

Einführung in die lineare Regression

– DRAFT VERSION –

Uwe Ziegenhagen

8. Juni 2023

Im Rahmen meiner Lehrtätigkeiten an der Humboldt-Universität zu Berlin und der FOM (Fachhochschule für Oekonomie und Management) durfte ich auch einige Male erklären, wie man die Formeln zur Linearen Regression herleitet. Daraus entstand dieses Dokument, das ich hier mit den \LaTeX -Quellen zum Download anbiete.

Vielen Dank an diejenigen, die mir Fehler und Ungenauigkeiten im Dokument melden. Am besten dazu ein Issue im github (https://github.com/UweZiegenhagen/Introduction_Linear_Regression) einstellen, ich versuche dann, zeitnah eine neue Version bereitzustellen.

Inhaltsverzeichnis

1 Einführung	2
2 Einfache lineare Regression	2
3 Herleitung der Parameter-Gleichungen	6
4 Beispiel	9
5 Maße für die Güte der Linearen Regression	11
5.1 Standardfehler der Schätzung	11
5.2 Korrelationskoeffizient	12
5.3 Bestimmtheitsmaß	13
6 Berechnung mit dem Taschenrechner	14
6.1 Schultaschenrechner	14
7 Anhang	15

8 „BLUE“ oder der Satz von Gauß-Markov	15
8.1 Code-Beispiele	15
8.1.1 Python	15
8.1.2 R	16
8.1.3 Microsoft Excel	17

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 *linear* 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 eine X -Variable gibt – spricht man von einer einfachen linearen Regression, in höheren Dimensionen – wenn es mehrere X_i -Variablen gibt – von der multiplen linearen Regression.

2 Einfache lineare Regression

Im Folgenden nutzen wir die Werte aus Tabelle 1, um an ihnen die einfache lineare Regression zu erklären. Die Zahlen könnten beispielsweise die verkaufte Menge Cocktails (Y) in Abhängigkeit vom Verkaufspreis (X) sein. Die Zahlen wurden so gewählt, dass man gut mit ihnen rechnen kann.

X-Wert	Y-Wert
1	50
2	40
3	45
4	20
5	25
6	15

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

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 sinken.

