Einführung in die lineare Regression
– DRAFT VERSION–

Uwe Ziegenhagen

7. Juli 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.

v3.0 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. "

Allgemein wird eine metrische Variable Y betrachtet, die von ein oder mehreren Variablen X_i abhängt. Y nennt man daher auch die „abhängige Variable" und die X_i die „unabhängigen Variablen". Im eindimensionalen Fall – wenn es nur

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

eine X-Variable gibt – spricht man von einer einfachen linearen Regression, in höheren Dimensionen von der multiplen Regression.

2 Einfache lineare Regression

Im folgenden nutzen wir die Werte aus Tabelle 1, um an ihnen 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 (auf englisch „Scatterplot“) wie in Abbildung 1 dar, so erkennt man dass mit steigendem Wert von X die Werte von Y ebenfalls steigen.

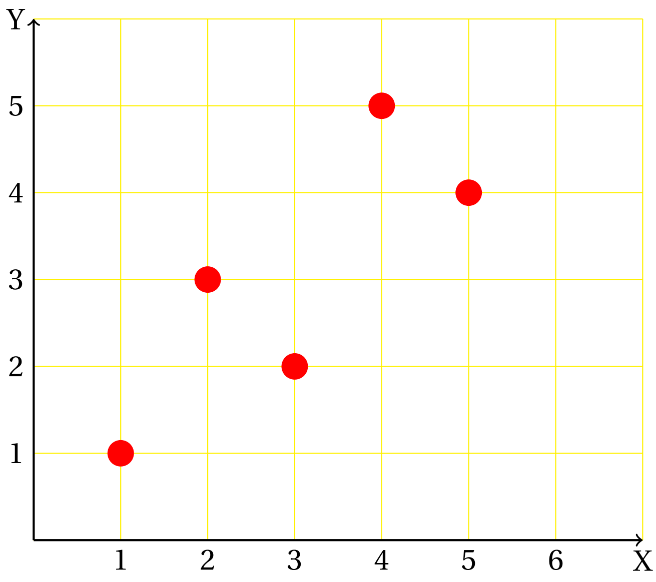

Abbildung 1: Scatterplot zur Darstellung der X-Y Wertepaare

Wenn wir den Zusammenhang dieser Punkte mittels Gerade (also „linear") 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 später noch zu sprechen).

Wir können wir nun die Regressionsgerade durch die Punkte zeichnen? Abbildung 2 zeigt zwei Beispiele für beliebig gewählte Regressionsgeraden. Im linken Plot erkennt man sehr deutlich, dass die Gerade nicht zu unseren Punkten passt, 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.

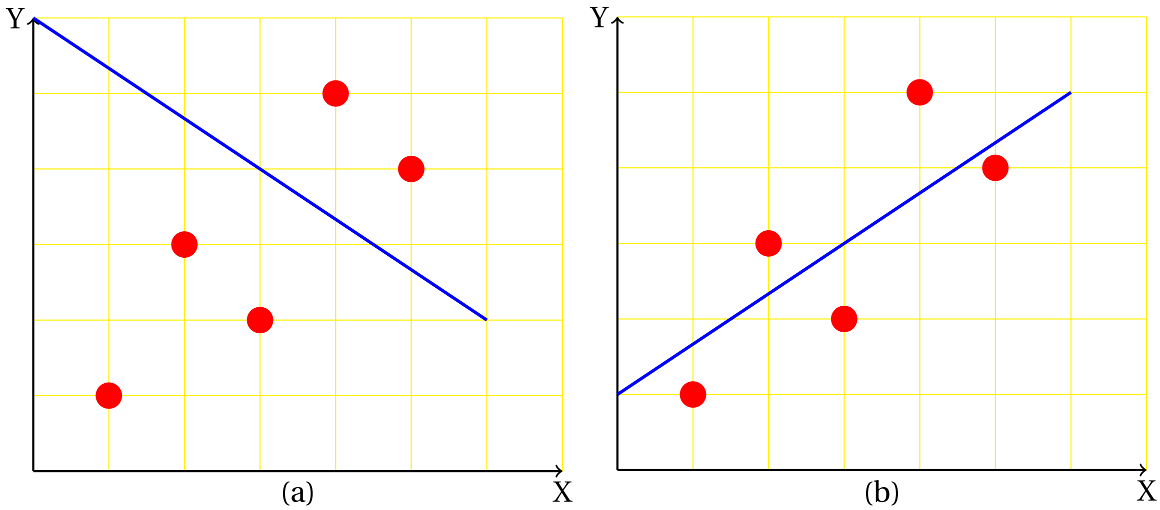

Abbildung 2: Zwei Regressionsgeraden

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.

