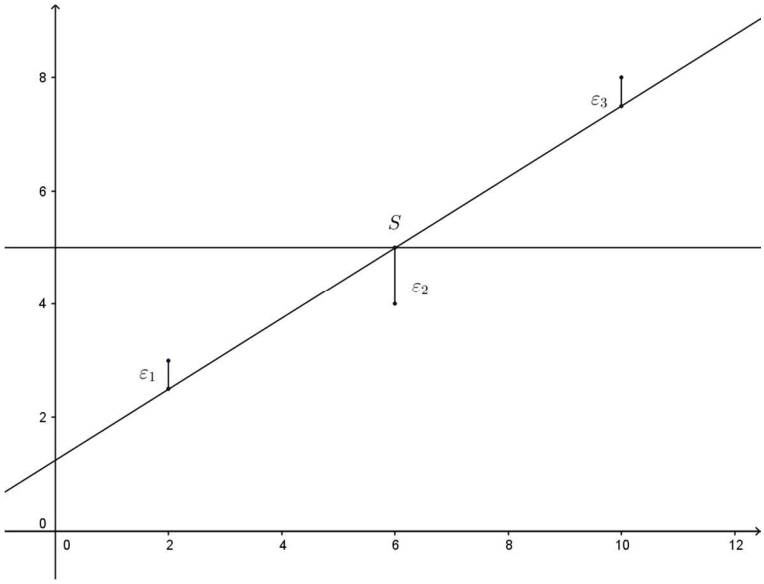
Lineare Regression und Matrizen

1. Einführendes Beispiel

Der im Kapitel Skalarprodukt gewählte Lösungsweg für das Problem der linearen Regression kann auch mit Matrizen formuliert werden. Die Idee wird zunächst am folgenden Beispiel erläutert, bei dem an n = 3 Stellen x_i die Werte y_i gemessen wurden. In diesem Spezialfall kann nämlich das Verfahren auch geometrisch interpretiert werden.

Beispiel:
Gegeben sind die drei Punkte
(2, 3), (6 ,4), (10, 8).

In der Abbildung sind dargestellt:
- der „Schwerpunkt“ $S(\bar{x}, \bar{y})$
- die Gerade $y = \bar{y}$
- die vermutete Ausgleichsgerade



Die Ausgleichsgerade kann in der folgenden Form angesetzt werden:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad 1)$$

Dabei bezeichnet x die Einflussgrösse (Regressor), y die Zielvariable (die eine Folge der Ursache x ist) und ε die Störung.
Die n = 3 Beobachtungen

im Beispiel:	allgemein
$3 = \beta_0 + \beta_1 \cdot 2 + \varepsilon_1$	$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$
$4 = \beta_0 + \beta_1 \cdot 6 + \varepsilon_2$	$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$
$8 = \beta_0 + \beta_1 \cdot 10 + \varepsilon_2$	$y_3 = \beta_0 + \beta_1 x_2 + \varepsilon_3$

können vektoriell in der folgenden Form geschrieben werden:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

Werden die Vektoren abkürzend mit $\vec{y}$, $\vec{1}$, $\vec{x}$ und $\vec{\varepsilon}$ bezeichnet, dann gilt also:

$$\vec{y} = \beta_0 \cdot \vec{1} + \beta_1 \cdot \vec{x} + \vec{\varepsilon}$$

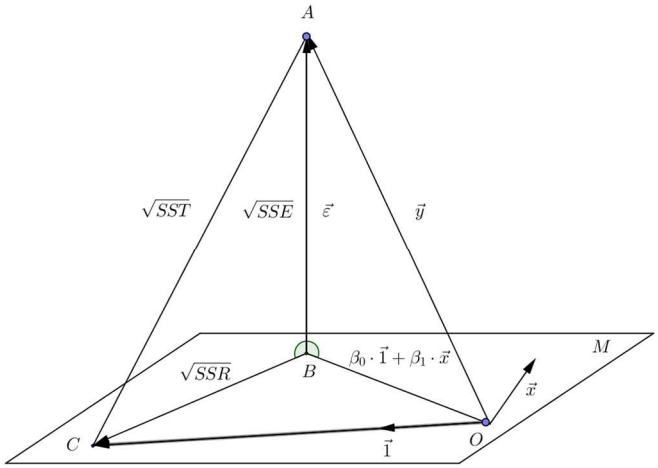
oder in Matrizenform

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_1 \\ 1 & x_1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} = (\vec{1} \quad \vec{x}) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \vec{\varepsilon} \text{ oder}$$

$$\vec{y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon} \text{ wobei } \vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \text{ und } \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_1 \\ 1 & x_1 \end{pmatrix} \qquad 2)$$

Die Aufgabe der linearen Regression besteht darin, die Parameter β_0 und β_1 aus den Daten zu schätzen und die „Güte“ der Approximation zu bewerten.

