Statistik

Vorlesung 7 (Lineare Regression)

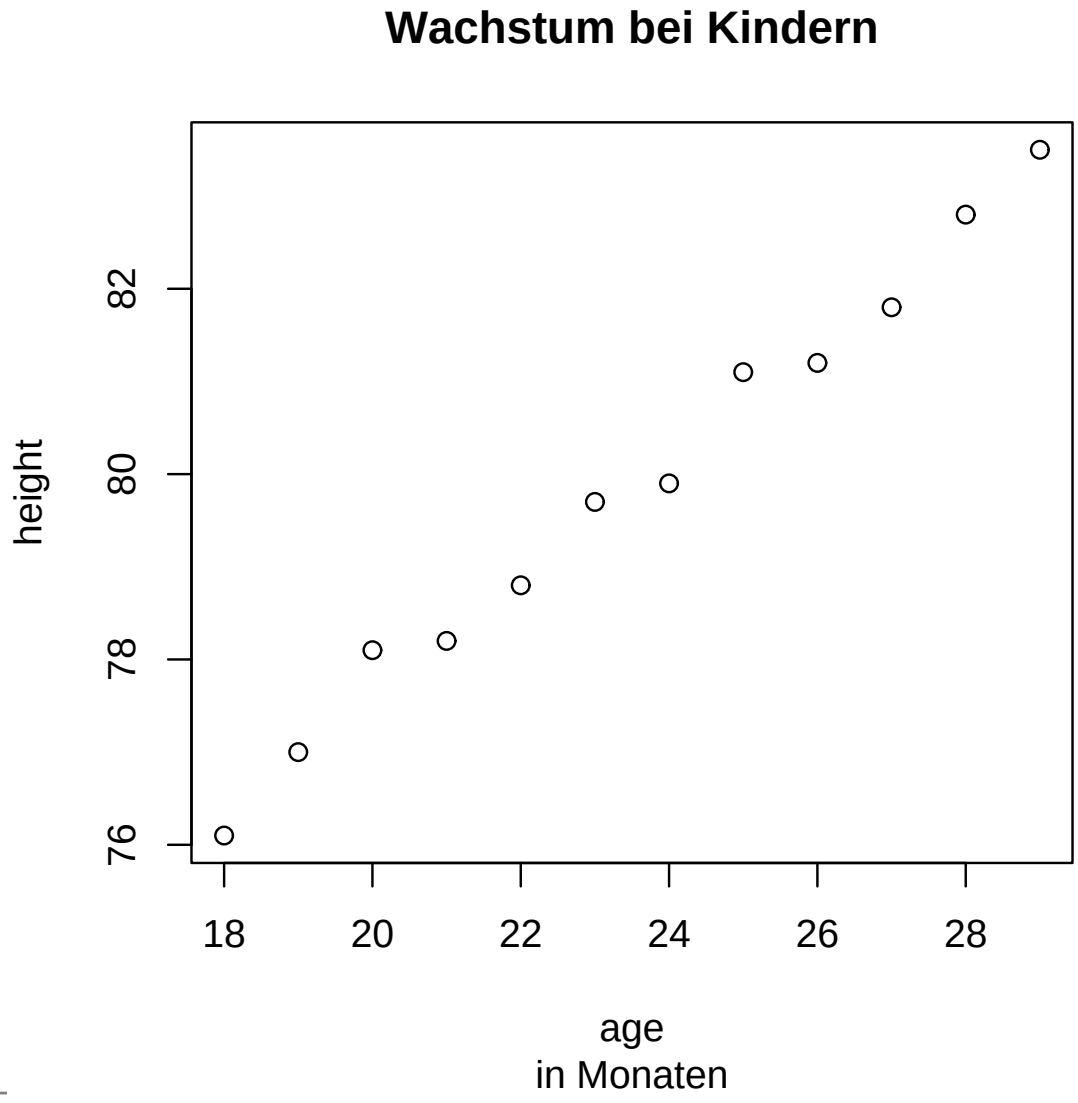
K.Gerald van den Boogaart
http://www.stat.boogaart.de/

```
> Brustkrebs <- read.table("BreastCancerData.txt",
+     sep = "\t", header = T)
> Chroma <- read.table("ChromatographyData.txt",
+     sep = "\t", header = T)
> Wachstum <- read.table("AgeandheightData.txt",
+     sep = "\t", header = T)
> Rauchen <- read.table("SmokingandCancerData.txt",
+     sep = "\t", header = T)
> data(anscombe)
> data(airquality)
> showGeraden <- function(form = Mortality ~ Smoking,
+     data = Rauchen, alpha = 0, outer = F, n = 0,
+     ..., pt = F, diag = F, g0 = F) {
+     par(pch = 20)
+     if (diag)
+         par(mfrow = c(2, 2))
+     plot(form, data = data, col = 2, pch = 20,
+         ...)
+     abline(m1 <- lm(form, data = data, x = T,
+         y = T), col = 2)
+     m <- m1
+     if (n > 0) {
+         s <- sample(1:NROW(data), n)
+         points(form, data = data[s, ], col = 3,
+             pch = 20)
+         abline(m <- lm(form, data = data[s, ],
+             x = T, y = T), col = 3)
+     }
+     a <- coef(m)[1]
+     b <- coef(m)[2]
+     if (alpha > 0) {
+         xr <- range(m1$x[, 2])
+         x <- seq(xr[1], xr[2], length.out = 200)
+         a <- coef(m)[1]
+         b <- coef(m)[2]
+         v <- vcov(m)
+         y <- a + b * x
+         sqy <- v[1, 1] + 2 * v[2, 1] * x + v[2,
+             2] * x * x + ifelse(outer, summary(m)$sigma^2,
```

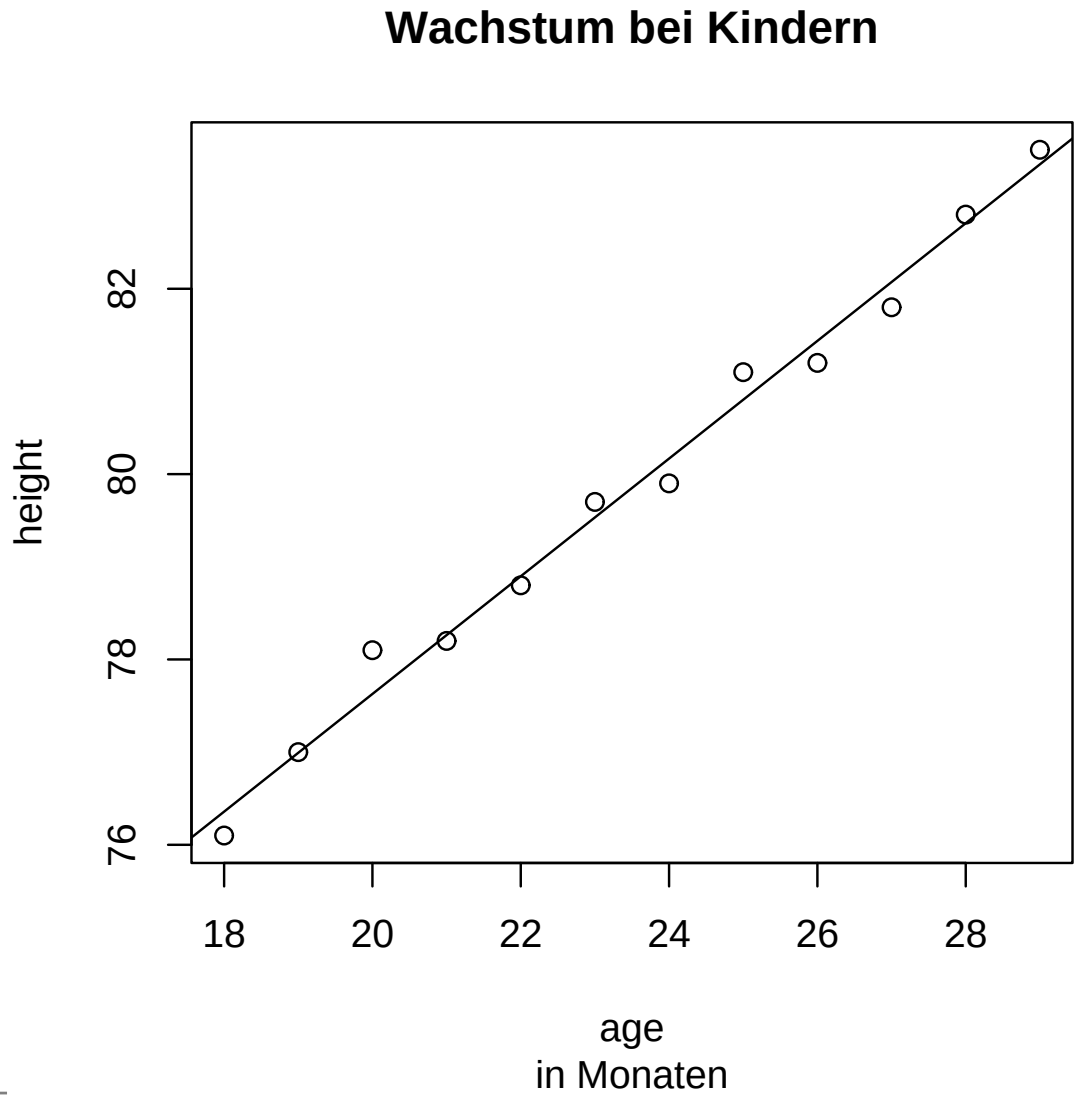
```
+             0)
+         tq <- qt(1 - alpha/2, df = m$df.resid)
+         lines(x, y + sqrt(sqy) * tq)
+         lines(x, y - sqrt(sqy) * tq)
+         if (outer)
+             title(main = "Konfidenzbereich fuer die Punkte")
+         else {
+             title(main = "Konfidenzbereich fuer die Gerade")
+             if (g0)
+                 abline(h = mean(m$y))
+         }
+     }
+     title(sub = call("==", form[[2]], call("+",
+         round(a, 3), call("*", round(b, 3), form[[3]]))))
+     if (pt) {
+         points(m$x[, 2], mean(m$y) + resid(m),
+             pch = 20)
+         abline(h = mean(m$y))
+     }
+     if (diag) {
+         plot(predict(m), resid(m))
+         plot(predict(m), influence(m)$hat, ylab = "Hebelwirkung",
+             ylim = c(0, 1))
+         plot(predict(m), cooks.distance(m), ylab = "Cook distance",
+             ylim = c(0, 1))
+     }
+     invisible(m)
+ }
> stepblock <- function(cc) {
+     pon <- T
+     cc <- as.list(cc[[2]][-1])
+     for (my.expression.A in cc) {
+         cat("> ")
+         cat(deparse(my.expression.A), sep = "\n+ ")
+         eval.parent(my.expression.A)
+         cat("\n")
+     }
+ }
> options(width = 60)
```

Konzepte der linearen Regression

Wie wachsen Kleinkinder?



Gerade als Vereinfachung



In Formeln

Das Modell

$$Y_i = a + bX_i + \epsilon_i$$

In Formeln

Das Modell

$$Y_i = a + bX_i + \epsilon_i$$

a, b sind unbekannt.

In Formeln

Das Modell

$$Y_i = a + bX_i + \epsilon_i$$

a, b sind unbekannt.
Der Computer schätzt die Werte als:

$$\hat{b} = \frac{\hat{\mathrm{cov}}(X, Y)}{\hat{\mathrm{var}}(X)}$$

$$\hat{a} = \bar{Y} - \bar{X}\hat{b}$$

Computerausgabe

```
> model <- lm(height ~ age, data = Wachstum)
> model

Call:
lm(formula = height ~ age, data = Wachstum)

Coefficients:
(Intercept)          age
     64.928        0.635
```

Definitionen

Das Modell

$$Y_i = a + bX_i + \epsilon_i$$

- a heißt Achsenabschnitt
  weil $a = a + b0$ der Wert der Geraden bei $X = 0$ ist.

Definitionen

Das Modell

$$Y_i = a + bX_i + \epsilon_i$$

- a heißt Achsenabschnitt
- b heißt Steigung
  weil b die Steigung der Geraden $a + bX$ ist.

Definitionen

Das Modell

$$Y_i = a + bX_i + \epsilon_i$$

- a heißt Achsenabschnitt
- b heißt Steigung
- Die X_i heißen Regressor
  weil X die das ansteigen bewirkt.

Definitionen

Das Modell

$$Y_i = a + bX_i + \epsilon_i$$

- a heißt Achsenabschnitt
- b heißt Steigung
- Die X_i heißen Regressor
- Die Y_i heißen Regressant
  weil Y erhöht wird.

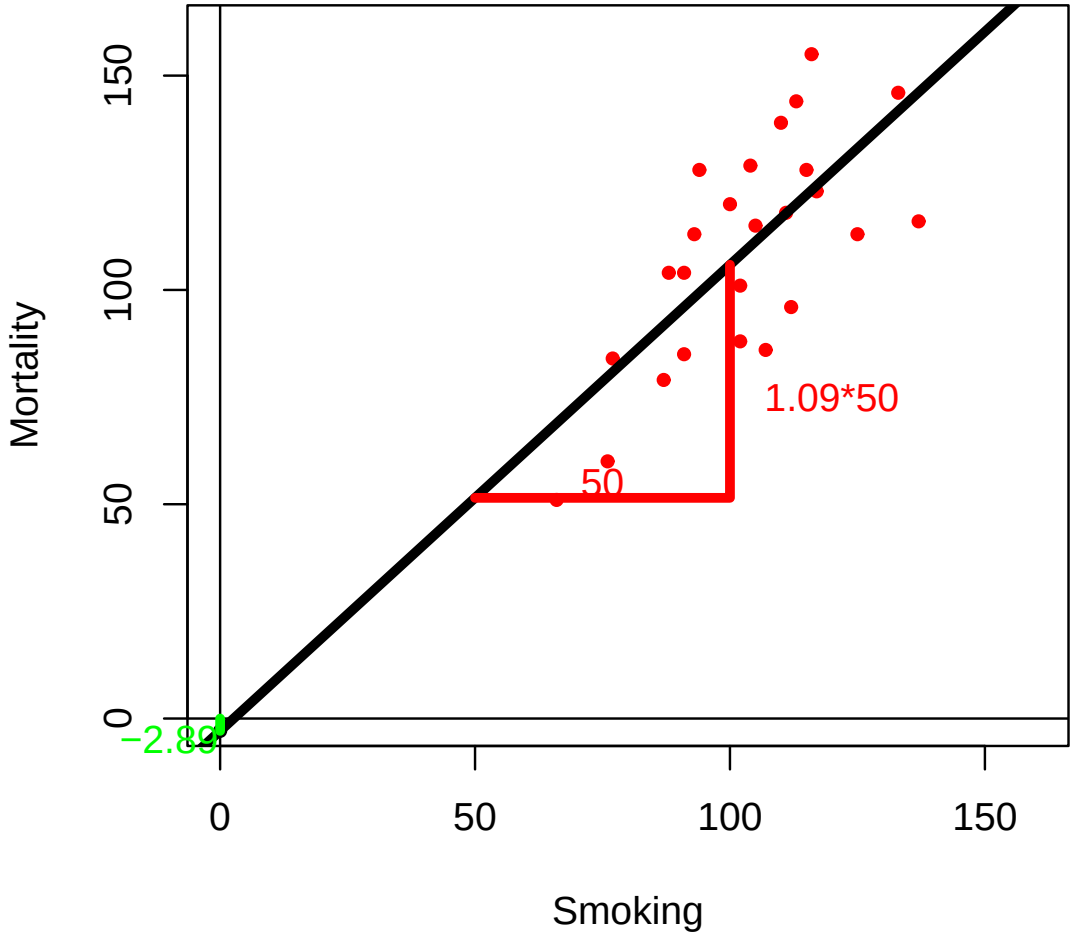
Definitionen

Das Modell

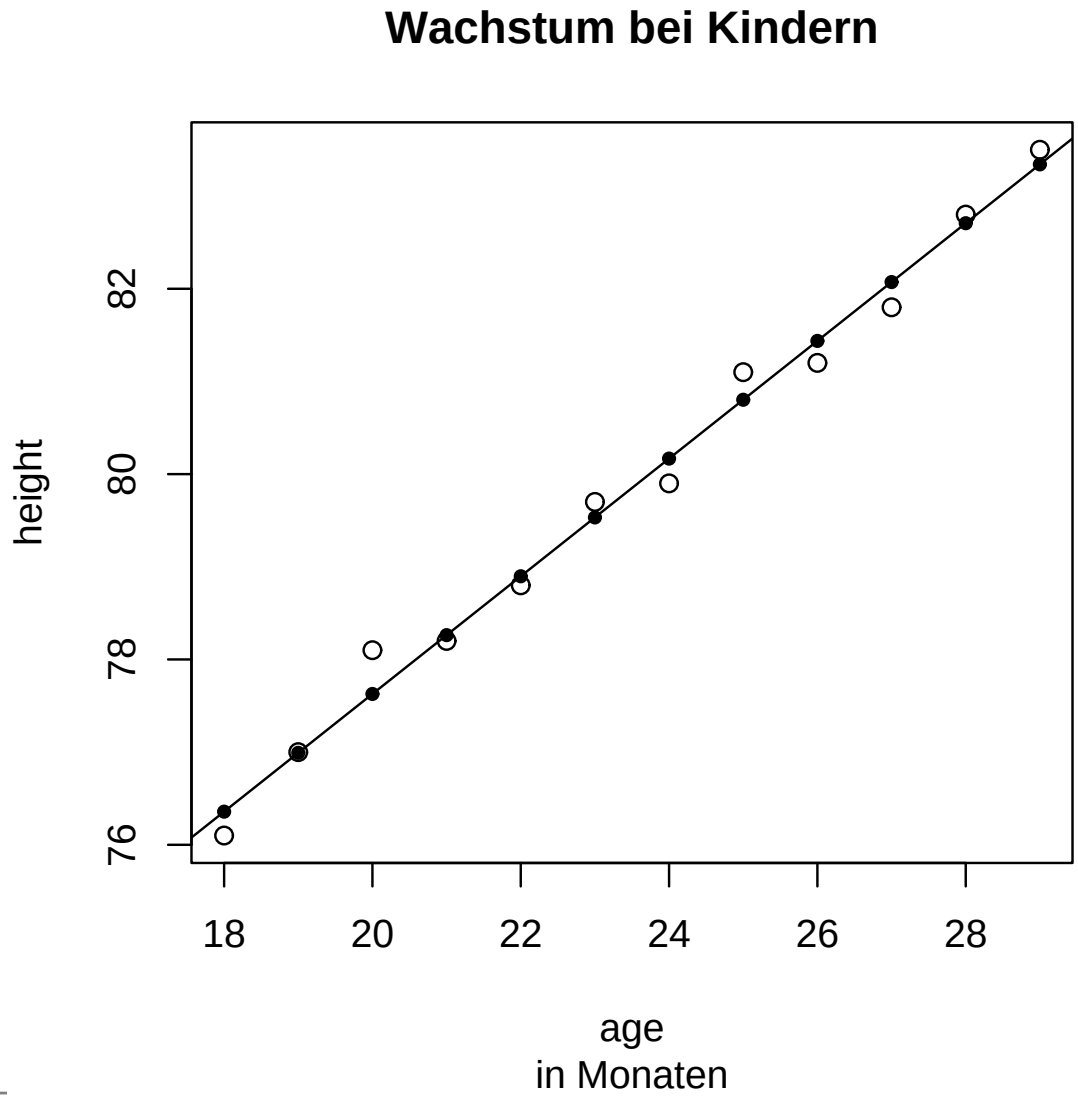
$$Y_i = a + bX_i + \epsilon_i$$

- a heißt Achsenabschnitt
- b heißt Steigung
- Die X_i heißen Regressor
- Die Y_i heißen Regressant
- Die ϵ_i heißen Fehlerterm.
  weil ϵ den Fehler der “Vereinfachung” Gerade beinhaltet.

Achsenabschnitt und Steigung



Vorhersagewerte



Die Vorhersagewerte

Das Regressionsmodel:

$$Y_i = a + bX_i + \epsilon_i$$

Wenn wir von den Unzulänglichkeiten

- Das a und b nur geschätzt werden können.
- Die Gerade nur bis auf den Fehler ϵ stimmt.

Die Vorhersagewerte

Das Regressionsmodel:

$$Y_i = a + bX_i + \epsilon_i$$

Wenn wir von den Unzulänglichkeiten

- Das a und b nur geschätzt werden können.
- Die Gerade nur bis auf den Fehler ϵ stimmt.

absehen würden wir also annehmen, dass für $X = x$ dann

$$y = a + bx$$

sein müßte.

Die Vorhersagewerte

Das Regressionsmodel:

$$Y_i = a + bX_i + \epsilon_i$$

Wenn wir von den Unzulänglichkeiten

- Das a und b nur geschätzt werden können.
- Die Gerade nur bis auf den Fehler ϵ stimmt.

absehen würden wir also annehmen, dass für $X = x$ dann

$$y = a + bx$$

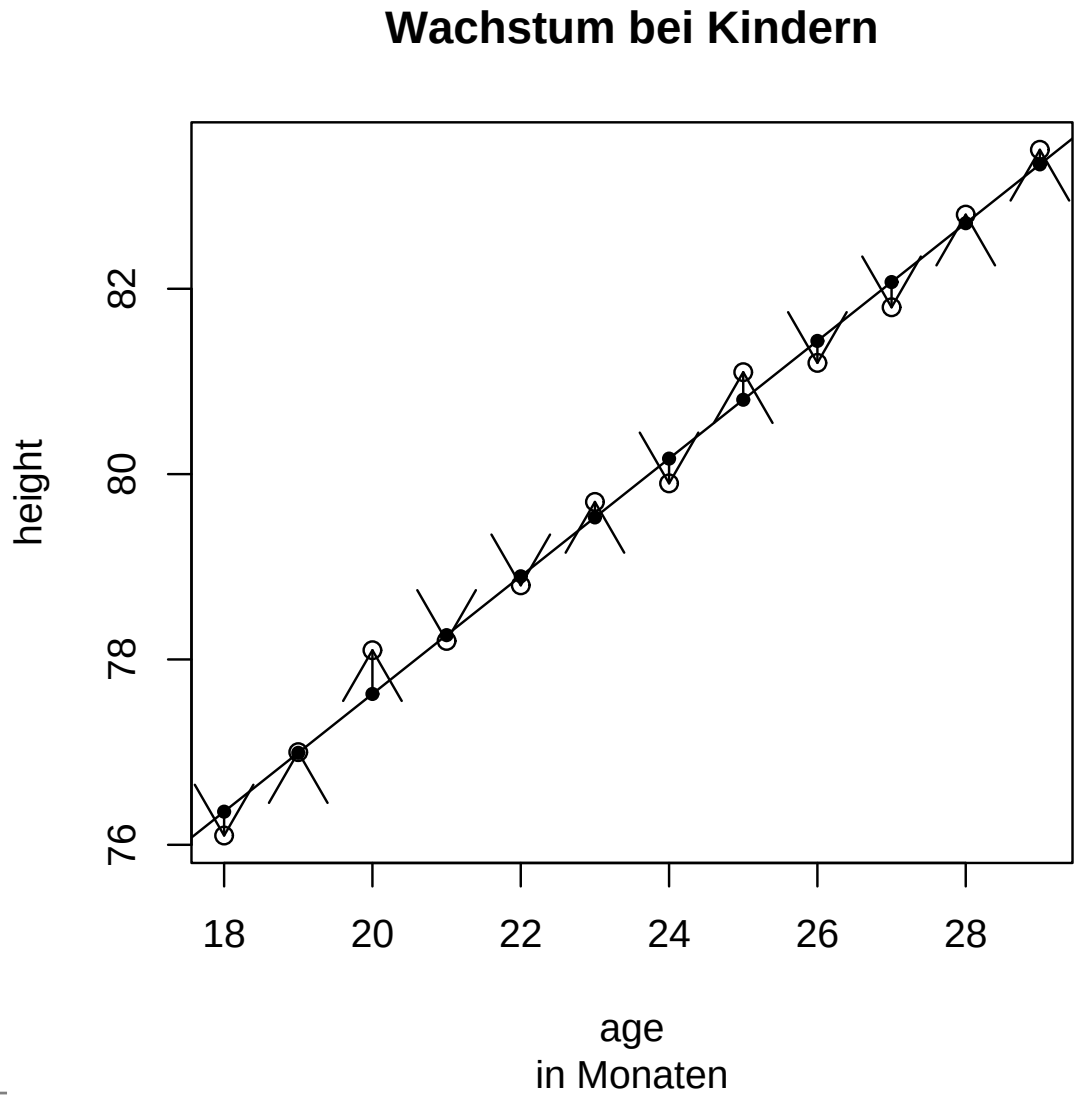
sein müßte.
Diesen Wert bezeichnen wir als die **vom Modell vorhergesagten Werte** $\hat{Y}_i$:

$$\hat{Y}_i := \hat{a} + \hat{b}X_i$$

Computerausgabe

```
> predict(model)

       1        2        3        4        5        6
76.35769 76.99266 77.62762 78.26259 78.89755 79.53252
       7        8        9       10       11       12
80.16748 80.80245 81.43741 82.07238 82.70734 83.34231
```

Residuen


Wachstum bei Kindern
height
76
78
80
82
18
20
22
24
26
28
age
in Monaten

Schätzung und Residuen

Das Regressionsmodel:

$$Y_i = a + bX_i + \epsilon_i$$

Wir kennen aber nur die Schätzwerte $\hat{a}$ und $\hat{b}$ und erhalten eine neue Gleichung:

$$Y_i = \hat{a} + \hat{b}X_i + r_i$$

- $\hat{b} = \frac{\hat{\mathrm{cov}}(X,Y)}{\hat{\mathrm{var}}(X)}$ der Schätzwert für b.

