3.3 Statistische Maßzahlen

Zwei Arten von statistischen Maßzahlen:

3.3 Statistische Maßzahlen

Zwei Arten von statistischen Maßzahlen:

Lagemaßzahlen:

In welchem Bereich der Zahlengeraden liegen die Werte (oder die "Mitte" der Werte) ?

3.3 Statistische Maßzahlen

Zwei Arten von statistischen Maßzahlen:

Lagemaßzahlen:

In welchem Bereich der Zahlengeraden liegen die Werte (oder die "Mitte" der Werte) ?

Streuungsmaßzahlen:

Wie groß ist der "Bereich", über den sich die Werte im wesentlichen erstrecken ?

Im Folgenden sei

$$x_1, \ldots, x_n$$

die Messreihe. Die der Größe nach aufsteigend sortierten Werte seien

$$x_{(1)}, \ldots, x_{(n)}.$$

Im Folgenden sei

$$x_1, \ldots, x_n$$

die Messreihe. Die der Größe nach aufsteigend sortierten Werte seien

$$x_{(1)}, \ldots, x_{(n)}.$$

In Beispiel 1 oben: Beschäftigungsquoten der Männer im Jahr 2006:

$x_1, \ldots, x_{26}$:
67, 63.3, 73.6, 80.6, 72.5, 71.3, 77.3, 74.6, 76, 68.5, 71.1, 79.6, 68.2, 66.4, 63.9,
73.8, 80.8, 77, 60.2, 74, 65.2, 70.8, 66.9, 71.7, 75.5, 77

Im Folgenden sei

$$x_1, \ldots, x_n$$

die Messreihe. Die der Größe nach aufsteigend sortierten Werte seien

$$x_{(1)}, \ldots, x_{(n)}.$$

In Beispiel 1 oben: Beschäftigungsquoten der Männer im Jahr 2006:

$x_1, \ldots, x_{26}$:
67, 63.3, 73.6, 80.6, 72.5, 71.3, 77.3, 74.6, 76, 68.5, 71.1, 79.6, 68.2, 66.4, 63.9,
73.8, 80.8, 77, 60.2, 74, 65.2, 70.8, 66.9, 71.7, 75.5, 77

$x_{(1)}, \ldots, x_{(26)}$:
60.2, 63.3, 63.9, 65.2, 66.4, 66.9, 67.0, 68.2, 68.5, 70.8, 71.1, 71.3, 71.7, 72.5,
73.6, 73.8, 74.0, 74.6, 75.5, 76.0, 77.0, 77.0, 77.3, 79.6, 80.6, 80.8

Beispiele für Lageparameter:

(empirisches arithmetisches) Mittel:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot (x_1 + x_2 + \cdots + x_n)$$

Beispiele für Lageparameter:

(empirisches arithmetisches) Mittel:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot (x_1 + x_2 + \cdots + x_n)$$

Bei den Beschäftigungsquoten für Männer: $\bar{x} = 71.8$

(Wert bei den Frauen: $\bar{x} = 58.2$)

Beispiele für Lageparameter:

(empirisches arithmetisches) Mittel:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot (x_1 + x_2 + \cdots + x_n)$$

Bei den Beschäftigungsquoten für Männer: $\bar{x} = 71.8$

(Wert bei den Frauen: $\bar{x} = 58.2$)

Problematisch bei nicht reellen Messgrößen oder falls Ausreißer in Stichprobe vorhanden.

In diesen Fällen besser geeignet:

(empirischer) Median:

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls} \quad n \text{ ungerade,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{falls} \quad n \text{ gerade.} \end{cases}$$

In diesen Fällen besser geeignet:

(empirischer) Median:

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls} \quad n \text{ ungerade,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{falls} \quad n \text{ gerade.} \end{cases}$$

Bei den Beschäftigungsquoten für Männer: $\tilde{x} = 72.10$

(Wert bei den Frauen: $\tilde{x} = 59.3$)

Beispiele für Streuungsparameter:

(empirische) Spannweite oder Variationsbreite:

$$r := x_{max} - x_{min} := x_{(n)} - x_{(1)}.$$

Beispiele für Streuungsparameter:

(empirische) Spannweite oder Variationsbreite:

$$r := x_{max} - x_{min} := x_{(n)} - x_{(1)}.$$

Bei den Beschäftigungsquoten für Männer: $r = 80.8 - 60.2 = 20.6$

(Wert bei den Frauen: $r = 73.2 - 34.6 = 29.6$)

(empirische) Varianz:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \cdot \Big((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \Big) .$$

(empirische) Varianz:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \cdot \left((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 \right).$$

Bei den Beschäftigungsquoten für Männer: $s^2 \approx 30.8$

(Wert bei den Frauen: $s^2 \approx 75.3$)

(empirische) Standardabweichung oder Streuung:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

(empirische) Standardabweichung oder Streuung:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Bei den Beschäftigungsquoten für Männer: $s \approx 5.55$

(Wert bei den Frauen: $s \approx 8.68$)