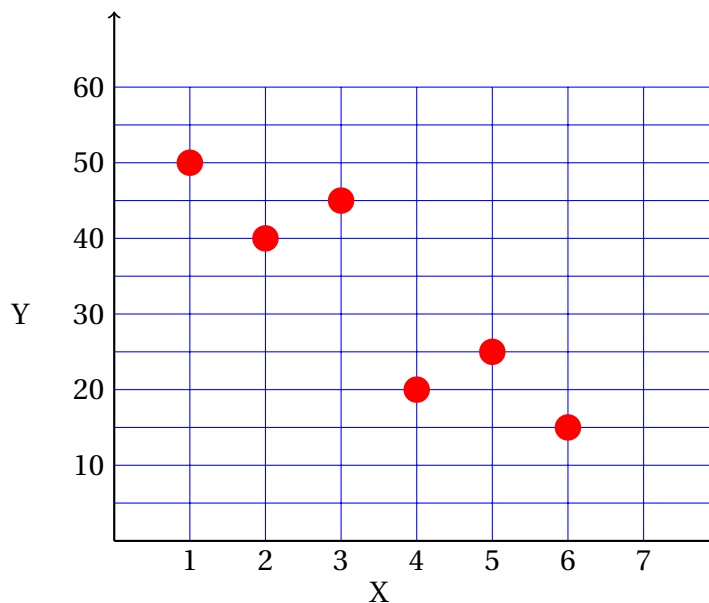


Abbildung 1: Scatterplot zur Darstellung der X-Y Wertepaare

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

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

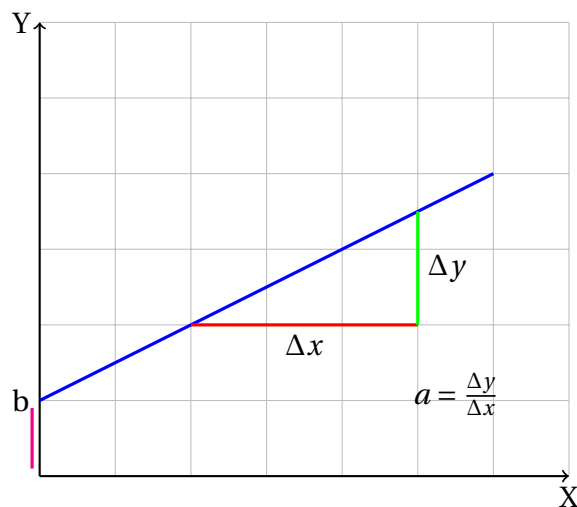


Abbildung 2: Grafische Erläuterung

- 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 Wertepaare zu bestimmen.
- ϵ_i steht für die Fehler, den wir bei der Modellierung machen, darauf kommen wir später noch zu sprechen.

Abbildung 2 auf Seite 3 beschreibt diesen Zusammenhang grafisch.

Wir können wir nun die Regressionsgerade durch die Punkte zeichnen? Abbildung 3 zeigt zwei Beispiele für beliebig gewählte Regressionsgeraden. Im linken Teil erkennt man, dass die Gerade schon recht gut zu unseren Punkten passt.

Im rechten Teil stimmt die Richtung überhaupt nicht, diese Gerade impliziert nämlich, dass mit steigendem X die Werte von Y ebenfalls steigen, also – angewandt auf unser Beispiel – bei höherem Preis mehr Cocktails verkauft werden, was normalerweise² Quatsch ist.

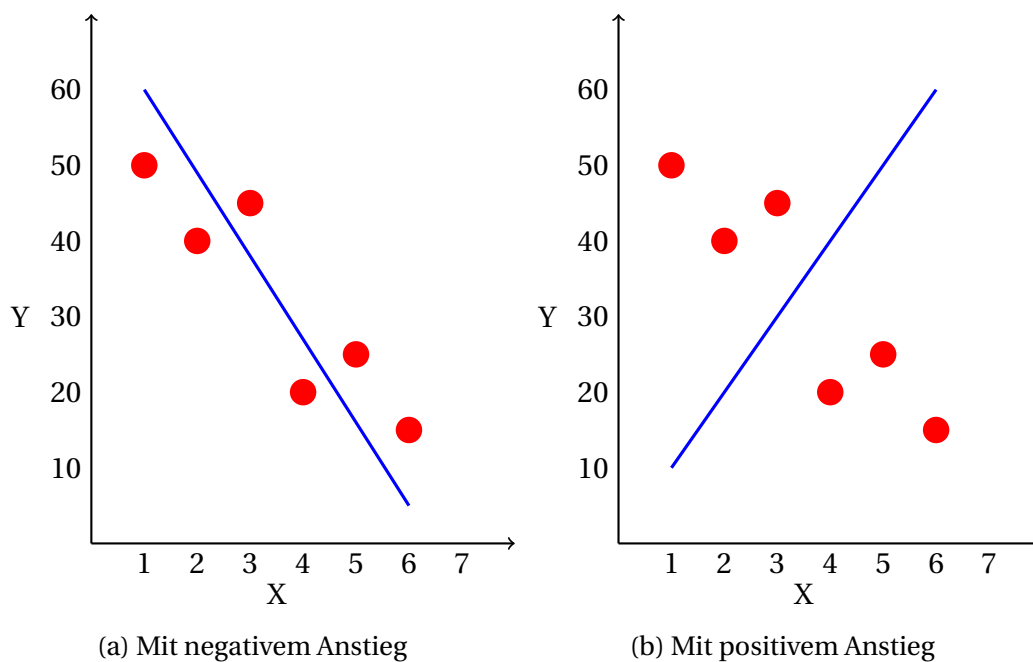


Abbildung 3: Zwei Regressionsgeraden

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

Betrachten wir dazu Abbildung 4. Hier wurden auf der blauen Geraden die Punkte in grün markiert, die eine Regressionsgleichung für den jeweiligen Wert von X vorhersagt, außerdem wurden die jeweiligen Abstände zwischen dem wahren Y -Wert und dem durch die Regressionsgerade geschätzten Y -Wert (den wir ab jetzt \hat{Y} nennen) markiert.

