Kapitel III

Regressionsanalyse

D. 3. 1. (*Regressionsfunktion*)
Gegeben sei die $n-$dimensionale Verteilung der metrisch messbaren Merkmale $X_1, X_2, ..., X_m$ und Y. Es sei Y statistisch abhängig von X_i, $i = 1, 2, ..., m$.
Eine Funktion $y^* = f(x_1, x_2, ..., x_m)$, die die Tendenz der Abhängigkeit beschreibt, heißt *Regressionsfunktion*.

B. 3. 1.
Da die Variable y durch eine Vielzahl weiterer nicht näher spezifizierter Erscheinungen sowie durch zufällige Einflüsse zustande kommt, existiert eine Abweichung zwischen der Variablen y und der durch die Regressionsfunktion berechneten mittleren Größen y^*, die wir mit u bezeichnen wollen:

$$y - y^* =: u\ .$$

u ist eine zufällige *Störvariable*, die somit alle nicht in den m erklärenden Variablen enthaltenen Einflüsse auf die Variable y und Zufallseinflüsse enthält.
Die Variable y ergibt sich somit als

$$y = y^* + u$$

bzw. als

$$y = f(x_1, x_2, ..., x_m) + u.$$

B. 3. 2.
Die Koeffizienten der Regressionsfunktion werden folgendermaßen berechnet:

$$S(.) = \sum_{i=1}^{n} (y_i - y_i^*)^2 \to Min!$$

(*Methode der kleinsten Quadratsummen*)

D. 3. 2. (*Einfache lineare Regression*)
Einfache lineare Regression liegt vor, wenn die Regressionsfunktion der Form

$$y^* = a_0 + a_1 x$$

ist.

<u>S. 3. 1.</u> (*Normalgleichungen der einfachen linearen Regression*)
Die *Normalgleichungen der einfachen linearen Regression* lauten:

$$\begin{cases} n \cdot a_0 \qquad + a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \cdot \sum_{i=1}^{n} x_i + a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

B e w e i s:
Die Behauptung ergibt sich aus:

$$S(a_0, a_1) \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i \right)^2 \rightarrow Min!$$

$$\begin{cases} \dfrac{\partial S\left(a_0, a_1\right)}{\partial a_0} = -2 \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i \right) \qquad = 0 \\ \dfrac{\partial S\left(a_0, a_1\right)}{\partial a_1} = -2 \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i \right) \cdot x_i = 0 \end{cases}$$

$$\frac{\partial^2 S\left(a_0, a_1\right)}{\partial a_0^2} = 2n > 0 \, , \qquad \qquad \frac{\partial^2 S\left(a_0, a_1\right)}{\partial a_1^2} = 2 \sum_{i=1}^{n} x_i^2 > 0$$

$$\frac{\partial^2 S\left(a_0, a_1\right)}{\partial a_0^2} \cdot \frac{\partial^2 S\left(a_0, a_1\right)}{\partial a_1^2} - \left(\frac{\partial^2 S\left(a_0, a_1\right)}{\partial a_0 \partial a_1} \right)^2 = \sum_{i=1}^{n} \left(x_i - \bar{x} \right) > 0.$$

<u>B. 3. 3.</u>
Das Normalgleichungssystem der einfachen linearen Regression lässt sich u. a. nach der *Cramer-Regel* lösen:

$$a_0 = \frac{\begin{vmatrix} \sum_i y_i & \sum_i x_i \\ \sum_i x_i \cdot y_i & \sum_i x_i^2 \end{vmatrix}}{\begin{vmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{vmatrix}} = \frac{\sum_i y_i \cdot \sum_i x_i^2 - \sum_i x_i \cdot \sum_i x_i \cdot y_i}{n \cdot \sum_i x_i^2 - \sum_i x_i \cdot \sum_i x_i}$$

$$a_1 = \frac{\begin{vmatrix} n & \sum_i y_i \\ \sum_i x_i & \sum_i x_i \cdot y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{vmatrix}} = \frac{n \cdot \sum_i x_i \cdot y_i - \sum_i x_i \cdot \sum_i y_i}{n \cdot \sum_i x_i^2 - \sum_i x_i \cdot \sum_i x_i}$$

