Herleitung der Parameter-Gleichungen für die einfache lineare Regression

– DRAFT –

Uwe Ziegenhagen

30. Juni 2018

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.

v3.0a auf Github gewechselt, Metapost gegen TikZ getauscht, einige Erläuterungen verbessert, LaTeX Code aufgeräumt

1 Einführung

Aus der Wikipedia[1]:

> „Die lineare Regression, die einen Spezialfall des allgemeinen Konzepts der Regressionsanalyse darstellt, ist ein statistisches Verfahren, mit dem versucht wird, eine beobachtete abhängige Variable durch eine oder mehrere unabhängige Variablen zu erklären. Das Beiwort ‚linear' ergibt sich dadurch, dass die abhängige Variable eine Linearkombination der Regressionskoeffizienten darstellt (aber nicht notwendigerweise der unabhängigen Variablen). Der Begriff Regression bzw. Regression zur Mitte wurde vor allem durch den Statistiker Francis Galton geprägt. "

[1] https://de.wikipedia.org/wiki/Lineare_Regression, Abruf: 24.06.2018

Allgemein wird eine metrische Variable Y betrachtet, die von ein oder mehreren Variablen X_i abhängt. Y nennt man die „abhängige Variable“, die X_i sind die „unabhängigen Variablen“. Im eindimensionalen Fall – wenn es nur eine X-Variable gibt – spricht man von einer einfachen linearen Regressionsanalyse, in höheren Dimensionen von der multiplen Regressionsanalyse.

2 Einfache linerare Regression

Im folgenden nutzen wir die Werte aus Tabelle 1, um die einfache lineare Regression zu erklären.

X-Wert	Y-Wert
1	1
2	3
3	2
4	5
5	4

Tabelle 1: Tabelle mit Wertepaaren

Stellt man die Punkte in einem Streu-Diagramm wie in Abbildung 1 dar, so erkennt man dass mit steigendem Wert von X die Werte von Y ebenfalls steigen.

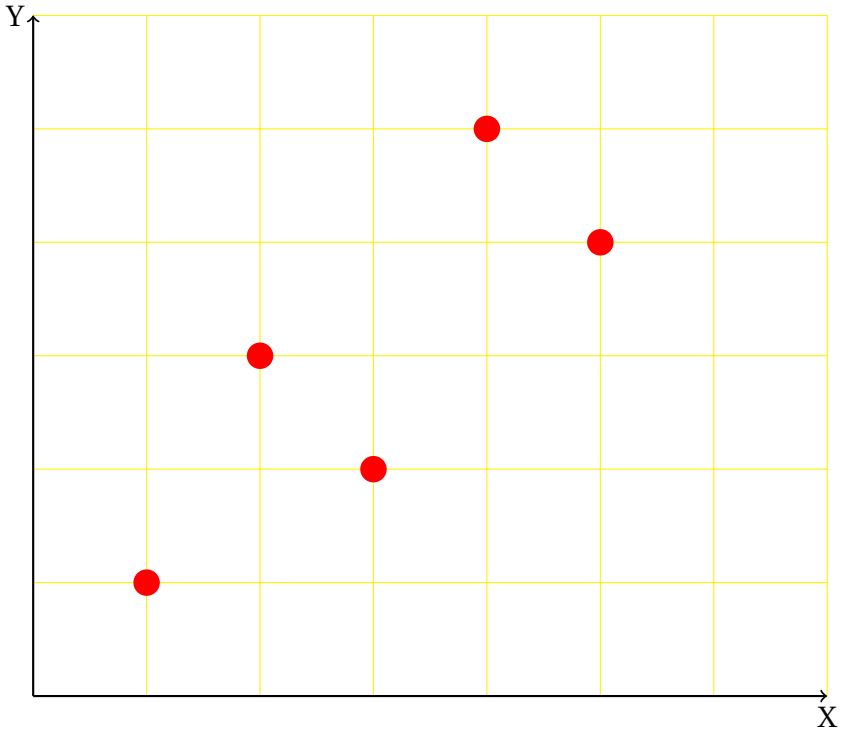

Abbildung 1: Scatterplot zur Darstellung der X-Y Wertepaare

Wenn wir den Zusammenhang dieser Punkte mittels Gerade modellieren wollen, unterstellen wir ein Modell der Form:

$$Y_i = b + a \cdot x_i + \epsilon_i \tag{1}$$

b ist dabei der Achsenabschnitt, also der Punkt $(0, b)$, an dem die X-Achse geschnitten wird. a hingegen ist der Parameter für die Steigung der Regressionsgeraden. a und b sind für unsere fünf Wertepaare zu bestimmen. (ϵ_i steht für die Fehler, den wir bei der Modellierung machen, darauf kommen wir gleich noch zu sprechen).

