Unterricht 1

Einführung in Regression

Stefan Schmidt

08.04.2020

Frage: Lässt sich die Zeit, die ein Serviceangestellter für die Bestückung und den Service des Getränkeautomaten braucht, quantitativ beschreiben? Wovon hängt die Zeit ab?

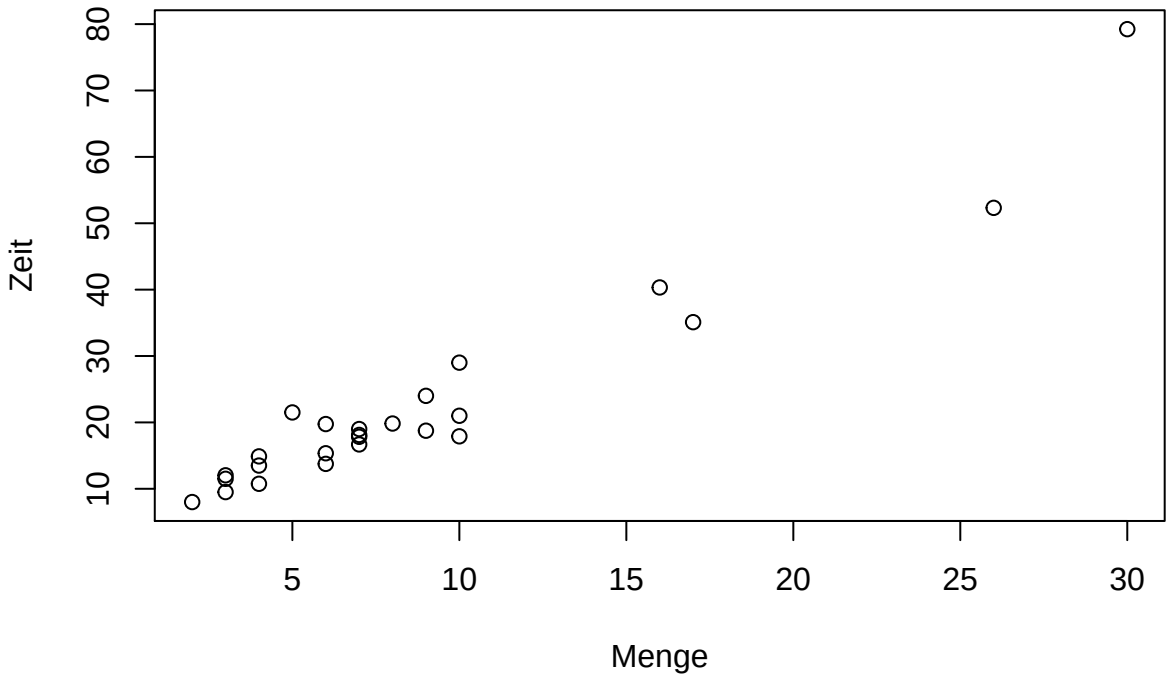
Daten: 25 zufällig ausgewählte Getränkeautomaten in 4 US-Grossstädten

```
ga <- read.table(paste0(path, "data/softdrink.dat"), header = TRUE)
str(ga)
```

```
## 'data.frame':    25 obs. of  4 variables:
##  $ Zeit   : num  16.7 11.5 12 14.9 13.8 ...
##  $ Menge  : int  7 3 3 4 6 7 2 7 30 5 ...
##  $ Distanz: num  168 66 102 24 45 ...
##  $ Ort    : chr  "San Diego" "San Diego" "San Diego" "San Diego" ...
```

Einfache Lineare Regression

```
plot(Zeit ~ Menge, data = ga)
```



Frage: Wie gross wäre die Servicezeit beim Nachfüllen von 20 Produkteinheiten?

Schätzung fuer α und β

```
beta_hat <- sum((ga$Zeit - mean(ga$Zeit)) * (ga$Menge - mean(ga$Menge))) /
  sum((ga$Menge - mean(ga$Menge))^2)

alpha_hat <- mean(ga$Zeit) - beta_hat * mean(ga$Menge)

c(alpha=alpha_hat, beta=beta_hat)
```

```
##    alpha     beta
## 3.320780 2.176167
```

Berechnung der Koeffizienten mit R:

