

6 Die Methode der kleinsten Quadrate

6.1 Einführung

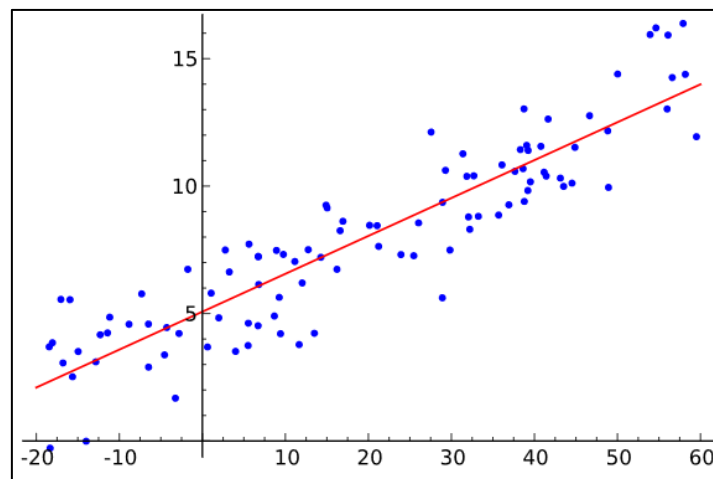
In diesem Kapitel geht es um eine weit verbreitete Optimierungsmethode zur Modellierung mathematischer Zusammenhänge in grossen Datenmengen. Dabei geht es darum, die optimalen Parameter zu finden, welche den funktionalen Zusammenhang zwischen den Messdaten am besten beschreiben. Bei der *linearen Regression* wird beispielsweise ein linearer Zusammenhang zwischen den Daten vermutet und versucht, eine optimale Gerade in die Datenmenge einzupassen.

6.2 Lineare Regression

Problem

Es liegt ein Streudiagramm von Messpunkten $(x_i; y_i)$ vor, die zwar nicht genau, aber doch näherungsweise auf einer Geraden liegen.

Nun soll die Gerade bestimmt werden, die «bestmöglich» auf die Datenpunkte passt.



Zunächst müssen wir festlegen, was in diesem Zusammenhang «bestmöglich» bedeuten soll.

Definition

Gegeben sind Datenpunkte $(x_i; y_i)$ mit $1 \leq i \leq n$.

Die *Residuen* oder *Fehler* $\epsilon_i = y_i - g(x_i)$ dieser Datenpunkte sind die Abstände in y -Richtung zwischen y_i und der Geraden g .

Die «bestmögliche» Gerade, die *Ausgleichs-* oder *Regressionsgerade*, sei diejenige Gerade, für die die Summe der quadrierten Residuen $\sum_{i=1}^n \epsilon_i^2$ am kleinsten ist. Gesucht ist also das Minimum der folgenden Summe:

$$\sum_{i=1}^n \underbrace{(y_i - g(x_i))^2}_{=\epsilon_i} = \sum_{i=1}^n \underbrace{(y_i - \hat{y}_i)^2}_{=\epsilon_i}$$

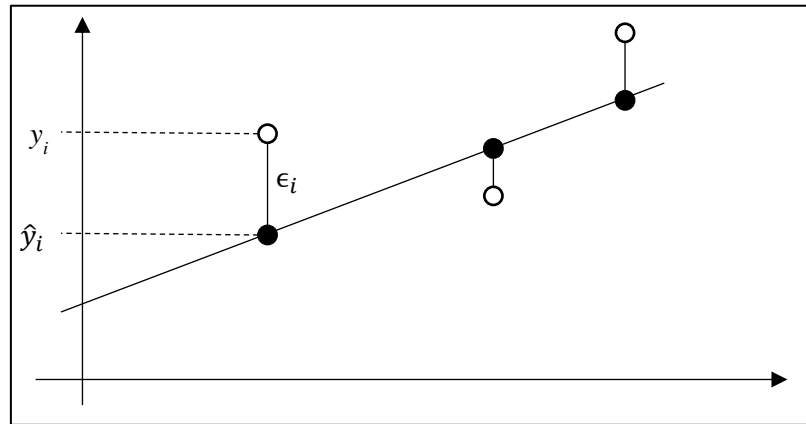
y_i : beobachtete y -Werte

\hat{y}_i : prognostizierte bzw. erklärte y -Werte

ϵ_i : Residuen (oder auch Fehler)

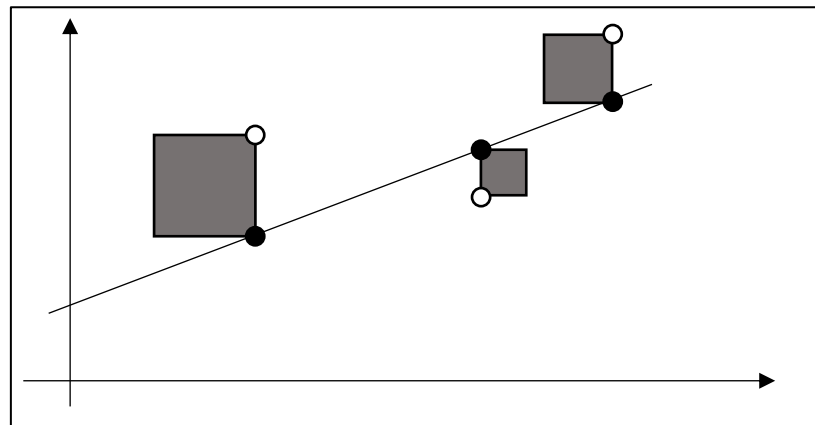
Man spricht daher von der.....

