Kovarianz, Korrelation, (lineare) Regression

Jonathan Harrington

BITTE NOCH EINMAL dframes.zip (Webseite 4.1)
herunterladen und in pfad auspacken

Kovarianz, Korrelation, (lineare) Regression

messen alle inwiefern es eine lineare Beziehung zwischen zwei Variablen gibt…

head(epg)

Vk-Reihenfolgen von einem deutschen Muttersprachler.

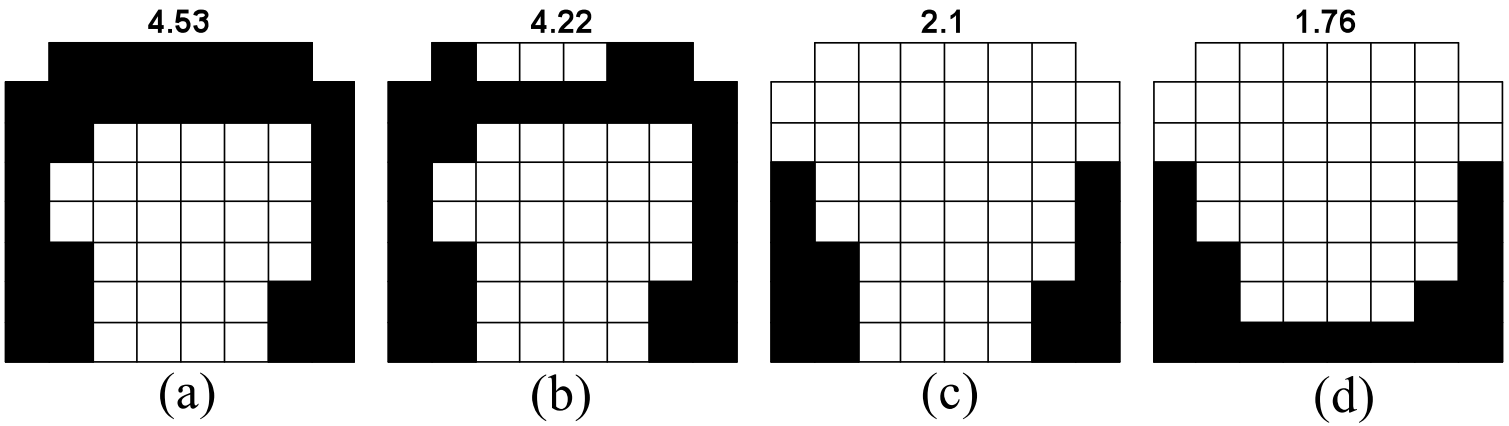
V = /a ɛ ɪ i ɔ ʊ/

F1, F2: F1 und F2-Werte zum Vokaloffset

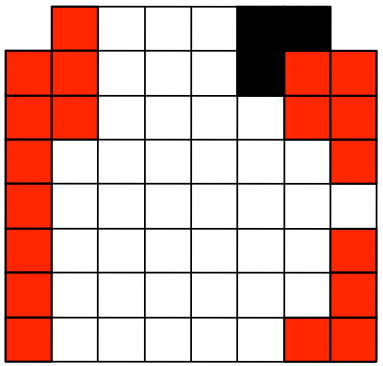
Zwei EPG-Parameter zum selben Zeitpunkt…

Die EPG Parameter

COG: Centre of gravity (Gewichtsschwerpunkt) Werte (ein Wert pro Vokal) elektropalatographische Daten.



