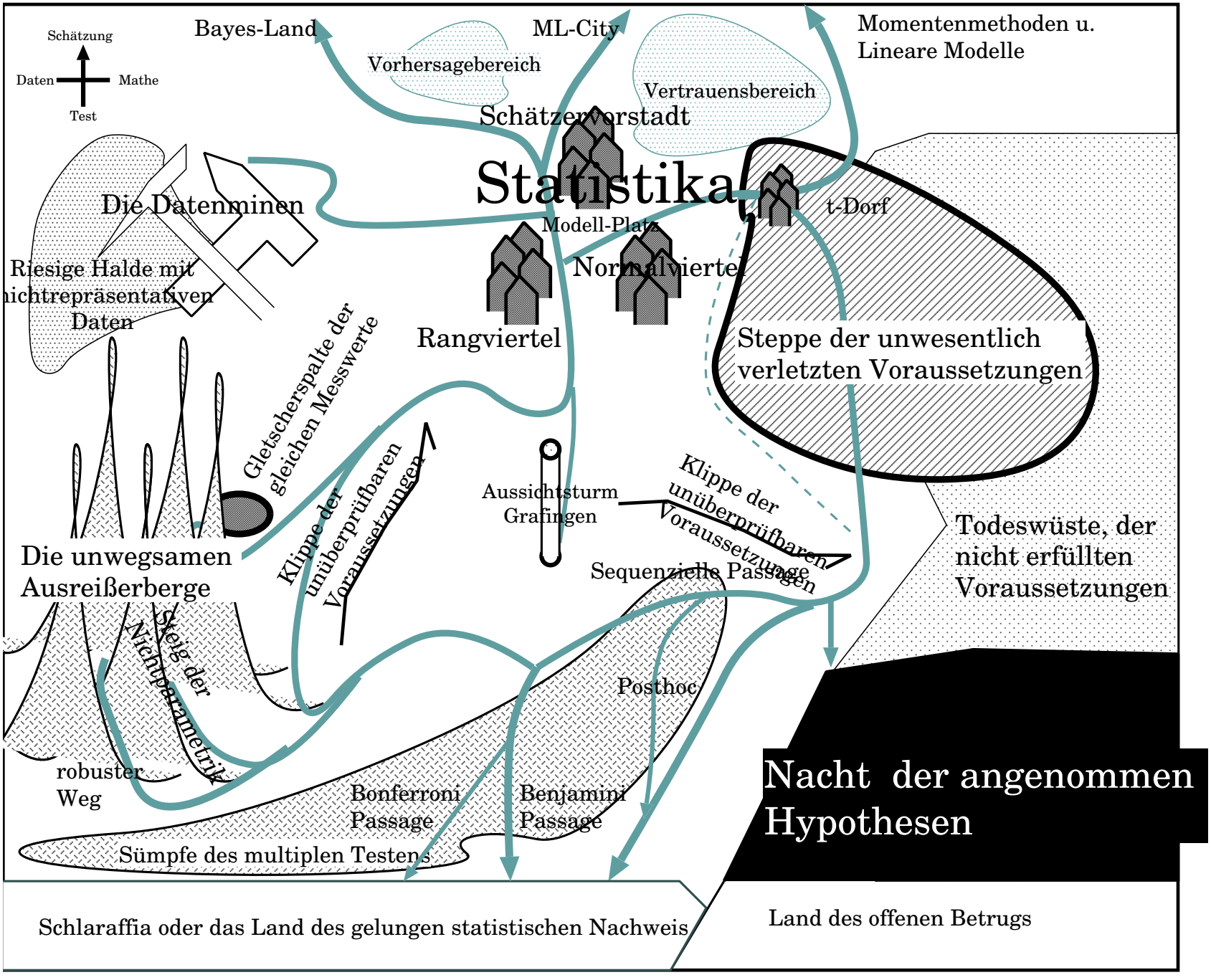
Datenanalyse und Statistik

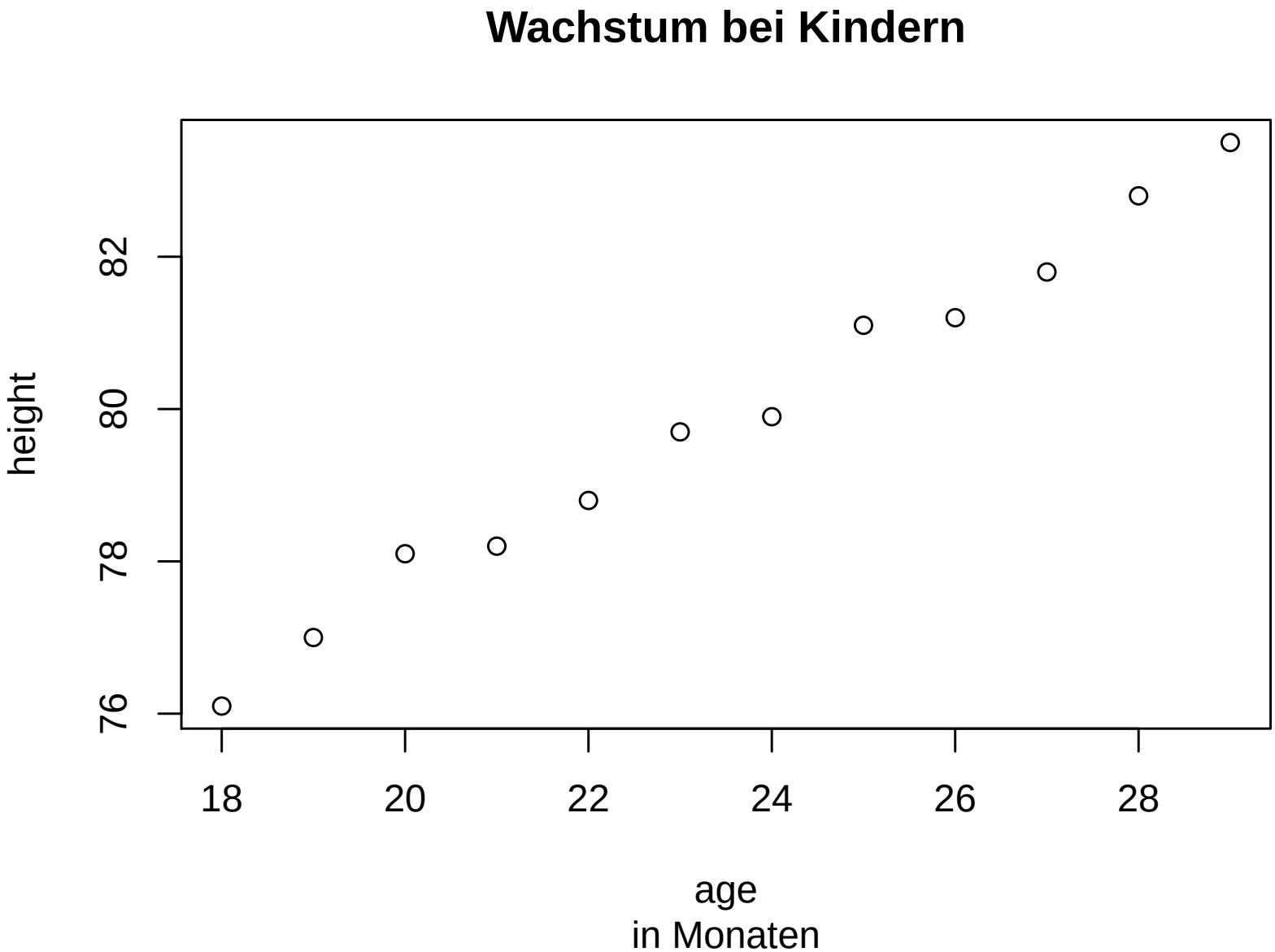
Vorlesung 7 (Lineare Regression)

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

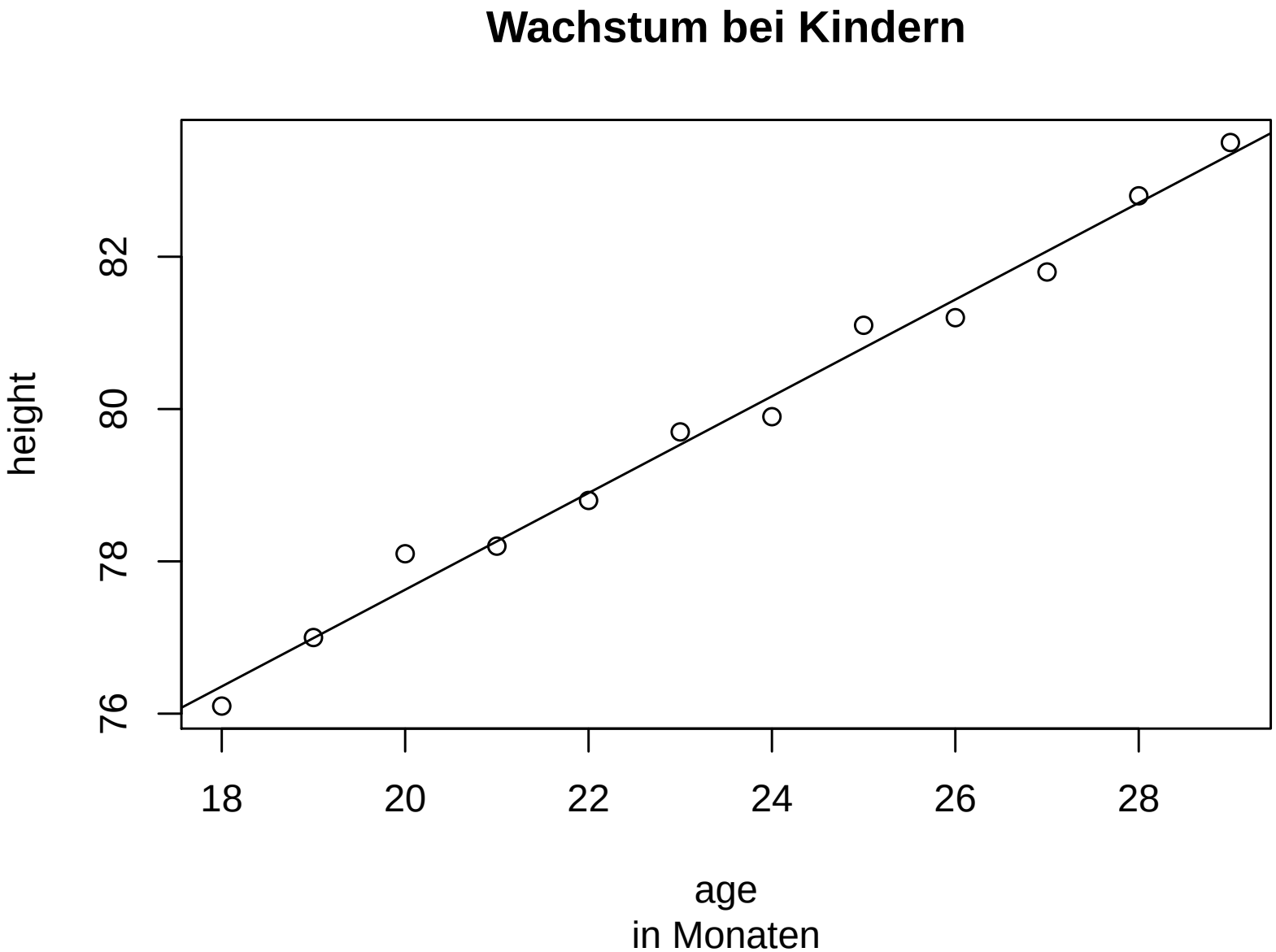
2. Dezember 2019


Schätzung
Daten
Mathe
Test
Bayes-Land
ML-City
Vorhersagebereich
Vertrauensbereich
Momentenmethoden u.
Lineare Modelle
Schätzervorstadt
Statistika
Die Datenminen
Modell-Platz
t-Dorf
Riesige Halde mit
nichtrepräsentativen
Daten
Normalviertel
Rangviertel
Steppe der unwesentlich
verletzten Voraussetzungen
Gletscherspalte der
gleichen Messwerte
Klippe der
unüberprüfbaren
Voraussetzungen
Aussichtsturm
Grafingen
Klippe der
unüberprüfbaren
Voraussetzungen
Todeswüste, der
nicht erfüllten
Voraussetzungen
Die unwegsamen
Ausreißerberge
Sequenzielle Passage
Steig der
Nichtparametrik
Posthoc
robuster
Weg
Bonferroni
Passage
Benjamini
Passage
Sümpfe des multiplen Testens
Nacht der angenommen
Hypothesen
Schlaraffia oder das Land des gelungen statistischen Nachweis
Land des offenen Betrugs

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**
  weil $a = a + b0$ der Wert der Geraden bei $X = 0$ ist.
- 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**
  weil $a = a + b0$ der Wert der Geraden bei $X = 0$ ist.
- b **heißt Steigung**
  weil b die Steigung der Geraden $a + bX$ ist.
- **Die X_i heißen Regressor**
  weil X die das ansteigen bewirkt.

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.
- b **heißt Steigung**
  weil b die Steigung der Geraden $a + bX$ ist.
- **Die X_i heißen Regressor**
  weil X die das ansteigen bewirkt.
- **Die Y_i heißen Regressant**
  weil Y erhöht wird.

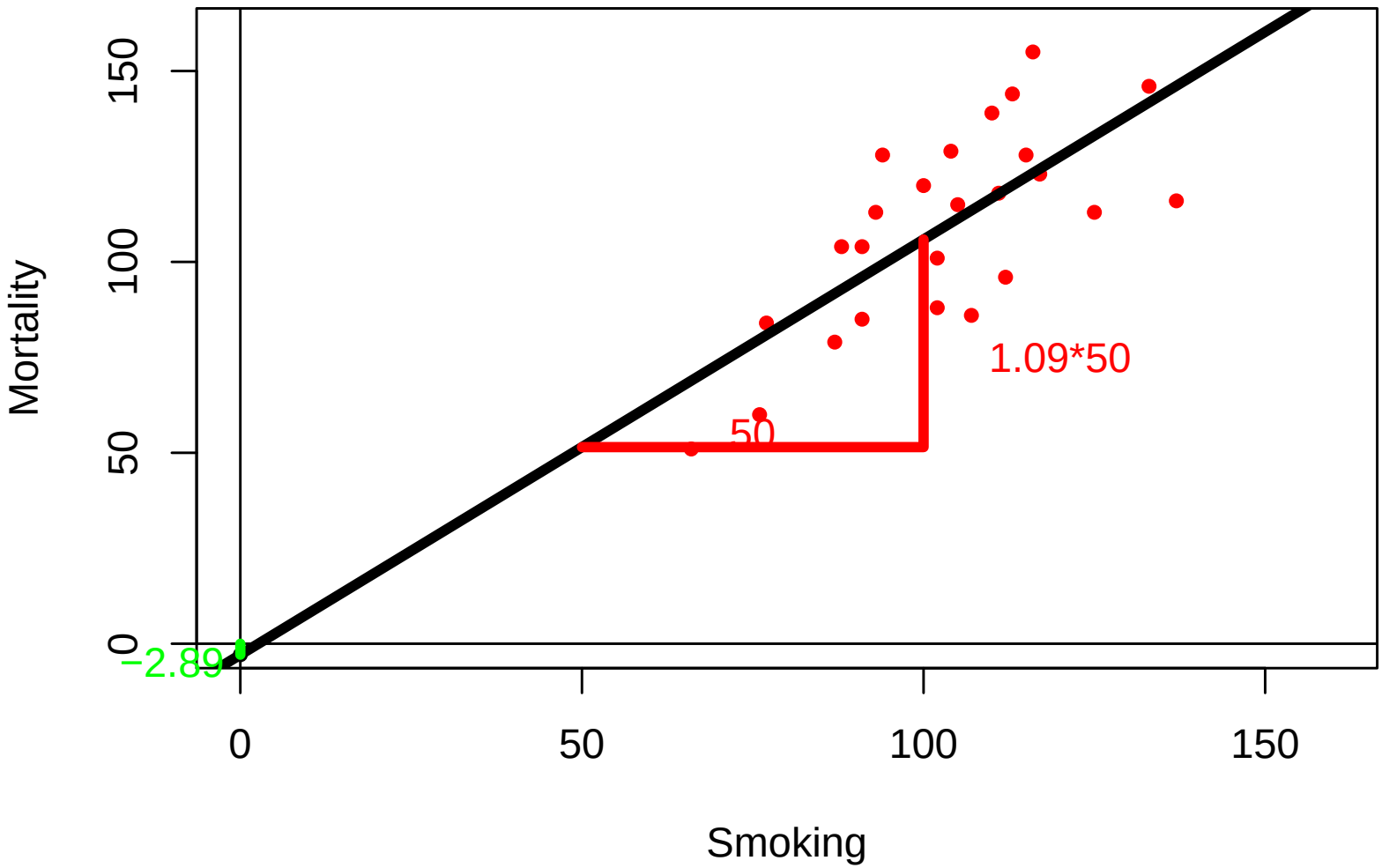
Definitionen

Das Modell

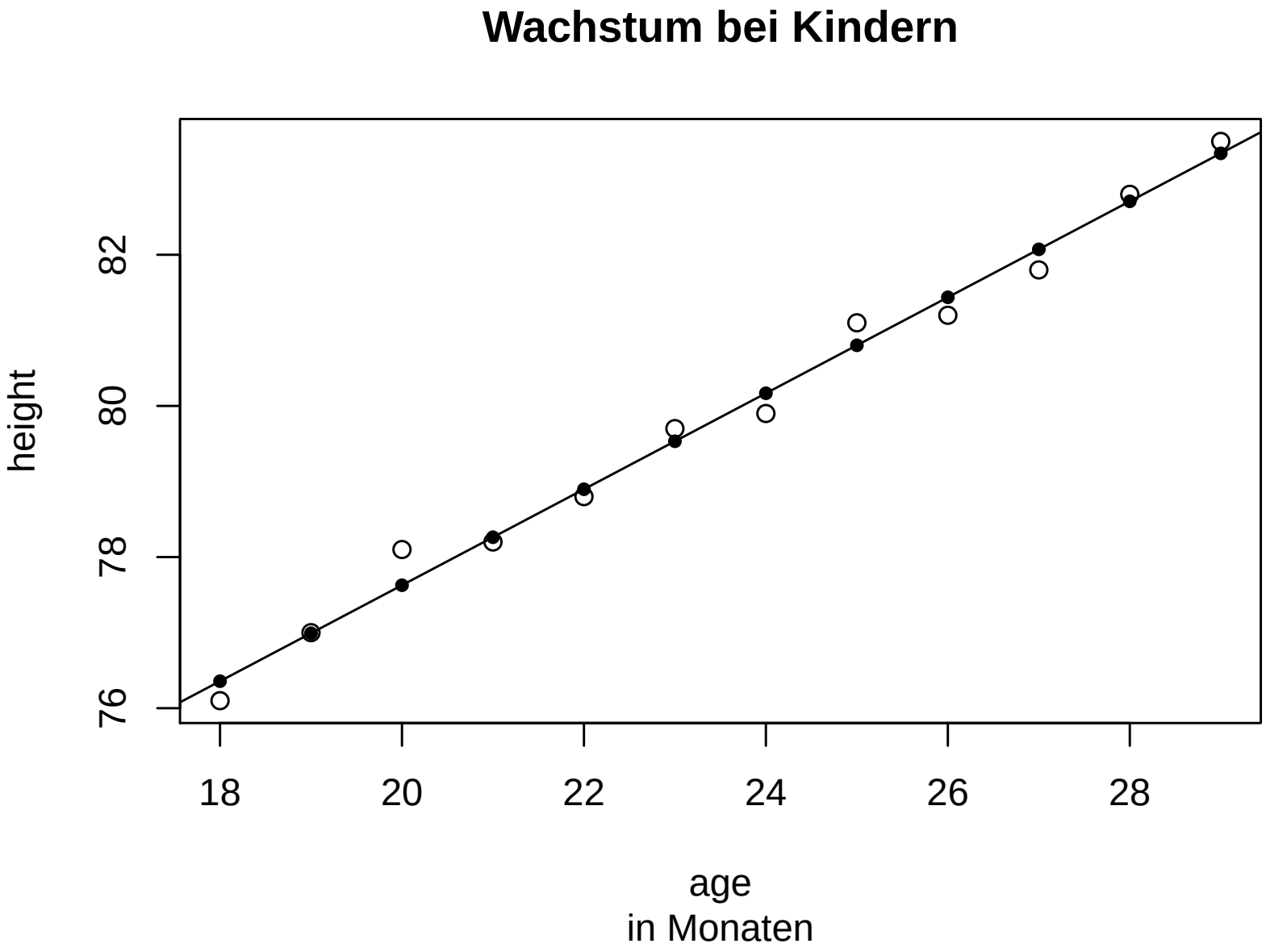
$$Y_i = a + bX_i + \epsilon_i$$

- a **heißt Achsenabschnitt**
  weil $a = a + b0$ der Wert der Geraden bei $X = 0$ ist.
- b **heißt Steigung**
  weil b die Steigung der Geraden $a + bX$ ist.
- **Die X_i heißen Regressor**
  weil X die das ansteigen bewirkt.
- **Die Y_i heißen Regressant**
  weil Y erhöht wird.
- **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

- dass a und b nur geschätzt werden können und dass
- die Gerade nur bis auf den Fehler ϵ die Werte vorhersagt

Die Vorhersagewerte

Das Regressionsmodel:

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

Wenn wir von den Unzulänglichkeiten

- dass a und b nur geschätzt werden können und dass
- die Gerade nur bis auf den Fehler ϵ die Werte vorhersagt

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

- dass a und b nur geschätzt werden können und dass
- die Gerade nur bis auf den Fehler ϵ die Werte vorhersagt

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

$$y = a + bx$$

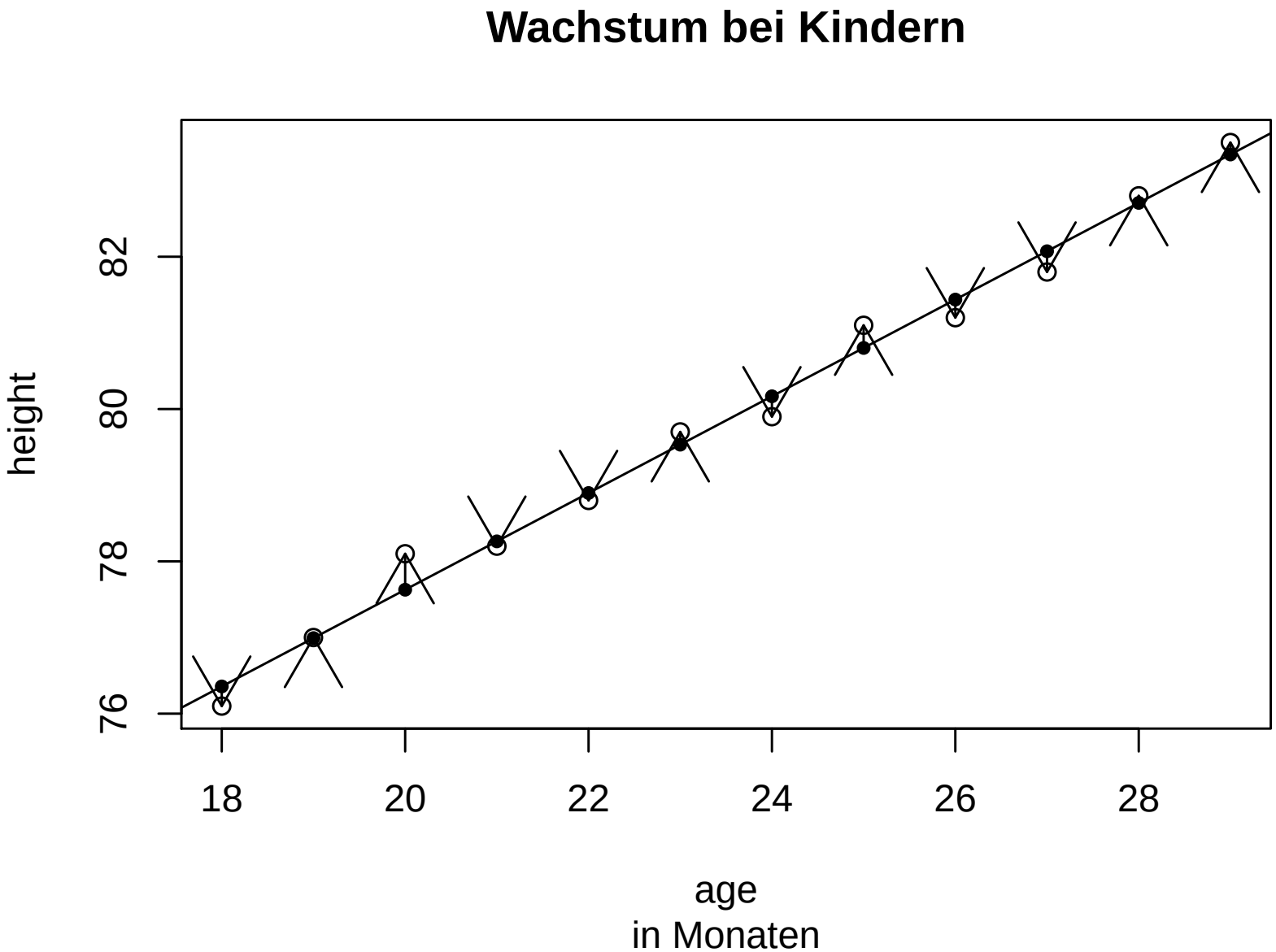
sein müßte.
Diesen Wert bezeichnen wir als die **vom Modell vorhergesagten Werte** $\hat{Y}_i$:

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

Computerausgabe

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

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

Coefficients:
(Intercept)          age
     64.928        0.635

> predict(model)

   1    2    3    4    5    6    7    8    9   10   11   12
76.4 77.0 77.6 78.3 78.9 79.5 80.2 80.8 81.4 82.1 82.7 83.3
```

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:

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

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

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:

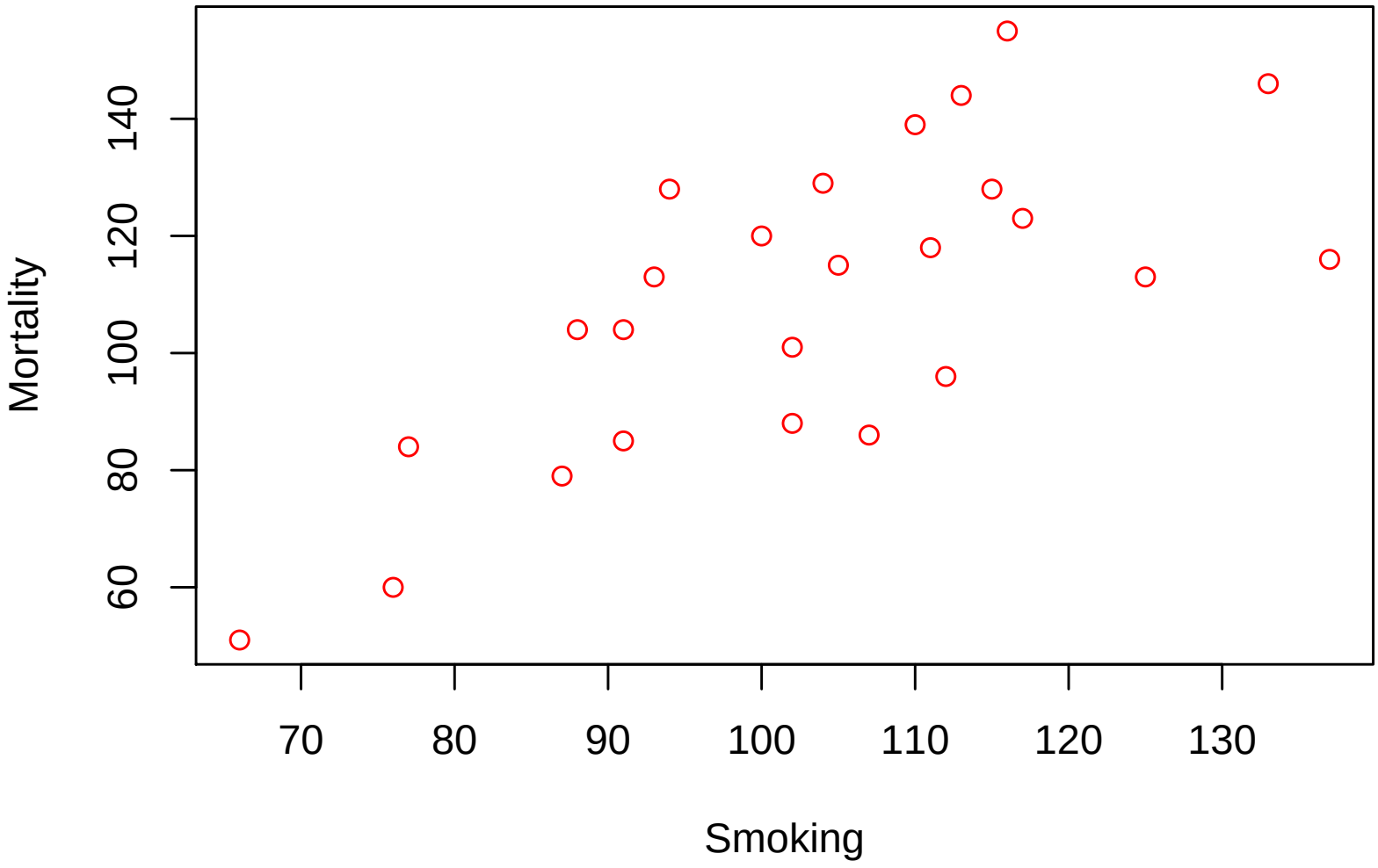
$$Y_i = \hat{a} + \hat{b}X_i + r_i$$

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

Computerausgabe

```
> resid(model)

       1        2        3        4        5        6        7        8
-0.25769  0.00734  0.47238 -0.06259 -0.09755  0.16748 -0.26748  0.29755
       9       10       11       12
-0.23741 -0.27238  0.09266  0.15769

> predict(model)

   1    2    3    4    5    6    7    8    9   10   11   12
76.4 77.0 77.6 78.3 78.9 79.5 80.2 80.8 81.4 82.1 82.7 83.3

> Wachstum$height

 [1] 76.1 77.0 78.1 78.2 78.8 79.7 79.9 81.1 81.2 81.8 82.8 83.5

> predict(model)+resid(model)-Wachstum$height

       1        2        3        4        5        6        7        8
8.53e-14 2.84e-14 1.42e-14 1.42e-14 1.42e-14 1.42e-14 1.42e-14 2.84e-14
       9       10       11       12
2.84e-14 2.84e-14 2.84e-14 1.42e-14
```

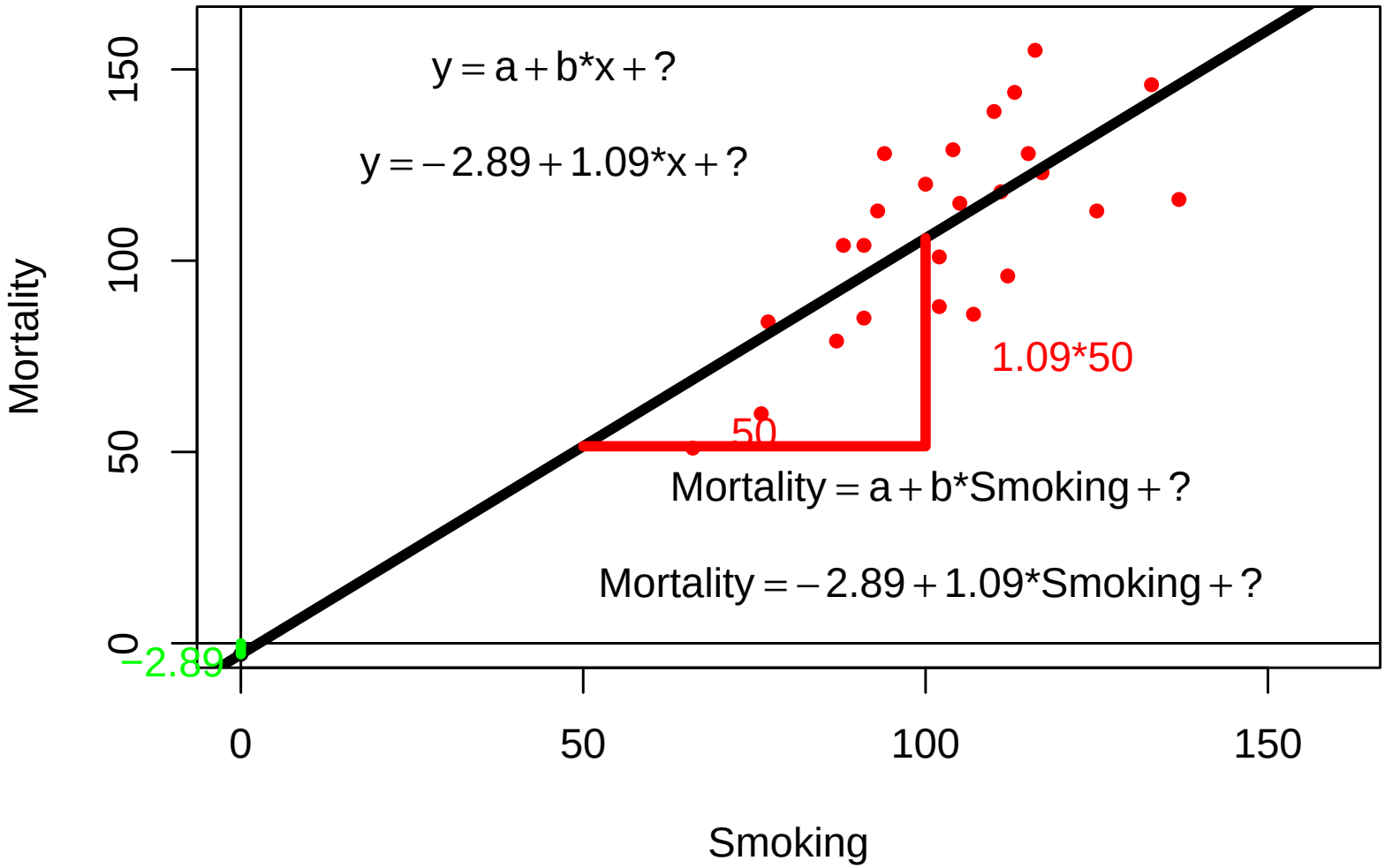
Beispiel II: Rauchen



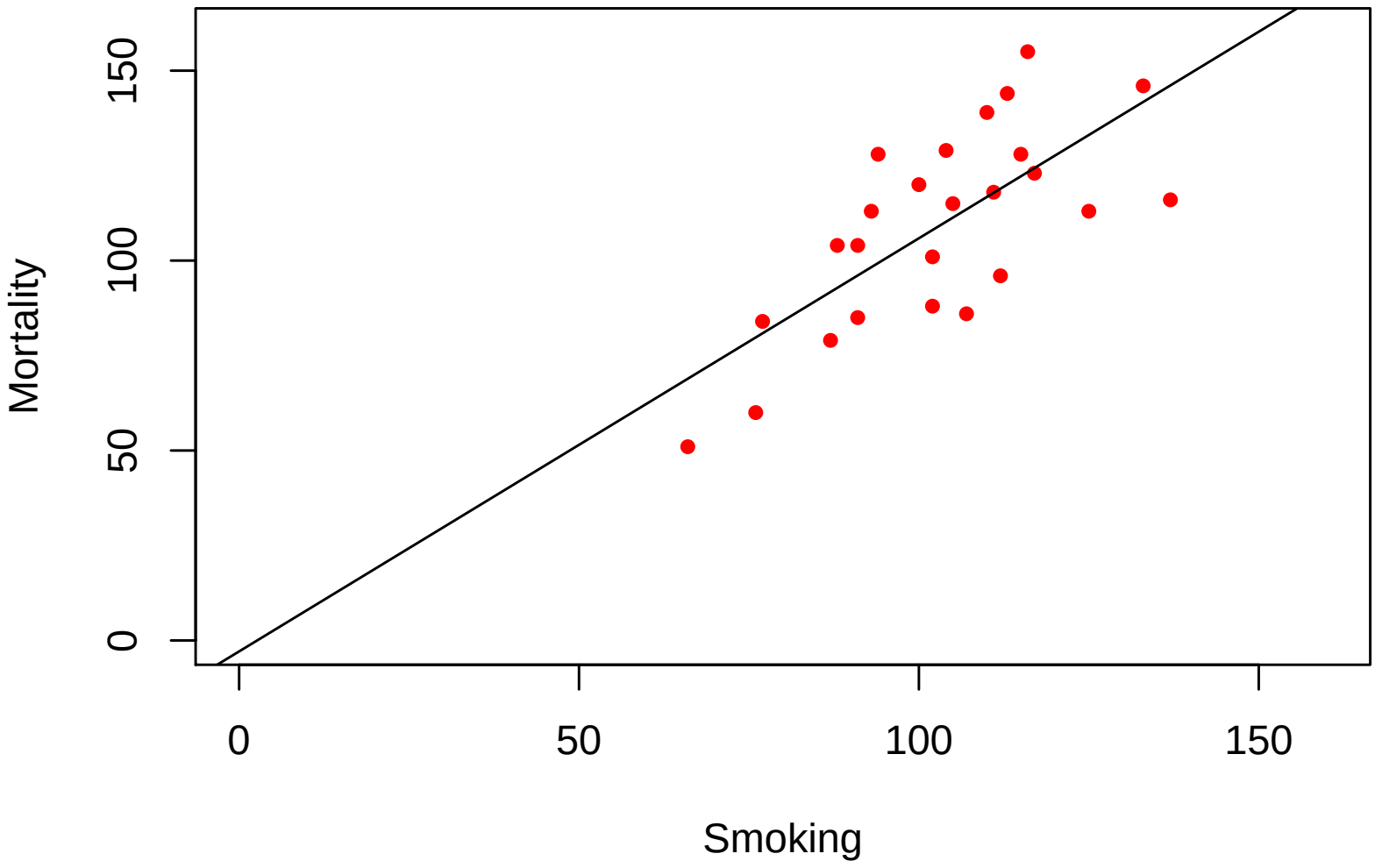
Computerausgabe

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

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

Coefficients:
(Intercept)      Smoking
      -2.89         1.09

> model

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

Coefficients:
(Intercept)      Smoking
      -2.89         1.09

> predict(model)
                Farmers, foresters, and fisherman
                                             80.9
                          Miners and quarrymen
                                            146.1
                   Gas, coke and chemical makers
                                            124.4
                       Glass and ceramics makers
                                             99.3
Furnace, forge, foundry, and rolling mill workers
                                            123.3
```

