

3 Regression

Oft interessiert uns der Zusammenhang zwischen zwei Merkmalen X und Y einer Einheit, etwa, wie Gewicht und Größe von Menschen voneinander abhängen. Einfach zu sagen, je größer, desto schwerer genügt wohl nicht, obwohl dies eine Hypothese sein könnte. Der einfachste Zusammenhang, den wir kennen, ist der lineare, d. h., eine Größe wächst oder fällt linear mit der anderen, etwa

$$Y = a + bX.$$

Als lineare Regressionsaufgabe, d. h., Regression von Y auf X , ergibt sich nun:

$(X, Y) \sim$ gemeinsame Verteilung

Finde Konstanten a und b so, dass $E(Y - (a + bX))^2$ minimal ist.

Wir sprechen von Einfachregression, im Unterschied zur multiplen oder Mehrfachregression, bei der die Beziehung zwischen drei oder mehr Variablen erklärt werden soll.

BEMERKUNG 3.1

Die Lösung der Regression erhalten wir durch folgende Optimierungsaufgabe

$$\begin{aligned} \frac{\partial}{\partial a} E(Y - (a + bX))^2 &= \\ &= 2E(Y - (a + bX)) = 0 \\ \Leftrightarrow E(Y) - a - bE(X) &= 0 \\ \Leftrightarrow a + bE(X) &= E(Y) \end{aligned} \tag{I}$$

$$\begin{aligned} \frac{\partial}{\partial b} E(Y - (a + bX))^2 &= \\ &= E(X(Y - (a + bX))) = 0 \\ \Leftrightarrow E(X \cdot Y) - aE(X) - bE(X^2) &= 0 \\ \Leftrightarrow aE(X) + bE(X^2) &= E(XY) \end{aligned} \tag{II}$$

$$aE(X) + b[E(X)]^2 = E(X) \cdot E(Y) \tag{I} \cdot E(X)$$

$$b[E(X^2) - [E(X)]^2] = E(X \cdot Y) - E(X) \cdot E(Y) \tag{II} - (I)$$

Lösung:
$$b = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$
$$a = E(Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} E(X)$$

Die Gleichungen (I) und (II) heißen auch *Normalgleichungen*.

BEMERKUNG 3.2 (MSE)

Wir sprechen auch davon, dass die Summe der Abweichungsquadrate minimiert wird, d. h., der sogenannte MSE (*mean square error*) wird minimiert. Daher sprechen wir auch von der sogenannten *Least squares-Approximation*.

DEFINITION 3.1 (REGRESSION)

Sind X und Y identisch verteilt, so ist die *Regressionsgerade*

$$\hat{Y} = a + bX$$

gegeben durch den *Regressionskoeffizienten* $b = b_{Y;X}$ (*regression coefficient*) und den *Interzept* $a = a_{Y;X}$ (*intercept*). X heißt auch unabhängige Variable, Regressor oder erklärende Variable, Y abhängige Variable, Regressand oder zu erklärende Variable.

$$b = b_{Y;X} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad a = a_{Y;X} = E(Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} E(X).$$

Die \hat{Y} heißen *Vorhersagen* (*predicted values*).

Die Fehler $e_i = \hat{Y}_i - Y_i$ heißen *Residuen* (es gilt: $\sum_{i=1}^n e_i = 0$).

Für eine Implementation ist die Vektorschreibweise sinnvoll:

$$x = (x_1, \dots, x_n)^t, \quad y = (y_1, \dots, y_n)^t \quad \text{und} \quad \tilde{x} = x - \bar{x} \cdot (1, \dots, 1)^t, \quad \tilde{y} = y - \bar{y} \cdot (1, \dots, 1)^t$$

$$\Rightarrow \quad b_{Y;X} = \frac{\tilde{x}^t \cdot \tilde{y}}{\|\tilde{x}\|^2}, \quad a_{Y;X} = \bar{y} - b\bar{x}.$$

BEISPIEL 3.1 (GRÖSSE UND GEWICHT)

Größe X	171	168	182	176	163	$\bar{X} = 172$
Gewicht Y	72	60	87	83	73	$\bar{Y} = 75$
$\tilde{X} = X - \bar{X}$	-1	-4	10	4	-9	$\text{Var}(X) = 53.5$
$\tilde{Y} = Y - \bar{Y}$	-3	-15	12	8	-2	$\text{Var}(Y) = 111.5$

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{4} ((-1)(-3) + (-4)(-15) + (10)(12) + (4)(8) + (-9)(-2)) \\ &= \frac{1}{4} \cdot 233 = 58.25 \end{aligned}$$

$$\text{Corr}(X, Y) = \frac{58.25}{7.31 \cdot 10.559} = 0.75419$$

$$\rho_{XY}^2 = \text{Corr}(X, Y)^2 = 0.5688$$

$$b_{Y;X} = \frac{58.25}{53.5} = 1.0888 \quad \text{und} \quad a_{Y;X} = 75 - 1.0888 \cdot 172 = -112.2710$$

$$\hat{Y} = -112 + 1.088 X$$

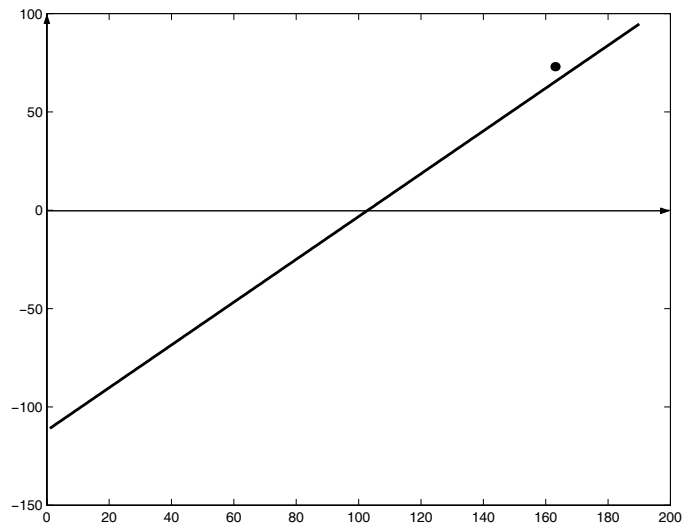


Abb. 1: Regressionsgerade, Erklärung siehe Text

Wenn wir nun die Regressionsgerade betrachten, fällt uns auf, dass

$$0 = -112 + 1.088X,$$

d. h., bei einer Größe von 102.94 cm würden 0 kg als Gewicht vorausgesagt!

Wir sehen also, dass dies nicht stimmen kann. Einerseits liegen für die lineare Regression in diesem Bereich keine Daten vor, d. h., wir haben extrapoliert, etwas das bei der Regression meistens nicht funktioniert, andererseits muss der Trend der Daten nicht linear sein. Im Abschnitt 3.1 werden wir darauf näher eingehen, indem wir Vertrauensbereiche für die Schätzungen berechnen und kurz die Residualanalyse betrachten. \square

BEISPIEL 3.2 (BREAKDOWN)

Betrachten wir den einfachen Datensatz

X	1	2	3	4	5
Y	1	2	3	4	5

so ist der lineare Zusammenhang $\hat{Y} = 0 + 1 \cdot X$ offensichtlich.