S. 3. 2.
Für eine einfache lineare Regressionsfunktion gilt

$$\bar{y} = a_0 + a_1 \bar{x} .$$

B e w e i s:
Wir dividieren beide Seiten der 1. Normalgleichung durch n:

$$\frac{\sum_{i=1}^{n} y_i}{n} = a_0 + a_1 \cdot \frac{\sum_{i=1}^{n} x_i}{n} ,$$

d. h.

$$\bar{y} = a_0 + a_1 \bar{x} .$$

B. 3. 4.
Es kann gezeigt werden, dass der Regressionskoeffizient a_1 auch folgendermaßen dargestellt werden kann:

$$a_1 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right) \cdot \left(y_i - \bar{y} \right)}{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2} .$$

Dividieren wir Zähler und Nenner durch $n-1$, so erhalten wir im Zähler die Kovarianz der Variablen x und y und im Nenner die Varianz der Variablen x, also

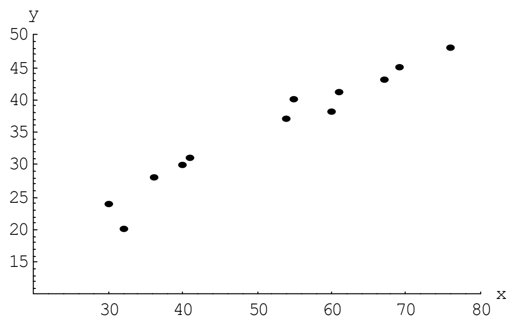
$$a_1 = \frac{s_{xy}}{s_x^2} .$$

<u>BS. 3. 1.</u>
Es soll die Abhängigkeit des Niveaus der Arbeitsproduktivität von dem Automatisierungsgrad der Arbeit in 14 Betrieben untersucht werden. Dazu liegt folgendes Datenmaterial vor:

Betrieb	Niveau der Arbeitsproduktivität t/Std.	Automatisierungsgrad der Arbeit %
1	20	32
2	24	30
3	28	36
4	30	40
5	31	41
6	33	47
7	34	56
8	37	54
9	38	60
10	40	55
11	41	61
12	43	67
13	45	69
14	48	76

Lösung:

Das <u>Streuungsdiagramm</u>:



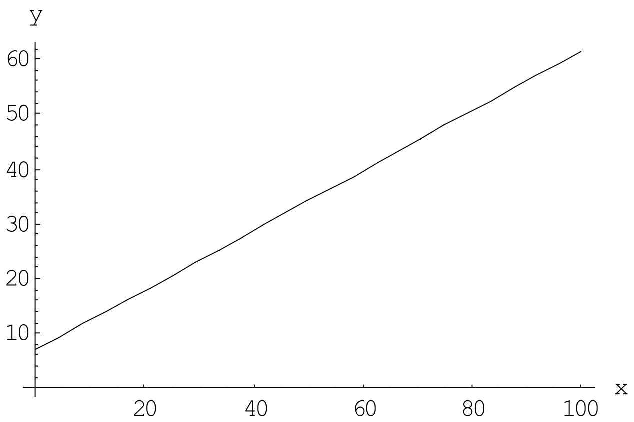
Arbeitstabelle

i	y_i	x_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	20	32	640	1024	400
2	24	30	720	900	576
3	28	36	1008	1296	784
4	30	40	1200	1600	900
5	31	41	1271	1681	961
6	33	47	1551	2209	1089
7	34	56	1904	3136	1156
8	37	54	1998	2916	1369
9	38	60	2280	3000	1444
10	40	55	2200	3025	1600
11	41	61	2501	3721	1681
12	43	67	2881	4489	1849
13	45	69	3105	4761	2025
14	48	76	3648	5776	2304
Summen	492	724	26907	40134	18138

$$\begin{aligned} 14a_0 \quad + 724a_1 &= 492 \\ 724a_0 + 40134a_1 &= 26907 \end{aligned} \quad \Rightarrow \quad a_0 = 7.0356, \quad a_1 = 0.5435 .$$

Damit lautet die gesuchte lineare Regressionsfunktion:

$$y^* = 7.0356 + 0.5435x .$$



Beispielsweise gilt:

$$y^*(42) = 7.0356 + 0.5435 \cdot 42 \approx 29.86 \text{ (Interpolation)}$$
$$y^*(80) = 7.0356 + 0.5435 \cdot 80 \approx 50.52 \text{ (Extrapolation).}$$

D. 3. 3. (*Multiple lineare Regression*)
*Multiple (*oder *mehrfache) lineare Regression* liegt vor, wenn die Regressionsfunktion der Form

$$y^* = a_0 + a_1 x + a_2 x_2 + \ldots + a_n x_n$$

ist.

