Arbeitsblatt 2

Einfache lineare Regression (Interferenz & Vorhersage)

Aufgabe 1 (Antike Uhren)

Diese Aufgabe ist eine Fortsetzung vom AB 1, Aufgabe 2

McClave und Benson haben Daten über das Alter (in Jahre) und den Preis (in US$) von antiken Uhren an Auktionen zusammengetragen. Sie stehen Ihnen im Ordner `Daten` im File `AntikeUhren.dat` auf Moodle zur Verfügung.

Einlesen der Daten

```
au <- read.table("data/AntikeUhren.dat", header=TRUE)
```

(a)

Passen Sie eine Gerade an die Datenpunkte an. Geben Sie die geschätzten Koeffizientenwerte an. Wie lautet die angepasste Geradengleichung?

```
au.fit <- lm(Preis ~ Alter, data = au)
```

Die beiden geschätzten Koeffizientenwerte sind:

```
coef(au.fit)
```

```
(Intercept)       Alter
 -191.65757    10.47909
```

Die angepasste Geradengleichung ist $\widehat{\texttt{Preis}} = -191.66 + 10.48 \cdot \texttt{Alter}$.

(b)

Hat das Alter einen signifikanten Einfluss auf den Preis? Führen Sie hierfür einen geeigneten Test auf dem 5% Niveau durch.

Die Frage kann mit einem t-Test zur Nullhypothese: $H_0 : \beta = 0, \quad H_A : \beta \neq 0$ beantwortet werden. Der Test kann (1) aus dem `summary()`-Output abgelesen werden oder (2) von Hand berechnet werden.

(1) Der `summary()`-Output sieht wie folgt aus

```
summary(au.fit)
```

```
Call:
lm(formula = Preis ~ Alter, data = au)

Residuals:
    Min      1Q  Median      3Q     Max
-485.29 -192.66   30.75  157.21  541.21
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -191.66     263.89  -0.726    0.473
Alter          10.48       1.79   5.854  2.1e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 273 on 30 degrees of freedom
Multiple R-squared:  0.5332,    Adjusted R-squared:  0.5177
F-statistic: 34.27 on 1 and 30 DF,  p-value: 2.096e-06
```

Der entsprechende t-Test verwirft die Nullhypothese (p-Wert < 0.05). Das Alter der antiken Uhr hat einen auf 5% signifikanten Einfluss auf den Auktionspreis.

(2) Berechnung von Hand:

Zuerst berechnen wir die realisierte Teststatistik $t = \hat{\beta}/(\sqrt{\hat{\sigma}^2/\mathrm{SS}_x})$, also

```
hat_beta <- coef(au.fit)[2]
ssx <- sum((au$Alter - mean(au$Alter))^2)
se_beta <- sqrt(summary(au.fit)$sigma^2 / ssx)
T <- hat_beta / se_beta
T
```

```
   Alter
5.854309
```

Unter H_0 ist die Teststatistik t-verteilt mit $\mathrm{df} = 32 - 2 = 30$. Der Verwerfungsbereich für das 5% Niveau ist mittels der folgenden Quantile definiert:

```
df <- nrow(au) - 2
c(qt(0.025, df), qt(0.975, df))
```

```
[1] -2.042272  2.042272
```

Der Verwerfungsbereich ist also $]-\infty, -2.04]$ und $[2.04, \infty[$. Da die realisierte Teststatistik ($t = 5.854$) im Verwerfungsbereich liegt, wird die Nullhypothese auf dem 5% Niveau verworfen. Das Alter hat einen signifikanten Einfluss auf den erzielten Preis bei der Auktion.

Alternativ kann das Verwerfen oder Annehmen der Nullhypothese auch mittels einem P-Wert für T festgestellt werden:

```
2 * (pt(-abs(T), df)) # p-Wert
```

```
       Alter
2.096498e-06
```

Der p-Wert ist deutlich kleiner als 0.05, daher wird die Nullhypothese auf dem 5% Niveau verworfen.

