

Kovarianz, Korrelation, (lineare) Regression

Jonathan Harrington & Ulrich Reubold

```
library(ggplot2)
```

```
epg = read.table(file.path(pfadu, "epg.txt"))
```

```
amp = read.table(file.path(pfadu, "dbdauer.txt"))
```

Kovarianz, Korrelation, (lineare) Regression

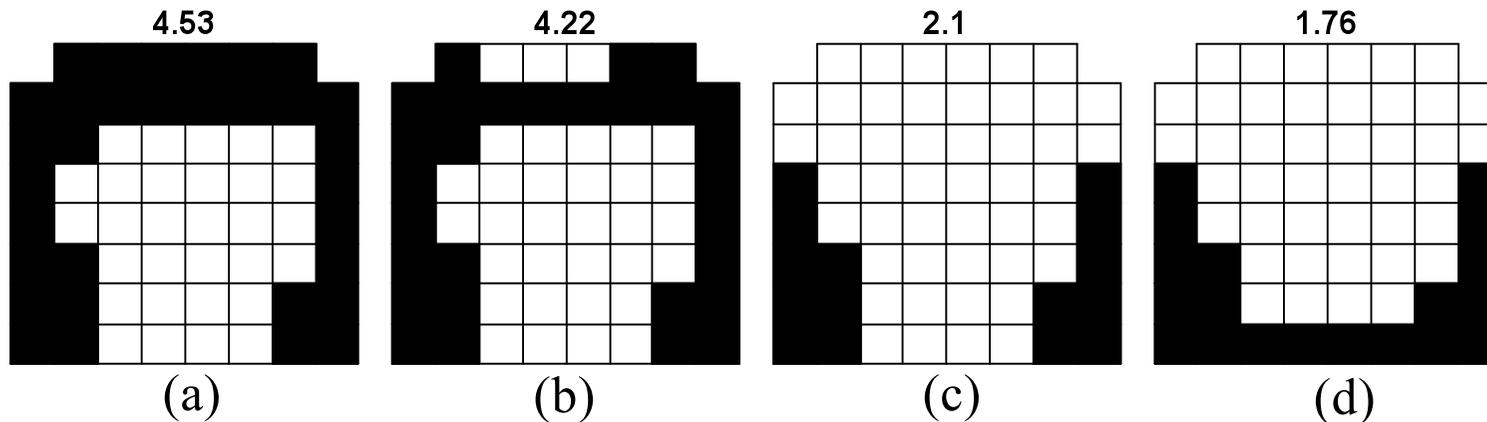
Eine Messung der Stärke der Beziehung zwischen 2 numerischen Variablen.

head(epg)

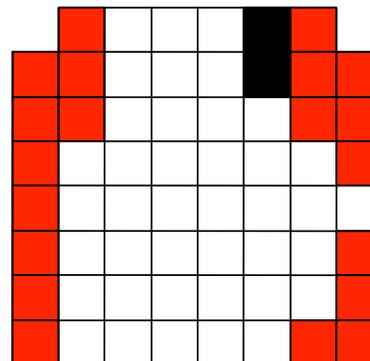
Vk-Reihenfolgen von einem deutschen Muttersprachler. $V = /a \ \varepsilon \ \text{ɪ} \ \text{ɔ} \ \text{ʊ}/$

