Lineare Regression

Pascal Wittmann

Inhaltsverzeichnis

Inhaltsverzeichnis

Inhaltsverzeichnis

Problemstellung

- Es sind Paare von Messwerten (x_i, y_i) mit $i \in \{1, \ldots, n\}$ und $n \geq 2$ gegeben. Diese stellen geometrisch eine Punktwolke im $\mathbb{R}^2$ dar.
- Diese Paare müssen (für die **lineare** Regression) annähernd auf einer Geraden liegen.
- Es soll der Zusammenhang zwischen interessierenden (endogene) und erklährenden (exogene) Variablen analysiert und beschrieben werden.
  - Dafür wird die Gerade gesucht, die am nächsten an den Messwerten liegt. Diese heißt Ausgleichs- oder Regressionsgerade.

Problemstellung

- Es sind Paare von Messwerten (x_i, y_i) mit $i \in \{1, \ldots, n\}$ und $n \geq 2$ gegeben. Diese stellen geometrisch eine Punktwolke im $\mathbb{R}^2$ dar.
- Diese Paare müssen (für die **lineare** Regression) annähernd auf einer Geraden liegen.
- Es soll der Zusammenhang zwischen interessierenden (endogene) und erklährenden (exogene) Variablen analysiert und beschrieben werden.
  - Dafür wird die Gerade gesucht, die am nächsten an den Messwerten liegt. Diese heißt Ausgleichs- oder Regressionsgerade.

Problemstellung

- Es sind Paare von Messwerten (x_i, y_i) mit $i \in \{1, \ldots, n\}$ und $n \geq 2$ gegeben. Diese stellen geometrisch eine Punktwolke im $\mathbb{R}^2$ dar.
- Diese Paare müssen (für die **lineare** Regression) annähernd auf einer Geraden liegen.
- Es soll der Zusammenhang zwischen interessierenden (endogene) und erklährenden (exogene) Variablen analysiert und beschrieben werden.
  - Dafür wird die Gerade gesucht, die am nächsten an den Messwerten liegt. Diese heißt Ausgleichs- oder Regressionsgerade.

Beispiel

Rauchen und Lungenkrebs

- In einer Studie (in England) wurde anhand von Berufsgruppen untersucht, ob diese mehr oder weniger als der Durchschnitt der Bevölkerung raucht und ob die Anzahl der Patienten (aus diesen Gruppen) häufiger an Lungenkrebs erkrankt.
  Quelle: Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales, 1970-1972, Her Majesty's Stationery Office, London, 1978

Occupational Group (i)	Smoking (x_i)	Mortality (y_i)
Farmers, foresters, and fisherman	77	84
Miners and quarrymen	137	116
Gas, coke and chemical makers	117	123
⋮	⋮	⋮
Service, sport, and recreation workers	100	120
Administrators and managers	76	60
Professionals, technical workers, and artists	66	51

Beispiel

Rauchen und Lungenkrebs

- In einer Studie (in England) wurde anhand von Berufsgruppen untersucht, ob diese mehr oder weniger als der Durchschnitt der Bevölkerung raucht und ob die Anzahl der Patienten (aus diesen Gruppen) häufiger an Lungenkrebs erkrankt.
  Quelle: Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales, 1970-1972, Her Majesty's Stationery Office, London, 1978

Occupational Group (i)	Smoking (x_i)	Mortality (y_i)
Farmers, foresters, and fisherman	77	84
Miners and quarrymen	137	116
Gas, coke and chemical makers	117	123
⋮	⋮	⋮
Service, sport, and recreation workers	100	120
Administrators and managers	76	60
Professionals, technical workers, and artists	66	51

Ansatz

Da die Punkte in der Punktwolke annähernd auf einer Geraden liegen, kann man allgemein einen ungefähren Zusammenhang zwischen der exogenen Variablen x und der endogenen Variablen y beschreiben:

$$y \approx \alpha + \beta x \tag{1}$$

$$\textit{Sterblichkeit} \approx \alpha + \beta * \textit{Rauchen} \tag{2}$$

Um die Abweichung der Punkte von der Geraden auszugleichen, wird eine Störgröße ϵ_i mit einbezogen:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{3}$$

Ansatz

Da die Punkte in der Punktwolke annähernd auf einer Geraden liegen, kann man allgemein einen ungefähren Zusammenhang zwischen der exogenen Variablen x und der endogenen Variablen y beschreiben:

