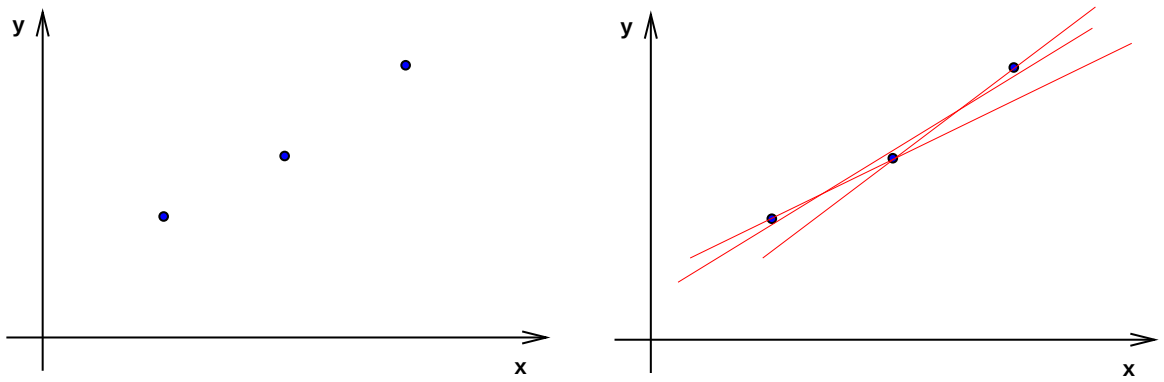
Statistik
Dr. Thomas Zehrt Regression

Motivation

Eine Firma will den Zusammenhang zwischen Werbungskosten und Absatz untersuchen. Dazu nimmt sie zunächst eine Stichprobe dieser beiden Merkmale

x_i	1	2	3
y_i	2	3	4.5

dabei sind x_i die Werbungskosten (in 1′000.−) und y_i der Absatz je Kunde (in 100′000.−). Welcher Absatz ist bei Werbungskosten von 8′000.− zu erwarten? Zunächst werde die Daten in einem Streuungsdiagramm (Punktwolke) dargestellt:



Es **scheint** zwischen beiden Merkmalen einen linearen Zusammenhang zu geben, der durch verschiedene Einflüsse leicht verfälscht ist. Deshalb wählen wir einen linearer Modellansatz:

$$y = f(x) \quad = \quad a\,x + b$$

Wie sollen die Zahlen a und b gewählt werden, d.h. welche Gerade kommt unserer Punktwolke am nächsten? Diese Frage wollen wir in diesem Kapitel beantworten, denn mit Hilfe dieser Funktion könnten wir eine Prognose für den Absatz bei Werbungskosten von 8′000.− wagen, indem wir $f(8)$ berechnen.

Benötigtes Schulwissen

- sicherer Umgang mit Summenzeichen
- ein wenig Differentialrechnung. Insbesondere müssen hier Funktionen abgeleitet werden, die von zwei oder drei Variablen abhängen, z.B. hängt die Funktion $S(a,b) = 2a^2 - 4ab + b^2$ von den beiden Variablen a und b ab und wir können S nach diesen beiden Variablen ableiten. Dabei wird die jeweils andere Variable, nach der nicht abgeleitet wird, wie eine Konstante behandelt. Es gilt:

$$\text{Ableitung von } S \text{ nach } a = \frac{\partial S(a,b)}{\partial a} = 4a - 4b$$
$$\text{Ableitung von } S \text{ nach } b = \frac{\partial S(a,b)}{\partial b} = -4a + 2b.$$

  Lassen Sie sich von der Notation nicht verwirren, sie ist willkürlich aber gebräuchlich.

- Die Potenz- und Logarithmengesetze: Sei $a > 0$ und $a \neq 1$ eine reelle Zahl. Dann gelten für alle reellen Zahlen r, s die folgenden Potenzgesetze bzw. für alle reellen Zahlen $u > 0$ und $v > 0$ die Logarithmengesetze:

1.	$a^r \cdot a^s$	$=$	a^{r+s}
2.	$\frac{a^r}{a^s}$	$=$	a^{r-s}
3.	$(a^r)^s$	$=$	$(a^s)^r = a^{r \cdot s}$

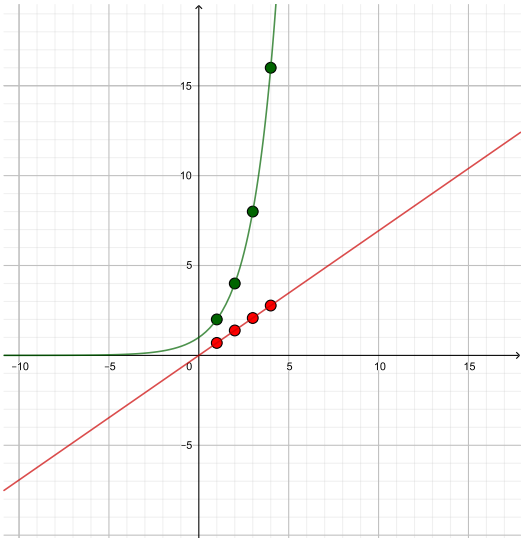
1.	$\log_a(u \cdot v)$	$=$	$\log_a(u) + \log_a(v)$
2.	$\log_a\left(\frac{u}{v}\right)$	$=$	$\log_a(u) - \log_a(v)$
3.	$\log_a(u^w)$	$=$	$w \cdot \log_a(u)$