²Siehe <https://de.wikipedia.org/wiki/Giffen-Paradoxon> für die Ausnahme „Giffen-Gut“

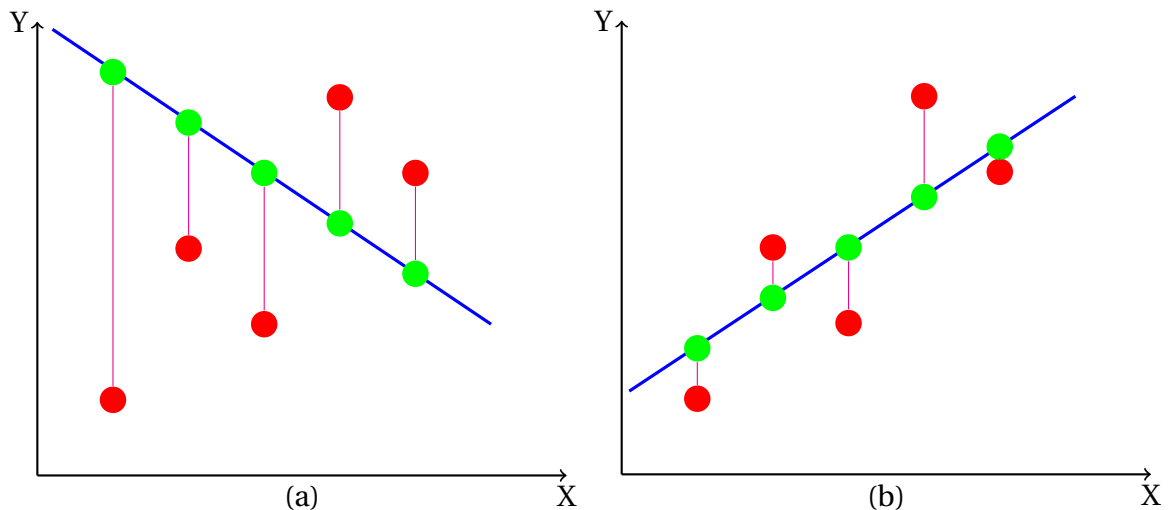


Abbildung 4: Zwei Regressionsgeraden

Wenn wir die Abweichungen der tatsächlichen Y -Werte von den geschätzten Y -Werten als Maß für die Güte der Regression nutzen, dann wird beim Betrachten der beiden Abbildungen schnell klar, dass die Regressionsgerade im linken Bild deutlich schlechter ist als die Regressionsgerade im rechten Bild: die Summe der Abstände zwischen den wahren Punkten (in rot) und den durch die Gerade geschätzten Punkten (in grün) ist deutlich größer.

Aus dieser Tatsache lässt sich ein sehr wichtiger Schluss ziehen: wenn wir diese Summe der Abstände mathematisch *minimieren* könnten, dann würden wir die *optimale* Gerade erhalten.

In Gleichungsform können wir dies wie folgt aufschreiben:

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

In Worten: S ist die Summe aller Differenzen von wahren 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 ein wenig mehr im Anhang unter Abschnitt 8.

Aus Gleichung 2 wird jetzt – da wir ja die Quadratsumme QS 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:

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

$$= \sum_{i=1}^n \left(y_i - \overbrace{[ax_i + b]}^{\hat{y}_i} \right)^2 \quad (6)$$

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 6 vereinfachen, denn hier lässt sich zweimal eine Binomische Formel anwenden, was in Gleichung 7 deutlich wird. Wenn wir den linken Teil des Binoms als s bezeichnen und den rechten Teil des Binoms als t , dann können wir die zweite Binomische Formel anwenden.

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

Mit Hilfe der 2. Binomischen Formel³ kommen wir von Gleichung 7 zu Gleichung 8:

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

Als nächstes können wir noch auf den Ausdruck $(ax_i + b)^2$ die 1. Binomische Formel⁴ anwenden:

$$QS(a, b) = \sum_{i=1}^n \left(y_i^2 - 2ax_i y_i - 2by_i + \overbrace{a^2 x_i^2}^{s^2} + \overbrace{2abx_i}^{2st} + \overbrace{b^2}^{t^2} \right) \quad (9)$$

Ausgehend von Gleichung 9 bilden wir jetzt die partiellen Ableitungen nach a und b .