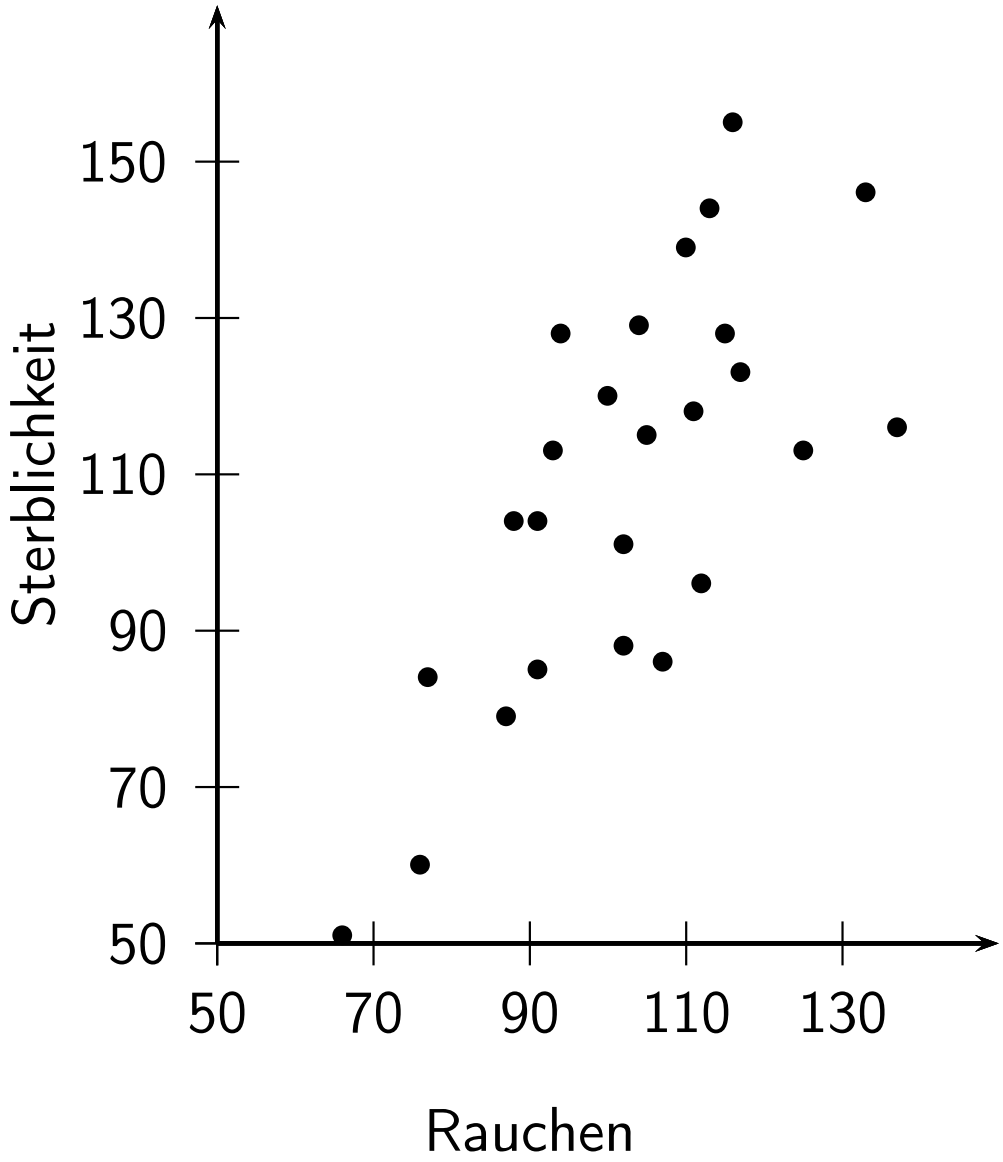
$$y \approx \alpha + \beta x \tag{1}$$

$$\textit{Sterblichkeit} \approx \alpha + \beta * \textit{Rauchen} \tag{2}$$

Um die Abweichung der Punkte von der Geraden auszugleichen, wird eine Störgröße ϵ_i mit einbezogen:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{3}$$

Punktwolke



Residuum

Die Störgröße $\hat{\epsilon}$ wird auch als Residuum bezeichnet und beschreibt den Abstand der Messwerte y_i von dem Schätzwert $\hat{y}_i$ der Regressionsgeraden. Die Regressionsgerade wird mit