- Machen Sie sich (an einem Beispiel) folgenden Sachverhalt klar:

  Liegen die Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ alle auf dem Graphen einer Exponentialfunktion (d.h. stets gilt $y_i = a^{x_i}$), dann liegen die Punkte $(x_1, \ln(y_1)), \ldots, (x_n, \ln(y_n))$ alle auf dem Graphen einer linearen Funktion (Geraden).

x_i	1	2	3	4
y_i	$2^1 = 2$	$2^2 = 4$	$2^3 = 8$	$2^4 = 16$

x_i	1	2	3	4
$\ln(y_i)$	$\ln(2)$	$2\ln(2)$	$3\ln(2)$	$4\ln(2)$

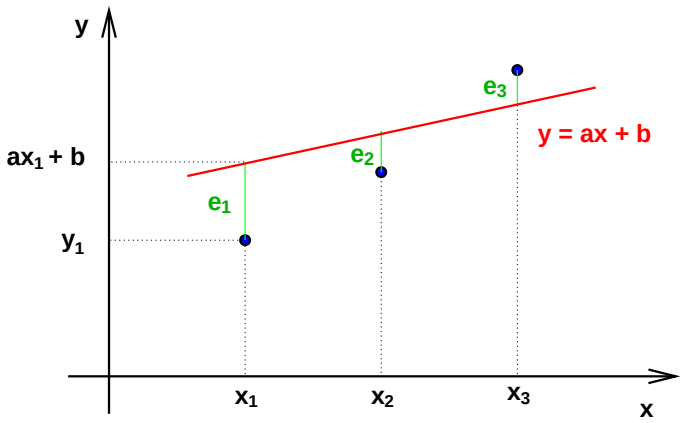


1 Einführung

Welche Gerade ist die Beste?

Wir wollen das allgemeine Vorgehen am Beispiel aus der Motivation durchspielen.

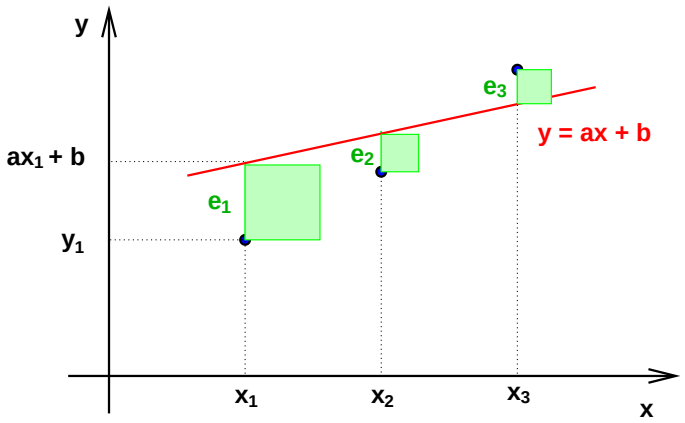
- Für jede Gerade $y = f(x; a, b) = a\ x + b$ führen wir in zwei Schritten ein Strafmass für deren Abweichung von der Punktwolke ein. Dieses Strafmass wird eine Funktion $S(a, b)$ sein, die von den beiden Parametern a und b (also von der Geraden) abhängt.
  - Abweichung der Geraden im Punkt x_i: $e_i = y_i - \underbrace{(a\ x_i + b)}_{f(x_i;a,b)} \in \mathbb{R}$



$$\begin{aligned} e_1 &= 2 - (a + b) = 2 - a - b \\ e_2 &= 3 - (2a + b) = 3 - 2a - b \\ e_3 &= 4.5 - (3a + b) = 4.5 - 3a - b \end{aligned}$$

  - Gesamtstrafe für die Gerade $y = a\ x + b$: $\underbrace{S(a, b)}_{\geq 0} = \sum_{i=1}^{n} \underbrace{e_i^2}_{\geq 0}$

    Diese Gesamtstrafe können wir uns als die Summe der Flächeninhalte aller Quadrate mit den Seitenlängen e_i vorstellen.



Für unser Beispiel bedeutet das:

$$\begin{aligned} S(a, b) &= (2 - a - b)^2 + (3 - 2a - b)^2 + (4.5 - 3a - b)^2 \\ &= 14a^2 + 3b^2 + 12ab - 43a - 19b + 33.25 \end{aligned}$$

- Wir minimieren diese Funktion S, d.h. wir suchen die Werte $\hat{a}$ und $\hat{b}$ die die Funktion (global) minimieren. Dazu bestimmen wir die Extremalstellen der Funktion $S(a, b)$ und notwendige Bedingungen sind das Verschwinden der Ableitungen von S nach beiden Variablen a und b:

$$0 = \text{Ableitung von } S(a,b) \text{ nach } a \quad \text{und} \quad 0 = \text{Ableitung von } S(a,b) \text{ nach } b$$

  Für unser Beispiel bedeutet das:

$$0 = 12a + 6b - 19 \quad \text{und} \quad 0 = 28a + 12b - 43$$

  mit den beiden Lösung: $\hat{b} = \frac{2}{3}$ und $\hat{a} = \frac{5}{4}$. Das sind die Koordinaten des einzigen lokalen (und globalen) Minimums der Funktion $S(a, b)$ und die optimale Gerade ist $y = \frac{5}{4}\, x + \frac{2}{3}$!