(empirische) Standardabweichung oder Streuung:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Bei den Beschäftigungsquoten für Männer: $s \approx 5.55$

(Wert bei den Frauen: $s \approx 8.68$)

Variationskoeffizient:

$$V = \frac{s}{\bar{x}}$$

(empirische) Standardabweichung oder Streuung:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Bei den Beschäftigungsquoten für Männer: $s \approx 5.55$

(Wert bei den Frauen: $s \approx 8.68$)

Variationskoeffizient:

$$V = \frac{s}{\bar{x}}$$

Bei den Beschäftigungsquoten für Männer: $V \approx 0.077$

(Wert bei den Frauen: $V \approx 0.149$)

Bei nicht reellen Messgrößen oder Vorhandensein von Ausreißern ist der sogenannte Interquartilabstand

$$IQR = x_{(\lceil \frac{3}{4}n \rceil)} - x_{(\lceil \frac{1}{4}n \rceil)}$$

günstiger.

Bei nicht reellen Messgrößen oder Vorhandensein von Ausreißern ist der sogenannte Interquartilabstand

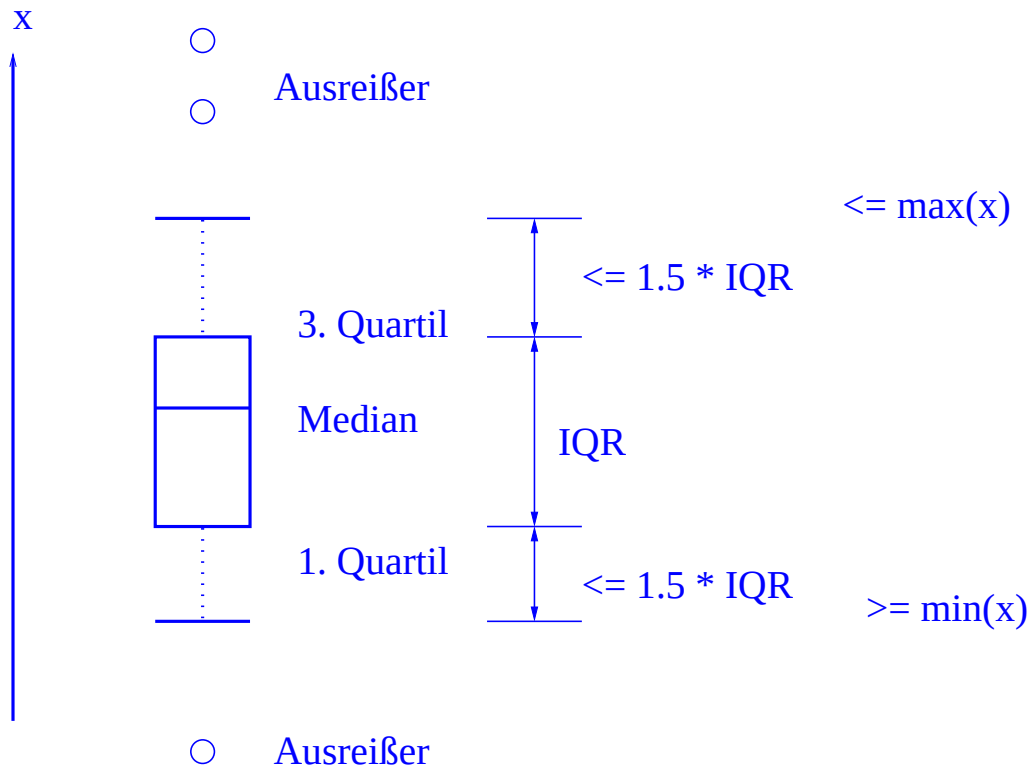
$$IQR = x_{(\lceil \frac{3}{4}n \rceil)} - x_{(\lceil \frac{1}{4}n \rceil)}$$

günstiger.

Bei den Beschäftigungsquoten für Männer:$IQR = 76 - 67 = 9$

(Wert bei den Frauen: $IQR = 63.3 - 53.2 = 10.1$)

Graphische Darstellung einiger dieser Lage- und Streuungsparameter im sogenannten Boxplot:



Boxplot zum Vergleich der Beschäftigungsquoten von Männern und Frauen:

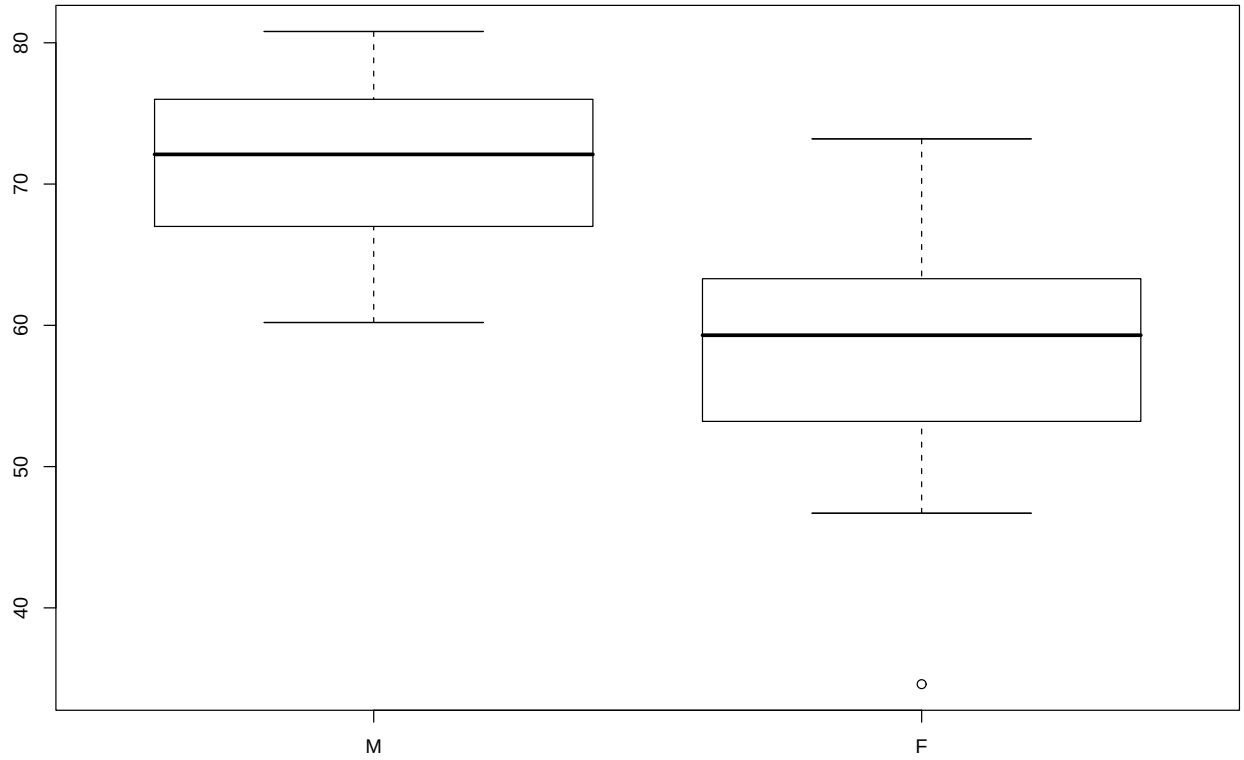
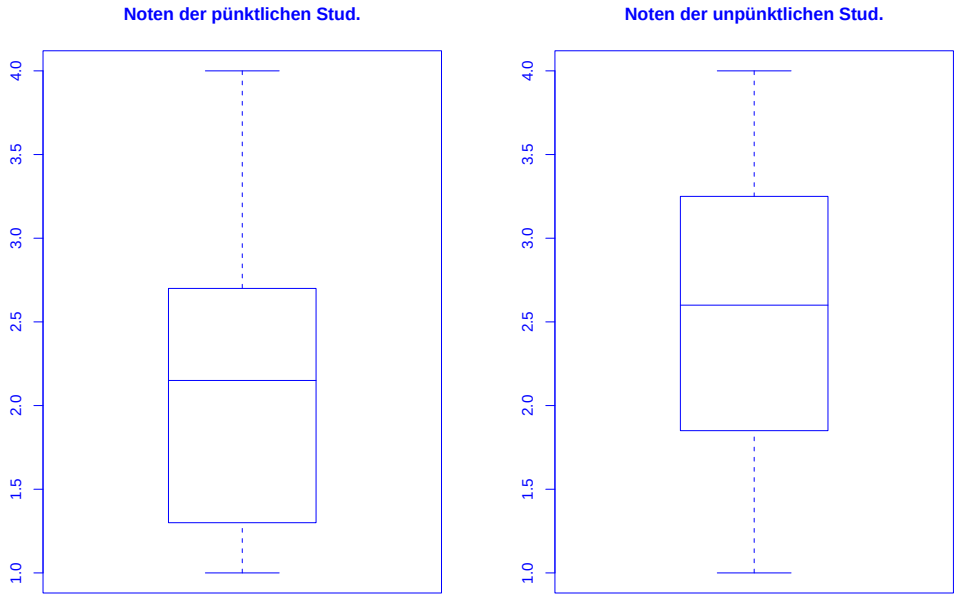
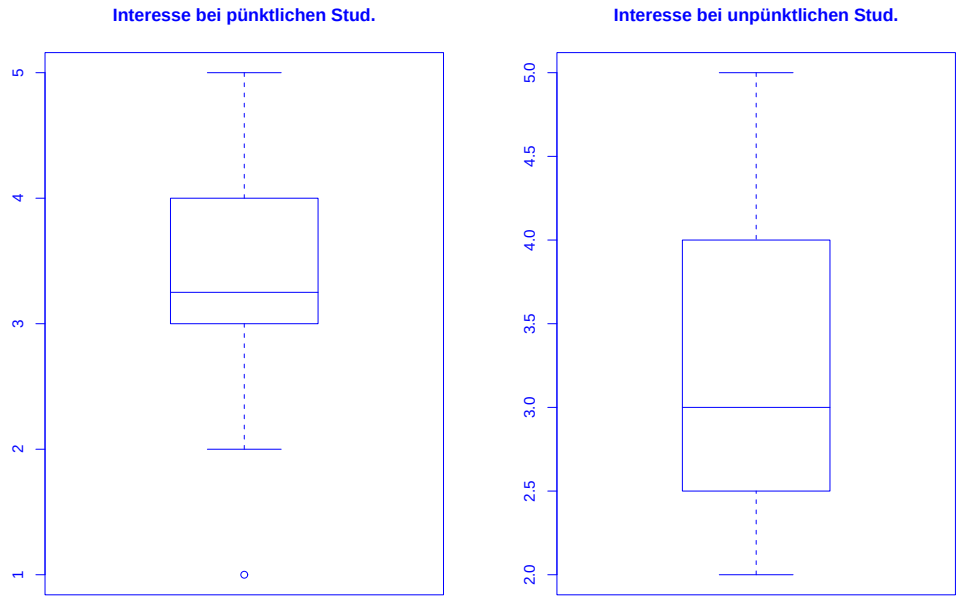