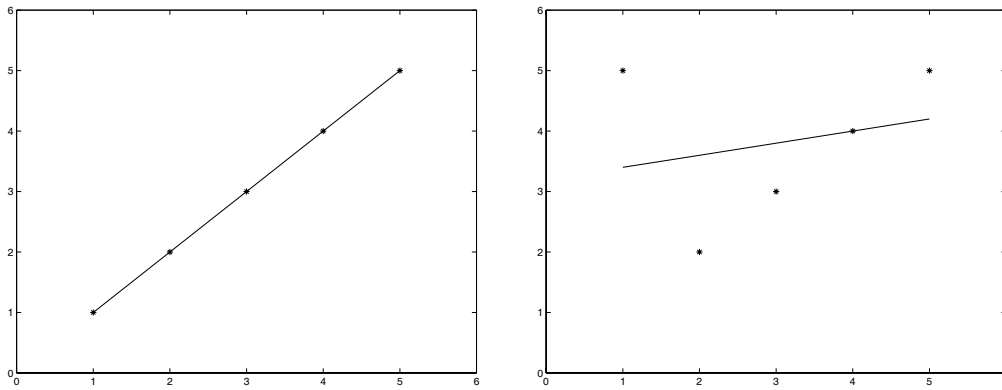


Abb. 2: Breakdown bei Regression

Verändern wir jedoch nur einen Wert,

X	1	2	3	4	5
Y	5	2	3	4	5

so erhalten wir die völlig andere Regressionsgerade $\hat{Y} = 3.2 + 0.2 \cdot X$. Wir sprechen von einem sogenannten *Regression Breakdown*, oder davon, dass die lineare Regression einen *Breakdown* Wert von $\frac{1}{n}$ hat, oder, mit $n \rightarrow \infty$, einen Breakdown Wert von 0% hat. Es genügt also ein einziger falscher Wert, um die Regression zusammenbrechen zu lassen.

Wir betrachteten bisher die Regression von Y auf X . Wir können natürlich auch die Rollen von erklärender Variable x und zu erklärender Variable y vertauschen, und die Regression von X auf Y betrachten.

BEISPIEL 3.3 (GRÖSSE UND GEWICHT)

Vertauschen wir nun die Regressionsvariablen, belassen aber die Bezeichnungen, so erhalten wir:

$b_{X;Y} = \frac{46.6}{89.2} = 0.5224$ und $a_{X;Y} = 172 - 0.5224 \cdot 75 = 132.8184$ und damit

$$\hat{X} = 132.8184 + 0.5224 \cdot Y \quad \text{oder} \quad Y = \frac{\hat{X} - 132.8184}{0.5224} = 1.9142\hat{X} - 254.2466.$$

Beide Regressionsgeraden gehen durch den Schwerpunkt (\bar{X}, \bar{Y}) .

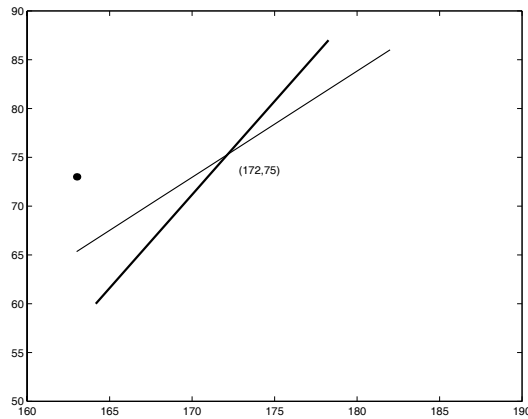


Abb. 3: Regression von Y auf X (dünn) und umgekehrt (dick).

Diese Gerade passt anscheinend besser zu den Werten. □

Aus $b_{Y;X} \cdot b_{X;Y} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \cdot \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} = [\text{Corr}(X, Y)]^2$ folgt, dass die Korrelation das geometrische Mittel aus den beiden Regressionskoeffizienten ist, d. h.,

$$\text{Corr}(X, Y) = \sqrt{b_{Y;X} b_{X;Y}} .$$

Weiters ist

$$b_{Y;X} = \text{Corr}(X, Y) \cdot \frac{\sqrt{\text{Var}(X)}}{\sqrt{\text{Var} Y}}$$

d. h., aus der Korrelation kann der Steigungsparameter $b_{Y;X}$ der Regression geschätzt werden.

Die Korrelation $\text{Corr}(X, Y)$ wird im Zusammenhang mit der Regression oft auch Korrelationskoeffizient genannt und mit ρ oder ρ_{XY} bezeichnet.

$$B_{XY} = \rho^2 = \text{Corr}^2(X, Y) = \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = \frac{s_{\hat{Y}}^2}{s_Y^2}$$

ist dann das *Bestimmtheitsmaß* (*coefficient of determination*), das 0 ist, falls kein Zusammenhang zwischen den Variablen besteht, und 1, falls ein linearer Zusammenhang besteht. $\rho^2 = 0$ bedeutet, dass die Regressionsgeraden für \hat{y} und \hat{x} achsenparallel sind und normal aufeinander stehen, $\rho^2 = 1$, dass diese übereinstimmen, und $0 < \rho^2 < 1$, dass diese eine Schere bilden (siehe Beispiel 3.3). $\rho = \text{Corr}$ ist der Cosinus des Winkels α , den die beiden Geraden zueinander haben.

Interpretation:

$$1 - \rho^2 = \frac{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{Var}(Y) - \sigma_{\hat{Y}}^2}{\text{Var}(Y)}$$

gibt den Anteil der Varianz von Y an, der durch die Regressionsbeziehung nicht erklärt wird.

BEMERKUNG 3.3

Um die Nachteile der linearen Regression bezüglich des MSE zu umgehen, gibt es verschiedene andere Regressionsverfahren, die auf einer geraden (symmetrischen) Funktion $r(x)$ basieren und $\sum_i r(e_i)$ minimieren. Es seien nur RM-Schätzer (*repeated median*), LMS-Schätzer (*least median of squares*) oder LTS-Schätzer (*least trimmed squares*) genannt, die alle einen Breakdown Wert von 50% haben, d. h., selbst dann noch funktionieren, wenn fast die Hälfte der Datensätze falsch ist.

BEMERKUNG 3.4

Viele Größen haben keinen linearen Zusammenhang, etwa Geschwindigkeit und Bremsweg (quadratischer Z.). Hier helfen wir uns, indem wir zuerst die Variablen so transformieren, dass ein linearer Zusammenhang möglich ist (etwa logarithmieren der Daten bei exponentiellen Wachstumsprozessen), dann eine lineare Regression durchführen, und das Ergebnis rücktransformieren.