F1, F2: F1 und F2-Werte zum Vokaloffset

EPG-Parameter COG (Centre of Gravity) zum selben Zeitpunkt



EPG-Parameter SUM1278

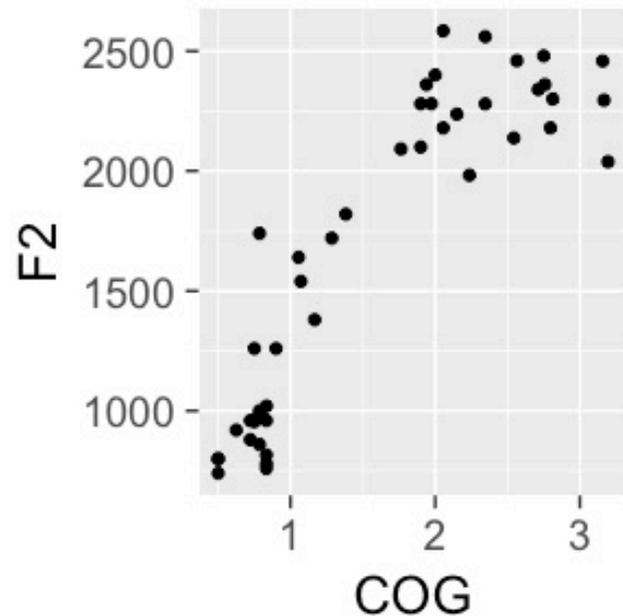


1. Kovarianz

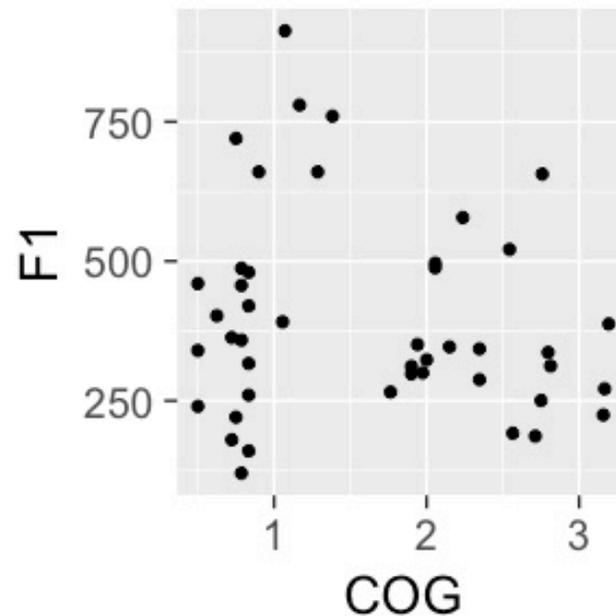
Je mehr die Kovarianz von 0 (Null) abweicht, umso deutlicher die lineare Beziehung zwischen den Variablen

Kovarianz-Werte

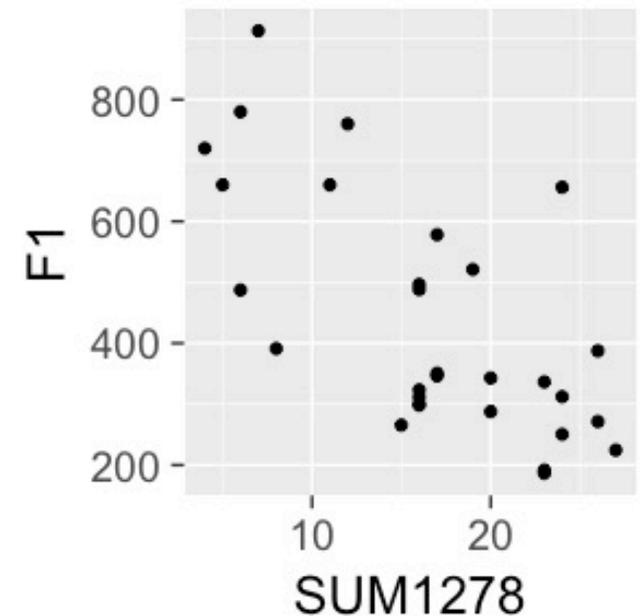
hoch und positiv
509.6908



nah an 0
-24.26598



mittel und negativ
-289.516



Berechnung der Kovarianz

Produkt-Summe der Abweichungen vom Mittelwert

$$y = \text{epg}\$F2$$

$$x = \text{epg}\$COG$$

$$n = \text{length}(y)$$

Mittelwert

$$mx = \text{mean}(x)$$

$$my = \text{mean}(y)$$

Abweichungen vom Mittelwert

$$dx = x - \text{mean}(x)$$

$$dy = y - \text{mean}(y)$$

Kovarianz = Produkt-Summe der Abweichungen dividiert durch n-1

$$\text{covxy} = \text{sum}(dx * dy) / (n - 1)$$

Funktion in R

$$\text{cov}(x, y)$$

Einige Merkmale der Kovarianz

$\text{cov}(x, y)$	gleich	$\text{cov}(y, x)$
$\text{cov}(x, x)$	gleich	$\text{var}(x)$
$\text{var}(x+y)$	gleich	$\text{var}(x) + \text{var}(y) + 2 * \text{cov}(x, y)$

daher: wenn es keine lineare Beziehung zwischen x und y gibt ist $\text{cov}(x, y)$ 0 (Null) sodass