2 Problemstellung

Seien X und Y zwei quantitative Merkmale. Die Regressionsrechnung beschäftigt sich mit der Frage der **Form** des (statistischen) Zusammenhangs beider Merkmale. Meistens ist diese Form von vorneherein durch ein theoretisches physikalisches Modell gegeben.

Definition 2.1 *Die Modellgleichung zwischen den Merkmalen X und Y lautet*

$$y \;=\; f(x; a, b, c, \ldots)$$

mit einer (an das Problem angepassten) Funktion f, mit noch zu bestimmenden Parametern $a, b, c, \ldots$. Das Merkmal X heisst Ursache das Merkmal Y die Wirkung. Die Variable x heisst frei und y abhängig.

Einige Beispiele:

Merkmal Y	Merkmal X	Modellgleichung	gesucht
Stromstärke I	Spannung U	$I = a \cdot U$	a
Energie E	Masse m	$E = a \cdot m$	a
Absatz(menge) eines Gutes	Werbungskosten	$y = ax + b$	a, b
Nachfrage nach einem Gut	Preis des Gutes	$y = ax + b$	a, b
Angebot eines Gutes	Preis des Gutes	$y = ax + b$	a, b
Nachfrage nach einem Gut	persönl. Einkommen	$y = a\left(1 - \frac{b}{x}\right)$	a, b
Nachfrage nach einem Gut	persönl. Einkommen	$y = \frac{a}{x}$	a
Produktionskosten	Output	$y = ax^3 + bx^2 + cx + d$	a, b, c, d

Nun führt man n Messungen beider Merkmale durch, erhält also n Messwertpaare oder Punkte $(x_1, y_1), \ldots, (x_n, y_n)$, die auf Grund von Messfehlern nicht genau auf einer dem Modell entsprechenden Kurve liegen. Mit Hilfe der Regressionsrechnung versucht man nun die unbekannten Modellparameter näherungsweise zu bestimmen, in dem man aus der Menge aller dem Modell entsprechenden Kurven die auswählt, welche die Punktwolke am Besten approximiert. Wir werden uns im Weiteren auf die einfachsten Modelle einschränken:

1. lineare Modelle und
2. exponentielle Modelle.

Wir suchen also die Kurven, die der Punktwolke am Besten (Was bedeutet das?) angepasst sind. Weiterhin wollen wir davon ausgehen, dass die Messwerte von X nicht alle gleich sind (Warum?).

3 Lineare Regression

3.1 Die Regressionsgerade

Die Modellgleichung zwischen den Merkmalen X und Y lautet $y = ax + b$ und wir nutzen die folgende klassische Methode, um die Parameter a und b zu bestimmen.

Die Methode der kleinsten Quadrate Diese Methode geht auf einen der bedeutendsten deutschen Mathematiker zurück: Carl Friedrich Gauß (1777-1855).

Zunächst denken wir uns eine beliebige Gerade $y = ax + b$ durch die Punktwolke gelegt. Als Strafmass für die Abweichung der Geraden von einem einzelnen Punkt (x_i, y_i) der Punktwolke, wählen wir das Quadrat des Abstandes von Punkt und Gerade in senkrechter Richtung. Dieser (noch vorzeichenbehaftete) Abstand beträgt:

$$e_i \;:=\; y_i - (ax_i + b)$$

und das folgende quadrieren beseitigt die negativen Vorzeichen. Als Gesamtstrafmass $S(a, b)$ für die Abweichung scheint es nun naheliegend zu sein, die Summe aller Abweichungsquadrate zu bilden:

$$S(a,b) \;=\; \sum_{i=1}^{n} e_i^2 \;=\; \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Als gute Übung für den Umgang mit Summenzeichen formen wir die rechte Seite etwas um.

$$\begin{aligned}
S(a,b) &= \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 \\
&= \sum_{i=1}^{n} [y_i^2 - 2y_i(ax_i + b) + (ax_i + b)^2] \\
&= \sum_{i=1}^{n} [y_i^2 - 2ax_iy_i - 2by_i + a^2x_i^2 + 2abx_i + b^2] \\
&= \sum_{i=1}^{n} y_i^2 - 2a\sum_{i=1}^{n} x_iy_i - 2b\sum_{i=1}^{n} y_i + a^2\sum_{i=1}^{n} x_i^2 + 2ab\sum_{i=1}^{n} x_i + \underbrace{\sum_{i=1}^{n} b^2}_{=b^2\cdot n}
\end{aligned}$$

Definition 3.1 *Die Regressionsgerade $y = \hat{a}x + \hat{b}$ zu den Wertepaaren $(x_1, y_1), \ldots, (x_n, y_n)$ ist die Gerade, für die die Summe $S(a, b)$ minimal wird. Wir suchen also Zahlen $\hat{a}$ und $\hat{b}$, so dass*

$$S(\hat{a}, \hat{b}) \;\leq\; S(a, b)$$

für alle reellen Zahlen a und b. Weiterhin sei $\hat{y}_i = \hat{a}x_i + \hat{b}$ für alle $i = 1, 2, \ldots n$.