Wir können wir nun die Regressionsgerade durch die Punkte zeichnen? Abbildung 2 zeigt zwei Beispiele für eher zufällige Regressionsgeraden. Im linken Plot erkennt man sehr deutlich, dass die Gerade nicht zu unseren Punkten passen kann, sie zeigt in die falsche Richtung und unterstellt damit, dass mit steigendem X die Werte für Y sinken. Im rechten Plot stimmt die Richtung, die Gerade sieht schon „recht gut“ aus.

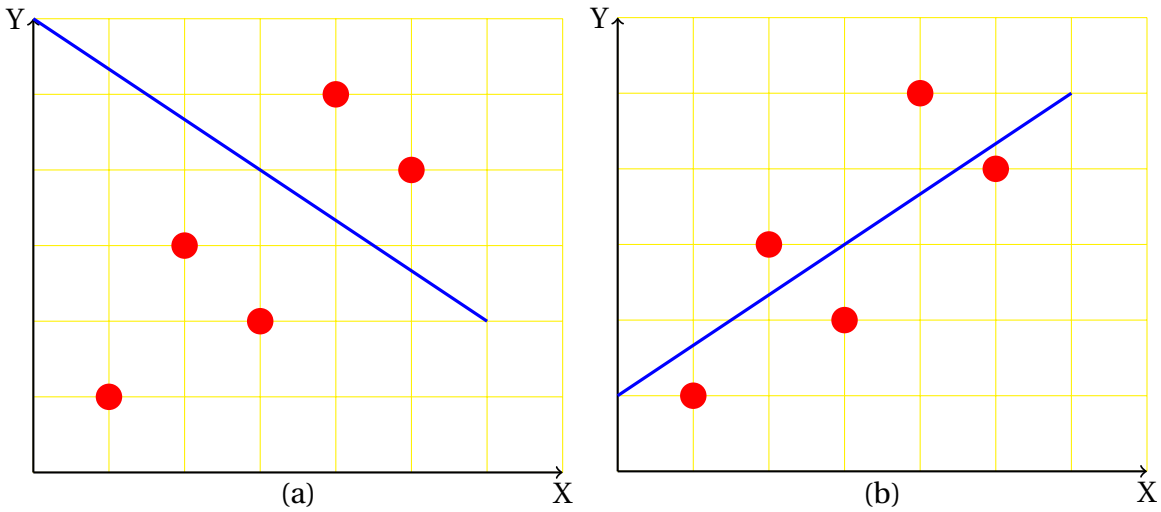

Abbildung 2: Zwei Regressionsgeraden

Da aber eine Einschätzung wie „recht gut“ nicht wirklich mathematisch exakt ist, müssen wir diesen Punkt ein wenig genauer betrachten.

Betrachten wir dazu Abbildung 3. Hier wurden auf der blauen Geraden die Punkte in grün markiert, die die Regressionsgleichung für den jeweiligen Wert von X vorhersagt, außerdem wurden die jeweiligen Abstände zwischen dem wahren Y-Wert und dem geschätzten Y-Wert (den wir ab jetzt $\hat{Y}$ nennen) markiert.

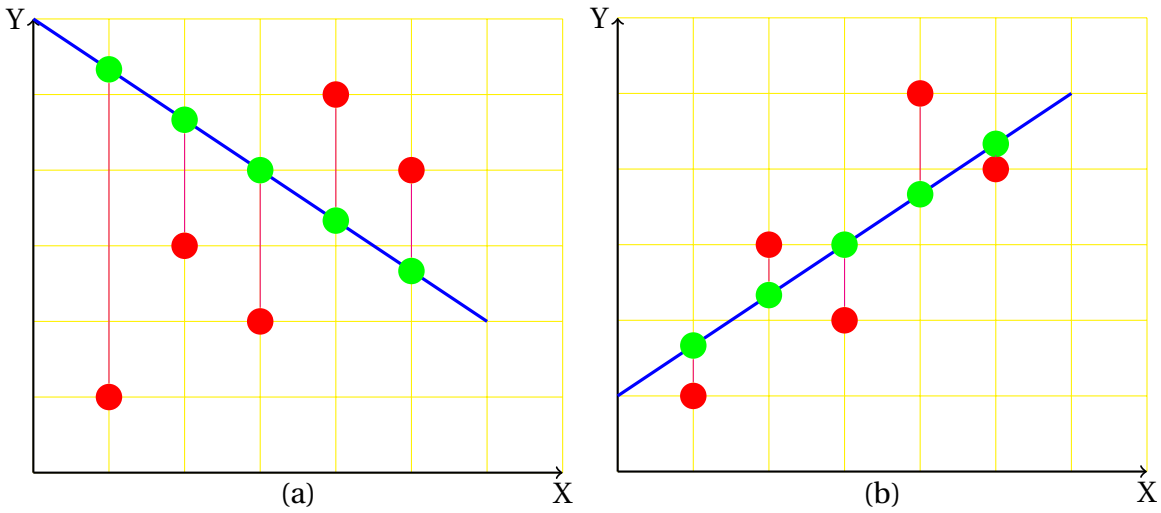

Abbildung 3: Zwei Regressionsgeraden

Wenn wir beide Grafiken betrachten, so ist schnell sichtbar, dass die Regressionsgerade im linken Bild deutlich schlechter ist als die Regressionsgerade im rechten Bild: die Summe der Abstände zwischen den wahren, roten, Punkten und den durch die Gerade geschätzten Punkten ist viel größer.