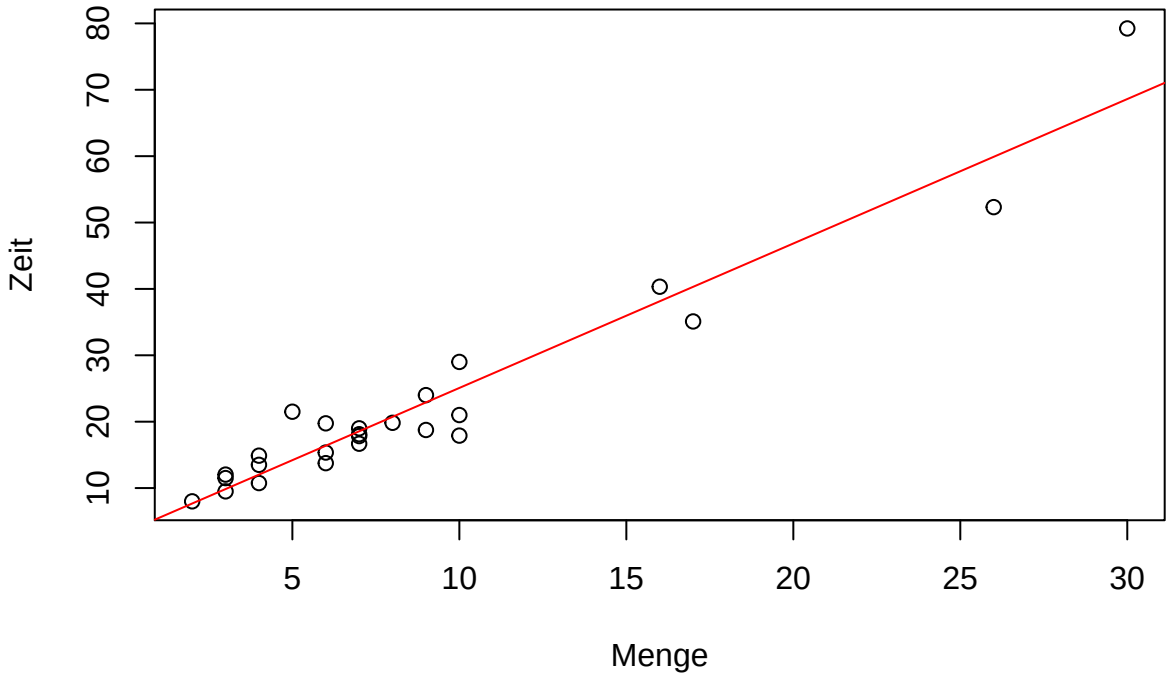
```
ga.fit <- lm(Zeit ~ Menge, data = ga)
coef(ga.fit)
```

```
## (Intercept)       Menge
##    3.320780    2.176167
```

Die geschätzte Gerade lautet somit: Zeit $= 3.32078 + 2.17617 \cdot$ Menge

Einzeichnen der Geraden:

```
plot(Zeit ~ Menge, data = ga)
abline(ga.fit, col = 'red')
```



Summe der Abweichungen addiert sich zu Null

```
ga.ri <- ga$Zeit - (alpha_hat + beta_hat * ga$Menge) # Residuen
sum(ga.ri) # Summe der Residuen: ~ 0
```

```
## [1] -2.309264e-14
```

Die Gerade geht durch den Datenschwerpunkt(x_quer, y_quer)

```
alpha_hat + beta_hat * mean(ga$Menge)
```

```
## [1] 22.384
```

```
mean(ga$Zeit)
```

```
## [1] 22.384
```

Koeffizienten, angepasste Werte, Residuuen

```
ga.fit <- lm(Zeit ~ Menge, data = ga)
coef(ga.fit) # Koeffizenten
```

```
## (Intercept)       Menge
##    3.320780    2.176167
```

```
fitted(ga.fit)
```

```
##         1         2         3         4         5         6         7         8
## 18.553947  9.849280  9.849280 12.025447 16.377780 18.553947  7.673113 18.553947
##         9        10        11        12        13        14        15        16
## 68.605780 14.201613 38.139447 25.082447 12.025447 16.377780 22.906280 25.082447
##        17        18        19        20        21        22        23        24
## 16.377780 18.553947  9.849280 40.315613 25.082447 59.901114 22.906280 20.730113
##        25
## 12.025447
```

```
# Alternative von "Hand"
coef(ga.fit)[1] + coef(ga.fit)[2] * ga$Menge
```

```
##  [1] 18.553947  9.849280  9.849280 12.025447 16.377780 18.553947  7.673113
##  [8] 18.553947 68.605780 14.201613 38.139447 25.082447 12.025447 16.377780
## [15] 22.906280 25.082447 16.377780 18.553947  9.849280 40.315613 25.082447
## [22] 59.901114 22.906280 20.730113 12.025447
```

```
resid(ga.fit)
```

```
##          1          2          3          4          5          6          7
## -1.8739466  1.6507201  2.1807201  2.8545534 -2.6277800 -0.4439466  0.3268867
##          8          9         10         11         12         13         14
## -0.7239466 10.6342198  7.2983867  2.1905532 -4.0824467  1.4745534  3.3722200
##         15         16         17         18         19         20         21
##  1.0937200  3.9175533 -1.0277800  0.4460534 -0.3492799 -5.2156134 -7.1824467
##         22         23         24         25
## -7.5811135 -4.1562800 -0.9001133 -1.2754466
```

```
# Alternative von "Hand"
y_hat <- coef(ga.fit)[1] + coef(ga.fit)[2] * ga$Menge
ga$Zeit - y_hat
```

```
##  [1] -1.8739466  1.6507201  2.1807201  2.8545534 -2.6277800 -0.4439466
##  [7]  0.3268867 -0.7239466 10.6342198  7.2983867  2.1905532 -4.0824467
## [13]  1.4745534  3.3722200  1.0937200  3.9175533 -1.0277800  0.4460534
## [19] -0.3492799 -5.2156134 -7.1824467 -7.5811135 -4.1562800 -0.9001133
## [25] -1.2754466
```

Schätzung der Fehlervarianz $var(E_i) = \sigma^2$

Der Standardfehler der Residuen ($\hat{\sigma}$)gibt an, wie stark die Beobachtungen um die Regressionsgerade streuen.