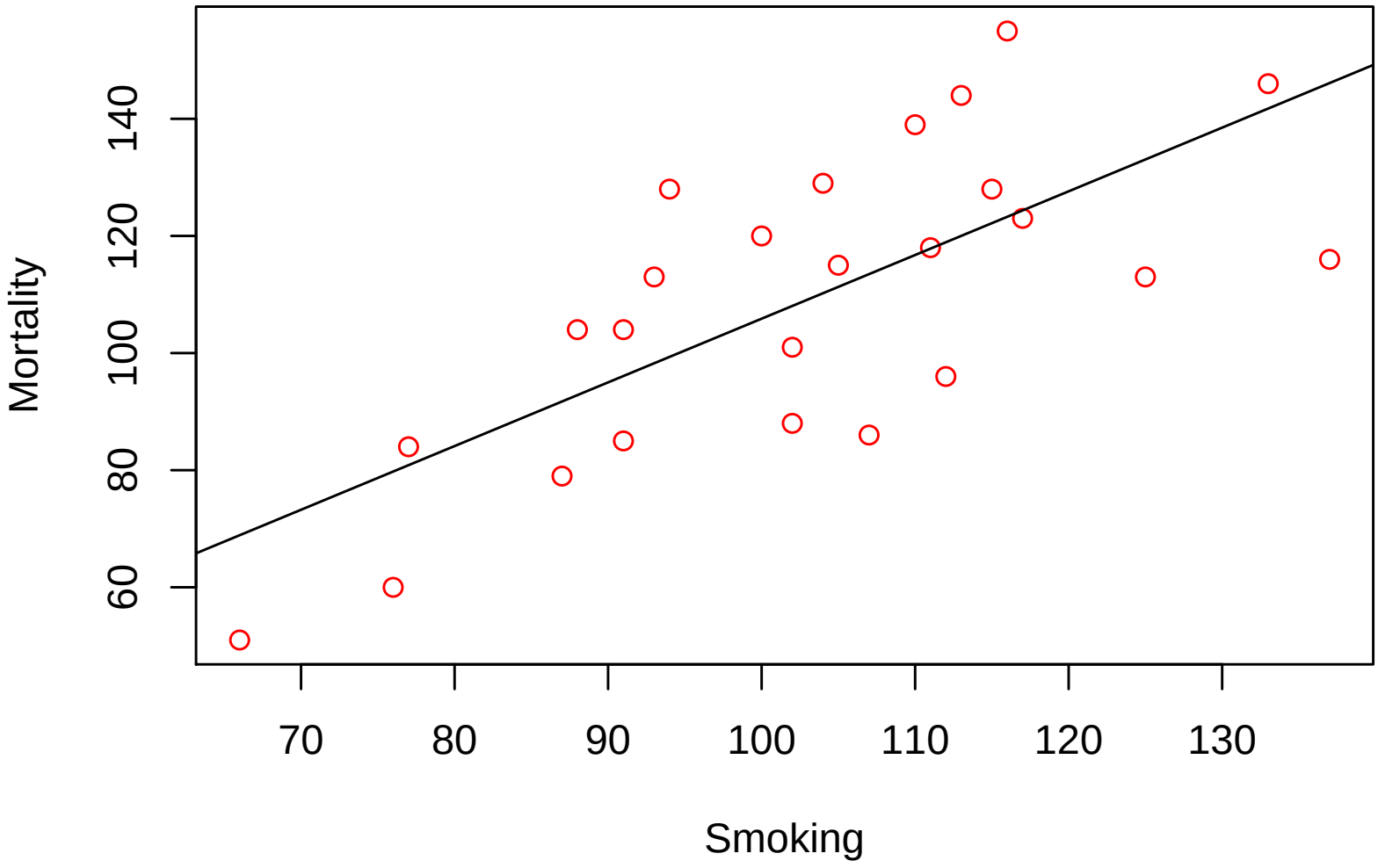
Graphische Interpretation



Normalansicht Schritt 1



Regressionsgerade



Modellvorstellung: Ungenaue Beschreibung

Das Regressionsmodel:

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

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

Modellvorstellung: Ungenaue Beschreibung

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 (bzw. a und b) wird so bestimmt, dass

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

so klein wie möglich wird. (SS=Sums of Squares)

Bestimmung der Geraden

- Die Regressionsgerade (bzw. a und b) wird so bestimmt, dass

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

  so klein wie möglich wird. (SS=Sums of Squares)
- Dieses Verfahren nennt man: Kleinste Quadrate

Bestimmung der Geraden

- Die Regressionsgerade (bzw. a und b) wird so bestimmt, dass

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

  so klein wie möglich wird. (SS=Sums of Squares)
- Dieses Verfahren nennt man: Kleinste Quadrate
- Dieses Vefahren ist besonders gut, wenn gewisse Annahmen erfüllt sind.

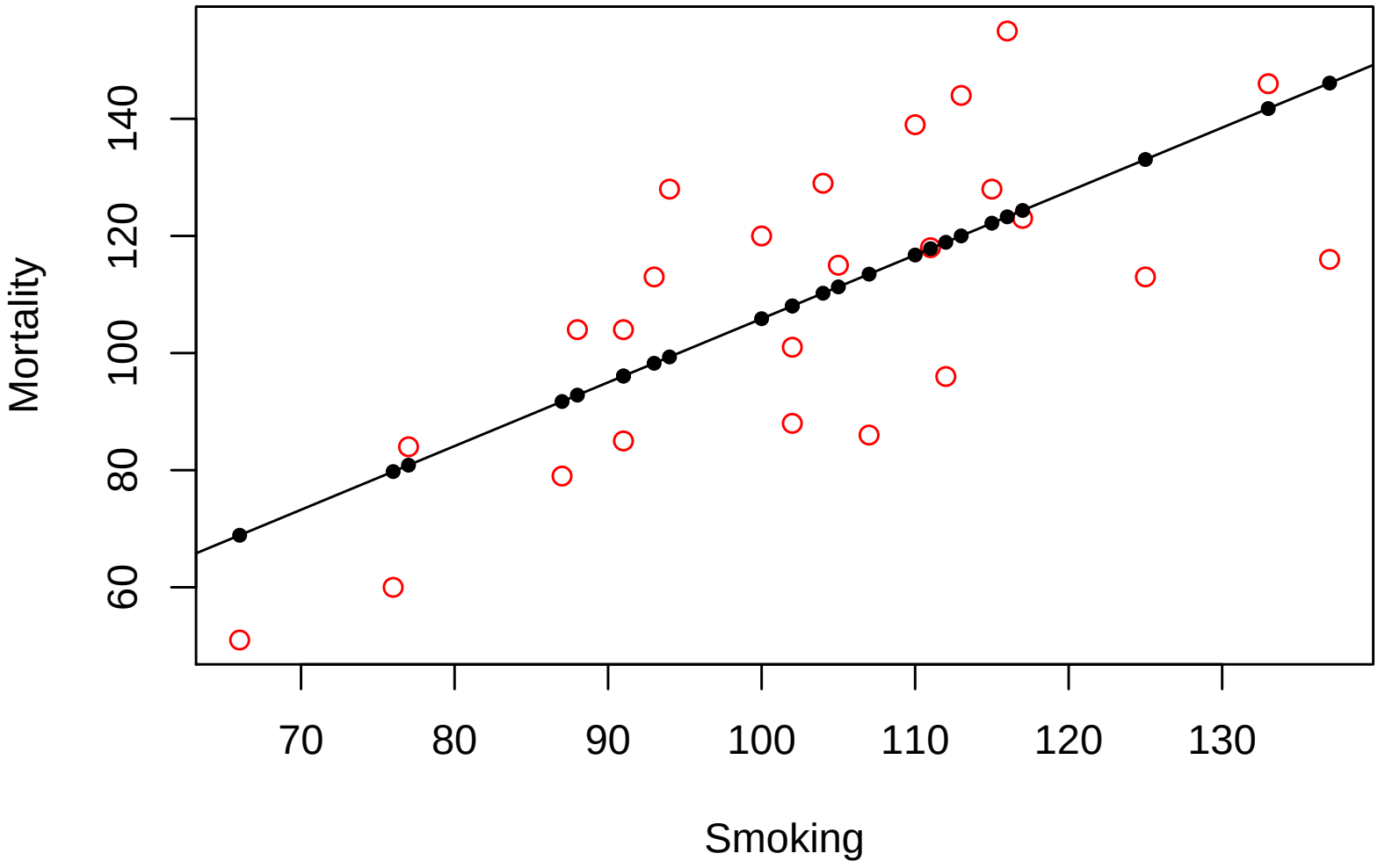
Bestimmung der Geraden

- Die Regressionsgerade (bzw. a und b) wird so bestimmt, dass

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

  so klein wie möglich wird. (SS=Sums of Squares)
- 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.

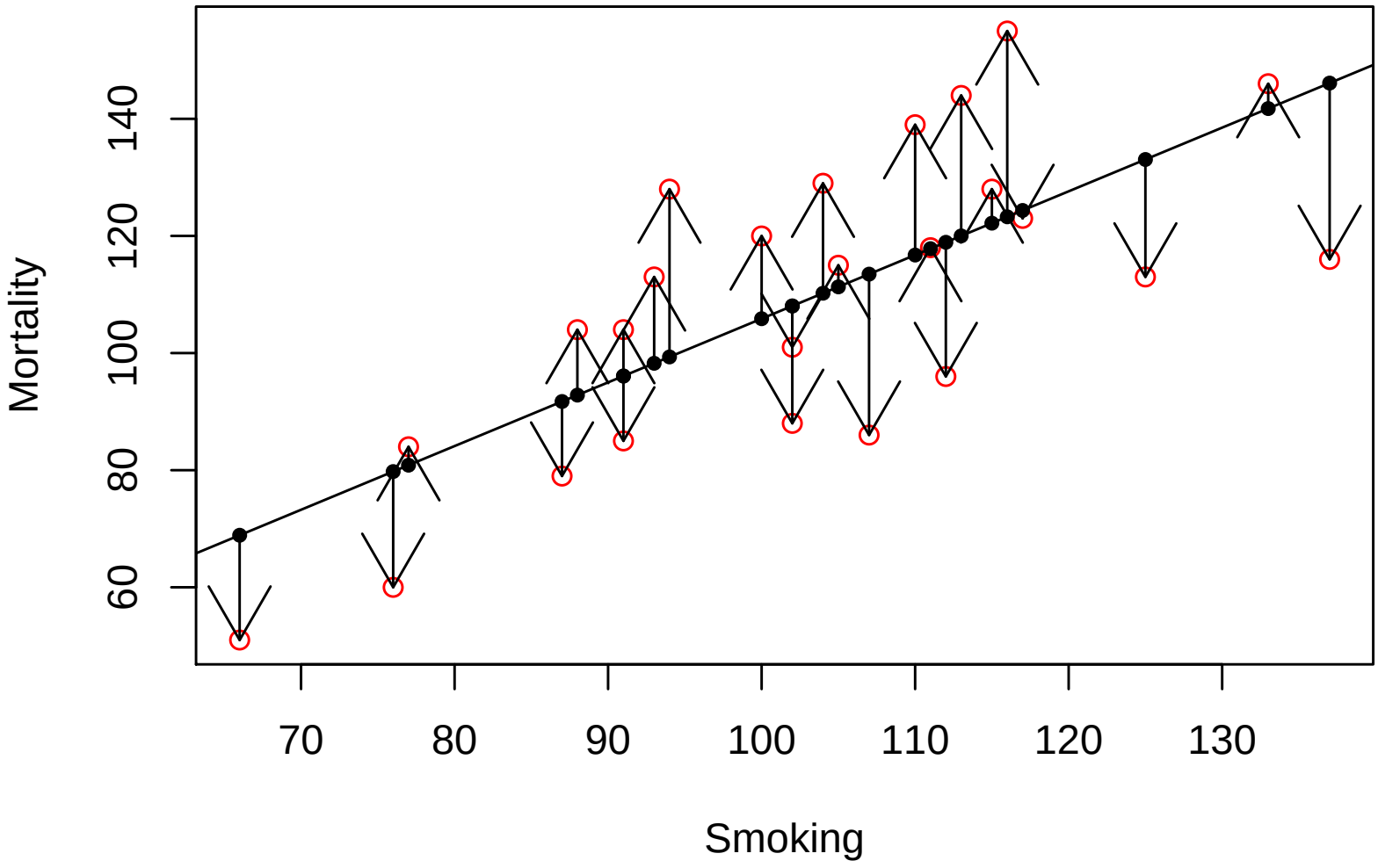
Vorhersagewerte



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

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

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

$$r_i = Y_i - \hat{Y}_i = Y_i - (\hat{a} + \hat{b}X_i)$$

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

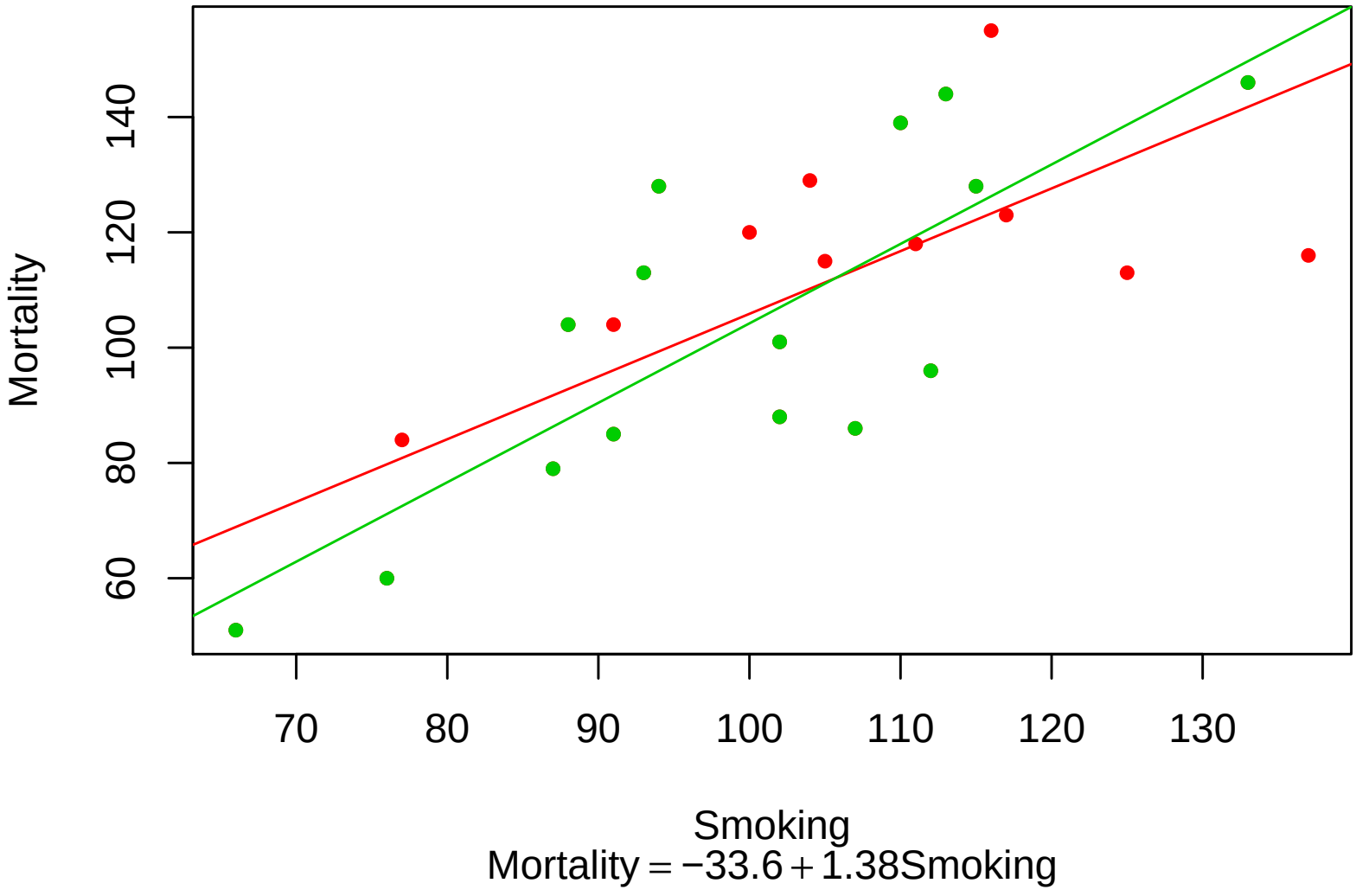
Sinn oder Unsinn,...

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

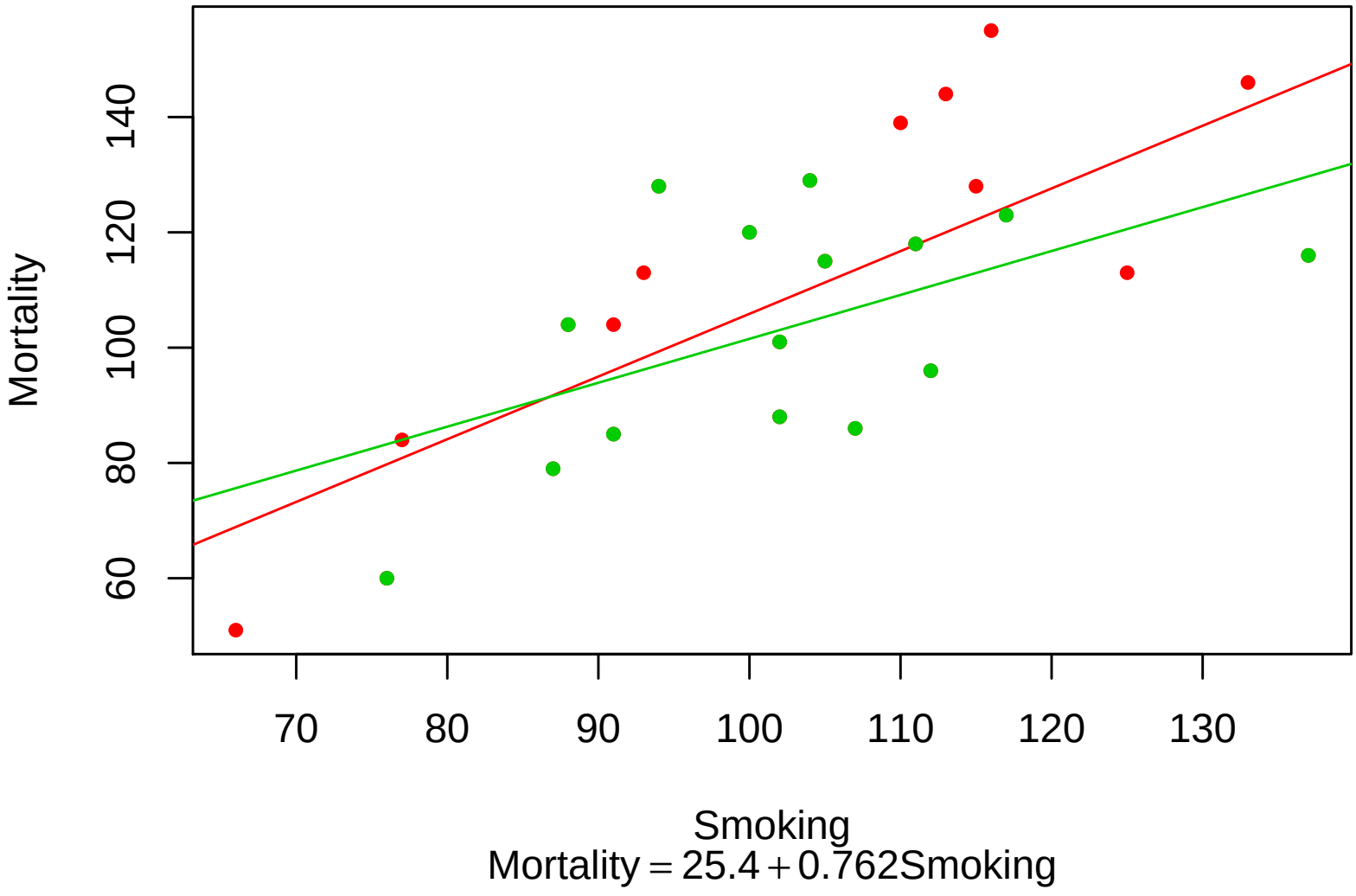
Demonstration: Stichprobenwahl

- 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.

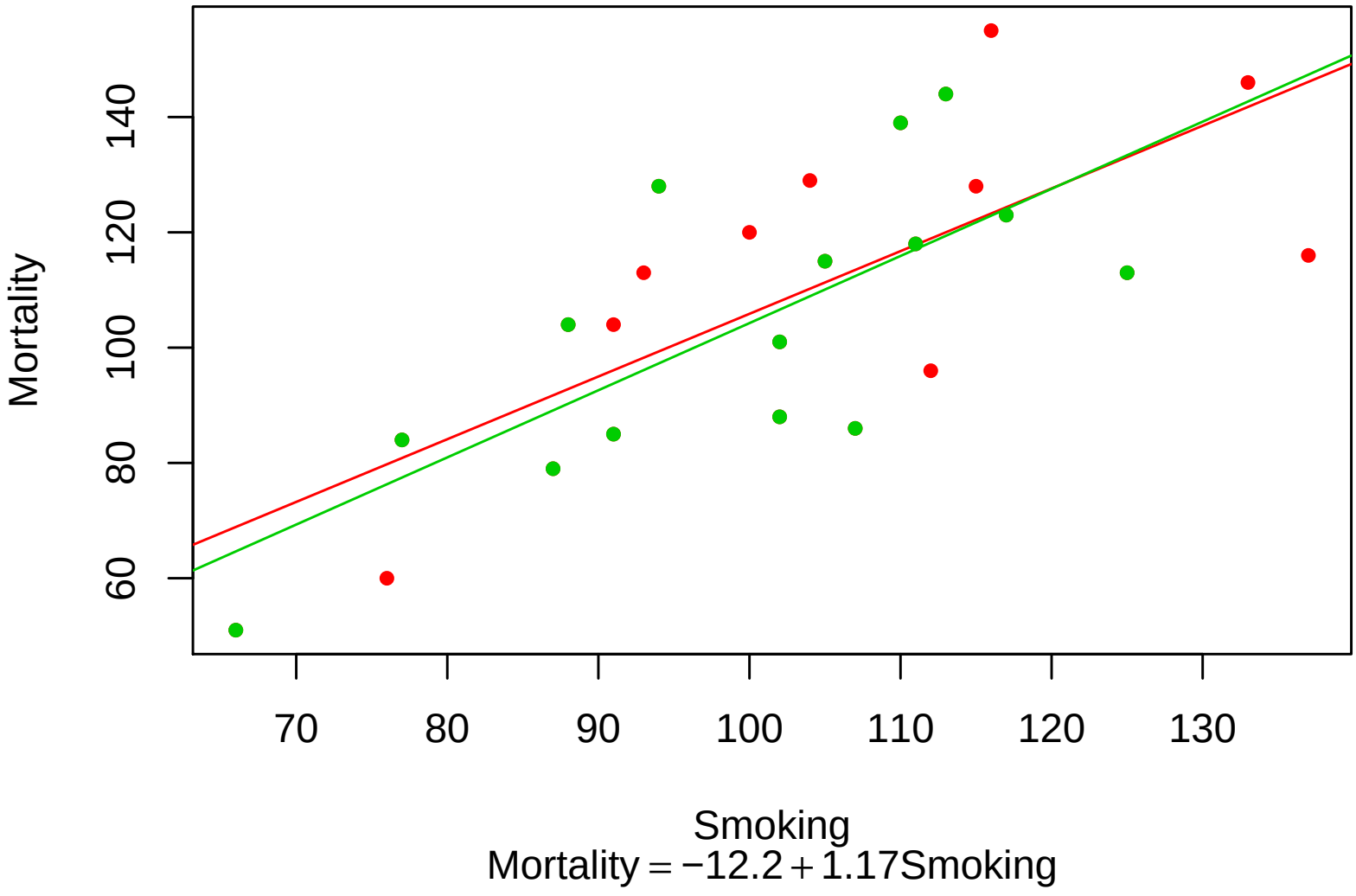
Demonstration: Stichprobenwahl



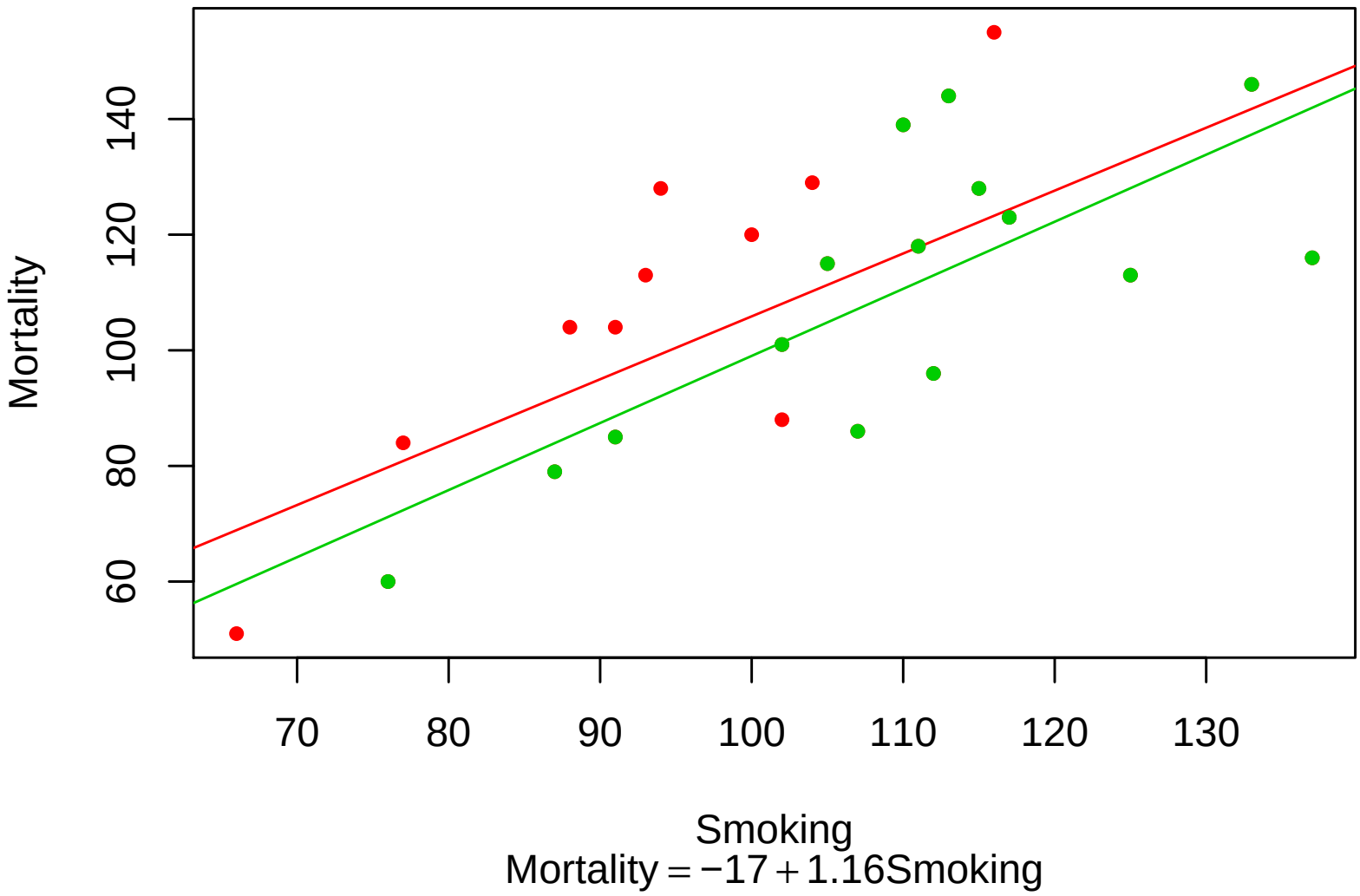
Demonstration: Stichprobenwahl



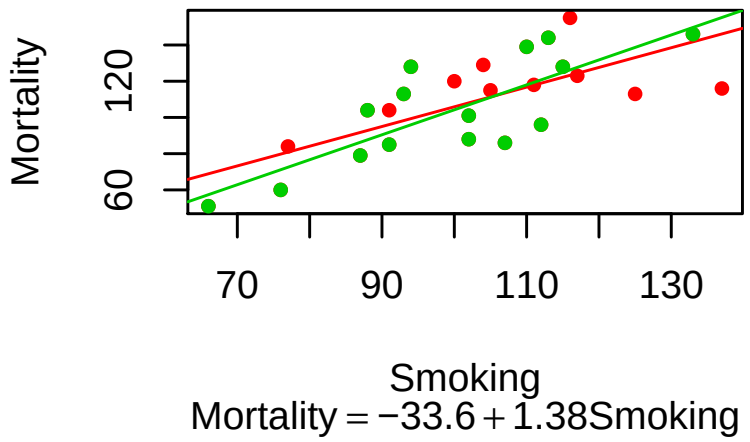
Demonstration: Stichprobenwahl



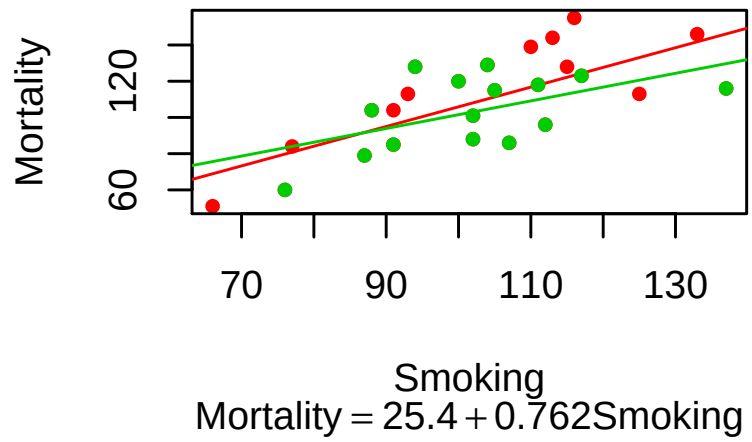
Demonstration: Stichprobenwahl



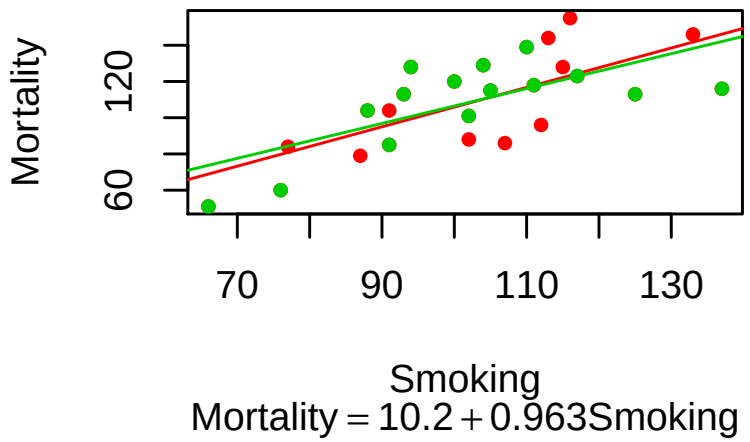
Demonstration: Stichprobenwahl



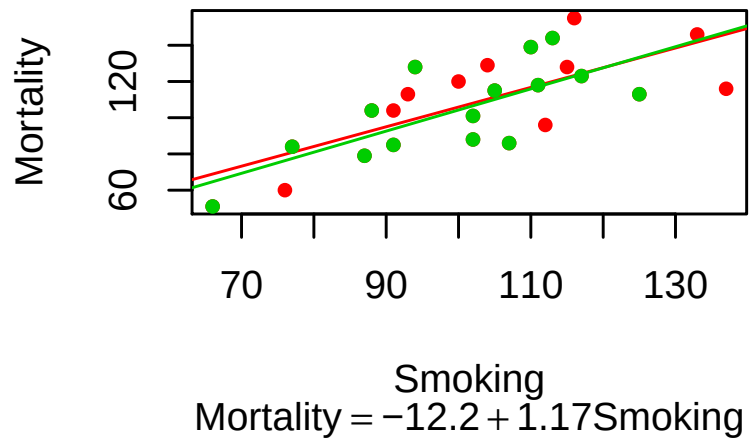
$$\text{Mortality} = -33.6 + 1.38\text{Smoking}$$



$$\text{Mortality} = 25.4 + 0.762\text{Smoking}$$



$$\text{Mortality} = 10.2 + 0.963\text{Smoking}$$



$$\text{Mortality} = -12.2 + 1.17\text{Smoking}$$

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:

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

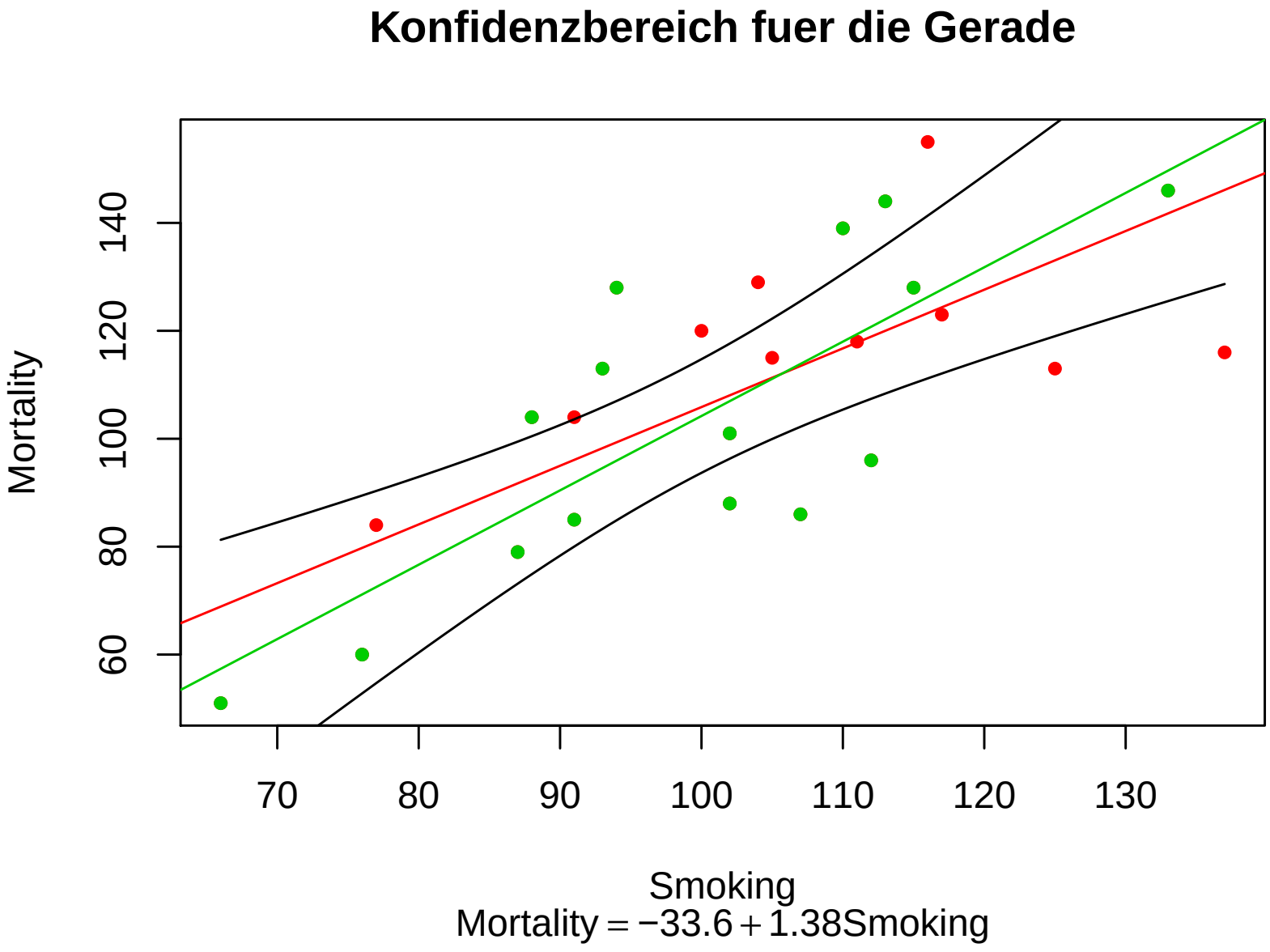
$$l(x) = \hat{a} + \hat{b}x - \hat{sd}(\epsilon) q_{t_{n-2},\alpha/2} (c_1 + 2c_2 x + c_3 x^2)$$

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

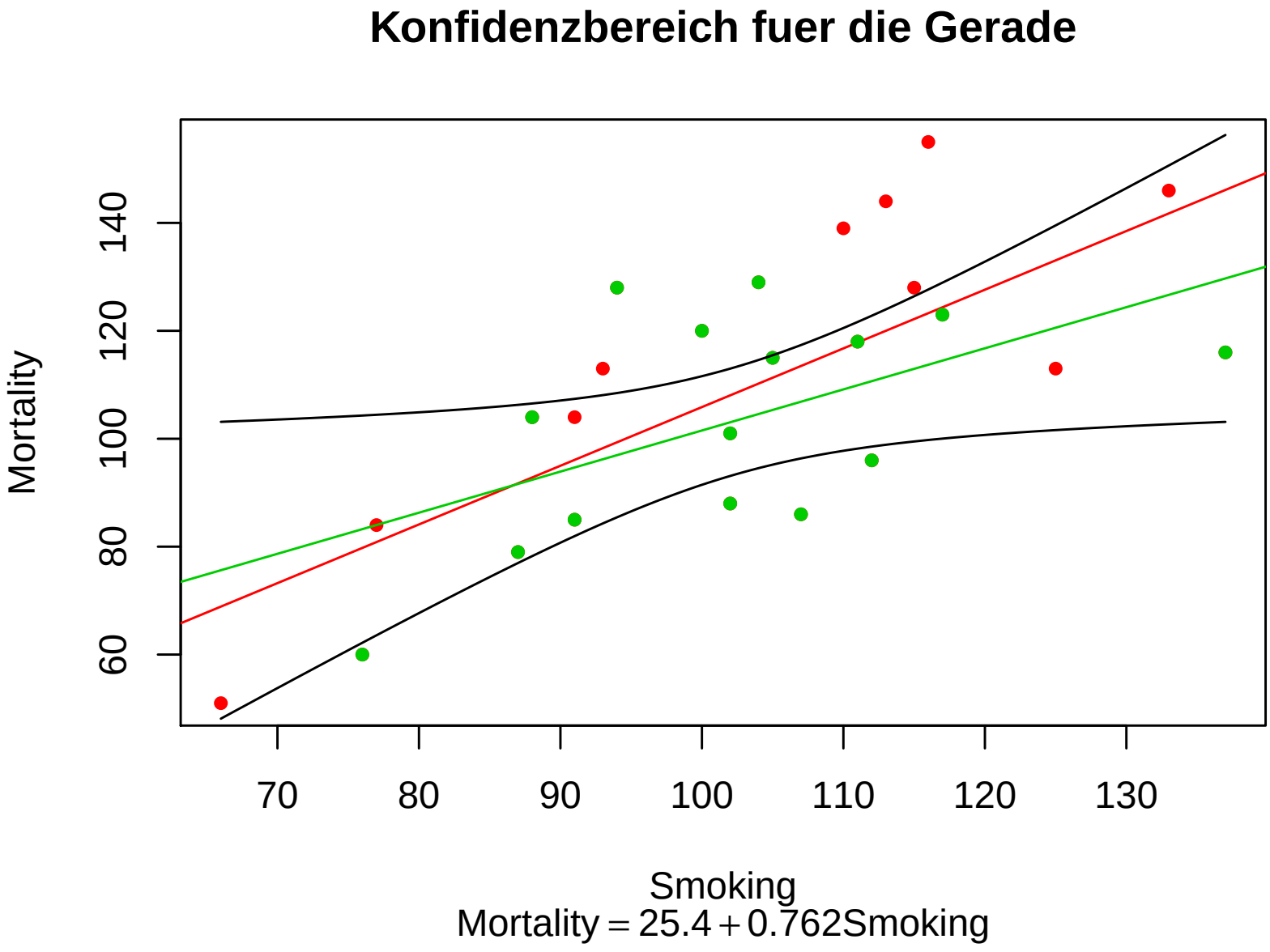
$$\begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}^{-1}$$

ohne Gewähr.

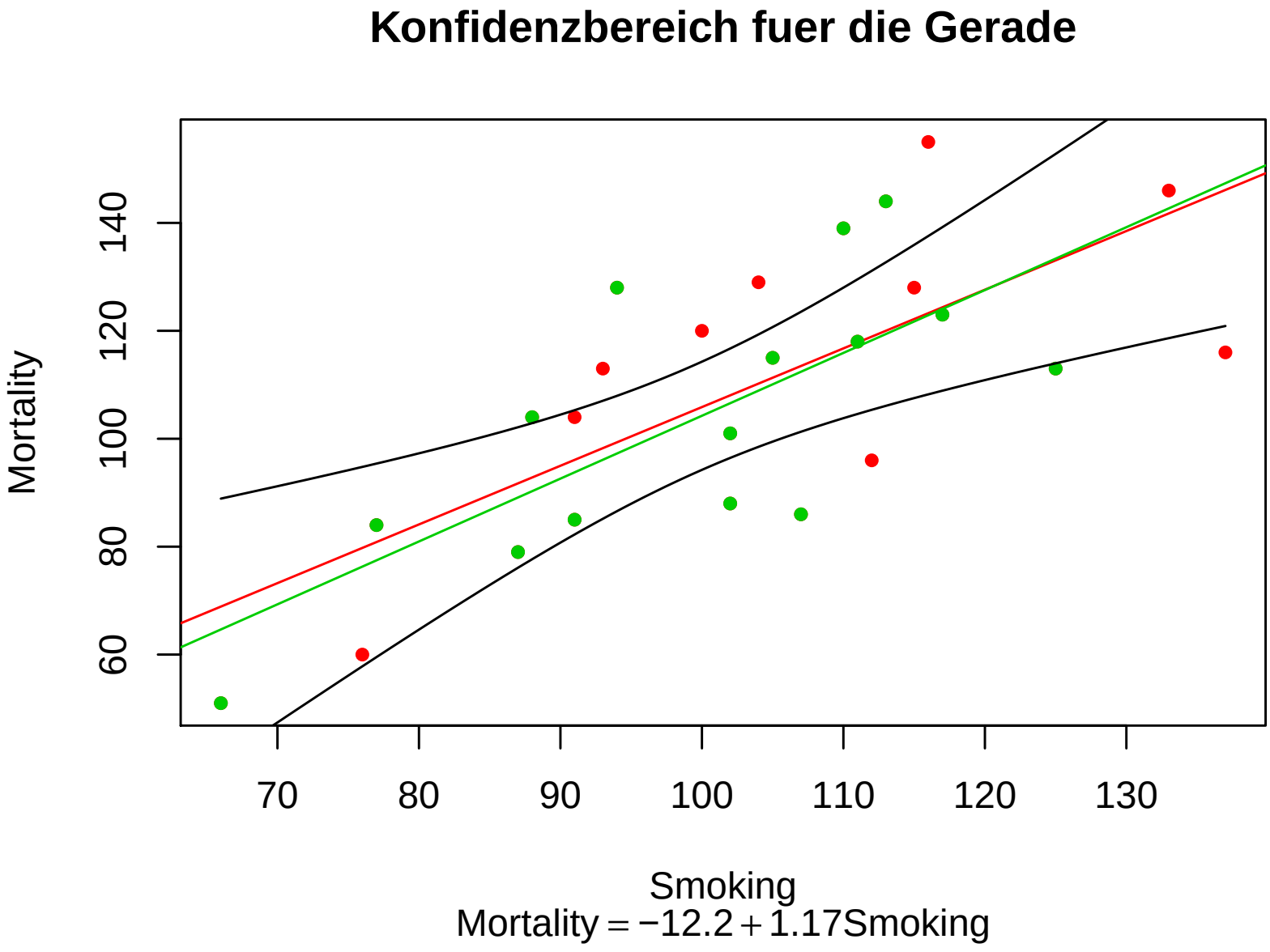
Demonstration des Konfidenzintervalls



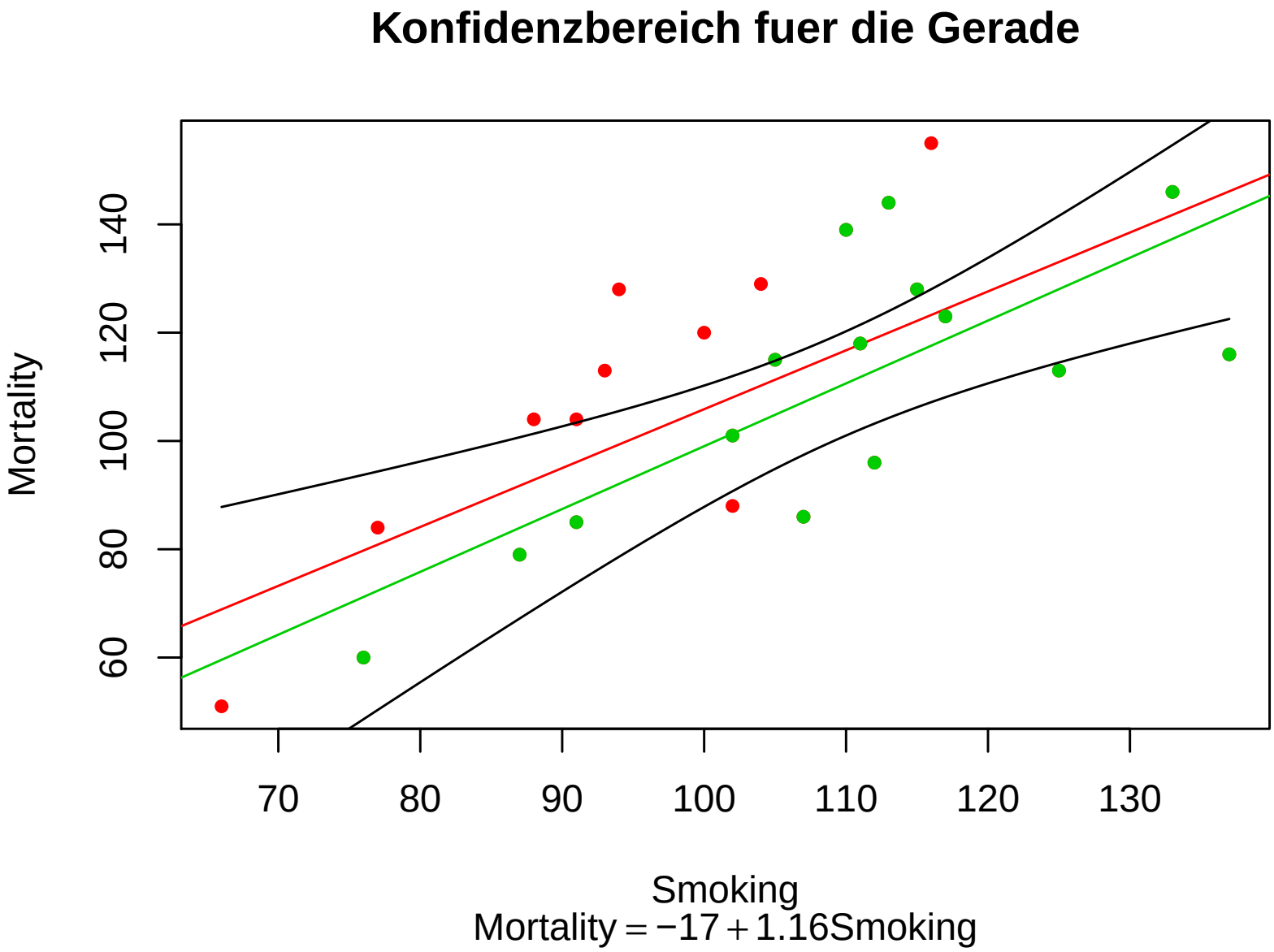
Demonstration des Konfidenzintervalls



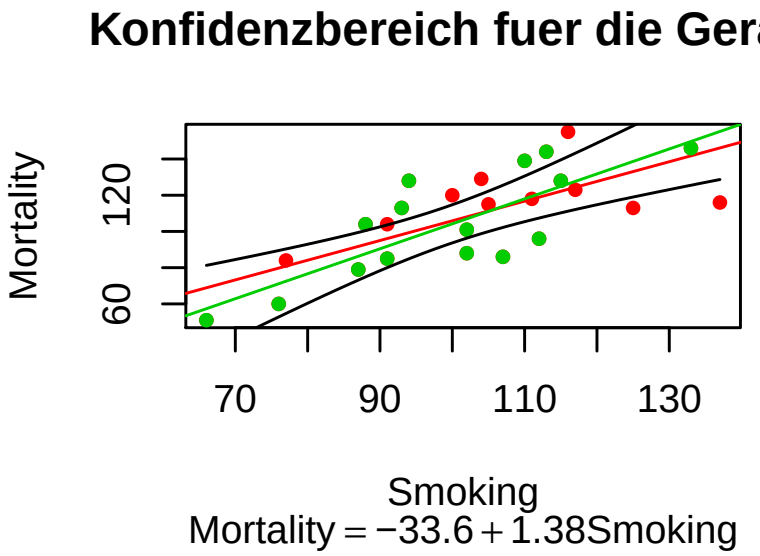
Demonstration des Konfidenzintervalls

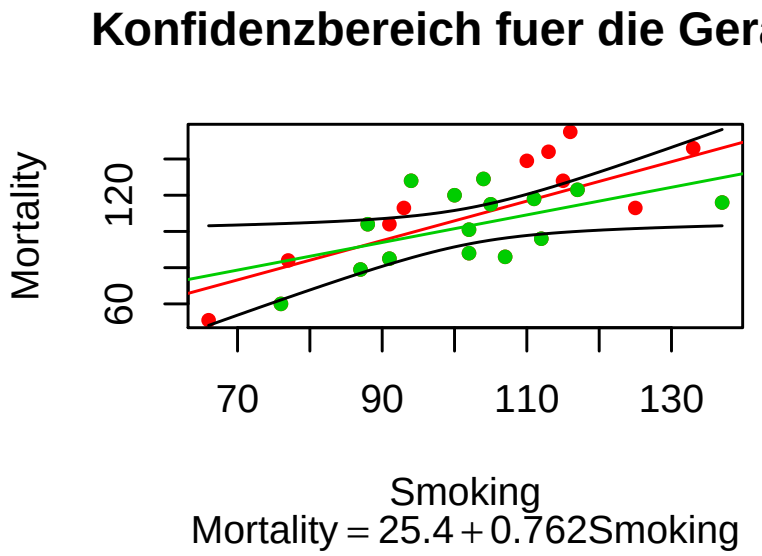


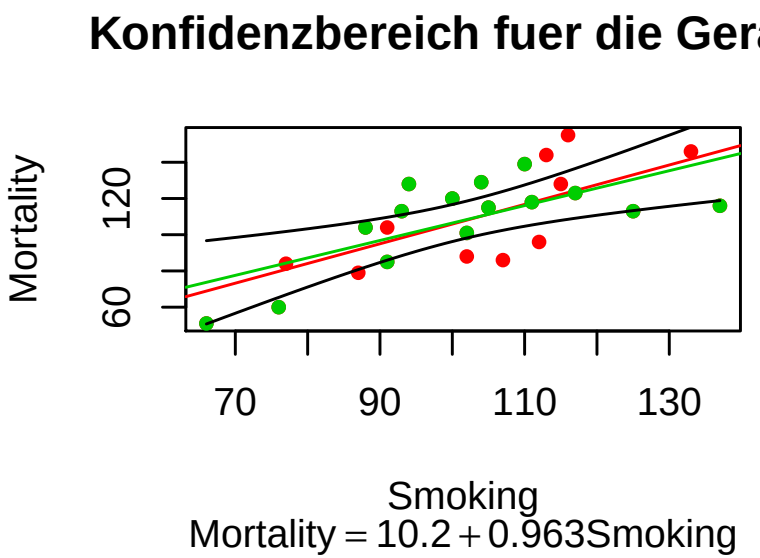
Demonstration des Konfidenzintervalls

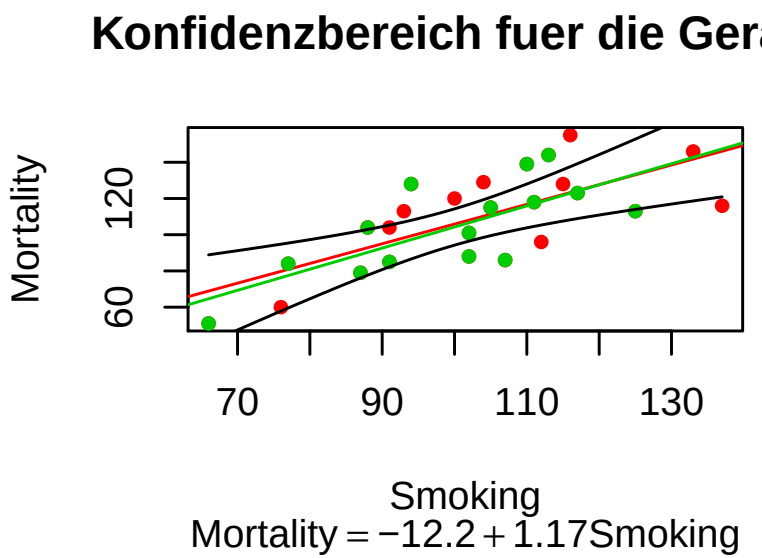


Demonstration des Konfidenzintervalls







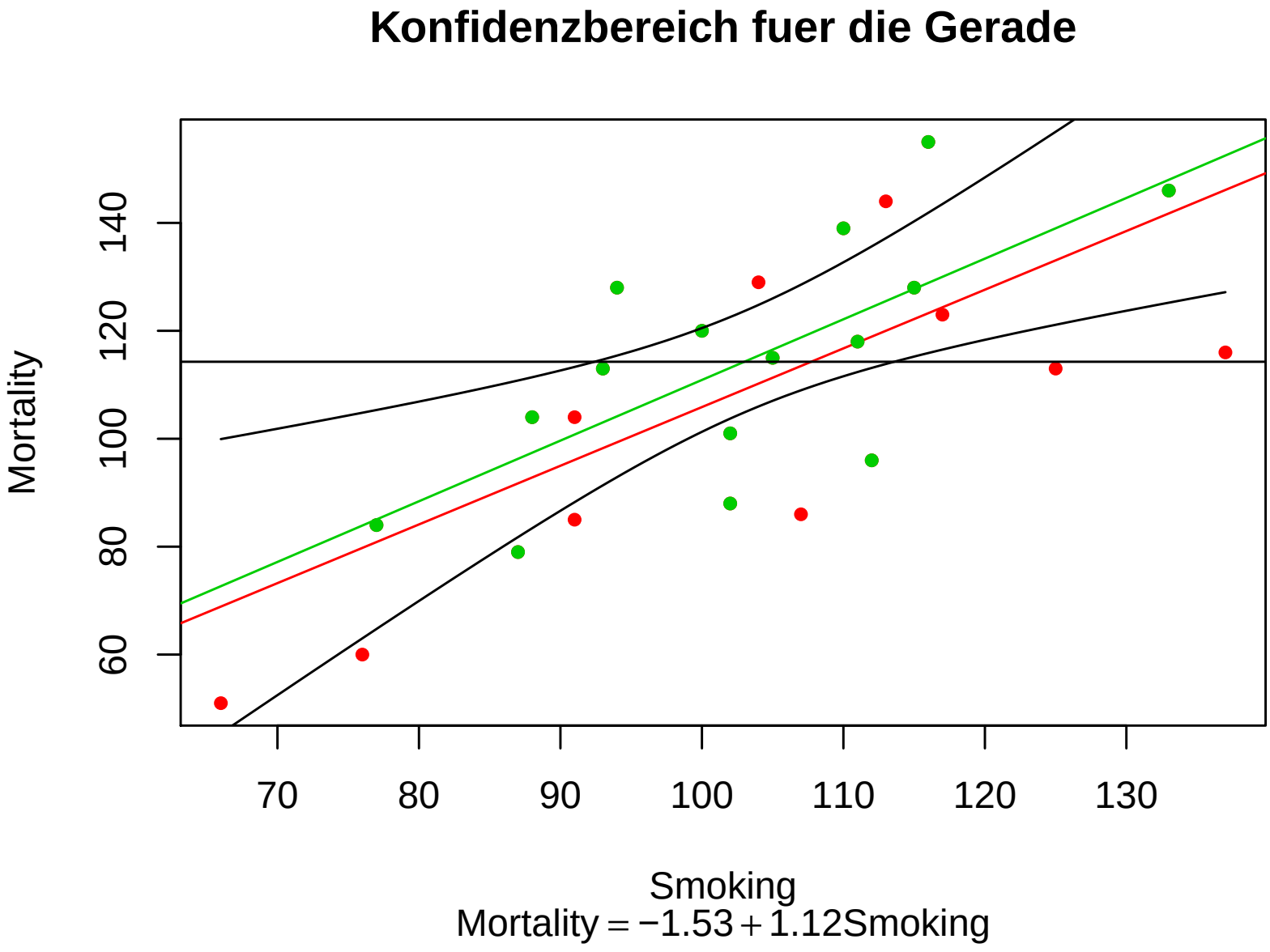


Sinn oder Unsinn,...

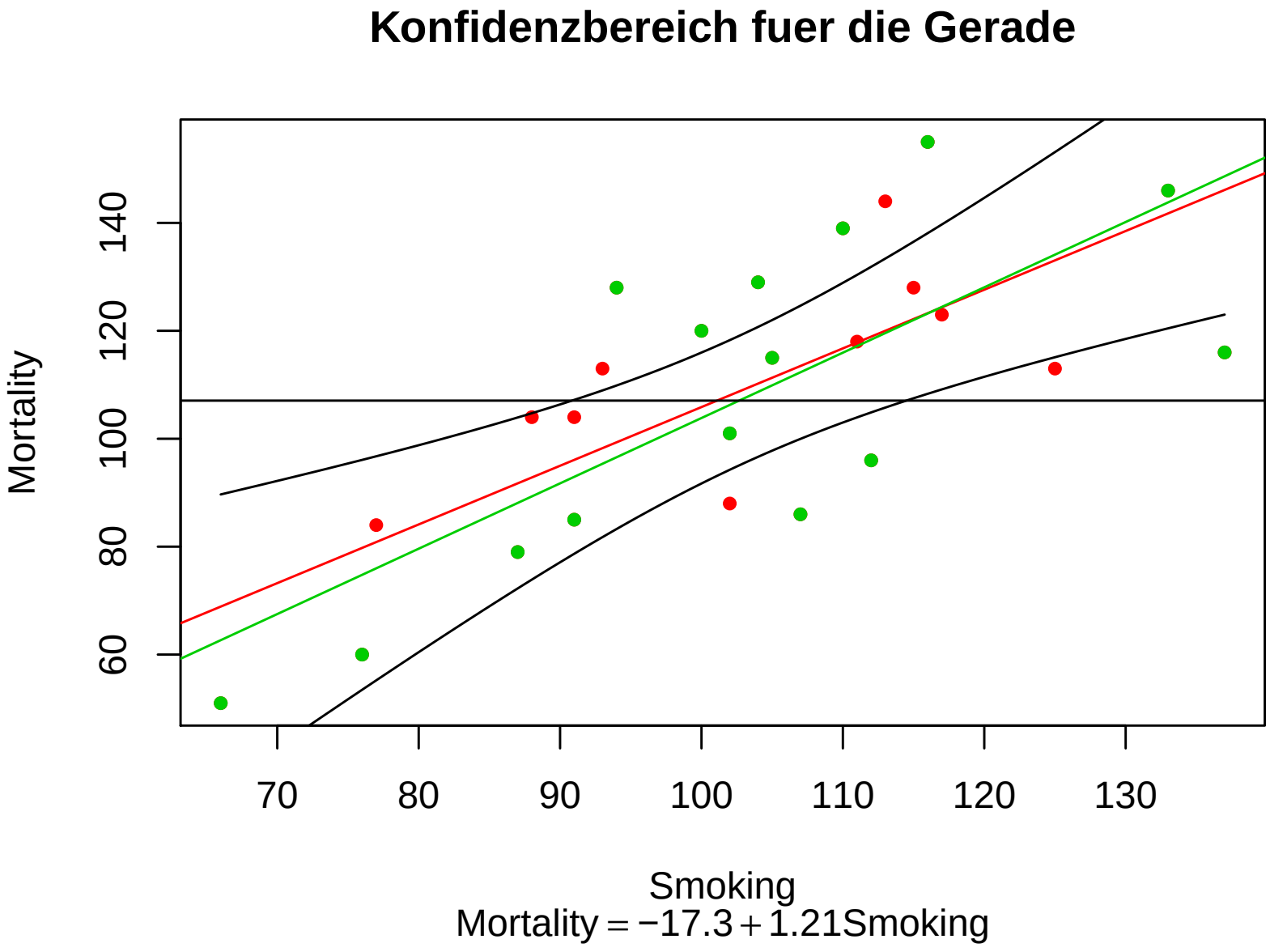
Eine wichtige Frage ist also immer, ob eventuell kein Zusammenhang bestehen konnte.
Wenn X und Y unabhängig sind, dann ändert sich der Erwartungswert von Y nicht in Abhängigkeit von X:

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

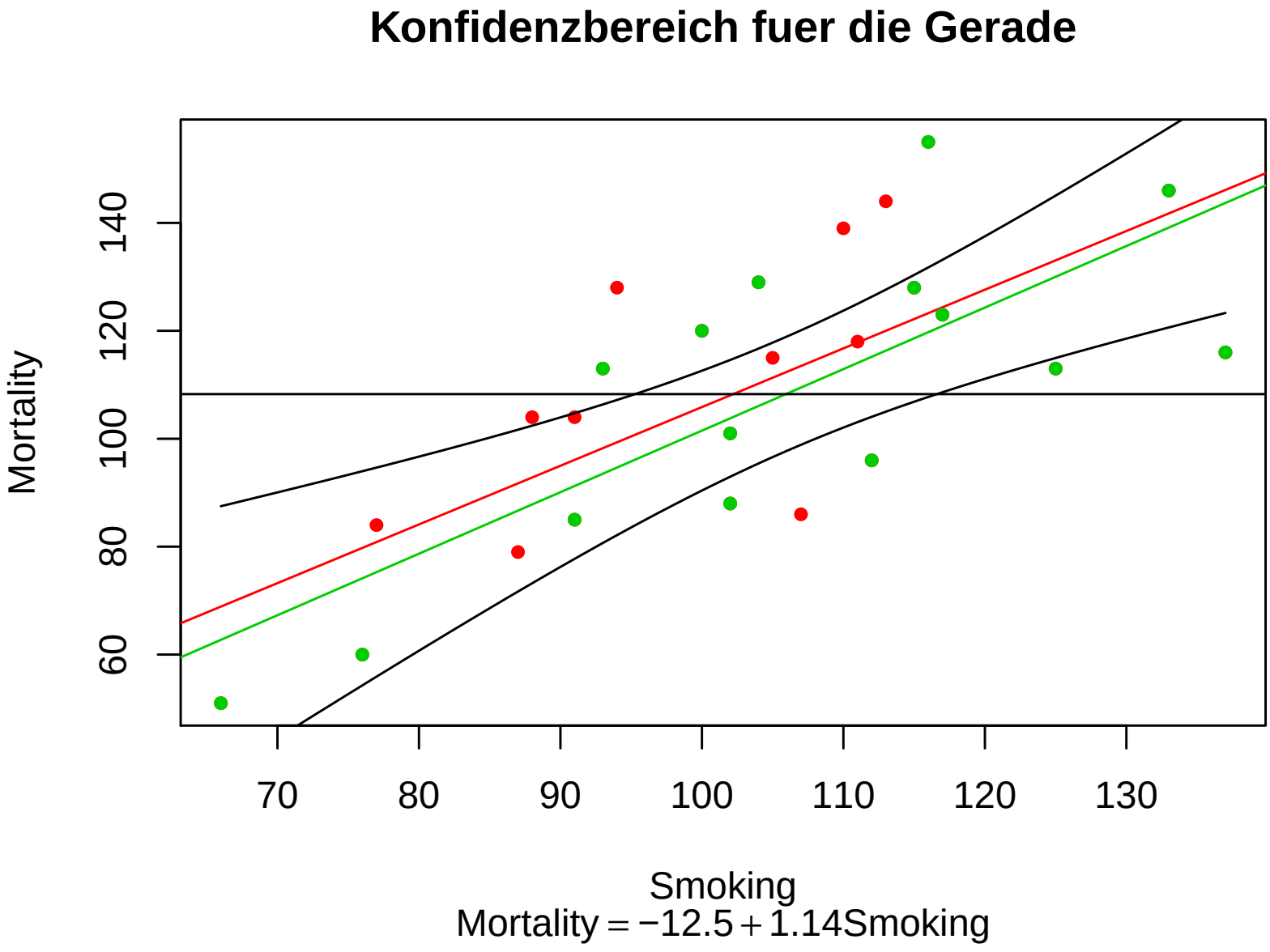
Könnte kein Zusammenhang bestehen?