Illustration der Verzerrung durch Auswahl durch Boxplots im Zusammenhang mit Umfrage in Statistik-Vorlesung am 26.10.01:

Illustration der Verzerrung durch Auswahl durch Boxplots im Zusammenhang mit Umfrage in Statistik-Vorlesung am 26.10.01:




Interesse bei pünktlichen Stud.
1
2
3
4
5
Interesse bei unpünktlichen Stud.
2.0
2.5
3.0
3.5
4.0
4.5
5.0

3.4 Regressionsrechnung

Geg.: 2–dimensionale Messreihe

$$(x_1, y_1), \ldots, (x_n, y_n)$$

vom Umfang n.

Frage: Zusammenhang zwischen den x– und den y–Koordinaten ?

3.4 Regressionsrechnung

Geg.: 2–dimensionale Messreihe

$$(x_1, y_1), \ldots, (x_n, y_n)$$

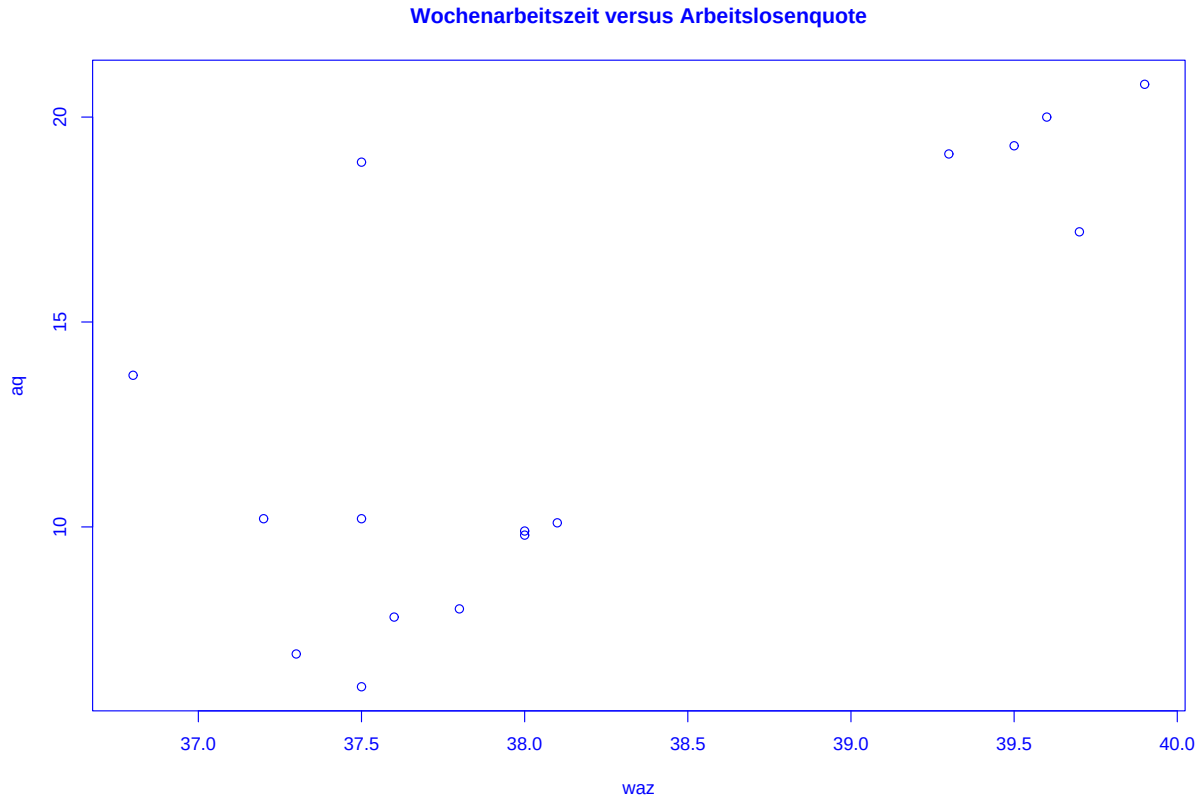
vom Umfang n.