Das ist ein klassisches Minimierungsproblem, wir müssen also die Nullstellen der (partiellen) Ableitungen von $S(a,b)$ (nach a und b) bestimmen.

1.

$$\begin{aligned} 0 \;=\; \frac{\partial}{\partial a}S(a,b) \;&=\; 2\sum_{i=1}^{n}(y_i - ax_i - b)(-x_i) \quad \Big| : 2 \\ \Longrightarrow \quad 0 \;&=\; -\sum_{i=1}^{n} x_i y_i + a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i \end{aligned}$$

2.

$$\begin{aligned} 0 \;=\; \frac{\partial}{\partial b}S(a,b) \;&=\; 2\sum_{i=1}^{n}(y_i - ax_i - b)(-1) \quad \Big| : 2 \\ \Longrightarrow \quad 0 \;&=\; -\sum_{i=1}^{n} y_i + a\sum_{i=1}^{n} x_i + b\cdot n \end{aligned}$$

Nun müssen wir dieses Gleichungssystem, bestehend aus zwei Gleichungen für zwei Unbekannte a und b, lösen.

Satz 1 *Die Regressionsgerade zu den Wertepaaren $(x_1,y_1),\ldots,(x_n,y_n)$ hat die Gleichung $\hat{y} = \hat{a}x + \hat{b}$ mit den Regressionsparametern $\hat{a}$ und $\hat{b}$, die auf eine der folgenden Arten bestimmt werden können.*

1. *Als Lösung des Gleichungssystems:*

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i \;&=\; a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i \;&=\; a\sum_{i=1}^{n} x_i + b\cdot n. \end{aligned}$$

2. *Direkt:*

$$\hat{a} = \frac{cov(X,Y)}{var(X)} \qquad und \qquad \hat{b} = \bar{y} - \frac{cov(X,Y)}{var(X)}\,\bar{x}.$$

Aus der zweiten Gleichung lässt sich sofort die Zentraleigenschaft herleiten:

$$\sum_{i=1}^{n} e_i \;=\; \sum_{i=1}^{n}\left(y_i - \hat{a}x_i - \hat{b}\right) \;=\; \sum_{i=1}^{n} y_i - \hat{a}\sum_{i=1}^{n} x_i - \hat{b}\,n \;=\; 0.$$

Die Regressionsgerade liegt also so, dass die Summe aller Residuen verschwindet.

Hat man nur die Messwerte zur Verfügung, so ist der Weg über das Gleichungssystem effizienter. Falls man die Messwerte schon weiter untersucht hat, also schon Mittelwerte, Varianzen und Kovarianz bestimmt hat, sollte man den direkten Weg bevorzugen.

Beweis:

Die beiden Modellparameter ergeben sich direkt als Lösung des obigen linearen Gleichungssystems, das wir in der zweiten Zeile vereinfachen:

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i &= a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i \qquad \textit{und} \qquad \sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} 1 \\ \sum_{i=1}^{n} x_i y_i &= a \sum_{i=1}^{n} x_i^2 + b \cdot n\bar{x} \qquad \textit{und} \qquad \bar{y} = a \cdot \bar{x} + b \end{aligned}$$

Lösen wir die zweite Gleichung nach b auf, erhalten wir zunächst $b = \bar{y} - a \cdot \bar{x}$, was wir in die erste Gleichung einsetzen:

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i &= a \sum_{i=1}^{n} x_i^2 + b \cdot n\bar{x} \\ &= a \sum_{i=1}^{n} x_i^2 + (\bar{y} - a \cdot \bar{x}) \cdot n\bar{x} \\ &= a \sum_{i=1}^{n} x_i^2 + \bar{y} \cdot n \cdot \bar{x} - a \cdot \bar{x} \cdot n \cdot \bar{x} \\ &= a \left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right) + \bar{y} \cdot n \cdot \bar{x} \end{aligned}$$

oder eben

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2} = \frac{cov(X,Y)}{var(X)}$$

Um den optimalen Wert für b zu berechnen, setzen wir unser Resultat für $\hat{a}$ in die zweite Gleichung ein und erhalten:

$$\hat{b} = \bar{y} - \frac{cov(X,Y)}{var(X)} \cdot \bar{x}$$

Sie sind somit die Koordinaten der (einzigen) Nullstelle der partiellen Ableitungen. Es ist noch zu klären, dass diese Stelle wirklich ein Minimum ist. Das ergibt sich aber daraus, dass $S(a,b)$ offensichtlich beliebig gross werden kann, also sicher kein globales Maximum besitzt. Andererseits ist $S(a,b)$ durch Null nach unten beschränkt, sollte also mindestens ein lokales Minimum haben. □

Aufgabe 3.1 *Berechnen Sie für die Wertepaare*

x_i	1	2	3	4
y_i	6	7	9	10

die Regressionsgerade.

Lösung: Es bietet sich das folgende Vorgehen an. Zunächst bestimmen wir alle Koeffizienten der beiden linearen Gleichungen:

	x_i	y_i	x_i^2	$x_i y_i$
	1	6	1	6
	2	7	4	14
	3	9	9	27
	4	10	16	40
$\sum$	10	32	30	87

Die beiden linearen Gleichungen sind also

$$87 \;=\; 30a + 10b \qquad \text{und} \qquad 32 \;=\; 10a + 4b$$

und deren Lösung ist $\hat{a} = 1.4$ und $\hat{b} = 4.5$. Somit ergibt sich die gesuchte Regressionsgerade: $\hat{y} = 1.4x + 4.5$.