Schätzung und Residuen

Das Regressionsmodel:

$$Y_i = a + bX_i + \epsilon_i$$

Wir kennen aber nur die Schätzwerte $\hat{a}$ und $\hat{b}$ und erhalten eine neue Gleichung:

$$Y_i = \hat{a} + \hat{b}X_i + r_i$$

- $\hat{b} = \frac{\hat{\text{cov}}(X,Y)}{\hat{\text{var}}(X)}$ der Schätzwert für b.
- $\hat{a} = \bar{Y} - \bar{X}\hat{b}$ ist der Schätzwert für b

Schätzung und Residuen

Das Regressionsmodel:

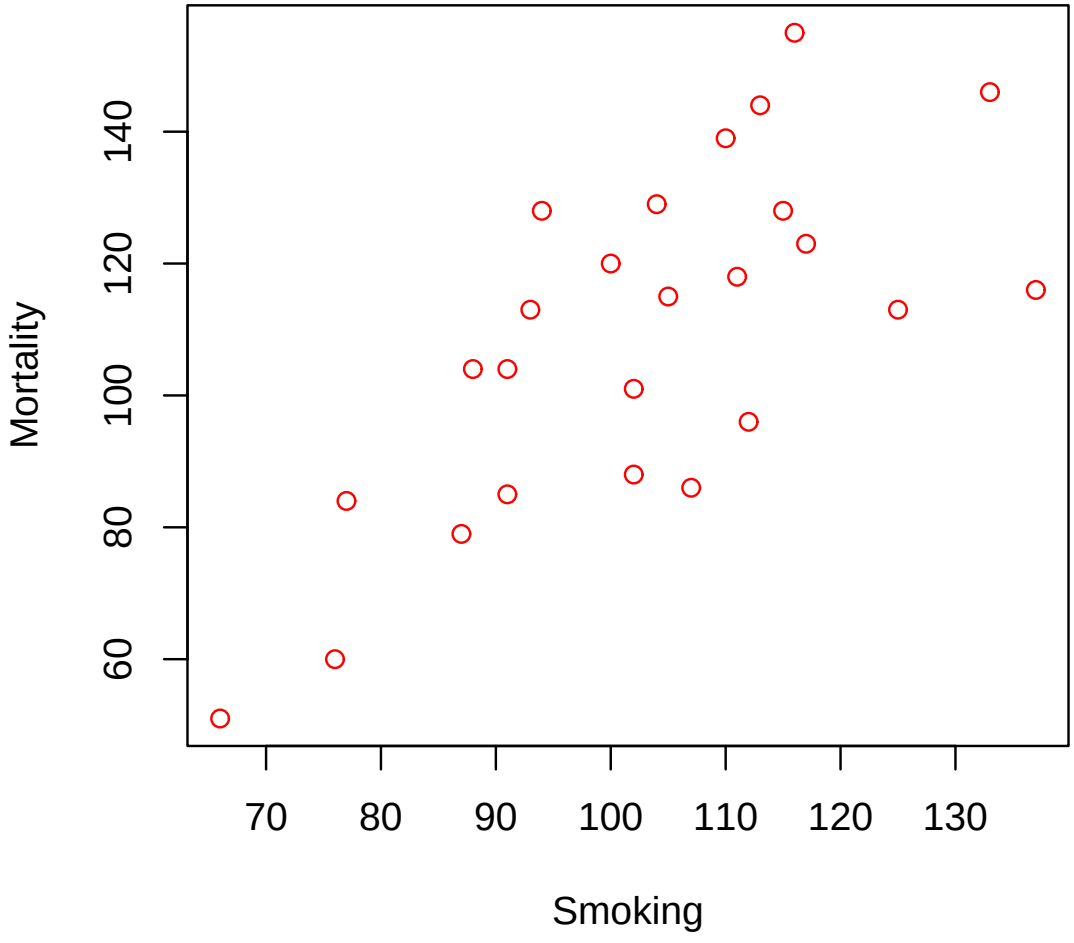
$$Y_i = a + bX_i + \epsilon_i$$

Wir kennen aber nur die Schätzwerte $\hat{a}$ und $\hat{b}$ und erhalten eine neue Gleichung:

$$Y_i = \hat{a} + \hat{b}X_i + r_i$$

- $\hat{b} = \frac{\hat{\mathrm{cov}}(X,Y)}{\hat{\mathrm{var}}(X)}$ der Schätzwert für b.
- $\hat{a} = \bar{Y} - \bar{X}\hat{b}$ ist der Schätzwert für b
- Die r_i heißen Residuen
  residuum: lat. für das Übriggebliebene

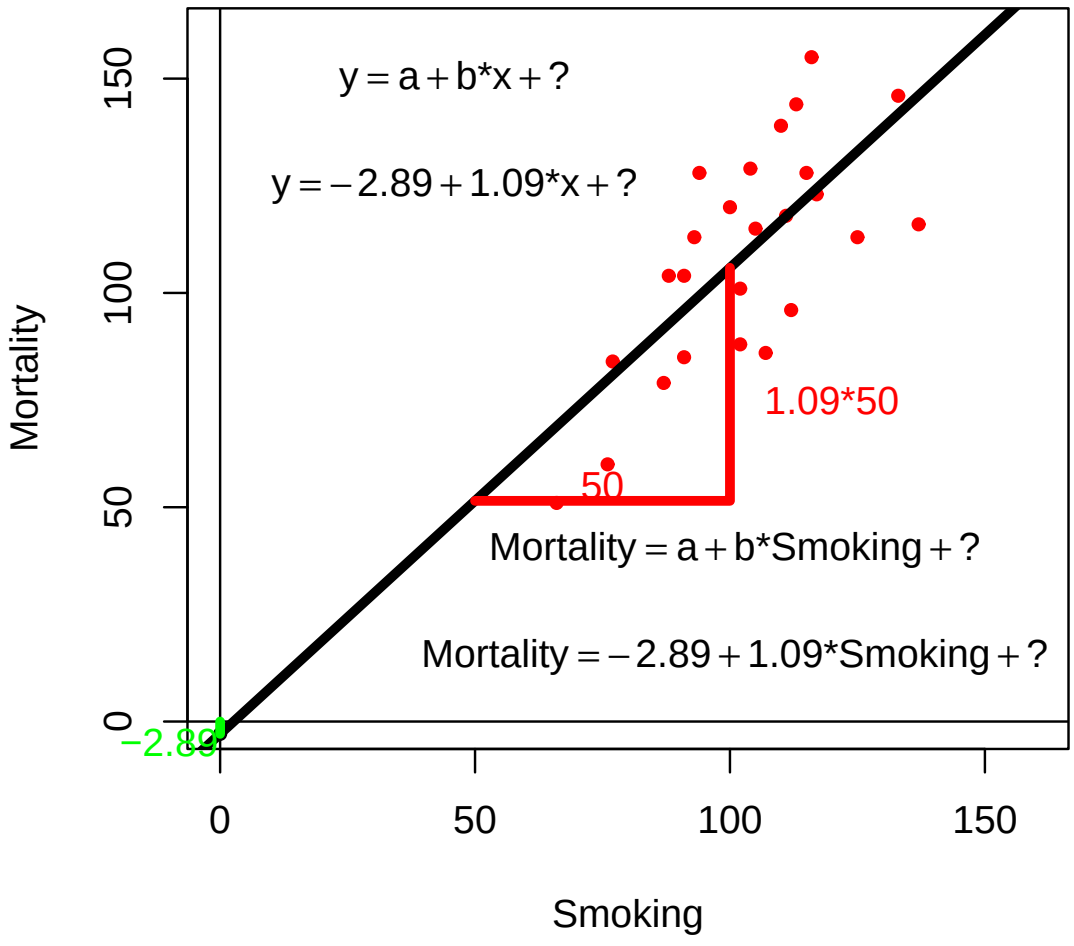
Beispiel II: Rauchen



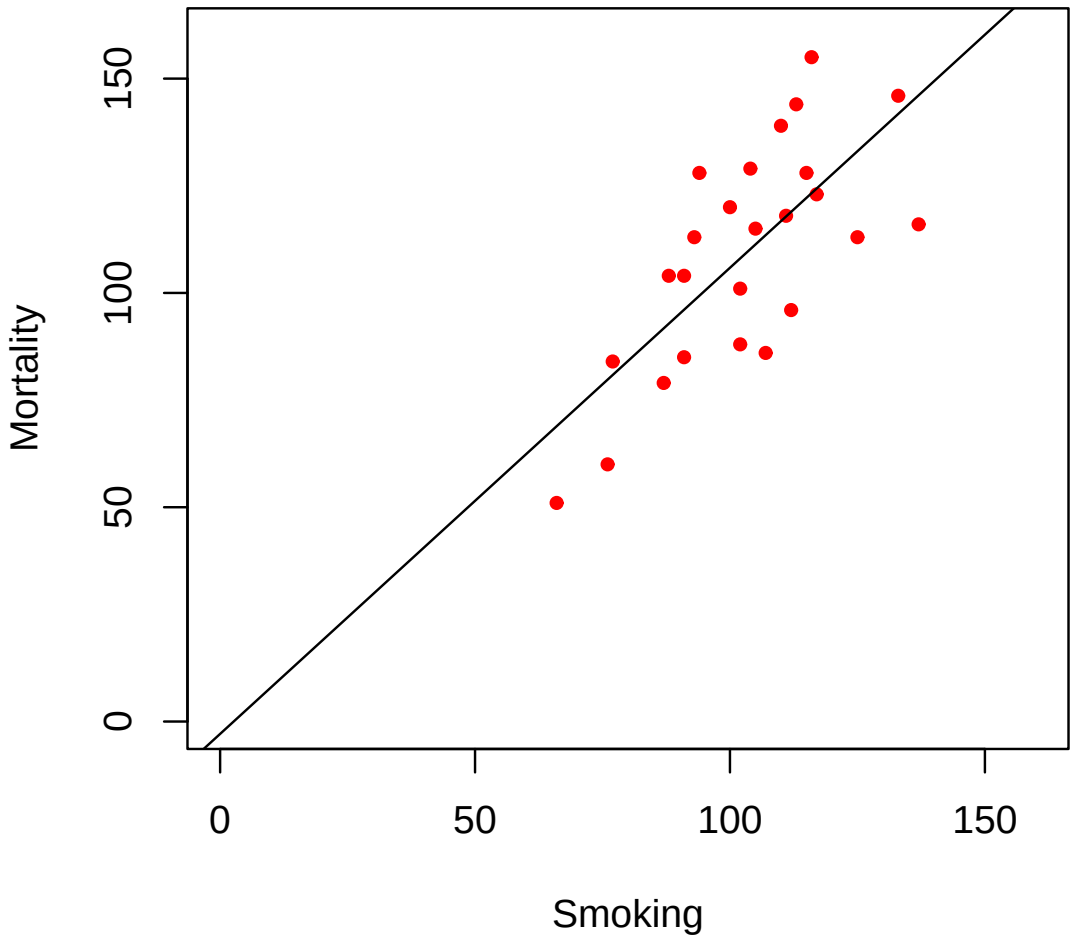
Computerausgabe

```
> model <- lm(Mortality ~ Smoking, data = Rauchen)
> model

Call:
lm(formula = Mortality ~ Smoking, data = Rauchen)

Coefficients:
(Intercept)      Smoking
     -2.885        1.088

> predict(model)

                Farmers, foresters, and fisherman
                                         80.85467
                        Miners and quarrymen
                                        146.10660
                  Gas, coke and chemical makers
                                        124.35596
```

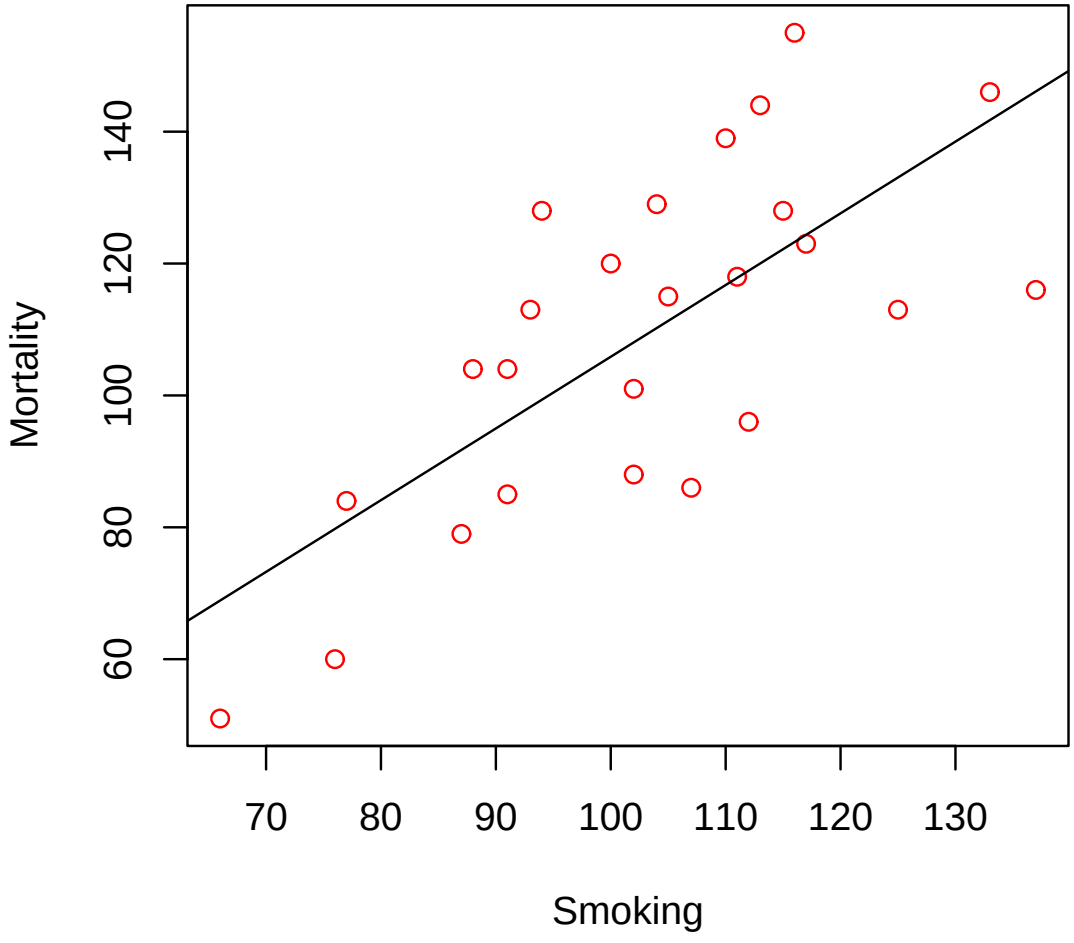
Graphische Interpretation



Normalansicht Schritt 1



Regressionsgerade



Modellvorstellung: Ungenaue Beschreibun

Das Regressionsmodel:

$$Y_i = a + bX_i + \epsilon_i$$

ist sich darüber bewußt, dass die Gerade die Wirklichkeit nicht genau beschreibt.

Modellvorstellung: Ungenaue Beschreibun

Das Regressionsmodel:

$$Y_i = a + bX_i + \epsilon_i$$

ist sich darüber bewußt, dass die Gerade die Wirklichkeit nicht genau beschreibt.
Man geht davon aus, dass die Abweichungen von der Gerade zufällig, unabhängig voneinander und im Mittel 0 sind.

Bestimmung der Geraden

- Die Regressionsgerade wird so bestimmt, dass

$$\sum_{i=1}^{n}(Y_i - (\hat{a} + \hat{b}X_i))^2$$

so klein wie möglich wird.

Bestimmung der Geraden

- Die Regressionsgerade wird so bestimmt, dass

$$\sum_{i=1}^{n}(Y_i - (\hat{a} + \hat{b}X_i))^2$$

  so klein wie möglich wird.
- Dieses Verfahren nennt man: Kleinste Quadrate

Bestimmung der Geraden

- Die Regressionsgerade wird so bestimmt, dass

$$\sum_{i=1}^{n}(Y_i - (\hat{a} + \hat{b}X_i))^2$$

  so klein wie möglich wird.
- Dieses Verfahren nennt man: Kleinste Quadrate
- Dieses Vefahren ist besonders gut, wenn gewisse Annahmen erfüllt sind.

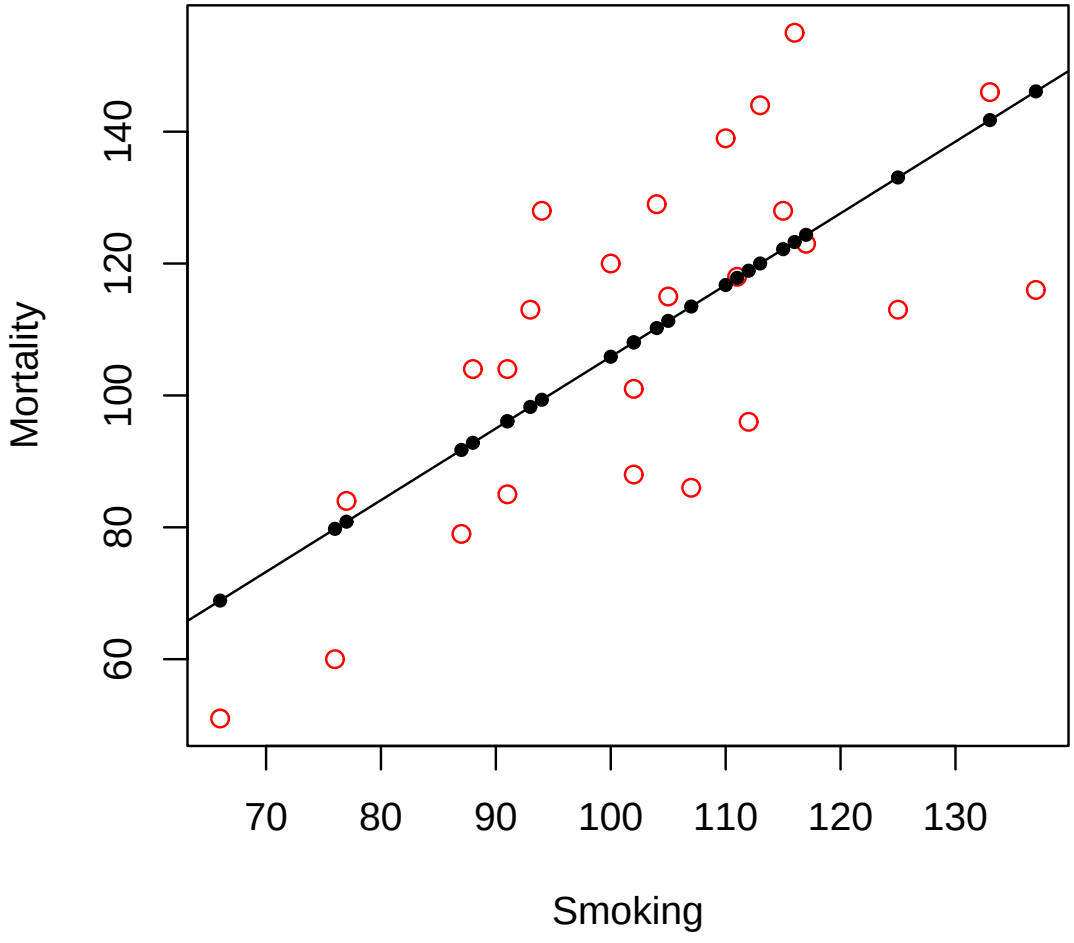
Bestimmung der Geraden

- Die Regressionsgerade wird so bestimmt, dass

$$\sum_{i=1}^{n}(Y_i - (\hat{a} + \hat{b}X_i))^2$$

  so klein wie möglich wird.
- Dieses Verfahren nennt man: Kleinste Quadrate
- Dieses Vefahren ist besonders gut, wenn gewisse Annahmen erfüllt sind.
- Annahmen: Die ϵ_i sind normalverteilt mit Erwartungswert 0 und einer unbekannten Varianz σ^2.

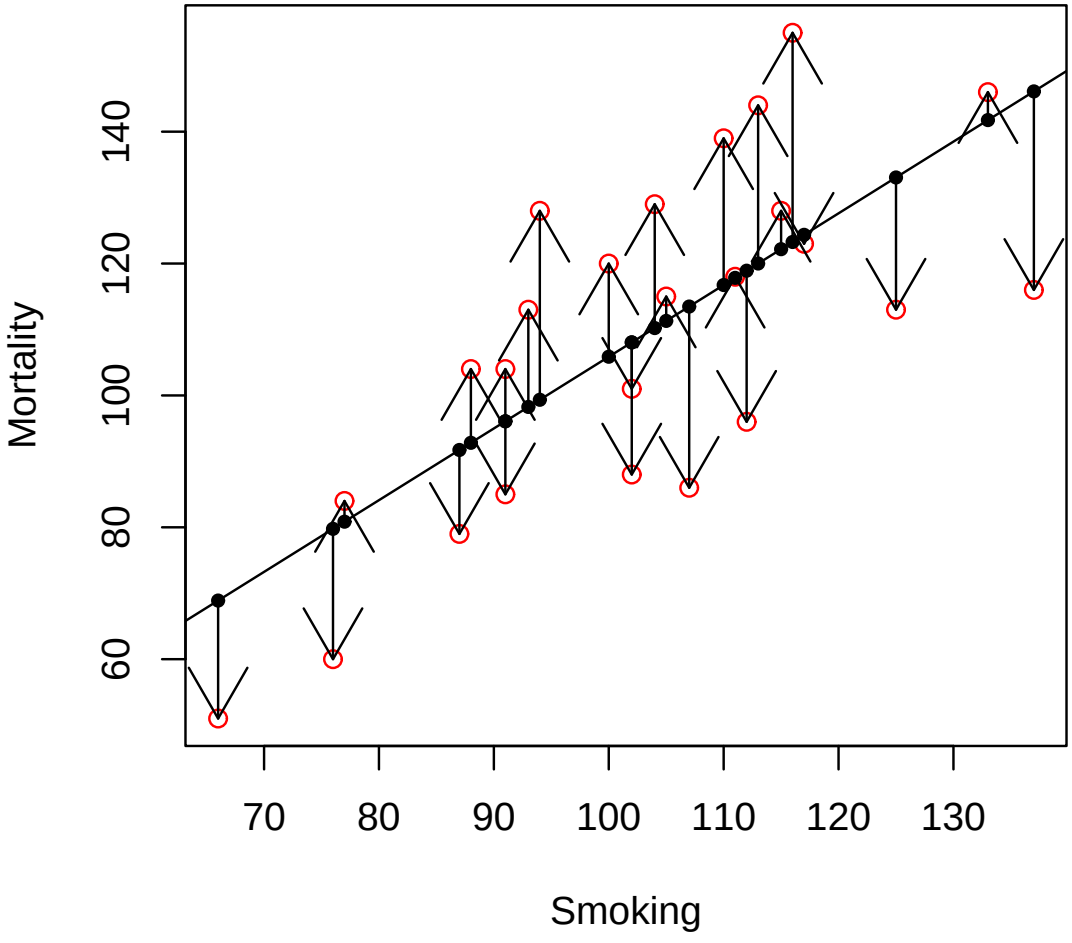
Vorhersagewerte



Statistische Vorhersagen

- Die Vorhersagewerte sind also keineswegs eine feststehende Wahrheit, sondern geben lediglich eine Tendenz an.
- Sie lassen Raum für andere Einflüsse.
- Im Mittel werden die Vorhersagewerte allerdings richtig sein.

Residuen



Vorhersagewerte und Residuen

Wir haben also:

$$\begin{aligned} Y_i &= a + bX_i + \epsilon_i \\ Y_i &= \hat{a} + \hat{b}X_i + r_i \\ \hat{Y}_i &= \hat{a} + \hat{b}X_i \end{aligned}$$

Vorhersagewerte und Residuen

Wir haben also:

$$\begin{aligned} Y_i &= a + bX_i + \epsilon_i \\ Y_i &= \hat{a} + \hat{b}X_i + r_i \\ \hat{Y}_i &= \hat{a} + \hat{b}X_i \end{aligned}$$

und somit

$$r_i = Y_i - \hat{Y}_i$$

Brustkrebs

	Mortality	Temperature
1	102.5	51.3
2	104.5	49.9
3	100.4	50.0
4	95.9	49.2
5	87.0	48.5
6	95.0	47.8
7	88.6	47.3
8	89.2	45.1
9	78.9	46.3
10	84.6	42.1
11	81.7	44.2
12	72.2	43.5
13	65.1	42.3
14	68.1	40.2
15	67.3	31.8
16	52.5	34.0

Sinn oder Unsinn,...

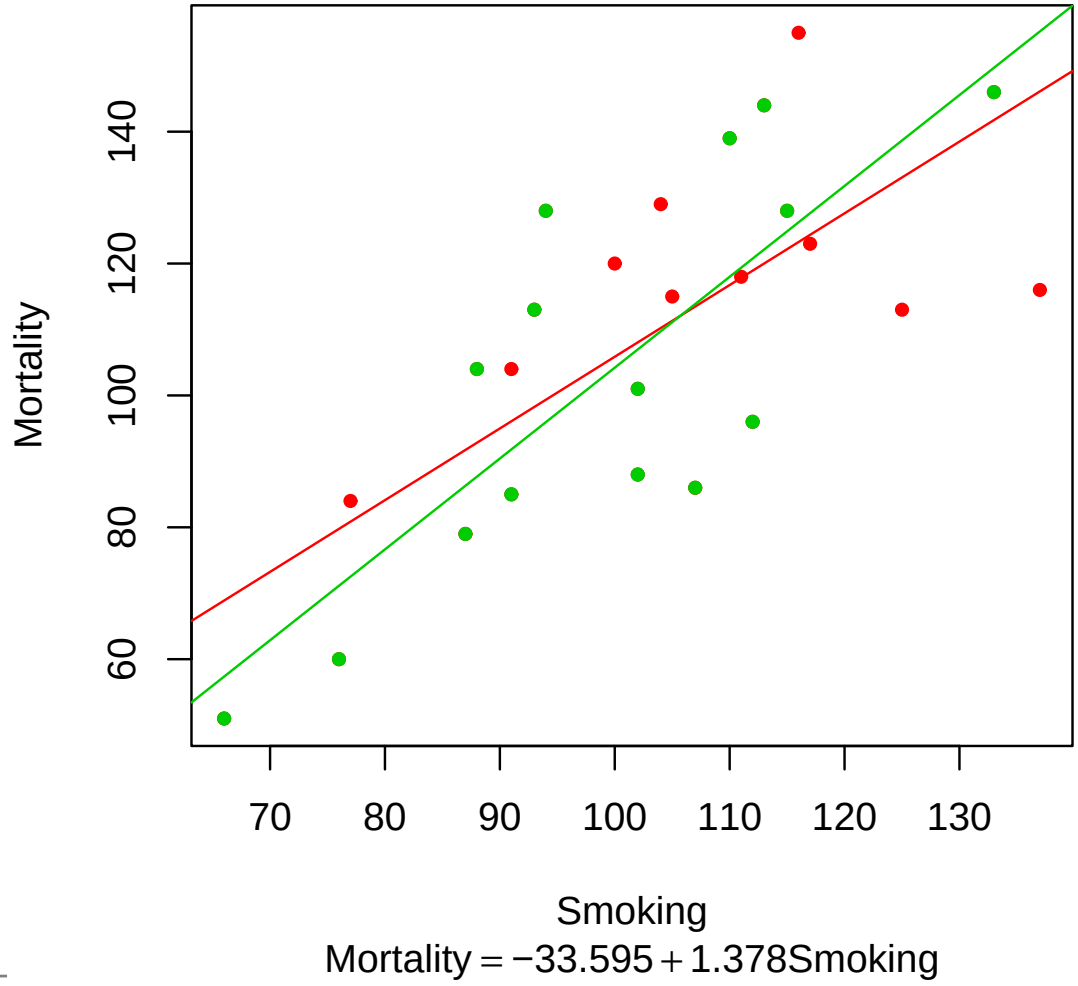
- Die Regression kann auf jeden Datensatz angewendet werden
- . . . , aber ist das auch immer sinnvoll?

Stichprobeneinfluß

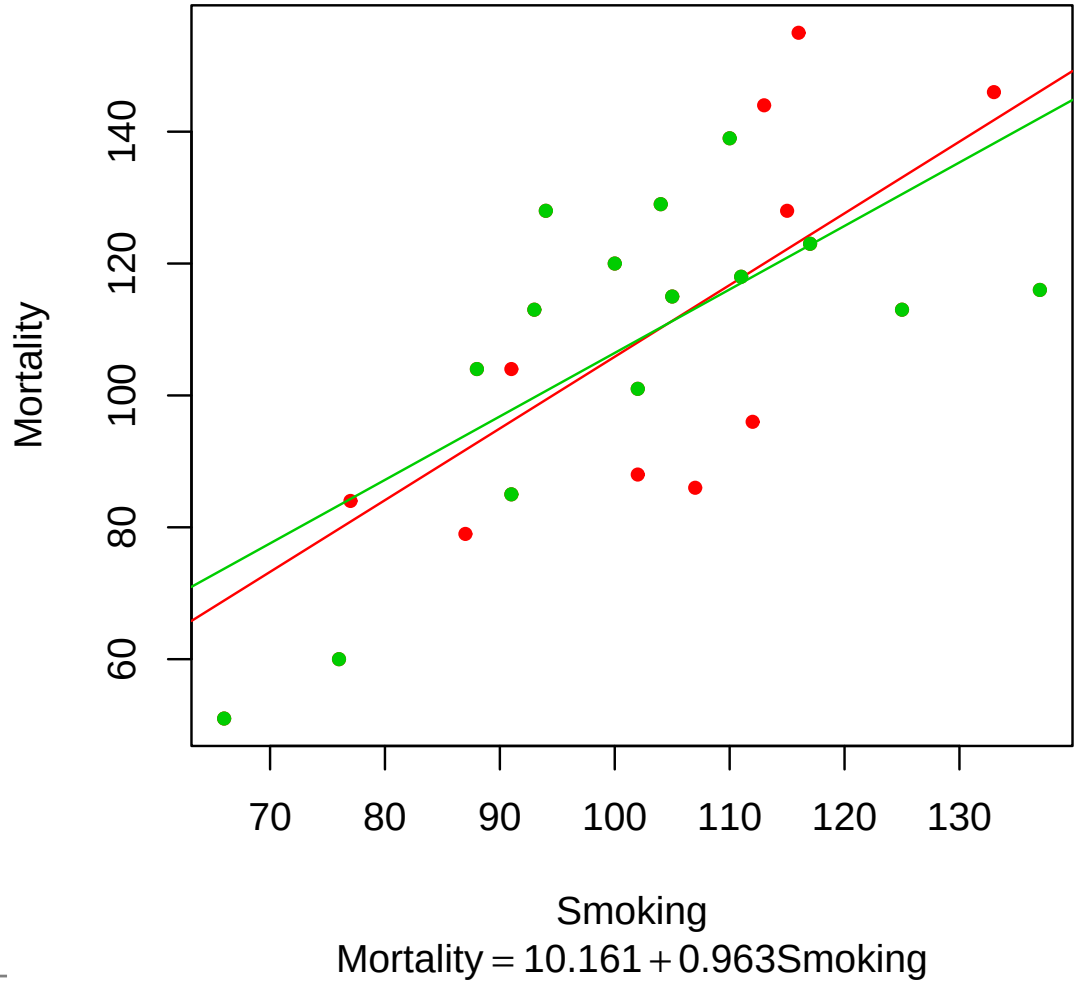
Demonstration: Stichprobenwahl

- Abhängig von der Wahl der Stichprobe werden die Geraden verschieden geschätzt.
- Das werden wir auf der nächsten Folie demonstrien indem wir jeweils 15 der Datenpunkte zufällig auswählen und die Gerade nur aufgrund dieser Punkte schätzen lassen.