B. 3. 5. (*Lineare Regression mit zwei erklärenden Variablen*)
Eine lineare Regressionsfunktion mit zwei erklärenden Variablen lässt sich folgendermaßen darstellen:

$$y^* = a_0 + a_1 x + a_2 x_2 \,.$$

Wendet man hierauf die Methode der kleinsten Quadratsummen

$$S(a_0, a_1, a_2) = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_{i1} - a_2 x_{i2} \right)^2 \to Min!,$$

so erhält man die Normalgleichungen

$$\begin{cases} n \cdot a_0 & + a_1 \sum_{i=1}^{n} x_{i1} & + a_2 \sum_{i=1}^{n} x_{i2} & = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_{i1} & + a_1 \sum_{i=1}^{n} x_{i1}^2 & + a_2 \sum_{i=1}^{n} x_{i1} x_{i2} & = \sum_{i=1}^{n} x_{i1} y_i \\ a_0 \sum_{i=1}^{n} x_{i2} & + a_1 \sum_{i=1}^{n} x_{i1} x_{i2} & + a_2 \sum_{i=1}^{n} x_{i2}^2 & = \sum_{i=1}^{n} x_{i2} y_i \end{cases}$$

B. 3. 6.
Analog zur einfachen linearen Regression können auch die multiplen Regressionskoeffizienten als Beziehungen zwischen den Varianzen und Kovarianzen dargestellt werden.

Zunächst dividieren wir die 1. Normalgleichung durch n, multiplizieren mit $\sum_{i=1}^{n} x_{i1}$ und subtrahieren das Ergebnis von der2. Normalgleichung und erhalten:

$$\sum_{i=1}^{n} x_{i1} y_i - \bar{y} \sum_{i=1}^{n} x_{i1} = a_1 \left(\sum_{i=1}^{n} x_{i1}^2 - \bar{x}_1 \sum_{i=1}^{n} x_{i1} \right) + a_2 \left(\sum_{i=1}^{n} x_{i1} x_{i2} - \bar{x}_2 \sum_{i=1}^{n} x_{i1} \right).$$

Dann multiplizieren wir die durch n dividierte 1. Normalgleichung mit $\sum_{i=1}^{n} x_{i1}$ und subtrahieren das Ergebnis von der 3. Normalgleichung, wodurch sich ergibt:

$$\sum_{i=1}^{n} x_{i2} y_i - \bar{y} \sum_{i=1}^{n} x_{i2} = a_1 \left(\sum_{i=1}^{n} x_{i1} x_{i2} - \bar{x}_1 \sum_{i=1}^{n} x_{i2} \right) + a_2 \left(\sum_{i=1}^{n} x_{i2}^2 - \bar{x}_2 \sum_{i=1}^{n} x_{i2} \right).$$

Die letzten beiden Gleichungen lassen sich auch folgendermaßen darstellen:

$$\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(y_i-\bar{y}\right)=a_1\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)^2+a_2\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(x_{i2}-\bar{x}_2\right)$$

$$\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_2\right)\left(y_i-\bar{y}\right)=a_1\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(x_{i2}-\bar{x}_2\right)+a_2\sum_{i=1}^{n}\left(x_{i2}-\bar{x}_2\right)^2 .$$

Dividieren wir nun die letzten beiden Gleichungen jeweils durch $n-1$ und lösen unter Beachtung der Definitionen der Varianzen und Kovarianzen nach den Regressionskoeffizienten a_1 und a_2, so erhalten wir:

$$a_1=\frac{s_{1y}s_2^2-s_{2y}s_{12}}{s_1^2s_2^2-s_{12}^2}, \qquad a_2=\frac{s_1^2s_{2y}-s_{12}s_{1y}}{s_1^2s_2^2-s_{12}^2}.$$