Grafisch dargestellt ergeben sich also die Residuen folgendermassen:



Nach der KQM wird die Gerade g mit $g(x_i) = \hat{y}_i$ so bestimmt, dass die Summe der Residuenquadrate minimal wird:

Diese Summe entspricht der Gesamtfläche der im folgenden Bild eingefärbten Quadrate mit Kantenlänge ϵ_i :



Vorgehensweise:

Für die Gleichung der Regressionsgeraden $g(x) = mx + d$ gilt es also, m und d so zu bestimmen, dass die Residualvarianz

minimal wird.

Das Minimum (nach den Parametern m und d) kann man u.a. finden, indem man die partiellen Ableitungen nach m und d bildet und deren Nullstellen sucht:

$$\frac{\partial}{\partial m} \tilde{s}_\epsilon^2 = -\frac{2}{n} \sum_{i=1}^n x_i (y_i - mx_i - d) = 0 \quad (1)$$

$$\frac{\partial}{\partial d} \tilde{s}_\epsilon^2 = -\frac{2}{n} \sum_{i=1}^n (y_i - mx_i - d) = 0 \quad (2)$$

Umgeformt ergibt sich aus der zweiten Bedingung:

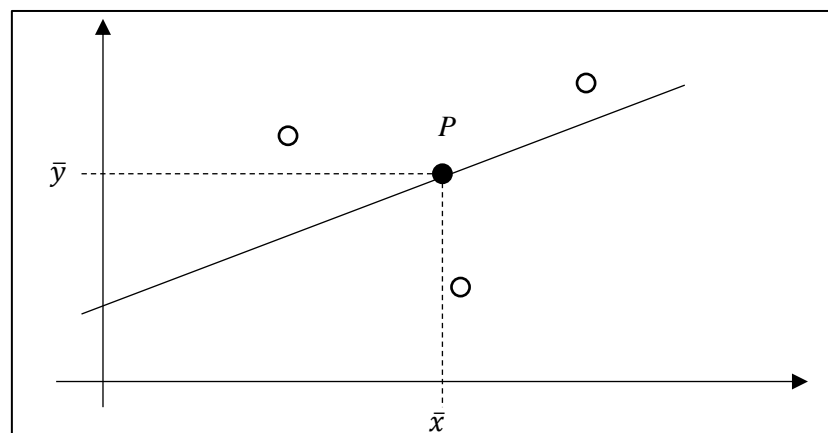
$$0 = \frac{1}{n} \sum_{i=1}^n (y_i - mx_i - d) = \frac{1}{n} \sum_{i=1}^n y_i - m \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n d = \bar{y} - m\bar{x} - d \quad (3)$$

Diese Formel deutet bereits die spezielle Lage der gesuchten Geraden an, denn es folgt daraus

$$\bar{y} = m\bar{x} + d$$

d.h. die Regressionsgerade verläuft durch.....der Punktmenge.
Wir können mit (3) bereits einen Parameter der Geraden bestimmen:

(4)



Aus der ersten Bedingung ergibt sich:

$$0 = \frac{1}{n} \sum_{i=1}^n x_i (y_i - mx_i - d) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \frac{m}{n} \sum_{i=1}^n x_i^2 - \frac{d}{n} \sum_{i=1}^n x_i = \overline{xy} - m\overline{x^2} - d\bar{x} \quad (5)$$

Setzen wir nun (4) in (5) ein und lösen nach m auf, erhalten wir für den zweiten Parameter m :

$$m = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{x^2} - \bar{x}^2} \quad (6) \quad \begin{array}{l} \tilde{s}_{xy}: \text{Kovarianz} \\ \tilde{s}_x^2: \text{Varianz der } x_i\text{-Werte} \end{array}$$

Damit sind die Parameter der Regressionsgerade mit Hilfe bereits bekannter Kennwerte bestimmbar.

Die (minimale) sogenannte Varianz der Residuen ergibt sich dabei zu:

$$\tilde{s}_\epsilon^2 = \frac{1}{n} \sum_{i=1}^n (y_i - mx_i - d)^2 \quad \text{mit } d = \bar{y} - m\bar{x} \quad (4)$$

$$= \frac{1}{n} \sum_{i=1}^n ((y_i - \bar{y}) - (mx_i - m\bar{x}))^2$$

$$= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 - \frac{2m}{n} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + \frac{m^2}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$= \tilde{s}_y^2 - 2m\tilde{s}_{xy} + m^2\tilde{s}_x^2 \quad \text{mit } m = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2} \quad (6)$$

$$= \tilde{s}_y^2 - 2 \frac{\tilde{s}_{xy}}{\tilde{s}_x^2} \tilde{s}_{xy} + \left(\frac{\tilde{s}_{xy}}{\tilde{s}_x^2} \right)^2 \tilde{s}_x^2$$

$$\tilde{s}_\epsilon^2 = \tilde{s}_y^2 - \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}$$

Zusammenfassung Regressionsgerade:

Die *Regressionsgerade* $g(x) = mx + d$ mit den Parametern m und d ist die Gerade, für die die *Residualvarianz* \tilde{s}_ϵ^2 minimal ist.