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

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

$$\frac{\partial \text{QS}(a, b)}{\partial a} = \sum_{i=1}^n (-2x_i y_i + 2ax_i^2 + 2bx_i) \quad (10)$$

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

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

$$\frac{\partial \text{QS}(a, b)}{\partial b} = \sum_{i=1}^n (-2y_i + 2ax_i + 2b) \quad (13)$$

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

Wenn wir Gleichung 14 nullsetzen und auflösen, erhalten wir

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

$$2 \sum_{i=1}^n ax_i + 2nb - 2 \sum_{i=1}^n y_i = 0 \quad (16)$$

$$2nb = 2 \sum_{i=1}^n y_i - 2 \sum_{i=1}^n ax_i \quad (17)$$

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

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

$$= \frac{1}{n} \sum_{i=1}^n y_i - a \frac{1}{n} \sum_{i=1}^n x_i \quad (19)$$

$$= \bar{y} - a\bar{x} \quad (20)$$

Setzen wir nun $b = \bar{y} - a\bar{x}$ in Gleichung 12 ein, erhalten wir

$$2 \sum_{i=1}^n x_i (ax_i + (\bar{y} - a\bar{x}) - y_i) = 0 \quad (21)$$

Durch Ausmultiplizieren und Vereinfachen ergibt sich:

$$0 = \sum_{i=1}^n x_i (ax_i + (\bar{y} - a\bar{x}) - y_i) \quad (22)$$

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

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

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

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

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

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.

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

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

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

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

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

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

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

Jetzt zerlegen wir die Summe unter dem Bruchstrich in Einzelsummen

$$a = \frac{\sum_{i=1}^n x_i y_i - \bar{y} n\bar{x}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{x}} \quad (32)$$

und ziehen \bar{x} vor das zweite Summenzeichen, denn \bar{x} ist ja ein konstanter Term:

$$a = \frac{\sum_{i=1}^n x_i y_i - \bar{y} n \bar{x}}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} \quad (33)$$

Über Formeln zu Varianz und Kovarianz⁵ erhalten wir

$$a = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{n \left(\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \right)}{n \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \right)} = \frac{n \text{Cov}(x, y)}{n \text{Var}(x)} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad (34)$$

Anmerkung

Die Kovarianz ist definiert als der Erwartungswert der Produkte der Abweichungen zweier Zufallsvariablen von ihren jeweiligen Populations-Mittelwerten. Sie ist ein Maß für den *linearen* Zusammenhang, diese Einschränkung ist wichtig. Eine positive Kovarianz bedeutet, dass sich beide Zufallsvariablen in eine Richtung gemeinsam bewegen (positiv-positiv), eine negative Kovarianz bedeutet, dass sich beide Variablen entgegengesetzt bewegen (positiv-negativ).

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

$$b = \bar{y} - a \bar{x} \quad (35)$$

$$a = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad (36)$$

Wie lassen sich die beiden Parameter interpretieren?

a , die Steigung, gibt den durchschnittlichen Betrag wieder, um den sich y ändert, wenn x um eine Einheit verändert wird.

b gibt den Betrag von y an, wenn x den Wert 0 hat.

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

⁵Verschiebungssatz:

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

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

Spalte	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	50	-2.5	17.5	-43.75	6.25
	2	40	-1.5	7.5	-11.25	2.25
	3	45	-0.5	12.5	-6.25	0.25
	4	20	0.5	-12.5	-6.25	0.25
	5	25	1.5	-7.5	-11.25	2.25
	6	15	2.5	-17.5	-43.75	6.25
Σ	21	195			-122.5	17.5

Mit Hilfe der Werte aus der Tabelle lassen sich a und b jetzt einfach bestimmen.
Hinweis: $\bar{x} = 21/6 = 3.5$, $\bar{y} = 195/6 = 32.5$

$$a = \frac{\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{6}}{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{6}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{-122.5}{17.5} = -7$$