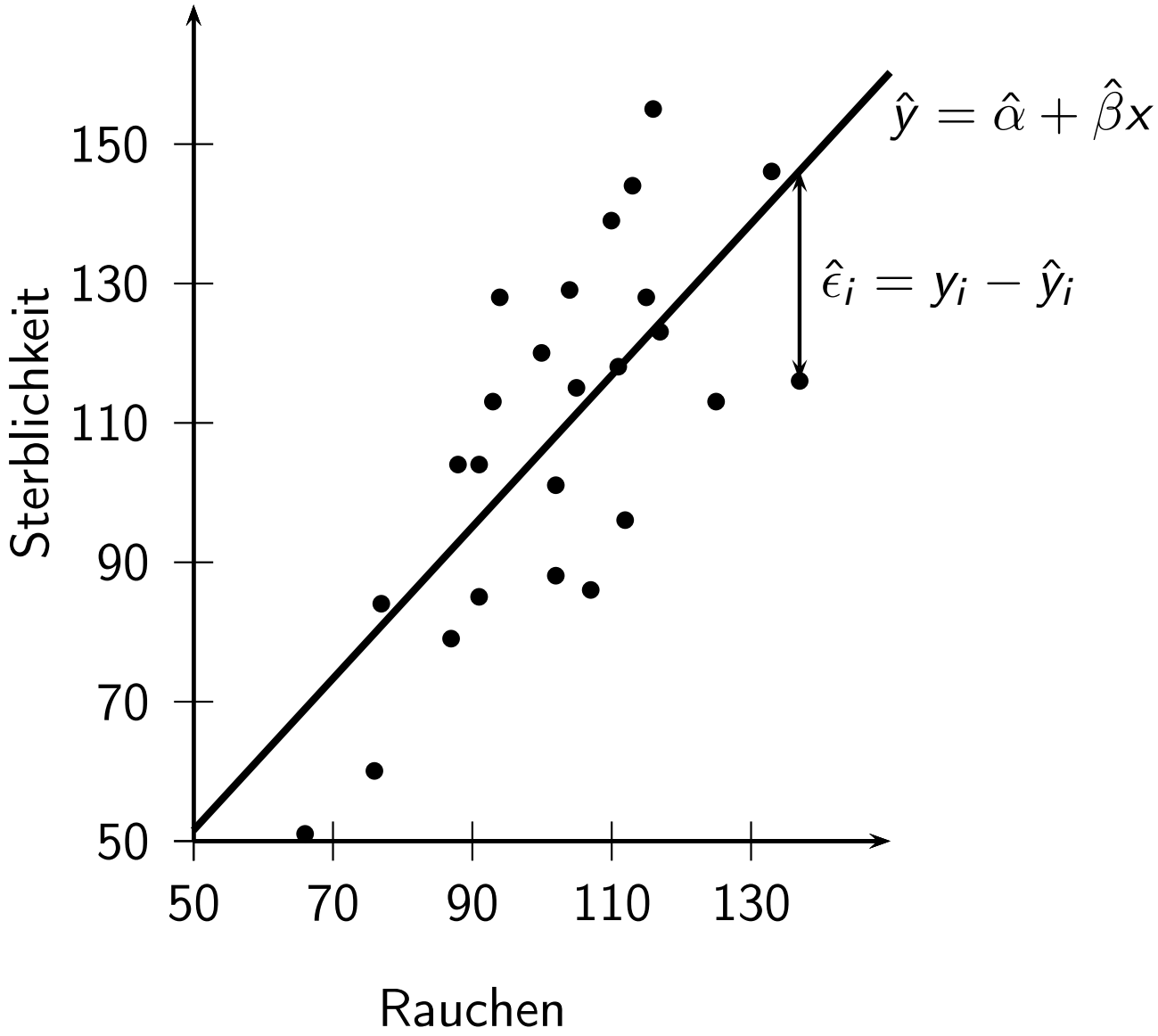
$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{4}$$

angegeben. Nun kann man das Residuum $\hat{\epsilon}_i$ allgemein angeben:

$$\hat{\epsilon}_i = y_i - \hat{y}_i \tag{5}$$

$$= y_i - (\hat{\alpha} + \hat{\beta}x_i) \tag{6}$$

Punktwolke & Residuum & Gerade



kleinste Quadrate

- Nun sollen die beiden unbekannten Parameter α und β geschätzt werden.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{7}$$

- Um eine Möglichst genaue Schätzung der Parameter zu erhalten, muss die Summe der Abstände minimal sein, daher müssen diese als Funktion D beschrieben werden.
- Die Funktion D besitzt die zwei Parameter α und β.

kleinste Quadrate

- Nun sollen die beiden unbekannten Parameter α und β geschätzt werden.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{7}$$

- Um eine Möglichst genaue Schätzung der Parameter zu erhalten, muss die Summe der Abstände minimal sein, daher müssen diese als Funktion D beschrieben werden.
- Die Funktion D besitzt die zwei Parameter α und β.

kleinste Quadrate

- Nun sollen die beiden unbekannten Parameter α und β geschätzt werden.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{7}$$

- Um eine Möglichst genaue Schätzung der Parameter zu erhalten, muss die Summe der Abstände minimal sein, daher müssen diese als Funktion D beschrieben werden.
- Die Funktion D besitzt die zwei Parameter α und β.

kleinste Quadrate

Um einen Funktionsterm zu finden, betrachten wir den Abstand der Vektoren

- $\vec{v} \in \mathbb{R}^n$ mit den Einträgen $v_i = \hat{y}_i$
- $\vec{y} \in \mathbb{R}^n$ mit den Einträgen y_i

$$D(\alpha, \beta) = \|\vec{v} - \vec{y}\| = \sqrt{\sum_{i=1}^{n} (v_i - y_i)^2} \tag{8}$$