Falls die Modellannahmen erfüllt sind, so können wir davon ausgehen, dass sich rund 95% der Punkte in einem Intervall von $\pm 2 * \hat{\sigma}$ um die Regressionsgerade befinden.

```
ga.fit <- lm(Zeit ~ Menge, data = ga)
summary(ga.fit)$sigma
```

```
## [1] 4.181397
```

Falls die Annahmen an die Fehler im Beispiel für die Getränkeautomaten erfüllt sind, so befinden sich rund 95% der Punkte in einem Intervall von $\pm 2 * 4.18 = \pm 8.36$ Minuten um die Regressionsgerade.

Achtung: Standard-Fehler σ NICHT Fehlervarianz σ^2!

Übung: Analyse des Effekts von Isolierungssanierung in 56 Häusern

```
library(MASS)
data(whiteside)
str(whiteside)
```

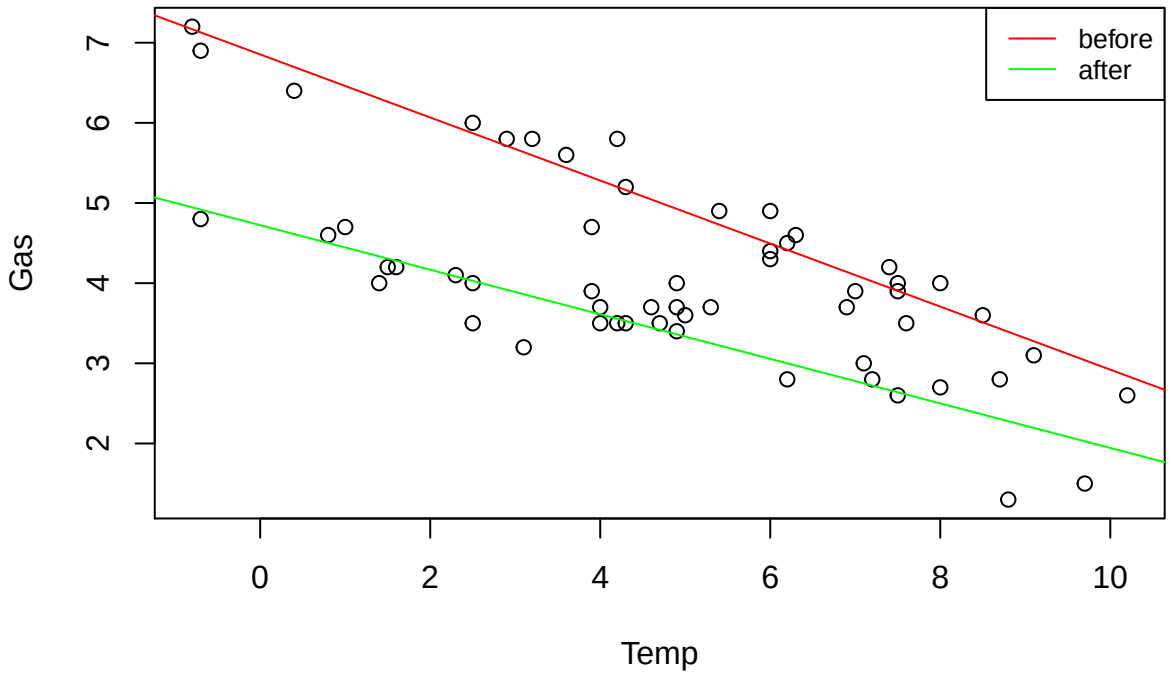
```
## 'data.frame':    56 obs. of  3 variables:
##  $ Insul: Factor w/ 2 levels "Before","After": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Temp : num  -0.8 -0.7 0.4 2.5 2.9 3.2 3.6 3.9 4.2 4.3 ...
##  $ Gas  : num  7.2 6.9 6.4 6 5.8 5.8 5.6 4.7 5.8 5.2 ...
```

Daten VOR und NACH der Sanierung

```
before <- whiteside[whiteside$Insul == "Before",]
b.fit <- lm(Gas ~ Temp, data = before)

after <- whiteside[whiteside$Insul == "After",]
a.fit <- lm(Gas ~ Temp, data = after)

plot(Gas ~ Temp, data = whiteside)
abline(b.fit, col = "red")      # before
abline(a.fit, col = "green")    # after
legend("topright", legend=c("before", "after"), col=c("red", "green"), lty=c(1, 1), cex=0.8)
```


before
after
Gas
Temp
2
3
4
5
6
7
0
2
4
6
8
10