Folgende Tabelle enthält einige Transformationen, die notwendig sind, um $\hat{y} = a + bx$ über die entsprechend transformierten Variablen y' und x' und die zu diesen gehörigen Regressionskoeffizienten a' und b' aus $\hat{y}' = a' + b'x'$ zu berechnen (Quellen: Hartung, Statistik).

Zusammenhang	Transformation		Rücktransformation	
	$y' =$	$x' =$	$a' =$	$b' =$
$y = a + \frac{b}{x}$	y	$\frac{1}{x}$	a	b
$y = \frac{a}{b + x}$	$\frac{1}{y}$	x	$\frac{b}{a}$	$\frac{1}{a}$
$y = \frac{1}{a + bx}$	$\frac{1}{y}$	x	a	b
$y = \frac{x}{a + bx}$	$\frac{x}{y}$	x	a	b
$y = ab^x$	$\log y$	x	$\log a$	$\log b$
$y = ax^b$	$\log y$	$\log x$	$\log a$	b
$y = ae^{bx}$	$\log y$	x	$\log a$	b
$y = ae^{b/x}$	$\log y$	$\frac{1}{x}$	$\log a$	b
$y = \frac{1}{a + be^{-x}}$	$\frac{1}{y}$	$\exp x$	a	b
$y = a + bx^n$	y	x^n	a	b

3.1 Test der Regressionsparameter

Nach einer kurzen Wiederholung der Einfachregression und einer leichten Erweiterung der Notation dieses Schätzers werden wir die geschätzten Parameter testen und Konfidenzintervalle für diese angeben und berechnen.

Wir haben Werte $Y_1, \dots, Y_n \sim N(a + bX_i, \sigma^2)$ i.i.d., d. h.,

$$\hat{Y} = a + bX.$$

Wir kennen bereits die folgenden Schätzwerte für die unbekannt Parameter Regressionskoeffizient b und Interzept a

$$\hat{b} = b_{Y,X} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{a} = a_{Y,X} = \bar{Y} - b_{Y,X} \cdot \bar{X} = E(Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \cdot E(X)$$

Außerdem kennen wir den Korrelationskoeffizienten $\rho = \rho_{X,Y}$ bzw. das Bestimmtheitsmaß ρ^2 (siehe Seite 5).

$$\rho \doteq \text{Corr}(X, Y) = \frac{\text{Cov}(Y, X)}{\sqrt{\text{Var}(Y)} \sqrt{\text{Var}(X)}}.$$

Als Schätzer erhalten wir

$$\hat{\rho} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \cdot \sum_{i=1}^n (X_i - \bar{X})^2}}$$

3.1.1 Die Regressionsparameter

Es wird angenommen, dass die Residuen e_i unabhängig nach $N(0, \sigma^2)$ verteilt sind. Dann sind auch $\hat{a} \sim N(a, \sigma_a^2)$ und $\hat{b} \sim N(b, \sigma_b^2)$ normalverteilt.

Der Schätzer für die Varianz der Residuen ist

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y})^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - a - bX_i)^2$$

BEMERKUNG 3.5

Aus s_e^2 können wieder die Parameter \hat{a} und \hat{b} berechnet werden.

$$\frac{\partial s_e^2}{\partial b} = -2 \sum X_i(Y_i - a - bX_i) = 0 \Rightarrow \hat{a} = \bar{Y} - \hat{b}\bar{X}$$

$$\frac{\partial s_e^2}{\partial a} = -2 \sum (Y_i - a - bX_i) = 0 \Rightarrow \hat{b} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Ein **Konfidenzintervall für σ^2** ergibt sich zu

$$\hat{I} = \left[\frac{(n-2)s_e^2}{\chi_{n-2,1-\alpha/2}^2}; \frac{(n-2)s_e^2}{\chi_{n-2,\alpha/2}^2} \right].$$

Die Varianz der Residuen hängt wiederum mit dem Bestimmtheitsmaß zusammen.

THEOREM 3.1

Es gilt:

$$s_e^2 = \hat{\sigma}^2 = [1 - \rho^2] \frac{1}{n-2} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

BEWEIS.

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}X_i)^2 \\ &= \frac{1}{n-2} \sum_{i=1}^n [Y_i - \bar{Y} - \hat{b}(X_i - \bar{X})]^2 \\ &= \frac{1}{n-2} \sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{2}{n-2} \hat{b} \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) + \frac{1}{n-2} \hat{b}^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n-2} \sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{2}{n-2} \frac{[\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})]^2}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{1}{n-2} \frac{[\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})]^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{1}{n-2} \sum_{i=1}^n (Y_i - \bar{Y})^2 [1 - \rho^2] \end{aligned}$$

□

Wir wollen nun testen, ob überhaupt ein Zusammenhang zwischen den Zufallsvariablen X_i und Y_i besteht, oder ob diese unabhängig sind. Dazu testen wir, ob der Regressionskoeffizient b signifikant von 0 verschieden ist oder nicht.

$$H_0 : b = 0 \qquad H_A : b \neq 0$$

Da $b = \rho \cdot \frac{\sqrt{\text{Var}(X)}}{\sqrt{\text{Var}(Y)}}$ ist, können wir für dieses Problem denselben Test wie für das Prüfen auf Unabhängigkeit verwenden (F -Test).

F-Test auf Unabhängigkeit (verb. Stichproben)

Seien $(X_1, Y_1) \dots (X_n, Y_n)$ Paare von Beobachtungen, wobei die $X_i \sim N(\mu_X, \sigma_X^2)$ und die $Y_i \sim N(\mu_Y, \sigma_Y^2)$ jeweils i.i.d. sind.

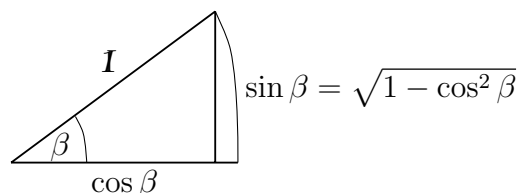
Der *Korrelationskoeffizient*

$$\rho_{XY} = \text{Corr}(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}$$

hängt nicht mehr von $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, sondern nur mehr von n ab.

Wie sieht die Verteilung von $\rho = \cos \beta$ aus?

Dazu führen wir die Cotangententransformation



$$\cot \beta = \cot(\arccos \rho) = \frac{\cos \beta}{\sqrt{1 - \cos^2 \beta}} = \frac{\rho}{\sqrt{1 - \rho^2}}$$

durch.

THEOREM 3.2

Falls X_i und Y_i unabhängig sind, so gilt:

$$\sqrt{n-2} \frac{\rho}{\sqrt{1-\rho^2}} \sim t(n-2) - \text{Verteilung}$$

oder

$$(n-2) \frac{\rho^2}{1-\rho^2} \sim F(1, n-2) - \text{Verteilung}$$

◇

Daraus entwickeln wir den F -Test auf Unabhängigkeit der beiden Stichproben

F-TEST auf Unabhängigkeit

Voraussetzung: $X_1, \dots, X_n \sim N(\mu_x, \sigma_x^2)$ i.i.d.
 $Y_1, \dots, Y_n \sim N(\mu_y, \sigma_y^2)$ i.i.d.