Um die Extremstellen der Funktion D zu bestimmen, muss gelten:

$$\nabla D = 0 \tag{9}$$

kleinste Quadrate

Um einen Funktionsterm zu finden, betrachten wir den Abstand der Vektoren

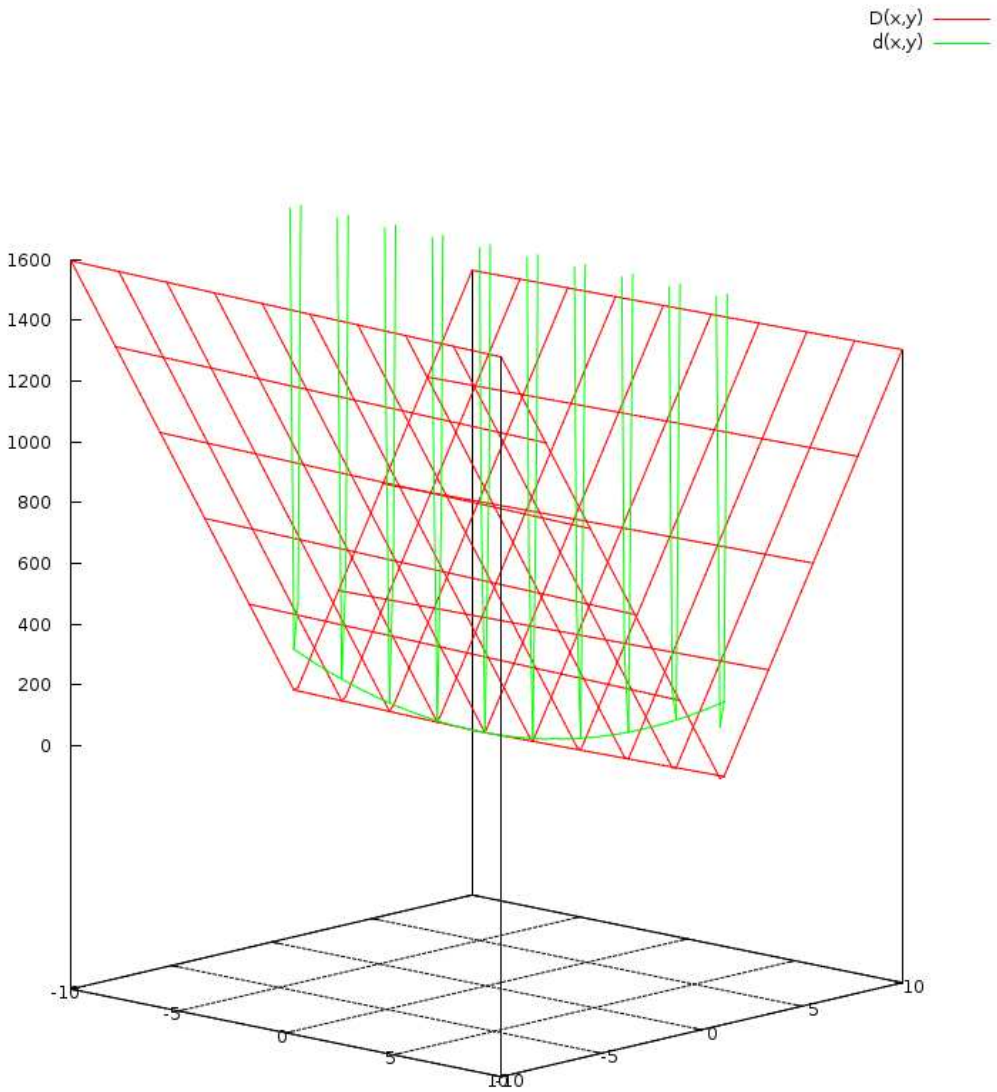
- $\vec{v} \in \mathbb{R}^n$ mit den Einträgen $v_i = \hat{y}_i$
- $\vec{y} \in \mathbb{R}^n$ mit den Einträgen y_i

$$D(\alpha, \beta) = \|\vec{v} - \vec{y}\| = \sqrt{\sum_{i=1}^{n} (v_i - y_i)^2} \tag{8}$$

Um die Extremstellen der Funktion D zu bestimmen, muss gelten:

$$\nabla D = 0 \tag{9}$$

Die Funktion D kann weiter vereinfacht werden, da das Minimum von $D(\alpha, \beta)$ und $d(\alpha, \beta) = D(\alpha, \beta)^2$ gleich ist.