Im Fall n = 3 kann die Idee folgendermassen geometrisch interpretiert werden:
In der Abbildung sind die Vektoren $\vec{y}, \vec{1}, \ \vec{x}\ und\ \vec{\varepsilon}$ dargestellt.
Die Vektoren , $\vec{1}\ und\ \vec{x}$ spannen eine Ebene (den zweidimensionalen Modellraum M) auf.
Mit anderen Worten:
$\overrightarrow{OB}$ ist die Normalprojektion des Vektors dreidimensionalen Vektors $\vec{y}$ in den zweidimensionalen Modellraum M.



Multiplizier t man die Modellgleichung 2) mit der Matrix X^T so folgt

$$\boldsymbol{X^T\vec{y} = X^TX\vec{\beta} + X^T\vec{\varepsilon}}$$

Da das Residuum $\vec{\varepsilon}$ senkrecht auf den Vektoren $\vec{1}\ und\ \vec{x}$ steht, gilt $\boldsymbol{X^T\vec{\varepsilon} = 0}$.
Dies führt auf das folgende Gleichungssystem der sogenannten Normalgleichungen

$$\boldsymbol{X^T\vec{y} = X^TX\vec{\beta}}.$$

Löst man diese Matrizengleichung nach $\vec{\beta}$ auf, dann ergibt sich der Kleinst-Quadrat-Schätzer für $\vec{\beta}$ zu

$$\vec{\beta} = (\boldsymbol{X^TX})^{-1}\boldsymbol{X^T}\vec{y} \qquad 3)$$

Wird dieses Ergebnis auf das einführende Beispiel angewendet, so ergibt sich

$$\vec{1} = \begin{pmatrix}1\\1\\1\end{pmatrix} \qquad \vec{x} = \begin{pmatrix}2\\6\\10\end{pmatrix} \qquad Y = \vec{y} = \begin{pmatrix}3\\4\\8\end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix}1 & 2\\1 & 6\\1 & 10\end{pmatrix}$$

$$X^T \cdot X = \begin{pmatrix}1 & 1 & 1\\2 & 6 & 10\end{pmatrix}\begin{pmatrix}1\ 2\\1\ 6\\110\end{pmatrix} = \begin{pmatrix}3 & 18\\18 & 140\end{pmatrix}$$

Det($X^T \cdot X$) = 420 – 324 = 96

$$(X^T \cdot X)^{-1} = \frac{1}{96} \cdot \begin{pmatrix}140 & -18\\-18 & 3\end{pmatrix} = \begin{pmatrix}\frac{35}{24} & -\frac{3}{16}\\-\frac{3}{16} & \frac{1}{32}\end{pmatrix}$$

$$(X^T \cdot X)^{-1} \cdot X^T = \begin{pmatrix}\frac{35}{24} & -\frac{3}{16}\\-\frac{3}{16} & \frac{1}{32}\end{pmatrix} \cdot \begin{pmatrix}1 & 1 & 1\\2 & 6 & 10\end{pmatrix} = \begin{pmatrix}\frac{13}{12} & \frac{1}{3} & -\frac{5}{12}\\-\frac{1}{8} & 0 & \frac{1}{8}\end{pmatrix}$$

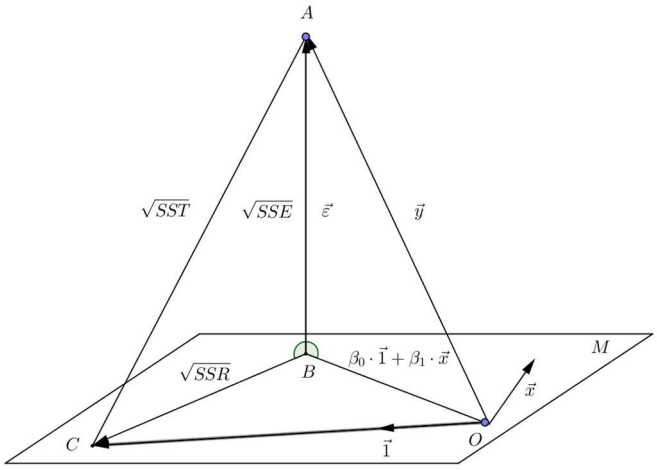
$$\vec{\beta} = (\boldsymbol{X^TX})^{-1}\boldsymbol{X^T}\vec{y} = \begin{pmatrix}\frac{13}{12} & \frac{1}{3} & -\frac{5}{12}\\-\frac{1}{8} & 0 & \frac{1}{8}\end{pmatrix} \cdot \begin{pmatrix}3\\4\\8\end{pmatrix} = \begin{pmatrix}\frac{5}{4}\\\frac{5}{8}\end{pmatrix}$$