Testgröße: $T = (n-2) \frac{\rho^2}{1-\rho^2} \sim F(1, n-2)$

Hypothesen: $H_0 : X_i, Y_i$ unabhängig, $H_A : X_i, Y_i$ abhängig
 H_0 ablehnen, falls $T > F_{1, n-1; 1-\alpha}$

Wir testen also, ob der Regressionskoeffizient b signifikant von 0 verschieden ist oder nicht.

$$H_0 : b = 0 \qquad H_A : b \neq 0$$

Unsere Testgröße ist:

$$T = \sqrt{n-2} \frac{\rho}{\sqrt{1-\rho^2}} \sim t(n-2) \quad \text{oder} \quad T^2 = (n-2) \frac{\rho^2}{1-\rho^2} \sim F(1, n-2)$$

H_0 wird abgelehnt, falls $T > t_{n-2, 1-\frac{\alpha}{2}}$ oder $T^2 > F_{1, n-2; 1-\alpha}$.

Eine andere Methode stützt sich direkt auf die Verteilung des Regressionskoeffizienten. Sie dient auch gleichzeitig dazu, ein Konfidenzintervall für b anzugeben.

Regressionskoeffizient b

Varianz $s_b^2 = s_e^2 \frac{1}{\sum (X_i - \bar{X})^2}$

Testgröße $\frac{\hat{b} - b_0}{s_b} \sim t(n-2)$

Kritischer Wert zweiseitig: $c_\alpha = t_{n-2, 1-\frac{\alpha}{2}}$ einseitig: $c_\alpha = t_{n-2, 1-\alpha}$

Konfidenz-I. $\hat{I} = [\hat{b} - s_b t_{n-2, 1-\frac{\alpha}{2}}; \hat{b} + s_b t_{n-2, 1-\frac{\alpha}{2}}]$

Interzept a

Varianz $s_a^2 = s_e^2 \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}$

Testgröße $\frac{\hat{a} - a_0}{s_a} \sim t(n-2)$

Kritischer Wert zweiseitig: $c_\alpha = t_{n-2, 1-\frac{\alpha}{2}}$ einseitig: $c_\alpha = t_{n-2, 1-\alpha}$

Konfidenz-I. $\hat{I} = [\hat{a} - s_a t_{n-2, 1-\frac{\alpha}{2}}; \hat{a} + s_a t_{n-2, 1-\frac{\alpha}{2}}]$

Ein äquivalenter zweiseitiger Test für den Interzept gegen 0 ist folgender

$$H_0 : a = 0 \qquad H_A : a \neq 0$$

Mit der F -verteilten Testgröße

$$T = n \frac{\hat{a}^2}{s_a^2} \sim F(1, n-2)$$

der sich direkt aus obigem zweiseitigen Test ableitet.

H_0 wird abgelehnt, falls $T > F_{1, n-2; 1-\alpha}$ ist.

BEISPIEL 3.4 (DÜNGEMITTEL)

Zusammenhang zwischen Düngemittelleinsatz X und Ertrag Y

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$
5.2	8.1	0	0.1
4.7	7.6	-0.5	-0.4
6.2	8.4	1.0	0.4
3.9	7.7	-1.3	-0.3
6.0	8.2	0.8	0.2
$\bar{X} = 5.2$	$\bar{Y} = 8.0$	$\text{Var}(X) = 0.895$	$\text{Var}(Y) = 0.115$

$$\sum(X_i - \bar{X})^2 = 3.58, \quad \sum(Y_i - \bar{Y})^2 = 0.46$$

$$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = (0 + 0.2 + 0.4 + 0.39 + 0.16) = 1.15 \Rightarrow \text{Cov}(X, Y) = 0.2875$$

$$\rho = \frac{1.15}{\sqrt{3.58 \cdot 0.46}} = 0.8961 \quad \rho^2 = \frac{0.2875}{\sqrt{0.895 \cdot 0.115}} = 0.80307$$

$$b = \frac{1.15}{3.58} = \frac{0.2875}{0.895} = 0.3212 \quad a = 8.0 - 0.3212 \cdot 5.2 = 6.3296$$

Die Regressionsgleichung lautet daher

$$\hat{Y} = 6.3296 + 0.3212 \cdot X$$

X_i	Y_i	\hat{Y}_i	$e_i = Y_i - \hat{Y}_i$
5.2	8.1	8.0	0.10
4.7	7.6	7.83	-0.2394
6.2	8.4	8.32	0.07877
3.9	7.7	7.58	0.1176
6.0	8.2	8.26	-0.05698

Die Summe der Residuen $Y_i - \hat{Y}_i$ muss gleich 0 sein, Rundungsungenauigkeiten können aber auftreten.

Die Varianz der Residuen ist $s_e^2 = \frac{1}{3} \sum(Y_i - \hat{Y}_i)^2 = 0.0153$.

Wir testen, ob $b = 0$ (d. h., $H_0 : b = 0$, $H_A : b \neq 0$)

$$T = \sqrt{n-2} \frac{\rho}{\sqrt{1-\rho^2}} = \sqrt{3} \frac{0.896}{\sqrt{1-0.8}} = 3.46 < t_{3;0.99} = 4.5404,$$

daher kann H_0 nicht abgelehnt werden, d. h., b ist nicht signifikant von 0 verschieden, es ist daher anzunehmen, dass X_i und Y_i unabhängig sind.

Wir testen nun, ob $a = 0$ (d. h., $H_0 : a = 0$, $H_A : a \neq 0$)