Nun muss die Funktion d minimiert werden:

$$d(\alpha, \beta) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (\alpha + \beta x - y_i)^2 \tag{10}$$

$$\frac{\partial d}{\partial \alpha} = \sum_{i=1}^{n} 2(\hat{\alpha} + \hat{\beta} x_i - y_i) \tag{11}$$

$$\frac{\partial d}{\partial \beta} = \sum_{i=1}^{n} 2(\hat{\alpha} + \hat{\beta} x_i - y_i) x_i \tag{12}$$

Nun muss die Funktion d minimiert werden:

$$d(\alpha, \beta) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (\alpha + \beta x - y_i)^2 \tag{10}$$

$$\frac{\partial d}{\partial \alpha} = \sum_{i=1}^{n} 2(\hat{\alpha} + \hat{\beta} x_i - y_i) \tag{11}$$

$$\frac{\partial d}{\partial \beta} = \sum_{i=1}^{n} 2(\hat{\alpha} + \hat{\beta} x_i - y_i) x_i \tag{12}$$

Um das Minimum zu bestimmen muss gelten $\nabla h = 0$. Aus dieser Bedingung kann ein LGS aufgestellt werden. Die Lösungen des LGS sind:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{13}$$

$$\hat{\alpha} = \bar{y} - m\bar{x} \tag{14}$$

mit

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \tag{15}$$

$$\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i \tag{16}$$

Güte der Regressionsgeraden

- Nachdem nun eine Regressionsgerade gefunden wurde, soll überprüft werden, wie genau diese Gerade die tatsächlichen Werte schätzt.
- Es muss nun zwischen der Streuung die von der Regressionsgeraden beschrieben wird und anderen Einflüssen getrennt werden.
- Die gesamte Streuung (sum of squares total)lässt so beschreiben:

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{17}$$

Güte der Regressionsgeraden

- Nachdem nun eine Regressionsgerade gefunden wurde, soll überprüft werden, wie genau diese Gerade die tatsächlichen Werte schätzt.
- Es muss nun zwischen der Streuung die von der Regressionsgeraden beschrieben wird und anderen Einflüssen getrennt werden.
- Die gesamte Streuung (sum of squares total)lässt so beschreiben:

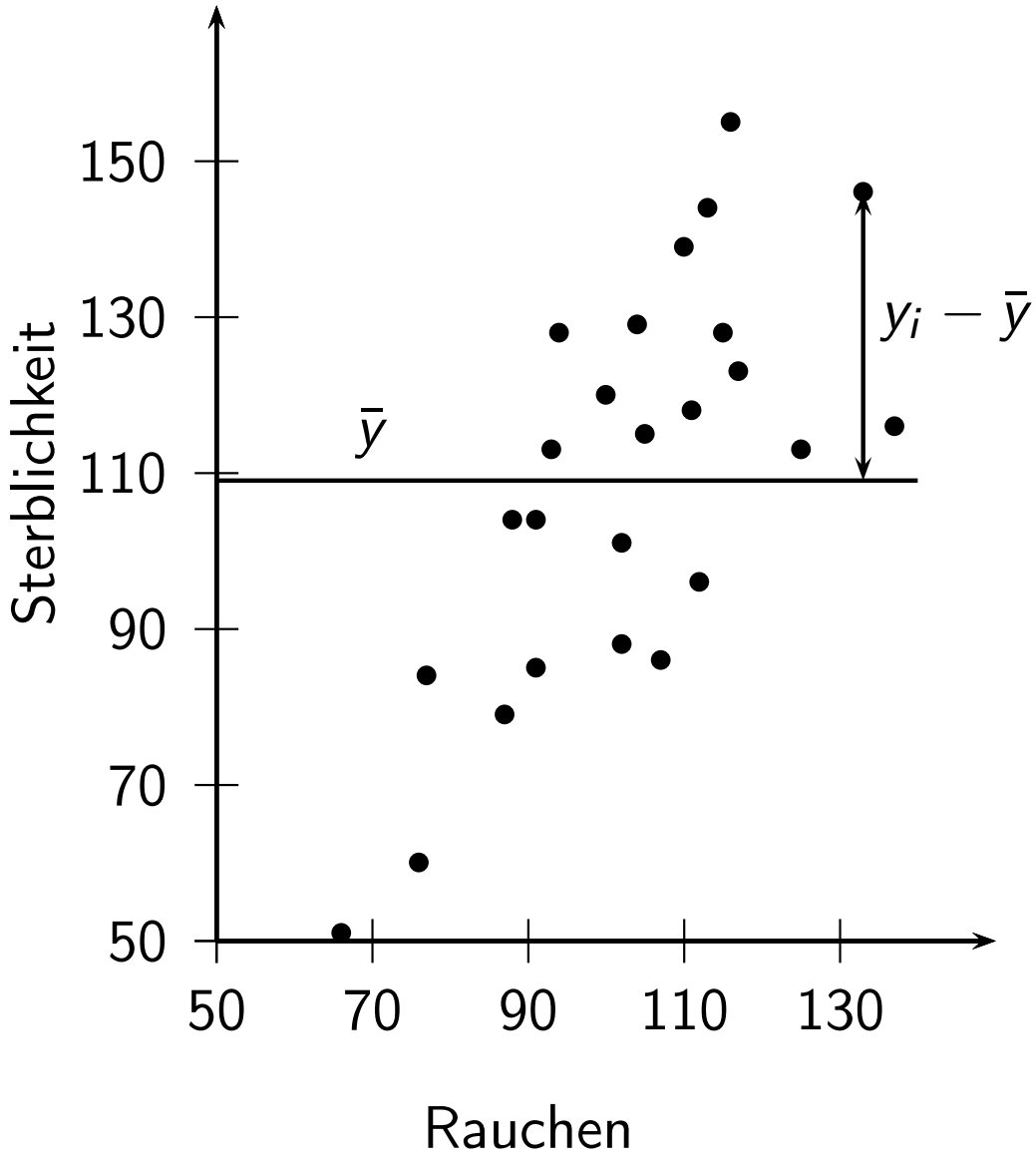
$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{17}$$

