Hinweise zu Blatt 3

Christoph Becker

2024-05-02

Hier ist der R-Code zur Berechnung der linearen Regression anhand einfacher Beispieldaten. Wir starten mit Daten (Punktepaaren) $(x_1, y_1), \ldots, (x_n, y_n)$ und verwenden als Modell

$$y = a + bx + \epsilon,$$

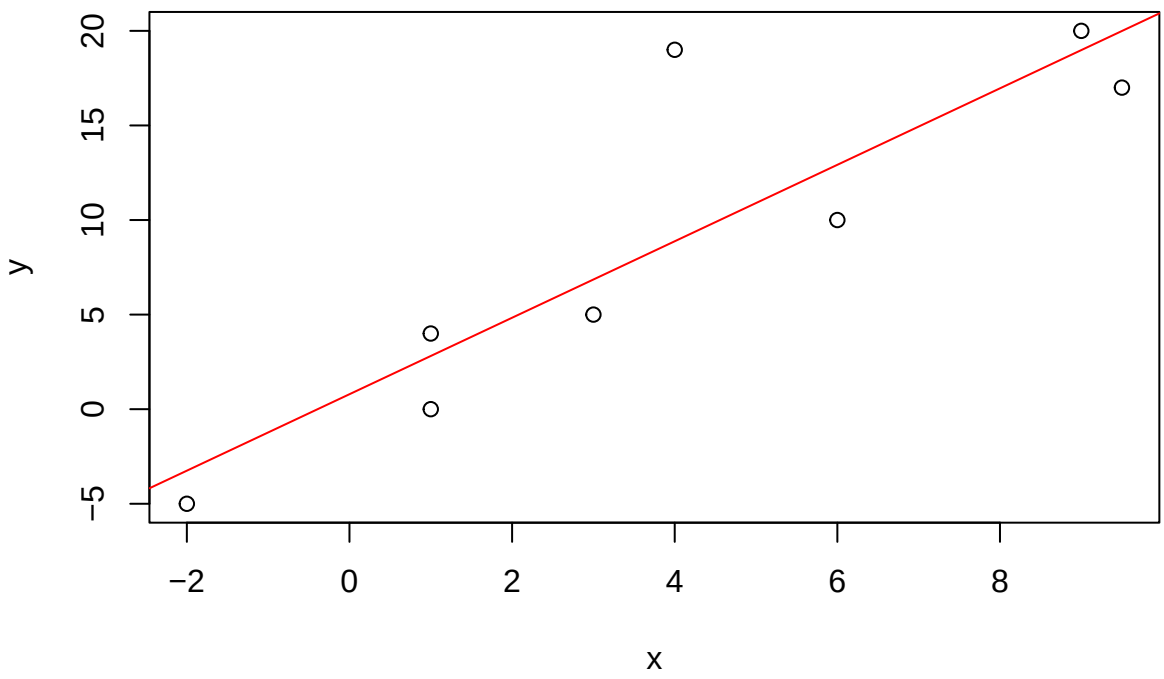
wobei a den Intercept und b die Steigung ('slope') der Regressionsgeraden bezeichnet. Der Ausdruck ϵ bezeichnet den Fehler, also die Abweichung der Daten von der Regressionsgeraden. Im Durchschnitt ist der Fehler $\epsilon = 0$.

Im folgenden R-Code nutzen wir den Befehl 'lm' zur Schätzung der Regression, d.h. eines **l**inearen **M**odells. Der Befehl 'lm' gibt ein sogenanntes Datenobjekt zurück, das ich anschließend an andere Befehle weiterreiche, um die geschätzten Parameter anzuzeigen (Befehl summary) oder die Regressionsgerade zu zeichen (Befehl abline).

```
x = c(9,-2, 3, 1, 9.5,6,4,1)
y = c(20, -5, 5, 4, 17, 10, 19, 0)
plot(x,y)
regressionsErgebnis = lm(y~x)
summary(regressionsErgebnis)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.994 -2.839 -1.802  1.060 10.124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7910     2.4384   0.324    0.757
## x             2.0213     0.4468   4.524    0.004 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.775 on 6 degrees of freedom
## Multiple R-squared:  0.7733, Adjusted R-squared:  0.7355
## F-statistic: 20.46 on 1 and 6 DF,  p-value: 0.004001
```

```
abline(regressionsErgebnis, col="red")
```



Die geschätzte Steigung der Geraden (Parameter b) ist 2.02. Wenn also x um eine Einheit wächst, dann wird y im Durchschnitt (also wenn wir entlang der Regressionsgeraden wandern) um 2.02 wachsen. Der Intercept a ist hier 0.79. Wenn also x den Wert Null annimmt (d.h. wir berechnen die Prognose des Modells für $x = 0$), dann erhalten wir als Ergebnis für den Wert auf der Regressionsgeraden 0.79. Manchmal lässt sich so ein Zusammenhang passend interpretieren, manchmal nicht. Wir werden Beispiele dazu besprechen. Alle weiteren Ausgaben des summary-Befehls besprechen wir später in der Vorlesung.

Noch ein Hinweis zur Aufgabe 2a). Hier sollen Sie aus den angegebenen Daten die Korrelation zwischen x und y berechnen und dann den Wert dieser Korrelation interpretieren. Zur Berechnung der Korrelation denken Sie bitte an den Zusammenhang

$$\hat{b} = \frac{\text{Corr}(x, y)s_y}{s_x}.$$

Und als Hinweis für die Aufgabe 1a: So berechnen Sie in R Varianz, Standardabweichung (in R: Befehl sd für standard deviation) Kovarianz und Korrelation:

```
x = c(9,-2, 3, 1, 9.5,6,4,1)
y = c(20, -5, 5, 4, 17, 10, 19, 0)
var(x)
```

```
## [1] 16.31696
```

```
sd(x)
```

```
## [1] 4.039426
```

```r
cov(x,y)
```

```
## [1] 32.98214
```

```r
cor(x,y)
```

```
## [1] 0.8793658
```

Ich hoffe, mit diesen Hinweisen kommen Sie weiter. Emailen Sie mir, wenn Sie noch Fragen haben.

Herzliche Grüße,

Christoph Becker