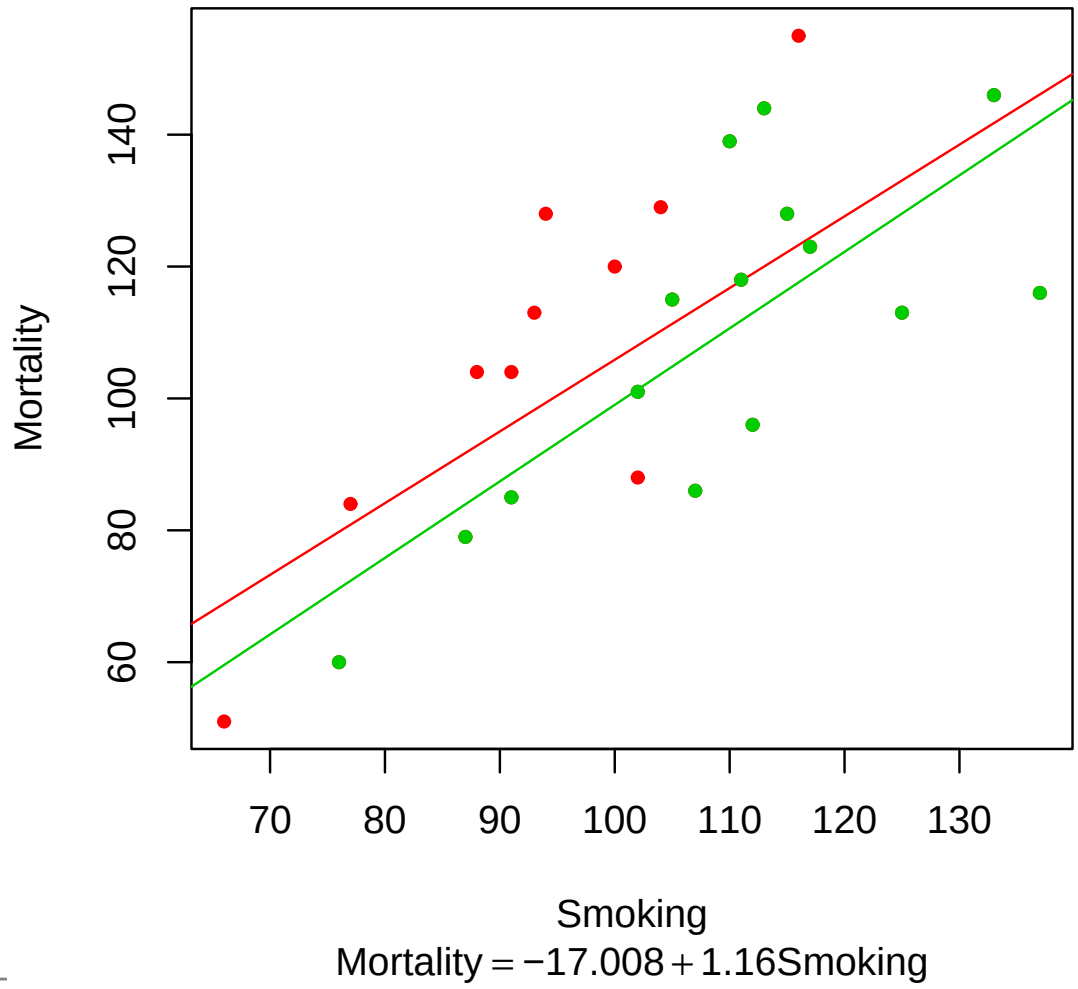
Demonstration: Stichprobenwahl



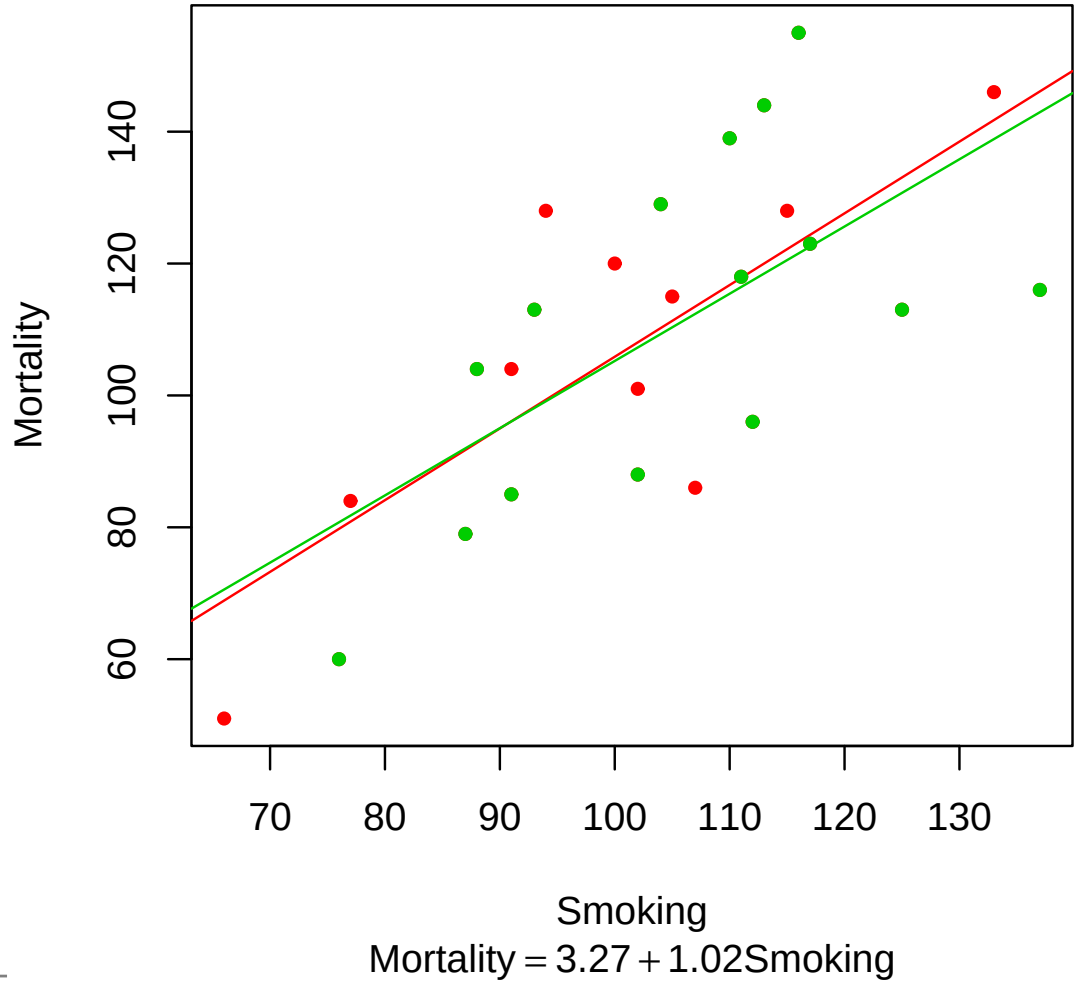
Demonstration: Stichprobenwahl



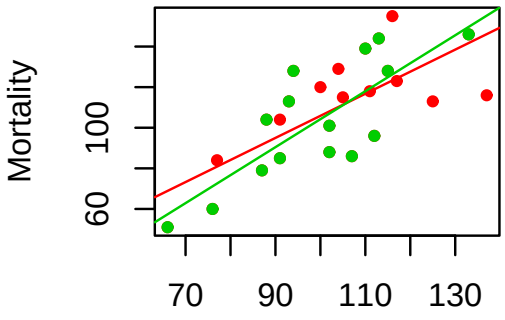
Demonstration: Stichprobenwahl



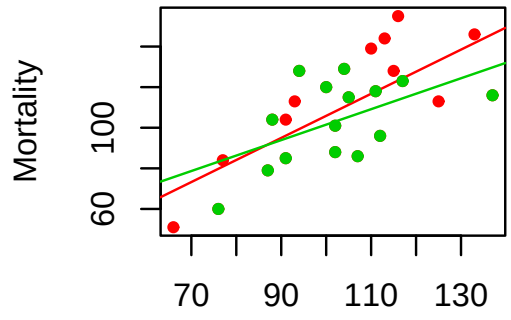
Demonstration: Stichprobenwahl



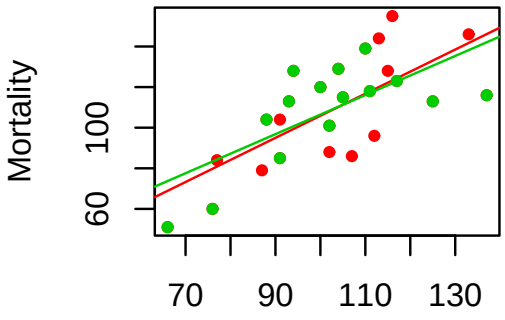
Demonstration: Stichprobenwahl



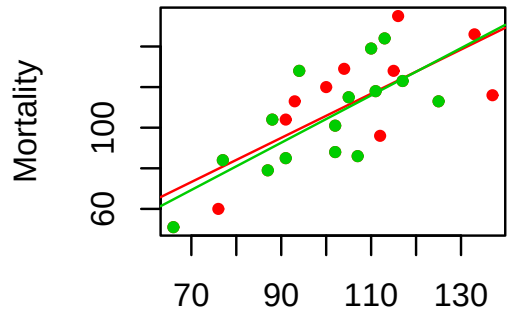
Mortality = −33.595 + 1.378Smoking



Mortality = 25.353 + 0.762Smoking



Mortality = 10.161 + 0.963Smoking



Mortality = −12.234 + 1.165Smoking

Konfidenzintervall für die Vorhersage

- Wir können einen Bereich einzeichnen in dem der wahre Wert $E[Y] = a + bX$ mit $1 - \alpha = 95\%$ Wahrscheinlichkeit zu liegen kommt:
- Die Formel ist kompliziert:

$$u(x) = \hat{a} + \hat{b}x + \hat{sd}(\epsilon) q_{t_{n-2},1-\alpha/2} (c_1 + 2c_2 x + c_3 x^2)$$

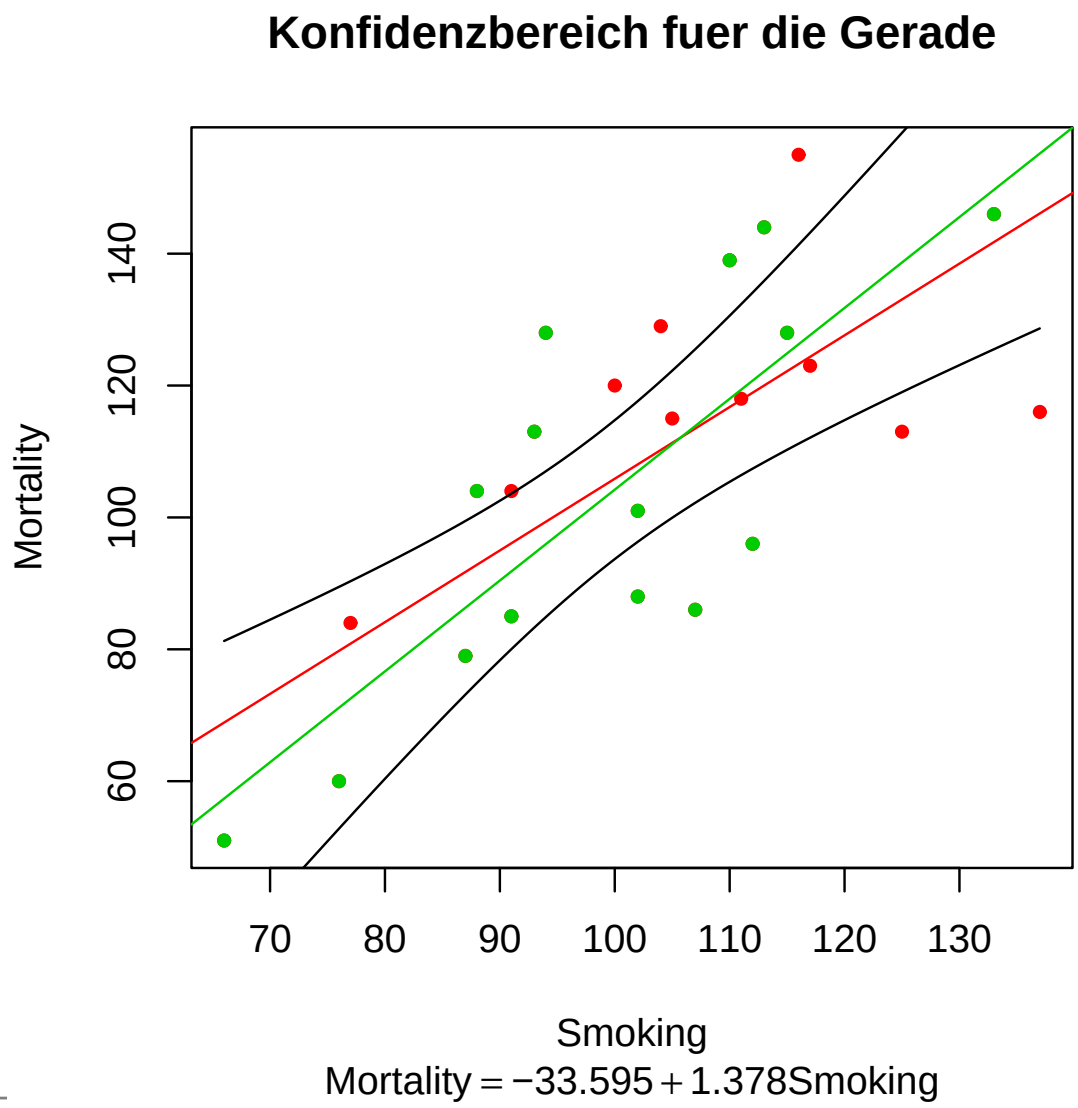
$$l(x) = \hat{a} + \hat{b}x - \hat{sd}(\epsilon) q_{t_{n-2},\alpha/2} (c_1 + 2c_2 x + c_3 x^2)$$

mit $q_{t_{n-2},1-\alpha/2}$ dem $1 - \alpha/2$-Quantil der t-Verteilung und

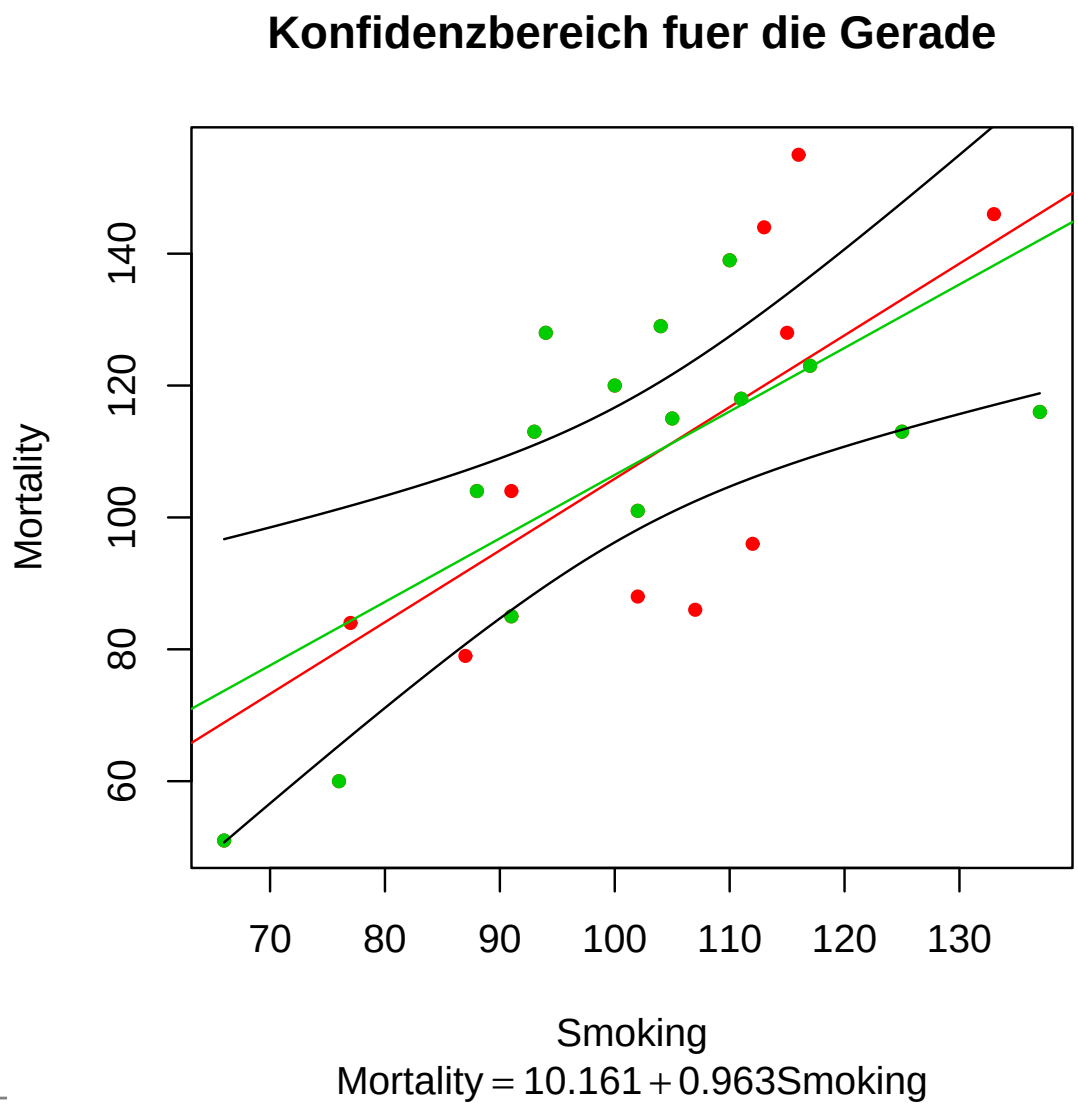
$$\begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}^{-1}$$

ohne Gewähr.

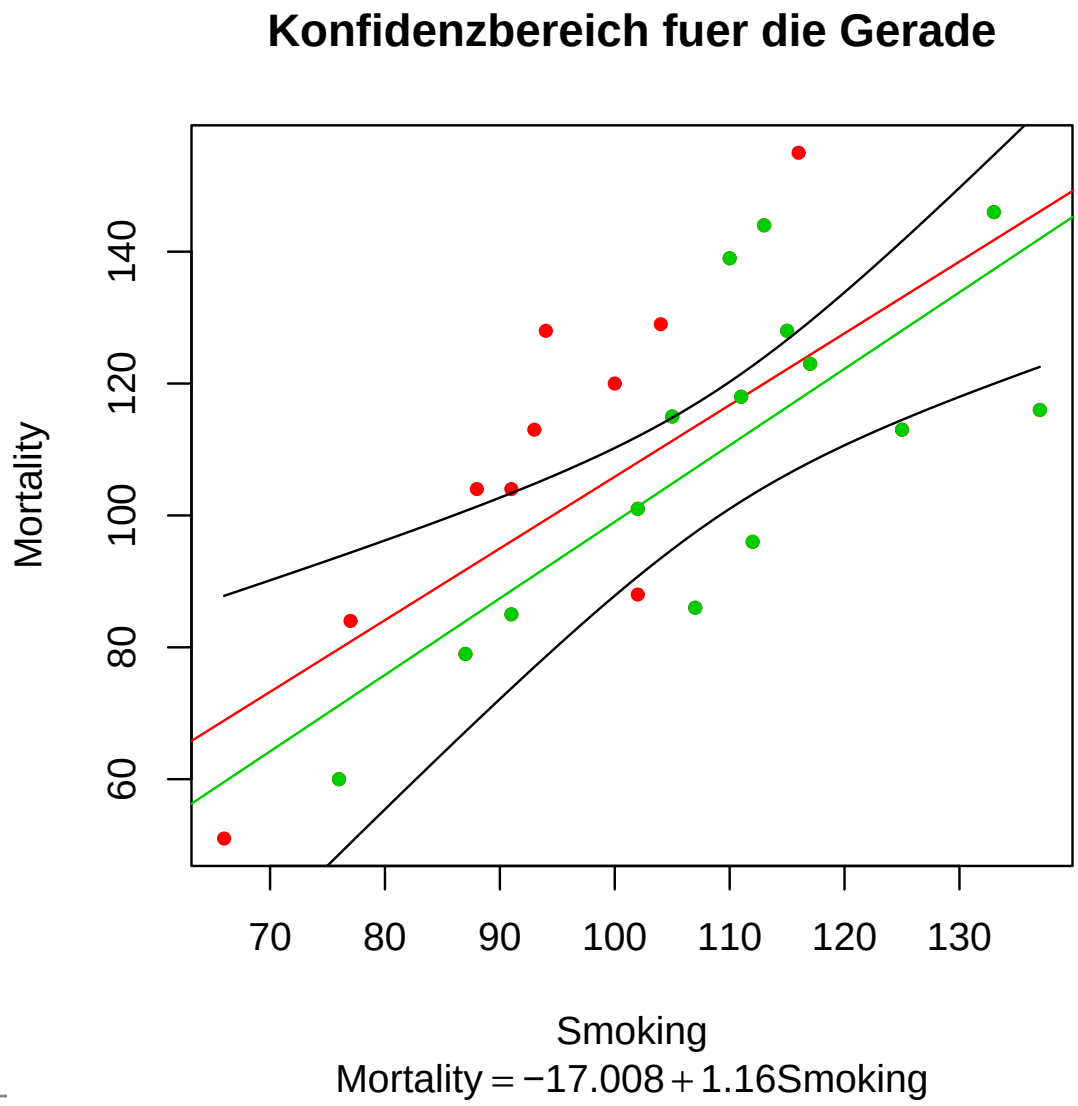
Demonstration des Konfidenzintervalls



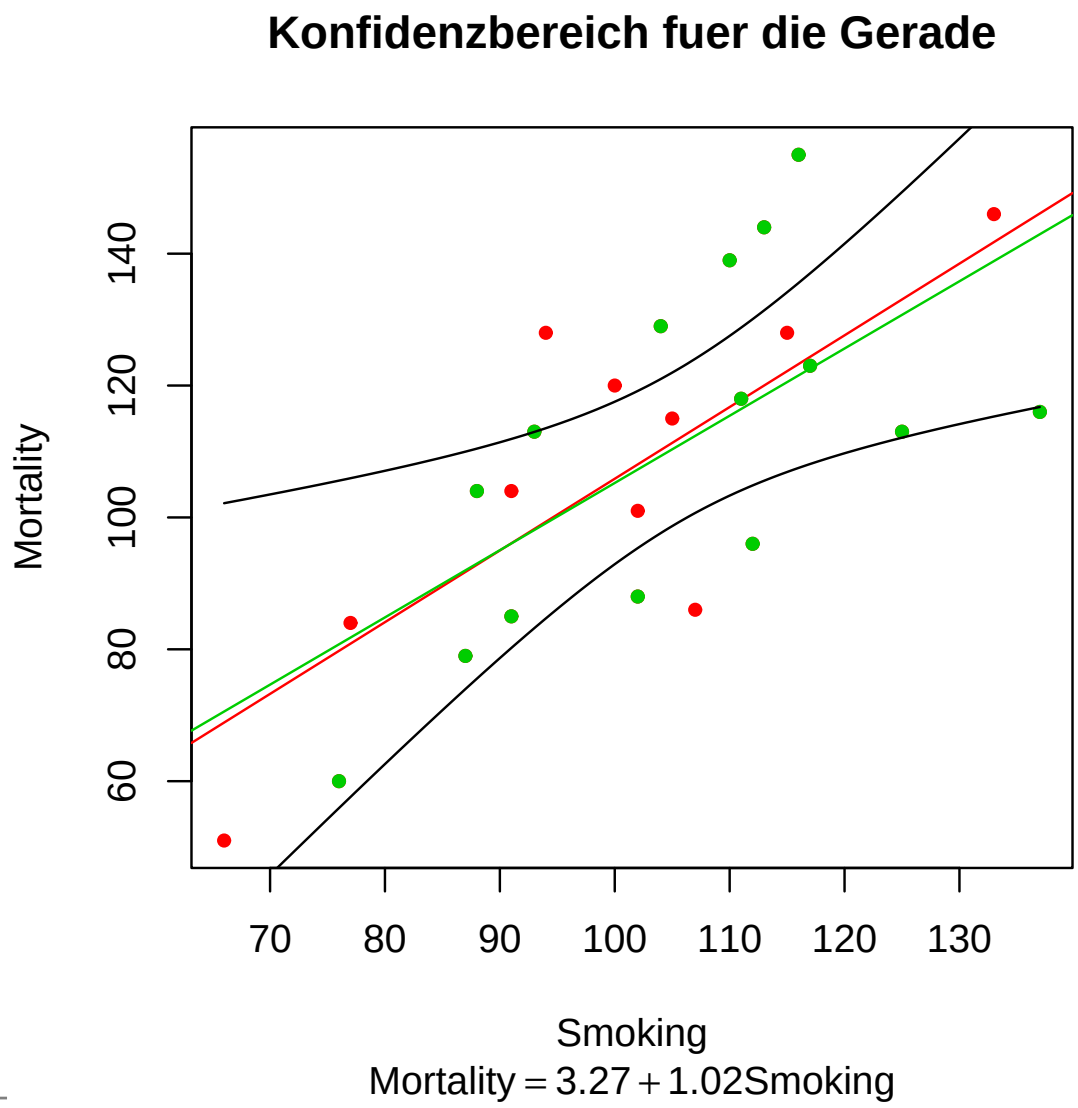
Demonstration des Konfidenzintervalls



Demonstration des Konfidenzintervalls

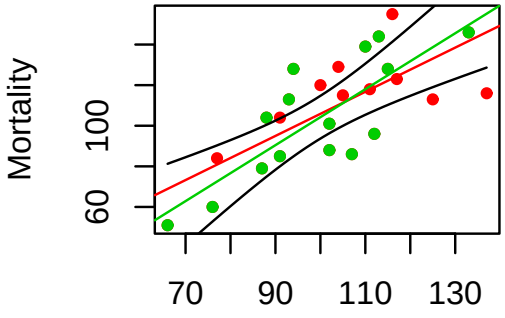


Demonstration des Konfidenzintervalls



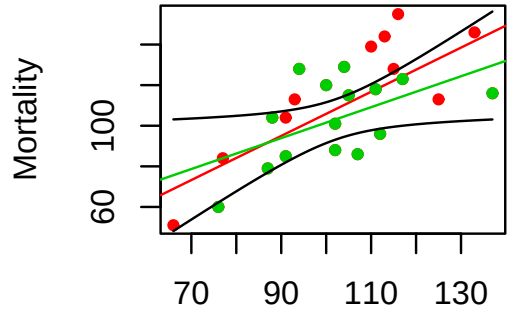
Demonstration des Konfidenzintervalls

Konfidenzbereich fuer die Gera



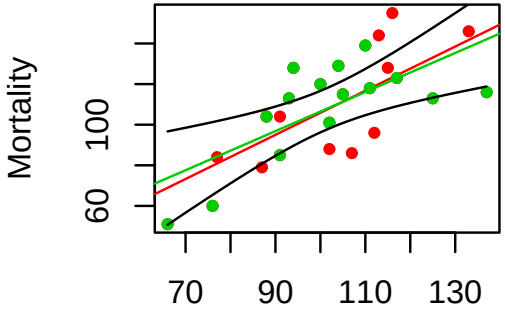
Mortality = −33.595 + 1.378Smoking

Konfidenzbereich fuer die Gera



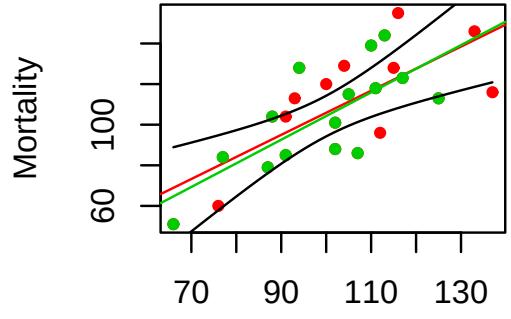
Mortality = 25.353 + 0.762Smoking

Konfidenzbereich fuer die Gera



Mortality = 10.161 + 0.963Smoking

Konfidenzbereich fuer die Gera



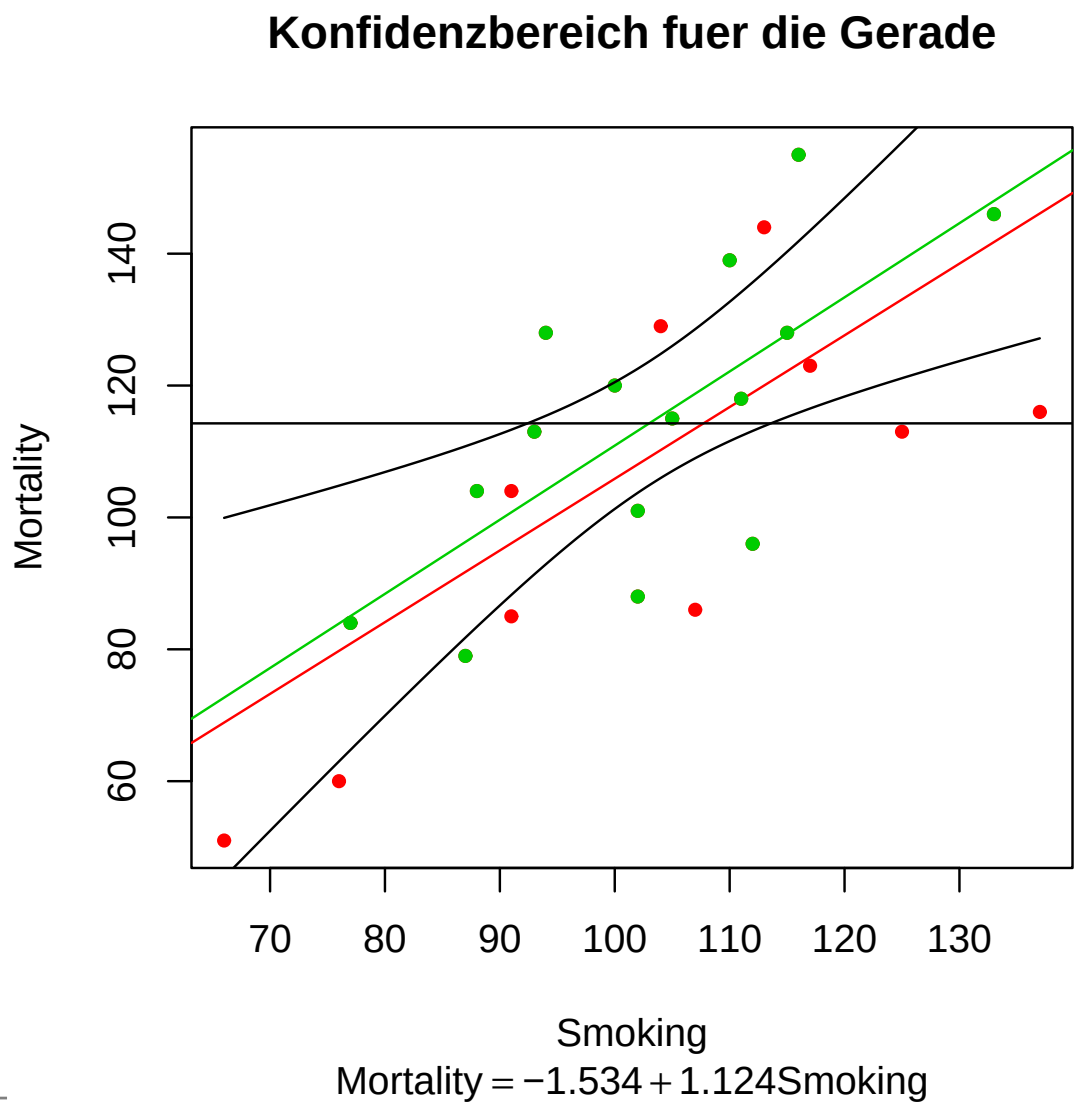
Mortality = −12.234 + 1.165Smoking

Sinn oder Unsinn,...