Güte der Regressionsgeraden

- Nachdem nun eine Regressionsgerade gefunden wurde, soll überprüft werden, wie genau diese Gerade die tatsächlichen Werte schätzt.
- Es muss nun zwischen der Streuung die von der Regressionsgeraden beschrieben wird und anderen Einflüssen getrennt werden.
- Die gesamte Streuung (**s**um of **s**quares **t**otal)lässt so beschreiben:

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{17}$$

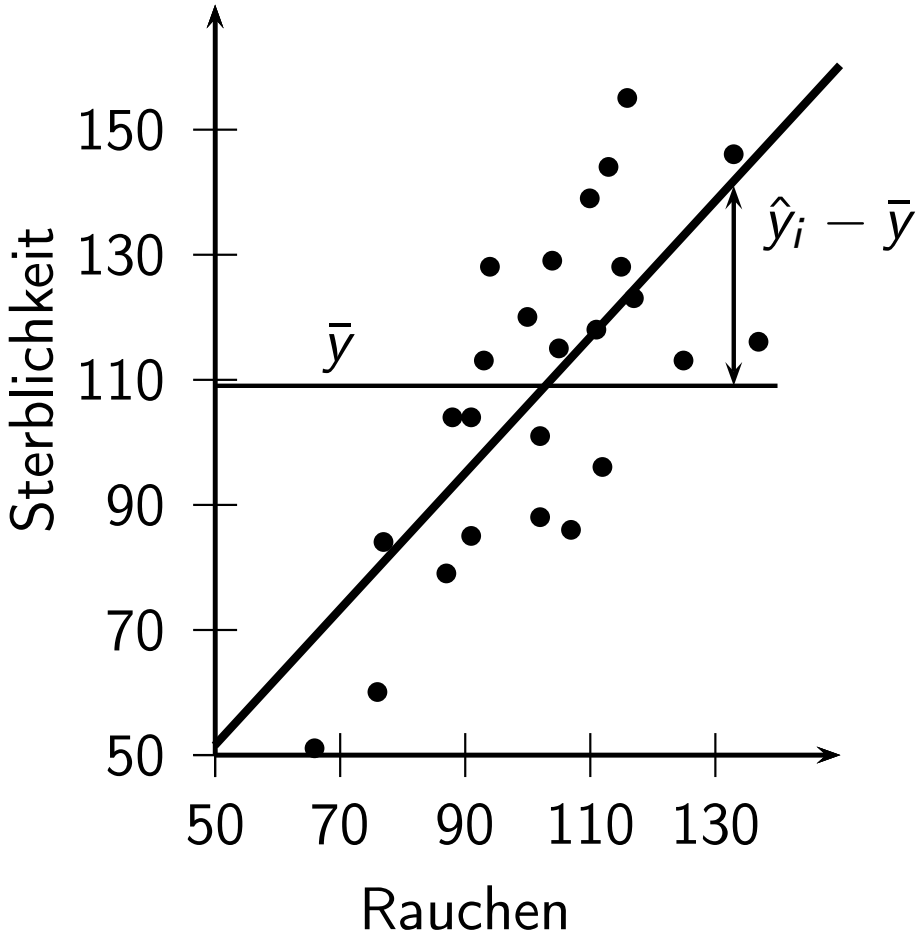
Darstellung *SQT*



SQE

Die erklärte Streuung (sum of squares explained):

$$SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{18}$$



Bestimmtheitsmass

Die Residualstreuung (**s**um of **s**quares **r**esiduals):

$$SQR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{19}$$

Das Maß der Genauigkeit der Regressionsgeraden, wird als Bestimmtheitsmaß oder auch Determinationskoeffizient bezeichnet. Es beschreibt den Anteil der erklärten Streuung an der Gesamtstreuung:

$$R^2 = \frac{SQE}{SQT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{20}$$