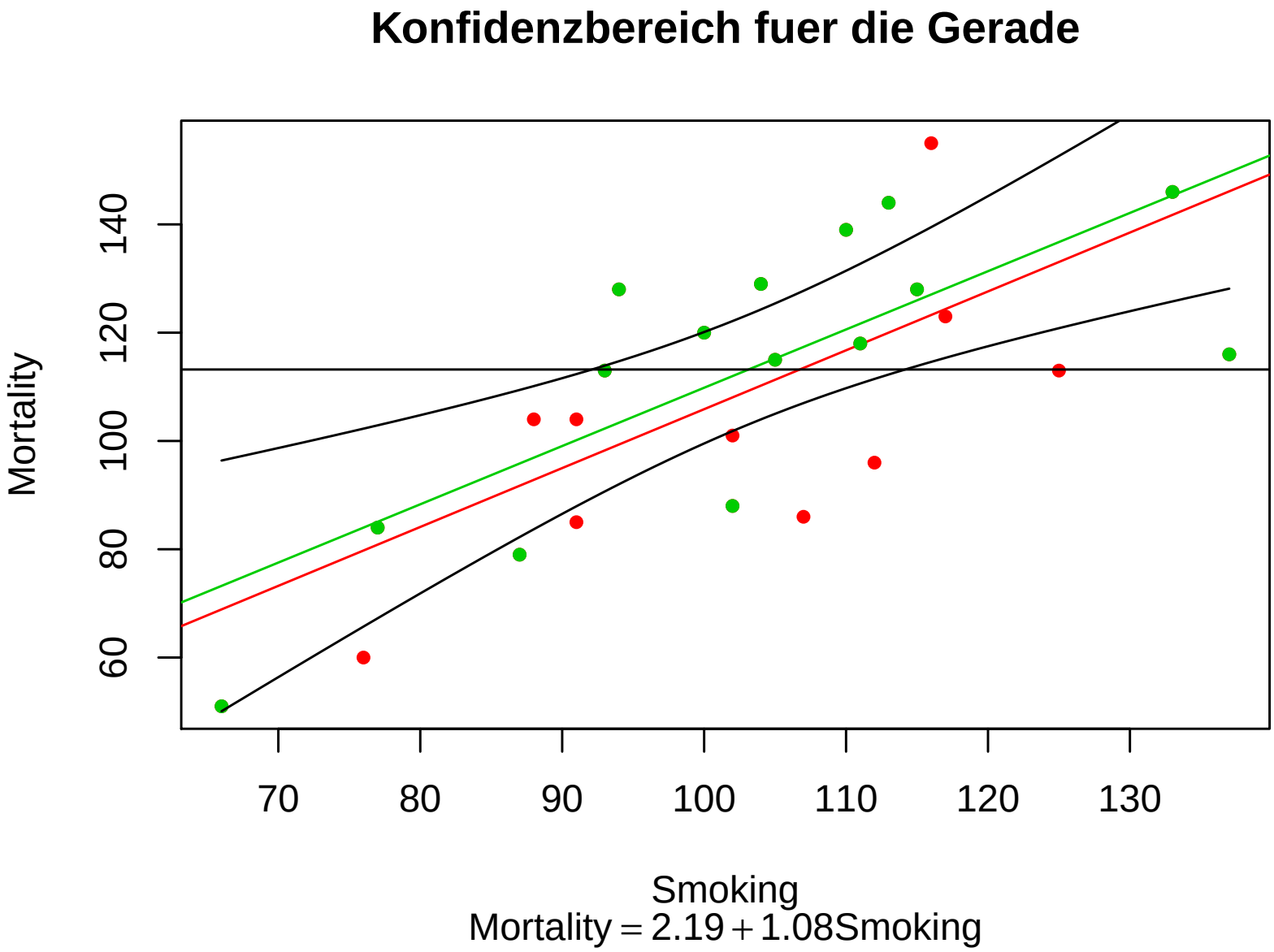
Könnte kein Zusammenhang bestehen?



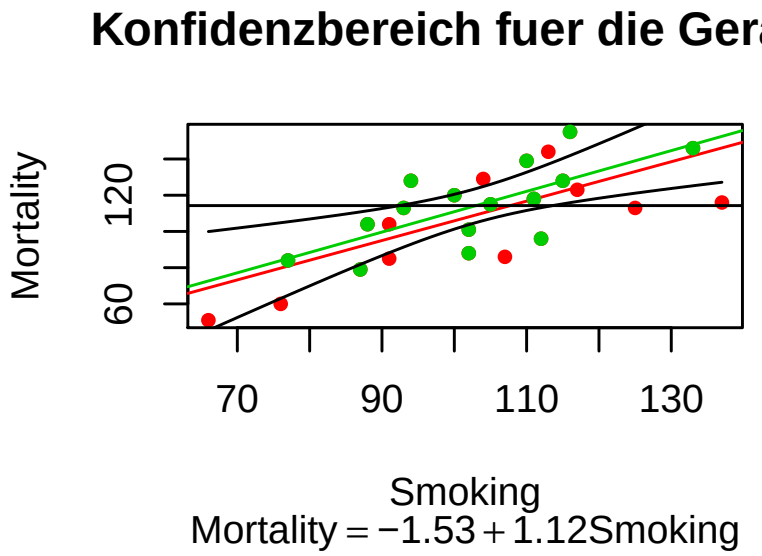
Könnte kein Zusammenhang bestehen?

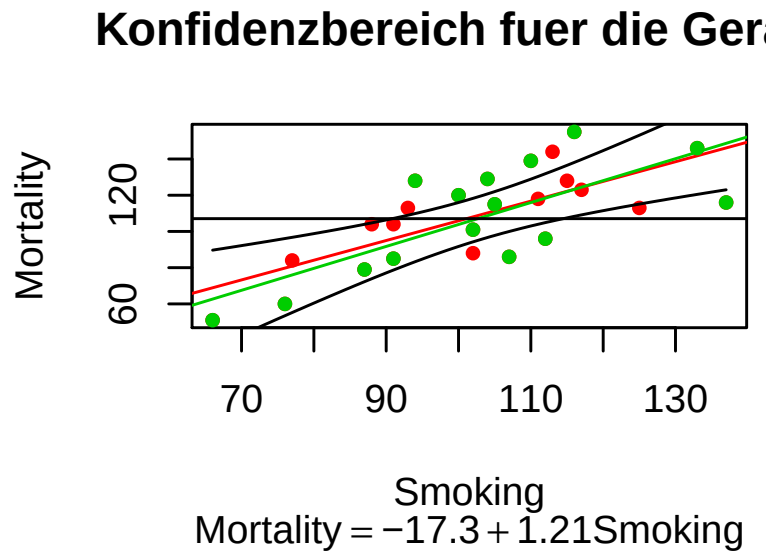


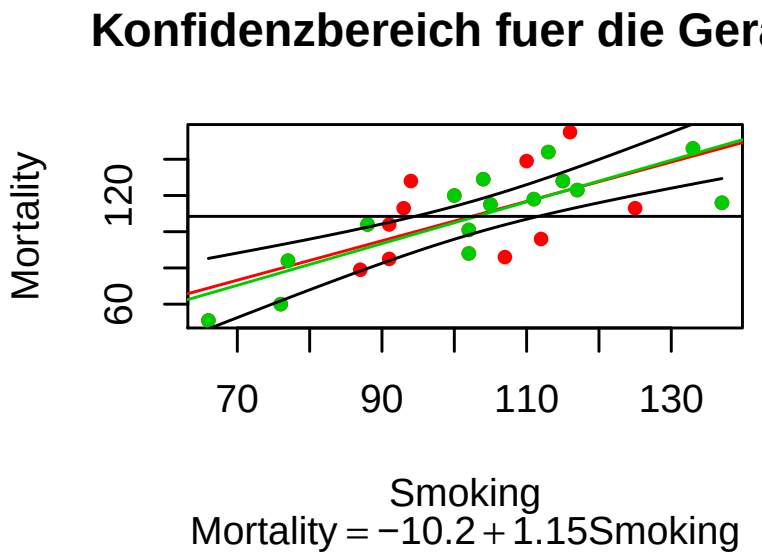
Könnte kein Zusammenhang bestehen?

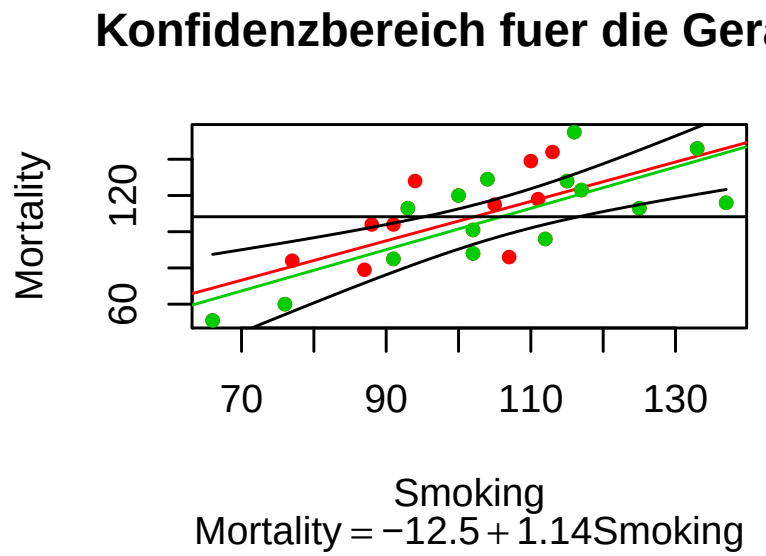


Könnte kein Zusammenhang bestehen?





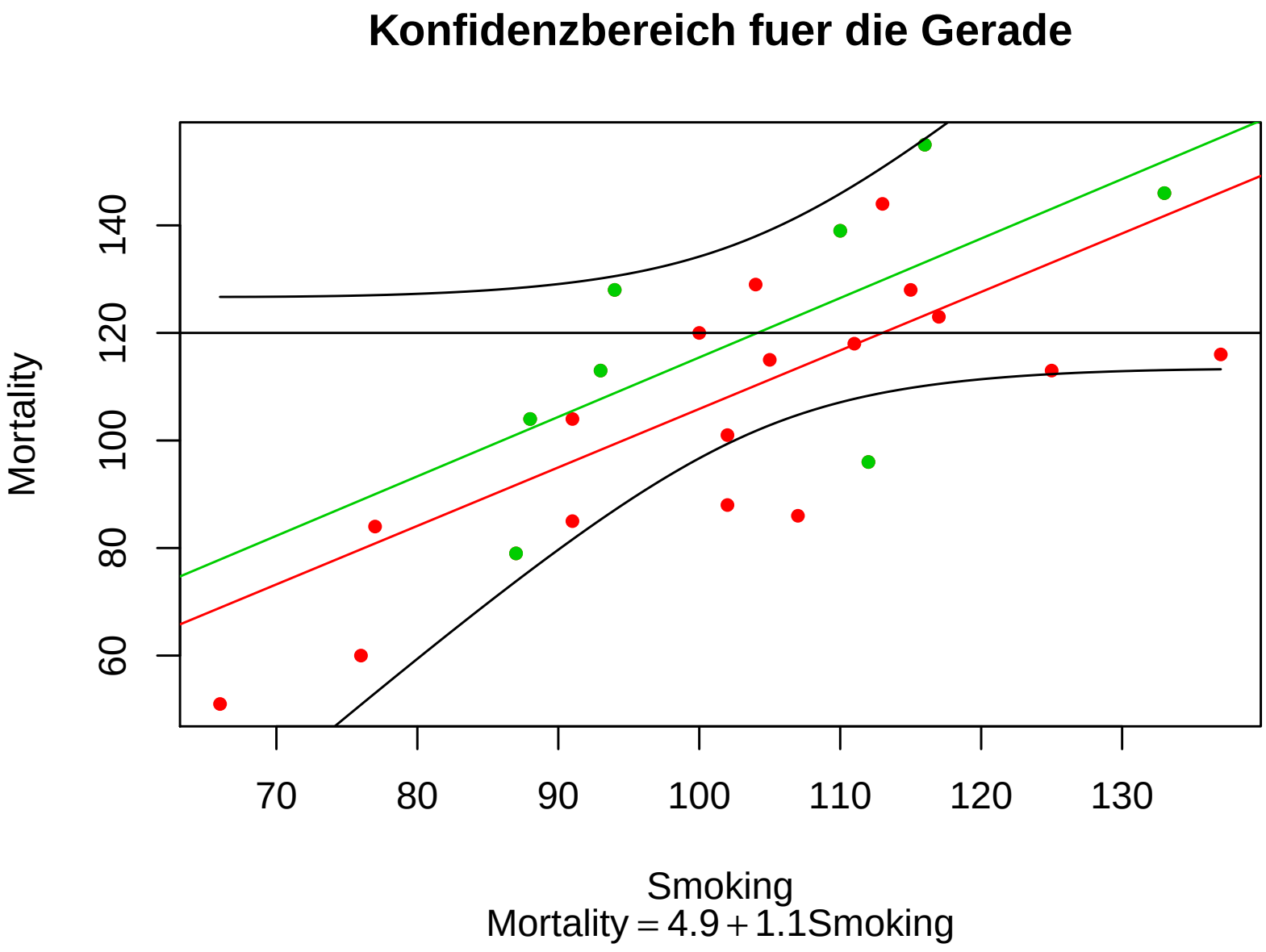




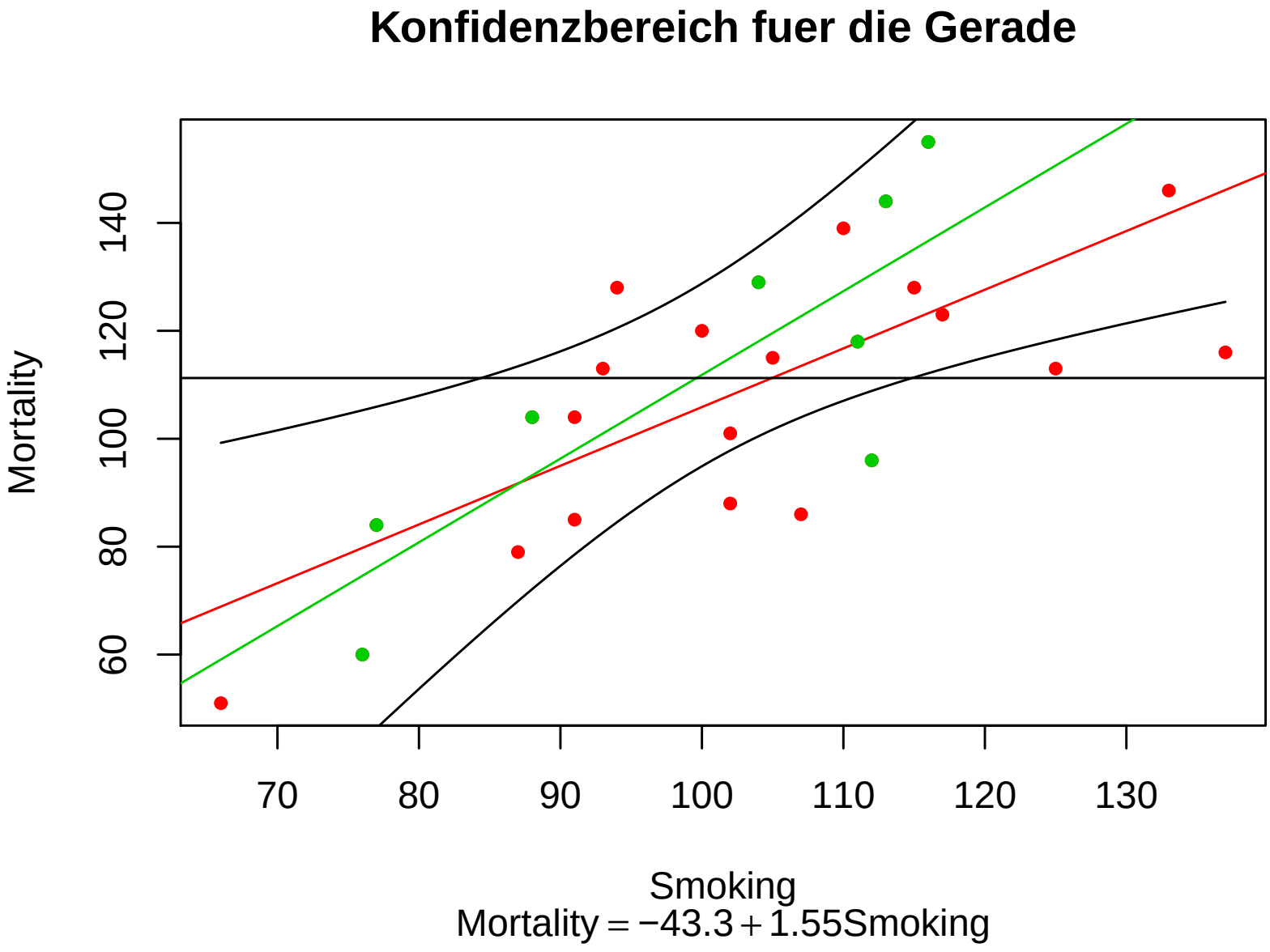
Ergebnisse

- 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 .

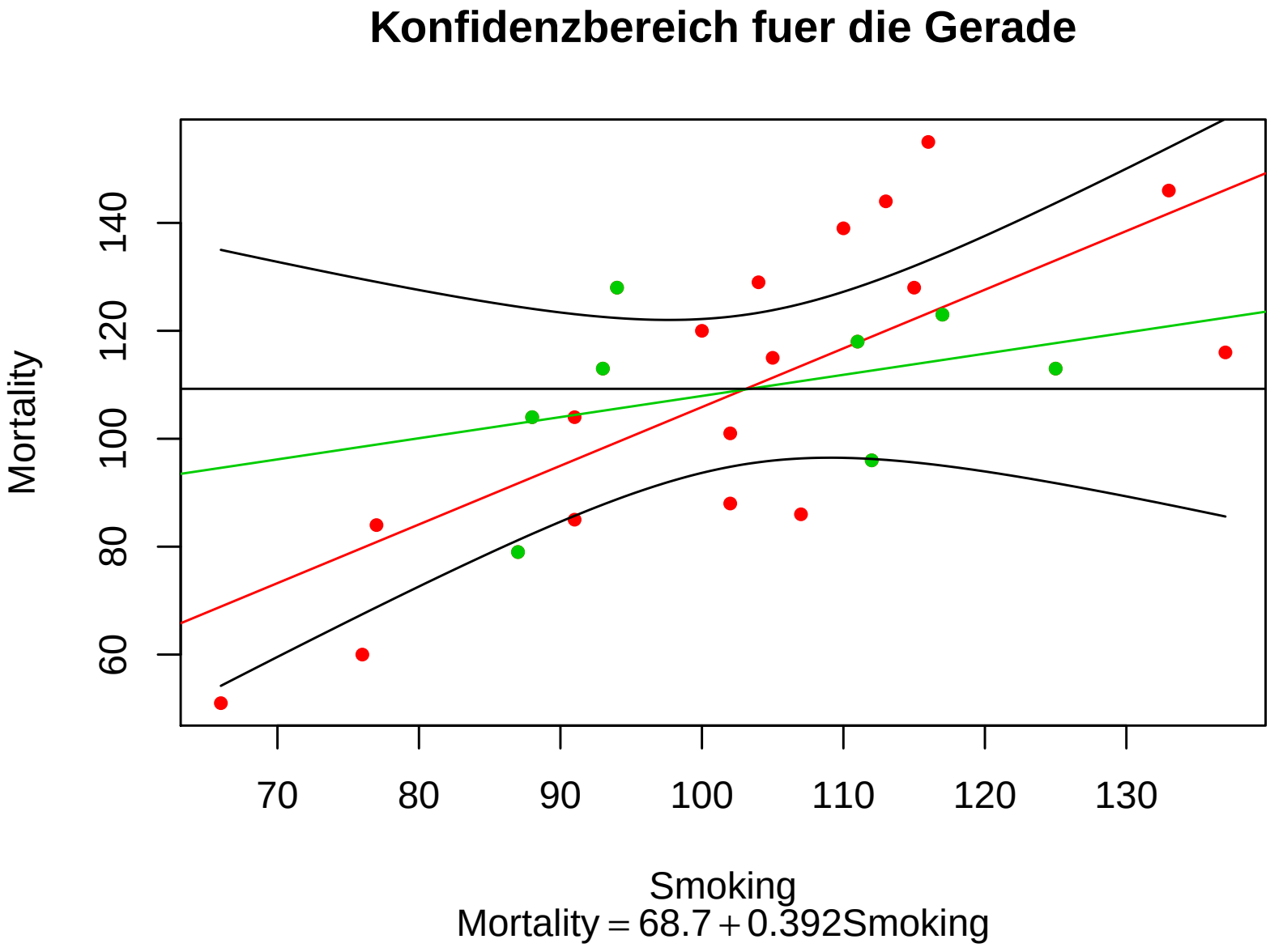
Nochmal mit weniger Daten



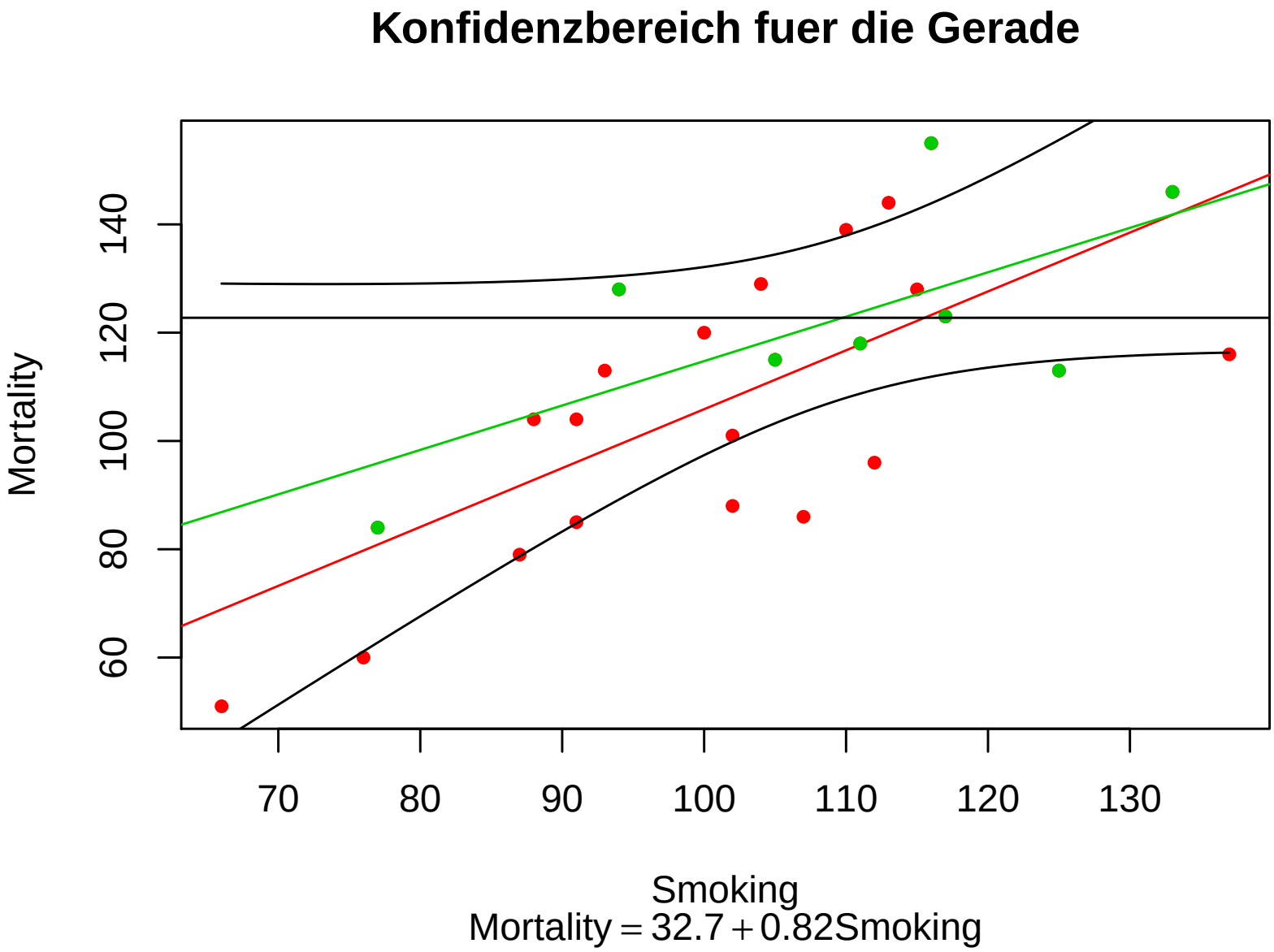
Nochmal mit weniger Daten



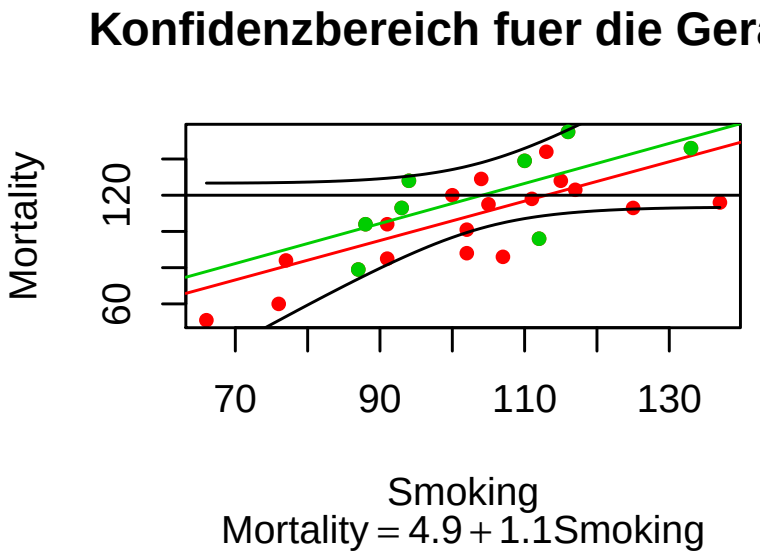
Nochmal mit weniger Daten

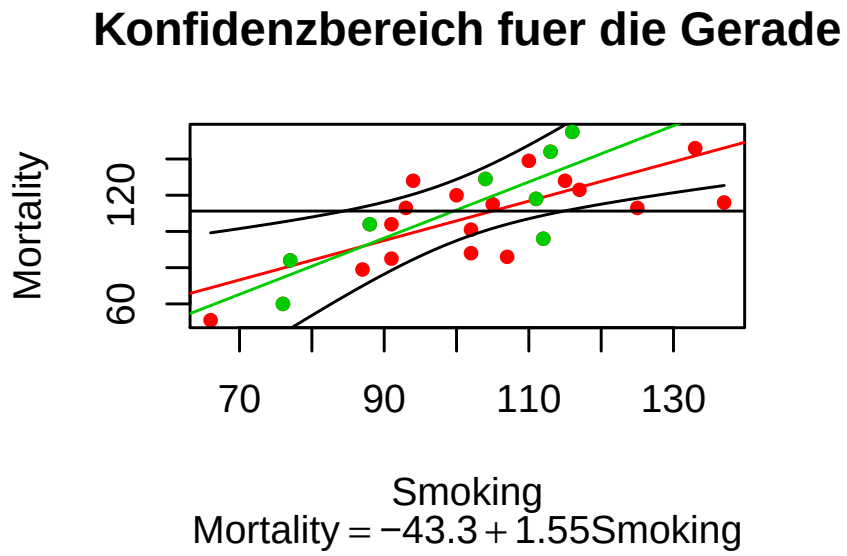


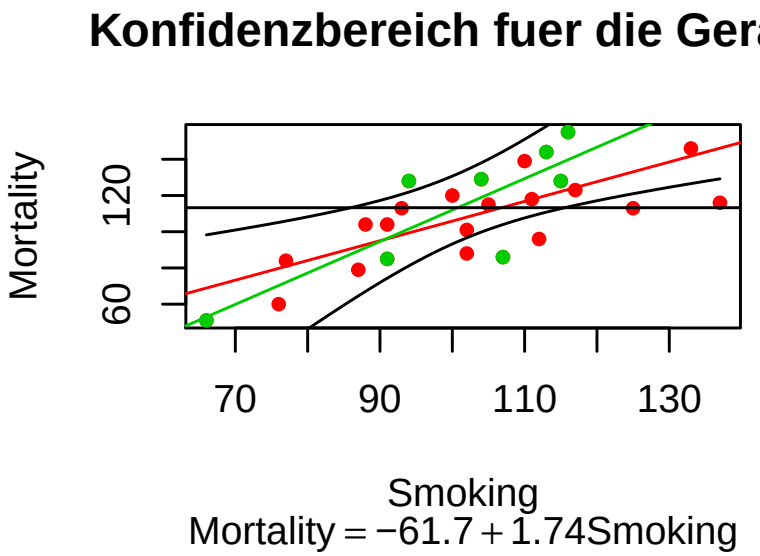
Nochmal mit weniger Daten

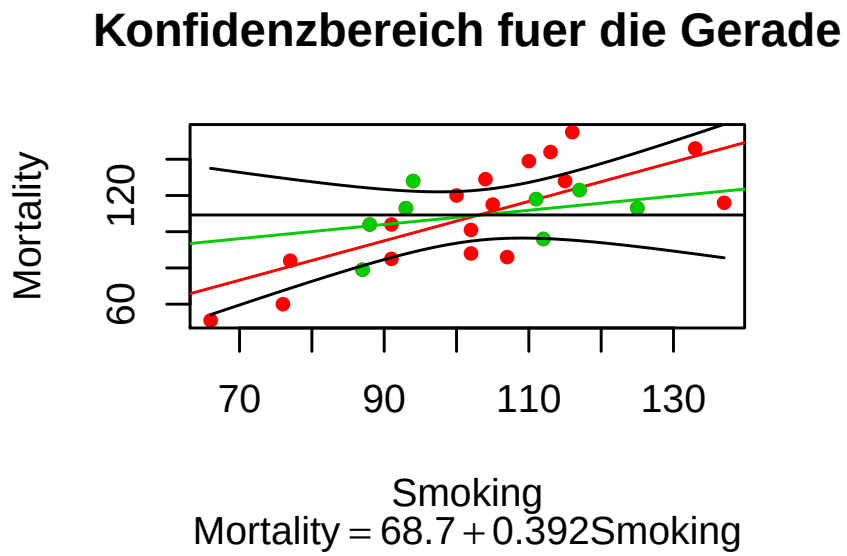


Nochmal mit weniger Daten









Regressionstest

- **Regressionstest**

  Situation:

  Test ob die Steigung in der Regression von 0 verschieden ist.

  Voraussetzungen:

  repräsentative Daten
  Regressions Modell: $Y_i \sim a + b * X_i + \epsilon_i$
  $\epsilon_i \sim N(0, \sigma^2)$ *stochastisch unabhängig*
  σ^2 *ist unbekannt*

  Testproblem:
  $H_0 : b = 0$
  $H_1 : b \neq 0$
  Bemerkung:

  Voraussetzungen mit Diagnostik prüfen

  Befehl: `anova(lm(Y X))`

Computerausgabe

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

Analysis of Variance Table

Response: Mortality
          Df Sum Sq Mean Sq F value  Pr(>F)
Smoking    1   8396    8396    24.2 5.7e-05 ***
Residuals 23   7970     347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

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

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

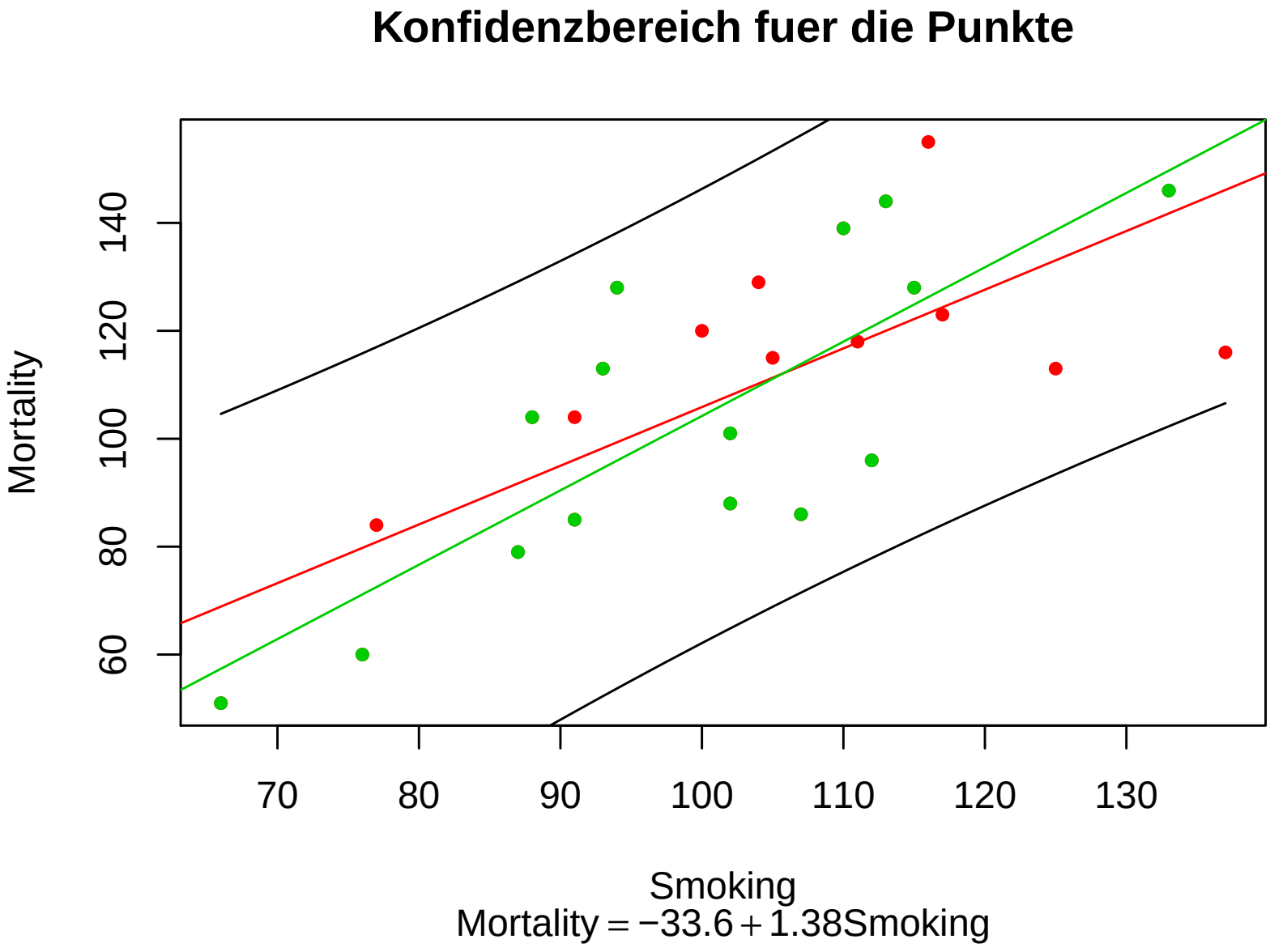
ohne Gewähr.

