Herleitung der Parameter-Gleichungen für die einfache lineare Regression

Uwe Ziegenhagen

2. März 2013

Historie

v1.0 16.03.2009, erste Version hochgeladen

v2.0 02.03.2013, einen Vorzeichenfehler beseitigt, diverse Gleichungen und Erläuterungen zum besseren Verständnis hinzugefügt.

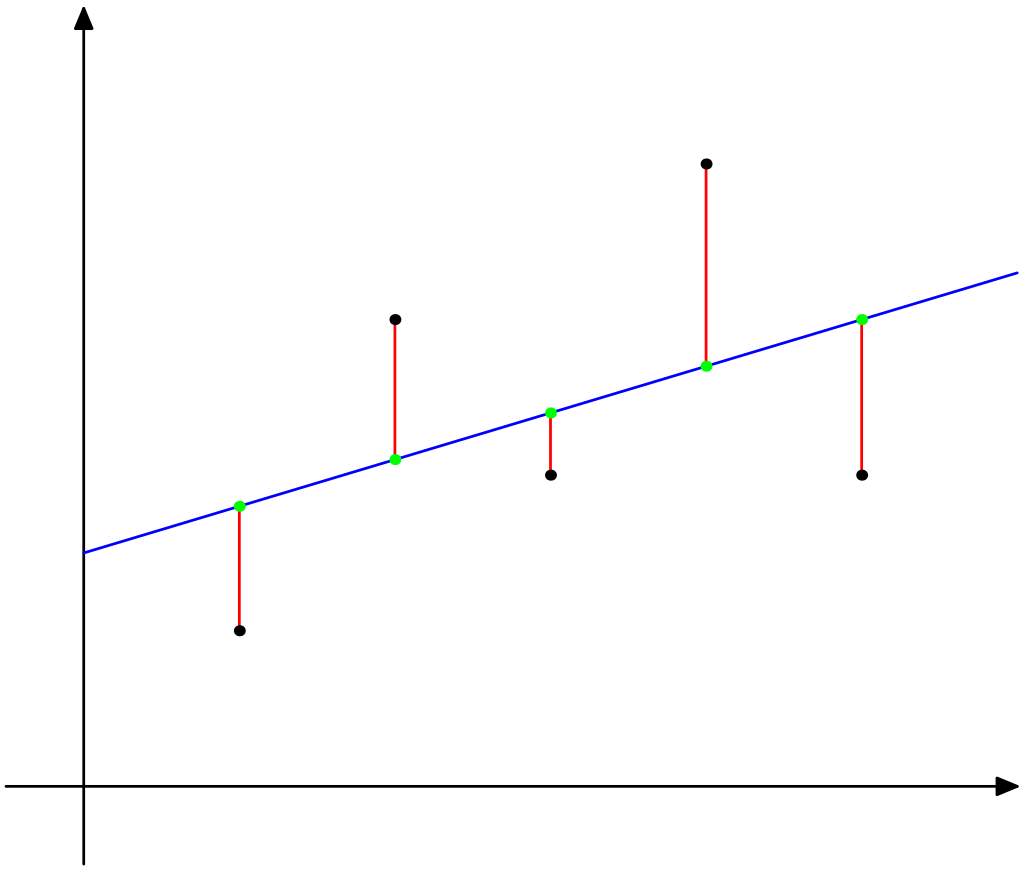
1 Einführung

Aus der Wikipedia: Die Regressionsanalyse ist ein statistisches Analyseverfahren mit dem Ziel, Beziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen festzustellen.
Allgemein wird eine metrische Variable Y betrachtet, die von einer zweiten Variablen x abhängt. Üblicherweise ist $x = (x_1, ..., x_n)^T$ ein n-dimensionaler Vektor, wobei die einzelnen x-Werte untereinander unabhängig sind. Im eindimensionalen Fall spricht man von einer einfachen linearen Regressionsanalyse, in Dimensionen größer gleich zwei von einer multiplen Regressionsanalyse.
Bei der einfachen linearen Regression liegen Wertepaare der Form (x_i, y_i) mit $i = 1, \ldots, n$ vor. Als Modell wählt man