$\text{var}(x+y)$	gleich	$\text{var}(x) + \text{var}(y)$
-------------------	--------	---------------------------------

2. Kovarianz und Korrelation

Die Korrelation (Pearson's product-moment correlation), r , ist dasselbe wie die Kovarianz, aber sie normalisiert für die Größe von x und y

- r ist die Kovarianz von x , y , dividiert durch deren Standardabweichungen
- r variiert zwischen -1 und +1

`cov(x,y)`

[1] 509.6908

`xgross = x*1000`

`cov(xgross,y)`

[1] 509690.8

`r = cov(x,y)/(sd(x) * sd(y))`

`cor(x,y)`

[1] 0.8917474

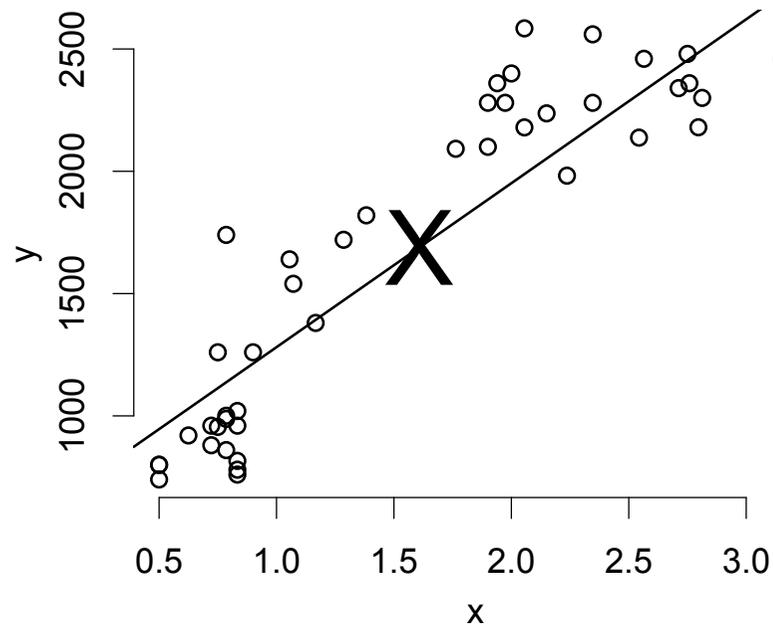
`cor(xgross,y)`

[1] 0.8917474

3. Regression

y-auf-x Regression: y (abhängige Variable) soll durch x (unabhängige Variable) modelliert werden, also **durch die Werte von x eingeschätzt werden.**

Regressionslinie: Eine gerade Linie durch die Verteilung, sodass der Abstand der Stichproben zu der Linie **minimiert** wird.



Diese Regressionslinie durchschneidet (m_x, m_y) den Mittelwert (\bar{X}) der Verteilung

3. Regression

Die Regressionslinie:

$$\hat{y} = bx + k$$

b ist die Die Steigung

$$b = r * sd(y)/sd(x) \quad \text{oder} \quad b = cov(x,y)/var(x)$$

k ist das Intercept (y-Achsenabschnitt)