Bestimmtheitsmass

Die Residualstreuung (**s**um of **s**quares **r**esiduals):

$$SQR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{19}$$

Das Maß der Genauigkeit der Regressionsgeraden, wird als Bestimmtheitsmaß oder auch Determinationskoeffizient bezeichnet. Es beschreibt den Anteil der erklärten Streuung an der Gesamtstreuung:

$$R^2 = \frac{SQE}{SQT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{20}$$

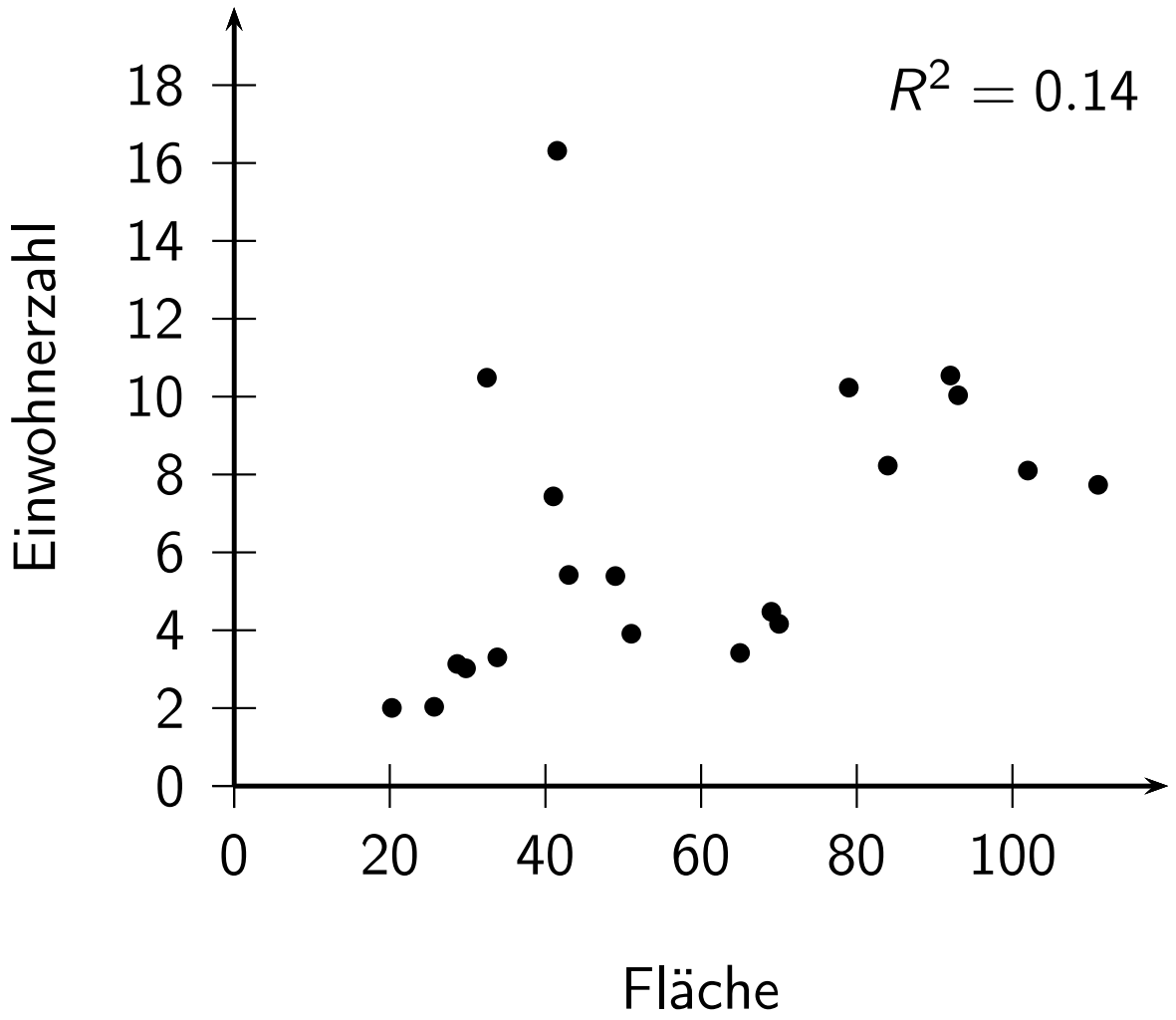
R^2 des Beispiels

- Für das Beispiel ist $R^2 = 0,51$. Der Zusammenhang zwischen Rauchen und Lungenkrebs wurde zu ca. 50% erklärt.
- Allein auf das Bestimmungsmaß sollte man sich jedoch nicht verlassen, wie dieses Beispiel zeigt:
  - Es wurde versucht einen Zusammenhang zwischen der Fläche eines Landes und dessen Einwohnerzahl zu beschreiben.