SUM1278



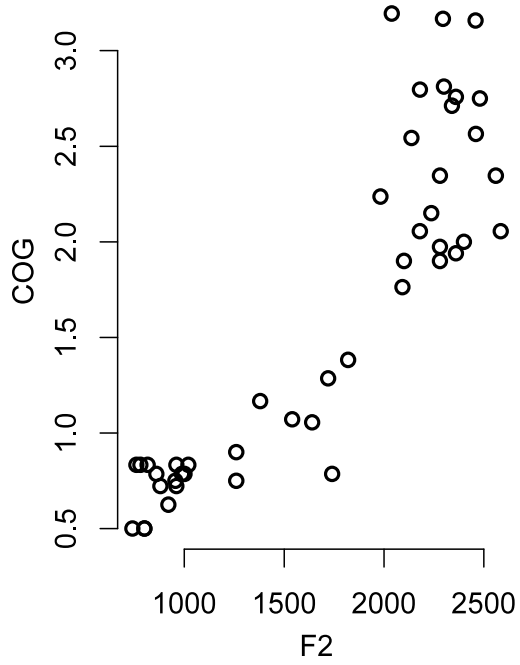
Kontaktsummen, Spalten 1+2+7+8

19

1. Kovarianz

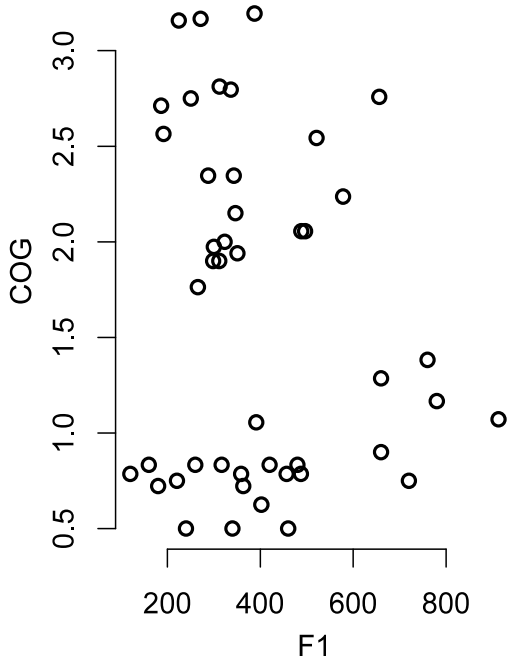
Je höher die Kovarianz, umso deutlicher die lineare Beziehung zwischen den Variablen

hoch und +ve



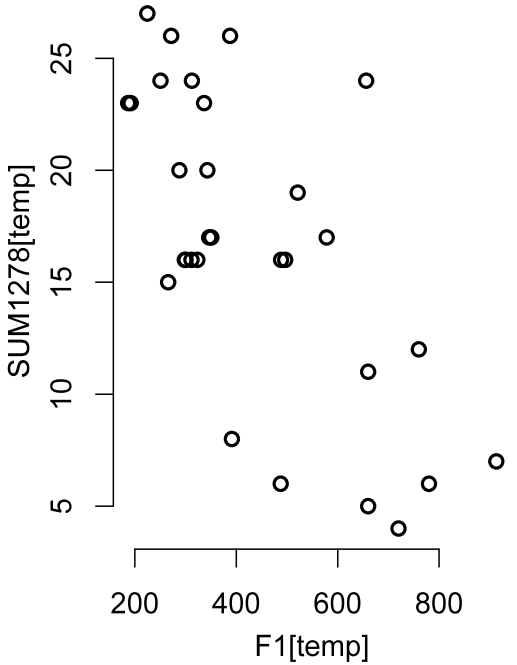
509.6908

nah an 0



-24.26598

mittel und -ve



-289.516

Berechung der Kovarianz

Produkt-Summe der Abweichungen vom Mittelwert

```
y = with(epg, F2)
x = with(epg, COG)
n = length(y)
```

Mittelwert

```
mx = mean(x)
```

```
my = mean(y)
```

Abweichungen vom Mittelwert

```
dx = x - mean(x)
```

```
dy = y - mean(y)
```

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

```
covxy = sum(dx*dy)/(n-1)
```