$$T = 5 \frac{(6.3296)^2}{0.0153} = 13064,294 \gg F_{1,3;0.99} = 34.116,$$

daher wird H_0 abgelehnt. a ist hochsignifikant von 0 verschieden. \square

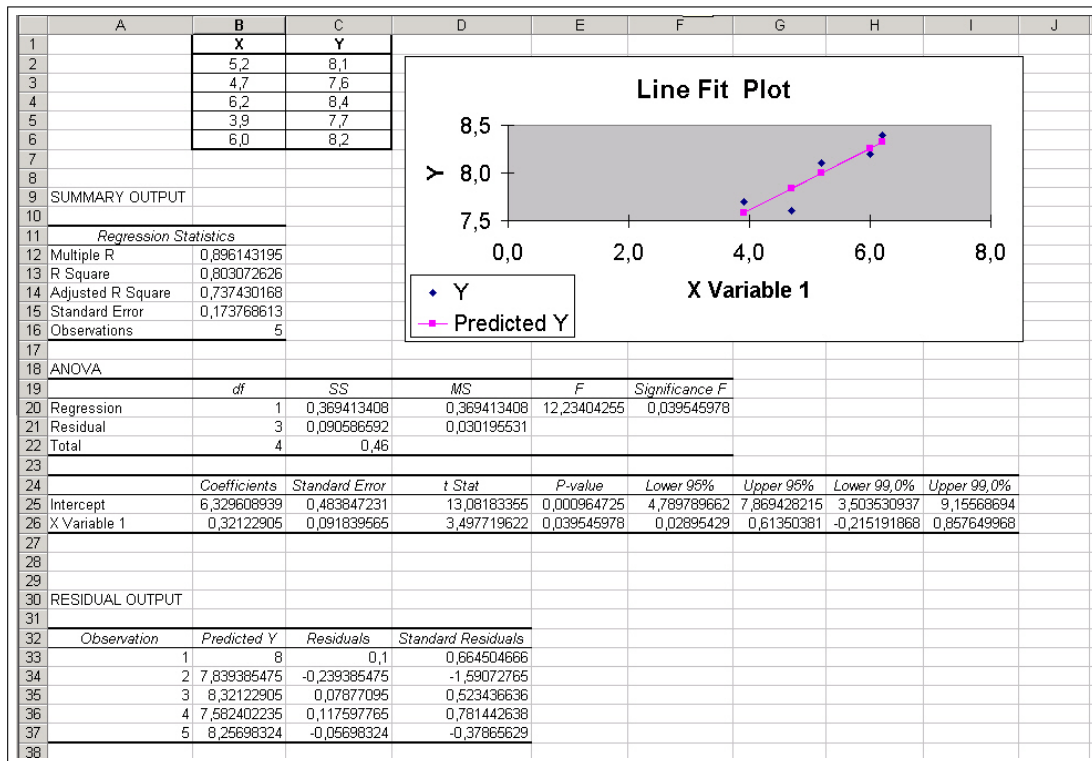


Abb. 4: Spreadsheet: Lineare Einfachregression

3.1.2 Prognoseintervalle

Wie wir Konfidenzintervalle für die Regressionsparameter angeben können, so können wir auch Konfidenz- und Prognoseintervalle für die Prognosen angeben.

Prognoseintervall für Y_0 :

$$s_Y^2 = s_e^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right)$$

$$\hat{I} = \left[\hat{Y}_0 - s_Y t_{n-2, 1-\frac{\alpha}{2}}; \hat{Y}_0 + s_Y t_{n-2, 1-\frac{\alpha}{2}} \right]$$

Konfidenzintervall für $E(Y_0)$:

$$s_{EY}^2 = s_e^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right)$$

$$\hat{I} = \left[\hat{Y}_0 - s_{EY} t_{n-2, 1-\frac{\alpha}{2}}; \hat{Y}_0 + s_{EY} t_{n-2, 1-\frac{\alpha}{2}} \right]$$

Wir sehen leicht, dass das Prognoseintervall für Y_0 größer ist, als das das Konfidenzintervall für $E(Y_0)$.

Simultanes Konfidenzintervall das zugleich an allen Stellen X_0 den Erwartungswert $E(Y|X_0)$ von Y mit der Wahrscheinlichkeit $1 - \alpha$ überdeckt.

$$Y_0 \pm \sqrt{2s_e^2 F_{2, n-2; 1-\alpha} \left(\frac{1}{n} + \frac{(X - X_0)^2}{\sum_i (X_i - \bar{X})^2} \right)}$$

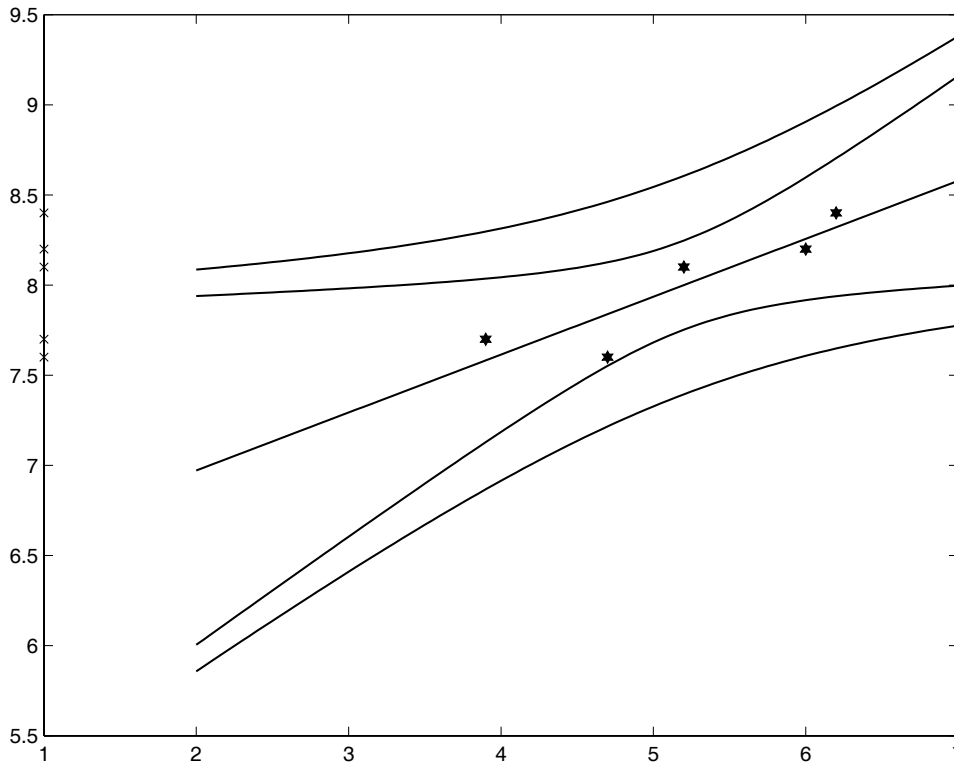


Abb. 5: Konfidenz- und Prognoseintervalle bei der Regression.

3.1.3 Regression durch einen gegebenen Punkt

Oft ist es notwendig, dass die Regressionsgerade durch einen bestimmten Punkt (X_0, Y_0) verläuft (oft ist dies der Ursprung). Wir wissen, dass die Regressionsgerade immer durch den Schwerpunkt (\bar{X}, \bar{Y}) verläuft, wir brauchen daher in den Regressionsgleichungen nur die Koordinaten der Mittelwerte durch die des Fixpunktes (X_0, Y_0) zu ersetzen, um das Gewünschte zu erhalten.

Wir haben daher

$$\hat{b} = \frac{\sum (Y_i - Y_0)(X_i - X_0)}{\sum (X_i - X_0)^2} \quad \text{und} \quad \hat{a} = Y_0 - \hat{b}X_0.$$

3.1.4 Residualanalyse

Wir müssen überprüfen, ob unsere Annahme, dass die Residuen $N(0, \sigma_e^2)$ -verteilt sind, auch stimmt – auf ihr basieren alle Tests der Parameter. Dazu bedienen wir uns standardisierter Tests.

