

Einführung in die lineare Regression

– DRAFT VERSION –

Uwe Ziegenhagen

3. März 2019

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
- v3.1 Erweiterung um multiple Regression, Quellcodes, Tests, etc. (DRAFT)

1 Einführung

Aus der Wikipedia¹:

„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

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

die X_i die „unabhängigen Variablen“. Im eindimensionalen Fall – wenn es nur 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.

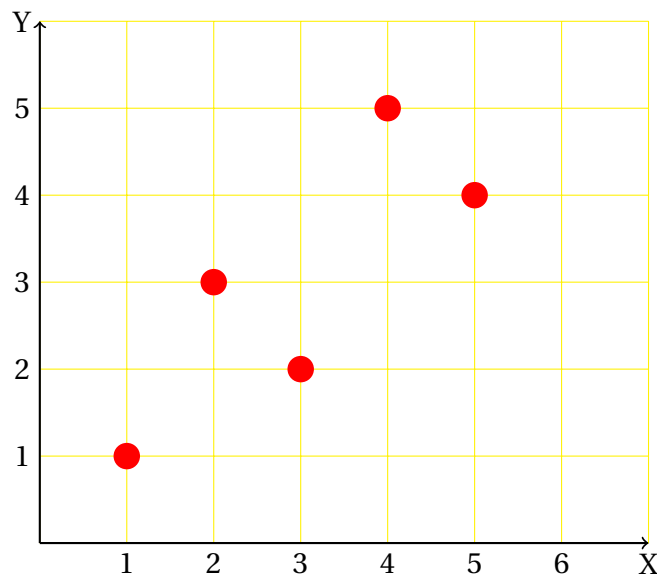


Abbildung 1: Scatterplot zur Darstellung der X-Y Wertepaare

Wenn wir den Zusammenhang dieser Punkte mittels Gerade (also „linear“) modellieren wollen, unterstellen wir ein Modell der Form:

$$Y_i = b + a \cdot x_i + \epsilon_i \quad (1)$$

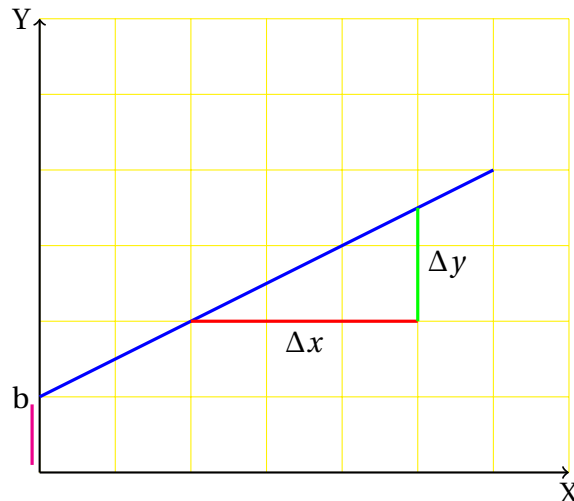


Abbildung 2: Grafische Erläuterung

b ist dabei der Achsenabschnitt, also der Punkt $(0, b)$, an dem die Y-Achse geschnitten wird. a hingegen ist der Parameter für die Steigung der Regressionsgeraden, also das Verhältnis von Δy und Δx . 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.

Abbildung 2 beschreibt diesen Zusammenhang grafisch: der Achsenabschnitt ist der Schnittpunkt der Geraden mit der Y-Achse, der Anstieg das Verhältnis von Δy zu Δx .

Wir können wir nun die Regressionsgerade durch die Punkte zeichnen? Abbildung 3 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.