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

Einige Merkmale der Kovarianz

$cov(x, y)$ gleicht $cov(y, x)$

$cov(x,x)$ gleicht $var(x)$

$var(x+y)$ gleicht $var(x)+var(y) + 2 * cov(x,y)$

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

$var(x+y)$ gleicht $var(x) + var(y)$

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
```

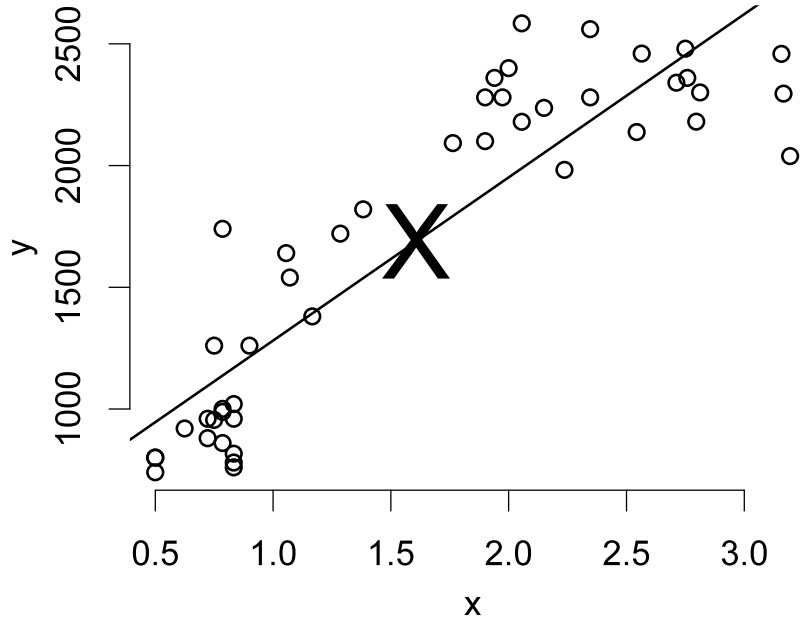
```
cor(xgross,y)
```

```
[1] 0.8917474
```

3. Regression

y-auf-x Regression: *y* soll durch *x* modelliert werden, also durch die Werte von *x* eingeschätzt werden.

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



Diese Regressionslinie durchschneidet (mx, my) den Mittelwert (X) der Verteilung

Die Regressionslinie: $\hat{y} = bx + k$

b ist die Die Neigung

b = r * sd(y)/sd(x) oder b = cov(x,y)/var(x)

k ist das y-Achsenabschnitt

k = my - b*mx

$\hat{y}$ die eingeschätzten Werte, die auf der R-Linie liegen

yhut = b*x + k

Abbildung

plot(x,y)

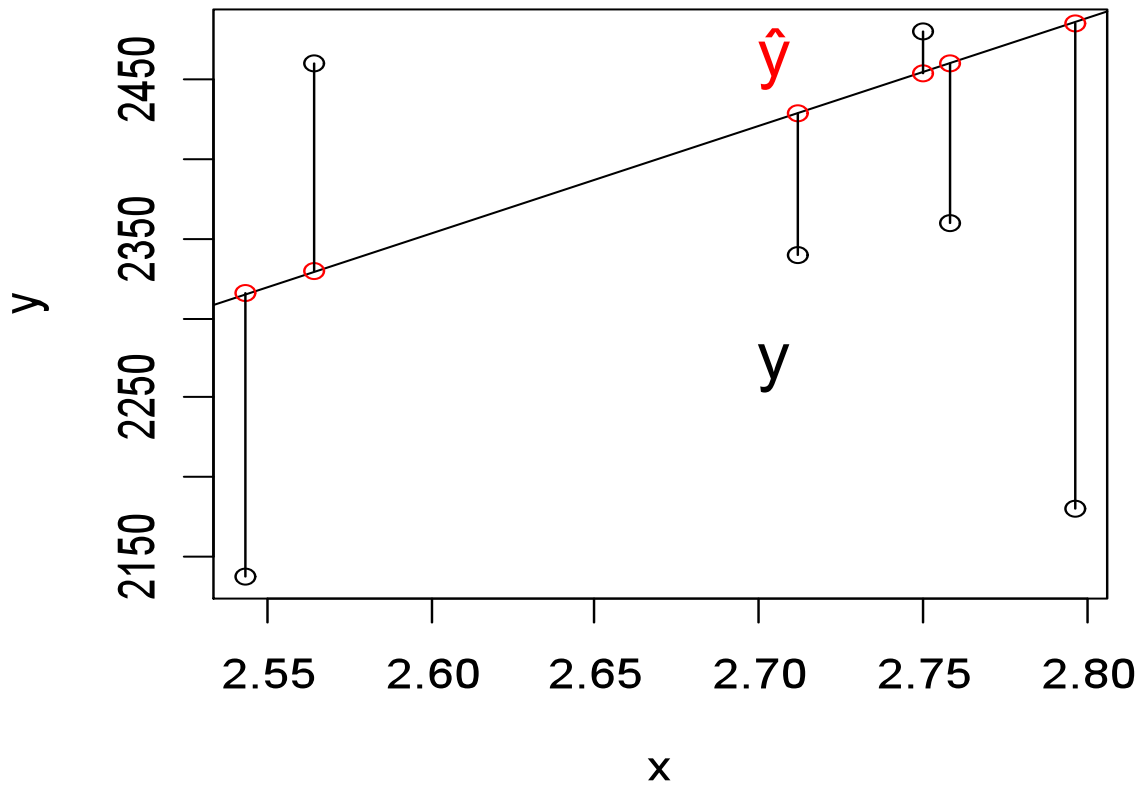
Regressionslinie überlagern

abline(k, b)

Regression und residuals

Der **residual** oder **error** ist der Unterschied zwischen den tatsächlichen und eingeschätzten Werten.

error = y - yhut



Regression, residuals, SSE

SSE = sum-of-the-squares of the error*

SSE = sum((y - yhut)^2)

oder

error = (y – yhut)

SSE = sum(error^2)

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

*wird auch manchmal RSS residual sum of squares genannt

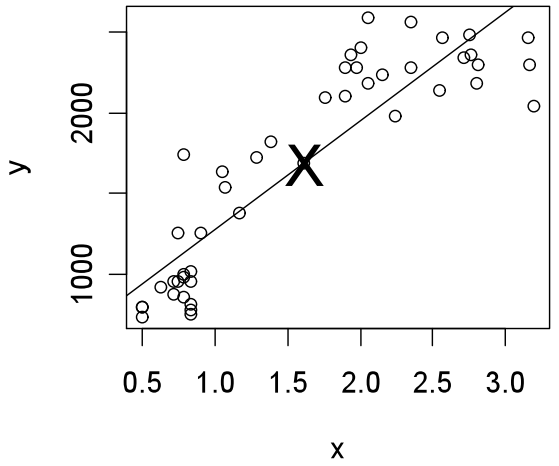
Die lm() Funktion