<u>BS. 3. 2.</u> (*Erweiterung des Beispiels BS.3.1.*)
Es sei angenommen, dass das Niveau der Arbeitsproduktivität nicht nur vom Automatisierungsgrad der Arbeit x_1, sondern auch vom Alter der Beschäftigten x_2 abhängt. Damit haben wir folgende Informationen:

Betrieb	Niveau der Arbeitsproduktivität t/Std.	Automatisierungsgrad der Arbeit %	Alter der Beschäftigten Jahre
1	20	32	33
2	24	30	31
3	28	36	41
4	30	40	39
5	31	41	46
6	33	47	43
7	34	56	34
8	37	54	38
9	38	60	42
10	40	55	35
11	41	61	39
12	43	67	44
13	45	69	40
14	48	76	41

Arbeitstabelle

i	y_i	x_{i1}	x_{i2}	x_{i1}^2	x_{i2}^2	$x_{i1} \cdot y_i$	$x_{i2} \cdot y_i$	$x_{i1} \cdot x_{i2}$
1	20	32	33	1024	1089	640	660	1056
2	26	30	31	900	961	720	744	930
3	28	36	41	1296	1681	1008	1148	1476
4	30	40	39	1600	1521	1200	1170	1560
5	31	41	46	1681	2116	1271	1426	1886
6	33	47	43	2209	1849	1551	1419	2021
7	34	56	34	3136	1156	1904	1156	1904
8	37	54	38	2916	1444	1998	1406	2052
9	38	60	42	3000	1764	2280	1596	2520
10	40	55	35	3025	1225	2200	1400	1925
11	41	61	39	3721	1521	2501	1599	2379
12	43	67	44	4489	1936	2881	1892	2948
13	45	69	40	4761	1600	3105	1800	2760
14	48	76	41	5776	1681	3648	1968	3116
Summen	492	724	546	40134	21544	26907	19384	28533

$$\begin{cases} 14a_0 + 724a_1 + 546a_2 = 492 \\ 724a_0 + 40134a_1 + 28533a_2 = 26907 \\ 546a_0 + 28533a_1 + 21544a_2 = 19384 \end{cases}$$

Das Gleichungssystem hat die Lösung:

$$a_0 = 1.7408, \qquad a_1 = 0.5259, \qquad a_2 = 0.1591.$$

Damit lautet die gesuchte Regressionsfunktion:

$$y^* = 1.7408 + 0.5259x_1 + 0.1591x_2$$

<u>B. 3. 7.</u> (*Lineare Regression mit m+1 erklärenden Variablen*)
Betrachtet seien die Funktionen

$$y^* = a_0x_0 + a_1x_1 + \ldots + a_mx_m \qquad \text{bzw.} \quad y = a_0x_0 + a_1x_1 + \ldots + a_mx_m + u^*$$

Sei

$$
y=\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}, \quad X=\begin{pmatrix} x_{10} & x_{11} & . & . & . & x_{1m} \\ x_{20} & x_{21} & . & . & . & x_{2m} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{n0} & x_{n1} & . & . & . & x_{nm} \end{pmatrix}, \quad u^*=\begin{pmatrix} u_1^* \\ u_2^* \\ . \\ . \\ . \\ u_n^* \end{pmatrix}, \quad b=\begin{pmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_m \end{pmatrix}.
$$

Damit erhält man die Matrizengleichungen

$$
y^* = Xa \qquad \text{bzw.} \qquad y = Xa + u^*.
$$

Unter Anwendung der Methode der kleinsten Quadratsummen ergibt sich:

$$
S(a) = (y - y^*)^T (y - y^*) = u^{*T} u^* \to \underset{b}{Min}!
$$

Bzw.

$$
S(a) = u^{*T} u^* = (y - Xa)^T (y - Xa) \to \underset{a}{Min}
$$

Bzw.