Hinweis zu dieser Rechnung: Brüche werden dividiert, indem man mit dem Kehrwert multipliziert: $\frac{1/6}{1/6} = 1/6 \cdot 6/1 = 1$. Für die Berechnung von a braucht man die Anzahl der Beobachtungen also nicht mehr.

$$b = \bar{y} - a \cdot \bar{x} = 32.5 - (-7) \cdot 3.5 = 57.0$$

Unsere Regressionsgleichung lautet also

$$y = -7 \cdot x + 57$$

Mit den gefundenen Werten für unsere beiden Parameter können wir jetzt die Regressionsgerade zeichnen, siehe dazu Abbildung 5 auf Seite 10.

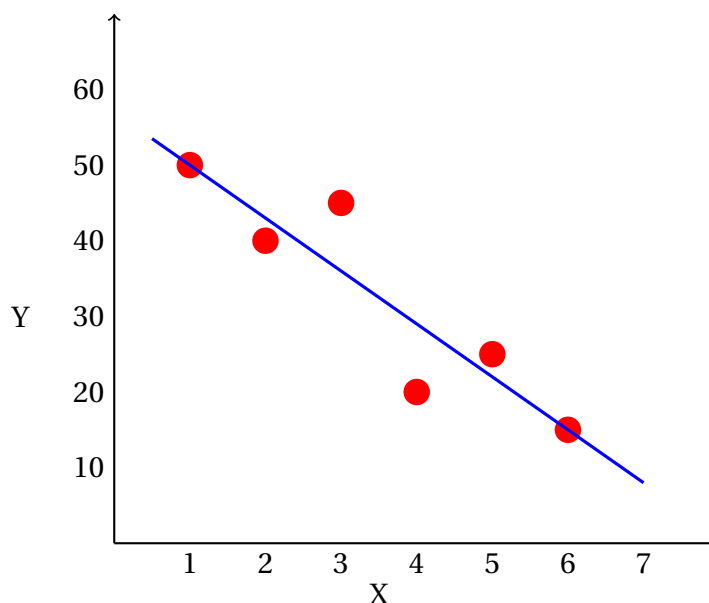


Abbildung 5: Scatterplot mit Regressionsgerade

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, „SEE“) ist ein Maß dafür, wie stark die tatsächlichen Y -Werte von den geschätzten Werten (\hat{Y}) abweichen, er repräsentiert die durchschnittliche Abweichung der beobachteten Werte von der Regressionsgeraden.

Je kleiner der Standardfehler, desto besser ist die Erklärung durch unser Modell. Berechnet wird der Standardfehler für eine Stichprobe wie folgt:

$$SEE = \sqrt{\frac{1}{N-2} \sum (y - \hat{y})^2} \quad (37)$$

$N - 2$ steht für die Anzahl der Freiheitsgrade, der Zahl der Beobachtungen minus der Zahl der geschätzten Parameter. In unserem Beispiel haben wir Anstieg und Achsenabschnitt geschätzt, daher beträgt die Zahl der Parameter 2.

Hinweis: Berechnet man SEE für eine Grundgesamtheit, dann wird aus $N - 2$ ein N , da keine Parameter mehr geschätzt werden, sondern aus der Population errechnet werden konnten.

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	50	50	0	0
2	40	43	-3	9
3	45	36	9	81
4	20	29	-9	81
5	25	22	3	9
6	15	15	0	0
Σ	180			

$$SEE = \sqrt{\frac{1}{4} \cdot 180} = \sqrt{45} = 6.7082 \quad (38)$$

Wie lässt sich dieser Wert interpretieren? Ich würde es wie folgt beschreiben: „Im Mittel weicht die prognostizierte Zahl der verkauften Cocktails von der gemessenen Anzahl um ungefähr sieben Stück ab.“

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:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (39)$$

$$r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \cdot \sqrt{\sum_{i=1}^n y_i^2 - n \bar{y}^2}} \quad (40)$$

Je nachdem, welche Werte man gegeben hat oder bereits berechnet hat, lässt sich besser mal der einen, mal mit der anderen Gleichung rechnen. Vom Ergebnis her ergeben beide Gleichungen das gleiche Ergebnis.