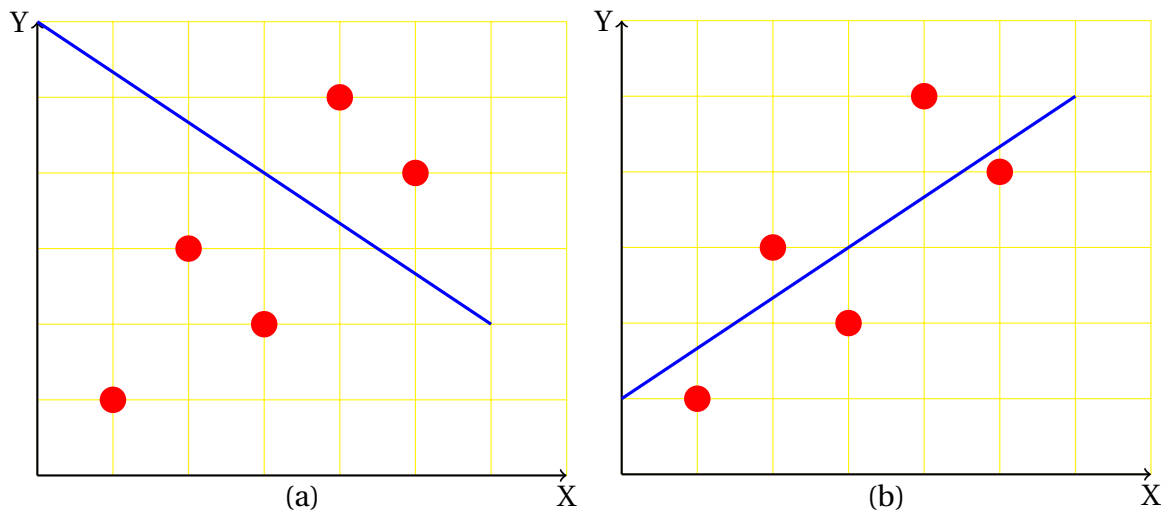


Abbildung 3: Zwei Regressionsgeraden

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

Betrachten wir dazu Abbildung 4. 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.

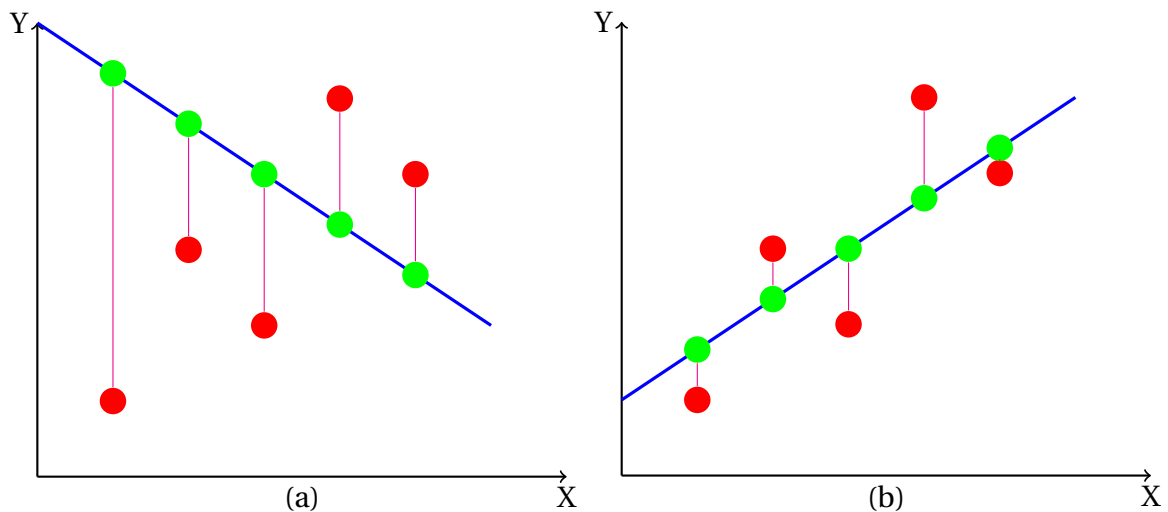


Abbildung 4: Zwei Regressionsgeraden

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. In Gleichungsform können wir schreiben:

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

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 2 wird jetzt – da wir ja die Quadratsumme minimieren wollen – die folgende Gleichung:

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

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

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

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:

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

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

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 Formel² lösen wir Gleichung 6 auf:

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

Da der Term $(ax_i + b)^2$ der 1. Binomischen Formel³ entspricht, lösen wir auch diesen auf und vereinfachen:

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

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

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

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

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

$$\frac{\partial QS(a, b)}{\partial b} = \sum_{i=1}^n (-2y_i + 2ax_i + 2b) \quad (12)$$

$$= 2 \sum_{i=1}^n (ax_i + b - y_i) \quad (13)$$

Wenn wir Gleichung 13 nullsetzen und auflösen, erhalten wir

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

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

$$2 \sum_{i=1}^n ax_i + 2 \sum_{i=1}^n b - 2 \sum_{i=1}^n y_i = 0 \quad (14)$$

$$2 \sum_{i=1}^n ax_i + 2nb - 2 \sum_{i=1}^n y_i = 0 \quad (15)$$

$$2nb = 2 \sum_{i=1}^n y_i - 2 \sum_{i=1}^n ax_i \quad (16)$$

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):