$$k = my - b*mx$$

\hat{y} sind die eingeschätzten Werte, die auf der Regressionslinie liegen

$$y_{\text{hut}} = b*x + k$$

Abbildung

`plot(y ~ x)`

Regressionslinie überlagern

`abline(k, b)`

Eingeschätzte Werte überlagern

`points(x, yhut, col = 2)`

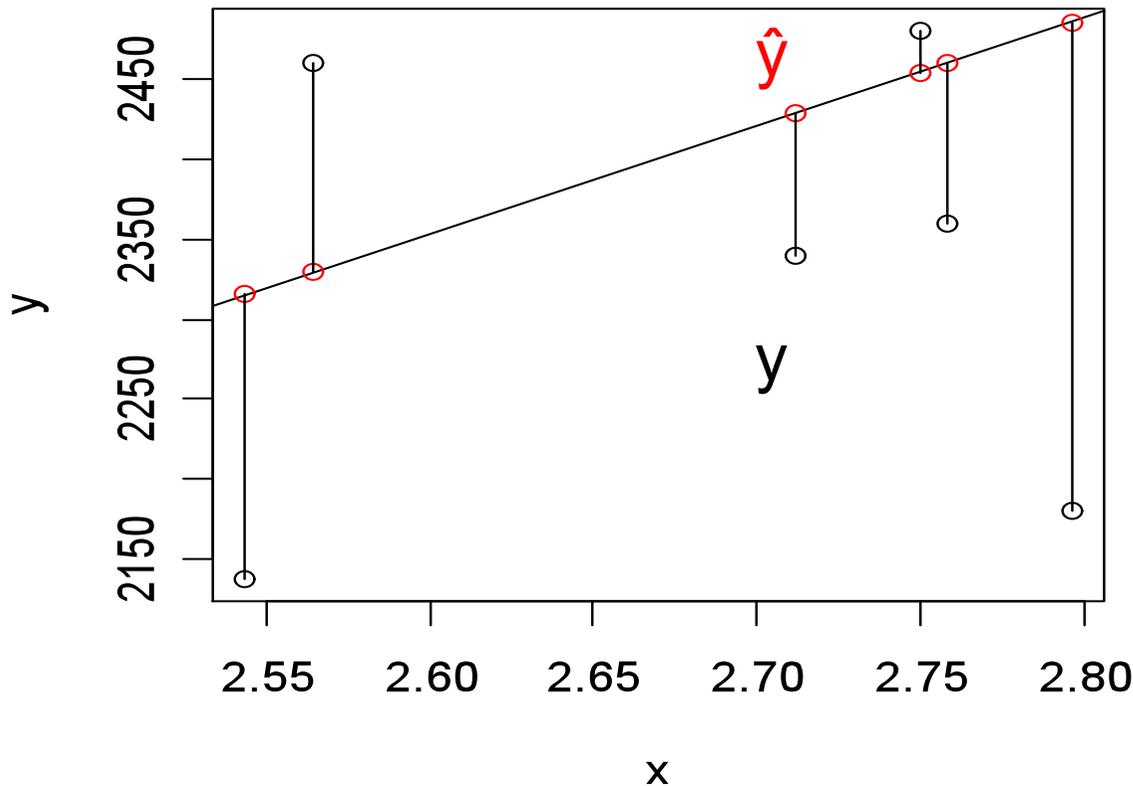
Error und SSE

Der **error** (auch residuals) ist der Unterschied zwischen den tatsächlichen und **eingeschätzten** Werten.

$$\text{error} = y - \hat{y}$$

SSE: sum of the squares of the error

$$\text{SSE} = \sum(\text{error}^2)$$



In der Regression wird die Linie auf eine solche Weise berechnet, dass SSE (RSS¹) **minimiert** wird.

1. wird auch manchmal RSS residual sum of squares genannt

Regression mit lm()