Frage: Zusammenhang zwischen den x– und den y–Koordinaten ?

Beispiel: Besteht ein Zusammenhang zwischen

- der Wochenarbeitszeit im produzierenden Gewerbe und der Arbeitslosenquote in den 16 Bundesländern der BRD im Jahr 2002 ?

Darstellung der Messreihe (Quelle: Statistisches Bundesamt) im Scatterplot (Streudiagramm):



Bei der sogenannten **linearen Regression** passt man eine Gerade

$$y = \mathbf{a} \cdot x + \mathbf{b}$$

an die Daten an.

Bei der sogenannten **linearen Regression** passt man eine Gerade

$$y = \mathbf{a} \cdot x + \mathbf{b}$$

an die Daten an.

Eine Möglichkeit dafür:

Wähle **a, b** $\in \mathbb{R}$ durch Minimierung von

$$\sum_{i=1}^{n} \left(y_i - (a \cdot x_i + b)\right)^2 .$$

Beispiel: Es sei $n = 3$ und

$$(x_1, y_1) = (0, 0), (x_2, y_2) = (1, 0), (x_3, y_3) = (-2, 1).$$

Beispiel: Es sei $n = 3$ und

$$(x_1, y_1) = (0, 0), (x_2, y_2) = (1, 0), (x_3, y_3) = (-2, 1).$$

Gesucht ist dann eine Gerade

$$y = a \cdot x + b,$$

für die der folgende Ausdruck möglichst klein ist:

$$(y_1 - (a \cdot x_1 + b))^2 + (y_2 - (a \cdot x_2 + b))^2 + (y_3 - (a \cdot x_3 + b))^2$$

Beispiel: Es sei $n = 3$ und

$$(x_1, y_1) = (0, 0), (x_2, y_2) = (1, 0), (x_3, y_3) = (-2, 1).$$

Gesucht ist dann eine Gerade

$$y = a \cdot x + b,$$

für die der folgende Ausdruck möglichst klein ist:

$$\begin{aligned}&(y_1 - (a \cdot x_1 + b))^2 + (y_2 - (a \cdot x_2 + b))^2 + (y_3 - (a \cdot x_3 + b))^2 \\ &= (0 - (a \cdot 0 + b))^2 + (0 - (a \cdot 1 + b))^2 + (1 - (a \cdot (-2) + b))^2\end{aligned}$$

Beispiel: Es sei $n = 3$ und

$$(x_1, y_1) = (0, 0), (x_2, y_2) = (1, 0), (x_3, y_3) = (-2, 1).$$

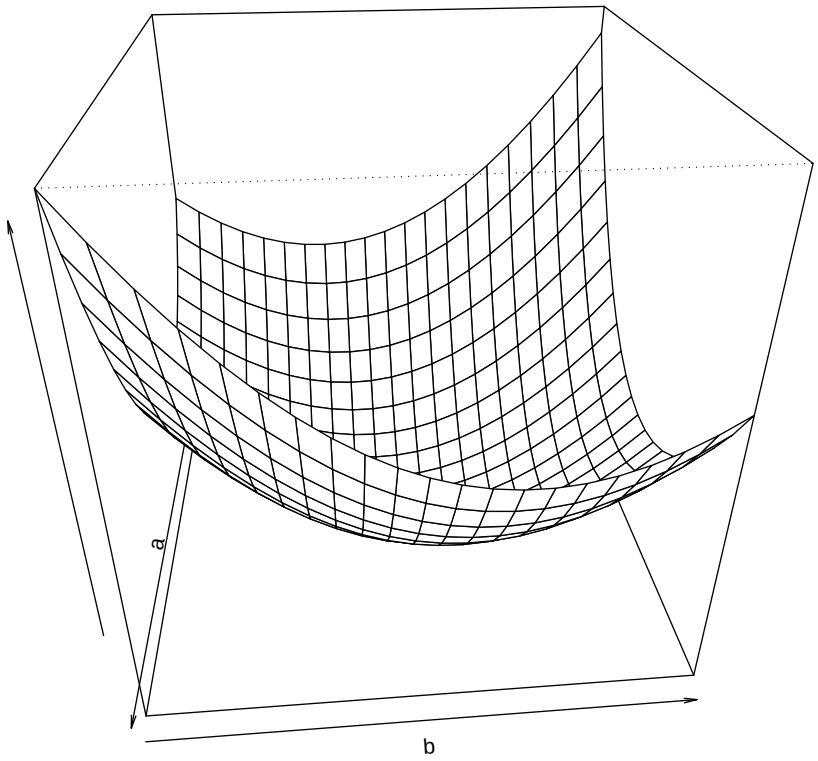
Gesucht ist dann eine Gerade

$$y = a \cdot x + b,$$

für die der folgende Ausdruck möglichst klein ist:

$$\begin{aligned}
&(y_1 - (a \cdot x_1 + b))^2 + (y_2 - (a \cdot x_2 + b))^2 + (y_3 - (a \cdot x_3 + b))^2 \\
&= (0 - (a \cdot 0 + b))^2 + (0 - (a \cdot 1 + b))^2 + (1 - (a \cdot (-2) + b))^2 \\
&= b^2 + (a + b)^2 + (1 + 2 \cdot a - b)^2.
\end{aligned}$$