Formeln

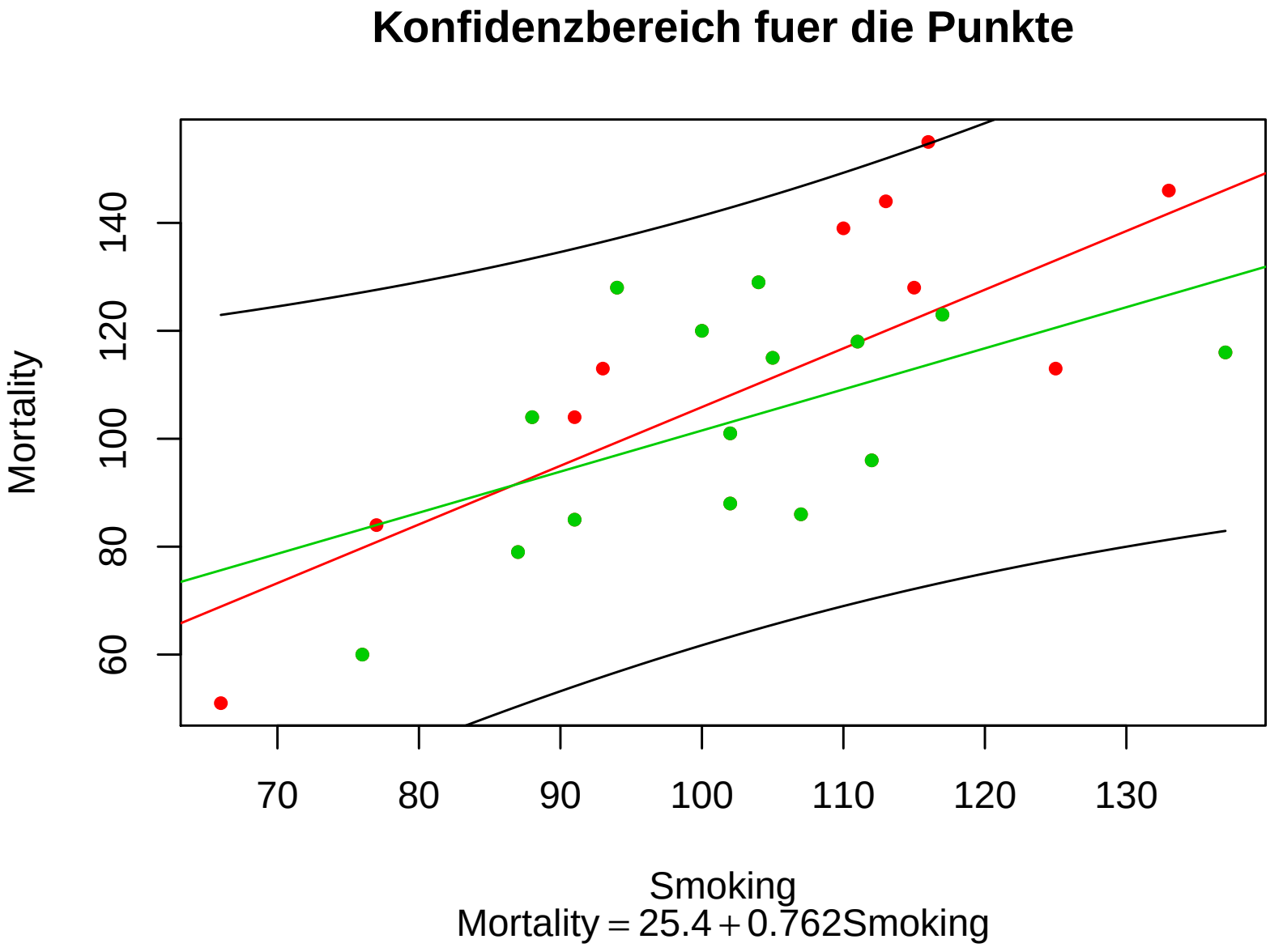
$$
\begin{aligned}
c_0 &= \frac{1}{n\sum X_i^2 - \left(\sum_i X_i\right)^2} \\
c_1 &= c_0 \sum X_i^2 \\
c_2 &= -c_0 \sum X_i \\
c_3 &= c_0 n \\
\hat{sd}(\epsilon) &= \frac{1}{n-2} \sum_{i=1}^{n} r_i^2
\end{aligned}
$$

ohne Gewähr

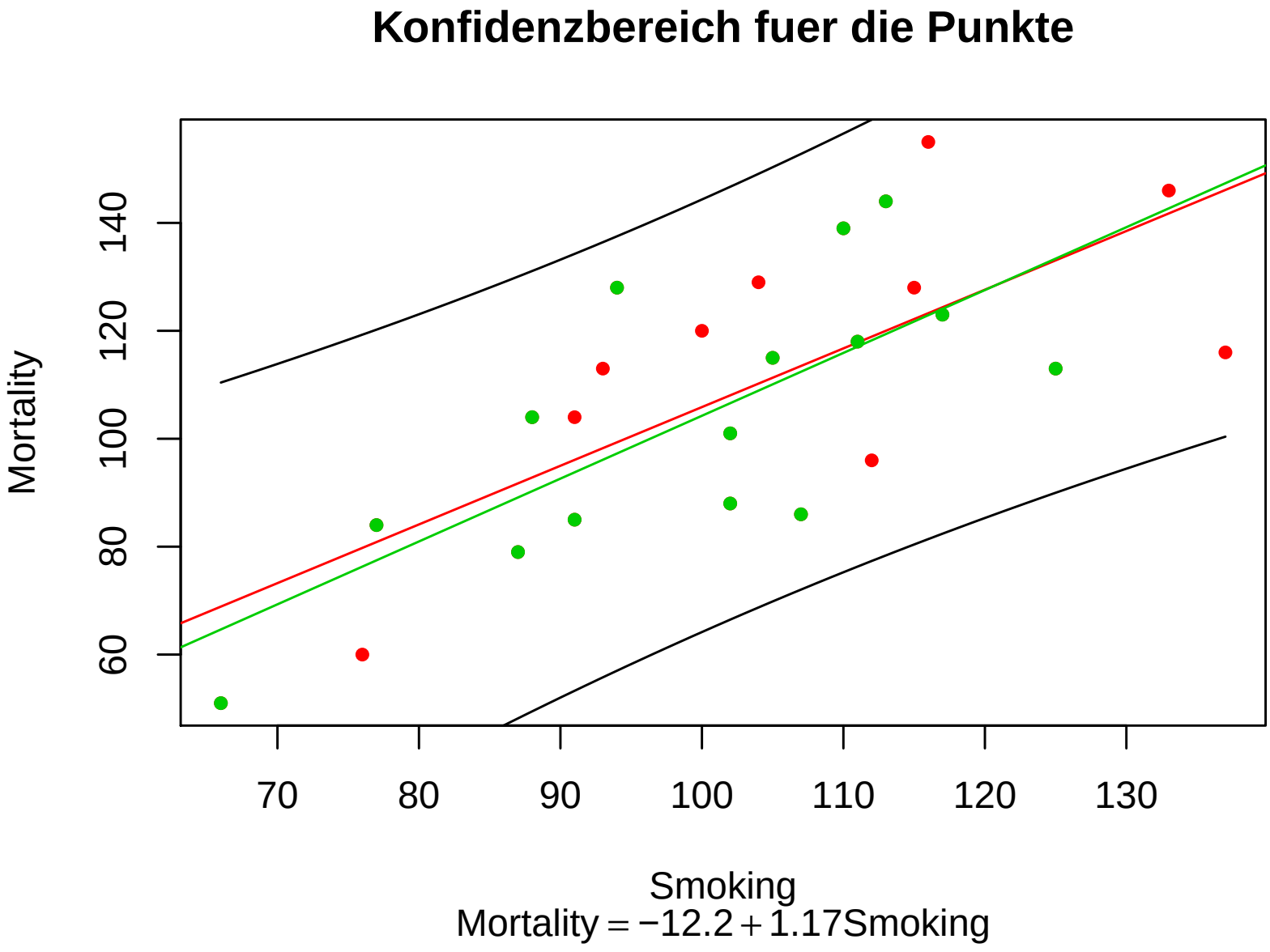
Demonstration des Vorhersageintervalls



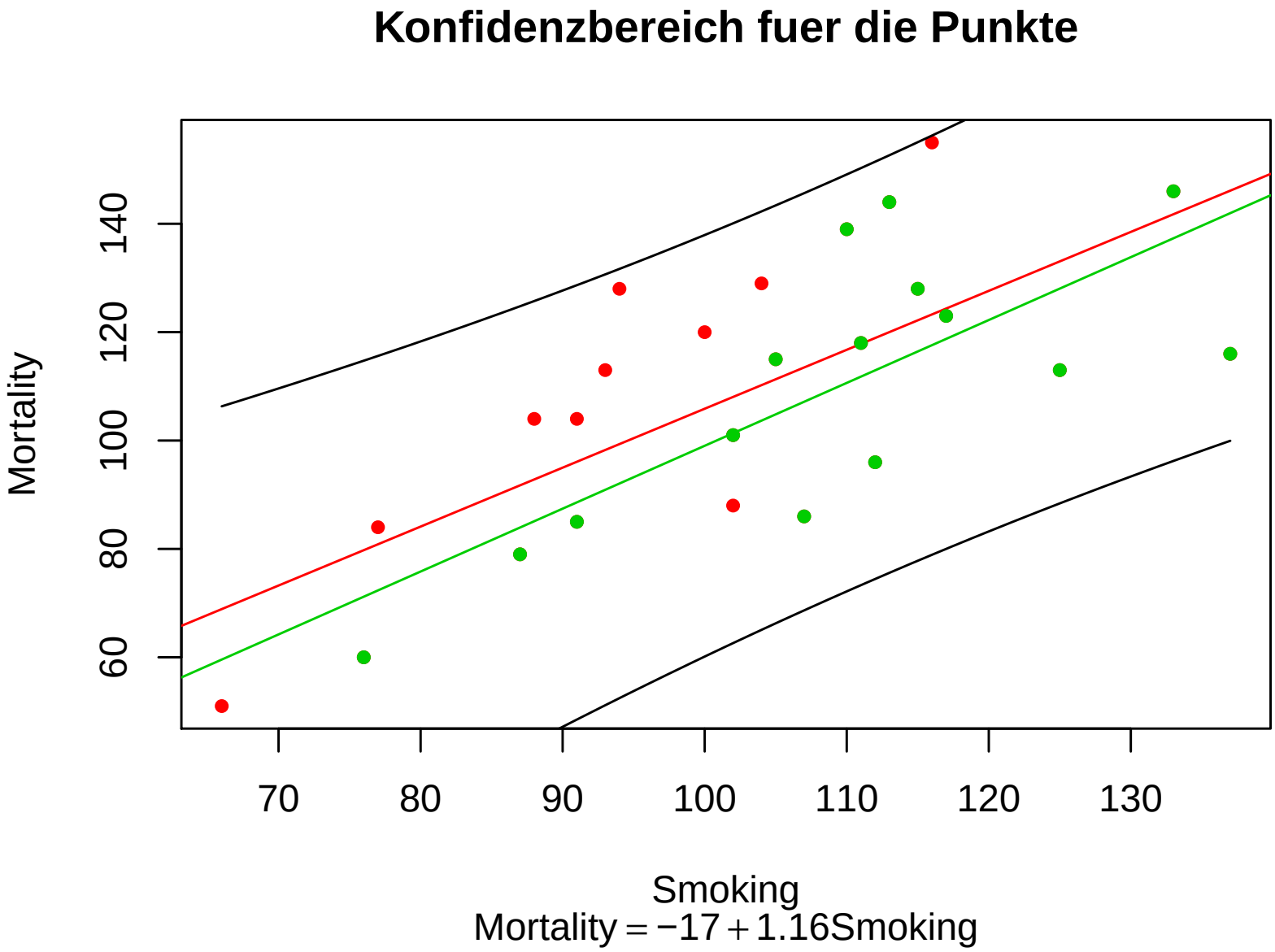
Demonstration des Vorhersageintervalls



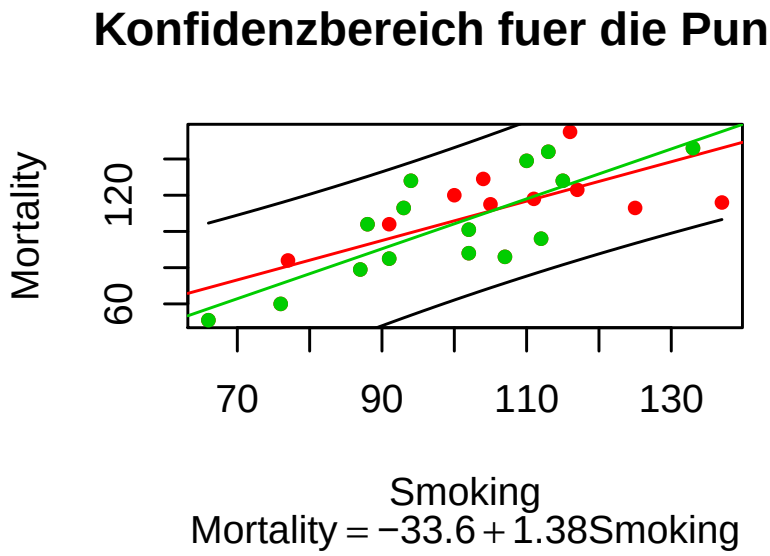
Demonstration des Vorhersageintervalls

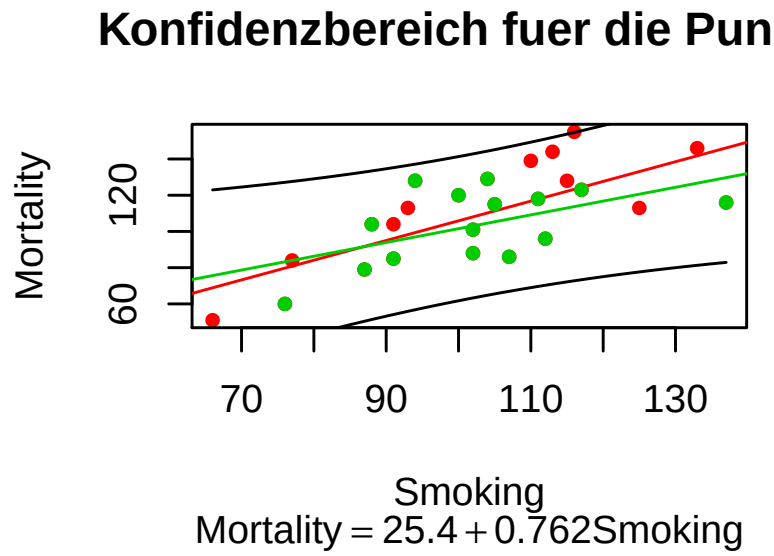


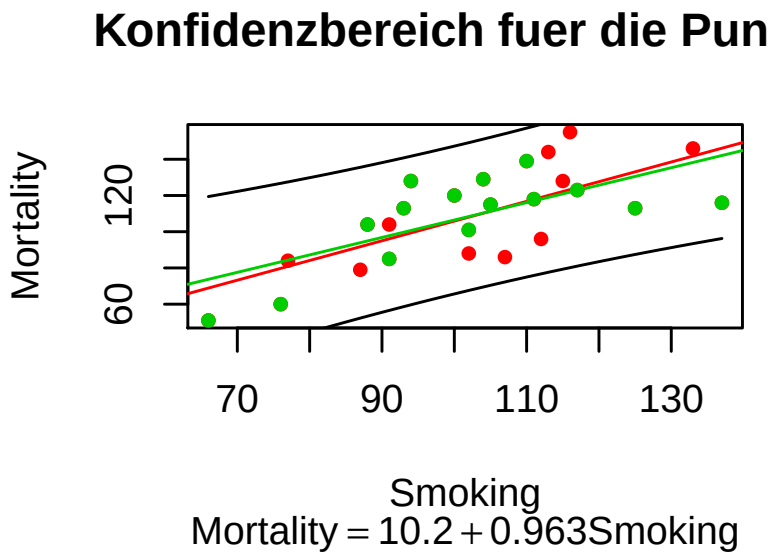
Demonstration des Vorhersageintervalls

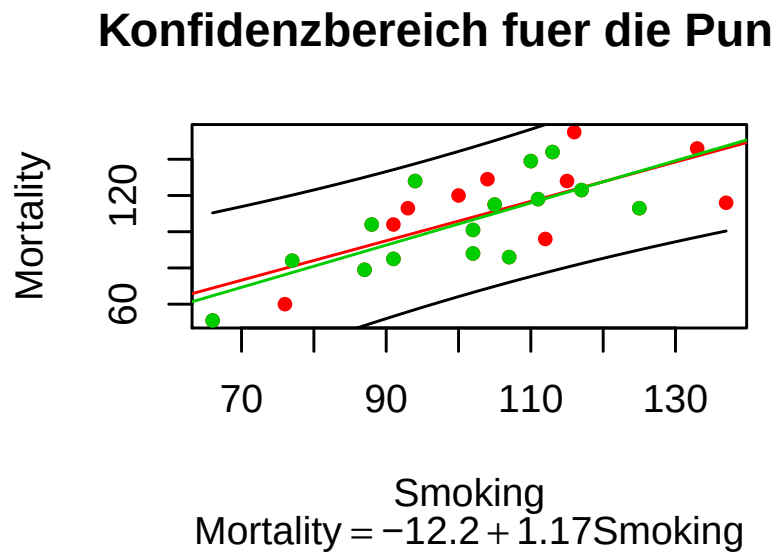


Demonstration des Vorhersageintervalls









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

$$\text{Steigung} = 0 \Leftarrow X, Y \text{ unabhängig.}$$

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:

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

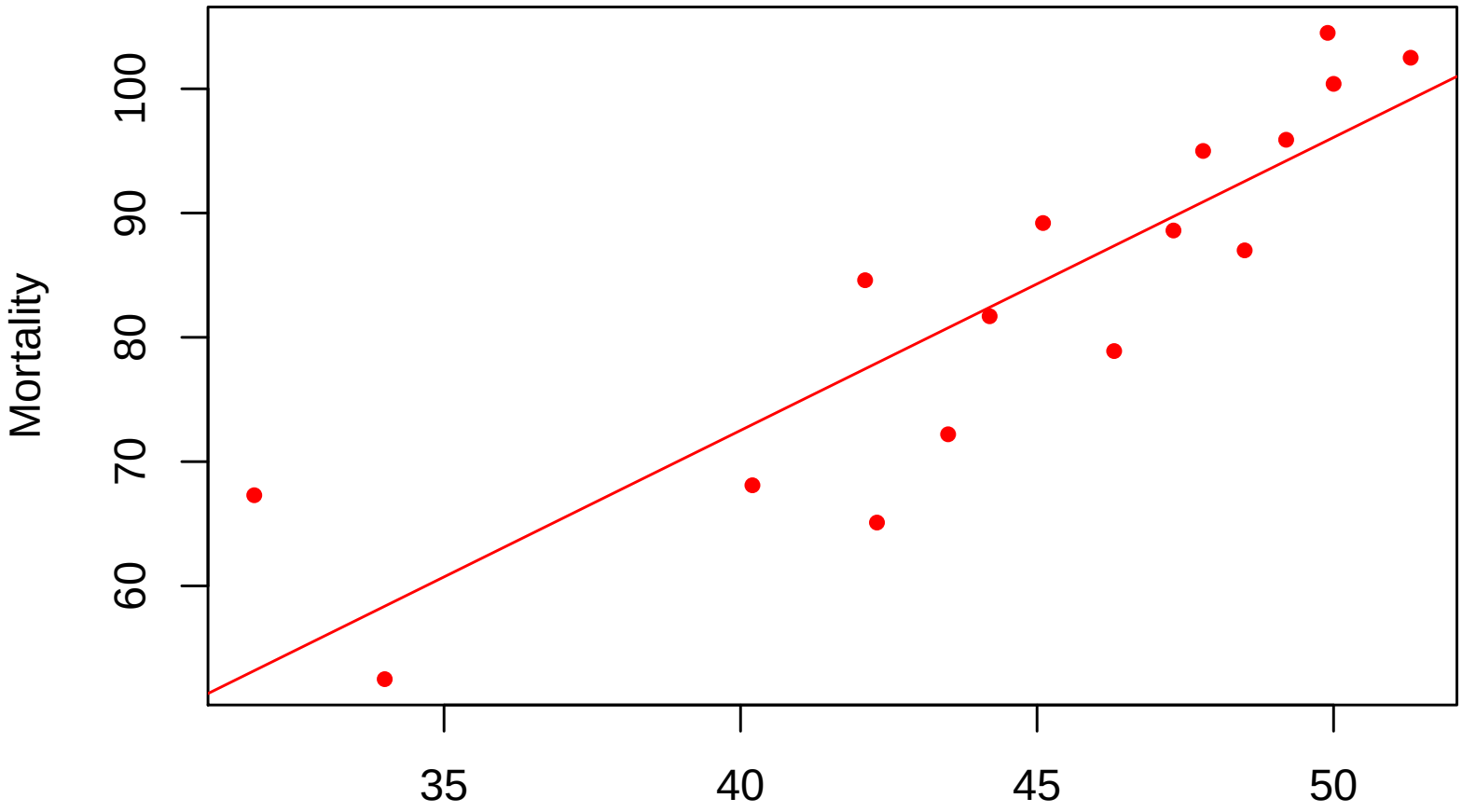
Voraussetzungen: wie Regression
Anwendung: Nachweis der Abhängigkeit