`reg = lm(y ~ x)` `~` wird modelliert c

Regressionslinie überlagern

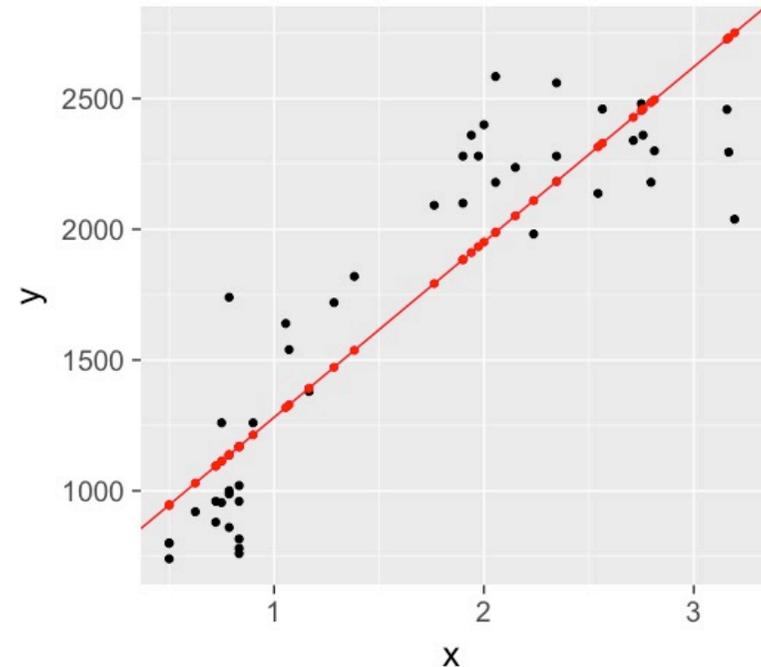
`plot(y ~ x)`

`abline(reg)`

Regressionskoeffiziente

`coef(reg)` **Intercept**
(Intercept)
610.6845

Steigung
x
670.2670



Eingeschätzte Werte

`yhut = predict(reg)`

$$yhut = b * x + k$$

Error

`residuals(reg)`

$$error = y - yhut$$

SSE

`deviance(reg)`

$$\text{sum}(error^2)$$

Regression: drei wichtige Quantitäten

1. **SSE** (oder RSS) sum of the squared errors

$$SSE = \text{sum}(\text{error}^2)$$

$$\text{oder } SSE = \text{deviance}(\text{reg})$$

2. **SSY** (oder SST): sum-of-the-squared deviations der tatsächlichen Werte

$$SSY = \text{sum}((y - \bar{y})^2)$$

3. **SSR**: sum of the squared-deviations von \hat{y} , also von den eingeschätzten Werten

$$SSR = \text{sum}((\hat{y} - \bar{y})^2)$$

$$SSY = SSR + SSE$$

R-squared (R^2)

$$SSY = SSR + SSE$$

Je besser die Werte durch die Regressionslinie modelliert werden (also je geringer der Abstand zwischen y und \hat{y}) umso kleiner SSE, sodass im besten Fall $SSE = 0$ und $SSY = SSR$ oder $SSR/SSY = 1$ (bedeutet: die tatsächlichen Werte sitzen auf der Linie).

R-squared = SSR/SSY beschreibt auch die Proportion der Varianz in y die **durch die Regressionslinie erklärt werden kann**

R-squared variiert zwischen 0 (keine 'Erklärung') und 1 (die Regressionslinie erklärt 100% der Varianz in y).

R-squared (fortgesetzt)

$$SSY = SSR + SSE$$

Diese Quantität SSR/SSY nennt man auch **R-squared** weil sie denselben Wert hat wie den Korrelationskoeffizient hoch zwei.

SSR/SSY

$cor(x, y)^2$

[1] 0.7952134

(und da r zwischen -1 und 1 variiert, muss R-squared zwischen 0 und 1 variieren)

Signifikanz-Test

Was ist die Wahrscheinlichkeit, dass ein lineares Verhältnis zwischen x und y besteht?

Dies kann mit einem t-test mit $n-2$ Freiheitsgraden berechnet werden:

$$tstat = r/rsb$$

$$rsb = \text{Standard-error von } r = \sqrt{\frac{1 - r^2}{n - 2}}$$

$$rsb = \text{sqrt}((1 - r^2)/(n-2))$$

$$tstat = r/rsb$$

$$[1] 12.92187$$

Signifikanz-Test

$$tstat = r/rsb$$

[1] 12.92187

$$fstat = tstat^2$$

[1] 166.9746

Ein t-test mit n-2
Freiheitsgraden

Ein F-test mit 1 und n-2
Freiheitsgraden

$$2 * (1 - pt(tstat, n-2))$$

$$1 - pf(fstat, 1, n-2)$$

bekommt man auch durch `cor.test(x,y)`

[1] 2.220446e-16

= 2.220446 x 10⁻¹⁶

Die Wahrscheinlichkeit, dass die Variablen nicht miteinander linear assoziiert sind, ist fast 0. (Hoch signifikant, $p < 0.001$).

summary(reg)

Es gibt eine signifikante lineare Beziehung zwischen COG und F2
($R^2 = 0.80$, $F[1, 43] = 167$, $p < 0.001$).