In Abhängigkeit von a und b lässt sich der zu minimierende Ausdruck graphisch wie folgt darstellen:



Man kann zeigen: Der Ausdruck

$$b^2 + (a+b)^2 + (1 + 2 \cdot a - b)^2$$

wird minimal für

$$a = -\frac{5}{14} \quad \text{und} \quad b = \frac{3}{14}.$$

Man kann zeigen: Der Ausdruck

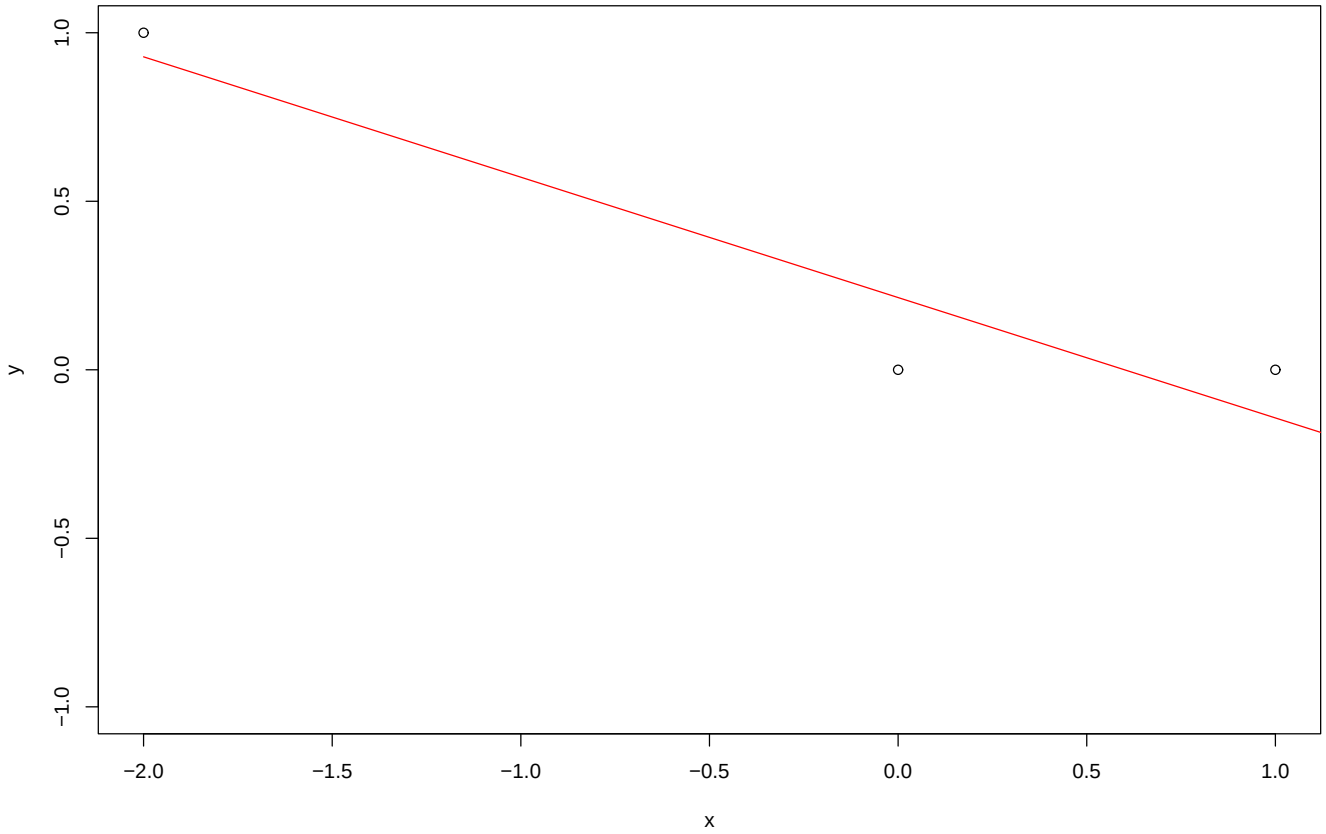
$$b^2 + (a+b)^2 + (1 + 2 \cdot a - b)^2$$

wird minimal für

$$a = -\frac{5}{14} \quad \text{und} \quad b = \frac{3}{14}.$$

Also ist die gesuchte Gerade hier gegeben durch

$$y = -\frac{5}{14} \cdot x + \frac{3}{14}.$$


1.0
0.5
0.0
−0.5
−1.0
y
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
x

Allgemein führt obige Minimierungsaufgabe auf die sogenannte **Regressionsgerade** gegeben durch

$$y = \hat{a} \cdot (x - \bar{x}) + \bar{y}$$

mit

Allgemein führt obige Minimierungsaufgabe auf die sogenannte **Regressionsgerade** gegeben durch

$$y = \hat{a} \cdot (x - \bar{x}) + \bar{y}$$

mit

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

und

Allgemein führt obige Minimierungsaufgabe auf die sogenannte **Regressionsgerade** gegeben durch

$$y = \hat{a} \cdot (x - \bar{x}) + \bar{y}$$

mit

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

und

$$\hat{a} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{x,y}}{s_x^2}$$

$(\frac{0}{0} := 0)$.