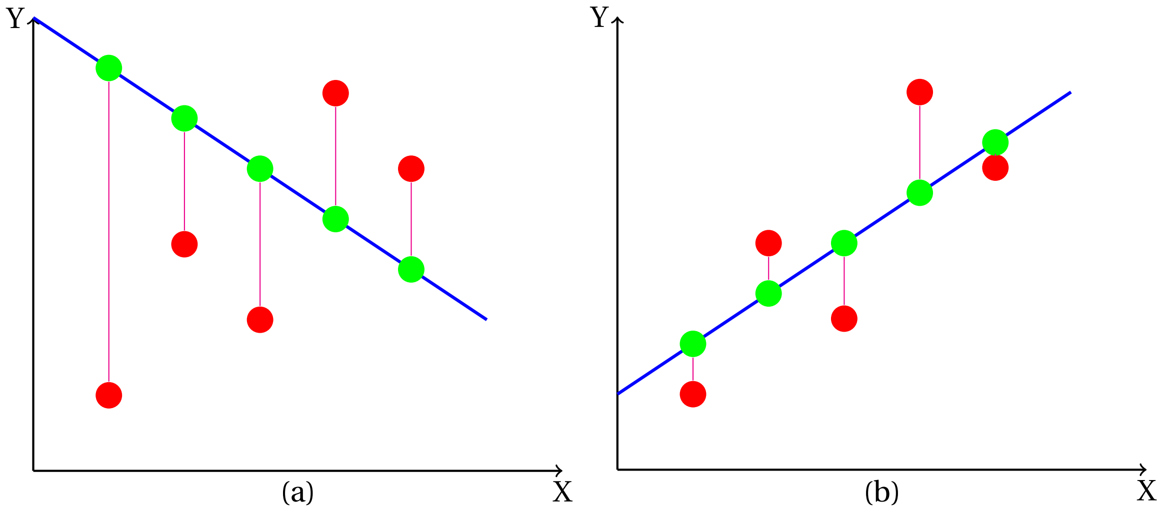

Abbildung 3: Zwei Regressionsgeraden

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 aller Differenzen von wahrem und geschätzten y-Wert.

Es hat sich als mathematisch sinnvoll herausgestellt, nicht einfach die Summe der Abstände zu minimieren, sondern die Summe der *quadrierten* Abstände. Es lässt sich nicht nur leicht damit rechnen, der Kleinste-Quadrate-Schätzer ist auch – sofern die Annahmen des klassischen linearen Regressionsmodells nicht verletzt sind – BLUE („Best Linear Unbiased Estimator"). Dazu später mehr…

Aus Gleichung 3 wird jetzt – da wir ja die Quadratsumme minimieren wollen – 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 partiell ableiten, um die Gleichungen für die optimalen a und b zu ermitteln.

3 Herleitung der Parameter-Gleichungen

Wir schreiben Gleichung 3 nochmals auf und ersetzen $\hat{y}$ durch die Modellgleichung:

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

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

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 Gleichung 6 auf:

$$\mathrm{QS}(a,b) = \sum_{i=1}^{n} \Bigg(\overbrace{y_i^2}^{s^2} - \overbrace{2y_i(ax_i+b)}^{-2st} + \overbrace{(ax_i+b)^2}^{t^2} \Bigg) \tag{7}$$

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

$$\mathrm{QS}(a,b) = \sum_{i=1}^{n} \Big(y_i^2 - 2ax_iy_i - 2by_i + \overbrace{a^2x_i^2}^{s^2} + \overbrace{2abx_i}^{2st} + \overbrace{b^2}^{t^2} \Big) \tag{8}$$

Ausgehend von Gleichung 8 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: $(s+t)^2 = s^2 + 2st + t^2$

$$\begin{aligned}
\frac{\partial \mathrm{QS}(a,b)}{\partial a} &= \sum_{i=1}^{n} (-2x_i y_i + 2ax_i^2 + 2bx_i) && (9)\\
&= 2\sum_{i=1}^{n} x_i(-y_i + ax_i + b) && (10)\\
&= 2\sum_{i=1}^{n} x_i(ax_i + b - y_i) && (11)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathrm{QS}(a,b)}{\partial b} &= \sum_{i=1}^{n} (-2y_i + 2ax_i + 2b) && (12)\\
&= 2\sum_{i=1}^{n} (ax_i + b - y_i) && (13)
\end{aligned}$$

Wenn wir Gleichung 13 nullsetzen und auflösen, erhalten wir

$$\begin{aligned}
2\sum_{i=1}^{n} ax_i + 2\sum_{i=1}^{n} b - 2\sum_{i=1}^{n} y_i &= 0 && (14)\\
2\sum_{i=1}^{n} ax_i + 2nb - 2\sum_{i=1}^{n} y_i &= 0 && (15)\\
2nb &= 2\sum_{i=1}^{n} y_i - 2\sum_{i=1}^{n} ax_i && (16)
\end{aligned}$$

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):

$$\begin{aligned}
b &= \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} ax_i}{n} && (17)\\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - a\frac{1}{n}\sum_{i=1}^{n} x_i && (18)\\
&= \bar{y} - a\bar{x} && (19)
\end{aligned}$$

Setzen wir nun $b = \bar{y} - a\bar{x}$ in Gleichung 11 ein, erhalten wir

$$2\sum_{i=1}^{n} x_i\big(ax_i + (\bar{y} - a\bar{x}) - y_i\big) = 0 \qquad (20)$$