Vergleich:

x_i	1	2	3	4
y_i	6	7	9	10
$\hat{y}_i = 1.4x_i + 4.5$	5.9	7.3	8.7	10.1
e_i	0.1	-0.3	0.3	-0.1

3.2 Beurteilung des Regressionsmodells

Mit dem im letzten Abschnitt hergeleiteten Verfahren kann man für **jede** Punktwolke eine Regressionsgerade bestimmen, selbst wenn aus der Punktwolke klar hervorgeht, dass zwischen den Wertepaaren kein (linearer) Zusammenhang besteht. Die Regressionsgerade (und die daraus abgeleiteten Prognosen) passt umso besser zur Punktwolke, je kleiner (in Relation zu einer geeigneten Grösse) die Summe der quadrierten Residuen, d.h. der Wert $S(\hat{a}, \hat{b})$, ist.

Wir wollen ein erstes Mass für die Güte der Anpassung der Regressionsgerade an die Punktwolke $P_i = (x_i, y_i)$ herleiten und analysieren dazu die Residuen:

$$e_i \;=\; y_i - \hat{y}_i \;=\; (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

Quadrieren und Summieren beider Seiten ergibt (das ist nicht offensichtlich, denn beim Quadrieren auf der rechten Seite muss man die binomische Formel anwenden):

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

oder

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{=} \;=\; \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{=} \;+\; \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{=}$$

$$SQ_{Total} \;=\; SQ_{Regression} \;+\; SQ_{Residual}$$

Alle drei Terme, die allgemein als Variabilitäten bezeichnet werden, sind fast als Varianzen (Streuungen) deutbar. Zumindest fehlt der Faktor $1/n$, den wir aber einfach in die Gleichung multiplizieren könnten.

Bemerkung 3.1 *Ich möchte hier nochmals darauf hinweisen, dass* **nicht**

$$[\,(y_i - \bar{y}) - (\hat{y}_i - \bar{y})\,]^2 \;=\; (y_i - \bar{y})^2 - (\hat{y}_i - \bar{y})^2$$

gilt! Dieser Typ von Fehler tritt sehr häufig auf und wird von mir meistens als **naives Quadrieren (von Summen oder Differenzen)** *bezeichnet. Natürlich muss man zunächst eine binomische Formel anwenden:*

$$[\,(y_i - \bar{y}) - (\hat{y}_i - \bar{y})\,]^2 \;=\; (y_i - \bar{y})^2 - 2(y_i - \bar{y})(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2$$

Die Gleichung beantwortet also die Frage, wieviel der (Total)Variabilität (Streuung) der Ausgangsdaten sich durch die Variabilität (Streuung) der Regressionsdaten erklären lässt. Somit liegt es wohl auch nahe, die Güte der Anpassung als Verhältnis aus der durch die Regression erklärten Variabilität zur Totalvariabilität zu erfassen.

Satz 2 *Die fundamentale Formel der Streuungszerlegung für die totale Variabilität SQ_{Total} lautet*

$$\boxed{SQ_{Total} \;=\; SQ_{Regression} + SQ_{Residual}}$$

Division dieser Gleichung durch SQ_{Total} ergibt sofort:

$$\frac{SQ_{Total}}{SQ_{Total}} = 1 = \frac{SQ_{Regression}}{SQ_{Total}} + \frac{SQ_{Residual}}{SQ_{Total}}$$

- $\frac{SQ_{Regression}}{SQ_{Total}}$: relativer Anteil der durch die Regression erklärten Variabilität
- $\frac{SQ_{Residual}}{SQ_{Total}}$: relativer Anteil der **nicht** durch die Regression erklärten Variabilität

Definition 3.2 *Das Bestimmtheitsmass R^2 der linearen Regression ist definiert als*

$$R^2 = \frac{SQ_{Regression}}{SQ_{Total}} = 1 - \frac{SQ_{Residual}}{SQ_{Total}}$$

Es scheint klar zu sein, dass die Regressionsgerade desto besser zur Punktwolke passt, je grösser das Bestimmtheitsmass ist. Oder genauer:

Satz 3 *Es gilt $0 \leq R^2 \leq 1$ und je grösser R^2 ist, desto stärker ist eine lineare Ursache-Wirkung-Beziehung zwischen den Merkmalen X und Y ausgeprägt. Es gilt der folgende Zusammenhang mit dem Korrelationskoeffizienten der beiden Messreihen:*

$$R^2 = kor(X,Y)^2$$

Beispiel 3.1 *Berechnen Sie für die Wertepaare*

x_i	1	2	3	4
y_i	2	4	6	8

$SQ_{Regression}$, SQ_{Total}, $SQ_{Residual}$ und R^2.

Lösung:

Regressionsgerade: $\hat{y} = 2x$ (ohne Rechnung!) also gilt stets $y_i = \hat{y}_i$

$SQ_{Total} = 20$

$SQ_{Residual} = 0$

$SQ_{Regression} = 20$

$R^2 = 1$ *perfekter linearer Zusammenhang*

Beispiel 3.2 *Berechnen Sie für die Wertepaare*

x_i	10	40	50	20
y_i	20	10	40	50

$SQ_{Regression}$, SQ_{Total}, $SQ_{Residual}$ und R^2.