Allgemein führt obige Minimierungsaufgabe auf die sogenannte **Regressionsgerade** gegeben durch

$$y = \hat{a} \cdot (x - \bar{x}) + \bar{y}$$

mit

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

und

$$\hat{a} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{x,y}}{s_x^2}$$

$(\frac{0}{0} := 0)$.

Hierbei wird

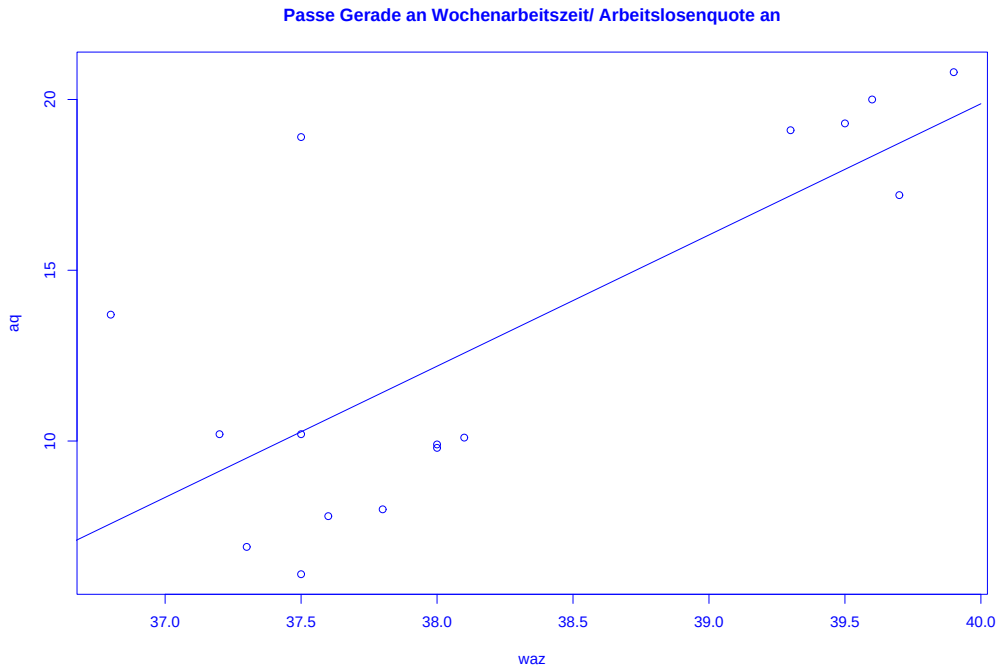
$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

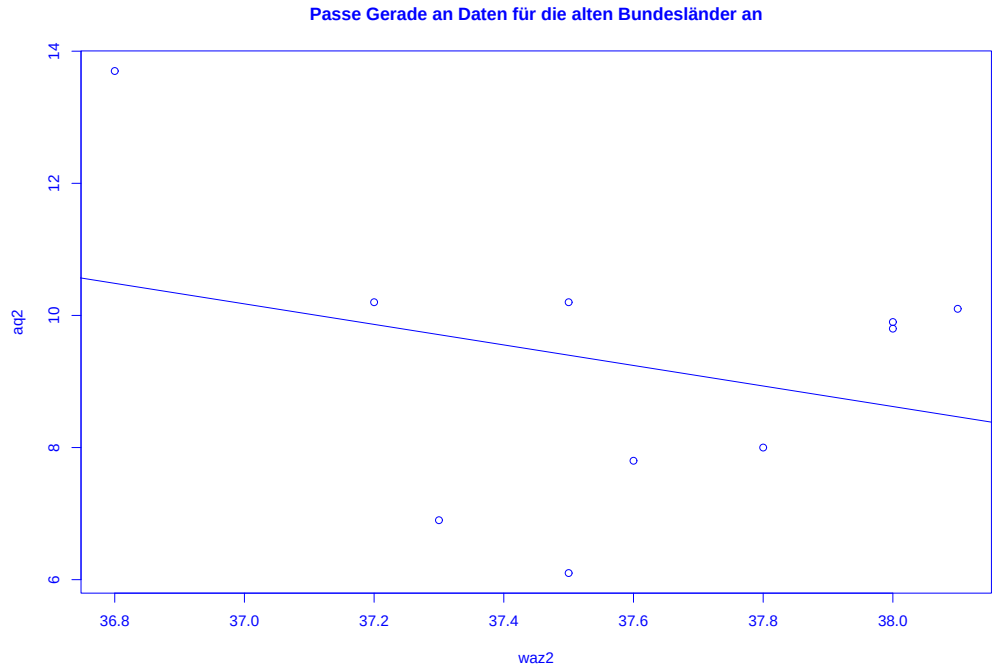
als empirische Kovarianz der zweidimensionalen Messreihe bezeichnet.