```
reg = lm(y ~ x)
```

~ wird modelliert durch

Regressionslinie überlagern

```
plot(x,y)
abline(reg)
```



Regressionskoeffiziente

```
coef(reg)
```

```
(Intercept)        x
  610.6845   670.2670
```

Eingeschätzte Werte

```
yhut = predict(reg)
```

```
yhut = b*x + k
```

Residuals

```
residuals(reg)
```

```
error = y - yhut
```

SSE

```
deviance(reg)
```

```
sum(error^2)
```

Regression: drei sehr wichtige Quantitäten

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

SSE = sum(error^2)

oder SSE = deviance(reg)

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

SSY = sum((y - my)^2)

3. **SSR**: sum of the squared-deviations in $\hat{y}$

SSR = sum((yhut - my)^2)

SSY = SSR + SSE

SSR + SSE gleicht SSY

R-squared

SSY = SSR + SSE

Je besser die Werte durch die Regressionlinie modelliert werden (also je geringer der Abstand zwischen y und ŷ) 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 Regressionlinie 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?

Signifikanz-Test

H0: $r = 0$

H1: r weicht signifikant ab von 0 (bedeutet: x und y sind miteineander mit einer hohen Wahrscheinlichkeit korreliert).

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

tstat = r/rsb

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

```
rsb = 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 = 2.220446×10^{-16}

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

Signifikanz-Test

Zwei wichtige Funktionen: summary(), anova()

```
reg = lm(y ~ x)

summary(reg)
anova(reg)
```

summary(reg)

2 * (1 - pt(tstat, n-2))
oder 1 - pf(fstat, 1, n-2)

sqrt(deviance(reg)/(n-2))

zB min(residuals(reg))

Call:
lm(formula = y ~ x)
Residuals:

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

Coefficients:

tstat

	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**

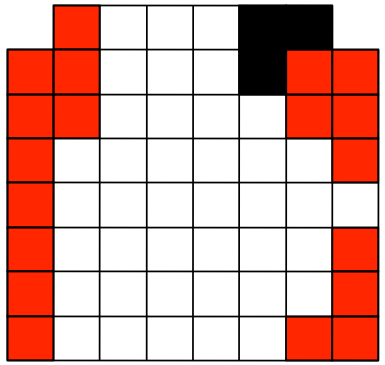
fstat

SSR/SSY oder cor(x,y)^2

Es gibt eine lineare Assoziation zwischen x und y, R^2 = 0.80, F[1, 43] = 167, p < 0.001.

Was sind die Erwartungen bezüglich der Beziehung zwischen F1 im Vokal und SUM1278?

SUM1278



Kontaktsummen, Spalten 1+2+7+8

19

y = F1; x = SUM1278