Oft genügt auch schon ein sogenannter *Indexplot* der Residuen – die Residuen $e_i = \hat{Y} - Y_i$ werden gegen ihren Index i geplottet – um zu sehen, ob die Residuen nicht normalverteilt sind, sondern dass es gewisse Regelmäßigkeiten gibt. Dies lässt auf einen anderen als einen

linearen Zusammenhang schließen. Oft hilft dann eine entsprechende Transformation der Daten.

Weiters hilft der Indexplot, Ausreißer zu eliminieren. Wenn wir die standardisierten Residuen $d_i = \frac{e_i}{s_e}$ verwenden, so können etwa Punkte mit $|d_i| > 3$ als Ausreißer eliminiert werden, sie liegen „zu weit von der Geraden weg“.

3.2 Mehrfachregression

Dieser Abschnitt wurde bereits ausführlich in der Mathematik 1 (Abschnitt Regression) behandelt und kann somit als bekannt vorausgesetzt werden. Er wird hier nur nochmals wiedergegeben, um ein vollständiges Skriptum zu liefern.

Seien n Datensätze $(y^{(i)}, x_1^{(i)}, \dots, x_k^{(i)})$, $i = 1, \dots, n$, gegeben, mit $k < n$, wobei die erste Koordinate $y^{(i)}$ jeweils von den restlichen Einträgen, den sogenannten unabhängigen Variablen, $x_1^{(i)}, \dots, x_k^{(i)}$ abhängt.

Es soll eine lineare Approximation in $x_j^{(i)}$ gefunden werden, sodass $y^{(i)}$ erklärt werden kann, d. h.,

$$y^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_k x_k^{(i)}, \quad \forall i = 1, \dots, n,$$

und gleichzeitig der Fehler der Approximation (Abstand von den tatsächlichen Werten) minimiert wird, d. h.

$$\sum_{i=1}^n [y^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_k x_k^{(i)})]^2 \rightarrow \min .$$

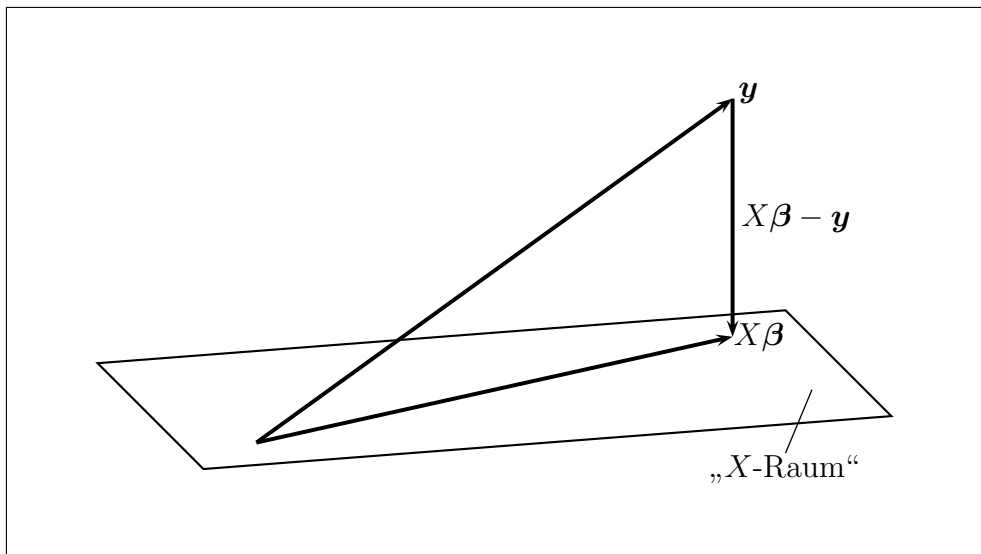
Das ist der sogenannte *kleinste Quadrate Schätzer* oder die *Gauss-Approximation* der Daten.

Wir suchen also eine Lösung des überbestimmten Gleichungssystems

$$X\boldsymbol{\beta} = \mathbf{y} \quad \text{und} \quad \|X\boldsymbol{\beta} - \mathbf{y}\|^2 \rightarrow \min$$

mit den Bezeichnungen

$$X = \begin{pmatrix} 1 & x_1^{(1)} & \cdots & x_k^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_k^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(n)} & \cdots & x_k^{(n)} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix} .$$



Da $(X\boldsymbol{\beta} - \mathbf{y})$ normal auf alle Vektoren $X\mathbf{z}$ im „ X -Raum“ steht, gilt

$$\langle X\mathbf{z}, (X\boldsymbol{\beta} - \mathbf{y}) \rangle = (X\mathbf{z})^t (X\boldsymbol{\beta} - \mathbf{y}) = 0$$

und da $(X\mathbf{z})^t = \mathbf{z}^t X^t$, gilt auch

$$(\mathbf{z}^t X^t)(X\boldsymbol{\beta} - \mathbf{y}) = \mathbf{z}^t (X^t X\boldsymbol{\beta} - X^t \mathbf{y}) = 0.$$

Da dies für alle \mathbf{z} gilt, muss der zweite Faktor gleich null sein, es gilt also

$$(X^t X)\boldsymbol{\beta} = X^t \mathbf{y}$$

bzw.

$$\boldsymbol{\beta} = (X^t X)^{-1} X^t \mathbf{y}.$$

BEMERKUNG 3.6

Wir wollten $X\boldsymbol{\beta} = \mathbf{y}$ lösen. Mit obigen Ausführungen gilt nun:

$$X\boldsymbol{\beta} = X(X^t X)^{-1} X^t \mathbf{y} \quad \text{oder} \quad X^t X\boldsymbol{\beta} = X^t X(X^t X)^{-1} X^t \mathbf{y} = X^t \mathbf{y}$$

und somit gilt auch

$$X^2 = (X(X^t X)^{-1} X^t)X = X(X^t X)^{-1}(X^t X) = X.$$

Wir sagen: \mathbf{y} wird auf $X\boldsymbol{\beta}$ projiziert, mittels der Projektion $X(X^t X)^{-1} X^t$.

DEFINITION 3.2

Eine *Projektion* P erfüllt $P^2 = P$ und damit auch $P^n = P$, $n > 0$.

Falls kein Designfehler (Messdaten) vorliegt, ist die $(k + 1) \times (k + 1)$ -Matrix $(X^t X)$ invertierbar, sodass dieses Regressionsproblem lösbar ist.

BEMERKUNG 3.7

Das Verfahren heißt lineare Regression, weil die Koeffizienten β_j Konstante sind und $y^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_k x_k^{(i)}$ eine Linearkombination von Funktionen 1 und x_j , nicht jedoch die Funktionen linear sind. Im Folgenden werden wir sehen, dass statt der Funktionen x_j beliebige Funktionen genommen werden können.