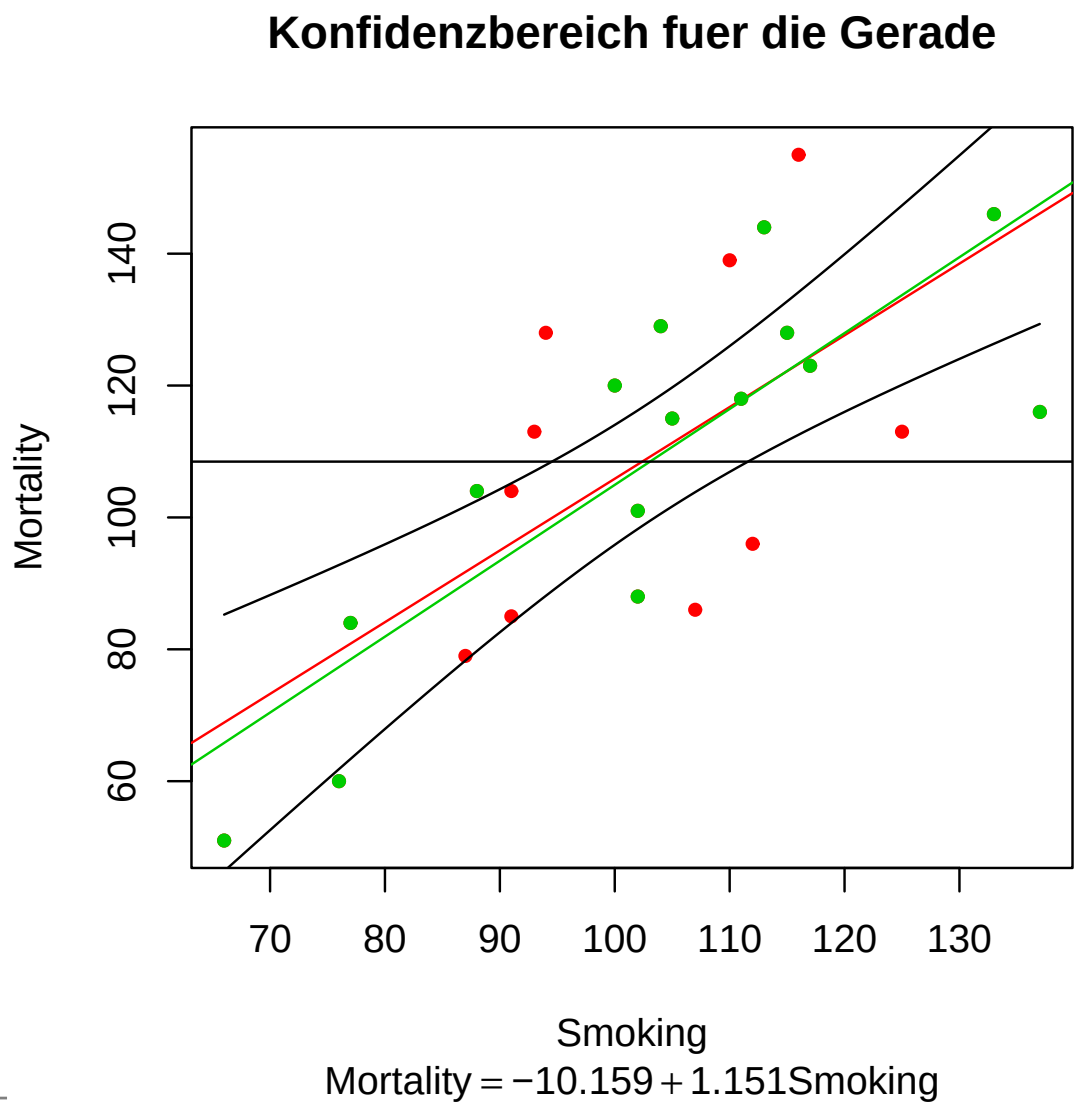
Eine wichtige Frage ist also immer, ob eventuell kein Zusammenhang bestehen konnte.
Wenn X und Y unabhängig sind, dann ändert sich der Erwartungswert von Y nicht in Abhängigkeit von X:

$$Y = a + 0X + \epsilon$$

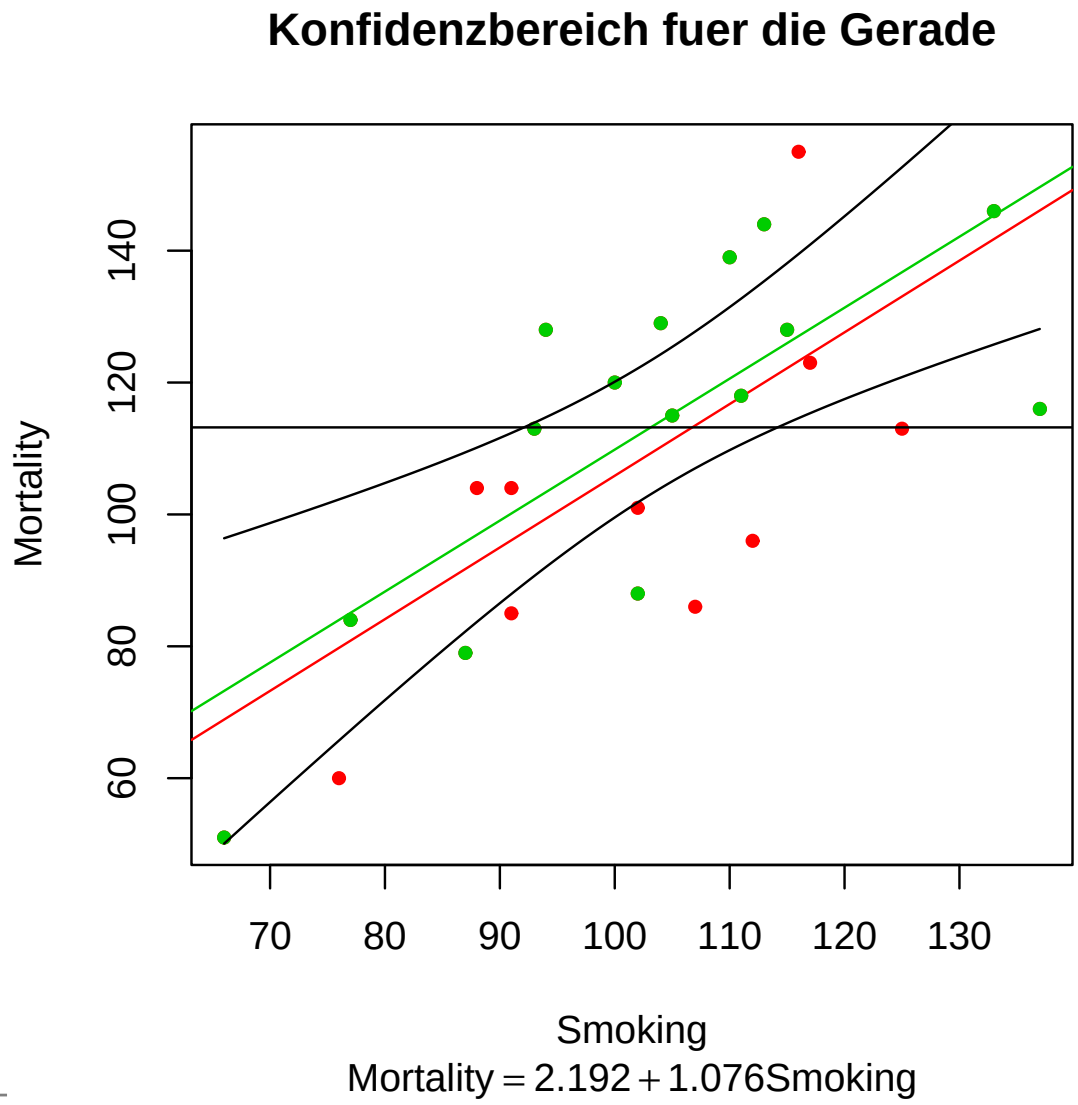
Könnte kein Zusammenhang bestehen?



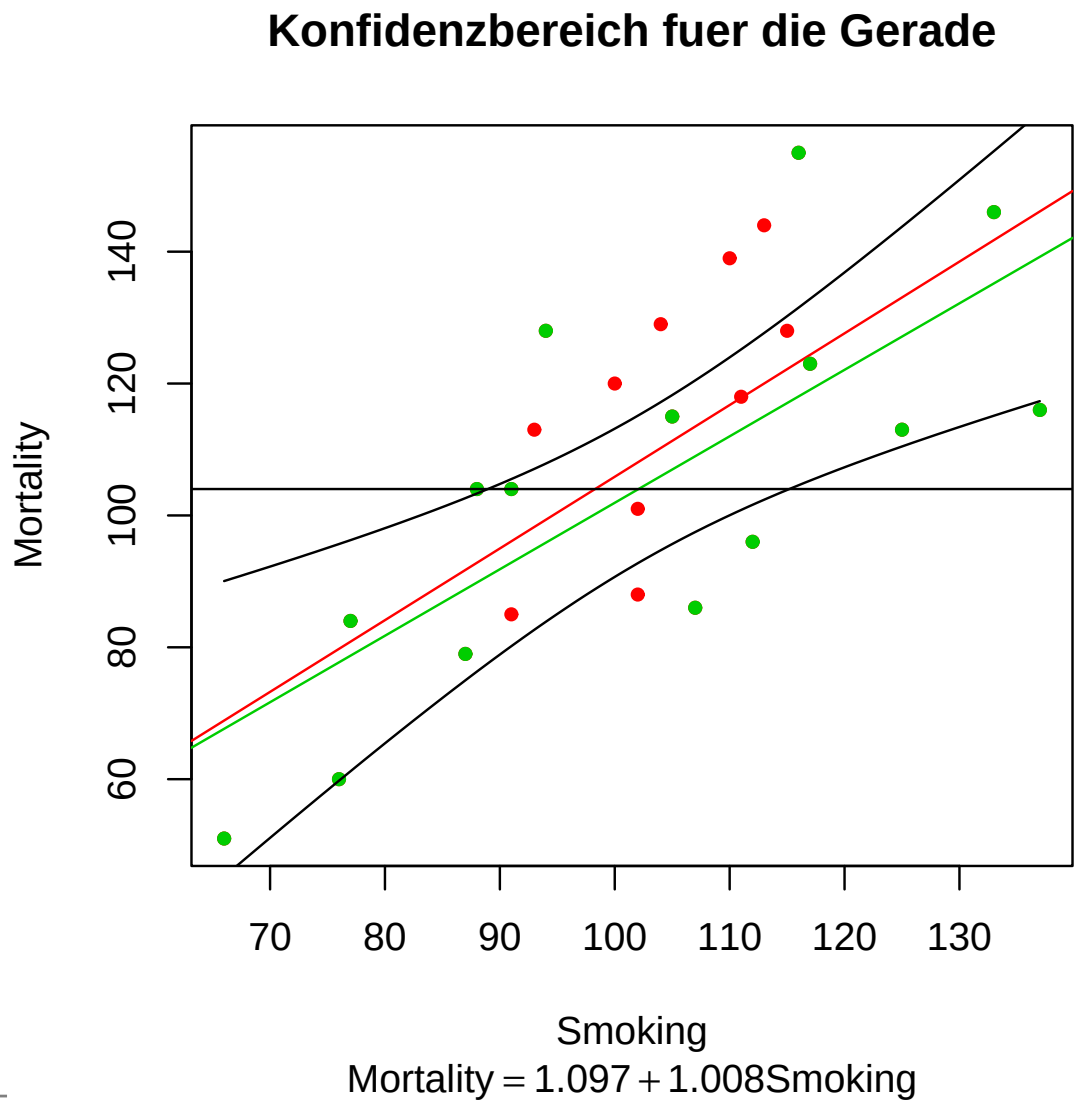
Könnte kein Zusammenhang bestehen?



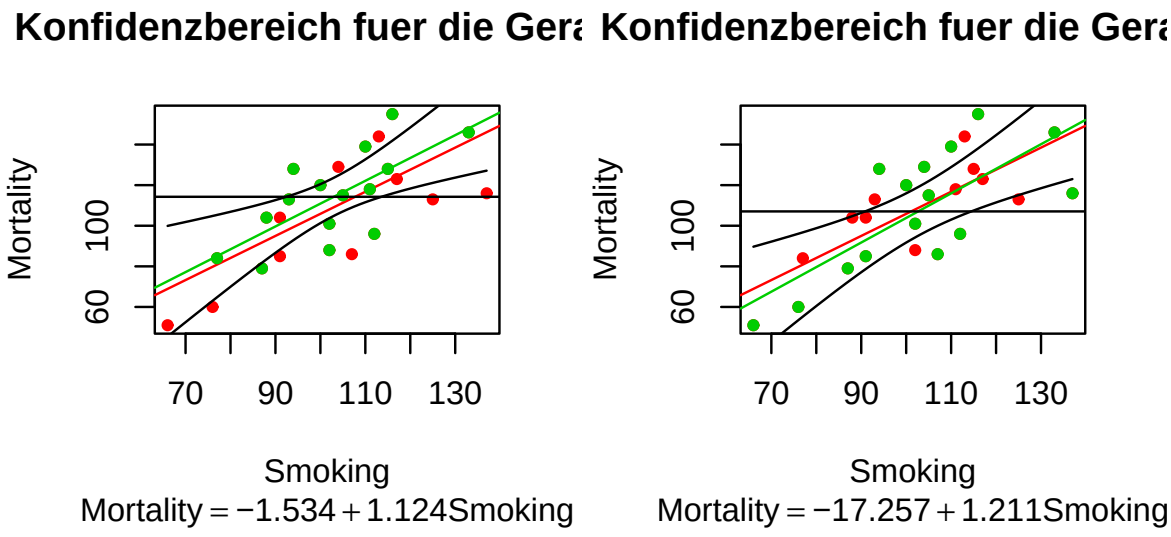
Könnte kein Zusammenhang bestehen?

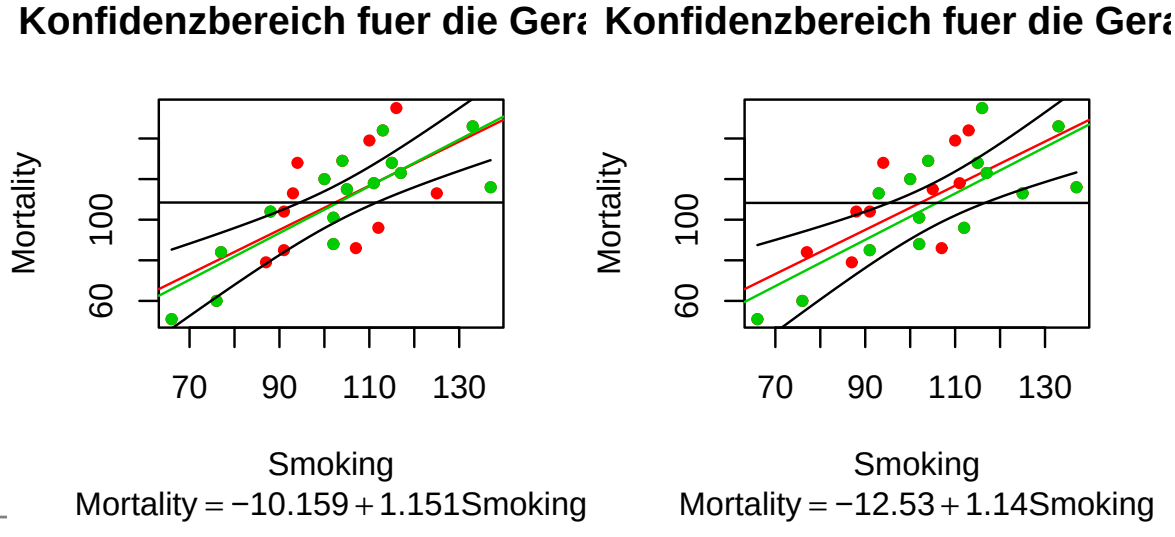


Könnte kein Zusammenhang bestehen?



Könnte kein Zusammenhang bestehen?

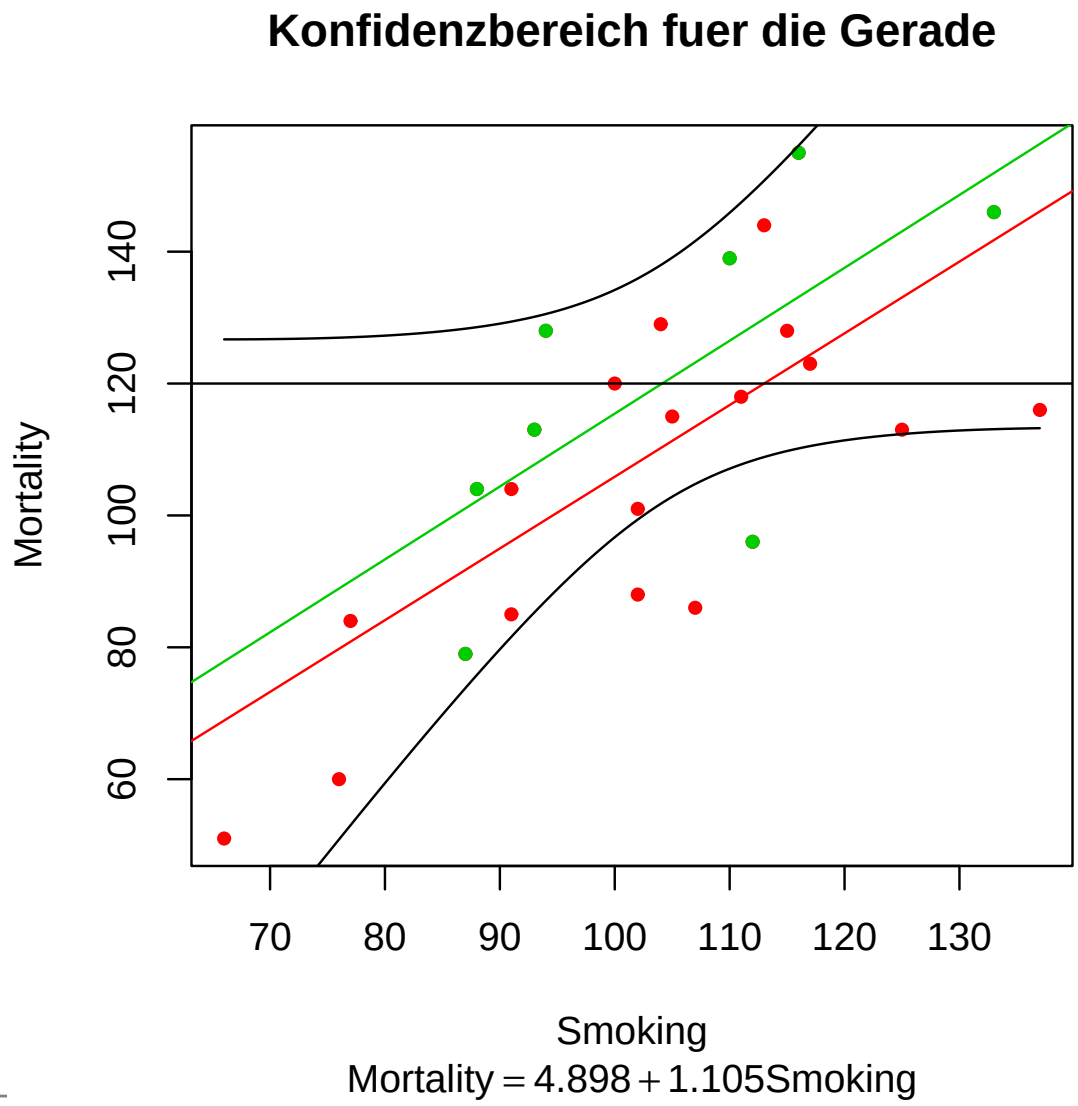




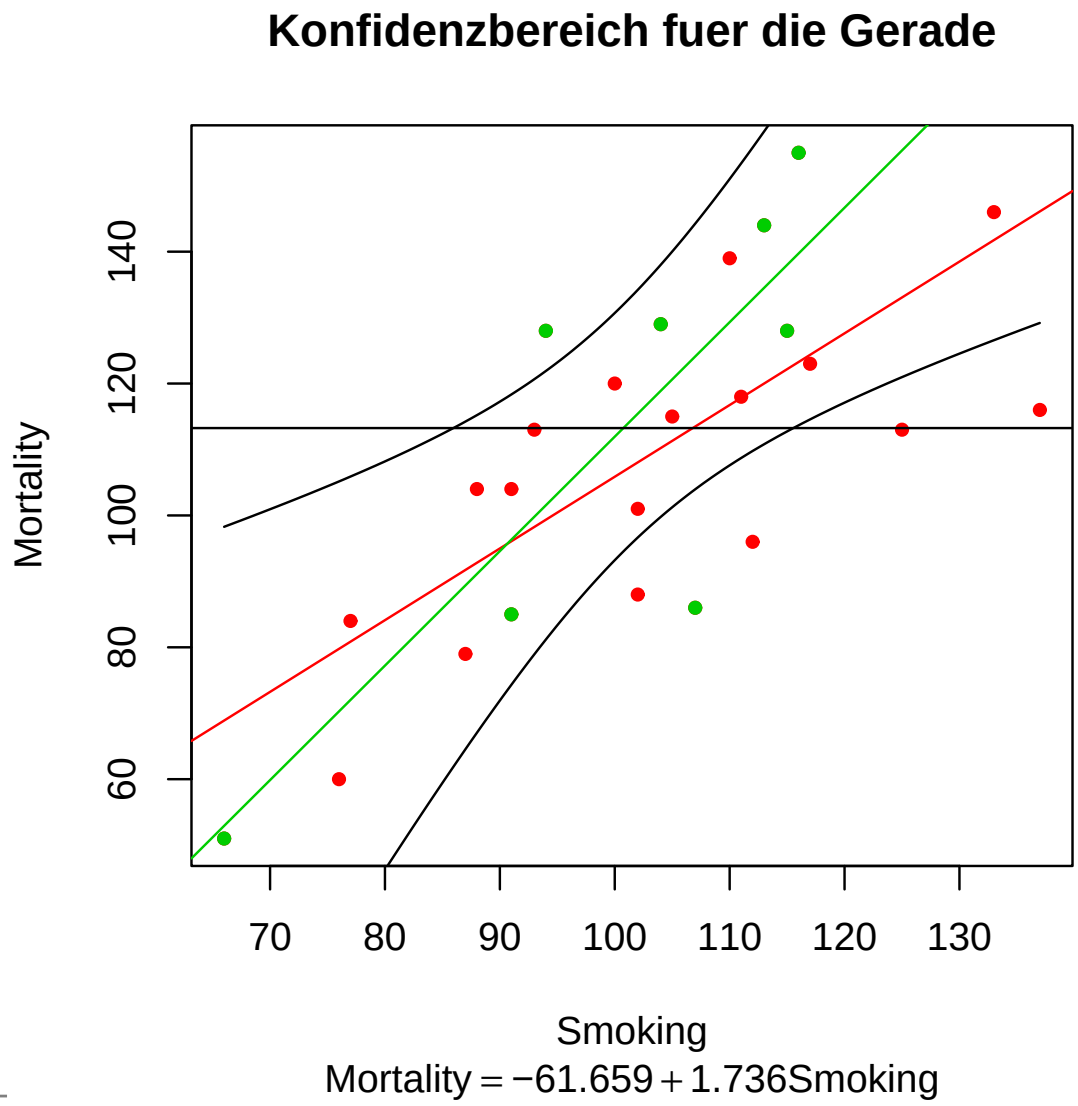
Ergebnisse

- Das 95%-Konfidenzintervalle (für die Gerade) ist ein zufälliges Intervall um die geschätzte Gerade, dass an jeder Stelle die wahren Gerade mit einer wahrscheinlichkeit von 95% umschließt .