Computerausgabe

```
Analysis of Variance Table

Response: Mortality
          Df Sum Sq Mean Sq F value  Pr(>F)
Smoking    1   8396    8396    24.2 5.7e-05 ***
Residuals 23   7970     347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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

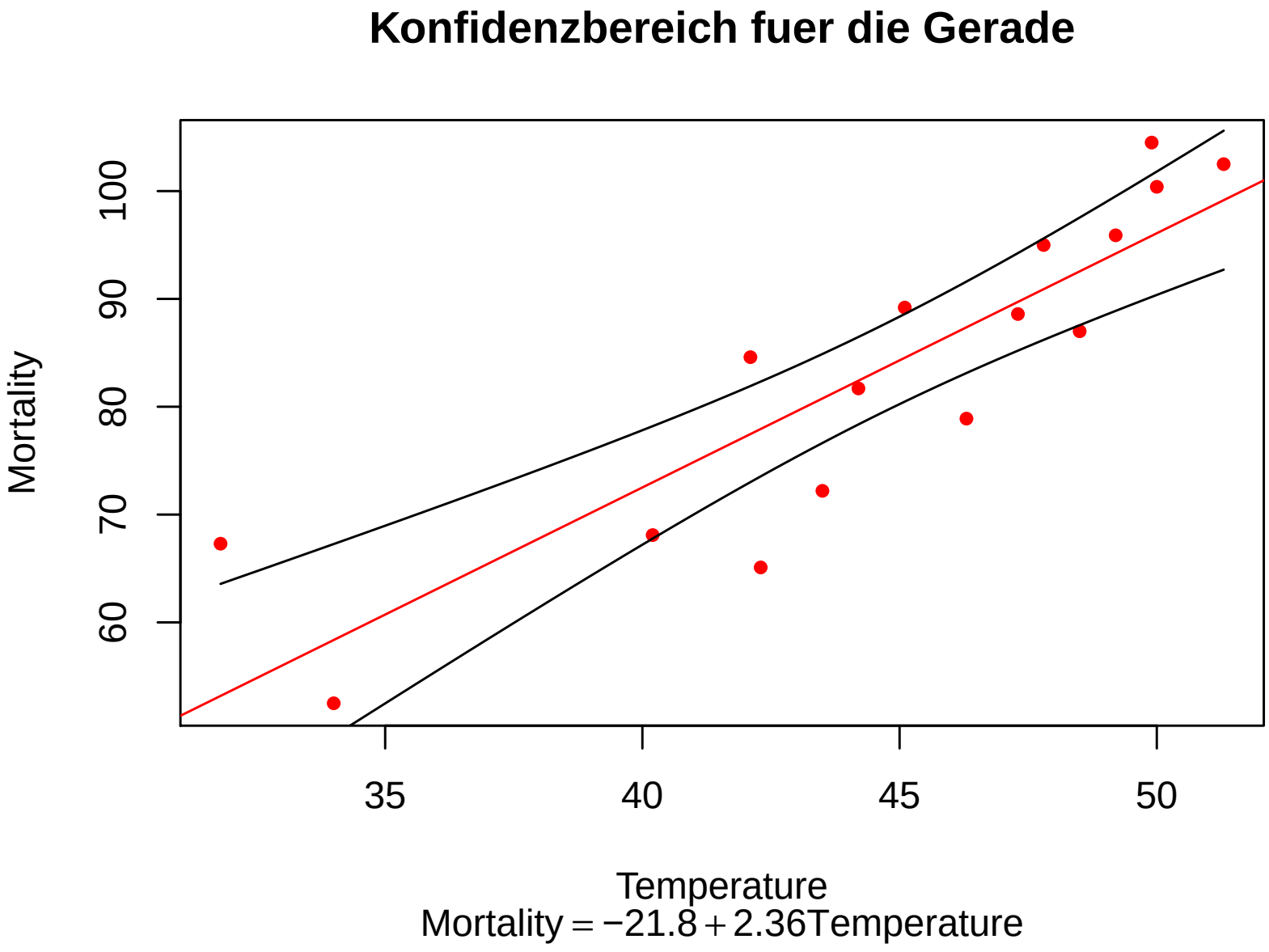
Beispiel: Brustkrebs



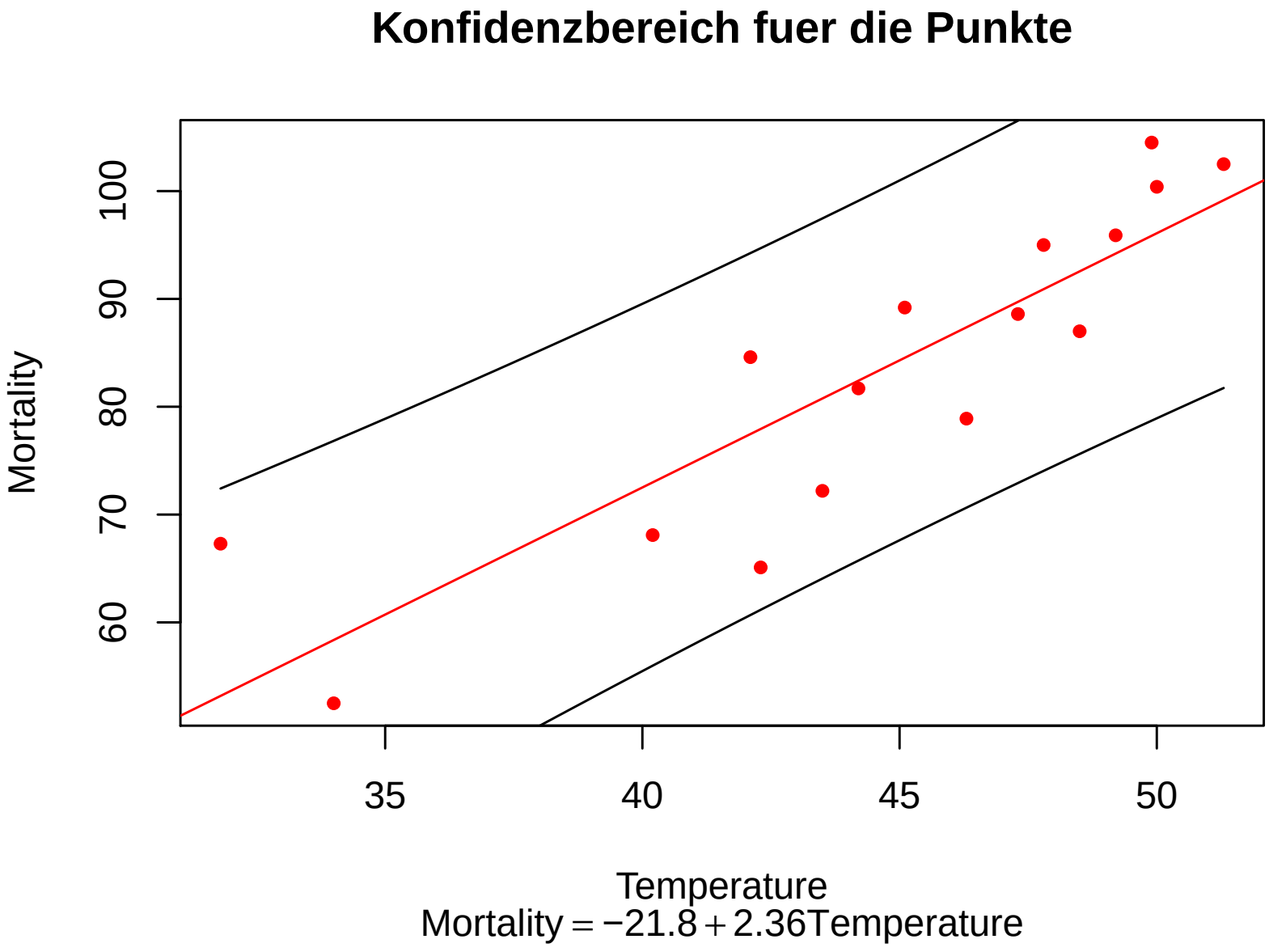
Computerausgabe

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

Coefficients:
(Intercept)  Temperature
     -21.79         2.36

Analysis of Variance Table

Response: Mortality
            Df Sum Sq Mean Sq F value  Pr(>F)
Temperature  1   2600    2600    45.7 9.2e-06 ***
Residuals   14    797      57
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

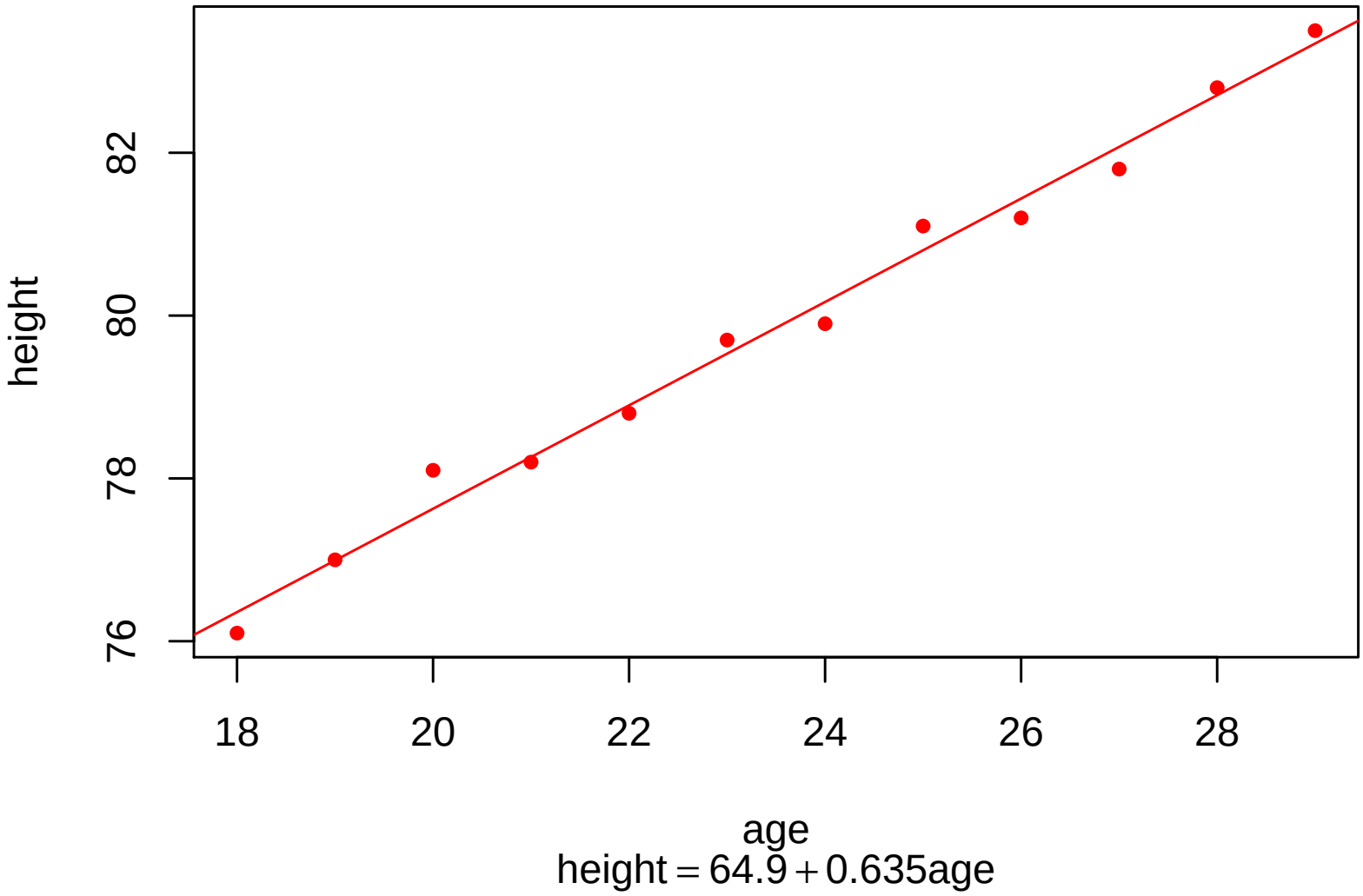
Konfidenzintervall



Vorhersageintervall



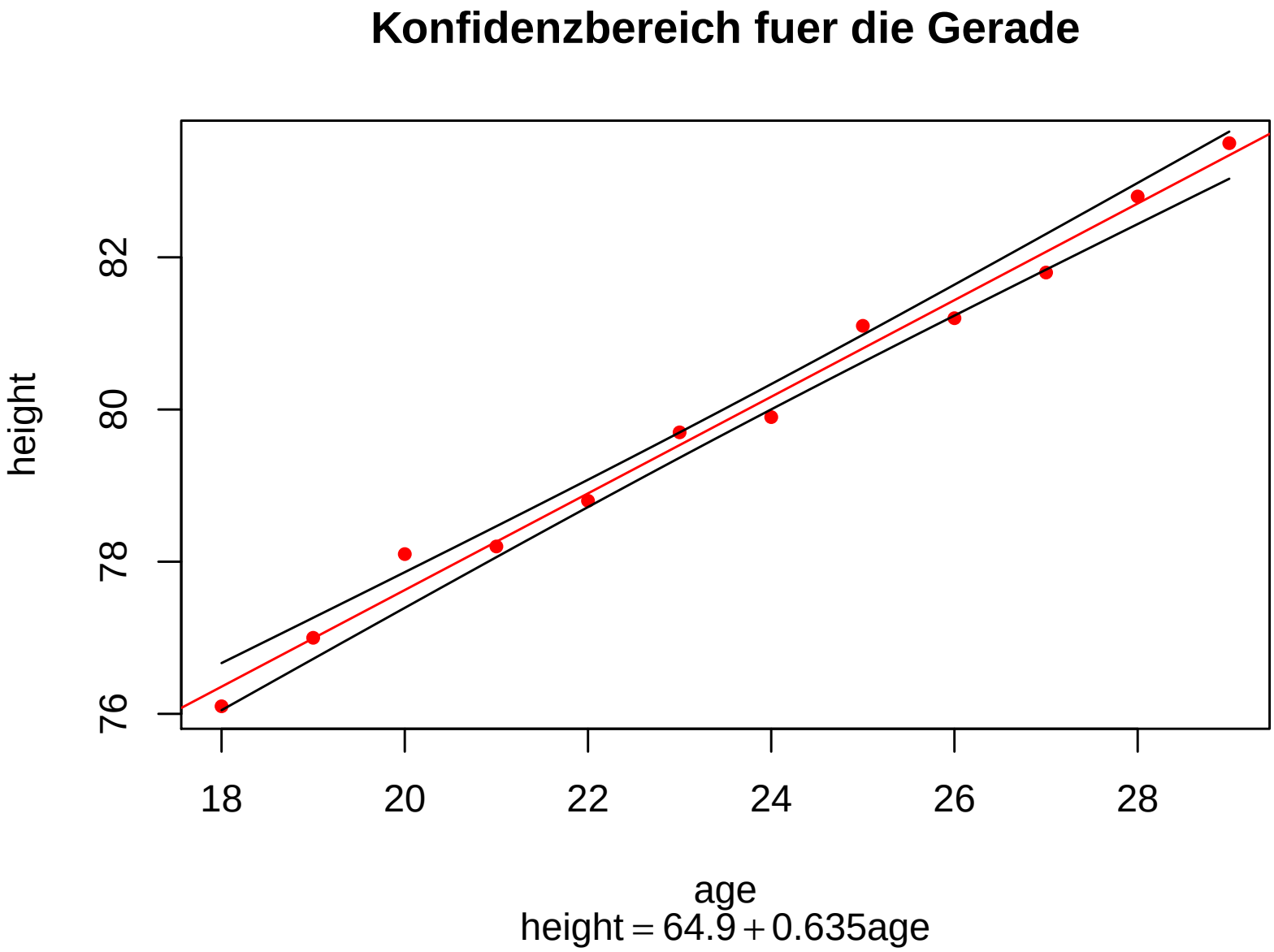
Beispiel: Wachstum



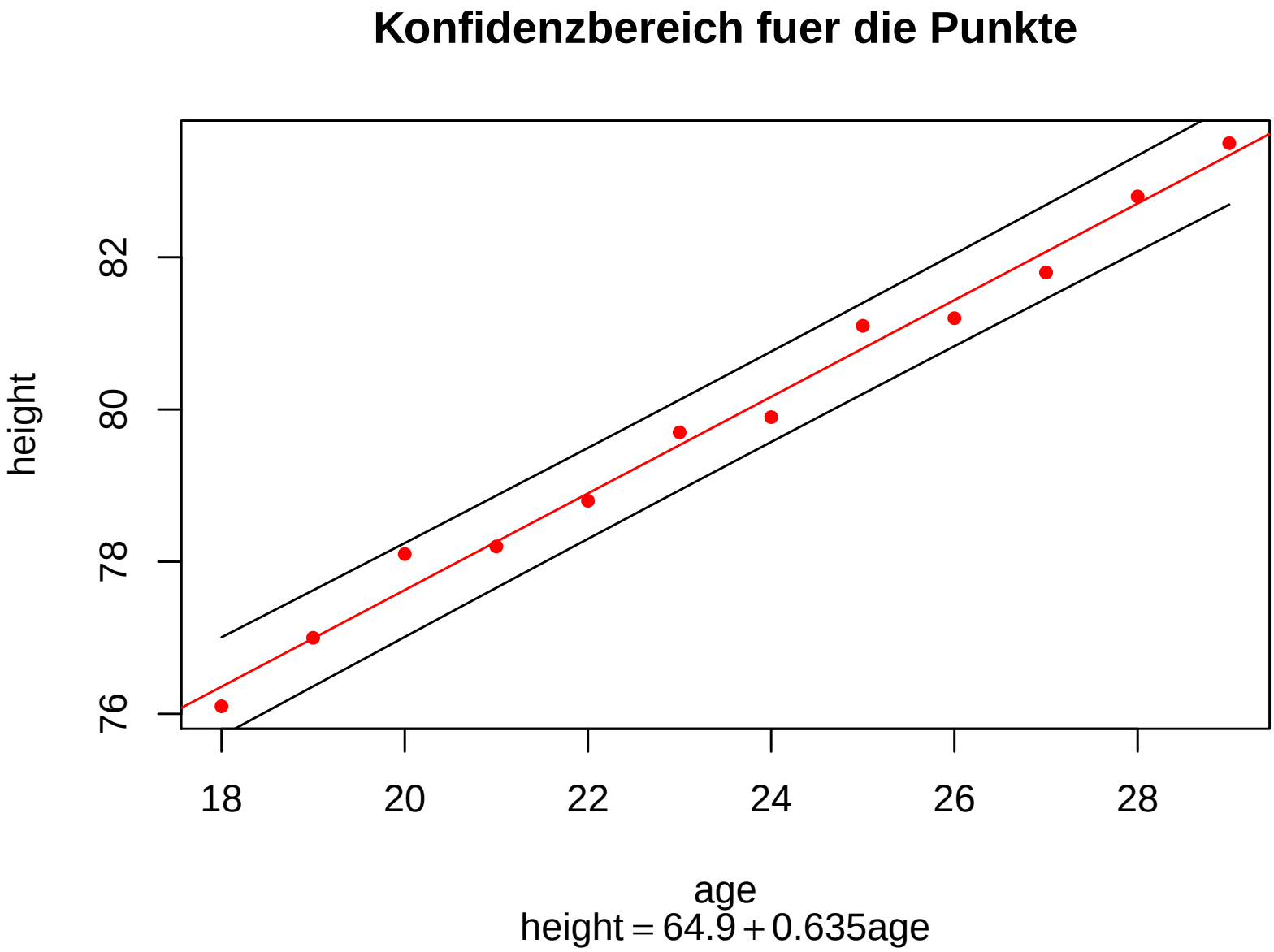
Computerausgabe

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

Coefficients:
(Intercept)          age
     64.928        0.635

Analysis of Variance Table

Response: height
          Df Sum Sq Mean Sq F value  Pr(>F)
age        1   57.7    57.7     880 4.4e-11 ***
Residuals 10    0.7     0.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

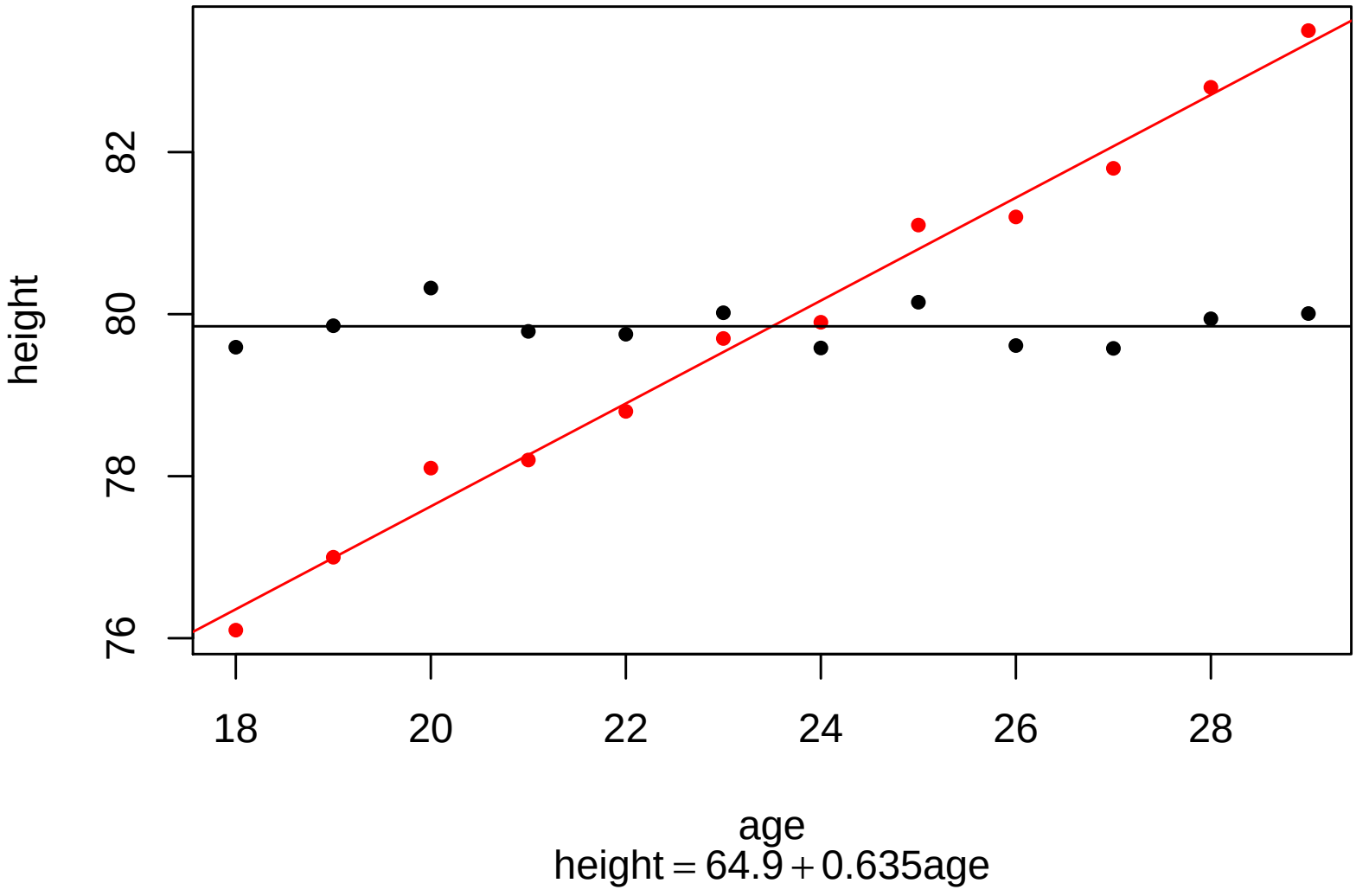
Konfidenzintervall



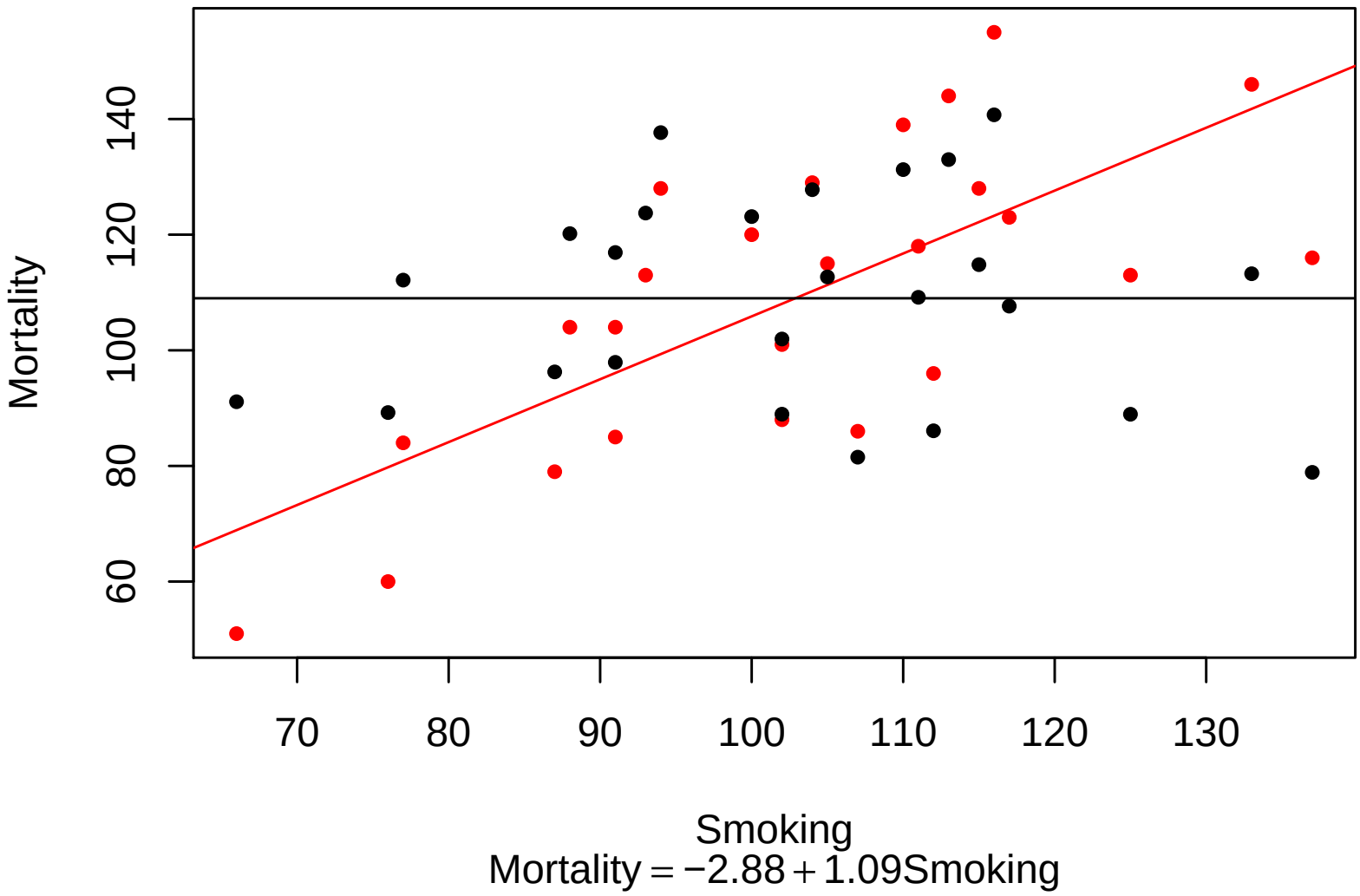
Vorhersageintervall



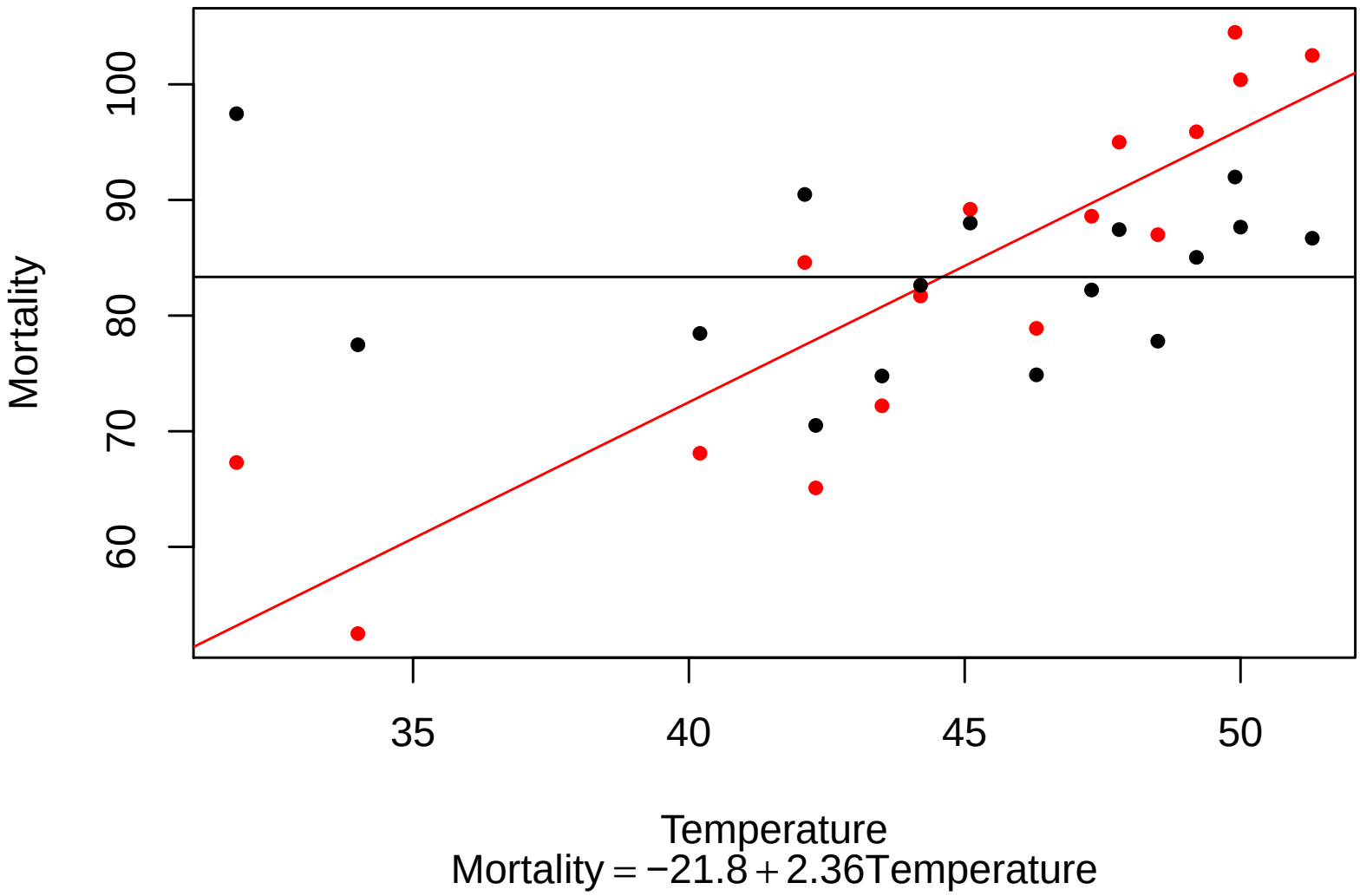
Residualstreuung



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.

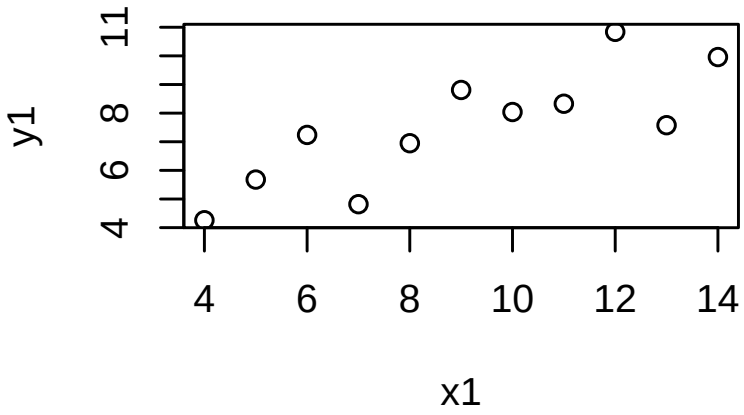
Interpretation von R^2

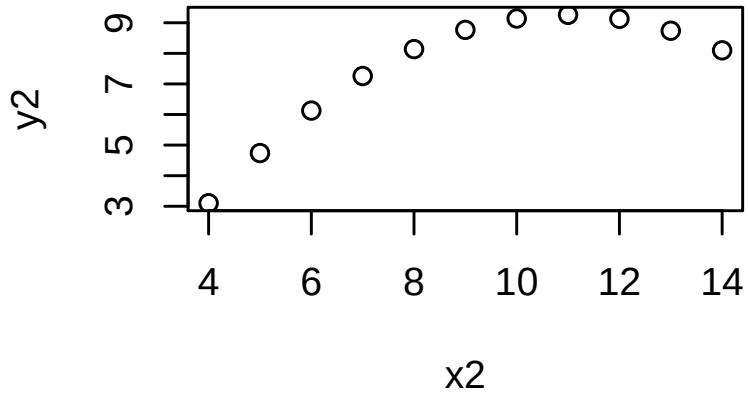
- $R^2 \in [0, 1]$
- $R^2 = 0 \Leftrightarrow$ keine Abhängigkeit erkennbar.
- $R^2 = 1 \Leftrightarrow \hat{Y}_i = Y_i$, Modell erklärt Daten perfekt.

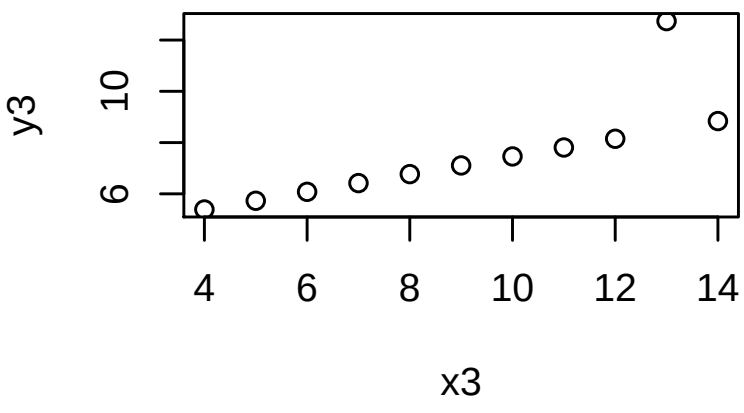
Computerausgabe

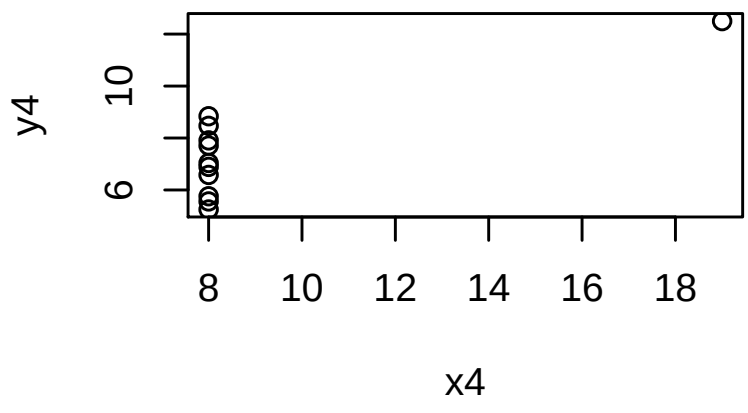
```
[1] 0.989