3.2.1 Moore-Penrose-Inverse oder Pseudoinverse

Wie lösen wir das System

$$(X^t X)\boldsymbol{\beta} = X^t \mathbf{y}?$$

Existiert die Inverse von $X^t X$, so ist das kein allzu großes Problem:

$$\boldsymbol{\beta} = (X^t X)^{-1} X^t \mathbf{y}.$$

Da $(X^t X)^{-1} X^t \mathbf{y}$ das System $X\boldsymbol{\beta} = \mathbf{y}$ im obigen Sinn (Gauss Approximation) löst, heißt die Matrix $(X^t X)^{-1} X^t$ die *Pseudoinverse* zu X und wird mit X^+ bezeichnet.

Die Pseudoinverse D^+ einer Diagonalmatrix D :

$$D = \begin{pmatrix} d_1 & & & & & \\ & \ddots & & & & \\ & & d_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \quad D^+ = \begin{pmatrix} \frac{1}{d_1} & & & & & \\ & \ddots & & & & \\ & & \frac{1}{d_k} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}.$$

Bei allgemeineren $m \times n$ -Matrizen A wird die Pseudoinverse über die sogenannte Singulärwertzerlegung (siehe später) berechnet.

DEFINITION 3.3

Sei A eine $m \times n$ -Matrix und ihre Singulärwertzerlegung

$$A = UDV^t.$$

Dann ist die *Pseudoinverse* oder *Moore-Penrose Inverse* A^+ gegeben durch

$$A^+ = VD^+U^t,$$

wobei die Pseudoinverse D^+ der Diagonalmatrix D wie oben gebildet wird.

Zur Berechnung der Singulärwertzerlegung siehe später.

Die Pseudoinverse A^+ erfüllt folgende Identitäten

$$AA^+A = A, \quad A^+AA^+ = A^+, \quad (AA^+)^t = AA^+, \quad (A^+A)^t = A^+A.$$

Ist die Matrix A invertierbar, so gilt $A^+ = A^{-1}$.

3.2.2 Regression bezüglich beliebiger Basisfunktionen

Es gebe wie oben einen funktionalen Zusammenhang

$$y = f(x_1, \dots, x_n) = \beta_0 g_0(x_1, \dots, x_n) + \dots + \beta_k g_k(x_1, \dots, x_n),$$

wobei die g_j bekannte Funktionen sind, und die Koeffizienten β_j wie oben konstant sind.

Wir wollen die Funktion $f(\mathbf{x}) = f(x_1, \dots, x_n)$ durch die bekannten Funktionen $g_j(\mathbf{x})$ *approximieren*. Durch Messung kennen wir die Funktionswerte von $y^{(i)} = f(\mathbf{x})$ an den Stellen $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_n^{(i)})$, $i = 1, \dots, n$. Mit den Bezeichnungen

$$X = \begin{pmatrix} g_0(\mathbf{x}^{(1)}) & g_1(\mathbf{x}^{(1)}) & \dots & g_k(\mathbf{x}^{(1)}) \\ g_0(\mathbf{x}^{(2)}) & g_1(\mathbf{x}^{(2)}) & \dots & g_k(\mathbf{x}^{(2)}) \\ \vdots & \vdots & \vdots & \vdots \\ g_0(\mathbf{x}^{(n)}) & g_1(\mathbf{x}^{(n)}) & \dots & g_k(\mathbf{x}^{(n)}) \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix},$$

analog zu oben, erhalten wir das Regressionsproblem

$$X\boldsymbol{\beta} = \mathbf{y} \quad \text{und} \quad \|X\boldsymbol{\beta} - \mathbf{y}\|^2 \rightarrow \min .$$

Falls die Lösung existiert, ist sie wiederum gegeben durch

$$(X^t X)\boldsymbol{\beta} = X^t \mathbf{y} \quad \text{bzw.} \quad \boldsymbol{\beta} = (X^t X)^{-1} X^t \mathbf{y} .$$

BEISPIEL 3.5

Sei $f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$,

so ist $g_0(\mathbf{x}) = 1$, $g_1(\mathbf{x}) = x_1$, $g_2(\mathbf{x}) = x_2$, $g_3(\mathbf{x}) = x_1^2$, $g_4(\mathbf{x}) = x_2^2$ und $g_5(\mathbf{x}) = x_1 x_2$.

BEISPIEL 3.6

Sei $f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ und

x_1	5	3	5	3
x_2	0.5	0.5	0.3	0.3
$f(x_1, x_2)$	1.5	3.5	6.2	3.2

und damit

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 1.5 \\ 3.5 \\ 6.2 \\ 3.2 \end{pmatrix} \quad X = \begin{pmatrix} 1 & 5 & 0.5 \\ 1 & 3 & 0.5 \\ 1 & 5 & 0.3 \\ 1 & 3 & 0.3 \end{pmatrix} .$$

Es folgt somit

$$\boldsymbol{\beta} = (X^t X)^{-1} X^t \mathbf{y} \quad \Rightarrow \quad \boldsymbol{\beta} = \begin{pmatrix} 7 \\ 0.25 \\ -11 \end{pmatrix} .$$

3.3 Singulärwertzerlegung

Gewisse quadratische Matrizen (symmetrische und normale) lassen sich auf Diagonalform bringen. Die Singulärwertzerlegung (*SVD-singular value decomposition*) ist eine entsprechende Darstellung für allgemeine $m \times n$ Matrizen:

$$A = U D V^t$$

$m \times n \quad m \times m \quad m \times n \quad n \times n$

mit orthogonalen Matrizen U und V und

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & \lambda_k & & & \vdots \\ \vdots & & & & 0 & & \vdots \\ \vdots & & & & & \ddots & \vdots \\ \vdots & & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix},$$

wobei $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_k|$.

Berechnet wird diese Darstellung durch Umformen von

$$A = U D V^t$$

mittels

$$A^t A = (U D V^t)^t (U D V^t) = V D^t U^t U D V^t = V D^t D V^t,$$

wobei $D^t D$ eine Diagonalmatrix mit folgender Form ist:

$$D^t D = \begin{pmatrix} \lambda_1^2 & & & \\ & \lambda_2^2 & & \\ & & \lambda_3^2 & \\ & & & \ddots \end{pmatrix}$$

und da $A^t A$ symmetrisch und V orthogonal ist, gilt weiters

$$D^t D = V^t (A^t A) V.$$

V kann somit aus den Eigenvektoren von $A^t A$ gebildet werden, die λ_i^2 sind die zugehörigen Eigenwerte α_i von $A^t A$ und somit $\lambda_i = \sqrt{\alpha_i}$.

Analog zu V erhalten wir U , da gilt

$$A A^t = (U D V^t) (U D V^t)^t = U D V^t V D U^t = U (D D^t) U^t,$$

d. h., U wird aus den Eigenvektoren von $A A^t$ zusammengesetzt.