(c)

Ein Händler behauptet, dass er mit seiner antiken Uhr im nächsten Jahr an der Auktion 15 US$ mehr erzielen kann. Ist dies plausibel? Führen Sie einen passenden Test auf dem 1% Niveau durch.

Seine Frage kann mit einem t-Test zur Nullhypothese: $H_0 : \beta = 15, \quad H_A : \beta \neq 15$ beantwortet

werden. Die Teststatistik lautet

$$T = \frac{\hat{\beta} - 15}{\text{se}(\hat{\beta})}$$

```
# Berechnung der realisierten Teststatistik:
hat_beta <- coef(au.fit)[2]
ssx <- sum((au$Alter - mean(au$Alter))^2)
se_beta <- sqrt(summary(au.fit)$sigma^2 / ssx)
T <- (hat_beta - 15)/ se_beta
T
```

```
    Alter
-2.525674
```

Unter H_0 ist die Teststatistik t-verteilt mit df $= 32 - 2 = 30$. Der Verwerfungsbereich für das 1% Niveau ist mittels der folgenden Quantile definiert:

```
df <- nrow(au) - 2
c(qt(0.005, df), qt(0.995, df))
```

```
[1] -2.749996  2.749996
```

Der Verwerfungsbereich ist somit $]-\infty, -2.75]$ und $[2.75, \infty[$. Da T nicht im Verwerfungsbereich liegt, wird die Nullhypothese auf dem 1% Niveau angenommen. Es kann also nicht ausgeschlossen werden, dass der Händler bei der Auktion im nächsten Jahr 15 US$ mehr erzielt.

Alternativ kann man das Verwerfen oder Annehmen der Nullhypothese auch wieder mittels dem P-Wert feststellen:

```
2 * (pt(-abs(T), df))
```

```
     Alter
0.01706362
```

Der p-Wert ist grösser als 0.01, daher wird die Nullhypothese auf dem 1% Niveau angenommen.

In dieser Aufgabe ist explizit nach einem Test gefragt. Äquivalent dazu und wesentlich schneller erhält man die Lösung über das Vertrauensintervall. Hierzu muss man prüfen, ob 15 US$ im 99% Vertrauensintervall enthalten ist. Dies ist der Fall, somit ist es durchaus plausibel, dass der Händler bei der Auktion im nächsten Jahr 15 US$ mehr erzielen kann.

```
confint(au.fit, parm = "Alter", level = 0.99)
```

```
         0.5 %   99.5 %
Alter 5.556658 15.40153
```

(d)

Der Händler beschliesst ein Jahr mit dem Verkauf seiner antiken Uhr zu warten. Welche Preiszunahme ist für die ein Jahr ältere antike Uhr plausibel? Geben Sie hierfür ein 95% Vertrauensintervall an.

Die Preiszunahme einer um ein Jahr älteren Uhr entspricht der geschätzten Steigung der Geraden:

```
coef(au.fit)[2]
```

```
   Alter
10.47909
```

Die plausiblen Werte für die Preiszunahme können mit dem 95% Vertrauensintervall zu β angegeben werden:

```
# Von "Hand"
df <- nrow(au) - 2
hat_beta <- coef(au.fit)[2]
se_beta <- summary(au.fit)$coefficients[2,2]
hat_beta + c(-1,1)*qt(0.975, df) * se_beta
```

```
[1]  6.823468 14.134721
```

```
# oder direkt:
confint(au.fit, parm = 2, level = 0.95)
```

```
         2.5 %   97.5 %
Alter 6.823468 14.13472
```

Der Händler kann mit 95% Wahrscheinlichkeit für eine ein Jahr ältere Uhr zwischen 6.82 US$ und 14.13 US$ mehr verdienen.