Als Schätzung für die Parameter des Modells erhält man also:
$\beta_0 = \frac{5}{4}$ (Intercept, y-Achsenabschnitt) und
$\beta_1 = \frac{5}{4}$ (Slope, Steigung)
Die Ausgleichsgerade hat damit die Gleichung
$$y = \frac{5}{8}x + \frac{5}{4}$$

2. Das Bestimmtheitsmass

Das sogenannte Bestimmtheitsmass macht eine Aussage über die Güte der Approximation.

In der Abbildung ist B der Punkt in der Ebene M, der von A den minimalen Abstand hat.
Für den Punkt C soll gelten: $\overrightarrow{OC} = \bar{y} \cdot \vec{1}$, wo $\bar{y}$ den Mittelwert der y-Werte bezeichnet.



Für den Verbindungsvektor $\overrightarrow{CA}$ gilt dann
$\overrightarrow{CA} = \vec{y} - \bar{y} \cdot \vec{1}$

Für die Quadrate der Seiten des rechtwinkligen Dreiecks ACB sind folgende Bezeichnungen gebräuchlich:

Hypotenuse AC: $\sqrt{SST}$ **S**um of **S**quares **T**otal
Kathete BC: $\sqrt{SSR}$ **S**um of **S**quares **R**egression
Kathete AB: $\sqrt{SSE}$ **S**um of **S**quares **E**rror

Nach Pythagoras gilt die folgende für die Varianzanalyse grundlegende Beziehung

$$SST^2 = SSR^2 + SSE^2 \qquad 4)$$

Der Winkel γ in C zwischen CB und CA ist ein Mass für die Güte der Approximation. Ist γ vergleichsweise klein, dann ist die Approximation gut. Als Mass verwendet man deswegen das sogenannte Bestimmheitsmass R^2 mit

$$R^2 = cos^2\gamma = \frac{SSR}{SST} \qquad 5)$$

Das Bestimmheitsmass ist also der Anteil der Streuung, der durch die Regression „erklärt“ wird. Der restliche Anteil
$\frac{SSE}{SST}$
verbleibt zufällig.
R = 1 bedeutet dabei, dass der y-Vektor im Modellraum liegt.
R = 0 bedeutet, dass die Modellerweiterung gegenüber dem Nullmodell $\bar{y} \cdot \vec{1} \cdot$ keine Verbesserung bringt.

Bemerkung:
Bestimmheitsmass stimmt formal mit dem Quadrat des Korrelationskoeffizienten aus der deskriptiven Statistik überein.

Im einführenden Beispiel ergeben sich die folgenden Ergebnisse

	x_i	y_i	$\tilde{y}_i$	x_i^2	$(y_i-\bar{y})^2$ SST	$(\tilde{y}_i-\bar{y})^2$ SSR	$(\tilde{y}_i-y_i)^2$ SSE
	2	3	2.5	4	4	6.25	0.25
	6	4	5	36	1	0	1
	10	8	7.5	100	9	6.25	0.25
Σ	18	15	15	140	14	12.5	1.5

Aus der Tabellen ergeben sich die folgenden Werte
$\bar{x} = 6 \qquad \bar{y} = 5 \qquad SSR = 12.5 \qquad SST = 14$
und daraus das Bestimmtheitsmass
$$R^2 = \frac{SSR}{SST} = \frac{12.5}{14} \approx 0.893$$

3. Allgemeiner Fall

Die bisher im Spezialfall n = 3 formulierten Ergebnisse gelten allgemein.
Werden an n Stellen x_i die Werte y_i gemessen, dann werden.
Allgemein gilt für den Kleinst-Quadrat-Schätzer für $\vec{\beta}$

