

Kovarianz, Korrelation, (lineare) Regression

Jonathan Harrington

Kovarianz, Korrelation, (lineare) Regression

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

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

```
head(epg)
```

Vk-Reihenfolgen von einem deutschen Muttersprachler.

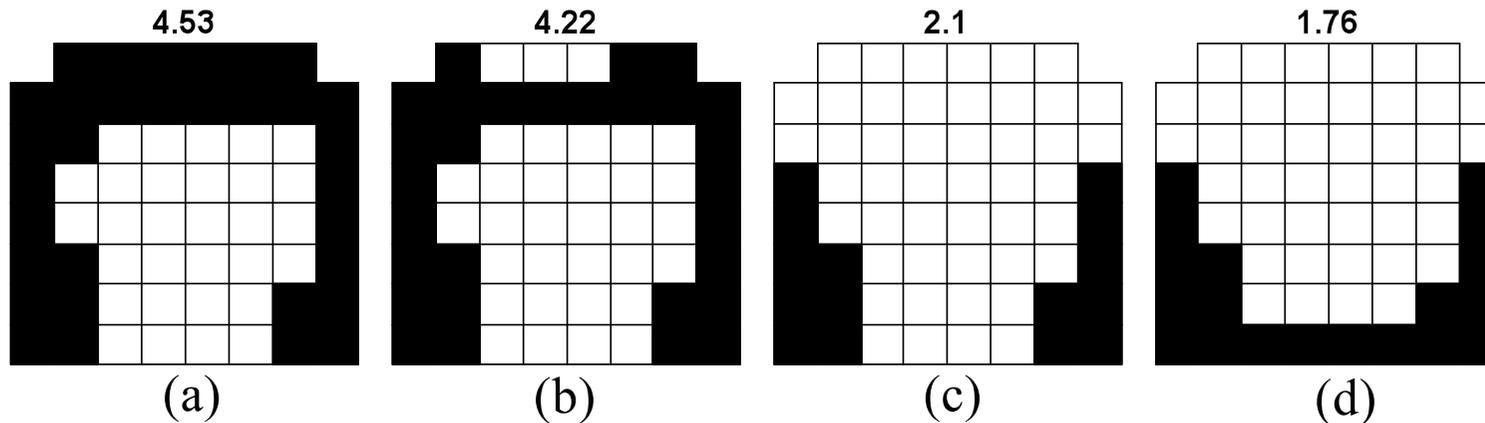
V = /a ε ɪ i ɔ ʊ/

F1, F2: F1 und F2-Werte zum Vokaloffset

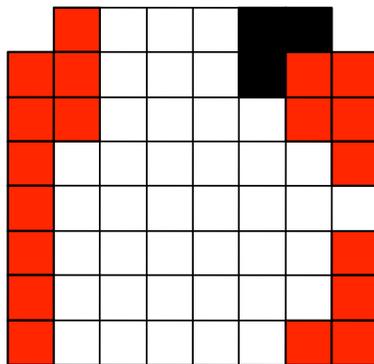
Zwei EPG-Parameter zum selben Zeitpunkt...