Die Regressionsgerade hat die Steigung

$$m = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}$$

und den y-Achsenabschnitt

$$d = \bar{y} - m\bar{x}$$

Für die zugehörige (minimale) Residualvarianz gilt:

$$\tilde{s}_\epsilon^2 = \tilde{s}_y^2 - \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}$$

mit:

Varianz der x_i -Werte

$$\tilde{s}_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2$$

Varianz. der y_i -Werte

$$\tilde{s}_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2$$

Kovarianz

$$\tilde{s}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{x} \cdot \bar{y}$$

arithmetische Mittelwerte

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

Hinweis: Die Berechnung kann auch mit den korrigierten Werten s_x^2 , s_y^2 , s_{xy} erfolgen

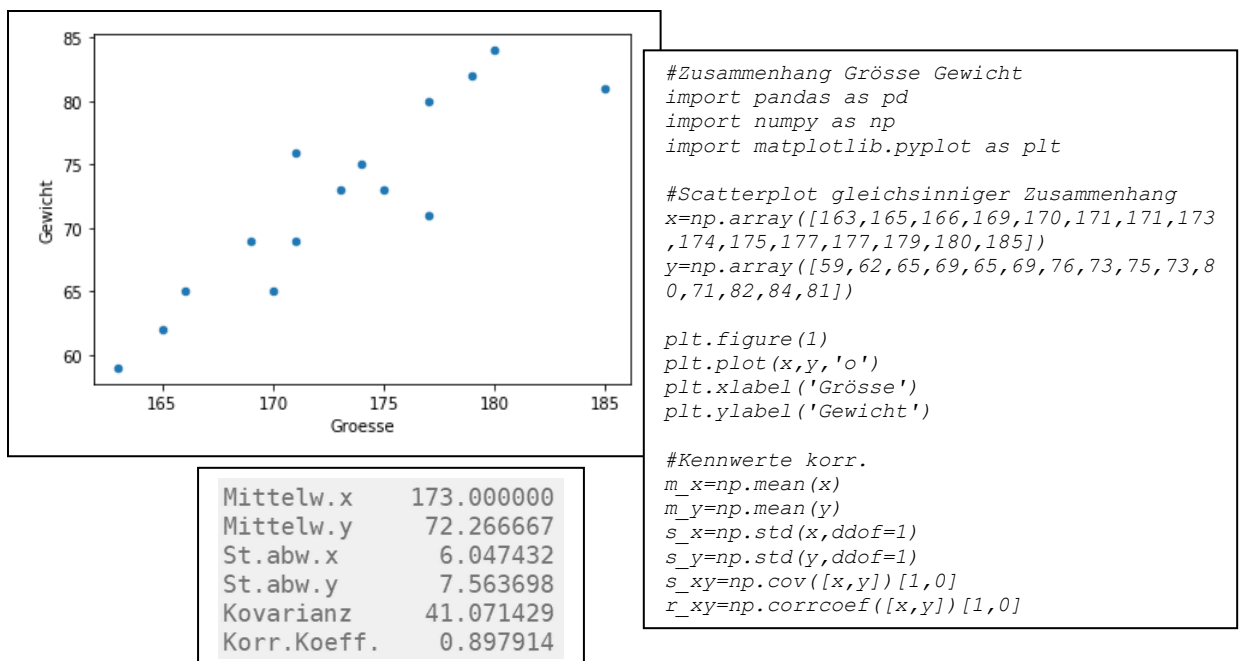
Kehren wir für die erste Anwendung wieder zu einem Beispiel zurück, das wir bereits vom Beginn des Semesters her kennen:

Beispiel 1: Zusammenhang Grösse und Gewicht

Von 15 zufällig ausgewählten erwachsenen Personen werden Körpergrösse x und Gewicht y gemessen. Es ergeben sich folgende Wertepaare (x_i, y_i) in cm bzw. kg:

(163,59), (165,62), (166,65), (169,69), (170,65), (171,69), (171,76), (173,73),
(174,75), (175,73), (177,80), (177,71), (179,82), (180,84), (185,81)

Wir haben bereits einen Korrelationskoeffizient $r_{xy} = 0.898$ gefunden und das folgende Streudiagramm:

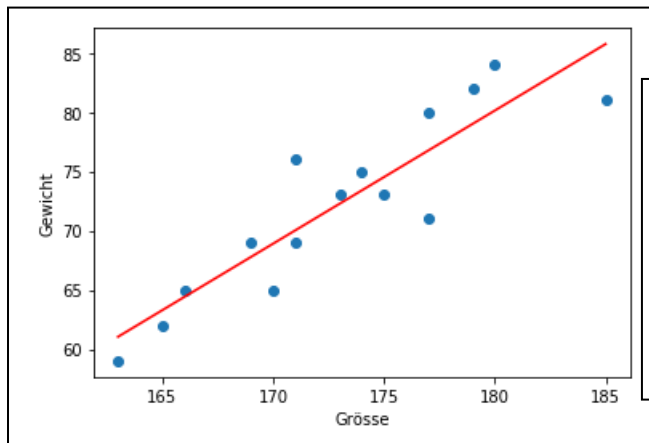


Beide Informationen lassen vermuten, dass wir die Werte durch eine Gerade annähern können.