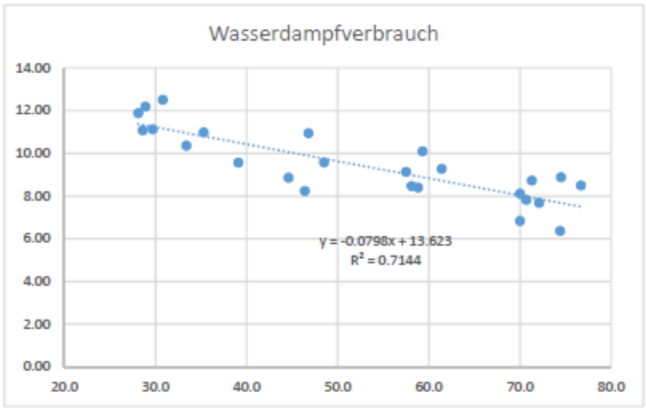
$$\vec{\beta} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}\vec{y} \qquad 3)$$

Die wichtigste Operation bei der Berechnung der Regressionskoeffizienten besteht somit in der Inversion der Matrix $X^T \cdot X$.
Es kann bewiesen werden, dass diese Matrix die folgende Form hat:

$$X^T \cdot X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \qquad 6)$$

Entsprechende Ergebnisse gelten, wenn die Werte y_i nicht nur von einer, sondern von mehreren Einflussgrössen abhängen.

Beispiel:
Quelle: Datensatz aus Draper, Smith (1966) applied regression analysis p. 351ff
In einem Chemiewerk wird heisser Wasserdampf benötigt. Untersucht wurde der Verbrauch von Wasserdampf in Abhängigkeit von 9 Einflussgrössen. In der Tabelle ist die Abhängigkeit des monatlichen Verbrauchs (Zielgrösse y) von der mittleren Tagestemperatur x in °F angegeben.



i	x_i	y_i	x_i^2	$x_i \cdot y_i$
1	35.3	10.98	1246.09	387.6
2	29.7	11.13	882.09	330.6
3	30.8	12.51	948.64	385.3
4	58.8	8.40	3457.44	493.9
5	61.4	9.27	3769.96	569.2
6	71.3	8.73	5083.69	622.4
7	74.4	6.36	5535.36	473.2
8	76.7	8.50	5882.89	652.0
9	70.7	7.82	4998.49	552.9
10	57.5	9.14	3306.25	525.6
11	46.4	8.24	2152.96	382.3
12	28.9	12.19	835.21	352.3
13	28.1	11.88	789.61	333.8
14	39.1	9.57	1528.81	374.2
15	46.8	10.94	2190.24	512.0
16	48.5	9.58	2352.25	464.6
17	59.3	10.09	3516.49	598.3
18	70.0	8.11	4900.00	567.7
19	70.0	6.83	4900.00	478.1
20	74.5	8.88	5550.25	661.6
21	72.1	7.68	5198.41	553.7
22	58.1	8.47	3375.61	492.1
23	44.6	8.86	1989.16	395.2
24	33.4	10.36	1115.56	346.0
25	28.6	11.08	817.96	316.9
Σ	**1315**	**235.6**	**76323.4**	**11821.4**

Aus der Tabelle ergibt sich die Matrix

$$X^T \cdot X = \begin{pmatrix} 25 & 1315 \\ 1315 & 76323.4 \end{pmatrix}$$

$$det(X^T \cdot X) \approx 1.7886 \cdot 10^5$$

$$(X^T \cdot X)^{-1} \approx \begin{pmatrix} \frac{76323.4}{1.7886 \cdot 10^5} & \frac{-1315}{1.7886 \cdot 10^5} \\ \frac{-1315}{1.7886 \cdot 10^5} & \frac{25}{1.7886 \cdot 10^5} \end{pmatrix} \approx \begin{pmatrix} 0.42672 & -0.007352 \\ -0.007352 & 0.0001398 \end{pmatrix}$$

$$X^T \cdot Y = \begin{pmatrix} 235.6 \\ 11821.4 \end{pmatrix}$$

$$(X^T \cdot X)^{-1} X^T \cdot Y = \begin{pmatrix} 0.42672 & -0.007352 \\ -0.007352 & 0.0001398 \end{pmatrix} \begin{pmatrix} 235.6 \\ 11821.4 \end{pmatrix} = \begin{pmatrix} 13.623 \\ -0.0798 \end{pmatrix}$$

Die Ausgleichsgerade hat also die Gleichung
$y = 13.624 - 0.0798x$