Aufgaben zur Regressionsanalyse

- 3.1 Im Rahmen einer Studie soll der Zusammenhang zwischen diastolischem Blutdruck X (mmHg) und Herzgewicht Y (g) von 10 an Gehirnblutung verstorbenen Männern geschätzt werden.

X	121	120	95	123	140	112	92	100	102	91
Y	521	465	352	455	490	388	301	395	375	418

Berechne die Regressionsgleichung und das Bestimmtheitsmaß!

Stelle die Daten graphisch dar und zeichne die geschätzte Regressionsgerade! ◁

- 3.2 In der Grazer Universitäts-Frauenklinik wurden die Größe X und der Kopfumfang Y neugeborener Knaben gemessen.

X	51	47	52	48	52	52	50	48	54	50
Y	34	35	36	34	37	36	35	33	38	34

- Berechne die Regressionsgleichung!
- Teste, ob b signifikant von 0 verschieden ist!
- Berechne das Bestimmtheitsmaß!
- Berechne ein 95 % Konfidenzintervall für Y_0 , wenn $X_0 = 49$!
- Berechne ein 95 % Konfidenzintervall für $E(Y_0)$, wenn $X_0 = 49$!
- Vergleiche d und e !

◁

- 3.3 Nachstehende Tabelle enthält die Belastbarkeit eines Materials in Abhängigkeit seines Alters in Jahren:

Belastbarkeit y	120	110	130	140	150	110
Alter x	12	8	3	2	1	12

- Schätze die Regressionsgleichung $\hat{y} = a + bx$!
- Gib ein 95 % Konfidenzintervall für den Regressionskoeffizienten b an!
- Prüfe, ob der Stichprobenregressionskoeffizient b statistisch gegen Null gesichert ist ($\alpha = 0.05$)!
- Gib ein 95 % Konfidenzintervall für den Erwartungswert der Belastbarkeit bei einem Alter von 10 Jahren!
- Berechne den Korrelationskoeffizienten und das Bestimmtheitsmaß!

◁

3.4 Gegeben ist eine Stichprobe folgender Wertepaare:

x	5.5	2.4	4.8	12.9	9.4	7.8	14.9	11.2	12.9	12.1	9.8	3.1
y	31.5	15.2	31.6	76.8	57.7	43.3	83.5	70.2	70.1	68.5	52.7	29.3

- Berechne die lineare Regressionsfunktion!
- Berechne ein 95 % Konfidenzintervall für die Steigung der Regressionsgeraden!
- Überprüfe, ob die Abweichungen von der Regressionsgeraden normalverteilt sind!

<

3.5 Gegeben sind folgende, an freiwilligen Versuchspersonen gemessene Daten

Armlänge	58	62	55	67	51
Beinlänge	88	91	84	90	81

- Berechne für die Armlänge 60 die zugehörige geschätzte Beinlänge y_{60} !
- Umgekehrte Regression: Berechne die zur Beinlänge y_{60} (wie berechnet) gehörige Armlänge! Warum ist diese nicht gleich 60?
- Zeichne beide Regressionsgleichungen aus (a) und (b) in ein gemeinsames Schaubild!

<

3.6 Die folgende Tabelle zeigt, wie die Stückkosten eines bestimmten Produktes von der Menge abhängen:

Menge	(x)	1	2	6	9	12
Stückkosten	(y)	120	90	75	45	30

- Bestimme die lineare Regressionsgleichung $\hat{y} = a + bx$!
- Wie ändern sich a und b , wenn sämtliche x -Werte um 20 % erhöht und sämtliche y -Werte um 5 % vermindert werden?
- Bestimme für die Regressionsgleichung $y = cx^2$ zunächst die Normalgleichungen und wende diese dann auf die obigen Daten an!

<

- 3.7 Es sei x_1 die Anzahl der in Gebrauch befindlichen Autos (in Millionen), x_2 die Anzahl der in Gebrauch befindlichen Lastwagen (in Millionen) und y der Benzinverbrauch in Millionen Barrel. Wir möchten die Regression von y auf x_1 und x_2 ermitteln, d.h.,

$$y = a + b_1x_1 + b_2x_2$$

x_1	x_2	y
36	8	990
40	8	1140
44	9	1230
47	9	1320
50	10	1370

- Schätze a , b_1 , b_2 !
- Schätze σ^2 !
- Schätze $\text{Var}(b_1)$ und $\text{Var}(b_2)$!

<

- 3.8 Gegeben sind folgende Werte:

x	0	1	2	3	4
y	130	145	150	165	170

Berechne:

- Bestimmtheitsmaß,
- Konfidenzintervall für die Regressionskoeffizienten der Grundgesamtheit !
- Teste in a) mit Hilfe eines t-Tests die Hypothese, dass in der Grundgesamtheit kein Zusammenhang zwischen den Variablen x und y besteht !

<

- 3.9 In einer sehr kleinen Stichprobe ($n = 17$) korrelieren zwei quantitative Variablen mit $r = 0.61$. Kann der Koeffizient durch Zufall entstanden sein ($\alpha = 0.05$) ?

<

- 3.10 In einer Stichprobe von $n = 259$ korrelieren zwei Variable mit $r = 0.58$. Gib ein 95 % Konfidenzintervall für ρ an !

<

3.11 Gegeben ist folgende Datenreihe:

x	8	10	11	14	15	17	19	20
y	110	150	144	180	210	220	250	265

- Bestimme $\hat{y} = a + bx$!
- Prüfe, ob der Stichprobenregressionskoeffizient b gegen 0 gesichert ist ($\alpha = 0.01$) !

◁

3.12 $X \dots$ Produktmenge $Y \dots$ Gesamtkosten

X	49	25	60	30	58	65	55	20	40	35	30	45
Y	214	112	265	142	250	275	240	104	175	165	128	205

Bestimme:

- $\hat{Y} = a + bX$,
- das Bestimmtheitsmaß,
- ein 99 % Konfidenzintervall für die durchschnittlichen Gesamtkosten mit einer Produktionsmenge von $X_0 = 52$ Stück,
- ein 95 % Prognoseintervall für die Gesamtkosten mit einer Produktionsmenge $x_0 = 70$ Stück!

◁

3.13 Bei der Messung von Hämoglobingehalt im Blut (X) und mittlerer Oberfläche der Erythrozyten (Y) bei 12 Personen ergaben sich folgende Daten:

	Person	Hämoglobin- gehalt (X)	mittl. Oberfläche d. Erythrozyten (Y)
Frauen	1	13.1	85.2
	2	12.9	92.4
	3	13.7	92.4
	4	14.5	90.8
	5	14.1	97.5
	6	12.7	88.6
Männer	7	16.5	103.1
	8	15.7	106.3
	9	17.0	99.8
	10	14.9	101.4
	11	15.8	98.8
	12	17.5	103.4

- a) Berechne die Korrelation für alle Daten!
- b) Berechne die Korrelation für Männer und Frauen getrennt!
- c) Stelle die Daten grafisch dar!

◀