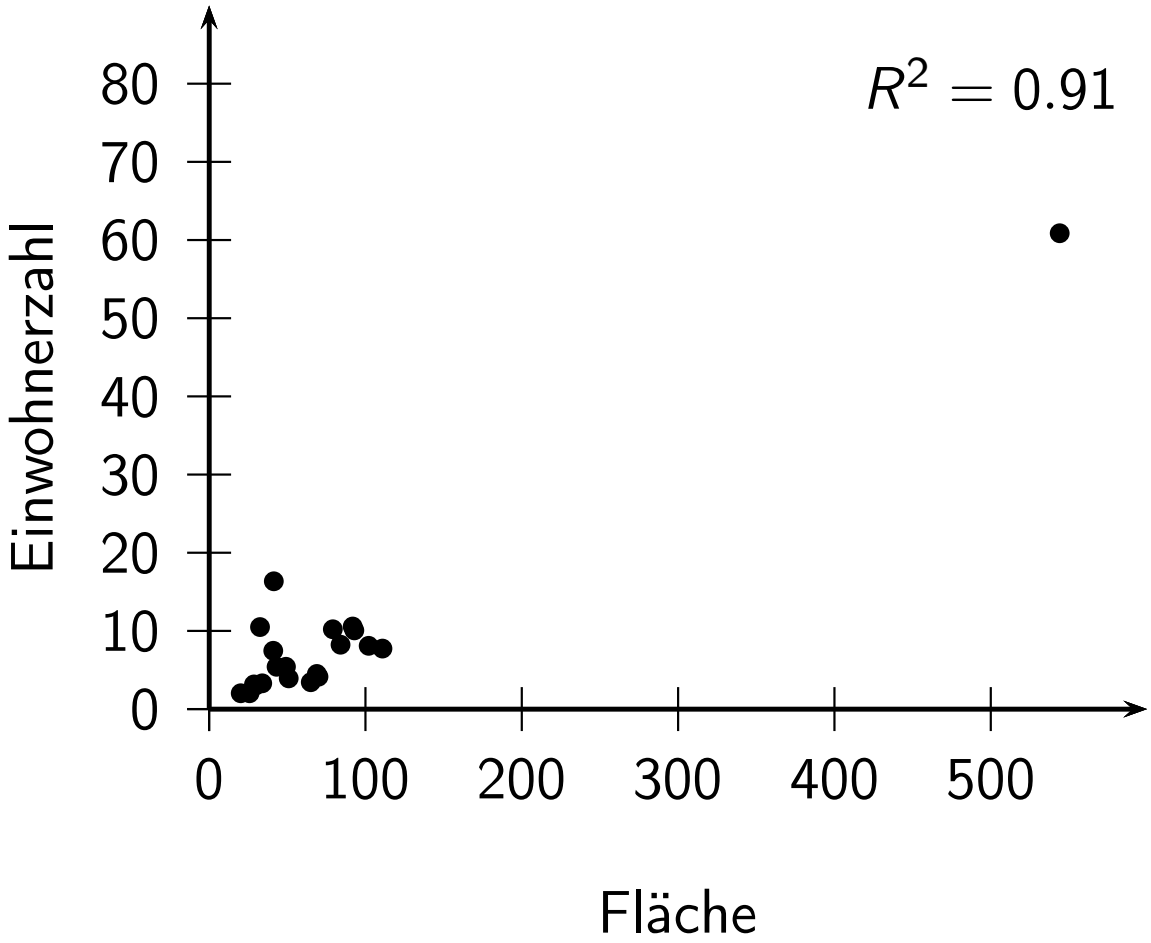
R^2 des Beispiels

- Für das Beispiel ist $R^2 = 0,51$. Der Zusammenhang zwischen Rauchen und Lungenkrebs wurde zu ca. 50% erklärt.
- Allein auf das Bestimmungsmaß sollte man sich jedoch nicht verlassen, wie dieses Beispiel zeigt:
  - Es wurde versucht einen Zusammenhang zwischen der Fläche eines Landes und dessen Einwohnerzahl zu beschreiben.

Punktwolke (ohne Deutschland und Frankreich)



Punktwolke


$R^2 = 0.91$
Einwohnerzahl
0
10
20
30
40
50
60
70
80
0
100
200
300
400
500
Fläche

Quellen

- **Keppeler**, Stefan: Lineare Regression (www.maphy.uni-tuebingen.de/lehre/ws-2007-08/m1bgg/scripts/14_lineare_Regression.pdf)
- **Fahrmeir**, Ludwig et.al.: Statistik, Springer Verlag
- **Stalder**, Peter: Einfache lineare Regression (www.forschung.snb.ch/files/stalder/update-nov-05/oekonometrieI/einfach-regression/SimpleReg_T1.pdf)
- **Stahel**, Werner: Linerare Regression (http://stat.ethz.ch/~stahel/courses/regression/reg-intro.pdf)

Verwendete Software: LaTeX, PSTricks und Gnuplot

Vielen Dank für ihre Aufmerksamkeit