Wir bestimmen also die Steigung m der Regressionsgeraden durch:

und den y-Achsabschnitt d :

Für die Ermittlung der Regressionsgeraden in Python existieren verschiedene Funktionen (z.B. in numpy oder scikit-learn).



```
#Regressionsgerade Variante 1
m1=s_xy/s_x**2
d1=m_y-m_x*m1

#Regressionsgerade Variante 2 mit numpy
m2,d2 = np.polyfit(x, y, 1)

#Scatterplot mit Ausgleichsgerade
plt.figure(2)
plt.plot(x,y,'o',x,m2*x+d2,'r')
plt.xlabel('Grösse')
plt.ylabel('Gewicht')
```

6.2.1 Varianzzerlegung und Bestimmtheitsmass

Mithilfe der *Residuen*

$$\epsilon_i = y_i - \hat{y}_i$$

zwischen den beobachteten y -Werten y_i und den durch die Regression *prognostizierten* y -Werten

$$\hat{y}_i = g(x_i) = mx_i + d$$

könnte man nun für jeden einzelnen Datenpunkt überprüfen, wie gut er aufgrund des Modells vorhergesagt worden wäre.

Damit haben wir aber noch kein Mass gefunden, mit dem wir die Güte des Modells insgesamt beurteilen können. Ein solches Mass können wir vielmehr über die Varianzzerlegung erhalten. Die Frage ist dabei: Welcher Anteil der Varianz der y_i lässt sich durch die Regression erklären?

Für die Varianz der prognostizierten Werte gilt:

$$\tilde{s}_y^2 = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2$$

Der Mittelwert der prognostizierten Werte $\bar{\hat{y}}$ stimmt dabei überein mit dem Mittelwert der beobachteten Werte \bar{y} , denn:

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n (mx_i + d) = m\bar{x} + d = \bar{y}$$

damit gilt:

$$\tilde{s}_y^2 = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (mx_i + d - \bar{y})^2$$

und mit $d = \bar{y} - m\bar{x}$ und $m = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}$ der Regressionsgeraden ergibt sich:

$$\tilde{s}_y^2 = \frac{1}{n} \sum_{i=1}^n (mx_i + \bar{y} - m\bar{x} - \bar{y})^2 = \frac{m^2}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = m^2 \tilde{s}_x^2 = \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}$$

Wir erhalten damit und mit der Gleichung für die Residualvarianz $\tilde{s}_\epsilon^2 = \tilde{s}_y^2 - \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}$ die Varianzzerlegung der totalen Varianz:

mit \tilde{s}_y^2 : Totale Varianz, \tilde{s}_y^2 : prognostizierte (erklärte) Varianz, \tilde{s}_ϵ^2 : Residualvarianz

Als Masszahl für die Güte der Modellanpassung verwendet man eine Grösse, die auf dieser Varianzzerlegung aufbaut. Diese gibt gerade den Anteil der Totalen Varianz \tilde{s}_y^2 an, der durch die Regression erklärt wird, und ist somit der Quotient aus erklärter Varianz \tilde{s}_y^2 und totaler Varianz \tilde{s}_y^2 .

Diese Masszahl wird *Bestimmtheitsmass* R^2 genannt.

Das Bestimmtheitsmass nimmt Werte zwischen Null und Eins an. Dabei bedeutet ein Wert von 0, dass die prognostizierte (erklärte) Varianz gleich Null und somit das Regressionsmodell denkbar schlecht ist.

Zum Beispiel bedeutet $R^2 = 0.75$,

Damit ergibt sich auch eine neue Interpretation des *Korrelationskoeffizienten* (nach Bravais-Pearson). Der quadrierte Korrelationskoeffizient entspricht dem Anteil der erklärten Varianz an der Totalen Varianz.