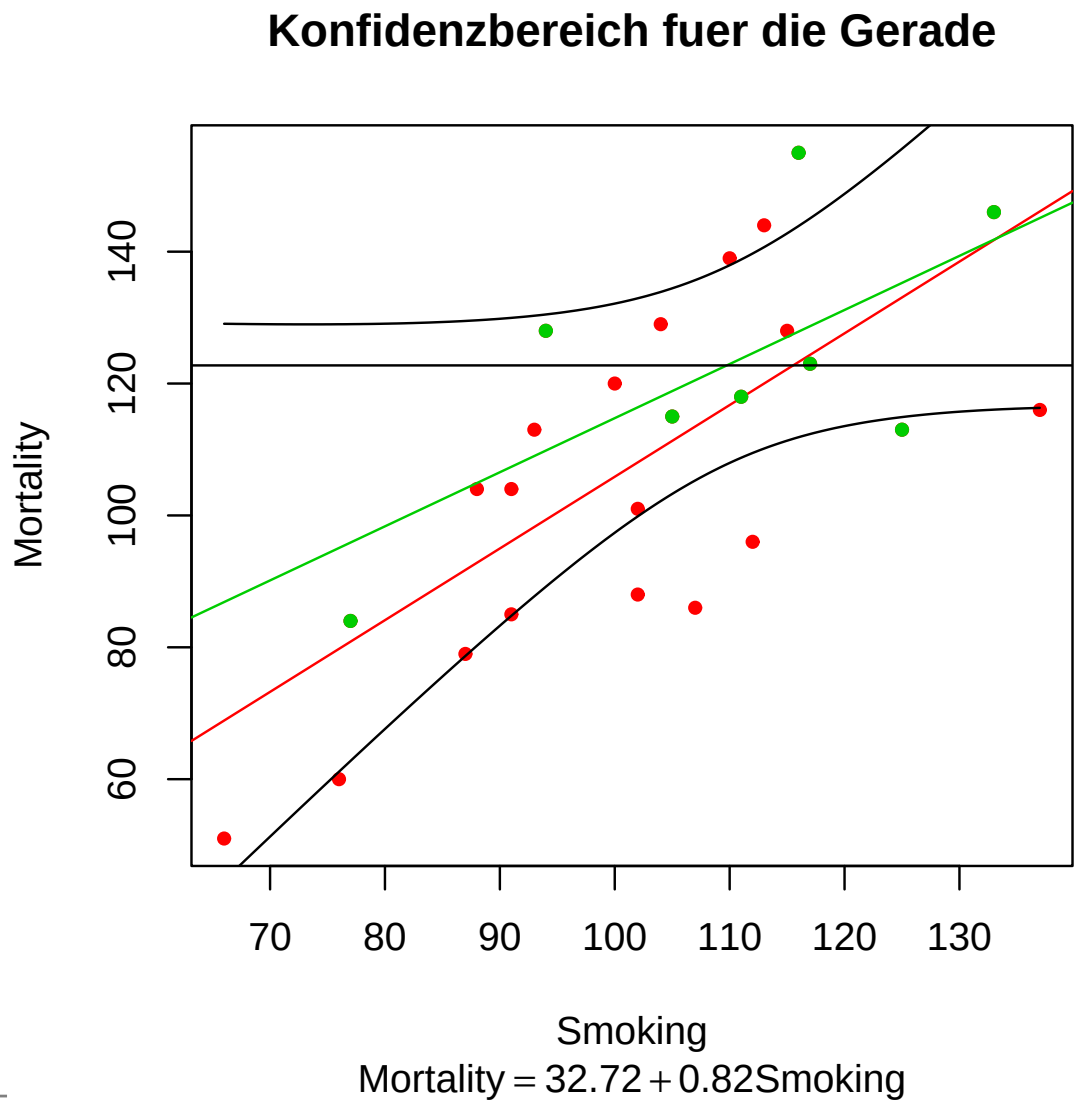
Nochmal mit weniger Daten



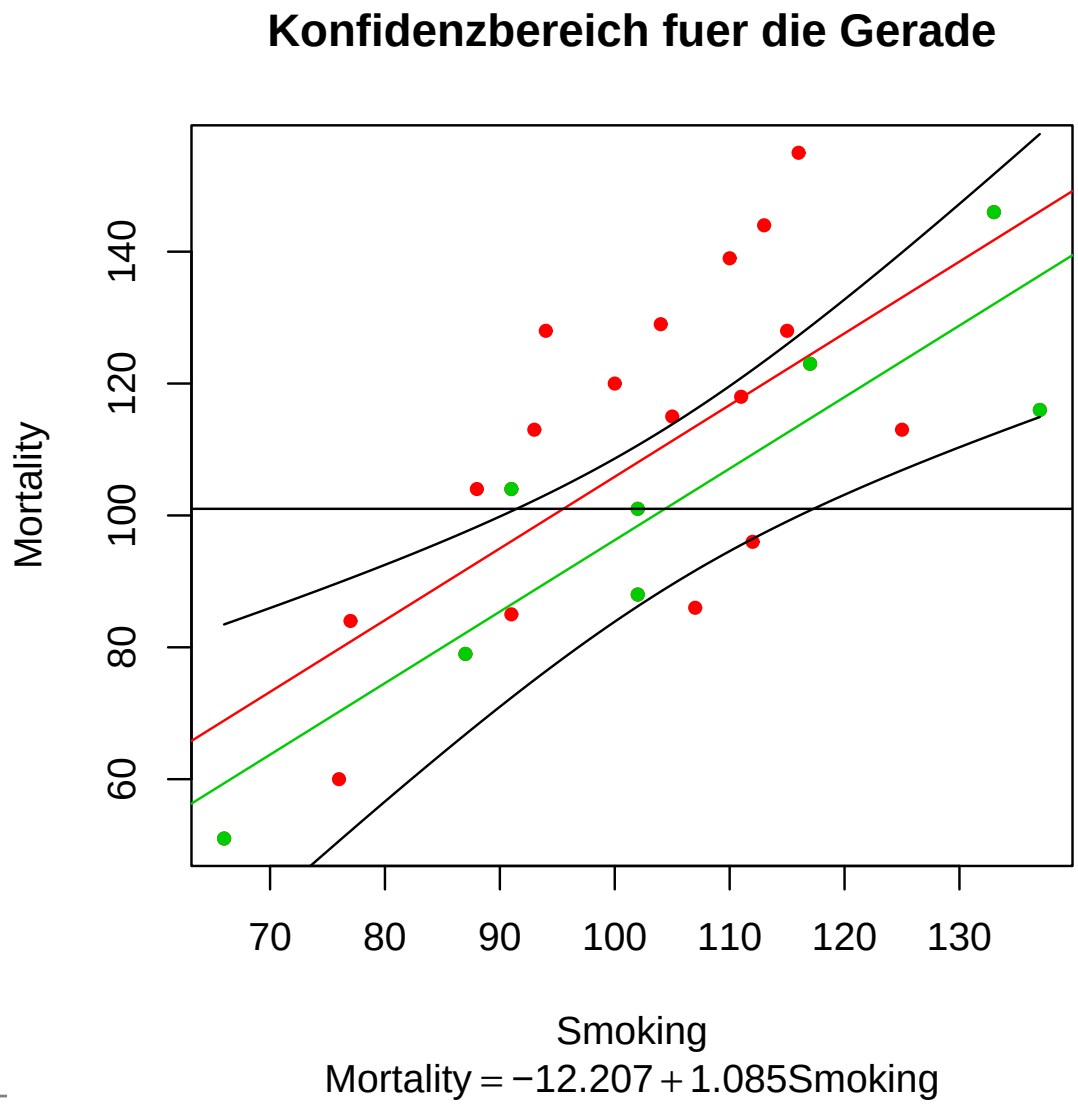
Nochmal mit weniger Daten



Nochmal mit weniger Daten

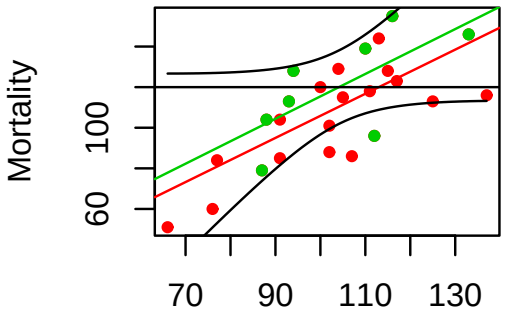


Nochmal mit weniger Daten



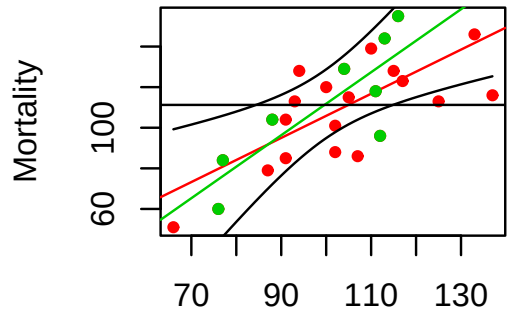
Nochmal mit weniger Daten

Konfidenzbereich fuer die Gera



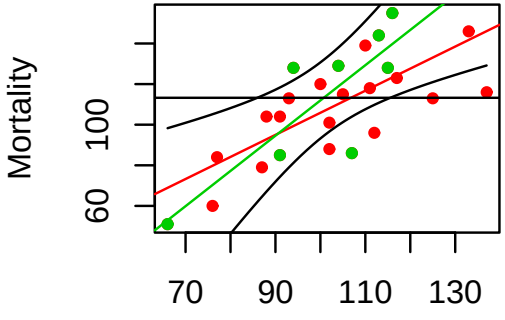
Mortality = 4.898 + 1.105Smoking

Konfidenzbereich fuer die Gera



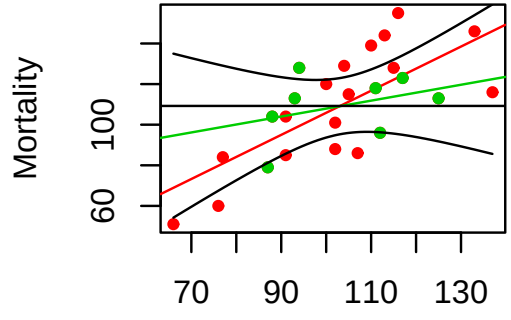
Mortality = −43.334 + 1.552Smoking

Konfidenzbereich fuer die Gera



Mortality = −61.659 + 1.736Smoking

Konfidenzbereich fuer die Gera



Mortality = 68.742 + 0.392Smoking

Das Vorhersageintervall

Frage: In welchem Bereich liegen die Sterblichkeit bei einer Berufsgruppe die 130% raucht?

Das Vorhersageintervall

Frage: In welchem Bereich liegen die Sterblichkeit bei einer Berufsgruppe die 130% raucht?
Lösung: Das Vorhersageintervall

Das Vorhersageintervall

Frage: In welchem Bereich liegen die Sterblichkeit bei einer Berufsgruppe die 130% raucht?
Lösung: Das Vorhersageintervall Die Formel ist kompliziert:

$$u(x) = \hat{a} + \hat{b}x + \hat{sd}(\epsilon) q_{t_{n-2},1-\alpha/2}(1 + c_1 + 2c_2 x + c_3 x^2)$$

$$l(x) = \hat{a} + \hat{b}x - \hat{sd}(\epsilon) q_{t_{n-2},\alpha/2}(1 + c_1 + 2c_2 x + c_3 x^2)$$

mit $q_{t_{n-2},p}$ dem p-Quantil der t-Verteilung und

$$\begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}^{-1}$$

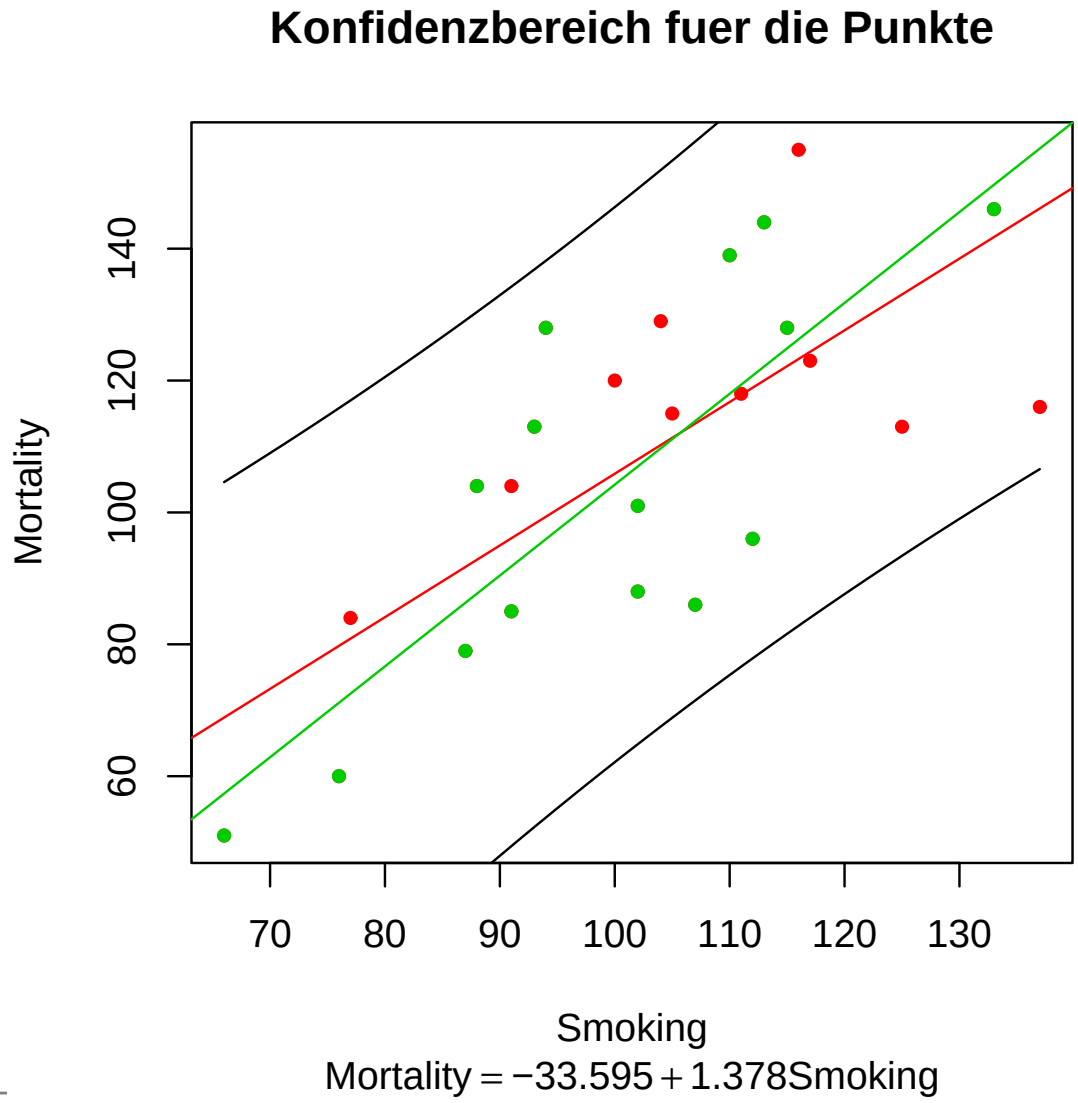
ohne Gewähr.

Formeln

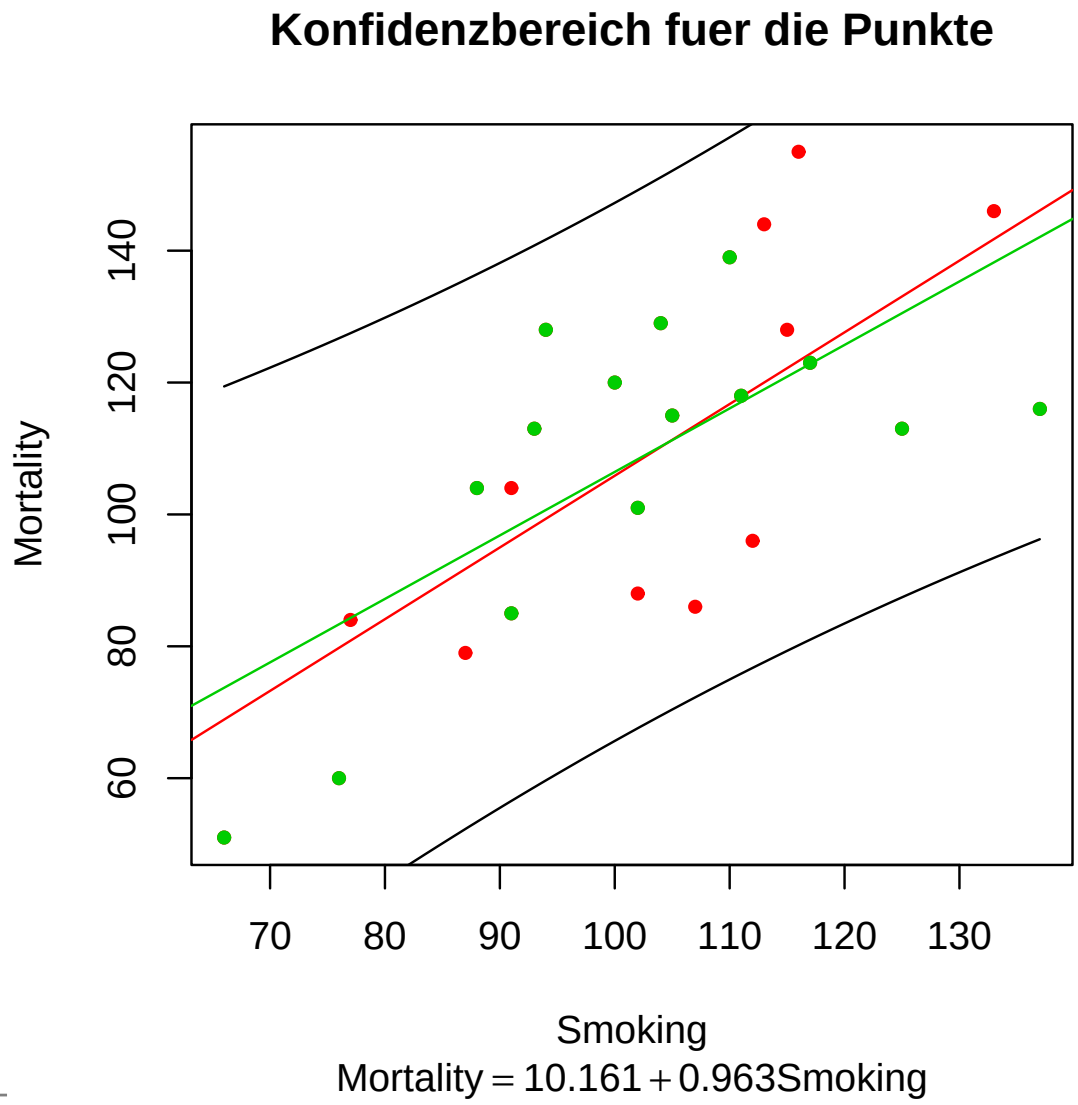
$$
\begin{aligned}
c_0 &= \frac{1}{n\sum X_i^2 - \left(\sum_i X_i\right)^2} \\
c_1 &= c_0 \sum X_i^2 \\
c_2 &= -c_0 \sum X_i \\
c_3 &= c_0 n \\
\hat{sd}(\epsilon) &= \frac{1}{n-2} \sum_{i=1}^{n} r_i^2
\end{aligned}
$$

ohne Gewähr

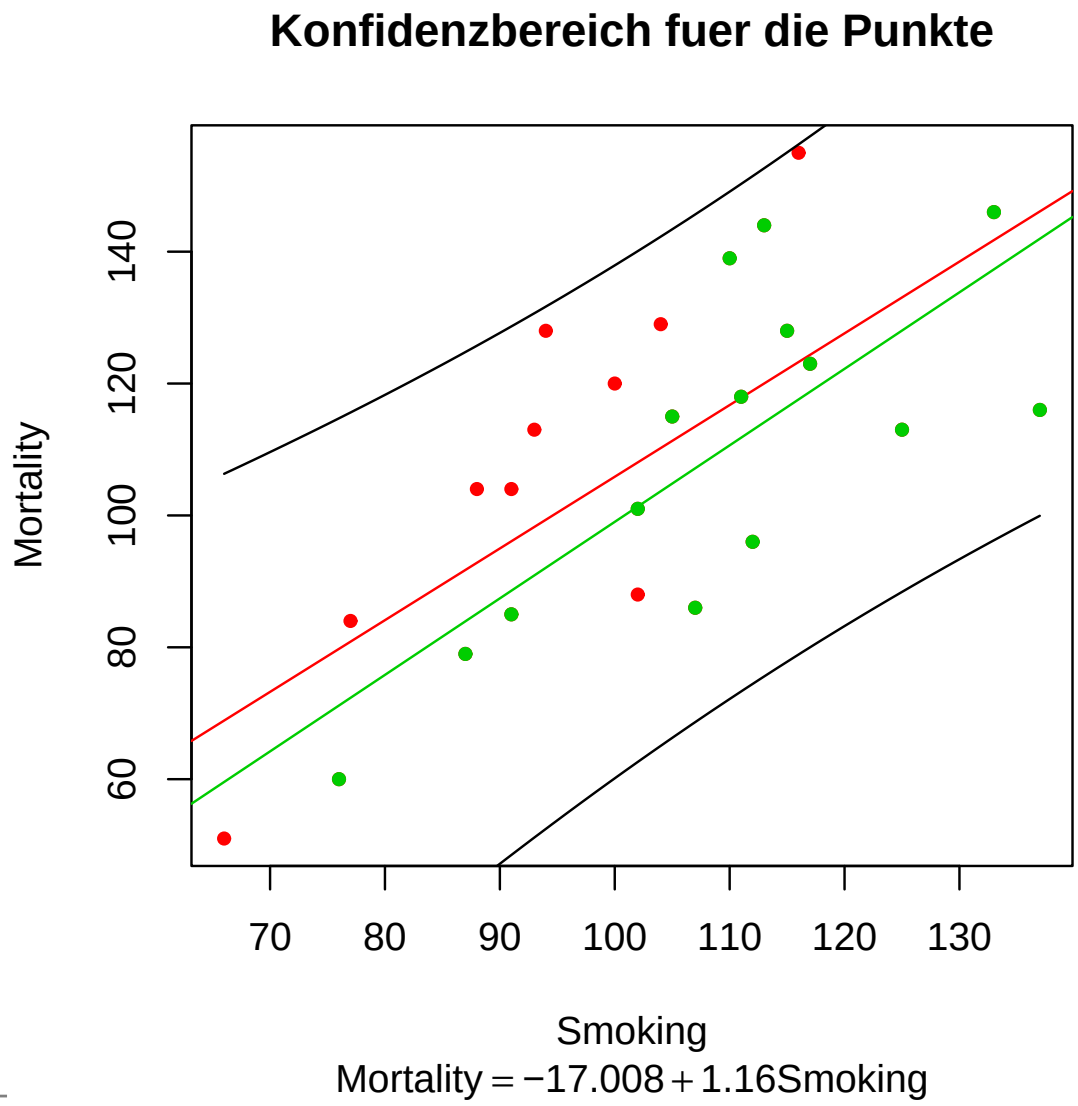
Demonstration des Vorhersageintervalls



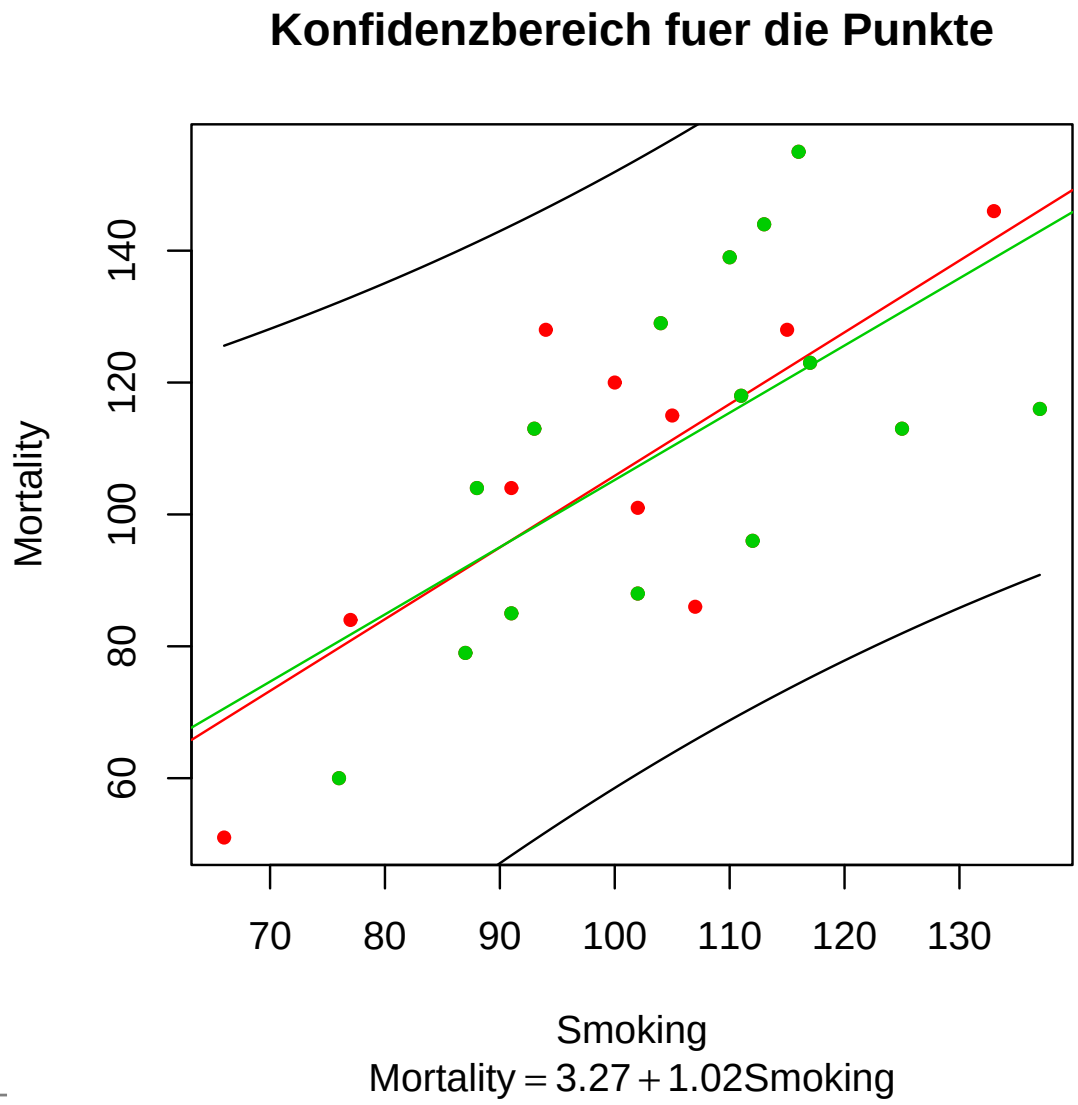
Demonstration des Vorhersageintervalls



Demonstration des Vorhersageintervalls

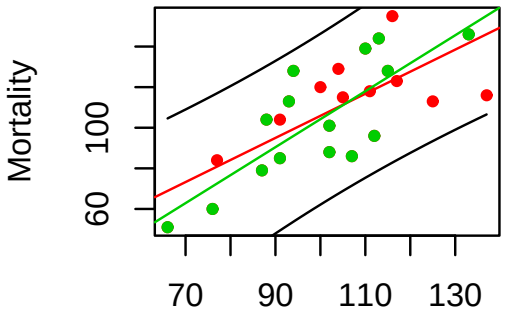


Demonstration des Vorhersageintervalls



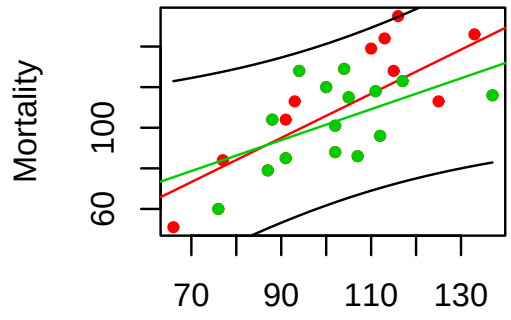
Demonstration des Vorhersageintervalls

Konfidenzbereich fuer die Pun



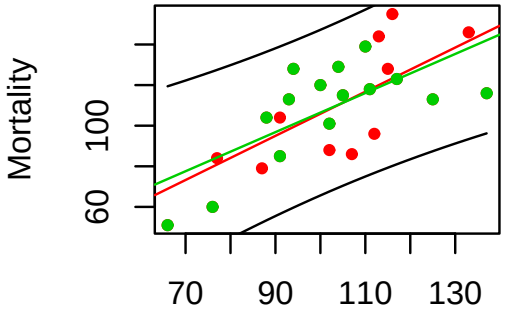
Mortality = −33.595 + 1.378Smoking

Konfidenzbereich fuer die Pun



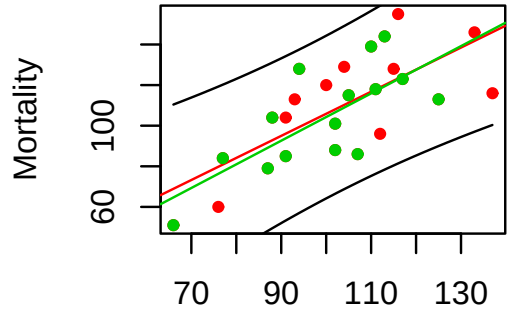
Mortality = 25.353 + 0.762Smoking

Konfidenzbereich fuer die Pun



Mortality = 10.161 + 0.963Smoking

Konfidenzbereich fuer die Pun



Mortality = −12.234 + 1.165Smoking

Ergebnisse

- Die Vorhersageintervalle sind bedeutend breiter als die Konfidenzintervalle.
- Echte Aussage über die einzelne Berufsgruppe sind offenbar nur für extremes Raucheverhalten zu treffen: Vorhersage von Werten ist bedeutend schwieriger als die Vorhersage von Tendenzen.

Der Regressionstest

Steigung = 0 $\Leftarrow X, Y$ unabhängig.