$$b = \frac{\sum_{i=1}^n y_i}{n} - \frac{\sum_{i=1}^n ax_i}{n} \quad (17)$$

$$= \frac{1}{n} \sum_{i=1}^n y_i - a \frac{1}{n} \sum_{i=1}^n x_i \quad (18)$$

$$= \bar{y} - a\bar{x} \quad (19)$$

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

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

Durch Ausmultiplizieren und Vereinfachen ergibt sich:

$$0 = \sum_{i=1}^n x_i (ax_i + (\bar{y} - a\bar{x}) - y_i) \quad (21)$$

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

Über Formeln zu Varianz und Kovarianz⁴ erhalten wir

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

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

$$b = \bar{y} - a \bar{x} \quad (33)$$

$$a = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad (34)$$

Wie lassen sich die beiden Parameter interpretieren? b , die Steigung, gibt den durchschnittlichen Betrag wider, um den sich y ändert, wenn x um eine Einheit verändert wird. a gibt den Betrag von y an, wenn x 0 ist. Je nachdem, welche Größen untersucht werden, kann ein x von 0 sinnvoll interpretiert werden oder nicht. Für eine Untersuchung des Körpergewichts in Abhängigkeit von der Körpergröße spielt die Körpergröße $x = 0$ sicherlich keine Rolle...

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
Σ	15	15			8	10

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

⁴Verschiebungssatz:

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

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

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

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:

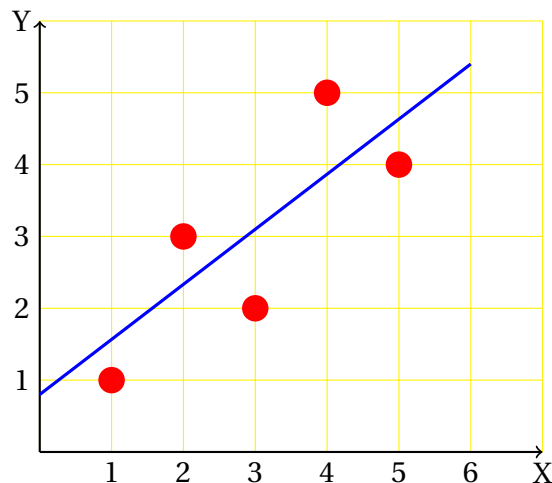


Abbildung 5: Scatterplot mit Regressionsgerade

5 Maße für die Güte der Linearen Regression

In diesem Abschnitt möchte ich erläutern, wie man die Stärke des linearen Zusammenhangs und die Güte des linearen Modells bestimmt. Unsere Regressionsparameter a und b sind die optimalen Modellparameter für die von uns genutzten Daten, aber sie erklären nicht, wie gut die Werte der unabhängigen Variablen die Werte unserer abhängigen Variablen erklären.

5.1 Standardfehler der Schätzung

Der Standardfehler der Schätzung (auf englisch: standard error of estimate, „SSE“) ist ein Maß dafür, wie stark die tatsächlichen Y -Werte von den geschätzten Werten

(\hat{Y}) abweichen. Je kleiner der Standardfehler, desto besser ist die Erklärung durch unser Modell.

Berechnet wird der Standardfehler wie folgt:

$$\text{SEE} = \sqrt{\frac{\sum (y - \hat{y})^2}{N - 2}} \quad (35)$$

$N - 2$ steht für die Anzahl der Freiheitsgrade, der Zahl der Observationen minus der Zahl der geschätzten Parameter (in unserem Beispiel Anstieg und Achsenabschnitt).

Für unser Beispiel ergibt sich SEE daher als:

Tabelle 3: Berechnung des Standardfehlers

x_i	y_i	\hat{y}_i	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	1	1.4	-0.4	0.16
2	3	2.2	0.8	0.64
3	2	3.0	-1.0	1.0
4	5	3.8	1.2	1.44
5	4	4.6	-0.6	0.36
Σ				3.6

$$\text{SEE} = \sqrt{\frac{3.6}{3}} = 1.0954451150103321 \quad (36)$$

5.2 Korrelationskoeffizient

Schauen wir uns zuerst den Korrelationskoeffizienten r an, auch Bravais-Pearson-Koeffizient genannt. r misst die Stärke und Richtung des linearen Zusammenhangs zwischen zwei metrischen Variablen. Wichtig ist hier das Wort „linearen“: r kann nicht sinnvoll bei nicht-linearen (wie z. B. quadratischen) Zusammenhängen genutzt werden.

