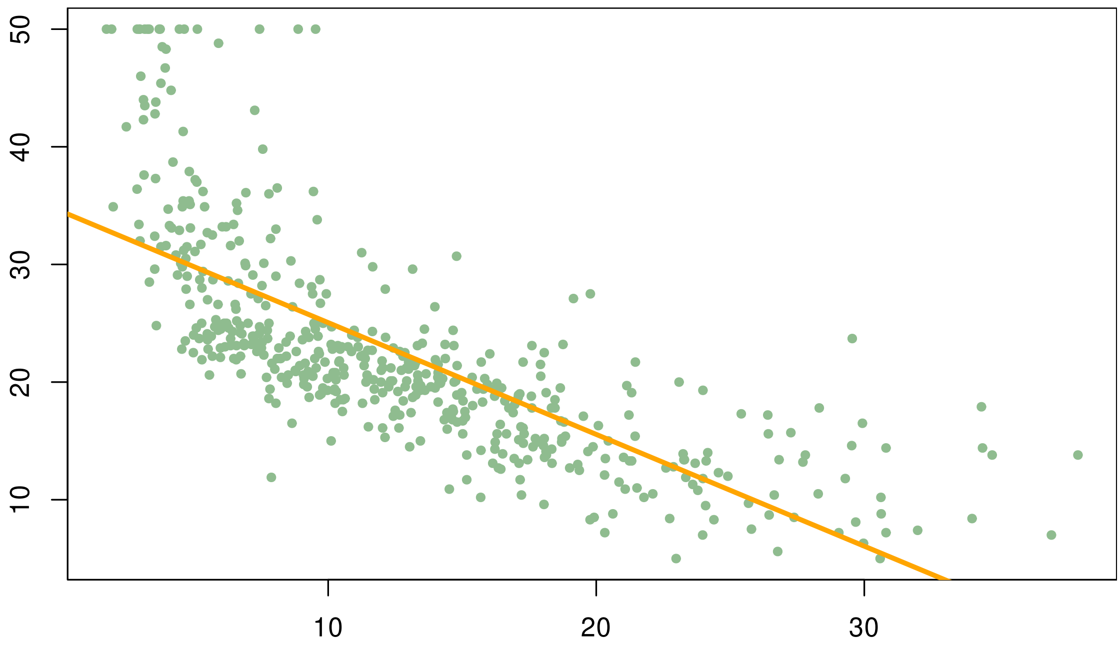
$$[33.45,\ 35,66]$$

Der wahre Wert der Steigung liegt mit 95 %iger Wahrscheinlichkeit im Intervall

$$[-1.02,\ -0.87]$$

h) Plot:

```
plot(lstat, medv, col = "darkseagreen", pch = 20)
abline(lm.fit, col = "orange", lwd = 3)
```



i) Der R^2-Wert ist 0.5441. Somit werden gut 54 % der Varianz durch das Modell (Regressionsgerade) erklärt.