Der Regressionstest

Steigung = 0 $\Leftarrow X, Y$ unabhängig.
Dieses Problem kann mit einem Test untersucht werden:

Der Regressionstest

Steigung = 0 $\Leftarrow X, Y$ unabhängig.
Dieses Problem kann mit einem Test untersucht werden:
Regressionstest:

$$H_0 : b = 0 \quad vs. \quad H_1 : b \neq 0$$

Voraussetzungen: wie Regression

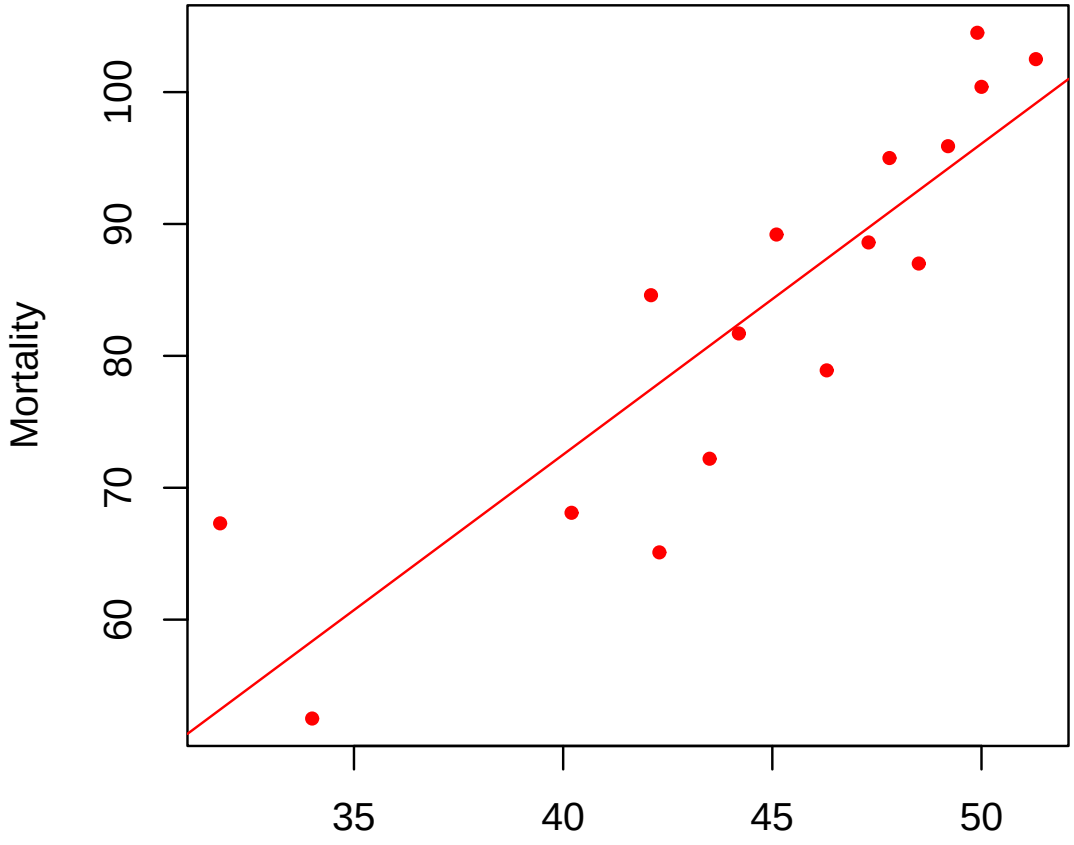
Anwendung: Nachweis der Abhängigkeit

Computerausgabe

```
> model <- lm(Mortality ~ Smoking, data = Rauchen)
> anova(model)

Analysis of Variance Table

Response: Mortality
          Df Sum Sq Mean Sq F value    Pr(>F)
Smoking    1 8395.7  8395.7  24.228 5.658e-05 ***
Residuals 23 7970.3   346.5
---
Signif. codes:  0 âÃŸ***âÃŹ 0.001 âÃŸ**âÃŹ 0.01 âÃŸ*âÃŹ 0.05 âÃŸ.â
```

Der p-Wert wird steht in der letzten Spalte der Tabelle.

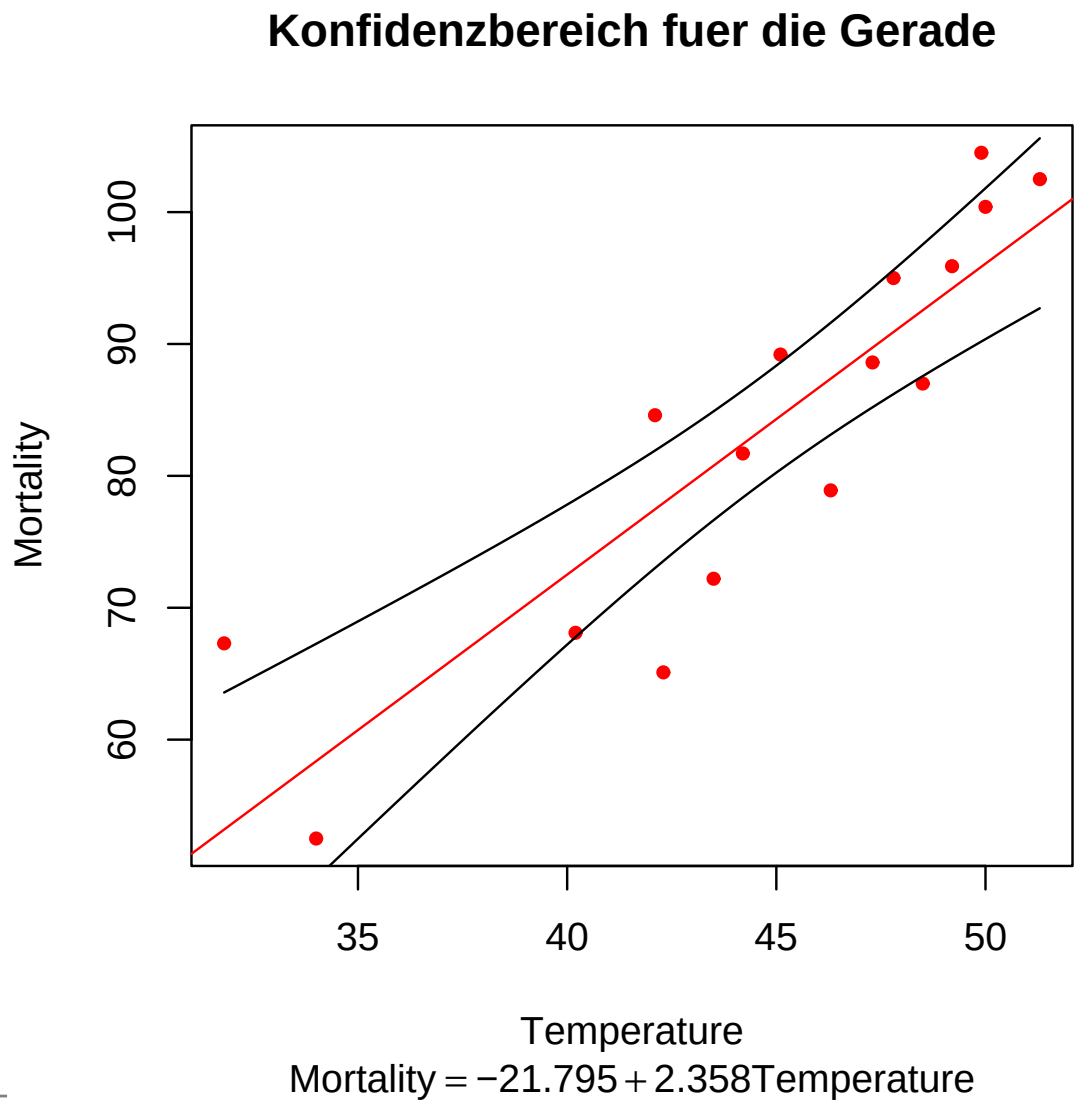
Beispiel: Brustkrebs



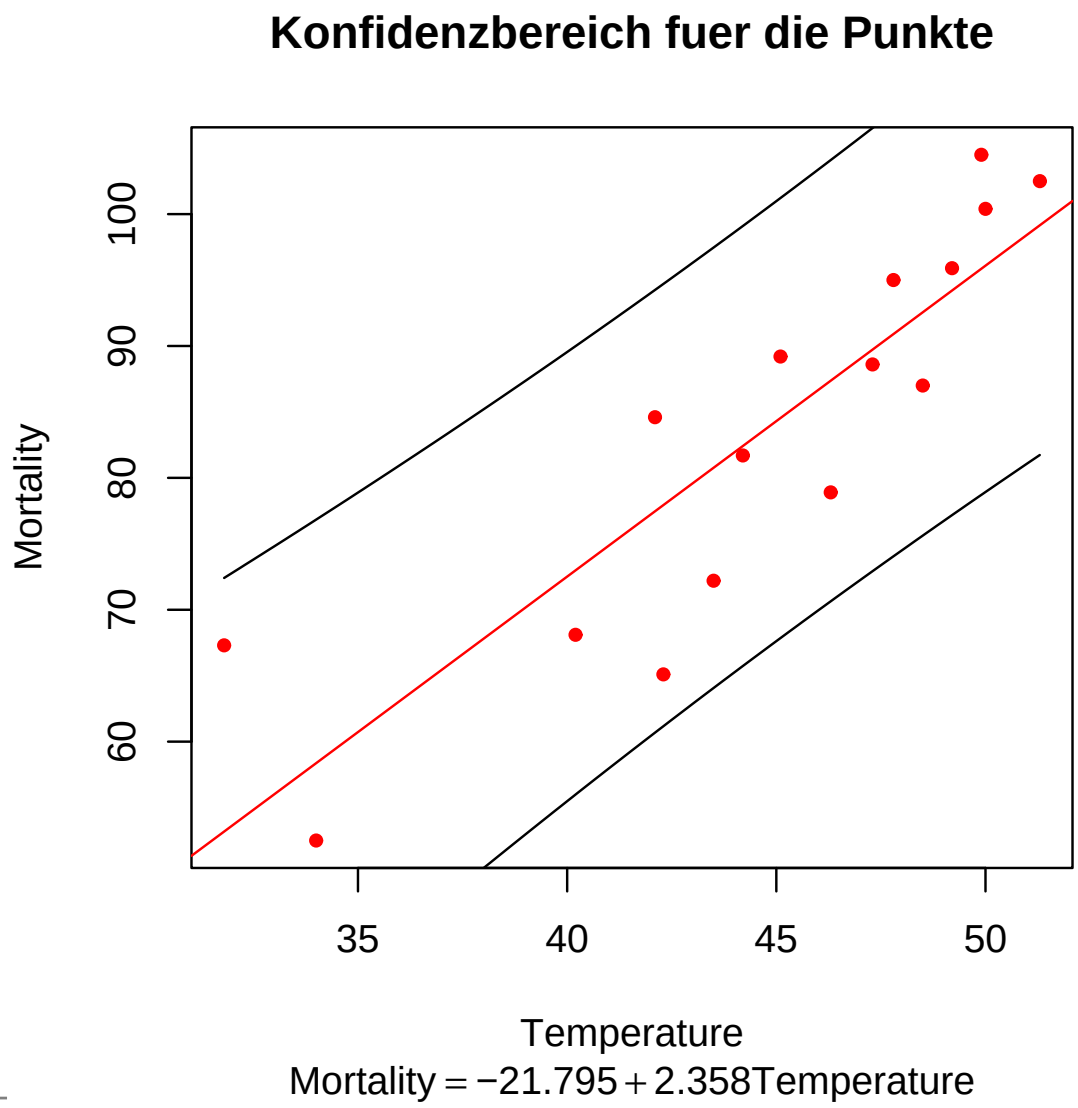
Computerausgabe

```
> model <- lm(Mortality ~ Temperature, data = Brustkrebs)
> model

Call:
lm(formula = Mortality ~ Temperature, data = Brustkrebs)

Coefficients:
(Intercept)  Temperature
    -21.795        2.358

> anova(model)

Analysis of Variance Table

Response: Mortality
            Df  Sum Sq Mean Sq F value    Pr(>F)
Temperature  1 2599.53 2599.53  45.669 9.202e-06 ***
Residuals   14  796.91   56.92
```

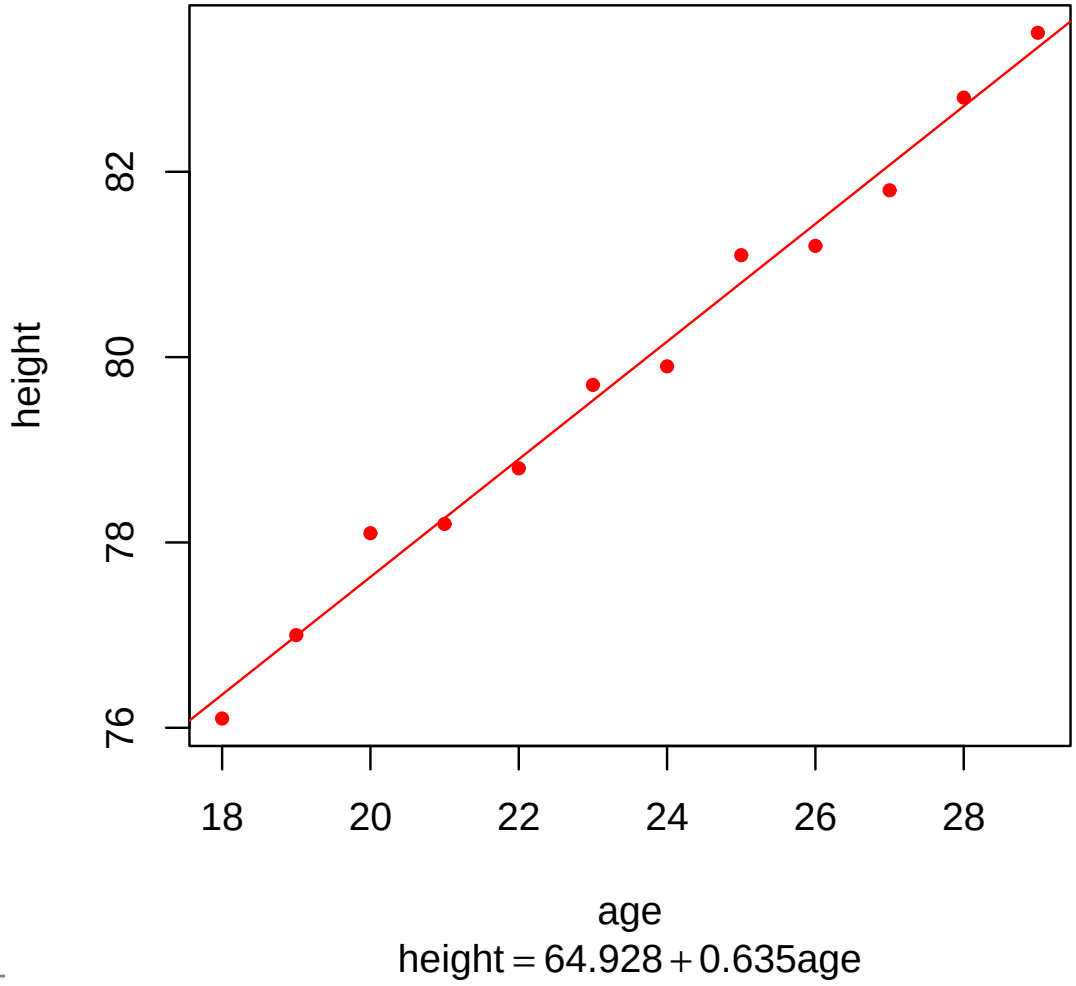
Konfidenzintervall



Vorhersageintervall



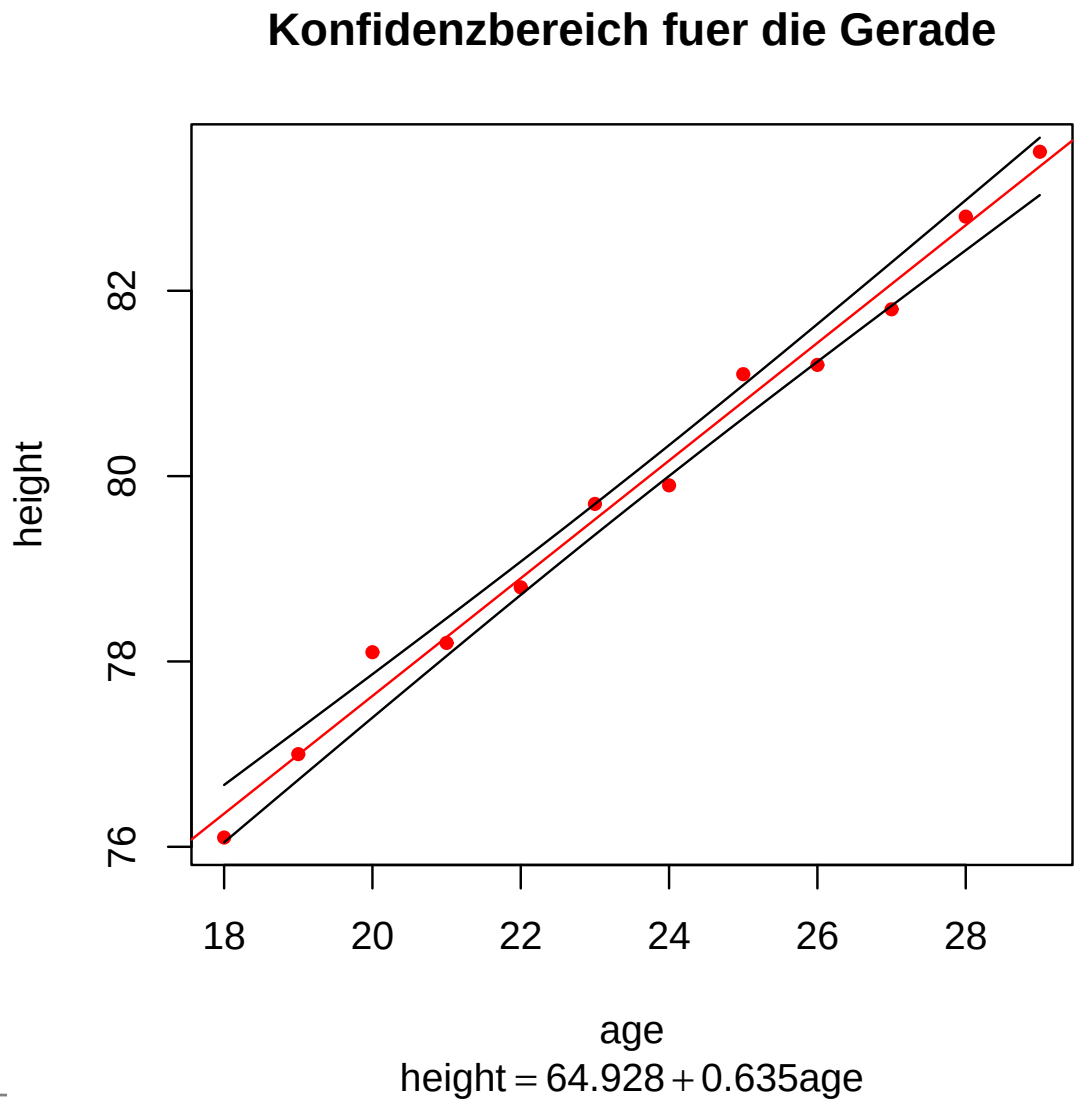
Beispiel: Wachstum



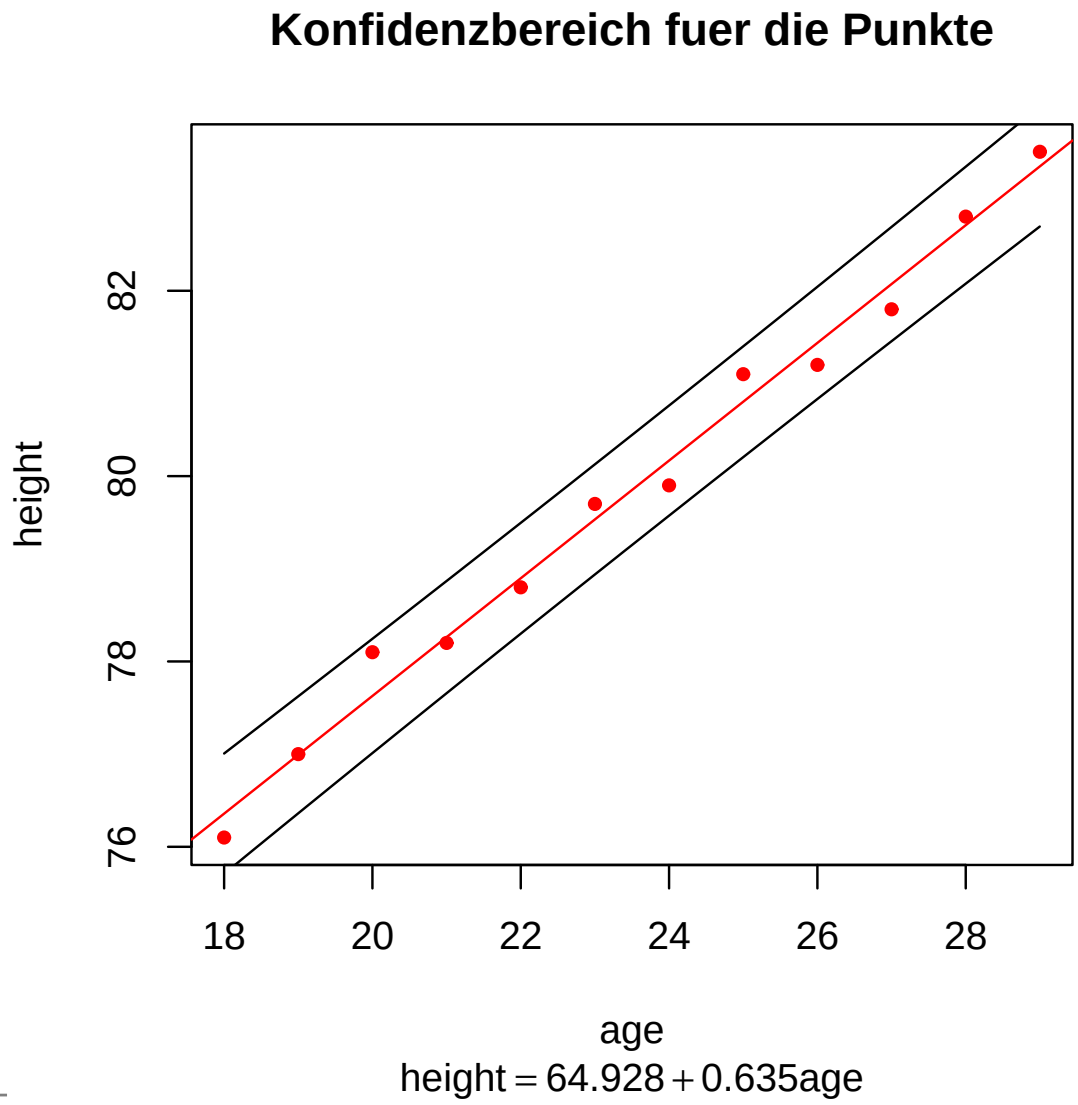
Computerausgabe

```
> model <- lm(height ~ age, data = Wachstum)
> model

Call:
lm(formula = height ~ age, data = Wachstum)

Coefficients:
(Intercept)          age
     64.928        0.635

> anova(model)

Analysis of Variance Table

Response: height
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 57.655  57.655     880 4.428e-11 ***
Residuals 10  0.655   0.066
```