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.

Mittels Signifikanztest kann man prüfen, ob der errechnete Korrelationskoeffizient tatsächlich signifikant ist oder nur ein Ergebnis des Zufalls unserer Stichprobe ist. Als Hypothesen formulieren wir

H_0 Die Korrelation in der Population ist 0 ($\rho = 0$).

H_1 Die Korrelation in der Population ist ungleich 0 ($\rho \neq 0$).

Als Teststatistik für diesen zweiseitigen Test erhalten wir

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (41)$$

t ist unter der Nullhypothese t -verteilt mit $n-2$ Freiheitsgraden. Den kritischen Wert entnehmen wir aus der Tabelle der t -Verteilung⁶ bei beispielsweise 95% und $n-2$ und lehnen die Nullhypothese ab, wenn unsere Teststatistik kleiner als -1 mal oder größer als der kritische Wert ist.

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:

$$R^2 = r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (42)$$

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

Tabelle 4: Hilfstabelle für die Berechnung von R^2

⁶z. B. von https://www.statistik.tu-dortmund.de/fileadmin/user_upload/Lehrstuehle/Oekonometrie/Lehre/WiSo0ekoSS16/tabelletV.pdf

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

$$R^2 = \frac{6.4}{10.0} = 64\% \quad (43)$$

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 Casio fx-991DE X CLASSWIZ, wie es funktioniert.

1. Über den „Menu“-Button wechselt man in das Statistik-Menü und wählt Punkt 2 für das Modell $y = a + bx$
2. Es erscheint das Tabellenblatt für die Eingabe der x - und y -Werte, hier gibt man jetzt alle 12 Werte ein. Mit „=“ wird ein Wert bestätigt, mit den Pfeiltasten wechselt man zwischen den Zellen.
3. Nach der Eingabe aller Werte drückt man die „OPTN“-Taste und gelangt in ein Menü.
4. Mit „4“ kommt man in das „Regression“-Untermenü, hier zeigt der Rechner dann an:
 - das Regressionsmodell $y = a + bx$
 - den Wert von a : $a=57$
 - den Wert von b : $b=-7$
 - sowie den Korrelationskoeffizienten r : $r=-0.909123767$
5. Drückt man wiederholt OPTN und dann die „3“, so gelangt man in das Menü für die Berechnung der Variablen. Hier zeigt der Rechner die verschiedenen Teilergebnisse wie \bar{x} , Σx , Σx^2 , etc.

7 Anhang

8 „BLUE“ oder der Satz von Gauß-Markov

Aus der Wikipedia: Der Satz von Gauß-Markov besagt, dass in einem linearen Regressionsmodell, in dem

1. die Störgrößen einen Erwartungswert von null
2. und eine konstante Varianz haben
3. sowie unkorreliert sind

der Kleinste-Quadrate-Schätzer ein bester linearer erwartungstreuer Schätzer, kurz „Best Linear Unbiased Estimator“ (BLUE) ist. Die drei genannten Punkte sind die Annahmen des klassischen Linearen Regressionsmodells,