Die EPG Parameter

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



SUM1278

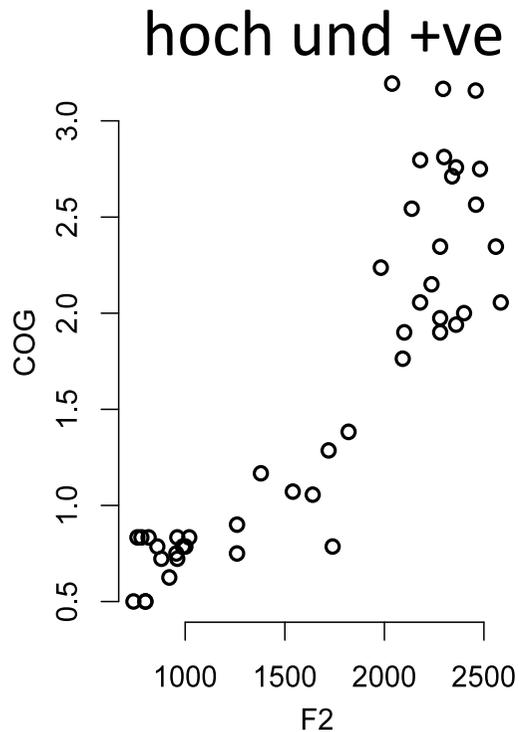


Kontaktsummen, Spalten 1+2+7+8

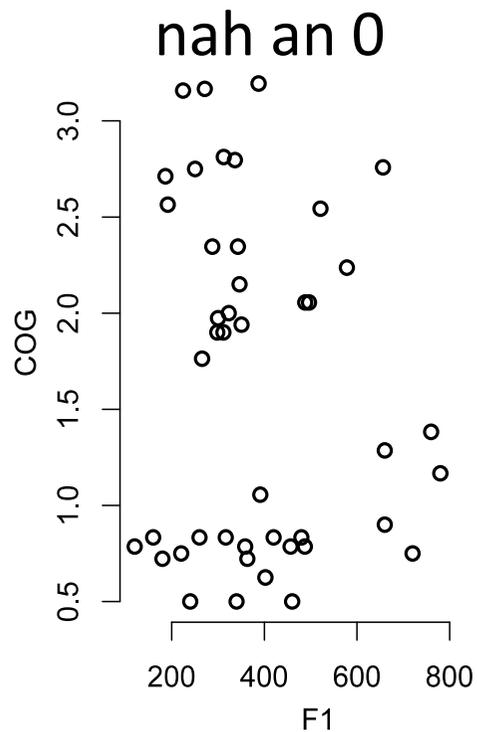
19

1. Kovarianz

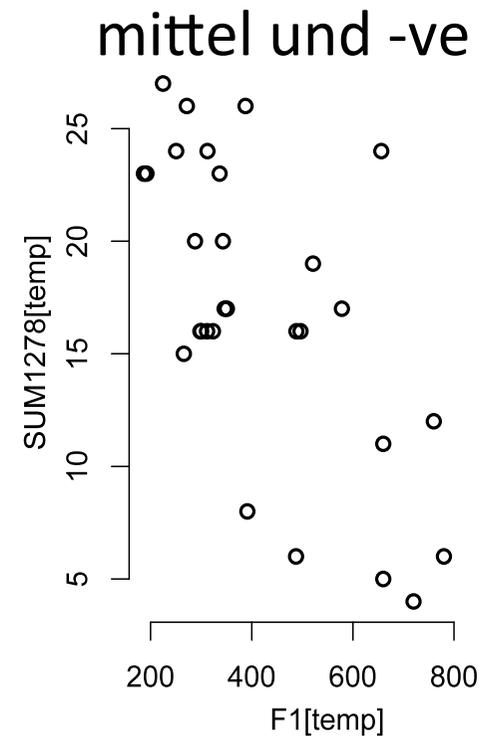
Je höher die Kovarianz, umso deutlicher die lineare Beziehung zwischen den Variablen