(e)

Ein etwas unerfahrener Käufer möchte bei der Auktion eine 160-Jahre alte, antike Uhr ersteigern. Geben Sie ihm, basierend auf dem Regressionsmodell aus Teilaufgabe (a), ein 95%-Vertrauens- und ein 95%-Prognoseintervall für den Preis einer solchen Uhr. Welches der beiden Intervalle ist für den unerfahrenen Käufer nützlicher?

```
x0 <- data.frame(Alter = 160)

## 95% Vertrauensintervall
vi95 <- predict(au.fit, newdata = x0, interval="confidence")
vi95[,c("lwr","upr")]
```

```
     lwr      upr
1372.090 1597.905
```

```
## 95% Prognose-Intervall
pi95 <- predict(au.fit, newdata = x0, interval="prediction")
pi95[,c("lwr","upr")]
```

```
      lwr       upr
 916.0829 2053.9124
```

Das Prognose-Intervall ist für den unerfahrenen Käufer wichtiger, weil es sinnvoller ist, auch die stochastische Variabilität des Beobachtungsfehlers E_i zu berücksichtigen und nicht nur die Ungenauigkeit, die sich aus der Schätzung des Modells ergibt. Der unerfahrene Käufer muss mit 95% Wahrscheinlichkeit mit einem Preis zwischen 916.08 US$ und 2053.91 US$ rechnen.

Aufgabe 2 (Conconi-Test)

Der Conconi-Test dient zur Messung der Ausdauer-Leistungsfähigkeit. Er findet auf der 400m-Bahn statt, wo man gemütlich (mit 9km/h) zu laufen beginnt. Alle 200m wird das Tempo um 0.5km/h

erhöht. Am Ende jedes 200m-Abschnitts wird die Herzfrequenz gemessen. Der Test geht so lange weiter, bis das Tempo nicht mehr erhöht werden kann. Die Daten eines Läufers stehen im File `conconi.rda` zu Verfügung.

Einlesen der Daten

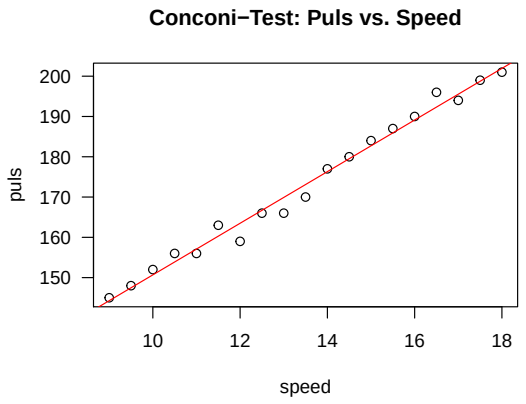
```
load("data/conconi.rda")
```

(a)

Stellen sie die Daten in einem Scatterplot dar, passen sie mit dem Befehl lm() die Regressionsgerade an und zeichnen sie diese ein.

```
## Scatterplot
plot(puls ~ speed, data = conconi, las = 1,
     main = "Conconi-Test: Puls vs. Speed")

## Regression und Einzeichnen
fit <- lm(puls ~ speed, data = conconi)
abline(fit, col="red")
```



(b)

Zu welchem Prozentanteil lassen sich die Schwankungen in den Pulswerten durch die Zunahme der Geschwindigkeit erklären?

Aus dem `summary()`-Output der Regression lässt sich das Bestimmtheitsmass R^2 ablesen:

```
summary(fit)
```

```
Call:
lm(formula = puls ~ speed, data = conconi)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4947 -1.0123  0.5228  1.1825  3.6737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  86.6105     2.5372   34.14   <2e-16 ***
speed         6.4070     0.1842   34.78   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.199 on 17 degrees of freedom
Multiple R-squared:  0.9861,    Adjusted R-squared:  0.9853
F-statistic:  1210 on 1 and 17 DF,  p-value: < 2.2e-16
```

```
# Oder direkt extrahiert
summary(fit)$r.squared
```

```
[1] 0.9861454
```

Das Multiple R-Squared ist 98.61%. Zu diesem Anteil wird die Schwankung im Puls durch die Geschwindigkeitszunahme erklärt.