Zusammenfassung Bestimmtheitsmass:

Die Totale Varianz setzt sich zusammen aus der Residualvarianz und der Varianz der prognostizierten Werte:

$$\tilde{s}_y^2 = \tilde{s}_\epsilon^2 + \tilde{s}_y^2 \text{ bzw. } s_y^2 = s_\epsilon^2 + s_y^2$$

Das *Bestimmtheitsmass* R^2 beurteilt die globale Anpassungsgüte einer Regression über den Anteil der prognostizierten (erklärten) Varianz \tilde{s}_y^2 an der totalen Varianz \tilde{s}_y^2 :

$$R^2 = \frac{\tilde{s}_y^2}{\tilde{s}_y^2} \text{ bzw. } R^2 = \frac{s_y^2}{s_y^2}$$

Das *Bestimmtheitsmass* R^2 stimmt überein mit dem Quadrat des *Korrelationskoeffizienten* (nach Bravais-Pearson)

$$R^2 = \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2 \tilde{s}_y^2} = r_{xy}^2 \text{ bzw. } R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = r_{xy}^2$$

6.2.2 Residuenbetrachtung

Neben dieser eher formalen Überprüfung der Güte des Modells mittels des Bestimmtheitsmasses kann man auch anhand grafischer Darstellungen einen guten Eindruck gewinnen. Denn selbst wenn die Punkte in der Nähe einer Geraden liegen, also $R^2 \approx 1$, kann man daraus nicht unbedingt einen linearen Trend folgern. Zudem kann ein linearer Trend auch bei kleinem R^2 bestehen.

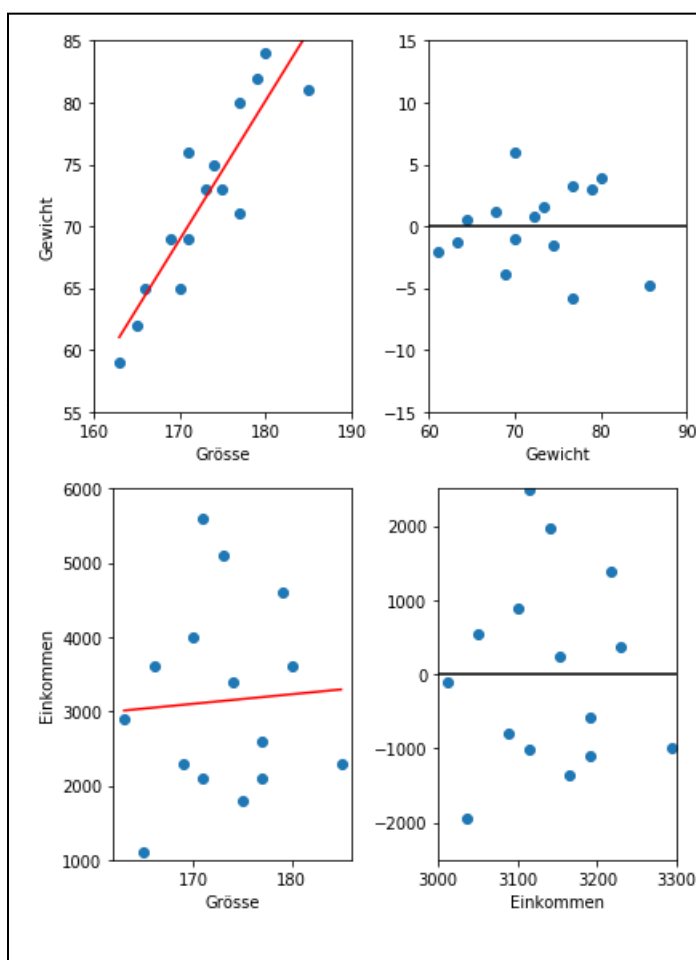
Zusätzlich kann man die Regression durch Betrachtung des Streudiagramms und des *Residuenplots* bewerten.

Die Residuen werden dabei bezogen auf diedargestellt. Auf der horizontalen Achse werden die prognostizierten y-Werte \hat{y} und auf der vertikalen Achse die Residuen angetragen.

Die Residuen sollten unsystematisch (d.h. zufällig) und überall etwa gleich um die horizontale Achse streuen, betragsmässig kleine Residuen sollten häufiger sein als grosse.