Call:

lm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-713.17	-195.81	-99.32	215.81	602.68

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	610.68	94.65	6.452	8.03e-08 ***
x	670.27	51.87	12.922	< 2e-16 ***

Residual standard error: **300** on 43 degrees of freedom

Multiple R-Squared: 0.7952, Adjusted R-squared: 0.7905

F-statistic: **167** on 1 and 43 DF, **p-value: < 2.2e-16**

Gültigkeit der Regression¹

Die Residuals sollen:

- (a) von einer Normalverteilung nicht signifikant abweichen.
- (b) eine konstante Varianz aufweisen.
- (c) keine Autokorrelation aufweisen.

1. siehe auch

http://scc.stat.ucla.edu/page_attachments/0000/0139/reg_1.pdf

(a) Sind die Residuals normalverteilt?

```
shapiro.test(resid(reg))
```

Shapiro-Wilk normality test

```
data: resid(regp)
```

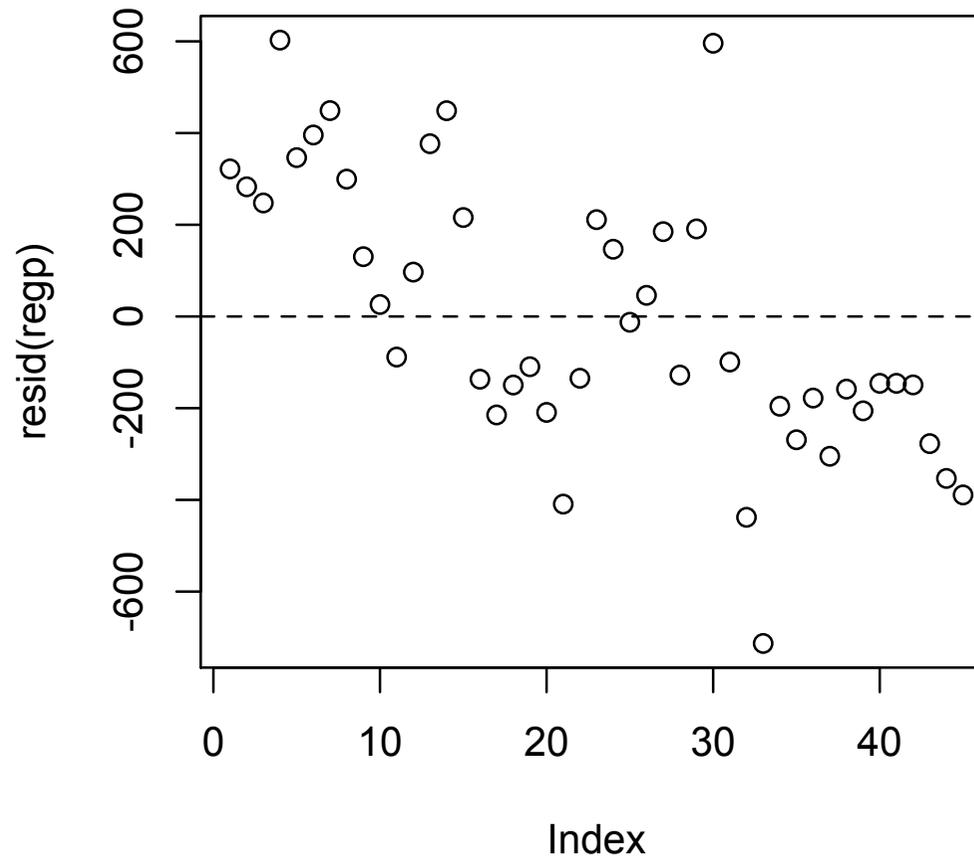
```
W = 0.9704, p-value = 0.2987
```

(b) Haben die Residuals eine konstante Varianz?