Konfidenzintervall

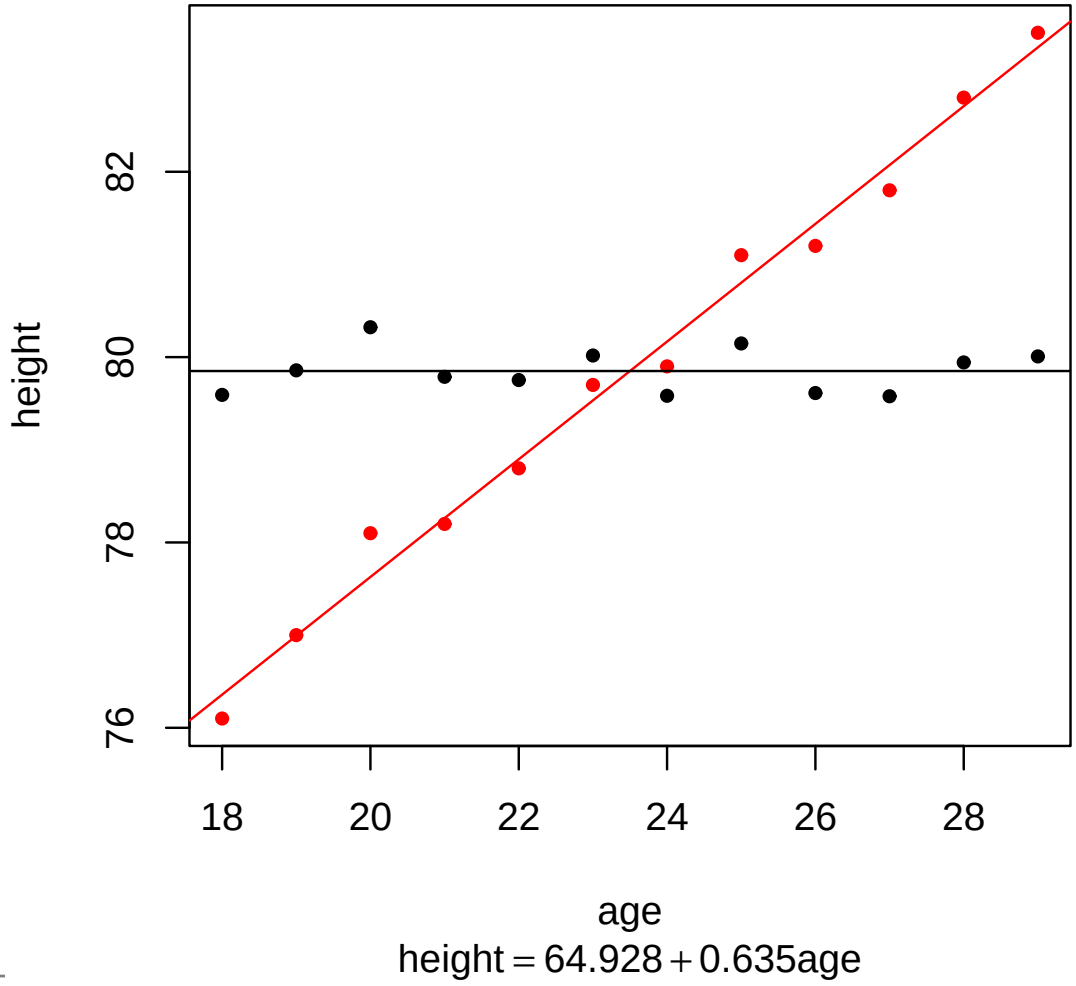


Vorhersageintervall

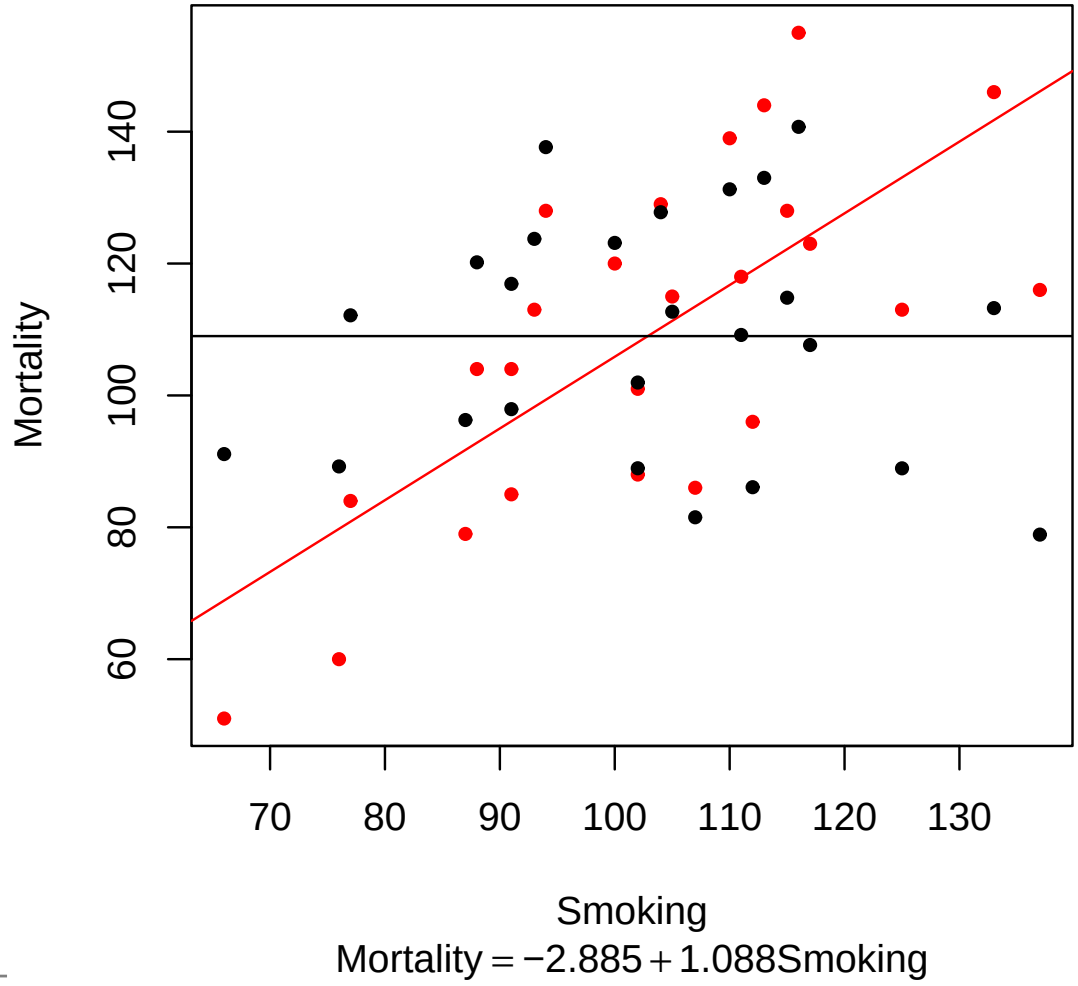


Stärke des Zusammenhangs bewerten: Korrelation und R^2

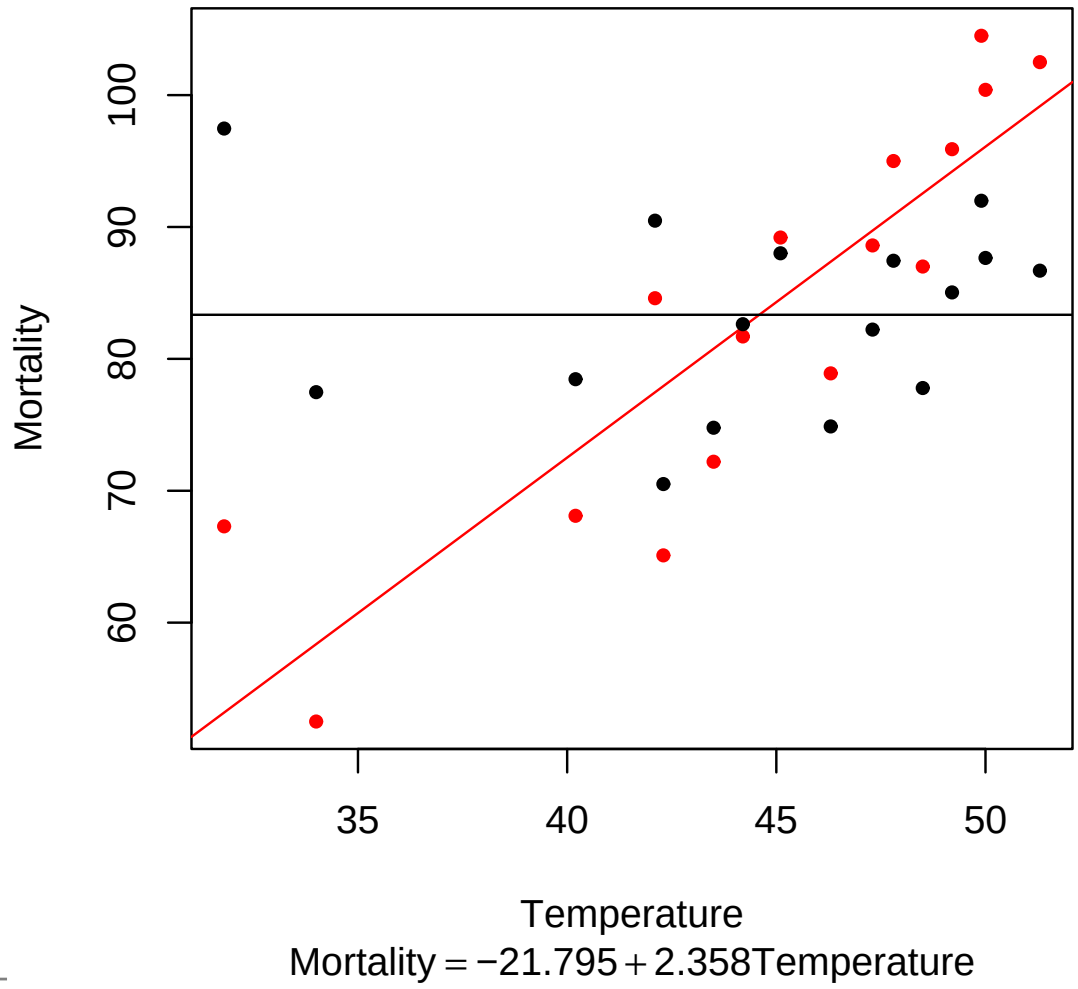
Residualstreuung



Residualstreuung



Residualstreuung



R^2

Def: Das Bestimmtheitsmaß R^2 ist gegeben durch

$$R^2 = \frac{\hat{\text{var}}(Y) - \hat{\text{var}}(r)}{\hat{\text{var}}Y}$$

also die erklärte Streuung, geteilt durch die Gesamtstreuung im Datensatz.
Es gilt:

$$R^2 = \hat{\text{cor}}(X, Y)^2$$

Da R^2 allgemein für ein Model definiert ist kann es weiter eingesetzt werden, als die Pearson Korrelation.

Interpretation von R^2

- $R^2 \in [0, 1]$

Interpretation von R^2

- $R^2 \in [0, 1]$
- $R^2 = 0 \Leftrightarrow$ keine Abhängigkeit erkennbar.

Interpretation von R^2

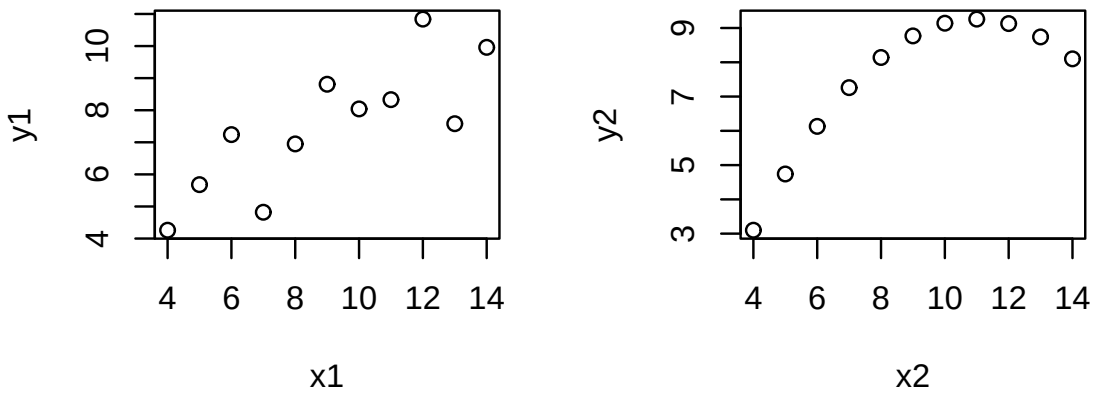
- $R^2 \in [0, 1]$
- $R^2 = 0 \Leftrightarrow$ keine Abhängigkeit erkennbar.
- $R^2 = 1 \Leftrightarrow \hat{Y}_i = Y_i$, Modell erklärt Daten perfekt.

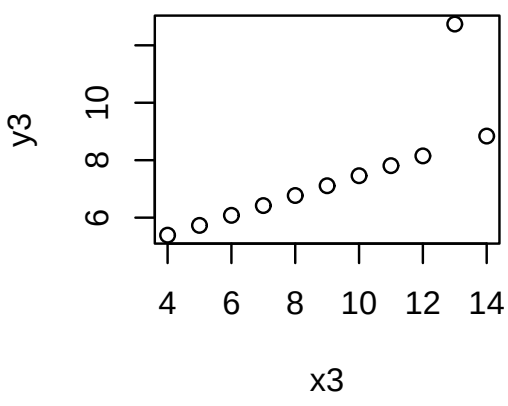
Computerausgabe

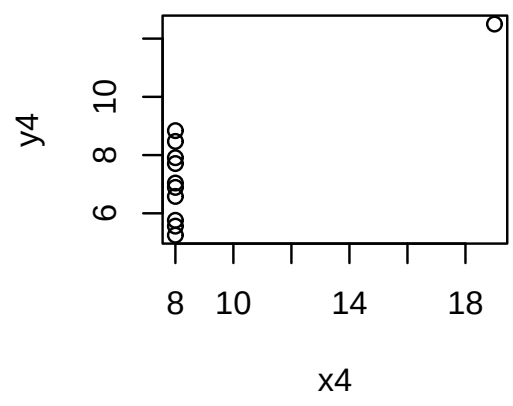
```
> R2 <- function(m) var(predict(m))/var(predict(m) +
+     resid(m))
> R2(lm(height ~ age, data = Wachstum))

[1] 0.988764

> R2(lm(Mortality ~ Smoking, data = Rauchen))

[1] 0.5129995

> R2(lm(Mortality ~ Temperature, data = Brustkrebs))

[1] 0.7653702
```

Der Blinde Fleck der Regression

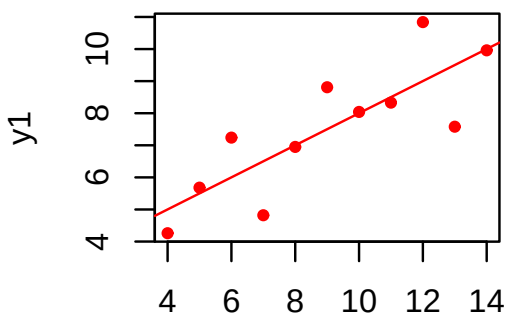
Das Anscombe Quartet

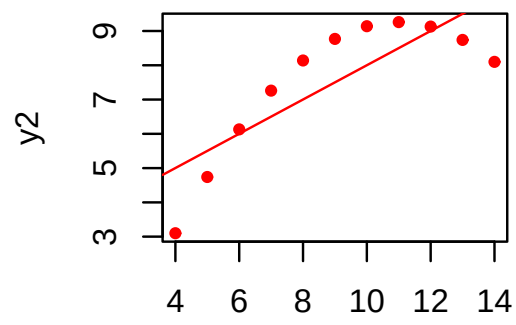


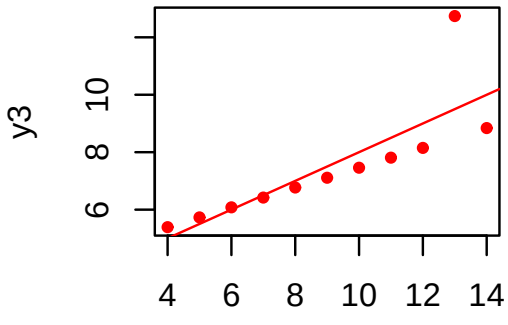


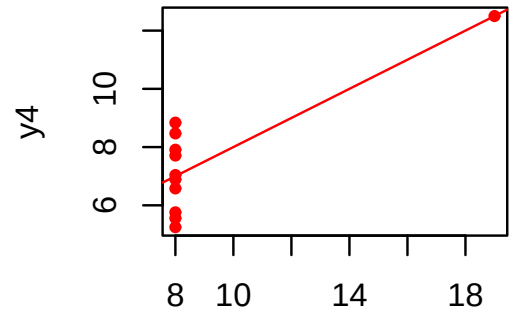


Regressionsgeraden

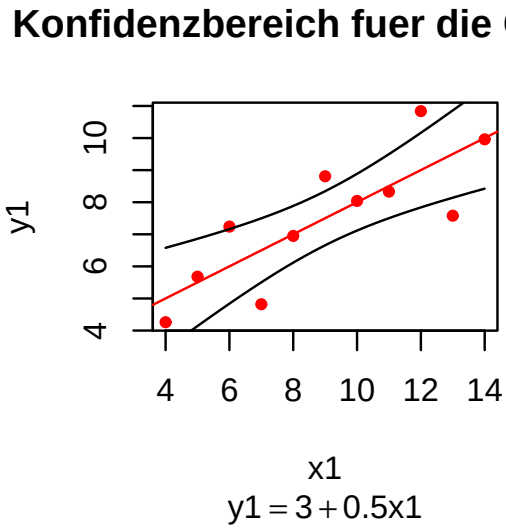


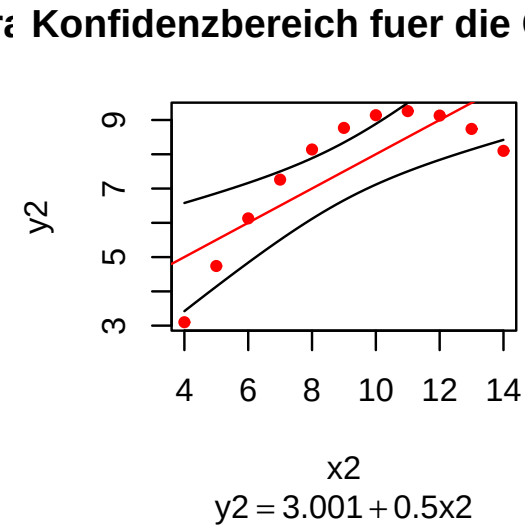


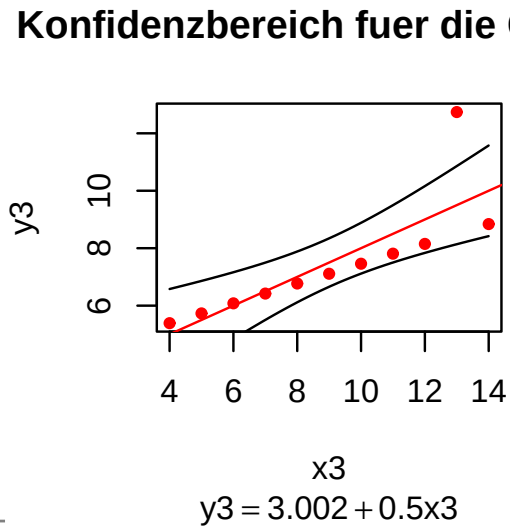


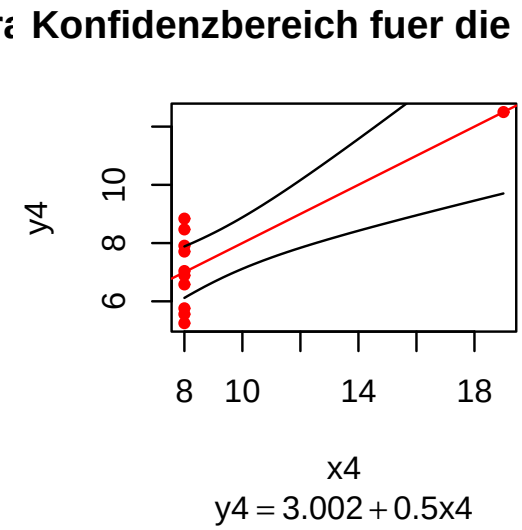


Konfidenzintervalle

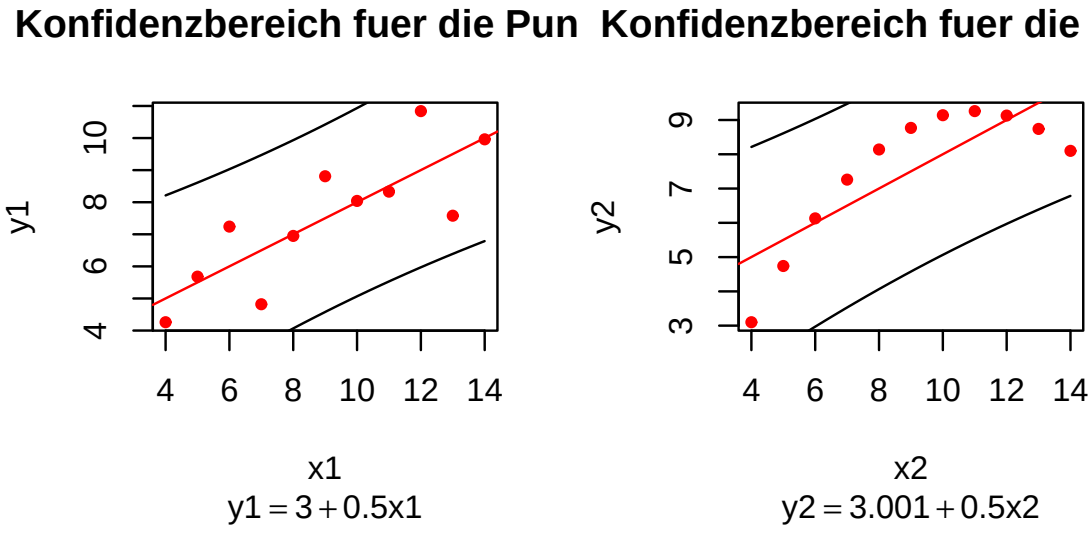


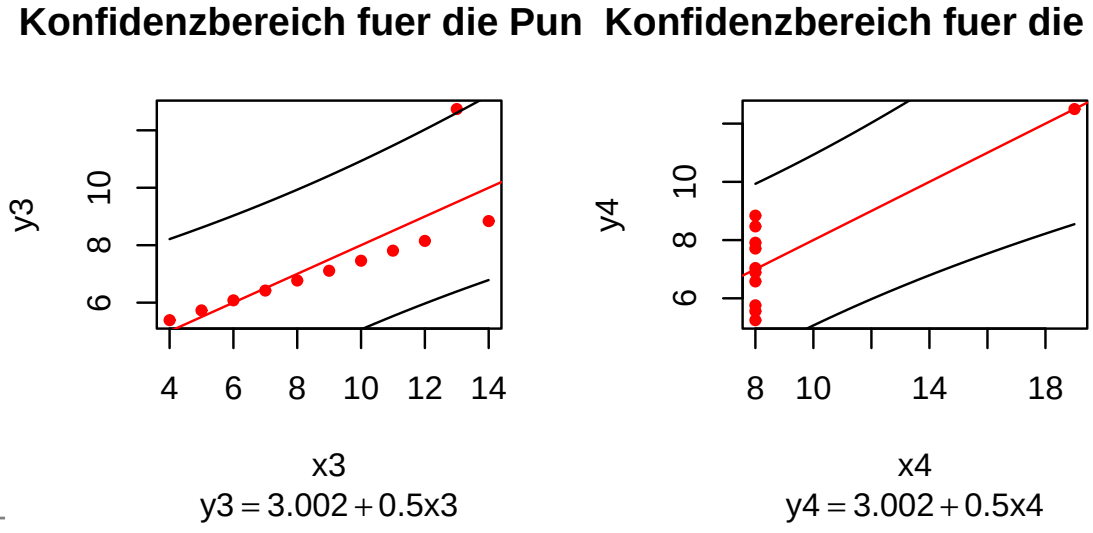






Vorhersageintervalle



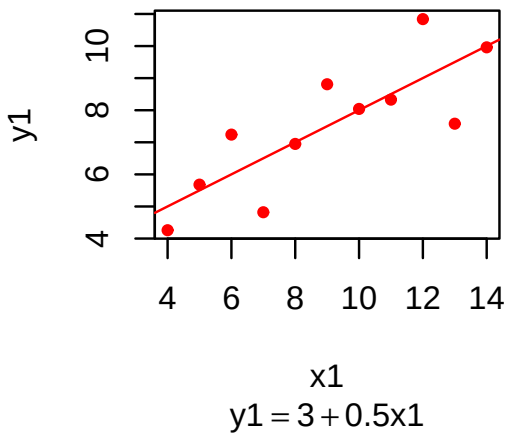


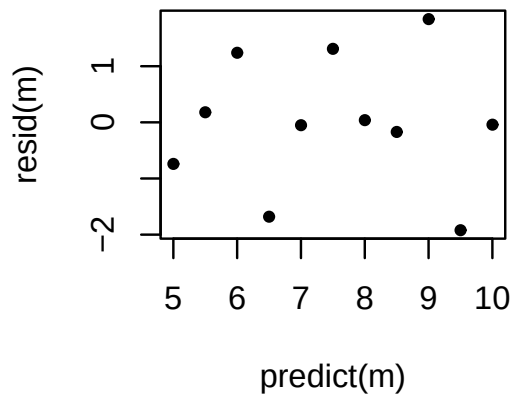
Was war das Problem?

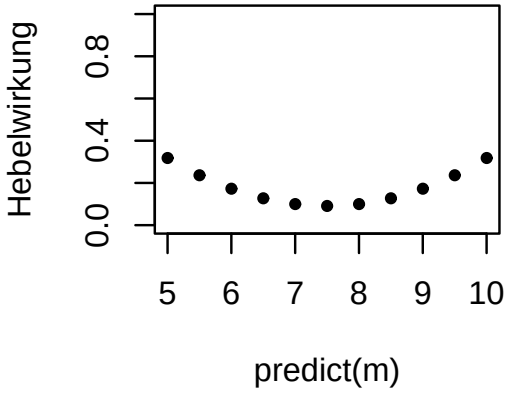
- Voraussetzungen nicht erfüllt
- Ausreißer in den Daten
- sehr einflußreiche Punkte

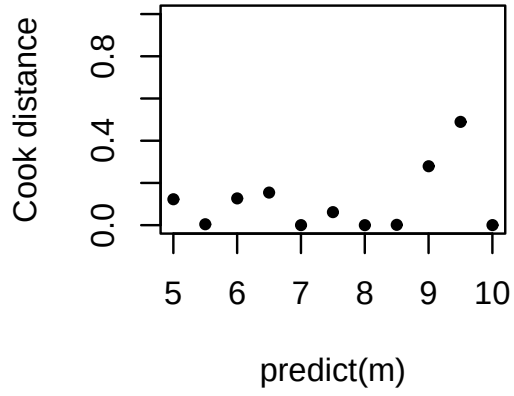
Regressionsdiagnostik

Anscombe 1


y1
4
6
8
10
4
6
8
10
12
14
x1
y1 = 3 + 0.5x1


resid(m)
−2
0
1
5
6
7
8
9
10
predict(m)


Hebelwirkung
0.0
0.4
0.8
5
6
7
8
9
10
predict(m)


Cook distance
0.0
0.4
0.8
5
6
7
8
9
10
predict(m)

residuals vs. predicted

- Hängt die Streuung vom Vorhersagewert ab?
- Gibt es extrem große Residuen?
- Treten die Vorhersagewerte gleichmäßig auf?
- Treten die Residuen gleichmäßig auf?
- Gibt es Strukturen?

Hebelwirkung

$\theta_i =$ Änderung von $\hat{Y}_i$ pro Änderung von Y_i

"Wer weniger Kraft braucht sitzt am längeren Hebel"

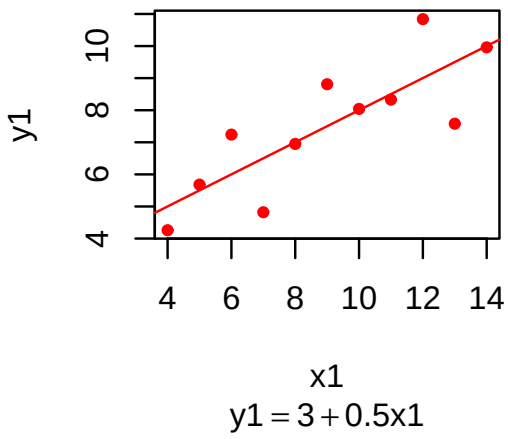
- Mißt die potentielle Wirkung eines Ausreißers im Regressant an dieser Stelle.
- Zeigt wie wichtig eine Beobachtung für die Schätzung ist.
- Große Werte deuten auf Unzuverlässige Schätzungen hin.

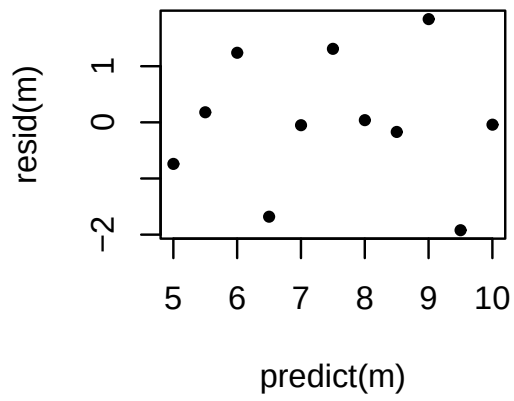
Cook's Distanz

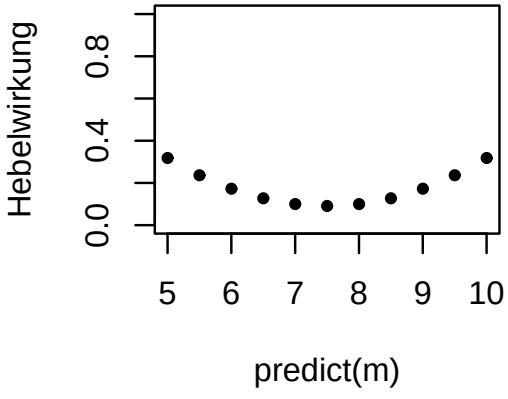
$c_i =$ Maß für tatsächlicher Einfluß der Beobachtung Y_i

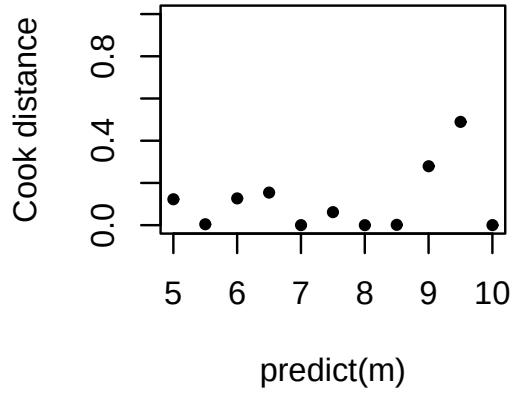
- Große Werten bedeuten, dass die Beobachtung vom dem abweicht, was die anderen Werte über diese Stelle aussagen würden.

