

# Herleitung der Parameter-Gleichungen für die einfache lineare Regression

– DRAFT –

Uwe Ziegenhagen

24. 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<sup>1</sup>:

„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.“

---

<sup>1</sup>[https://de.wikipedia.org/wiki/Lineare\\_Regression](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 lineare 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.

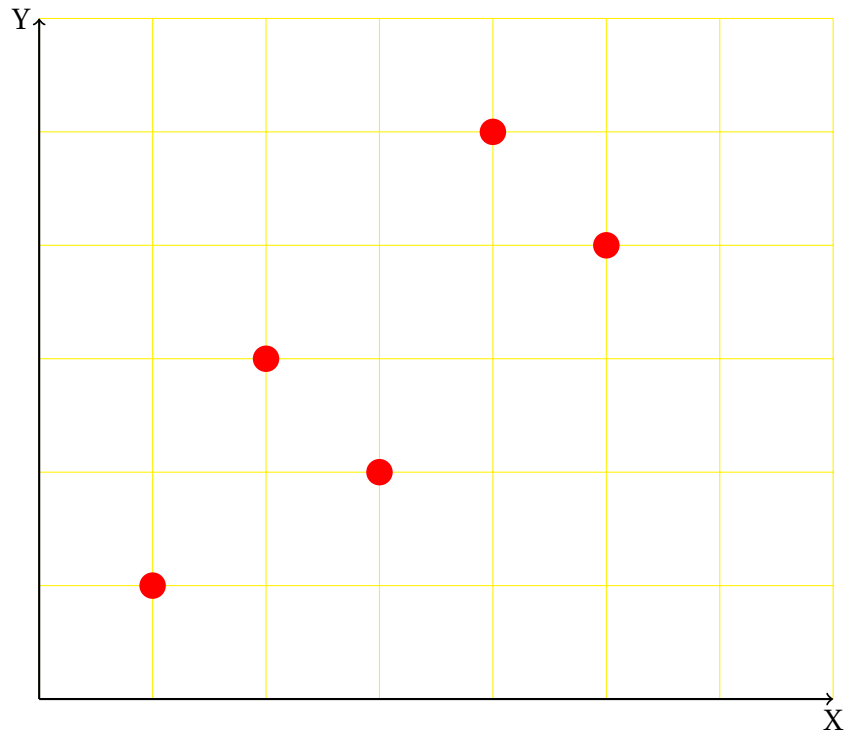


Abbildung 1: Scatterplot zur Darstellung der X-Y Wertepaare

Wenn wir jetzt den Zusammenhang zwischen  $X$  und  $Y$  mittels einfacher linearer Regression modellieren möchten, unterstellen wir ein Modell der Form:

$$Y_i = b + ax_i + \epsilon_i \quad (1)$$

$b$  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. (Die  $\epsilon_i$  stehen für die Schätzfehler, 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 dagegen die Richtung, die Gerade sieht schon recht gut aus.

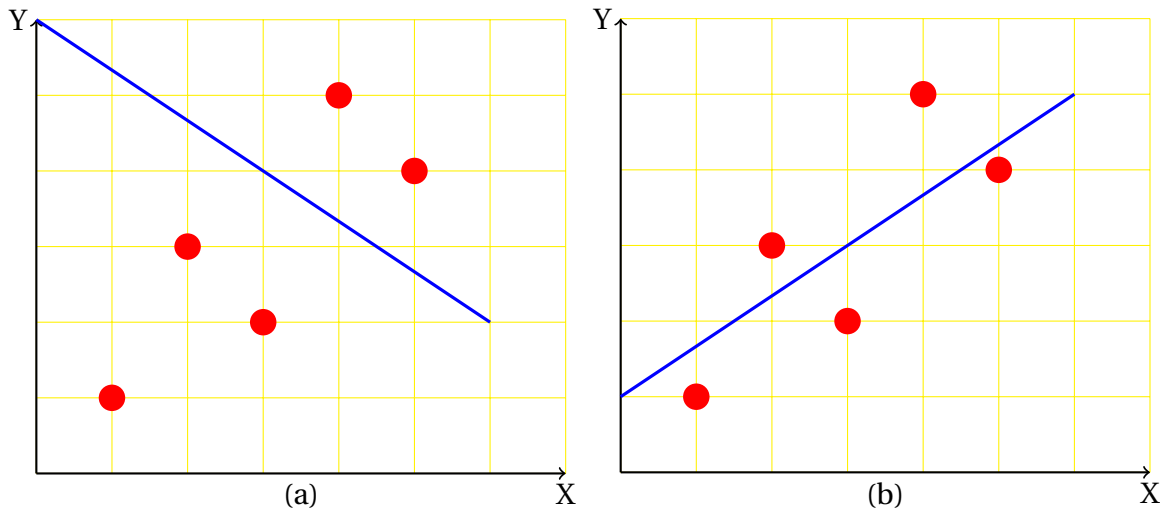


Abbildung 2: Zwei Regressionsgeraden

Da aber eine Einschätzung wie „recht gut“ nicht wirklich mathematisch exakt ist, müssen wir hier ein wenig tiefer einsteigen.