Ist die empirische Kovarianz positiv (bzw. negativ), so ist auch die Steigung der Regressionsgeraden positiv (bzw. negativ).

Ist die empirische Kovarianz positiv (bzw. negativ), so ist auch die Steigung der Regressionsgeraden positiv (bzw. negativ).

Beispiel:




Passe Gerade an Daten für die alten Bundesländer an
aq2
14
12
10
8
6
36.8
37.0
37.2
37.4
37.6
37.8
38.0
waz2

Man kann weiter zeigen, dass die sogenannte empirische Korrelation

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

im Intervall $[-1, 1]$ liegt.

Man kann weiter zeigen, dass die sogenannte empirische Korrelation

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

im Intervall $[-1, 1]$ liegt.

Die empirische Korrelation dient zur Beurteilung der Abhängigkeit der x– und der y–Koordinaten.

Sie macht Aussagen über die Regressionsgerade und die Lage der Punktwolke im Scatterplot:

- Ist die empirische Korrelation $+1$ oder -1, so liegen die Punkte (x_i, y_i) alle auf der Regressionsgeraden.

Sie macht Aussagen über die Regressionsgerade und die Lage der Punktwolke im Scatterplot:

- Ist die empirische Korrelation $+1$ oder -1, so liegen die Punkte (x_i, y_i) alle auf der Regressionsgeraden.

- Ist die empirische Korrelation positiv (bzw. negativ), so ist auch die Steigung der Regressionsgeraden positiv (bzw. negativ).

Sie macht Aussagen über die Regressionsgerade und die Lage der Punktwolke im Scatterplot:

- Ist die empirische Korrelation $+1$ oder -1, so liegen die Punkte (x_i, y_i) alle auf der Regressionsgeraden.

- Ist die empirische Korrelation positiv (bzw. negativ), so ist auch die Steigung der Regressionsgeraden positiv (bzw. negativ).

- Ist die empirische Korrelation Null, so verläuft die Regressionsgerade waagrecht.

3.5 Nichtparametrische Regressionsschätzung

Verallgemeinerung der linearen Regression:

Passe Funktionen allgemeinerer Bauart (z.B. Polynome) an Daten an. Z.B. wie bei linearer Regression durch Minimierung der Summe der quadratischen Fehler (Prinzip der Kleinsten-Quadrate).

3.5 Nichtparametrische Regressionsschätzung

Verallgemeinerung der linearen Regression:

Passe Funktionen allgemeinerer Bauart (z.B. Polynome) an Daten an. Z.B. wie bei linearer Regression durch Minimierung der Summe der quadratischen Fehler (Prinzip der Kleinsten-Quadrate).

Falls Bauart vorgegeben ist und diese nur von endlich vielen Parametern abhängt: **parametrische Regressionsschätzung**.

3.5 Nichtparametrische Regressionsschätzung

Verallgemeinerung der linearen Regression:

Passe Funktionen allgemeinerer Bauart (z.B. Polynome) an Daten an. Z.B. wie bei linearer Regression durch Minimierung der Summe der quadratischen Fehler (Prinzip der Kleinsten-Quadrate).

Falls Bauart vorgegeben ist und diese nur von endlich vielen Parametern abhängt: **parametrische Regressionsschätzung**.

Anderer Ansatz:

Nichtparametrische Regressionsschätzung.

Keine Annahme über die Bauart der anzupassenden Funktion.

Einfachstes Beispiel: lokale Mittelung

Versucht wird, den durchschnittlichen Verlauf der y-Koordinaten der Datenpunkte in Abhängigkeit der zugehörigen x-Koordinaten zu beschreiben.

Einfachstes Beispiel: lokale Mittelung

Versucht wird, den durchschnittlichen Verlauf der y-Koordinaten der Datenpunkte in Abhängigkeit der zugehörigen x-Koordinaten zu beschreiben.

z.B. durch sogenannten **Kernschätzer**:

$$m_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \cdot y_i}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}.$$

Hierbei ist $K : \mathbb{R} \to \mathbb{R}_+$ die sogenannte **Kernfunktion** und $h > 0$ die sogenannte **Bandbreite**.

z.B. naiver Kern

$$K(u) = \frac{1}{2} 1_{[-1,1]}(u)$$

oder Gauss-Kern

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2).$$

z.B. naiver Kern

$$K(u) = \frac{1}{2} 1_{[-1,1]}(u)$$

oder Gauss-Kern

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2).$$

Wie beim Kern-Dichteschätzer bestimmt die Bandbreite die Glattheit bzw. Rauheit der Schätzung.