(c)

Mit welcher Pulsfrequenz muss der Läufer rechnen, wenn er mit 10km/h unterwegs ist? Geben Sie ein 95% Prognoseintervall an.

```
predict(fit, newdata = data.frame(speed = 10), interval = "prediction")
```

```
       fit      lwr      upr
1 150.6807 145.7308 155.6306
```

(d)

Geben Sie an, wie hoch der Ruhepuls (d.h. keine Vorwärtsbewegung) geschätzt wird. In welchem 95% Intervall würden Sie den entsprechenden Messpunkt erwarten.

Der Schätzwert für die Regressionsgerade an der Stelle `speed` $= 0$ ist der Achsenabschnitt.

```
coef(fit)[1]
```

```
(Intercept)
   86.61053
```

Das 95% Vertrauensintervall zum Achsenabschnitt gibt die plausiblen Werte für den Ruhepuls an:

```
confint(fit, level = 0.95)
```

```
                2.5 %    97.5 %
(Intercept) 81.257578 91.963475
speed        6.018418  6.795617
```

Zum identischen Ergebnis kommt man auch bei einer Vorhersage für `speed` $= 0$ mit 95% Vertrauensintervall:

```
predict(fit, newdata = data.frame(speed = 0), interval = "confidence")
```

```
       fit      lwr      upr
1 86.61053 81.25758 91.96347
```

Zu beachten ist, dass es sich bei dieser Vorhersage um eine Extrapolation handelt. Sie ist daher grundsätzlich nicht als verlässlich anzusehen. Allerdings scheint der vorhergesagte Wert wie auch das

Intervall realistisch. Somit könnte es durchaus sein, dass die lineare Beziehung auch für langsamere Geschwindigkeiten bis hin zum Stillstand gilt. Um sicher zu sein, bräuchten wir aber zusätzliche Messpunkte.

(e)

Um wie viel nimmt der Puls im Schnitt zu, wenn die Geschwindigkeit um 1 km/h erhöht wird? Welche anderen Werte sind für die Pulszunahme ebenfalls plausibel?

Die Antwort entspricht der geschätzten Steigung. Der Puls nimmt um 6.4 zu, wenn die Geschwindigkeit um 1 km/h erhöht wird. Die plausiblen Werte erhalten wir mit einem 95%-Vertrauensintervall für die Steigung:

```
confint(fit, parm = 2, level = 0.95)
```

```
         2.5 %   97.5 %
speed 6.018418 6.795617
```

Mit 95% Wahrscheinlichkeit liegt die Pulszunahme pro zusätzlichen 1 km/h Geschwindigkeit zwischen 6.02 und 6.80.

(f)

Im File `conconi2.rda` stehen Ihnen die Daten eines zweiten Läufers zur Verfügung. Wessen Puls steigt bei Geschwindigkeitserhöhung langsamer an? Können sie eine Aussage treffen, ob zwischen den beiden ein auf 5% signifikanter Unterschied besteht? Lässt sich ableiten, wer der besser trainierte Läufer ist?

```
load("data/conconi2.rda")
fit2 <- lm(puls ~ speed, data = conconi2)
coef(fit2)
```

```
(Intercept)       speed
  84.238346    4.093233
```

```
confint(fit2, parm = 2, level = 0.95)
```

```
         2.5 %   97.5 %
speed 3.883739 4.302727
```

Beim zweiten Läufer ist der Schätzwert für die Steigung 4.093 und somit tiefer als beim ersten Läufer. Die beiden 95 % Vertrauensintervalle für die Steigung überlappen nicht, somit kann man auch ohne formellen Test die Aussage treffen, dass die beiden Werte signifikant unterschiedlich sind.

Lässt sich daraus ableiten, wer der bessere Läufer ist? Jein. Der langsamere Pulsanstieg deutet darauf hin, aber die maximal erreichbare Pulsobergrenze ist für jeden Menschen individuell. Deshalb kann man keine sichere Aussage treffen.