Lösung:

Regressionsgerade: $\hat{y} = 30$

$SQ_{Total} = 1000$

$SQ_{Residual} = 1000$

$SQ_{Regression} = 0$

$R^2 = 0$

4 Exponentielle Modellfunktionen

Die Modellgleichung ist von der Form $y = d \cdot e^{cx}$ mit $c > 0$ und wie in den vorherigen Abschnitten sollen die Parameter c und d bestimmt werden. Durch einen einfachen Trick, genauer gesagt durch logarithmieren der Modellgleichung kann dieses Problem auf ein lineares Regressionsproblem zurück geführt werden. Ich hoffe, dass Sie jeden Umformschritt nachvollziehen können!?

$$\begin{aligned} y &= d \cdot e^{cx} \\ \Longrightarrow \ln(y) &= \ln(d \cdot e^{cx}) \\ \Longrightarrow \ln(y) &= \ln(d) + \ln(e^{cx}) \\ \Longrightarrow \ln(y) &= cx + \ln(d). \end{aligned}$$

Man könnte das wie folgt ausdrücken: Falls es zwischen x- und y-Daten einen funktionalen Zusammenhang der Gestalt $y = d \cdot e^{cx}$ gibt, muss es zwischen den x- und den $\ln(y)$-Daten einen linearen Zuammmenhang geben!

Aufgabe 4.1 *Gegeben sind die Daten*

x_i	0	1	2	3	4
y_i	3	1	0.5	0.2	0.05

Bestimmen Sie mit den Techniken der linearen Regression eine Funktion der Form $f(x) = de^{cx}$, die diese Daten gut approximiert.

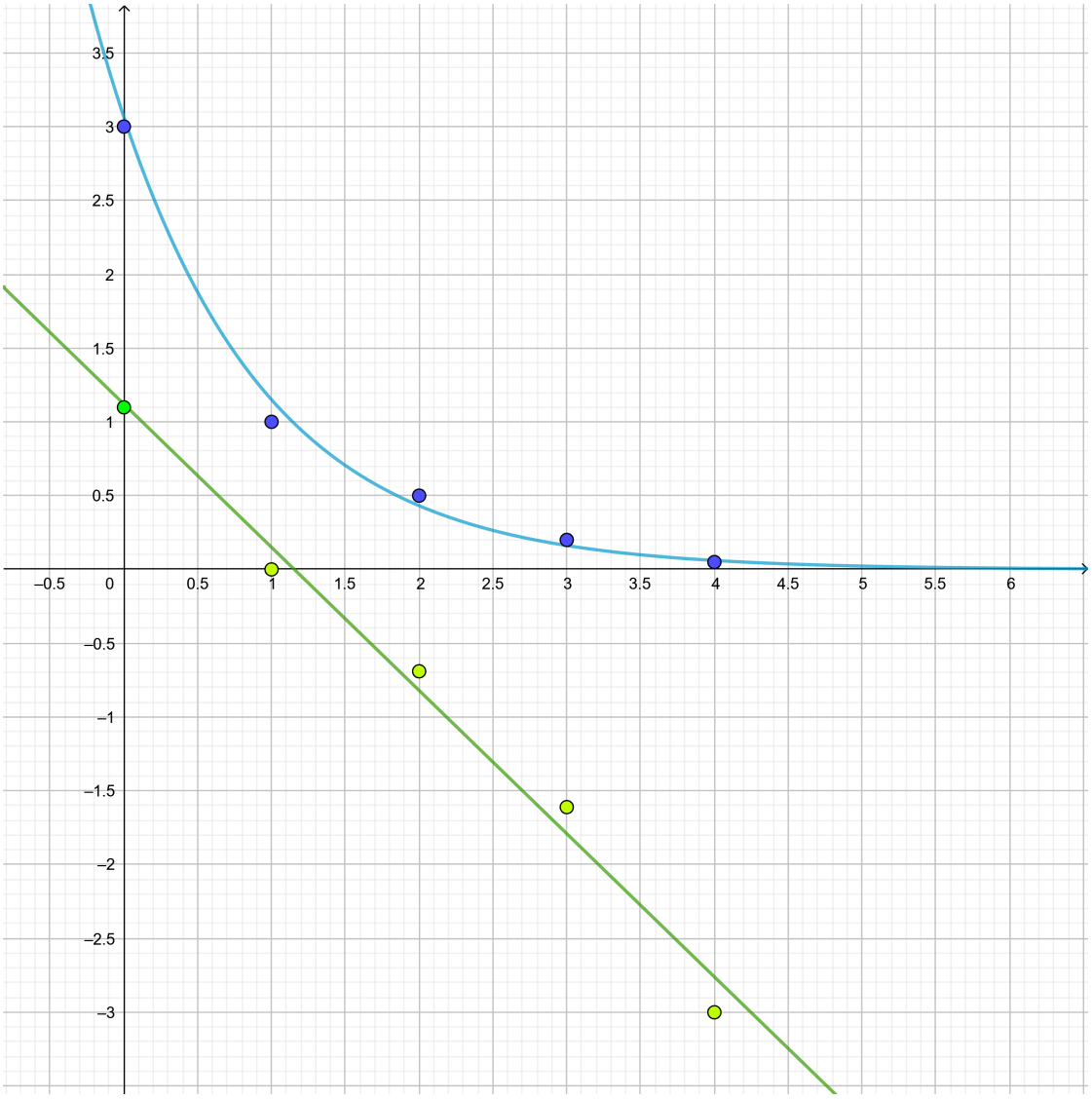
Lösung:

- $$y = f(x) = de^{cx} \quad \longrightarrow \quad \ln(y) = \ln(de^{cx}) = \ln(d) + cx = cx + \ln(d)$$

  d.h. zwischen $\ln(y_i)$ und x_i besteht ein linearer Zusammenhang

x_i	0	1	2	3	4
y_i	3	1	0.5	0.2	0.05
$\ln(y_i)$	1.0986	0	-0.6931	-1.6094	-2.9957

  In der Skizze sehen Sie die (x, y)-Punktwolke (blau) und die $(x, \ln(y))$-Punktwolke (grün).



- Löse das lineare Regressionsproblem für die x_i und $\ln(y_i)$ Daten:

	x_i	$\ln(y_i)$	x_i^2	$x_i \ln(y_i)$
	0	1.0986	1	0
	1	0	1	0
	2	−0.6931	4	−1.3862
	3	−1.6094	9	−4.8282
	4	−2.9957	16	−11.9828
$\sum$	10	−4.1996	30	−18.1972

  Die beiden linearen Gleichungen sind also

$$-4.1996 = 10a + 5b \qquad \text{und} \qquad -18.1972 = 30a + 10b$$

  und die zugehörige Regressionsgerade ist

$$\begin{aligned} \ln\hat{(y)} &= 1.1197 - 0.9798\, x \\ \ln(y) &= \ln(d) \quad + cx \end{aligned}$$

  In der zweiten Zeile steht der obige Ansatz (zum Vergleich).

- Rückrechnung:

$$\ln(d) = 1.1197 \quad \longrightarrow \quad \begin{aligned} d &= e^{1.1197} = 3.0639 \\ c &= -0.9798 \end{aligned}$$

- Regressionfunktion: $\hat{y} = 3.0639 \cdot e^{-0.9798x}$

5 Übungsaufgaben

1. Bestimmen Sie die Regressionsgeraden zu den folgenden Daten

(a)

x_i	-1	0	1	2	3
y_i	1	3.5	6.0	8.5	10

(b)

x_i	-1	0	1	2	3
y_i	1	0.5	-0.5	-2	-4

2. Es sei x_t die Anzahl der in Bayern registrierten Gästeübernachtungen im Monat t in Mio., y_t die Arbeitlosenzahl in Bayern am Ende des Monats t in 1000.

	Jan.	Feb.	März	Apr.	Mai	Juni	Juli	Aug.	Sept.	Okt.	Nov.	Dez.
x_t	4.55	4.36	4.67	5.01	5.55	6.87	6.99	7.16	7.76	6.33	5.02	4.16
y_t	431	427	428	403	423	347	337	321	349	357	366	377

   Bestimmen Sie die Regressionsgerade und beachten Sie, dass Sie mit diesen Daten schon in der letzten Übung gearbeitet haben.

3. Ein Autohersteller analysiert die Preisentwicklung (Preis y_i in $1'000.-$) im Gebrauchtwagenhandel für eines seiner Modelle in Abhängigkeit von den gefahrenen Kilometern (x_i in $1'000$ km):

x_i	10	20	50	100	150	200	250	300
y_i	40	35	28	15	10	6	4	2

   (a) Zeichnen Sie das Streudiagramm.

   (b) Ermitteln Sie mit dem Ansatz $f(x) = de^{cx}$ die Regressionsfunktion.

   (c) Welcher Preis ist für einen PKW mit $120'000$ km zu erwarten?

4. (***Zusatzaufgabe***) Gegeben sind die Daten

x_i	0	1	2	3	4
y_i	6	12	30	80	140

   Bestimmen Sie die Funktion der Form $y = f(x) = ce^x + d$, die diese Daten bestmöglich (im Sinne der kleinsten Quadrate) approximiert.