Betrachten wir dazu Abbildung 3. Hier wurden auf der blauen Geraden die Punkte in grün markiert, die die jeweilige Regressionsgleichung für den jeweiligen Wert von  $X$  vorhersagt, außerdem wurden die jeweiligen Abstände zwischen dem wahren  $Y$ -Wert und dem geschätzten  $Y$ -Wert markiert.

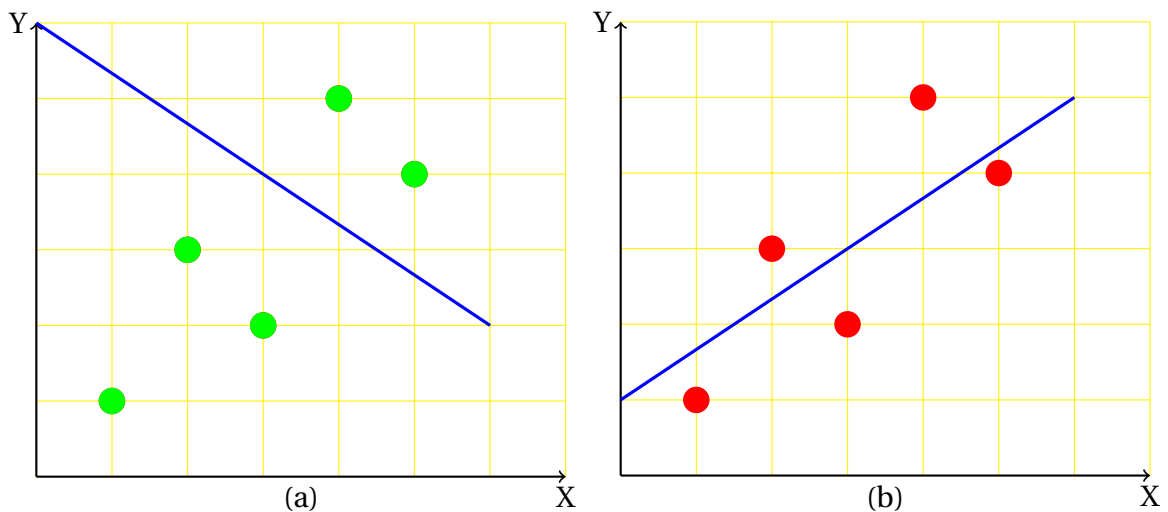


Abbildung 3: Zwei Regressionsgeraden

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<sup>2</sup>  $e_i = y_i - \hat{y}_i$  dieser durch die Gleichung

<sup>2</sup>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.

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.

### 3 Herleitung der Gleichungen

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

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

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<sup>3</sup> lösen wir 3 auf:

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

Da der Term  $(ax_i + b)^2$  der 1. Binomischen Formel<sup>4</sup> entspricht, lösen wir auch diesen auf und vereinfachen:

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

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

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

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

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

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

<sup>3</sup> 2. Binomische Formel:  $(s - t)^2 = s^2 - 2st + t^2$

<sup>4</sup> 1. Binomische Formel:  $(s + t)^2 = s^2 + 2st + t^2$

Wenn wir Gleichung 9 nullsetzen und auflösen, erhalten wir

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

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

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

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):

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

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

$$= \bar{y} - a\bar{x} \quad (15)$$

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

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

Durch Ausmultiplizieren und Vereinfachen ergibt sich:

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

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

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

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

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

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

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.

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

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

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

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

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

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

zen ergibt

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

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

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

Über alternative Formeln zu Varianz und Kovarianz<sup>5</sup> erhalten wir

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

## 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

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

lautet.

---

<sup>5</sup>Verschiebungssatz:

$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(X, Y) - E(X)E(Y)$



$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2$



	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
$\Sigma$	15	12				

## 5 Quelldateien

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

$\LaTeX$    
 Metapost   
 Metapost (kompiliert) 