[1] 0.513

[1] 0.765
```

Das Anscombe Quartet

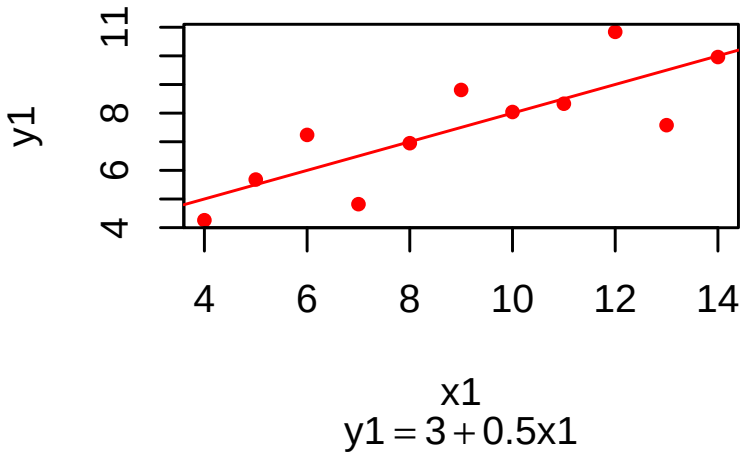


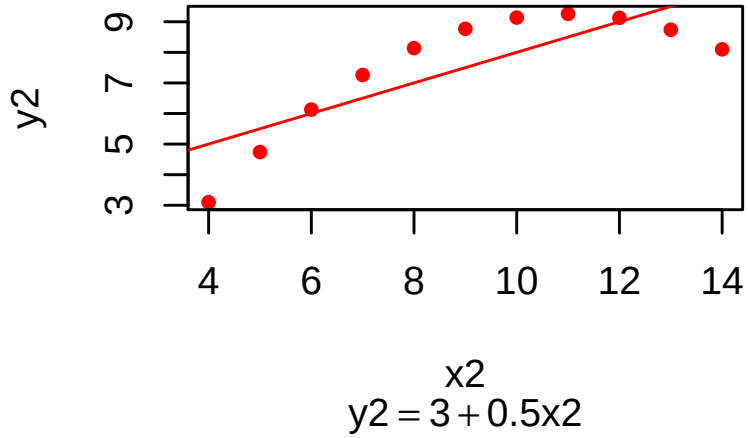


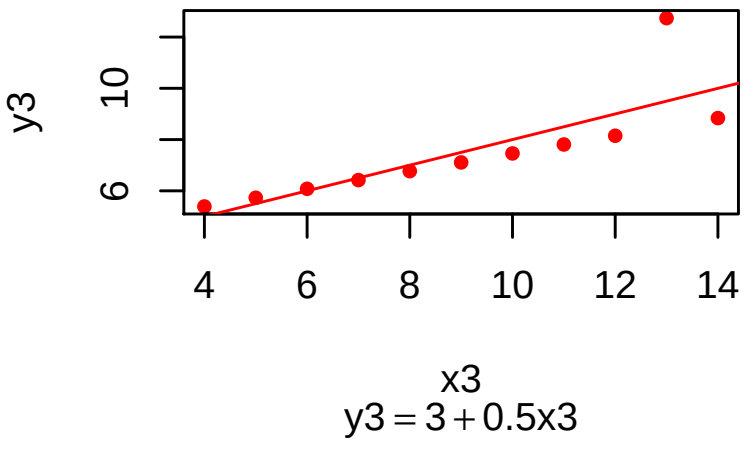


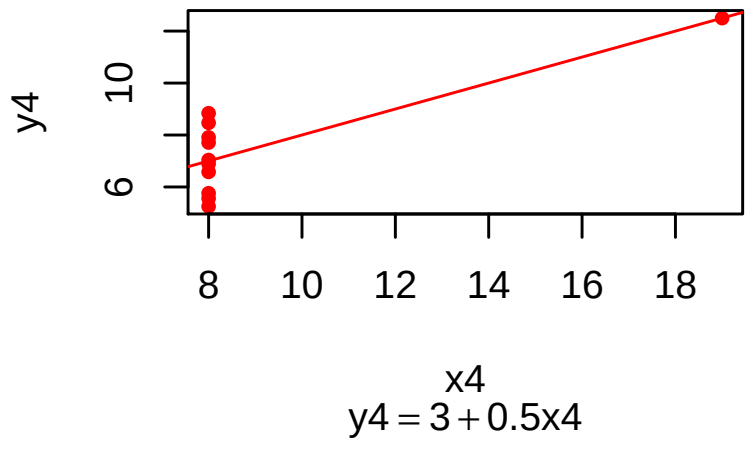


Regressionsgeraden

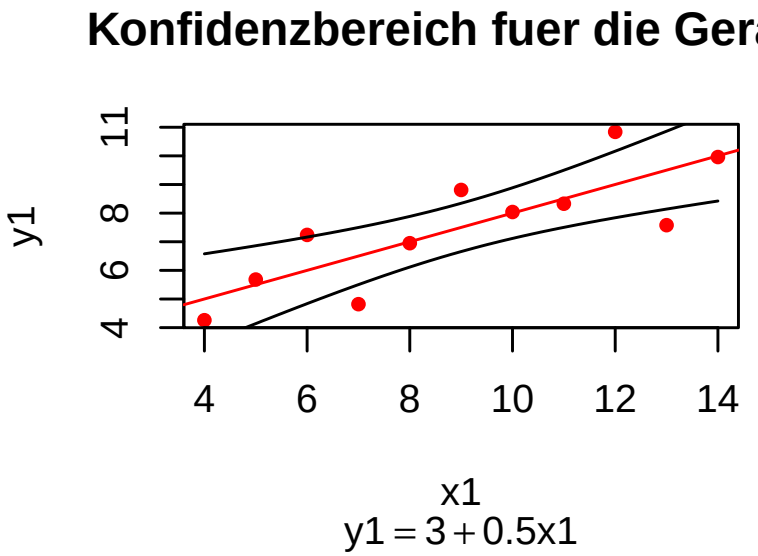


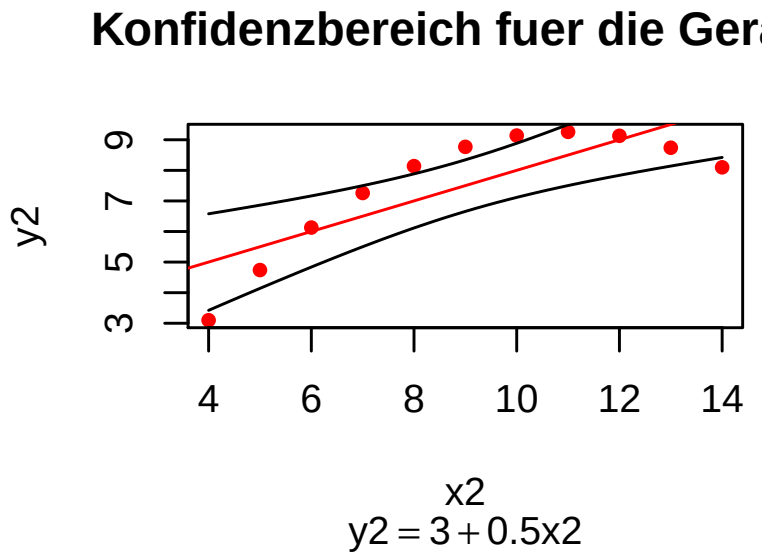


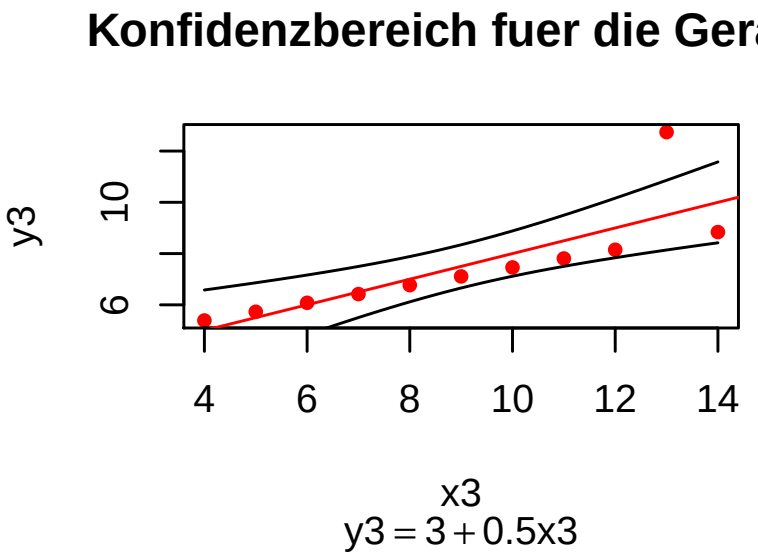


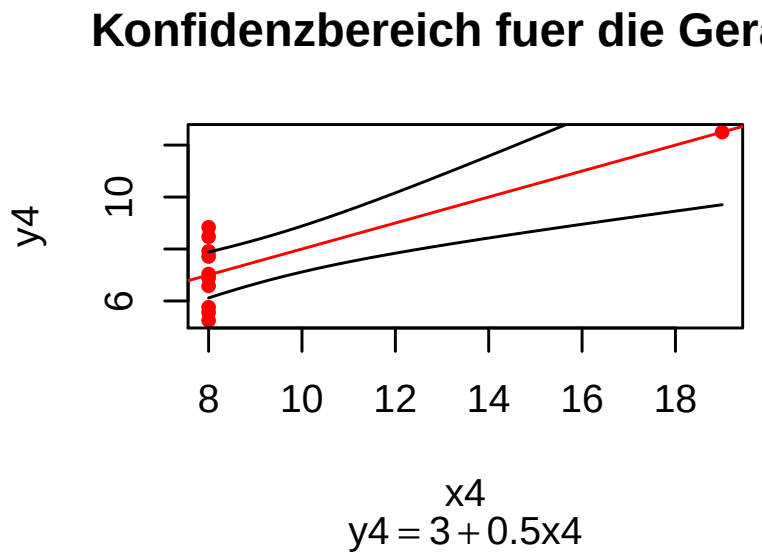


Konfidenzintervalle

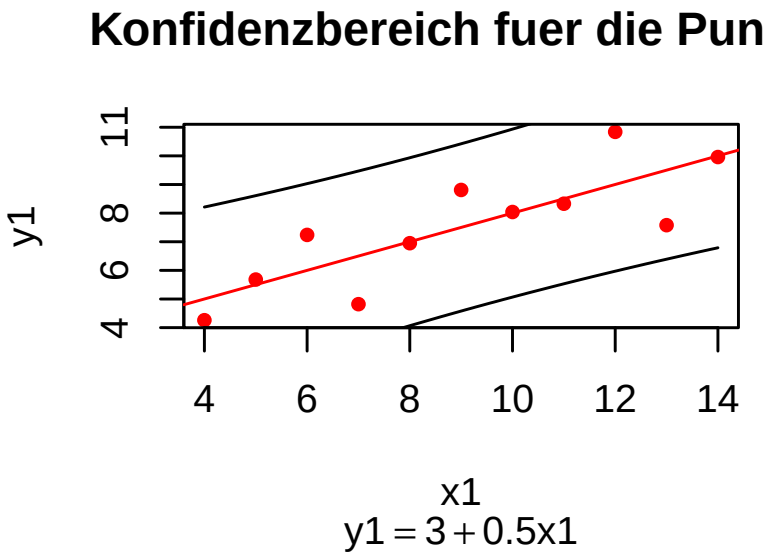


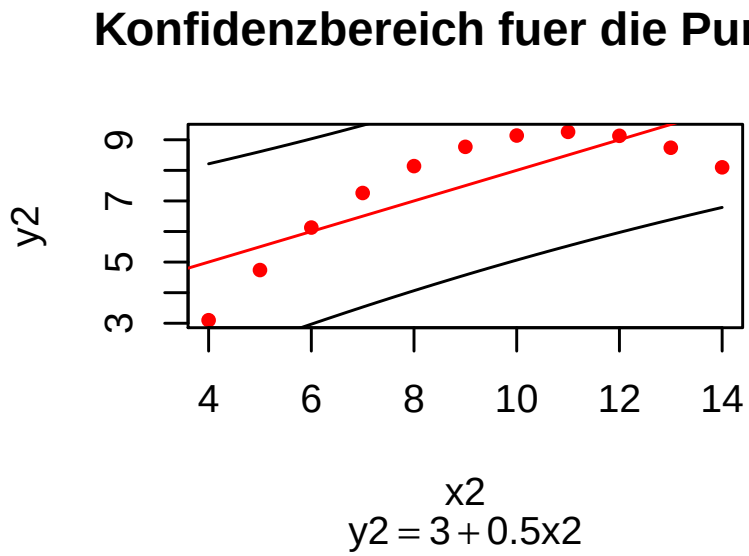


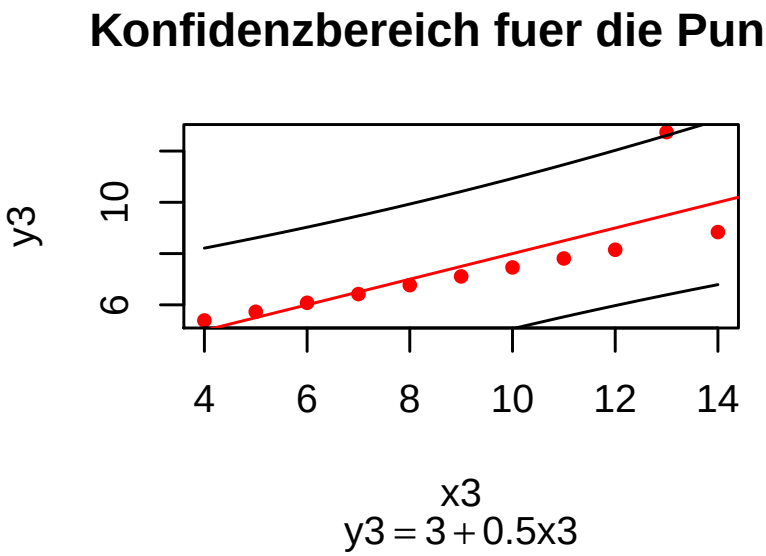


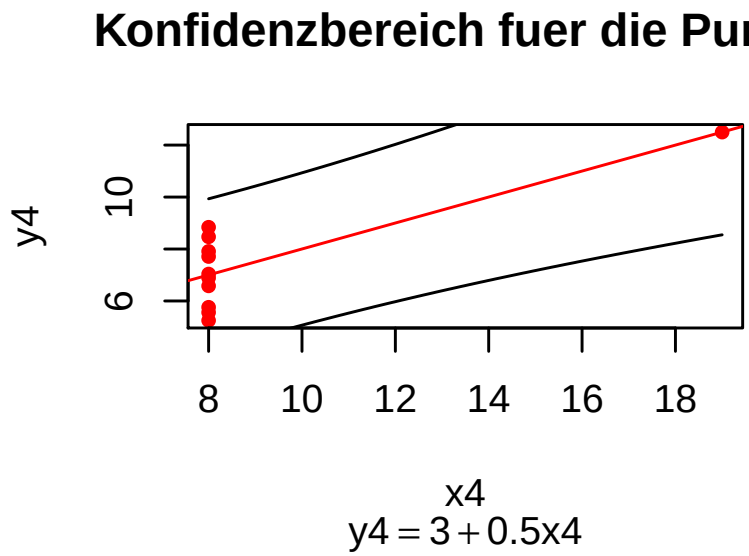


Vorhersageintervalle





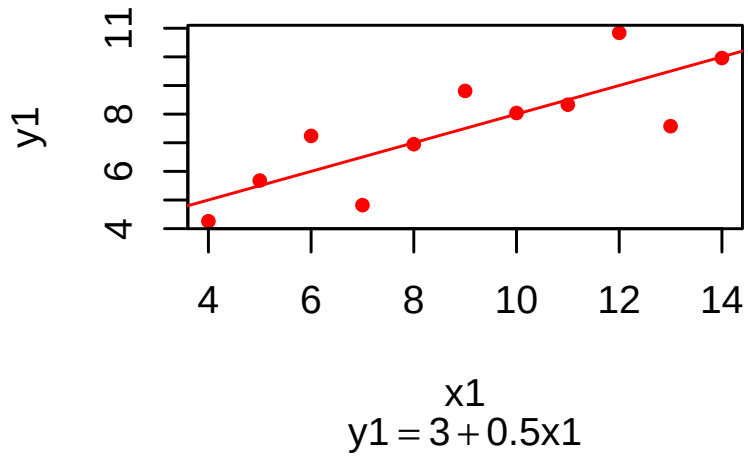


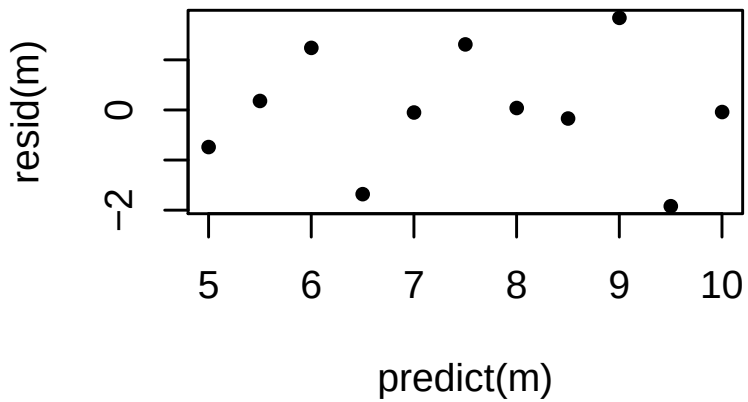


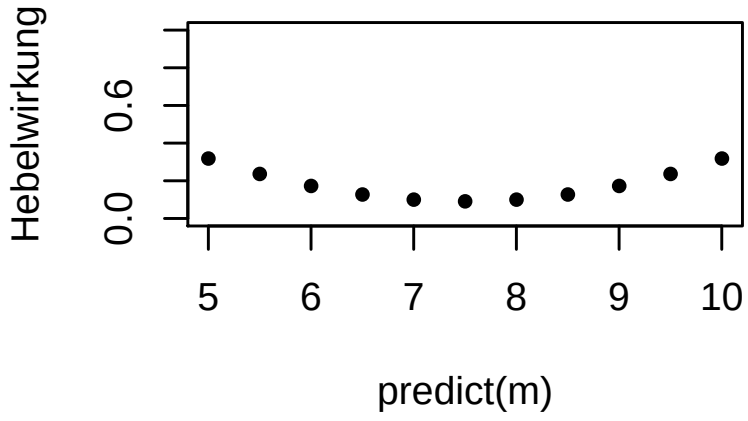
Was war das Problem?

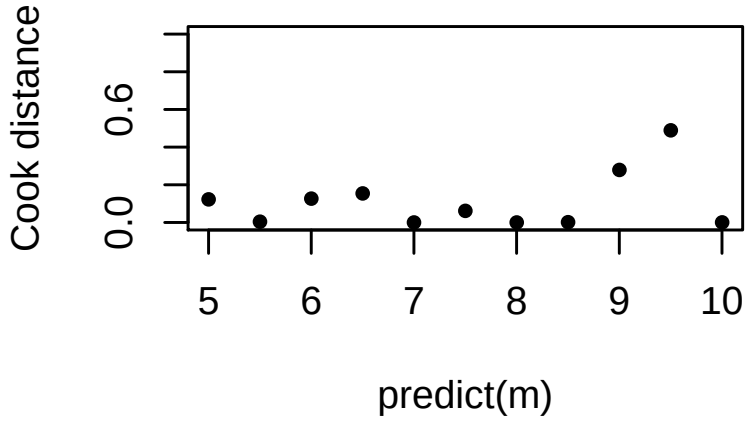
- Voraussetzungen nicht erfüllt
- Ausreißer in den Daten
- sehr einflußreiche Punkte

Anscombe 1


y1
4
6
8
11
4
6
8
10
12
14
x1
y1 = 3 + 0.5x1


resid(m)
−2
0
5
6
7
8
9
10
predict(m)


Hebelwirkung
0.0
0.6
5
6
7
8
9
10
predict(m)


Cook distance
0.0
0.6
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

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

"Wer weniger Kraft braucht sitzt am längeren Hebel"

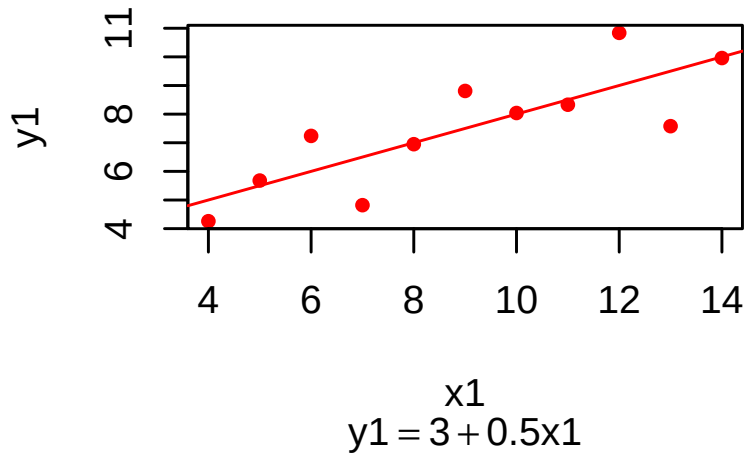
- Mißt die potentielle Wirkung eines Ausreißers im Regressant an dieser Stelle.
- Zeigt wie wichtig eine Beobachtung für die Schätzung ist.
- Große Werte deuten auf Unzuverlässige Schätzungen hin.

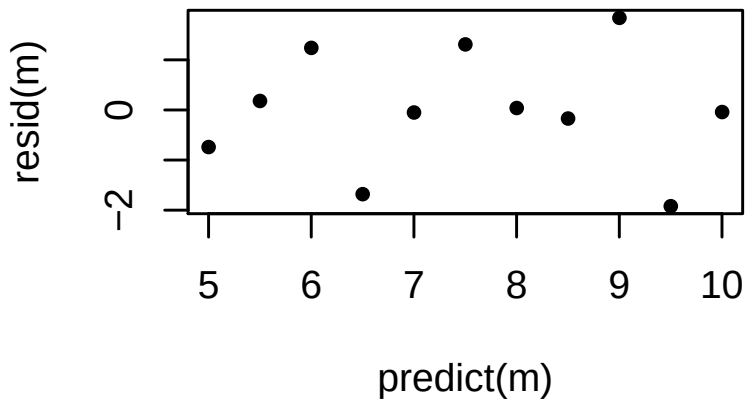
Cook's Distanz

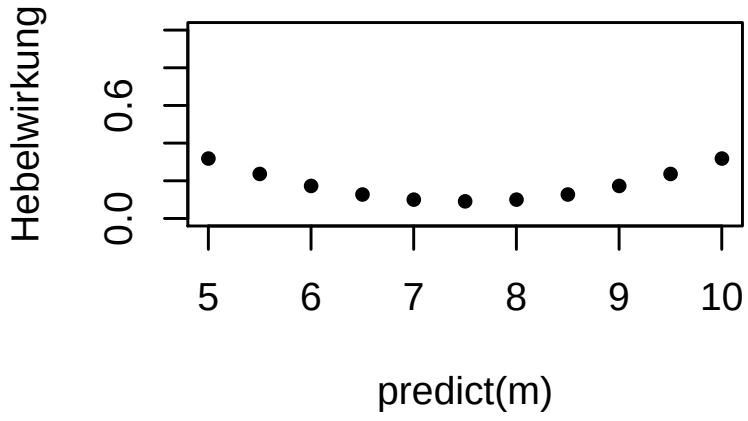
c_i = Maß für tatsächlicher Einfluß der Beobachtung Y_i

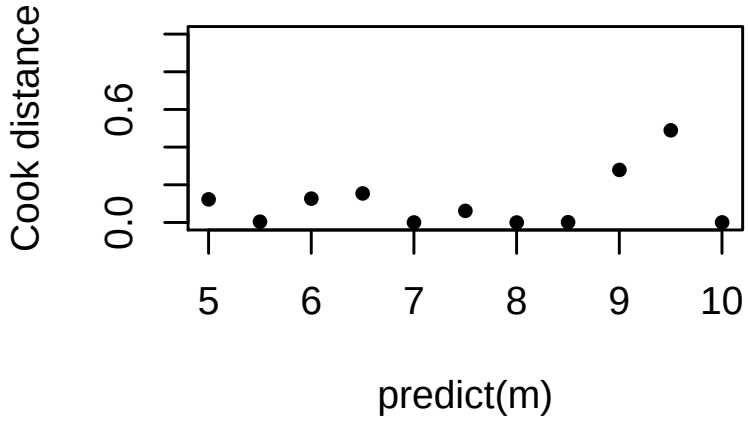
- Große Werten bedeuten, dass die Beobachtung vom dem abweicht, was die anderen Werte über diese Stelle aussagen würden.

Anscombe 1

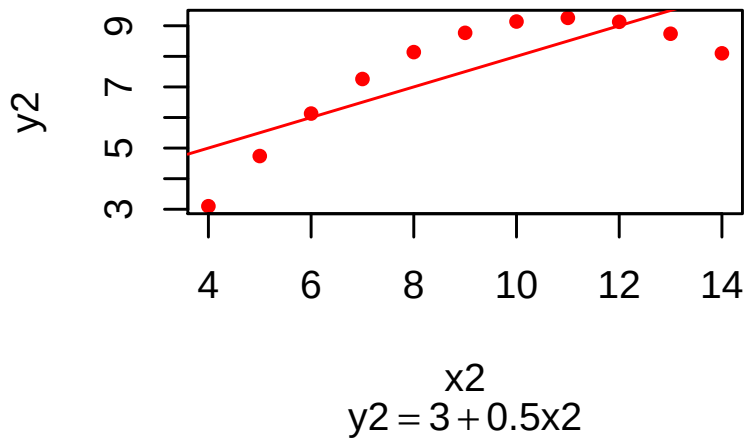

y1
4
6
8
11
4
6
8
10
12
14
x1
y1 = 3 + 0.5x1

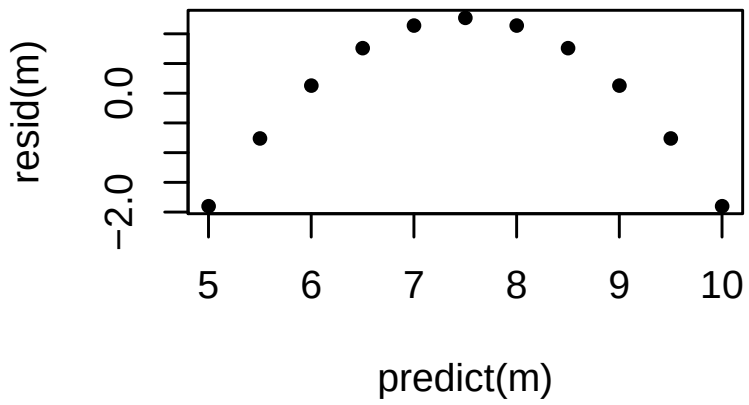

resid(m)
−2
0
5
6
7
8
9
10
predict(m)

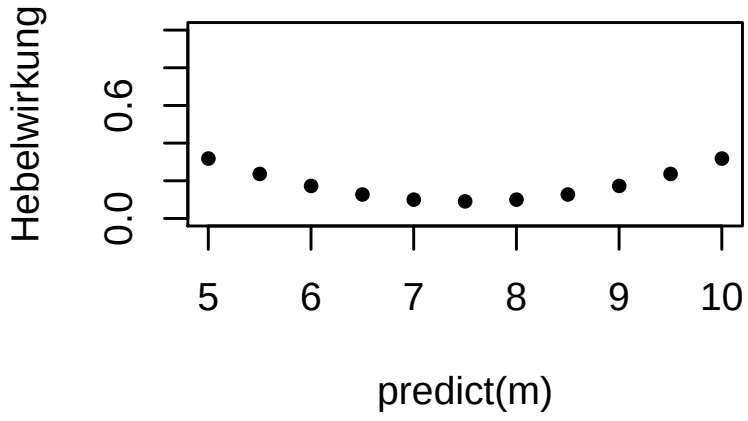

Hebelwirkung
0.0
0.6
5
6
7
8
9
10
predict(m)

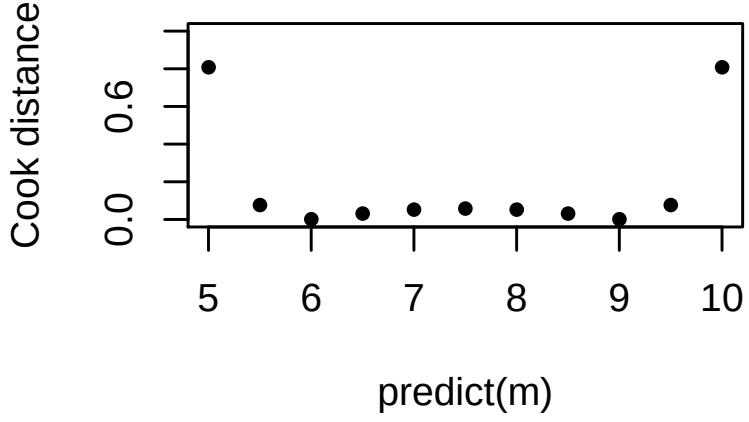

Cook distance
0.0
0.6
5
6
7
8
9
10
predict(m)

Anscombe 2

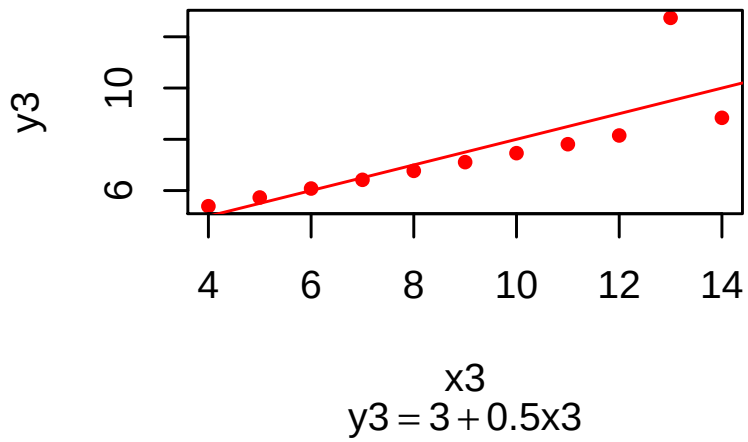

y2
3
5
7
9
4
6
8