Aus dieser Tatsache lässt sich ein sehr wichtiger Schluss ziehen: wenn wir diese

Summe der Abstände minimieren könnten, würden wir die optimale Gerade erhalten. Als Gleichung:

$$S = \sum_{i=1}^{n} y_i - \hat{y}_i \tag{2}$$

In Worten: S ist die Summe der Differenzen von wahrem und geschätzten y-Wert.

Es hat sich jedoch herausgestellt, dass es sinnvoll ist, nicht die einfache Summe der Abstände zu minimieren, sondern die Summe der quadrierten Abstände.

Daher wird aus Gleichung 3 die folgende Gleichung:

$$QS = \left(\sum_{i=1}^{n} y_i - \hat{y}_i \right)^2 \tag{3}$$

Diese Quadratsumme der Abweichungen ist nur abhängig von den Parametern a und b der Regressionsgleichung, daher können wir schreiben:

$$QS(a,b) = \left(\sum_{i=1}^{n} y_i - \hat{y}_i \right)^2 \tag{4}$$

Im Folgenden werden wir diese Funktion differenzieren, um die Gleichungen für die optimalen a und b zu ermitteln.

3 Herleitung der Parameter-Gleichungen

$$\mathrm{QS}(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{5}$$

$$= \sum_{i=1}^{n} \Big(y_i - [a x_i + b] \Big)^2 \tag{6}$$

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 5 vereinfachen. Mit Hilfe der 2. Binomischen Formel[2] lösen wir

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

Gleichung 6 auf:

$$\mathrm{QS}(a,b) = \sum_{i=1}^{n} \left(y_i^2 - 2y_i(ax_i + b) + (ax_i + b)^2 \right) \tag{7}$$

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

$$\mathrm{QS}(a,b) = \sum_{i=1}^{n} \left(y_i^2 - 2ax_iy_i - 2by_i + a^2x_i^2 + 2abx_i + b^2 \right) \tag{8}$$

Ausgehend von Gleichung 8 bilden wir jetzt die partiellen Ableitungen nach a und b:

$$\frac{\partial \mathrm{QS}(a,b)}{\partial a} = \sum_{i=1}^{n} (-2x_iy_i + 2ax_i^2 + 2bx_i) \tag{9}$$

$$= 2\sum_{i=1}^{n} x_i(-y_i + ax_i + b) \tag{10}$$

$$\frac{\partial \mathrm{QS}(a,b)}{\partial b} = \sum_{i=1}^{n} (-2y_i + 2ax_i + 2b) \tag{11}$$

$$= 2\sum_{i=1}^{n} (ax_i + b - y_i) \tag{12}$$

Wenn wir Gleichung 12 nullsetzen und auflösen, erhalten wir

$$2\sum_{i=1}^{n} ax_i + 2\sum_{i=1}^{n} b \; - 2\sum_{i=1}^{n} y_i \quad = 0 \tag{13}$$

$$2\sum_{i=1}^{n} ax_i + \; 2nb \; - 2\sum_{i=1}^{n} y_i \quad = 0 \tag{14}$$

$$2nb = 2\sum_{i=1}^{n} y_i - 2\sum_{i=1}^{n} ax_i \tag{15}$$

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):

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

$$b = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} a x_i}{n} \tag{16}$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i - a\frac{1}{n}\sum_{i=1}^{n} x_i \tag{17}$$

$$= \bar{y} - a\bar{x} \tag{18}$$

Setzen wir nun $b = \bar{y} - a\bar{x}$ in Gleichung 10 ein, erhalten wir

$$2\sum_{i=1}^{n} x_i\big(ax_i + (\bar{y} - a\bar{x}) - y_i\big) = 0 \tag{19}$$

Durch Ausmultiplizieren und Vereinfachen ergibt sich:

$$0 = \sum_{i=1}^{n} x_i\big(ax_i + (\bar{y} - a\bar{x}) - y_i\big) \tag{20}$$

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

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

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

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

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

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.

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

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

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

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

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n}(x_i^2 - x_i\bar{x})} \tag{28}$$

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

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}n\bar{x}}{\sum_{i=1}^{n}\left(x_i^2 - \sum_{i=1}^{n} x_i\bar{x}\right)} \tag{29}$$

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

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}n\bar{x}}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} \tag{30}$$

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

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{n\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}\right)}{n\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)} = \frac{n\,\mathrm{Cov}(x,y)}{n\,\mathrm{Var}(x)} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \tag{31}$$

4 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

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

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

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				

5 Quelldateien

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

LaTeX 
Metapost 
Metapost (kompiliert)