Für r gibt es zwei Formeln:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (37)$$

$$r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \cdot \sqrt{\sum_{i=1}^n y_i^2 - n \bar{y}^2}} \quad (38)$$

Je nach den gegebenen Zahlen lässt sich besser mal die eine, mal die andere Gleichung nutzen, vom Ergebnis her ergeben beide das gleiche.

r nimmt Werte zwischen -1 und 1 an, der Wert wird wie folgt interpretiert:

$r \approx 1$ positive Korrelation, bei steigenden Werten von X steigen auch die Werte von Y. Beispiele: Größe und Gewicht einer Person, Körpergröße und Schuhgröße, Außentemperatur und Eis-Absatz an der Eisdiele

$r \approx -1$ negative Korrelation, bei steigenden Werten von X sinken die Werte von Y. Beispiele: Außentemperatur und Absatz von Glühwein

$r \approx 0$ kein linearer Zusammenhang. Beispiel: Körpergröße und Postleitzahl. Hinweis: Ein r nahe 0 bedeutet nicht, dass es keinen Zusammenhang gibt, er ist halt nur nicht linear.

Wichtig ist in diesem Zusammenhang, dass r keine Aussagen über die Kausalität trifft! Man könnte vielleicht für die Korrelation der Anzahl der verkauften Smartphones in Hongkong und der Anzahl der verkauften Eistüten in Vancouver ein positives r messen, kausal gibt es aber zwischen beiden Variablen keinen Zusammenhang.

5.3 Bestimmtheitsmaß

Quadriert man r , so erhält man das sogenannte Bestimmtheitsmaß R^2 , je nach Literatur auch „Determinationskoeffizient“ bezeichnet. R^2 ist ein Anteilswert, der die erklärte Varianz in das Verhältnis zur Gesamtvarianz setzt. Daraus folgt, dass R^2 Werte zwischen 0 und 1 (0% und 100%) annehmen kann.

Die Formel dafür lautet:

$$R^2 = r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (39)$$

Rechnen wir im nächsten Schritt mal R^2 für unser Zahlenbeispiel aus:

Tabelle 4: Hilfstabelle für die Berechnung von R^2

i	y_i	\hat{y}_i	$(\hat{y}_i - \bar{y})^2$	$(y_i - \bar{y})^2$
1	1	1.4	2.56	4.0
2	3	2.2	0.64	0.0
3	2	3.0	0.0	1.0
4	5	3.8	0.64	4.0
5	4	4.6	2.56	1.0
Σ			6.4	10.0

Damit ergibt sich

$$R^2 = \frac{6.4}{10.0} = 64\% \quad (40)$$

Zu beachten ist, dass R^2 – wie r – nur eine Aussage über den *linearen* Zusammenhang trifft, für Daten mit einem nicht-linearen Zusammenhang ist es nicht geeignet.

6 Berechnung mit dem Taschenrechner

6.1 Schultaschenrechner

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!

7 Code-Beispiele

7.1 Python

```

1 from scipy import stats
2
3 x = [1, 2, 3, 4, 5]
4 y = [1, 3, 2, 5, 4]
5
6 slope, intercept, r_value, p_value, std_error = stats.linregress(x,y)
7
8 print('Slope', slope)
9 print('Intercept', intercept)

```

```

10 print('R_value', r_value)
11 print('P_value', p_value)
12 print('Std_error', std_error)

```

```

1 Slope 0.8
2 Intercept 0.5999999999999996
3 R_value 0.8
4 P_value 0.10408803866182778
5 Std_error 0.3464101615137754

```

7.2 R

```

1 x = c(1,2,3,4,5)
2 y = c(1,3,2,5,4)
3
4 model = lm(x~y)
5 summary(model)

```

```




1 Call:
2 lm(formula = x ~y)
3
4 Residuals:
5     1     2     3     4     5
6 -0.4 -1.0  0.8 -0.6  1.2
7
8 Coefficients:
9             Estimate Std. Error t value Pr(>|t|)
10 (Intercept)  0.6000     1.1489   0.522   0.638
11 y            0.8000     0.3464   2.309   0.104
12
13 Residual standard error: 1.095 on 3 degrees of freedom
14 Multiple R-squared:  0.64,    Adjusted R-squared:  0.52
15 F-statistic: 5.333 on 1 and 3 DF, p-value: 0.1041

```

7.3 Microsoft Excel

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 
Python 
R 
Excel 