Resultate einiger Übungsaufgaben

1. (a) Das zu lösende lineare Gleichungssystem ist

$$\begin{aligned} 52 &= 15a + 5b \\ 29 &= 5a + 5b \end{aligned}$$

   Regressionsgerade: $y = 2.3\ x + 3.5$

   (b) Das zu lösende lineare Gleichungssystem ist

$$\begin{aligned} -17.5 &= 15a + 5b \\ -5 &= 5a + 5b \end{aligned}$$

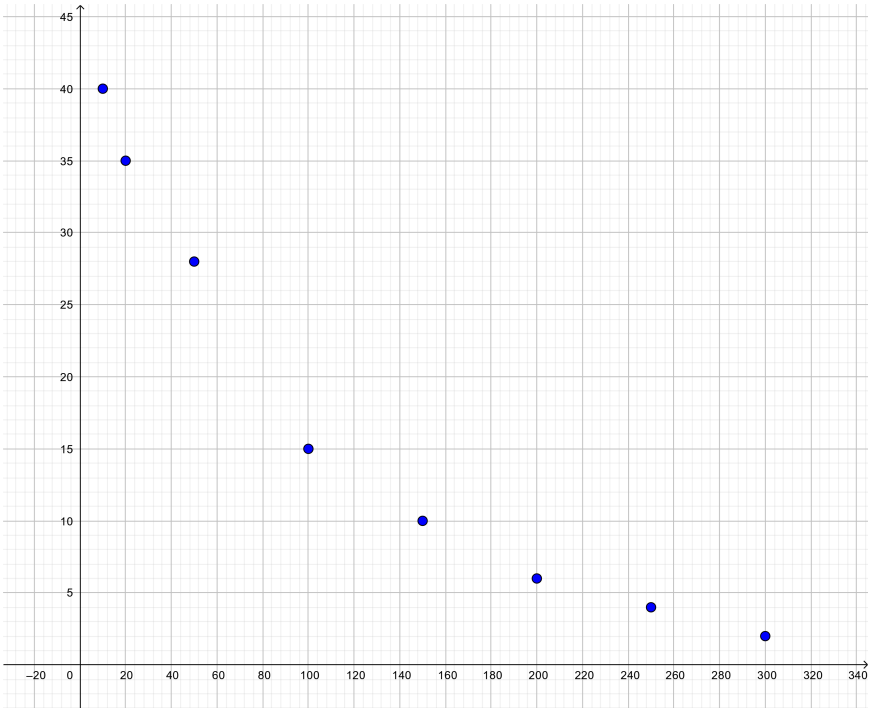
   Regressionsgerade: $y = -1.25\ x + 0.25$

2. Glücklicherweise haben wir diese Daten schon in der letzten Übung bearbeitet und kennen alle wichtigen Parameter.

   Regressionsgerade: $y = -25.34\ x + 533.94$

   Das Ergebnis kann etwas anders ausfallen, wenn Sie z.B. mit mehr als zwei Stellen nach dem Komma rechnen und erst am Ende runden. Das ist aber kein Problem und (mindestens) genauso richtig.

3. (a)



   (b) $y = f(x) = 44 \cdot e^{-0.01x}$

   (c) $f(120) = 44 \cdot e^{-0.01 \cdot 120} = 13.25$ oder $13'250,-$ CHF

4. Modellfunktion: $y = f(x) = ce^x + d$

   Straffunktion:

$$S(c,d) = \sum_{i=1}^{5} e_i^2 = \sum_{i=1}^{5} (y_i - f(x_i))^2 = \sum_{i=1}^{5} (y_i - ce^{x_i} - d)^2$$

   Notwendige Bedingungen für ein lokales Minimum:

$$\begin{aligned}
0 = \frac{\partial}{\partial c} S(c,d) &\iff 0 = 2\sum_{i=1}^{5} (y_i - ce^{x_i} - d) \cdot (-e^{x_i}) \\
&\iff 0 = \sum_{i=1}^{5} (y_i e^{x_i} - ce^{2x_i} - de^{x_i}) \\
&\iff 0 = \sum_{i=1}^{5} y_i e^{x_i} - c\sum_{i=1}^{5} e^{2x_i} - d\sum_{i=1}^{5} e^{x_i} \\
&\iff \sum_{i=1}^{5} y_i e^{x_i} = c\sum_{i=1}^{5} e^{2x_i} + d\sum_{i=1}^{5} e^{x_i} \\
0 = \frac{\partial}{\partial d} S(c,d) &\iff 0 = -2\sum_{i=1}^{5} (y_i - ce^{x_i} - d) \\
&\iff 0 = \sum_{i=1}^{5} (y_i - ce^{x_i} - d) \\
&\iff 0 = \sum_{i=1}^{5} y_i - c\sum_{i=1}^{5} e^{x_i} - 5d \\
&\iff \sum_{i=1}^{5} y_i = c\sum_{i=1}^{5} e^{x_i} + 5d
\end{aligned}$$

   Nun berechnen wir aus den Datenpaaren die Einträge dieses linearen Gleichungssystems für die gesuchten Unbekannten c und d.

x_i	y_i	e^{x_i}	$e^{2x_i} = (e^{x_i})^2$	$y_i e^{x_i}$
0	6	$e^0 = 1$	$e^0 = 1$	$6e^0 = 6$
1	12	$e^1 \approx 2.72$	$e^2 \approx 7.39$	$12e^1 \approx 32.62$
2	30	$e^2 \approx 7.39$	$e^4 \approx 54.60$	$30e^2 \approx 221.7$
3	80	$e^3 \approx 20.08$	$e^6 \approx 403.43$	$80e^3 \approx 1606.4$
4	140	$e^4 \approx 54.6$	$e^8 \approx 2980.96$	$140e^4 \approx 7644$
10	268	85.79	3447.38	9510.72

   Wir erhalten also das Gleichungssystem

$$\begin{aligned}
3447.38 \cdot c + 85.79 \cdot d &= 9510.72 \\
85.79 \cdot c + 5 \cdot d &= 268
\end{aligned}$$

   mit den beiden Lösungen $c \approx 2.486$ und $d \approx 10.92$. Die gesuchte Regressionsfunktion ist also $f(x) = 2.486e^x + 10.92$.