509.6908



-24.26598



-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

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

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

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

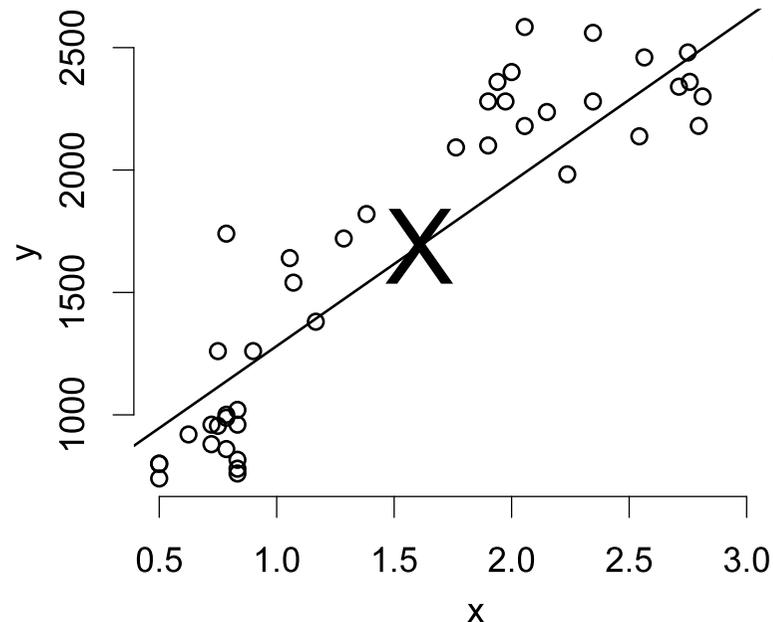
`cor(xgross,y)`

[1] 0.8917474

3. Regression

y-auf-x Regression: y 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 (m_x, m_y) den Mittelwert (\bar{X}) der Verteilung

Die Regressionslinie:

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

b ist die Die Neigung

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

k ist das y-Achsenabschnitt

$$k = my - b * mx$$

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

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

Abbildung

`plot(x,y)`

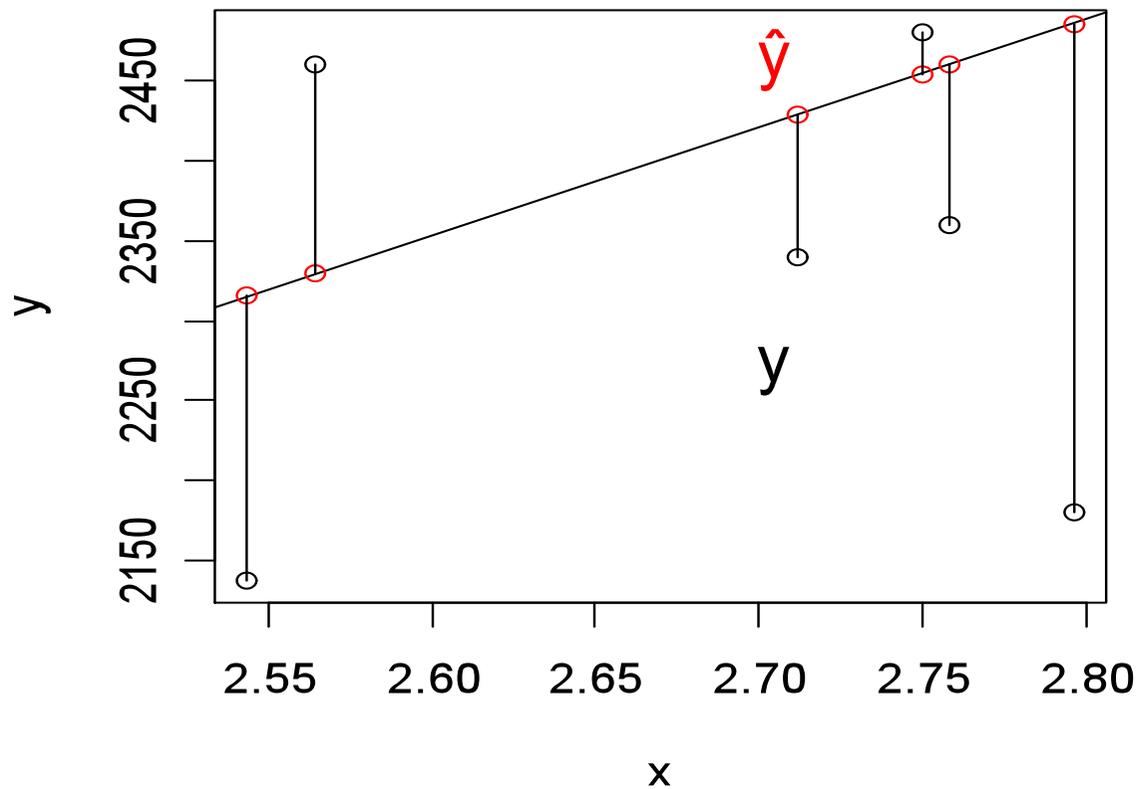
Regressionslinie überlagern

`abline(k, b)`

Regression und residuals

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

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



Regression, residuals, SSE

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

$$SSE = \sum((y - \hat{y})^2)$$

oder

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

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

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

<code>coef(reg)</code>	(Intercept)	x
	610.6845	670.2670

Eingeschätzte Werte

`yhut = predict(reg)`

$$yhut = b * x + k$$

Residuals

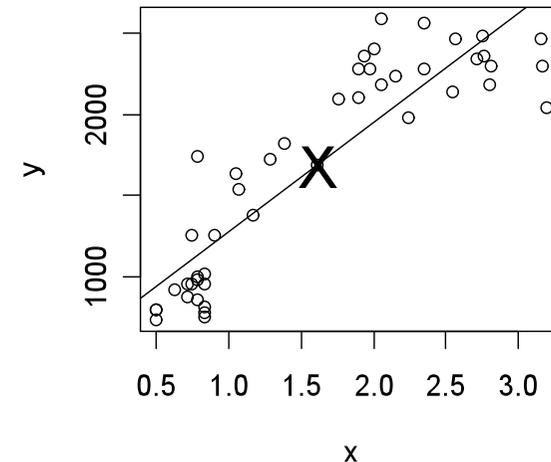
`residuals(reg)`

$$error = y - yhut$$

SSE

`deviance(reg)`

$$sum(error^2)$$



Regression: drei sehr wichtige Quantitäten

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

$$SSE = \text{sum}(\text{error}^2) \quad \text{oder } SSE = \text{deviance}(\text{reg})$$

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 in \hat{y}

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

$$SSY = SSR + SSE$$

SSR + SSE

gleich

SSY

R-squared

$$SSY = SSR + SSE$$

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, $p < 0.001$).

Signifikanz-Test

Zwei wichtige Funktionen: `summary()`, `anova()`

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

```
summary(reg)
```

```
anova(reg)
```

summary(reg)

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

$\sqrt{\text{deviance}(\text{reg})/(n-2)}$

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**

zB $\min(\text{residuals}(\text{reg}))$

tstat

fstat

SSR/SSY oder $\text{cor}(x,y)^2$

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

Auflistung von SSR, SSE, MSR, MSE

`anova(reg)`

MSR = mean-sum-of-squares due to regression

MSE = mean-sum-of squares of the error

$\sqrt{\text{MSE}}$ = residual standard error (vorige Seite)

$f_{\text{stat}} = \text{MSR}/\text{MSE}$

$\text{MSR} = \text{SSR}/1$

$2 * (1 - \text{pt}(t_{\text{stat}}, n-2))$

oder

$1 - \text{pf}(f_{\text{stat}}, 1, n-2)$

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	15031672	15031672	166.97	< 2.2e-16 ***
Residuals	43	3871019	90024		

SSR

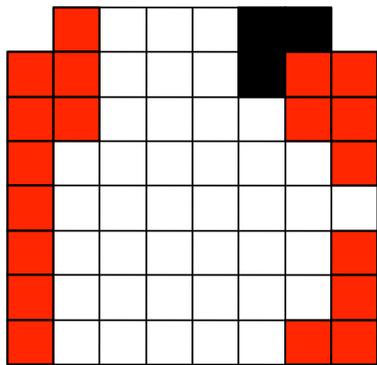
SSE

$\text{MSE} = \text{SSE}/(n-2)$

$\text{MSE} = \text{deviance}(\text{reg})/(n-2)$

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

SUM1278



Kontaktsummen, Spalten 1+2+7+8

19

$$y = F1; x = \text{SUM1278}$$