lineare Regression mit R

Benjamin Schlegel

29. März 2016

Ein lineares Modell kann in R mit dem Befehl lm() gerechnet werden. Was eine lineare Regression ist, kann im Artikel lineare Regression nachgelesen werden.

In diesem Artikel wird die lineare Regression in R anhand eines Beispiels mit dem Datensatz world_data.csv gezeigt. Der Datensatz kann unter Daten heruntergeladen werden.

Der Datensatz enthält die meisten Länder der Erde mit den durchschnittlichen Anzahl Kindern im Jahr 2012 (fertility_2012), den Anzahl Jahren obligatorische Schulzeit im Jahr 2012 (educ_2012) und einen Index, welcher die politischen Rechte der Frauen im Jahr 2011 misst (woman_right_2011). Der Index kann die Werte 0, 1, 2 und 3 annehmen (0: keine Rechte im Gesetz; 3: (alle) Rechte im Gesetz und in der Praxis).

In einem ersten Schritt müssen die Daten eingelesen werden.

```
data = read.csv2(file.choose(),stringsAsFactors = FALSE)
```

Als nächstes müssen die Variablen rekodiert werden.

```
data$fertility_2012 = as.numeric(data$fertility_2012)
data[which(data$woman_right_2011==-77),"woman_right_2011"] = 0
data$woman_right_2011 = as.factor(data$woman_right_2011)
```

Jetzt können wir uns einen Überblick über den Datensatz schaffen.

```
head(data)
```

```
                CTRY fertility_2012 educ_2012 woman_right_2011
1         Afghanistan          5.141         9                2
2             Albania          1.760         9                2
3             Algeria          2.820        11                2
4             Andorra          5.979         6                3
5              Angola          2.102        11                3
6 Antigua and Barbuda          2.188        13                3
```

```
summary(data)
```

```
     CTRY           fertility_2012    educ_2012      woman_right_2011
 Length:182         Min.   :1.260   Min.   : 0.000   0:  1
 Class :character   1st Qu.:1.792   1st Qu.: 7.000   1: 19
 Mode  :character   Median :2.405   Median : 9.000   2:133
```

```
Mean   :2.901   Mean   : 8.313   3: 29
3rd Qu.:3.842   3rd Qu.:10.000
Max.   :7.574   Max.   :15.000
```

Modell mit einer stetigen Variable

Nun sind wir soweit und können mit der linearen Regression beginnen. In einem ersten Schritt wollen wir die Hypothese überprüfen: "Je länger die obligatorische Bildung in einem Land dauert, desto weniger Kinder haben die Frauen."

Um die Hypothese zu überprüfen verwenden wir die Methode lm(). Als Parameter geben wir die Formel an, wobei zuerst die abhängige Variable kommt, dann ein Tilde und dann die unabhängigen Variablen. Der Datensatz wird mit dem Parameter data angegeben.

```
model1 = lm(fertility_2012 ~ educ_2012, data=data)
summary(model1)
```

```
Call:
lm(formula = fertility_2012 ~ educ_2012, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8819 -1.1009 -0.5064  1.1004  4.6031

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.34393    0.26128  12.798
```