10
12
14
x2
y2 = 3 + 0.5x2

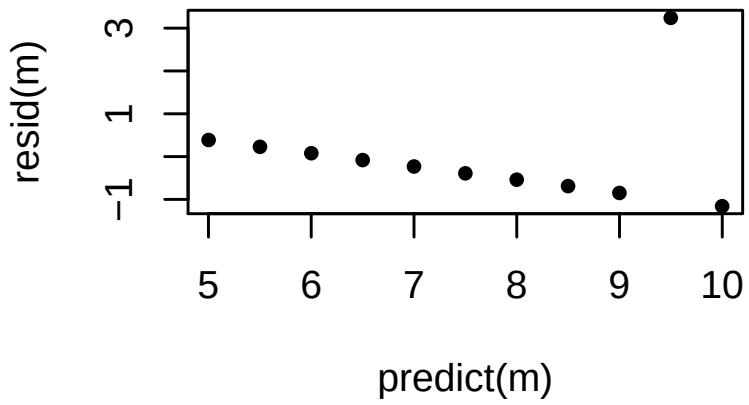

resid(m)
−2.0
0.0
5
6
7
8
9
10
predict(m)

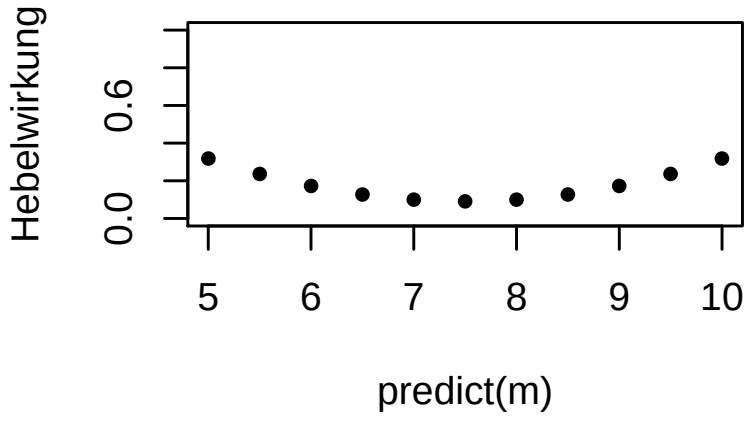

Hebelwirkung
0.0
0.6
5
6
7
8
9
10
predict(m)

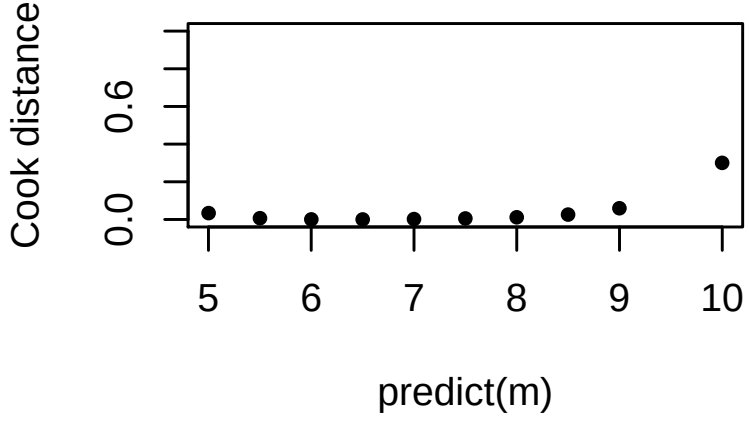

Cook distance
0.0
0.6
5
6
7
8
9
10
predict(m)

Anscombe 3

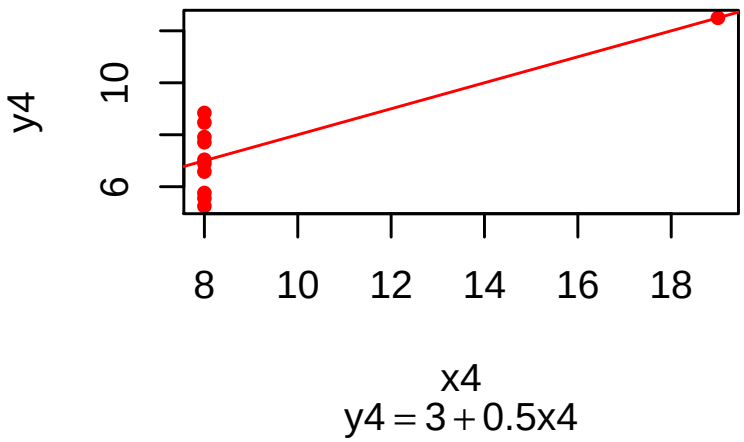

y3
6
10
4
6
8
10
12
14
x3
y3 = 3 + 0.5x3

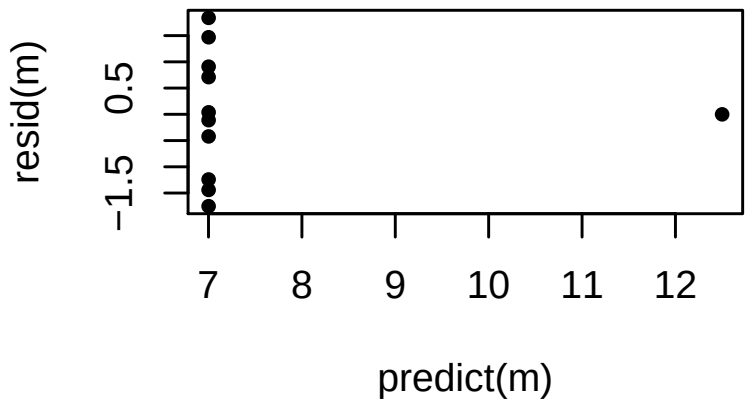

resid(m)
−1
1
3
5
6
7
8
9
10
predict(m)

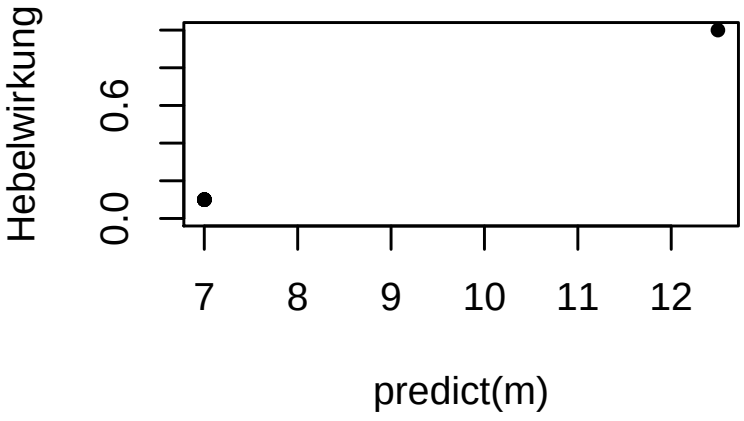

Hebelwirkung
0.0
0.6
5
6
7
8
9
10
predict(m)

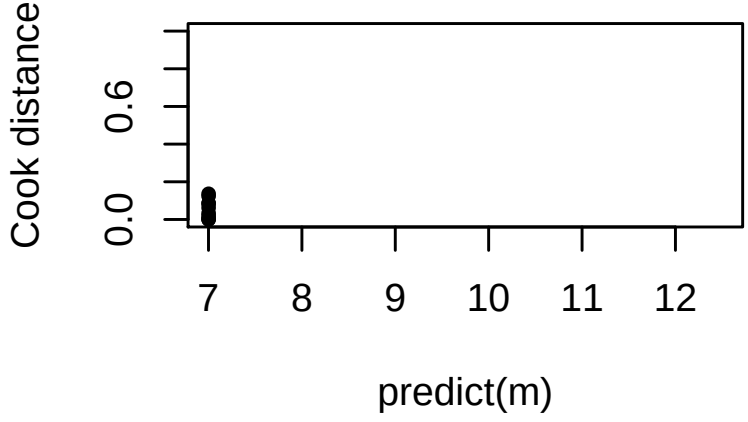

Cook distance
0.0
0.6
5
6
7
8
9
10
predict(m)

Anscombe 4

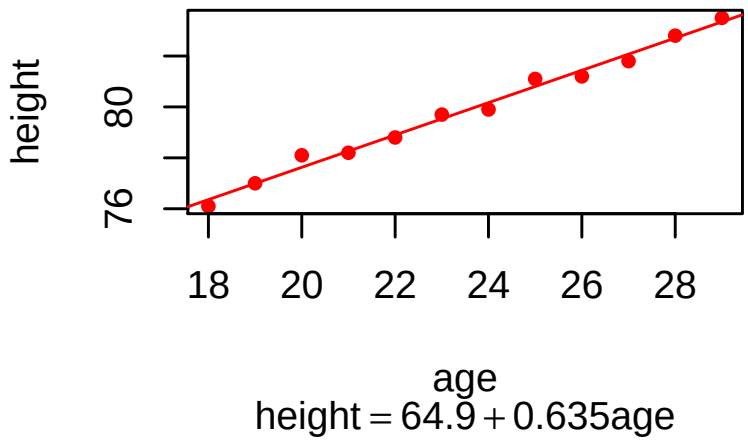

y4
6
10
8
10
12
14
16
18
x4
y4 = 3 + 0.5x4

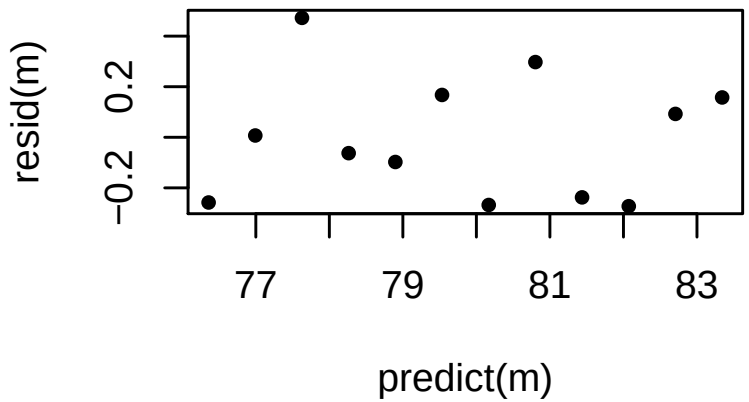

resid(m)
−1.5
0.5
7
8
9
10
11
12
predict(m)

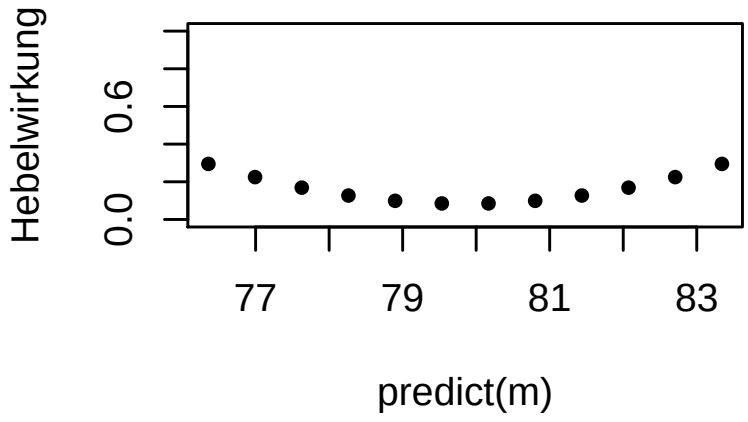

Hebelwirkung
0.0
0.6
7
8
9
10
11
12
predict(m)

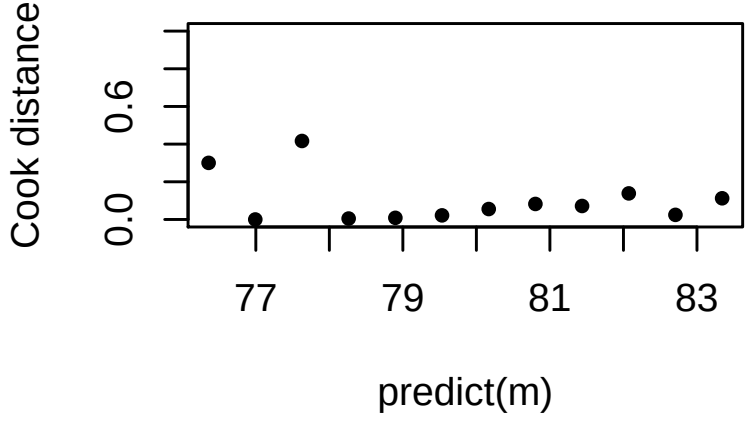

Cook distance
0.0
0.6
7
8
9
10
11
12
predict(m)

Wachstumsdaten

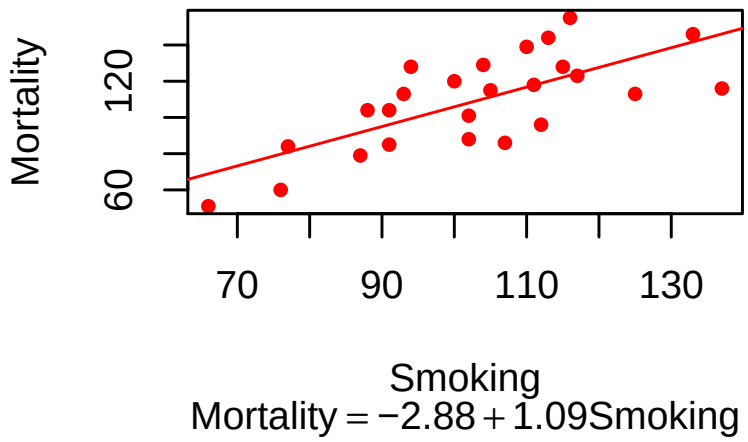


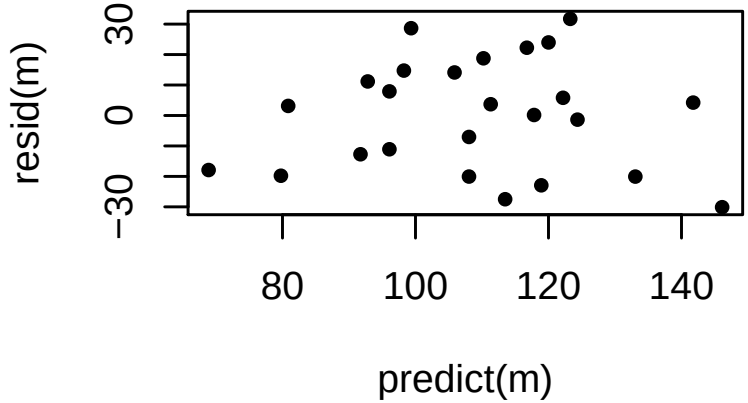


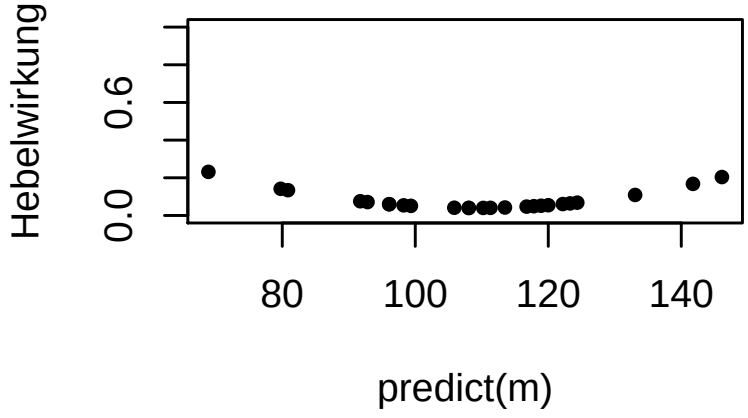


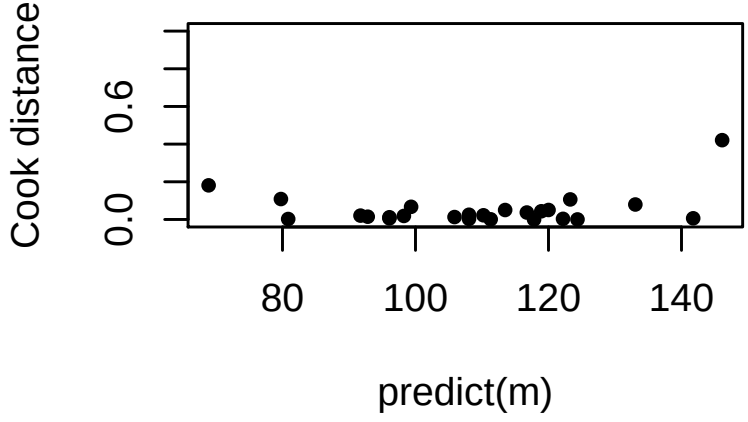


Raucherdaten

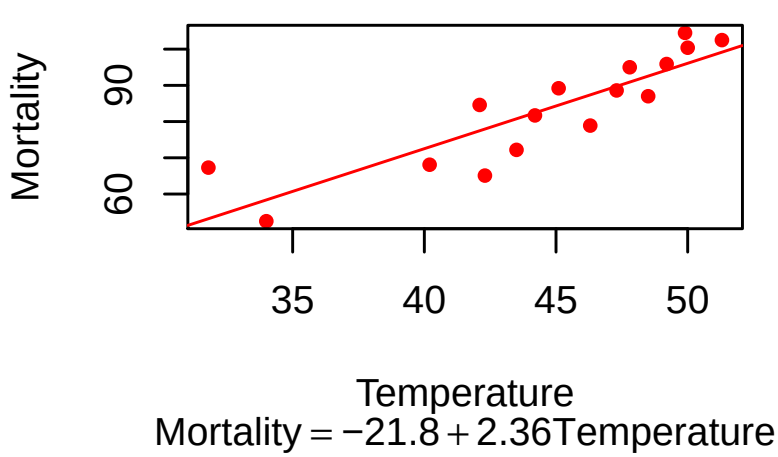


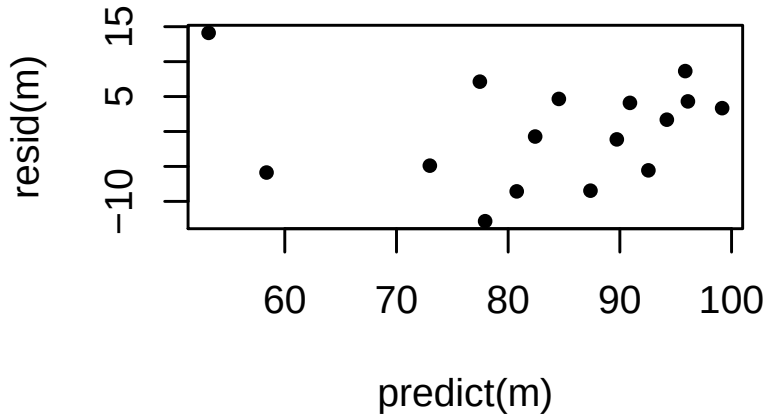


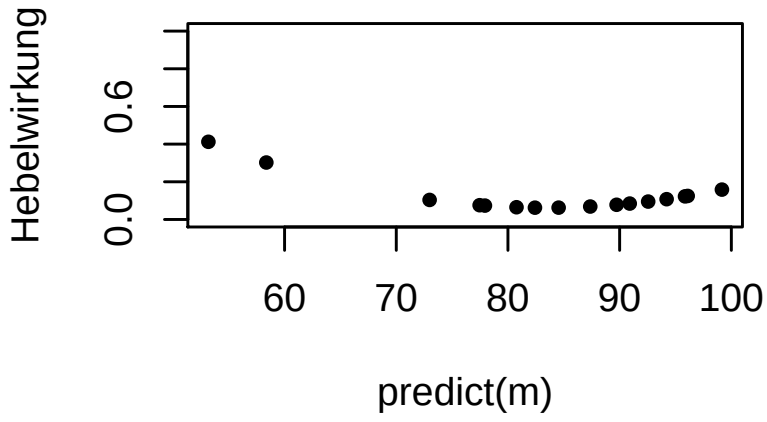


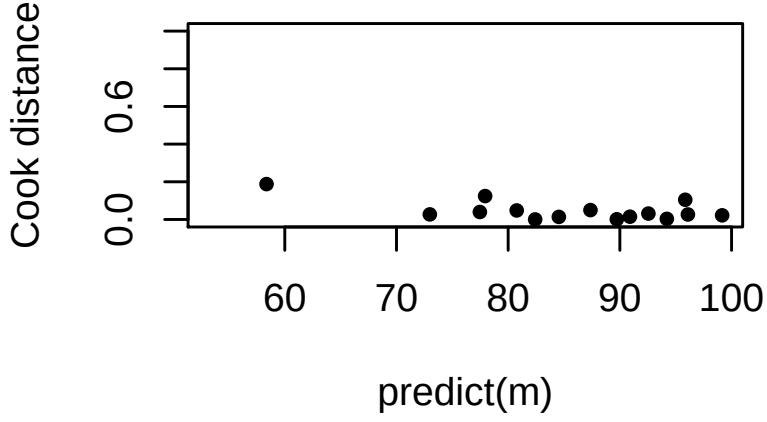


Brustkrebsdaten





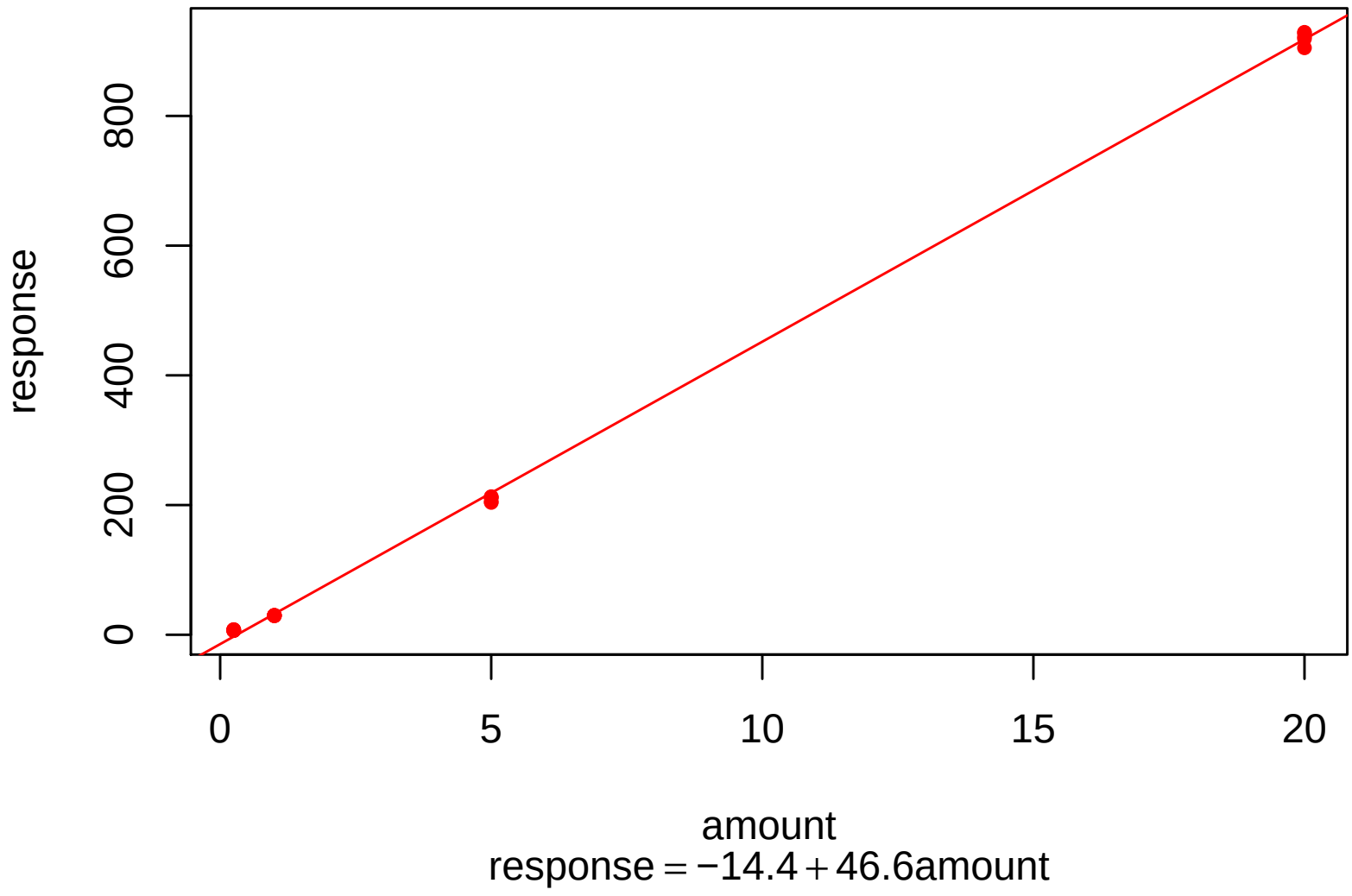




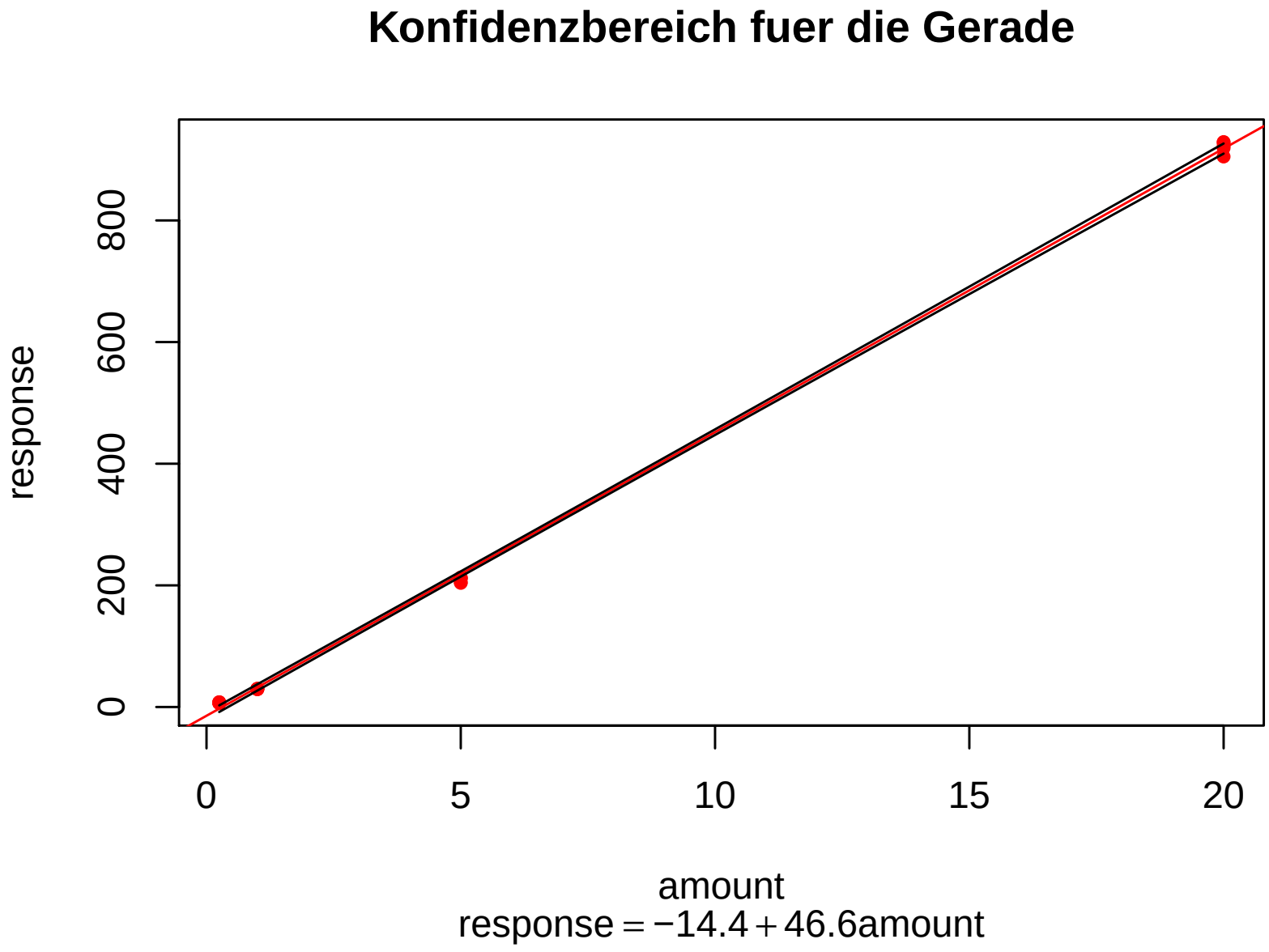
Idee der robusten Schätzung

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

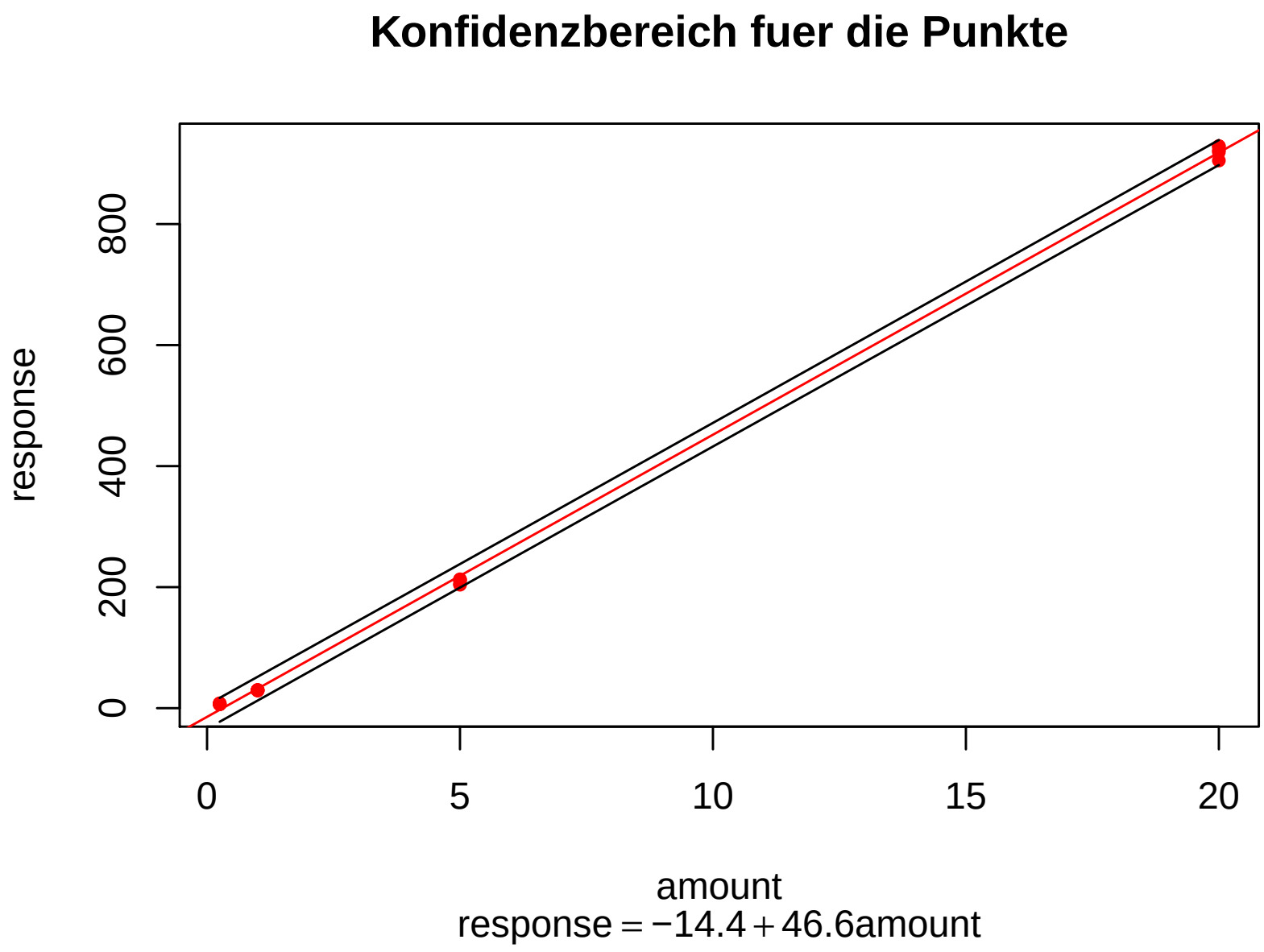
Chromatograph



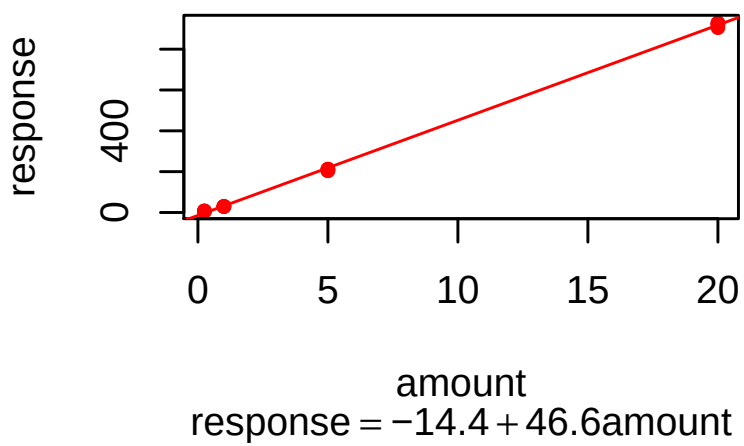
Chromatograph

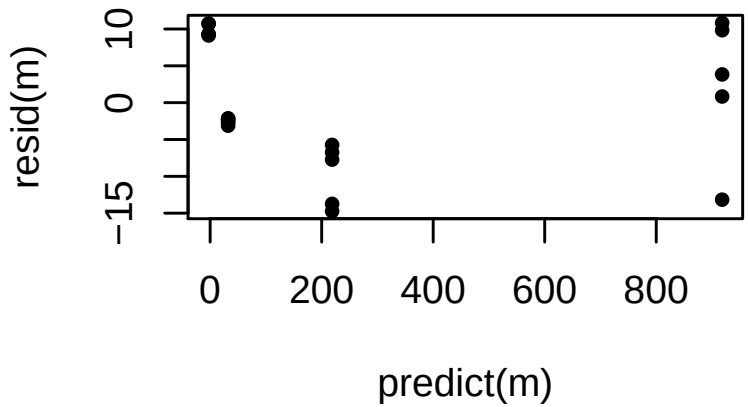


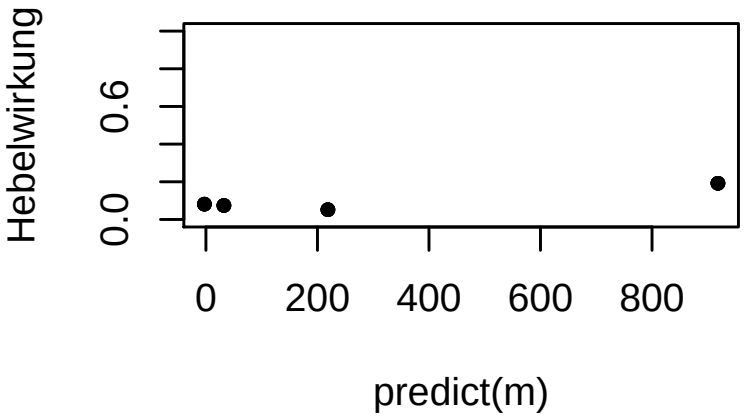
Chromatograph

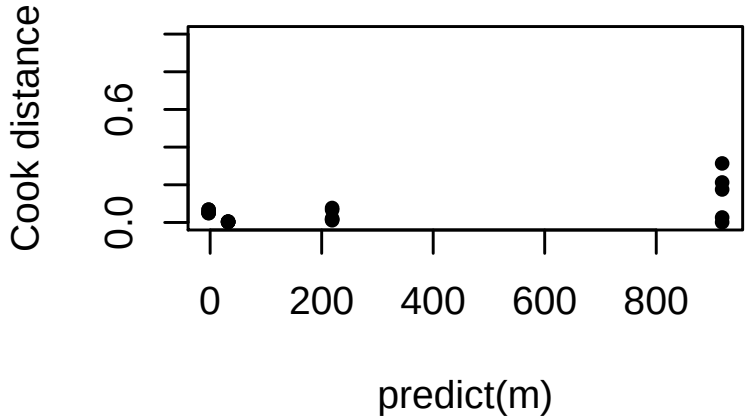


Chromatograph

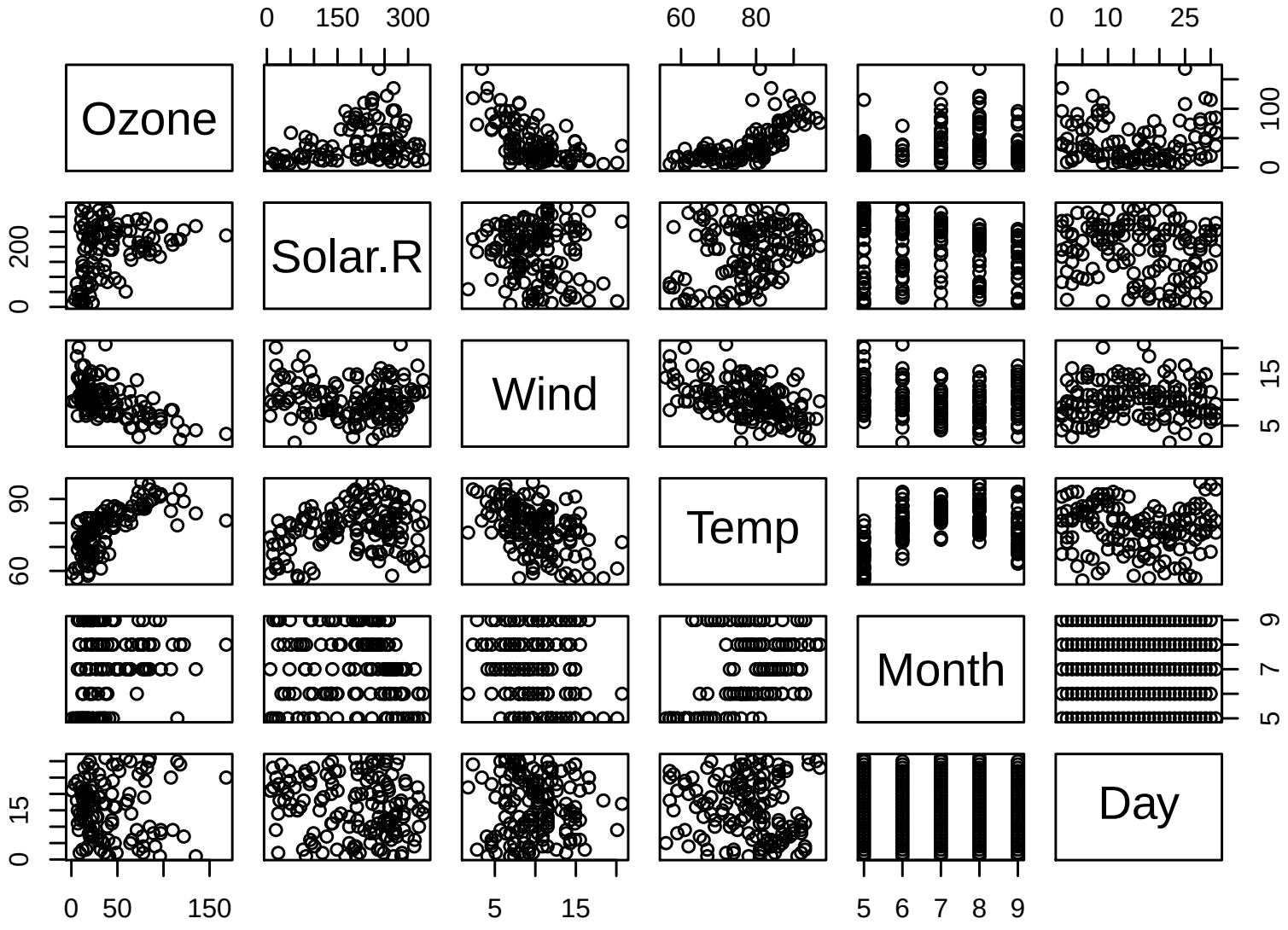




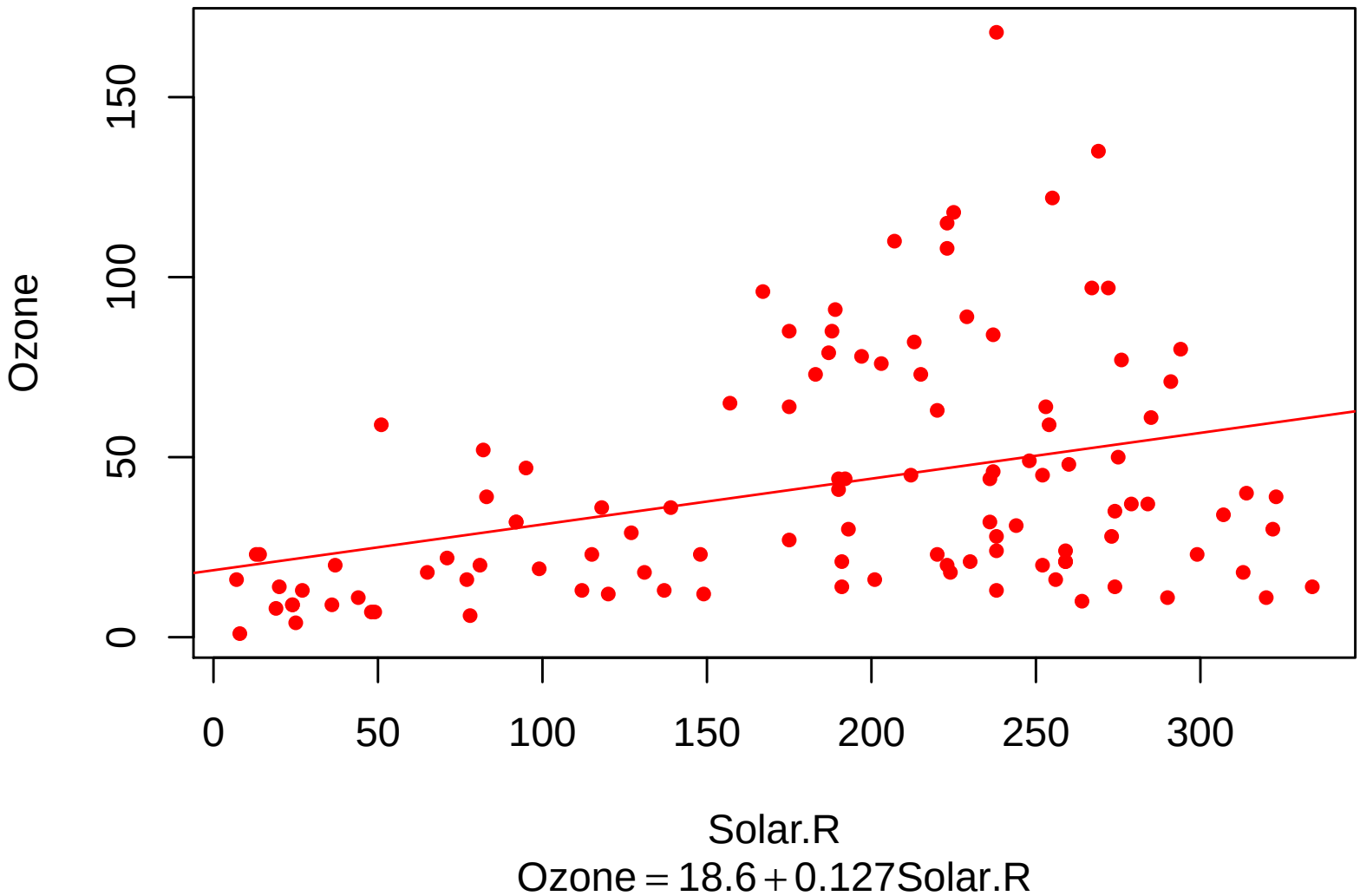




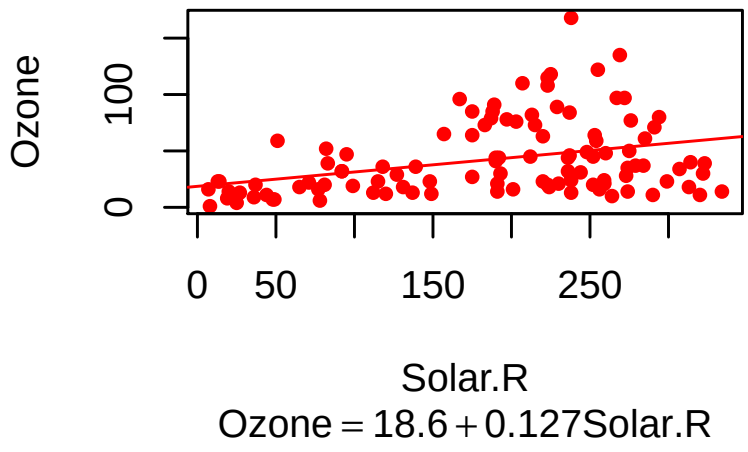
Luftqualität

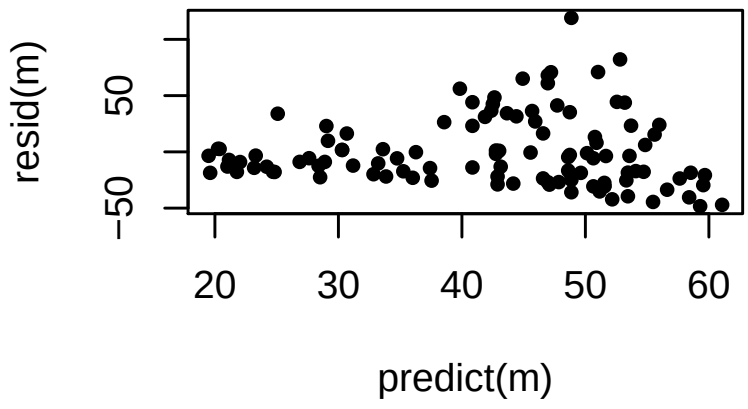


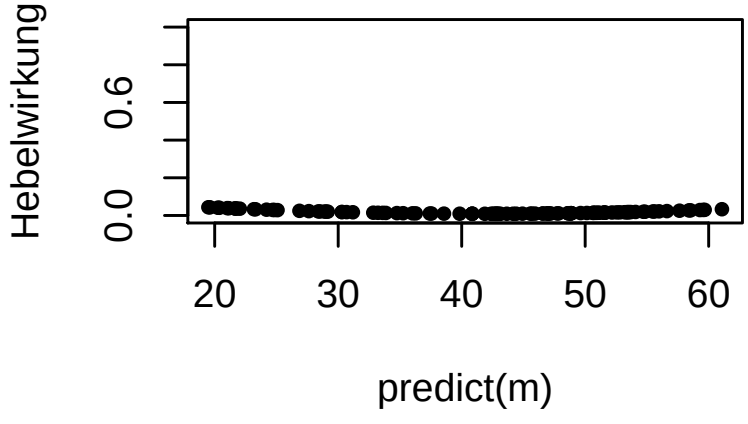
Luftqualität

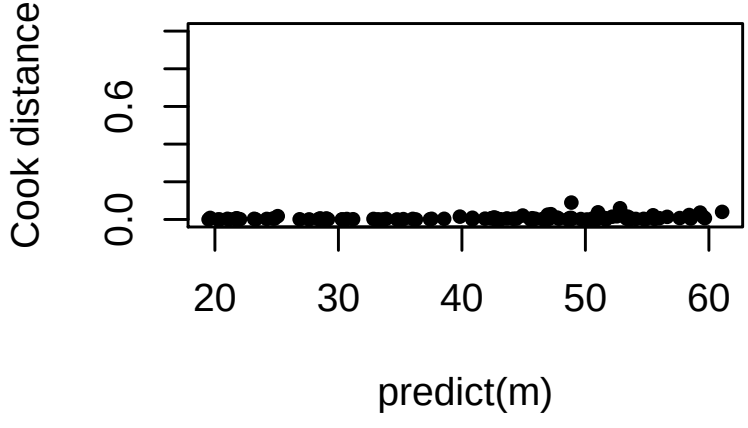


Luftqualität

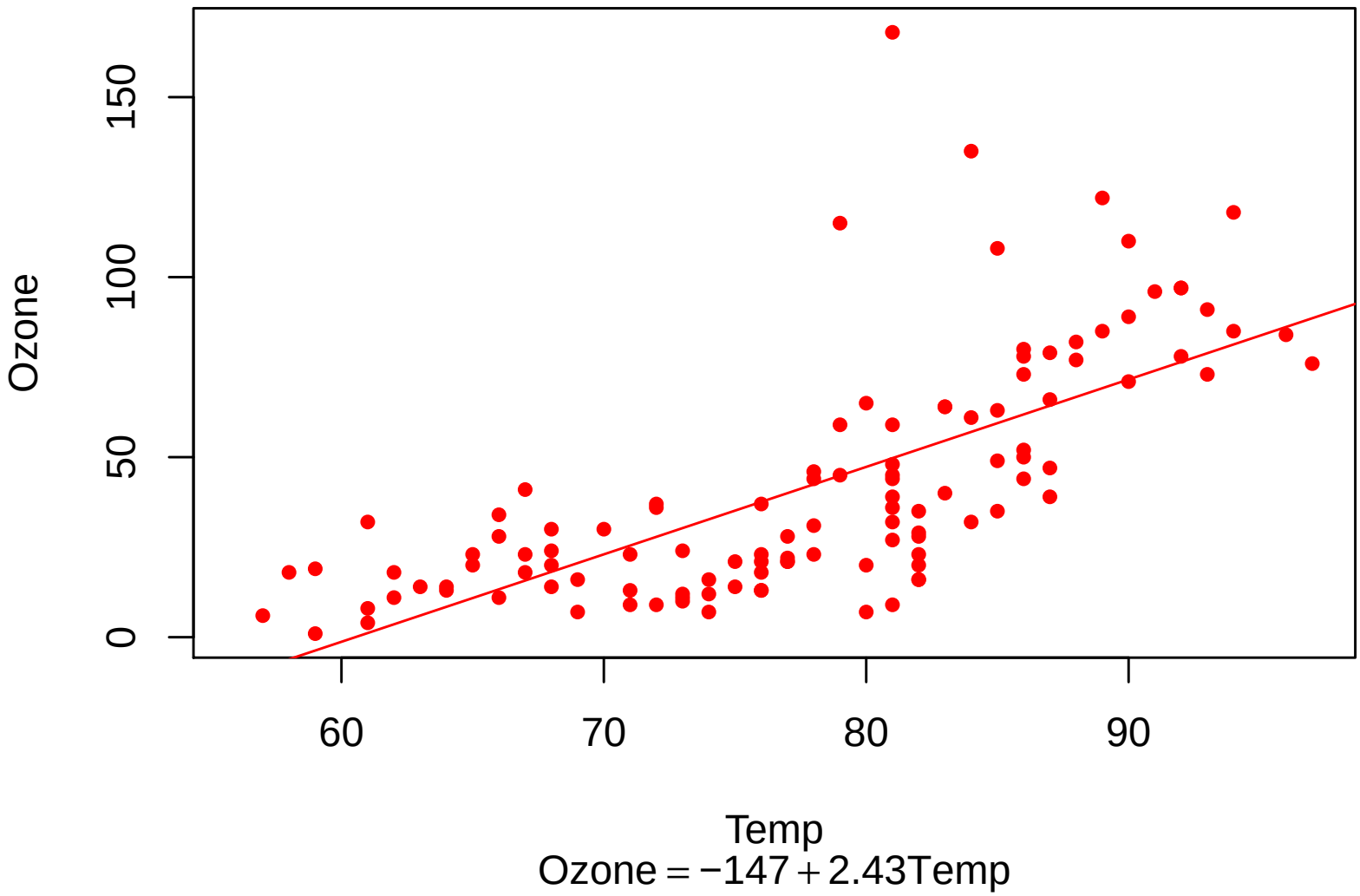








Luftqualität



Luftqualität