8.1 Code-Beispiele

8.1.1 Python

```
1 from scipy import stats
2 import pandas as pd
3 import numpy as np
4 import matplotlib.pyplot as plt
5
6 import seaborn as sns
7 sns.set(rc={'figure.figsize':(11.7,8.27)})
8
9 x = [1, 2, 3, 4, 5, 6]
10 y = [50, 40, 45, 20, 25, 15]
11
12 # Calculation via stats module
13 slope, intercept, r_value, p_value, std_error = stats.linregress(x,y)
14
15 df = pd.DataFrame(np.transpose(np.array([x,y])), columns = ['x', 'y'])
16
17 print('Slope', slope)
18 print('Intercept', intercept)
19 print('R_value', r_value)
20 print('P_value', p_value)
21 print('Std_error', std_error)
22
23 # manual computation
24 mean_x = df.mean()[0]
25 mean_y = df.mean()[1]
26
27 df['x-mean_x'] = df['x'] - mean_x
28 df['y-mean_y'] = df['y'] - mean_y
```

```

29
30 df['(x-mean_x)(y-mean_y)'] = df['x-mean_x'] * df['y-mean_y']
31
32 df['(x-mean_x)^2'] = df['x-mean_x']**2
33
34 sum_xmeanxymeany = df['(x-mean_x)(y-mean_y)'].sum()
35 sum_xminusxsq = df['(x-mean_x)^2'].sum()
36
37 print(sum_xmeanxymeany, sum_xminusxsq)
38
39
40 slope = sum_xmeanxymeany / sum_xminusxsq
41 intersect = mean_y - slope * mean_x
42
43 print(df, '\n\nSlope: ', slope, '\nIntersect: ', intersect)
44
45 # graphics
46 plt.scatter(x,y)
47
48 sns.regplot(x=df.x, y=df.y, ci=False);

```

```

1 Slope -6.999999999999999
2 Intercept 57.0
3 R_value -0.909123767204656
4 P_value 0.012012484314940097
5 Std_error 1.6035674514745457
6 -122.5 17.5
7      x  y  x-mean_x  y-mean_y  (x-mean_x)(y-mean_y)  (x-mean_x)^2
8 0 1 50      -2.5      17.5          -43.75          6.25
9 1 2 40      -1.5       7.5          -11.25          2.25
10 2 3 45      -0.5      12.5           -6.25          0.25
11 3 4 20       0.5     -12.5           -6.25          0.25
12 4 5 25       1.5      -7.5          -11.25          2.25
13 5 6 15       2.5     -17.5          -43.75          6.25
14
15 Slope: -7.0
16 Intersect: 57.0

```

8.1.2 R

```

1 x = c(1, 2, 3, 4, 5, 6)
2 y = c(50, 40, 45, 20, 25, 15)
3
4 model = lm(y~x)
5 summary(model)

```

```

1 Call:

```



```

2 lm(formula = y ~x)
3
4 Residuals:
5      1      2      3      4      5      6
6 9.974e-15 -3.000e+00 9.000e+00 -9.000e+00 3.000e+00 5.112e-15
7
8 Coefficients:
9             Estimate Std. Error t value Pr(>|t|)
10 (Intercept)  57.000      6.245   9.127 0.000799 ***
11 x           -7.000      1.604  -4.365 0.012012 *
12 ---
13 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
14
15 Residual standard error: 6.708 on 4 degrees of freedom
16 Multiple R-squared:  0.8265, Adjusted R-squared:  0.7831
17 F-statistic: 19.06 on 1 and 4 DF, p-value: 0.01201

```

8.1.3 Microsoft Excel

Die Excel-Datei in diesem Github-Repo enthält zwei Arbeitsblätter, einmal mit der Berechnung auf traditionelle Weise, einmal mit der Berechnung über die entsprechenden Excel-Formeln STEIGUNG und ACHSENABSCHNITT.

	A	B	C	D	E	F	G	H	I
1									
2		x	y	x-xbar	y-ybar	(x-xbar)(y-ybar)	(x-xbar)^2		
3		1	50	-2,5	17,5	-43,75	6,25		
4		2	40	-1,5	7,5	-11,25	2,25		
5		3	45	-0,5	12,5	-6,25	0,25		
6		4	20	0,5	-12,5	-6,25	0,25		
7		5	25	1,5	-7,5	-11,25	2,25		
8		6	15	2,5	-17,5	-43,75	6,25		
9		21	195			-122,5	17,5		
10		3,5	32,5						
11									
12					Steigung	-7			
13									
14									
15					Intersect	57			
16									


Abbildung 6: Manuelle Berechnung in Excel

	A	B	C	D	E	F	G	H	I
1									
2		Preis	Menge						
3		1	50						
4		2	40						
5		3	45					a	-7
6		4	20						
7		5	25					b	57
8		6	15						
9									
10								R^2	0,82650602
11									
12								r	0,90912377
13									
14									
15									
16									

Abbildung 7: Berechnung über Excel-Formeln

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 