Beispiel 2: Residuenplots

Nun werden wieder die Beispiele zu Personengrösse gegen Gewicht, Einkommen und Laufzeit (Bsp. 2 aus Kapitel 2) betrachtet. Dargestellt werden jeweils das Streudiagramm mit der Regressionsgeraden und der Residuenplot.



```
#Verschiedene Residuenplot
import numpy as np
import matplotlib.pyplot as plt
#x=Grösse
#y1=Gewicht
#y2=Einkommen
#y3=Laufzeit

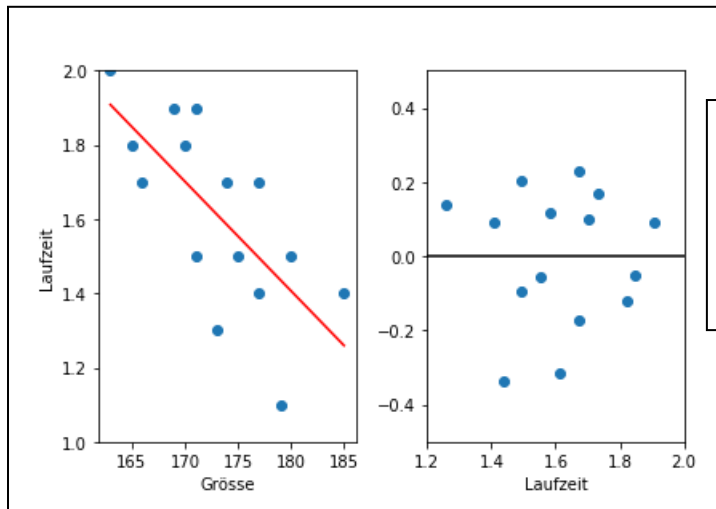
#Regression
m1,d1 = np.polyfit(x, y1, 1)
m2,d2 = np.polyfit(x, y2, 1)
m3,d3 = np.polyfit(x, y3, 1)

#Berechnete y-Werte
y1b=m1*x+d1
y2b=m2*x+d2
y3b=m3*x+d3

#Residuen
R1=y1-y1b
R2=y2-y2b
R3=y3-y3b

plt.figure(1)
plt.subplot(121)
plt.plot(x,y1,'o',x,m1*x+d1,'r')
plt.xlabel('Grösse')
plt.ylabel('Gewicht')
plt.subplot(122)
plt.plot(y1b,R1,'o',[60,90],[0,0],'k')
plt.xlabel('Gewicht')

plt.figure(2)
plt.subplot(121)
plt.plot(x,y2,'o',x,m2*x+d2,'r')
plt.xlabel('Grösse')
plt.ylabel('Einkommen')
plt.subplot(122)
plt.plot(y2b,R2,'o',[3000,3300],[0,0],'k')
plt.xlabel('Einkommen')
```

```
plt.figure(3)
plt.subplot(121)
plt.plot(x,y3,'o',x,m3*x+d3,'r')
plt.xlabel('Grösse')
plt.ylabel('Laufzeit')
plt.subplot(122)
plt.plot(y3b,R3,'o',[1, 2],[0, 0],'k')
plt.xlabel('Laufzeit')
```

Beim ersten Teil des Beispiels kann ein linearer Trend vermutet werden. Die Residuen streuen wenig und in etwa gleich für grosse und kleine Werte des Gewichtes.

Die Residuen streuen beim Teil zwei und drei des Beispiels sehr viel stärker. Daran kann man zusätzlich zum Korrelationskoeffizienten erkennen, dass die Daten durch das vorgegebene lineare Modell schlechter beschrieben werden.

6.3 Nichtlineares Verhalten

Das bislang vorgestellte Modell ging von der Anpassung einer Ausgleichsgeraden an die Daten aus. Oft zeigt bereits ein erster Blick auf das Streudiagramm der Daten aber, dass diese evtl. besser durch eine andere Funktion beschrieben werden können.

Auch für nichtlineares Verhalten können die im Ansatz enthaltenen Modellparameter mithilfe der Gauss'schen Methode der kleinsten Quadrate bestimmt werden.

Betrachten wir etwa die Situation, dass die y -Werte in etwa exponentiell mit den x -Werten anwachsen, also

$$y_i \approx \alpha e^{\beta x_i} \quad i = 1, \dots, n$$

so lassen sich die Schätzungen für α und β prinzipiell nach der KQM herleiten.

Aber nicht immer sind die Schätzungen explizit darstellbar. Dies macht den Einsatz von numerischen Verfahren zur Bestimmung der Schätzer erforderlich.

In manchen Fällen besteht jedoch die Möglichkeit, durch eine geschickte Transformation ein nichtlineares Regressionsmodell auf ein lineares Modell zurückzuführen.

Im obigen Beispiel liefert die Verwendung der Logarithmus-Funktion:

Nach Anwendung dieser Transformation lassen sich $\ln(\alpha)$ und β wie zuvor schätzen.

Für verschiedene Funktionen ist eine Linearisierung möglich, wie z.B. für:

Ausgangsfunktion	Transformation
$y = q \cdot x^m$	$\log(y) = \log(q) + m \cdot \log(x)$
$y = q \cdot m^x$	$\log(y) = \log(q) + \log(m) \cdot x$
$y = q \cdot e^{m \cdot x}$	$\ln(y) = \ln(q) + m \cdot x$
$y = \frac{1}{q + m \cdot x}$	$V = q + m \cdot x; \quad V = \frac{1}{y}$
$y = q + m \cdot \ln(x)$	$y = q + m \cdot U; \quad U = \ln(x)$
$y = \frac{1}{q \cdot m^x}$	$\log\left(\frac{1}{y}\right) = \log(q) + \log(m) \cdot x$

6.4 Allgemeines Vorgehen

In manchen Fällen soll die Abhängigkeit nicht allein von einer, sondern von mehreren Variablen betrachtet werden. Hier ist das Vorgehen vergleichbar und wir können dies als ein Problem der linearen Algebra betrachten.

Im Folgenden ist das Vorgehen für einen linearen Zusammenhang beschrieben.

Allgemein kann man dabei die KQM in der Matrixdarstellung formulieren.

Gegeben sind n Gleichungen und wir suchen k Regressionsparameter: p_1, p_2, \dots, p_k .