Anscombe 1

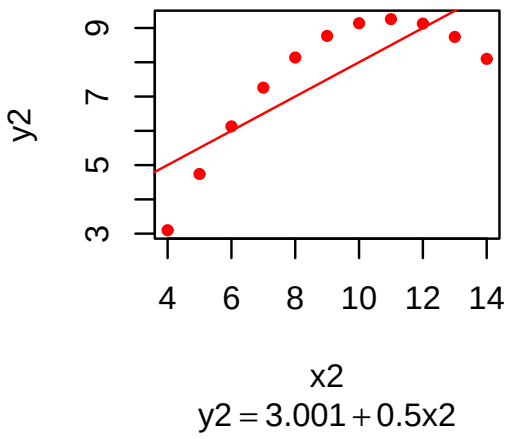

y1
4
6
8
10
4
6
8
10
12
14
x1
y1 = 3 + 0.5x1

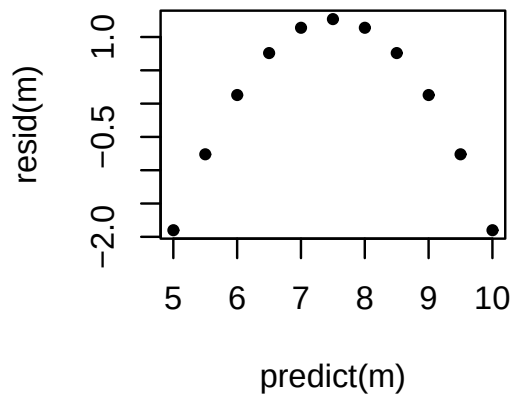

resid(m)
−2
0
1
5
6
7
8
9
10
predict(m)

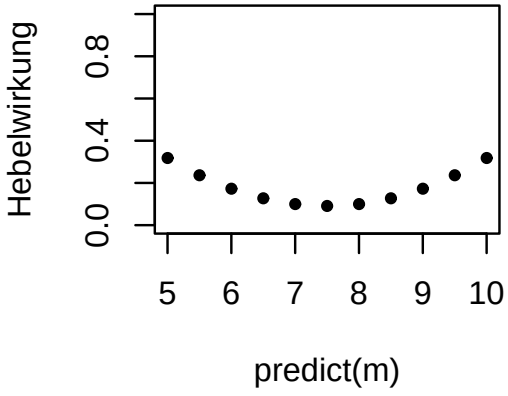

Hebelwirkung
0.0
0.4
0.8
5
6
7
8
9
10
predict(m)

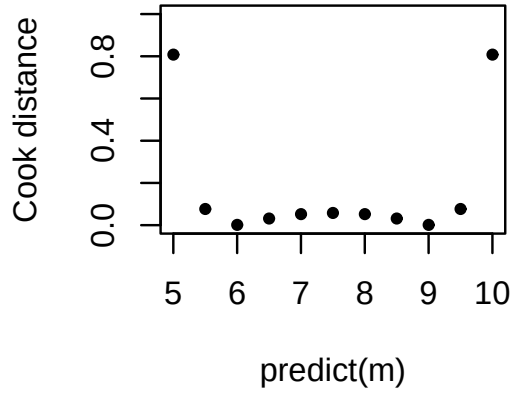

Cook distance
0.0
0.4
0.8
5
6
7
8
9
10
predict(m)

Anscombe 2

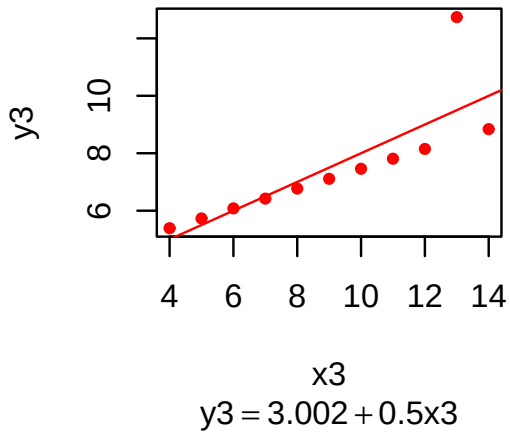

y2
3
5
7
9
4
6
8
10
12
14
x2
y2 = 3.001 + 0.5x2

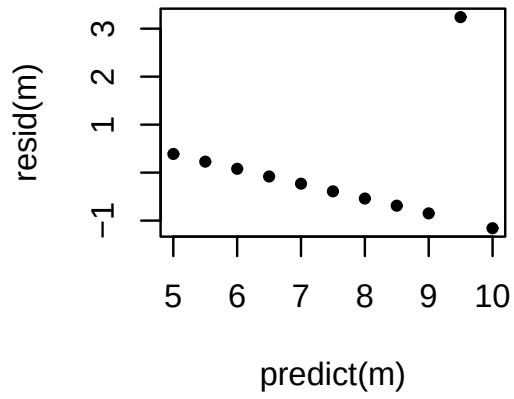

resid(m)
−2.0
−0.5
1.0
5
6
7
8
9
10
predict(m)

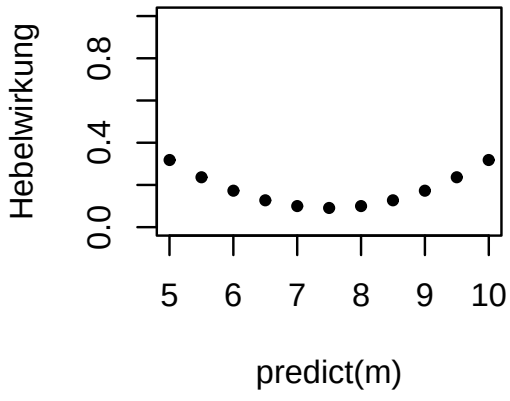

Hebelwirkung
0.0
0.4
0.8
5
6
7
8
9
10
predict(m)

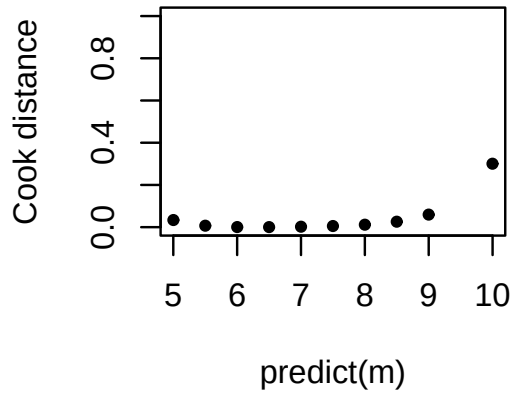

Cook distance
0.0
0.4
0.8
5
6
7
8
9
10
predict(m)

Anscombe 3

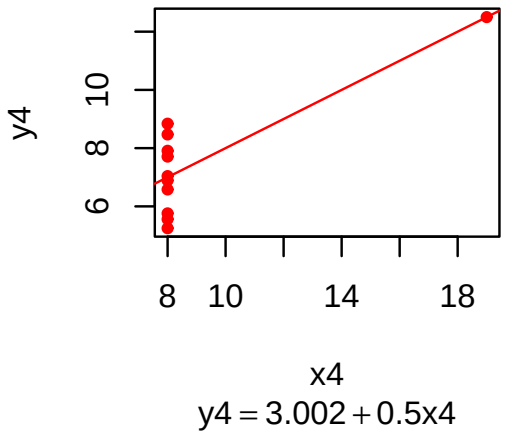

y3
6
8
10
4
6
8
10
12
14
x3
y3 = 3.002 + 0.5x3

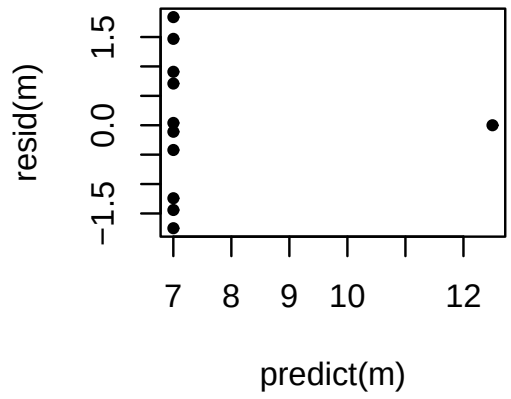

resid(m)
−1
1
2
3
5
6
7
8
9
10
predict(m)

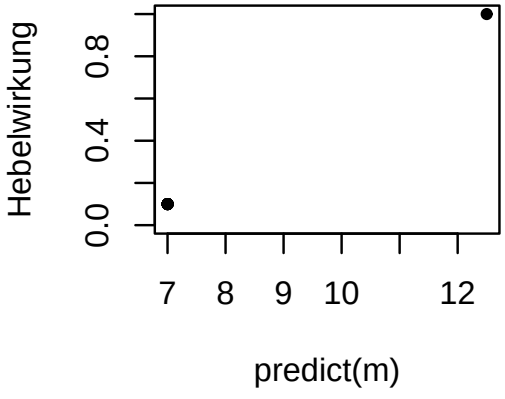

Hebelwirkung
0.0
0.4
0.8
5
6
7
8
9
10
predict(m)

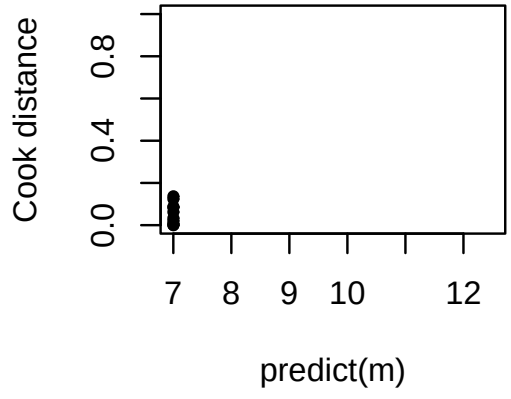

Cook distance
0.0
0.4
0.8
5
6
7
8
9
10
predict(m)

Anscombe 4

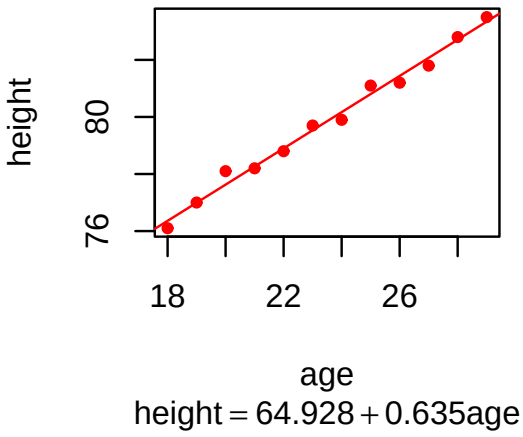

y4
6
8
10
8
10
14
18
x4
y4 = 3.002 + 0.5x4

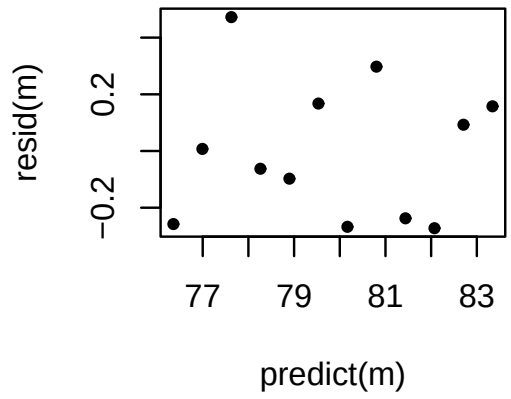

resid(m)
−1.5
0.0
1.5
7
8
9
10
12
predict(m)

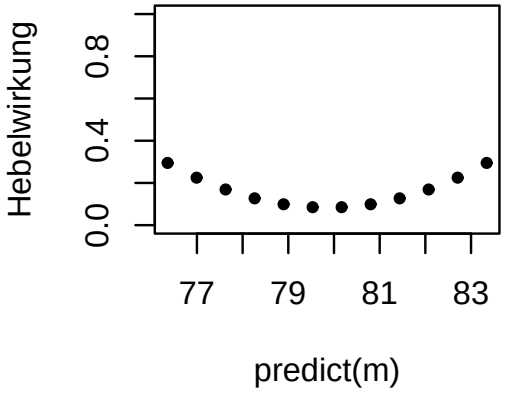

Hebelwirkung
0.0
0.4
0.8
7
8
9
10
12
predict(m)

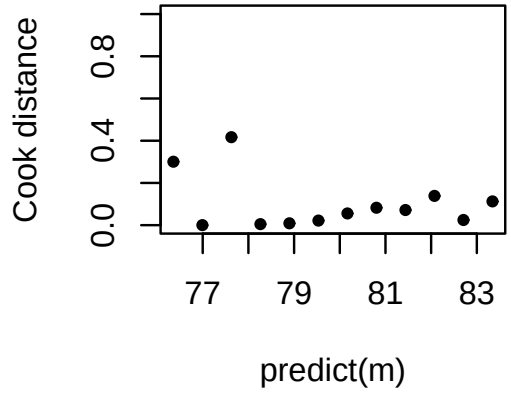

Cook distance
0.0
0.4
0.8
7
8
9
10
12
predict(m)

Wachstumsdaten

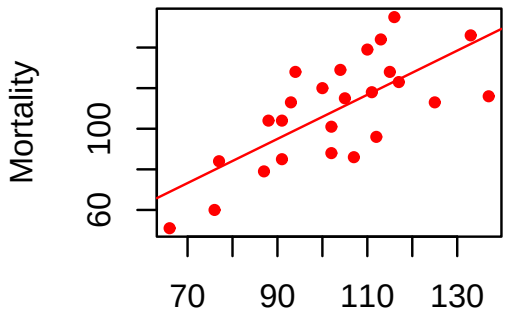




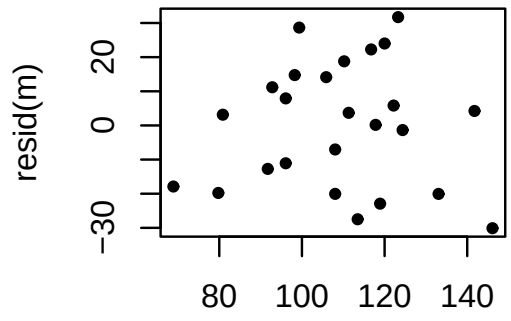


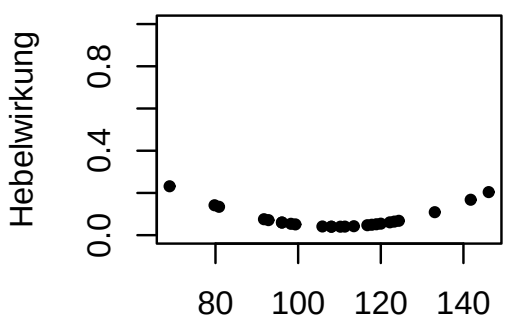


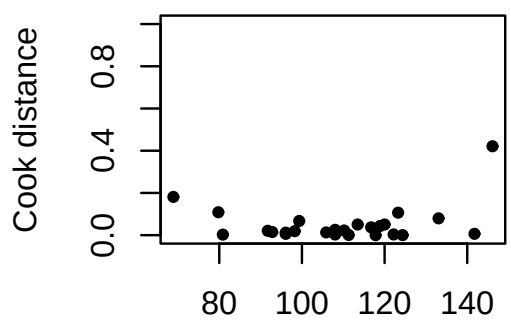
Raucherdaten



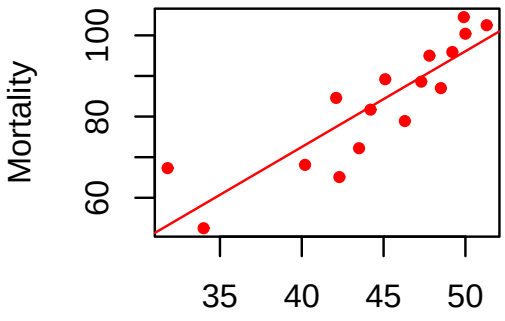
Mortality = −2.885 + 1.088Smoking



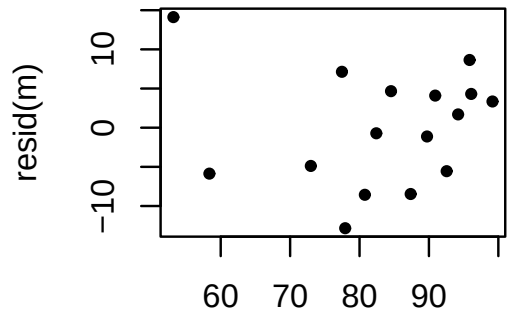


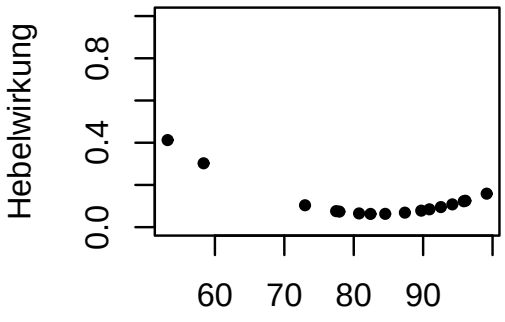


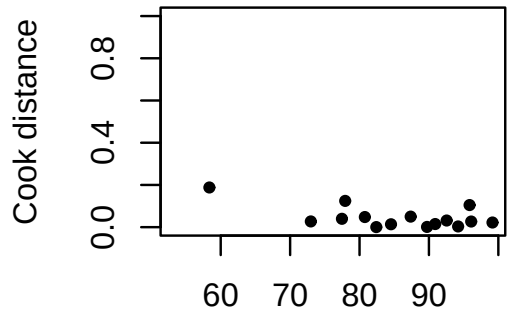
Brustkrebsdaten



Mortality = −21.795 + 2.358Temperatu







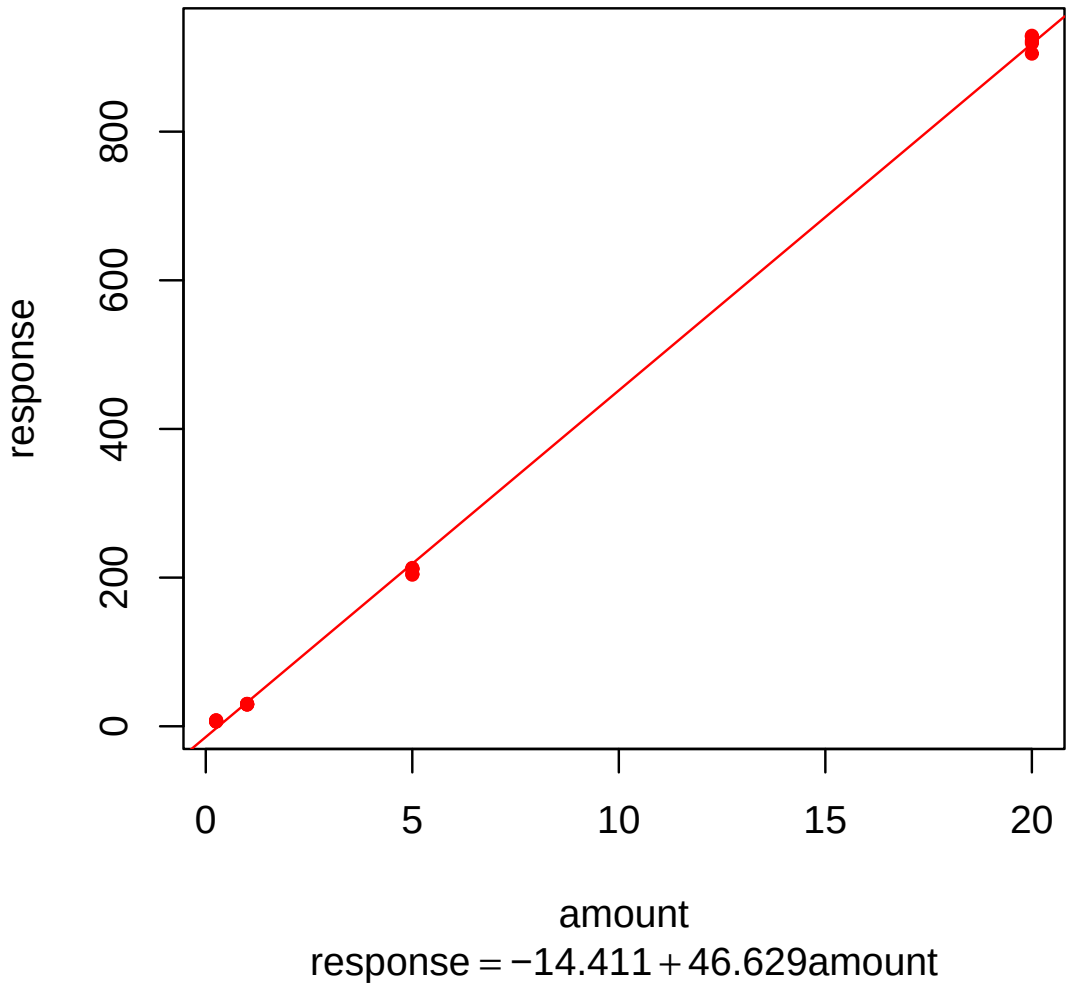
Robuste Schätzung

Idee der robusten Schätzung

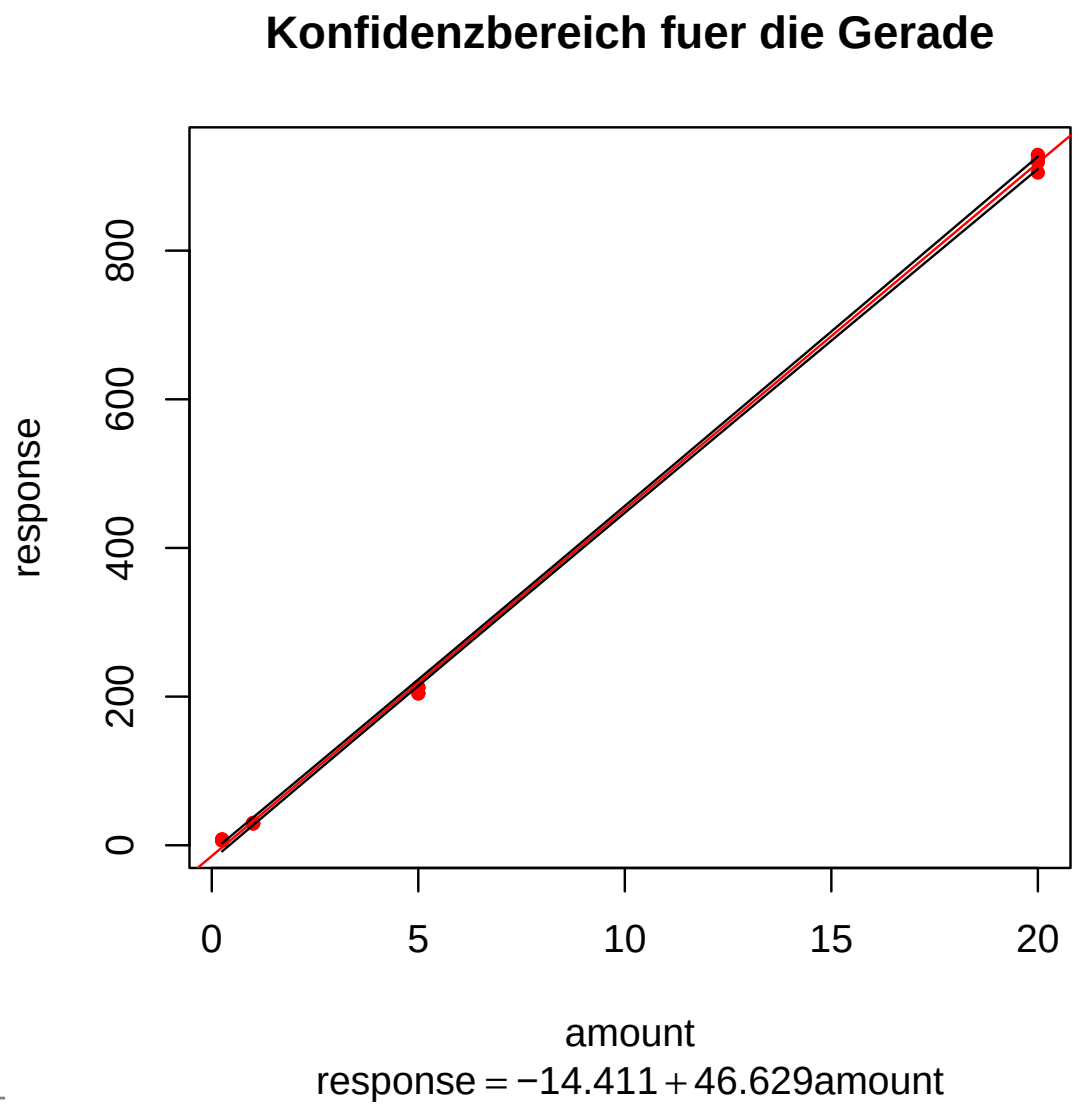
- Minimiere den mittleren quadratischen Abstand zu den $n - k$-nächsten Punkten.
- Dann können k-Ausreißer die Gerade nicht stark beeinflussen.

Weitere Beispiele

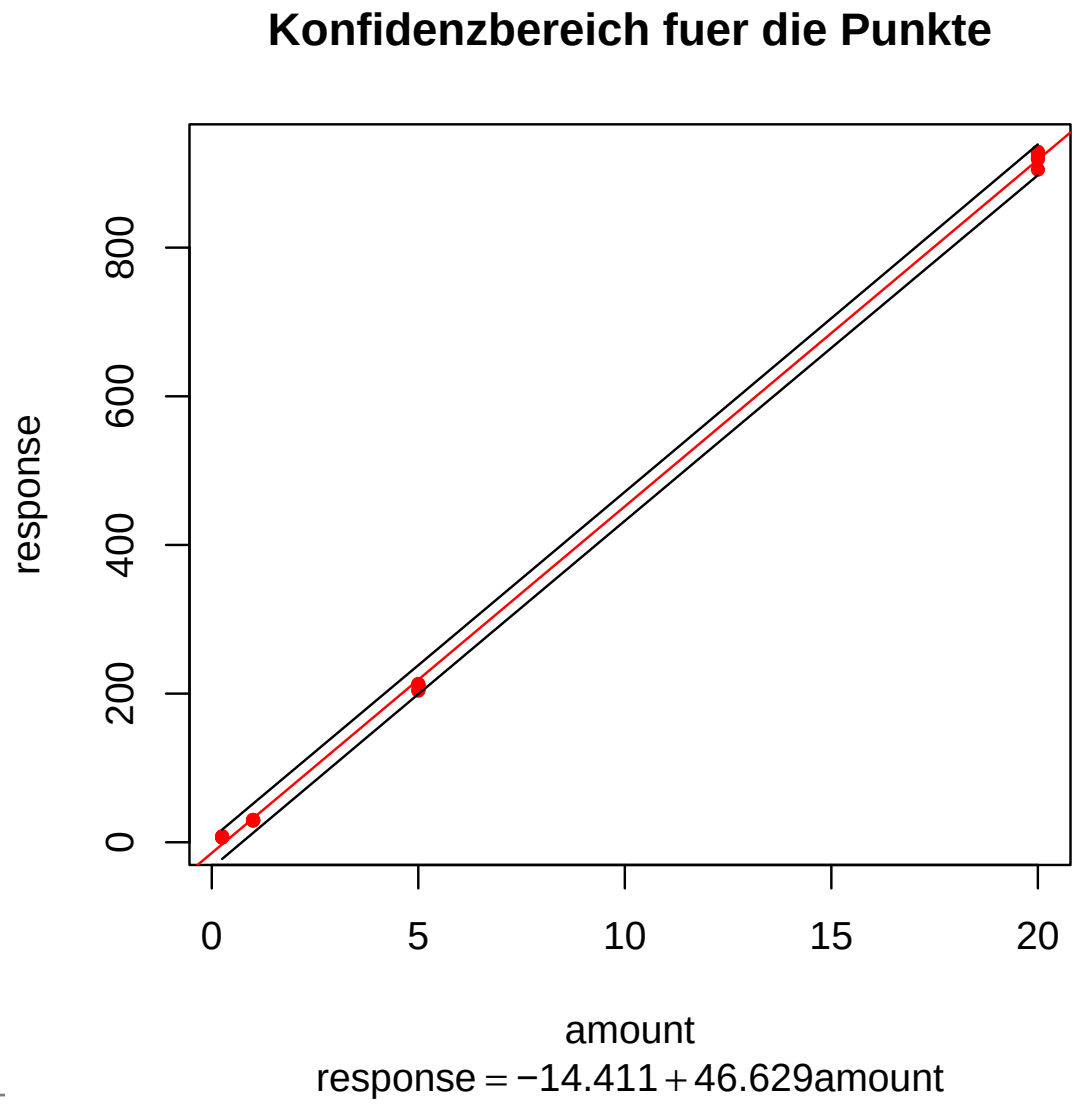
Chromatograph



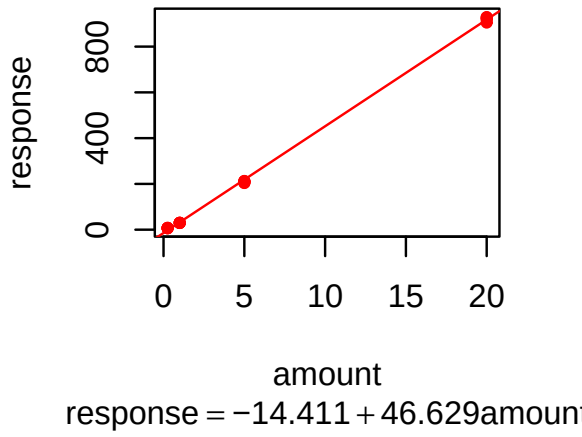
Chromatograph

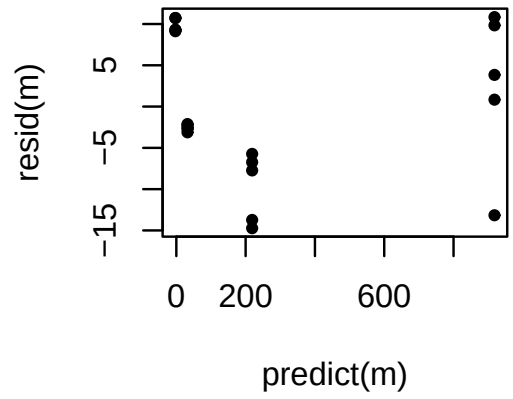


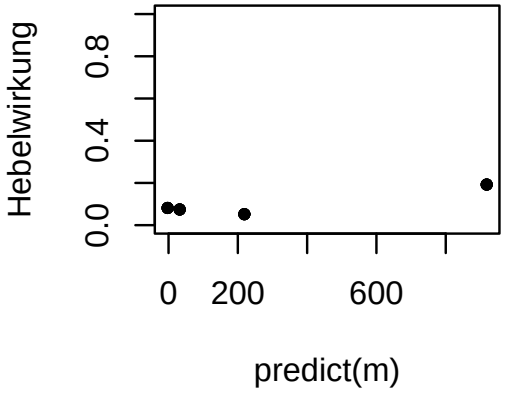
Chromatograph

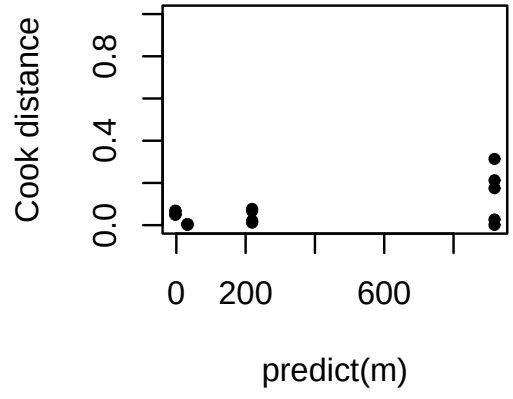


Chromatograph









Luftqualität

```
> pairs(airquality)
> pairs(airquality)
> showGeraden(Ozone ~ Solar.R, data = airquality)
> showGeraden(Ozone ~ Solar.R, data = airquality,
+     diag = T)
> pairs(airquality)
> par(mfrow = c(1, 1))
> showGeraden(Ozone ~ Temp, data = airquality)
> showGeraden(Ozone ~ Temp, data = airquality, alpha =
+     diag = T)
> par(mfrow = c(1, 1))
```