Durch Ausmultiplizieren und Vereinfachen ergibt sich:

$$0 = \sum_{i=1}^{n} x_i \big(a x_i + (\bar{y} - a\bar{x}) - y_i \big) \tag{21}$$
$$= \sum_{i=1}^{n} \big(a x_i^2 + x_i (\bar{y} - a\bar{x}) - x_i y_i \big) \tag{22}$$
$$= \sum_{i=1}^{n} \big(a x_i^2 + x_i \bar{y} - a\bar{x} x_i - x_i y_i \big) \tag{23}$$
$$= \sum_{i=1}^{n} \big(a x_i^2 - a\bar{x} x_i + x_i \bar{y} - x_i y_i \big) \tag{24}$$

$$= \sum_{i=1}^{n} \big((a x_i^2 - a\bar{x} x_i) + x_i \bar{y} - x_i y_i \big) \tag{25}$$
$$= \sum_{i=1}^{n} \big(a x_i^2 - a\bar{x} x_i \big) + \sum_{i=1}^{n} x_i \bar{y} - \sum_{i=1}^{n} x_i y_i \tag{26}$$

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

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

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

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

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

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

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

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

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

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} n \bar{x}}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} \tag{31}$$

Über Formeln zu Varianz und Kovarianz[4] erhalten wir

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2} = \frac{n \left(\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x} \bar{y} \right)}{n \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right)} = \frac{n \operatorname{Cov}(x, y)}{n \operatorname{Var}(x)} = \frac{\operatorname{Cov}(x, y)}{\operatorname{Var}(x)} \tag{32}$$

Damit haben wir die beiden Gleichungen hergeleitet, um die Regressionsgerade zu bestimmen:

$$b = \bar{y} - a \bar{x} \tag{33}$$

$$a = \frac{\operatorname{Cov}(x, y)}{\operatorname{Var}(x)} \tag{34}$$

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

4 Beispiel

Tabelle 2: Hilfstabelle

	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	-2	4	4
2	2	3	-1	0	0	1
3	3	2	0	-1	0	0
4	4	5	1	2	2	1
5	5	4	2	1	2	4
$\sum$	15	15			8	10

Mit Hilfe der Werte aus der Tabelle lassen sich a und b einfach bestimmen. Hinweis: $\bar{x} = 15/5 = 3$, $\bar{y} = 15/5 = 3$

$$a = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{5}}{\frac{\sum_{i=1}^{5}(x_i-\bar{x})^2}{5}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{8}{10} = 0.8$$

Hinweis zu dieser Rechnung: Brüche werden dividiert, indem man mit dem Kehrwert multipliziert: $\frac{1/5}{1/5} = 1/5 \cdot 5/1 = 1$. Für die Berechnung von a braucht man die Anzahl der Beobachtungen also nicht mehr.

$$b = \bar{y} - a \cdot \bar{x} = 3 - 0.8 \cdot 3 = 3 - 2.4 = 0.6$$

Mit den gefundenen Werten für unsere beiden Parameter können wir jetzt die Regressionsgerade zeichnen:

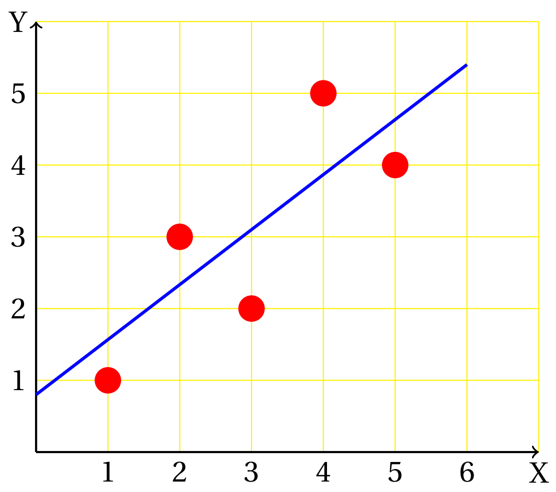

Abbildung 4: Scatterplot mit Regressionsgerade

5 Berechnung mit dem Taschenrechner

Moderne (Schul-)Taschenrechner haben alle entsprechende Funktionen eingebaut, um anhand von übergebenen Wertepaaren die Parameter a und b schnell zu bestimmen. Im Folgenden zeigen wir anhand eines Casio Schultaschenrechners vom Typ , wie es funktioniert. TODO!

6 Güte der Linearen Regression

In diesem Abschnitt möchten wir erläutern, wie man die Güte des linearen Modells bestimmt. Unsere Regressionsparameter a und b sind zwar die optimalen Modellparameter für die von uns genutzten Daten, aber

Quelldateien

Dieses Dokument wurde mit LaTeX, dem freien Textsatzsystem, erstellt. Die Quelldatei dieses Dokuments ist im PDF enthalten, klicken Sie einfach auf das Symbol. Sofern Ihr PDF-Betrachter Attachments unterstützt, sollten Sie auf die Quelldatei zugreifen können.

LaTeX