Dafür ergibt sich das folgende lineare Gleichungssystem für $(k - 1)$ Merkmale:

$$(*) \begin{cases} y_1 = \underbrace{p_1 x_{11} + p_2 x_{12} + \dots + p_{k-1} x_{1(k-1)}}_{=\hat{y}_1} + p_k + \epsilon_1 \\ \dots \\ y_n = \underbrace{p_1 x_{n1} + p_2 x_{n2} + \dots + p_{k-1} x_{n(k-1)}}_{=\hat{y}_n} + p_k + \epsilon_n \end{cases} \quad \text{mit den Residuen } \epsilon_i$$

Mit den Vektoren der Parameter $p = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{pmatrix}$, der Messwerte $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, der Residuen $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$

und der Matrix der Eingangswerte $X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1(k-1)} & 1 \\ x_{21} & x_{22} & \dots & x_{2(k-1)} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{n(k-1)} & 1 \end{pmatrix}$,

lautet das Gleichungssystem (*):

Die KQM verlangt dann, den Vektor p so zu bestimmen, dass die Summe der Fehlerquadrate $\sum_{i=1}^n \epsilon_i^2$ minimal ist.

Die Lösung ist

$$p = (X^T X)^{-1} X^T y$$

falls $(X^T X)$ invertierbar.

Analytische Herleitung:

Es gilt die Summe der Residuenquadrate zu minimieren:

$$S_{\epsilon^2}(p) = (y - Xp)^T \cdot (y - Xp) = (y^T y - 2p^T X^T y + p^T X^T X p) \rightarrow \min$$

$$\frac{\partial}{\partial p} S_{\epsilon^2}(p) = \begin{pmatrix} \frac{\partial}{\partial p_1} S(p) \\ \frac{\partial}{\partial p_2} S(p) \\ \vdots \\ \frac{\partial}{\partial p_k} S(p) \end{pmatrix} = 0$$

Die partiellen Ableitungen lauten:

$$\frac{\partial}{\partial p_1} S_{\epsilon^2}(p) = \frac{\partial}{\partial p_1} y^T y - \frac{\partial}{\partial p_1} 2p^T X^T y + \frac{\partial}{\partial p_1} p^T X^T X p = -2x_{(1)}^T y + 2x_{(1)}^T X p$$

$$\frac{\partial}{\partial p_2} S_{\epsilon^2}(p) = \frac{\partial}{\partial p_2} y^T y - \frac{\partial}{\partial p_2} 2p^T X^T y + \frac{\partial}{\partial p_2} p^T X^T X p = -2x_{(2)}^T y + 2x_{(2)}^T X p$$

\vdots

$$\frac{\partial}{\partial p_k} S_{\epsilon^2}(p) = \frac{\partial}{\partial p_k} y^T y - \frac{\partial}{\partial p_k} 2p^T X^T y + \frac{\partial}{\partial p_k} p^T X^T X p = -2x_{(k)}^T y + 2x_{(k)}^T X p$$

wobei x_j der Vektor der j -ten Spalte der Matrix X ist.

Damit lässt sich der Regressionsvektor darstellen als:

$$\frac{\partial}{\partial p} S_{\epsilon^2}(p) = -2X^T y + 2X^T X p = 0$$

Umgeformt ergibt sich:

Die Summe der Residuenquadrate ist also genau dann minimal, wenn p eine Lösung dieser sogenannten Normalengleichung ist.

Falls die Matrix $X^T X$ invertierbar ist, d.h. wenn die Matrix X vollen Rang hat, gibt es dabei genau eine Lösung:

Geometrische Herleitung:

Die KQM verlangt, den Vektor p so zu bestimmen, dass die Summe der Residuenquadrate minimal ist.

$$\sum_{i=1}^n \epsilon_i^2 \rightarrow \min$$

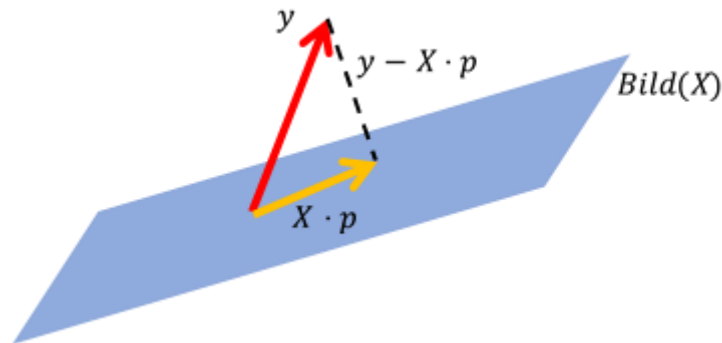
Dies ist gleichbedeutend mit dem Optimierungsproblem, denjenigen Vektor p zu bestimmen, für den die euklidische Norm.....

$$|y - X \cdot p|^2 = |\epsilon|^2 \rightarrow \min$$

Das ist geometrisch gesehen der Fall, wenn

.....
ist, d.h. auf den Vektorraum $\text{Bild}(X)$, welcher durch die Spaltenvektoren von X erzeugt wird.

Der Vektor p , wird so bestimmt, dass $X \cdot p$ und $\epsilon = y - X \cdot p$ sind.