$$Y_i = b + ax_i + \epsilon_i \tag{1}$$

man nimmt somit einen linearen Zusammenhang zwischen x_i und Y_i an. Die Daten y_i werden als Realisierungen der Zufallsvariablen Y_i angesehen, die x_i sind nicht stochastisch, sondern Messstellen. Ziel der Regressionsanalyse ist in diesem Fall die Bestimmung der unbekannten Parameter a und b.
Die Vorgehensweise bei der linearen Regression veranschaulicht folgende Grafik. Gegeben sind Wertepaare x_i, y_i, als schwarze Punkte eingezeichnet. Grün sind die Werte $(\hat{x}, \hat{y})$ die durch die lineare Regressionsfunktion errechnet werden. Die roten Linien symbolisieren die Abweichungen[1] $e_i = y_i - \hat{y}_i$ dieser durch die Gleichung bestimmten Punkte von den wahren Punkten. Aufgabe bei der Bestimmung der Parameter ist es nun, a und b so zu wählen, dass die Summe QS der quadrierten Abweichungen – also $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – minimal wird.

[1] Es ist egal, ob man $y_i - \hat{y}_i$ oder $\hat{y}_i - y_i$ schreibt, durch die Quadrierung heben sich eventuelle negative Vorzeichen auf.



2 Herleitung der Gleichungen

$$QS(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

$$= \sum_{i=1}^{n} \Big(y_i - [ax_i + b]\Big)^2 \tag{3}$$

Da wir die optimalen Werte für die Minimierung dieser Quadratsumme erhalten wollen, bilden wir die partiellen Ableitungen nach a und b. Vorher können wir jedoch Gleichung 2 vereinfachen. Mit Hilfe der 2. Binomischen Formel[2] lösen wir 3 auf:

$$QS(a,b) = \sum_{i=1}^{n} \Big(y_i^2 - 2y_i(ax_i + b) + (ax_i + b)^2\Big) \tag{4}$$

Da der Term $(ax_i + b)^2$ der 1. Binomischen Formel[3] entspricht, lösen wir auch diesen auf und vereinfachen:

$$QS(a,b) = \sum_{i=1}^{n} \Big(y_i^2 - 2ax_iy_i - 2by_i + a^2x_i^2 + 2abx_i + b^2\Big) \tag{5}$$

Ausgehend von Gleichung 5 bilden wir jetzt die partiellen Ableitungen nach a und b:

[2] 2. Binomische Formel: $(s-t)^2 = s^2 - 2st + t^2$
[3] 1. Binomische Formel: $(s+t)^2 = s^2 + 2st + t^2$

$$\frac{\partial QS(a,b)}{\partial a} = \sum_{i=1}^{n}(-2x_i y_i + 2ax_i^2 + 2bx_i) \tag{6}$$

$$= 2\sum_{i=1}^{n} x_i(-y_i + ax_i + b) \tag{7}$$

$$\frac{\partial QS(a,b)}{\partial b} = \sum_{i=1}^{n}(-2y_i + 2ax_i + 2b) \tag{8}$$

$$= 2\sum_{i=1}^{n}(ax_i + b - y_i) \tag{9}$$

Wenn wir Gleichung 9 nullsetzen und auflösen, erhalten wir

$$2\sum_{i=1}^{n} ax_i + 2\sum_{i=1}^{n} b \quad - 2\sum_{i=1}^{n} y_i \quad = 0 \tag{10}$$

$$2\sum_{i=1}^{n} ax_i + \quad 2nb \quad - 2\sum_{i=1}^{n} y_i \quad = 0 \tag{11}$$

$$2nb = 2\sum_{i=1}^{n} y_i - 2\sum_{i=1}^{n} ax_i \tag{12}$$

Auflösen nach b (durch $2n$ teilen) gibt (zusammen mit der Tatsache, dass das arithmetische Mittel allgemein als $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ definiert ist):

$$b = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} ax_i}{n} \tag{13}$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i - a\frac{1}{n}\sum_{i=1}^{n} x_i \tag{14}$$

$$= \bar{y} - a\bar{x} \tag{15}$$

Setzen wir nun $b = \bar{y} - a\bar{x}$ in Gleichung 7 ein, erhalten wir

$$2\sum_{i=1}^{n} x_i\big(ax_i + (\bar{y} - a\bar{x}) - y_i\big) = 0 \tag{16}$$

Durch Ausmultiplizieren und Vereinfachen ergibt sich:

$$0 = \sum_{i=1}^{n} x_i \big(a x_i + (\bar{y} - a\bar{x}) - y_i\big) \tag{17}$$

$$= \sum_{i=1}^{n} \big(a x_i^2 + x_i(\bar{y} - a\bar{x}) - x_i y_i\big) \tag{18}$$

$$= \sum_{i=1}^{n} \big(a x_i^2 + x_i\bar{y} - a\bar{x}x_i - x_i y_i\big) \tag{19}$$

$$= \sum_{i=1}^{n} \big(a x_i^2 - a\bar{x}x_i + x_i\bar{y} - x_i y_i\big) \tag{20}$$

$$= \sum_{i=1}^{n} \big((a x_i^2 - a\bar{x}x_i) + x_i\bar{y} - x_i y_i\big) \tag{21}$$

$$= \sum_{i=1}^{n} \big(a x_i^2 - a\bar{x}x_i\big) + \sum_{i=1}^{n} x_i\bar{y} - \sum_{i=1}^{n} x_i y_i \tag{22}$$

Jetzt subtrahiert man $\sum_{i=1}^{n} x_i y_i$ und addiert $\sum_{i=1}^{n} x_i\bar{y}$, um diese beiden Teile auf die andere Seite der Gleichung zu bekommen.

$$\sum_{i=1}^{n} (a x_i^2 - a x_i\bar{x}) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\bar{y} \tag{23}$$

Da a konstant ist, können wir es vor die Klammer ziehen.

$$a \sum_{i=1}^{n} (x_i^2 - x_i\bar{x}) = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i \tag{24}$$

Jetzt teilen wir durch $\sum_{i=1}^{n} (x_i^2 - x_i\bar{x})$

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} (x_i^2 - x_i\bar{x})} \tag{25}$$

Aus der Definition des arithmetischen Mittels $\bar{x} = \frac{1}{n} \sum x_i$ folgt $\sum_{i=1}^{n} x_i = n\bar{x}$. Einsetzen ergibt

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} n \bar{x}}{\sum_{i=1}^{n} \big(x_i^2 - \sum_{i=1}^{n} x_i\bar{x}\big)} \tag{26}$$

Jetzt zerlegen wir die Summe unter dem Bruchstrich in Einzelsummen und ziehen $\bar{x}$ vor das zweite Summenzeichen (Zur Erinnerung: konstanter Term!)

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} n \bar{x}}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} \tag{27}$$

Über alternative Formeln zu Varianz und Kovarianz[4] erhalten wir

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{n\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}\right)}{n\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)} = \frac{n\,\mathrm{Cov}(x,y)}{n\,\mathrm{Var}(x)} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \tag{28}$$

3 Beispiel

Für unser Beispiel vom Anfang hier die numerische Bestimmung der Parameter. Für $\bar{x}$ erhalten wir 3, für $\bar{y} = 2.4$, die Summe der $(x-\bar{x})(y-\bar{y})$ ergibt 3, die Summe der $(x-\bar{x})^2 = 10$. Durch Einsetzen dieser Werte erhalten wir dann als Parameterwert für b 1.5, als Parameterwert für a 0.3, sodass die Formel unseres linearen Modells

$$y = 0.3 \cdot x + 1.5$$

lautet.

	1	2	3	4	5	6
	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	1	1	-2	-1.4	2.8	4
2	2	3	-1	0.6	-0.6	1
3	3	2	0	-0.4	0.0	0
4	4	4	1	1.6	1.6	1
5	5	2	2	-0.4	-0.8	4
$\sum$	15	12				

4 Quelldateien

Dieses Dokument wurde mit LaTeX, dem freien Textsatzsystem, erstellt.

LaTeX 

Metapost 

Metapost (kompiliert) 

[4] Verschiebungssatz:
$\mathrm{Cov}(X,Y) = \mathrm{E}\left((X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right) = \mathrm{E}(X,Y) - \mathrm{E}(X)\,\mathrm{E}(Y)$
$\mathrm{Var}(X) = \mathrm{E}\left((X-\mathrm{E}(X))^2\right) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$