Insbesondere sollten die Residuals nicht wesentlich größer am Anfang/Ende sein, sondern auf eine randomisierte Weise um die 0 Linie verteilt sein. Das ist hier nicht der Fall.

```
plot(resid(reg))
```

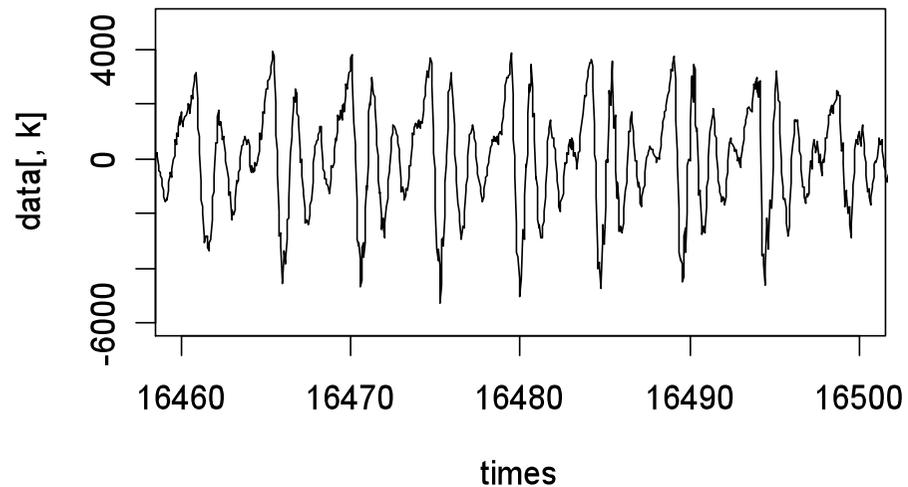
```
abline(h=0, lty=2)
```



(c) Keine Autokorrelation

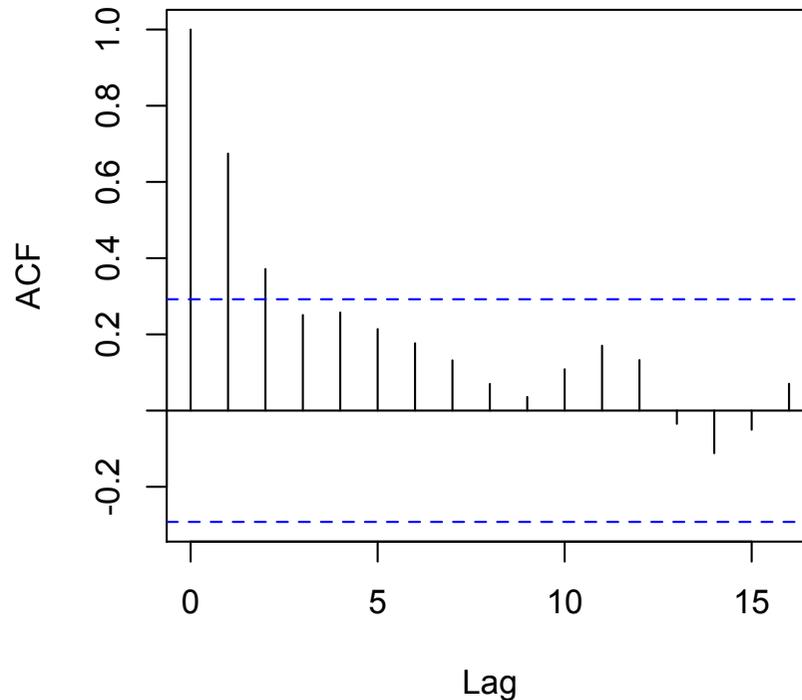
Autokorrelation ist wenn die Werte eines Signals mit sich selbst korreliert sind

Ein gutes Beispiel von autokorrelierten Daten: ein stimmfahtes Sprachsignal



(c) Keine Autokorrelation

Series resid(regp)



`acf(resid(reg))`

95% Konfidenzintervall um 0

Wenn die meisten ACF-Werte innerhalb der blauen Linien liegen, gibt es keine Autokorrelation.

Insbesondere die Werte bei lag 1 und 2 beobachten: diese sollten innerhalb des Vertrauensintervalls liegen (nicht der Fall).