Wir nehmen an, dass $y = X \cdot p + \epsilon$ eine solche orthogonale Zerlegung ist, d.h. $X \cdot p$ und ϵ stehen senkrecht aufeinander. Das Bild $\text{Bild}(X)$ wird durch die Spaltenvektoren von X erzeugt, welche senkrecht zu ϵ stehen. Wir folgern daraus für die Transponierte von X , dass gilt:

$$X^T \epsilon = 0$$

Wenn wir $y = X \cdot p + \epsilon$ nun mit X^T multiplizieren, so folgt:

$$X^T y = X^T (X \cdot p + \epsilon) = X^T X \cdot p + X^T \epsilon = X^T X \cdot p$$

Diese Gleichung kann man nach p auflösen, wenn $X^T X$ invertierbar ist. Dies ist der Fall, wenn die Spaltenvektoren von X linear unabhängig sind. Dann folgt die Berechnungsvorschrift für p :

Wegen der orthogonalen Zerlegung von y in die vektoriellen Komponenten $X \cdot p$ und ϵ gilt auch hier die Quadratsummenzerlegung

$$|y|^2 = |X \cdot p|^2 + |\epsilon|^2$$

Beispiel 3: Occasionfahrzeug

Der «Preis» eines Occasionsfahrzeugs soll durch den Jahrgang «Jg» und die abgelaufenen «Km» modelliert werden.

Preis	Jahrgang	Kilometerstand
28500	2017	28100
31500	2019	28200
28900	2019	28000
29500	2017	36200
24500	2019	41300
25500	2018	36600
38800	2019	22500
25400	2016	66000
23900	2017	92000
29800	2019	16000
22500	2016	8400

```
Occasionsfahrzeuge
import pandas as pd
import matplotlib.pyplot as plt

df=pd.DataFrame({'Preis':[28500,31500,28900,29500,
24500,25500,38800,25400,23900,29800,22500],
'Jg':[2017,2019,2019,2017,2019,2018,2019,2016,2017,2019,2016],
'Km':[28100,28200,28000,36200,41300,36600,22500,66000,92000,16000,8400],})
```

Mit den Vektoren der Parameter

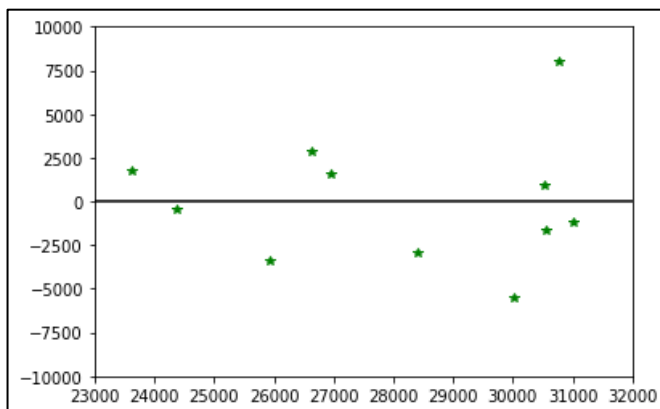
der Messwerte

und der Matrix der Eingangswerte

```
#Lineare Regression
#Minimieren Sum((Preis[i] - (a*Jg[i] + b*Km[i] + c))^2)
#Standard lineare Algebra Problem
#Es gibt auch noch diverse andere Herangehensweisen
Preis=df.iloc[:,0]
Jg=df.iloc[:,1]
Km=df.iloc[:,2]
X = np.column_stack([Jg, Km, np.ones_like(Jg)])
abc, residuals, rank, s = np.linalg.lstsq(X, Preis, rcond=-1)
a=abc[0]
b=abc[1]
c=abc[2]
```

Es ergeben sich die folgenden Modellparameter: $a = 1797.42$, $b = -0.0401$, $c = -3597323$ und einen Wert für die Summe der Residuenquadrate von $S_{\epsilon^2} = 1.34 \cdot 10^8$

Mit dem folgenden Residuenplot:



```
#Prognostizierte Preise
Preis_b=a*Jg+b*Km+c
Res=Preis-Preis_b
columns = ("Preis", "Berechneter
Preis", "Residuen")
cellText=np.column_stack([Preis,
Preis_b, Res])
```

```
#Residuenplot
plt.figure()
plt.plot(Preis_b,Res,'g*',
[23000,33000],[0,0], 'k')
plt.xlim(23000,32000)
plt.ylim(-10000,10000)
```

Der Residuenplot ist nicht ideal,

.....

.....

6.5 Lernziele für dieses Kapitel

- ☐ Ich kann die KQM am Beispiel einer Regressionsgerade erklären.
- ☐ Ich kann die Regressionsgerade zu gegebenen Punkten berechnen.
- ☐ Ich kann die lineare Regressionsmethode für mehrere Merkmale mit Hilfe von Matrizen durchführen.
- ☐ Ich verstehe wie die KQM als Problem der Linearen Algebra aufgefasst werden kann.
- ☐ Ich kann die Quadratsummenzerlegung mittels der Lösung der KQM und des Residuenvektors berechnen.
- ☐ Ich kann aus der Quadratsummenzerlegung das Bestimmtheitsmass ermitteln.
- ☐ Ich kann den Fehler berechnen, der bei der Approximation durch die KQM auftritt.
- ☐ Ich kann den Fehler grafisch bewerten mittels des Residuenplots