$$
S(a) = y^T y - 2a^T X^T y + a^T Xy + a^T X^T Xa \to \underset{a}{Min}!
$$

Die partielle Ableitung dieser Funktion wird nun gebildet und gleich Null gesetzt:

$$
\frac{\partial S(a)}{\partial a} = -2X^T y + 2X^T Xa = 0.
$$

Hieraus folgt die Lösung

$$
a = \left(X^T X\right)^{-1} X^T y
$$

Unter Berücksichtigung der Scheinvariablen $x_{i0} \equiv 1,\ \forall i,$ sind:

$$
X^T X = \begin{pmatrix} n & \sum_i x_{i1} & . & . & . & \sum_i x_{im} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & . & . & . & \sum_i x_{i1} x_{im} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sum_i x_{im} & \sum_i x_{im} x_{i1} & . & . & . & \sum_i x_{im}^2 \end{pmatrix}; \qquad X^T y = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i1} y_i \\ . \\ . \\ . \\ \sum_i x_{im} y_i \end{pmatrix}.
$$

<u>BS. 3. 3.</u> (*Erweiterung des Beispiels BS.3.2.*)

Betrieb	Niveau der Arbeitsproduktivität t/Std.	Automatisierungsgrad der Arbeit [%]	Durchschnittsalter der Beschäftigten [Jahre]	Durchschnittliches Wachstumstempo gegenüber Vorjahr [%]
1	20	32	33	127
2	24	30	31	120
3	28	36	41	116
4	30	40	39	117
5	31	41	46	106
6	33	47	43	128
7	34	56	34	109
8	37	54	38	114
9	38	60	42	115
10	40	55	35	121
11	41	61	39	110
12	43	67	44	111
13	45	69	40	108
14	48	76	41	113

Lösung:

Wir haben:

$$y=\begin{pmatrix}20\\24\\28\\30\\31\\33\\34\\37\\38\\40\\41\\43\\45\\48\end{pmatrix};\qquad X=\begin{pmatrix}1&32&33&127\\1&30&31&120\\1&36&41&116\\1&40&39&117\\1&41&46&106\\1&47&43&128\\1&56&34&109\\1&54&38&114\\1&60&42&115\\1&55&35&121\\1&61&39&110\\1&67&44&111\\1&69&40&108\\1&76&41&113\end{pmatrix}.$$

Hieraus folgt:

$$X^TX=\begin{pmatrix}14&724&546&1615\\724&40134&28533&82884\\546&28533&21544&62840\\1615&82884&62840&186891\end{pmatrix};\quad X^Ty=\begin{pmatrix}492\\26907\\19384\\56389\end{pmatrix};$$

$$a=\left(X^TX\right)^{-1}X^Ty=\begin{pmatrix}52.88929&-0.06869&-0.26929&-0.33603\\-0.06869&0.00052&-0.00034&0.00048\\-0.26929&-0.00034&0.00489&0.00083\\-0.33603&0.00048&0.00083&0.00242\end{pmatrix}\cdot\begin{pmatrix}492\\26907\\19384\\56389\end{pmatrix}=\begin{pmatrix}5.05729\\0.52123\\0.15092\\-0.02389\end{pmatrix}.$$

Damit lautet die gesuchte Regressionsfunktion:

$$y^*=5.05729+0.52123x_1+0.15092x_2-0.02389x_3$$

$$y^* = \begin{pmatrix} 23.6829 \\ 22.5058 \\ 27.2379 \\ 28.9971 \\ 30.8376 \\ 32.9866 \\ 36.7733 \\ 36.2151 \\ 39.9222 \\ 36.1163 \\ 40.1101 \\ 43.9682 \\ 44.4786 \\ 48.1587 \end{pmatrix}; \qquad u^* = y - y^* = \begin{pmatrix} -3.6829 \\ 1.4942 \\ 0.7621 \\ 1.0029 \\ 0.1624 \\ 0.0134 \\ -2.7733 \\ 0.7849 \\ -1.9222 \\ 3.8837 \\ 0.8899 \\ -0.9682 \\ 0.5214 \\ -0.1567 \end{pmatrix}.$$

B. 3. 8. (***Annahmen des linearen Regressionsmodells***)
1. Linearität in den Parametern.
2. Erwartungswert der Störgrößen gleich Null.
3. Berücksichtigung aller relevanten Variablen.
4. Homoskedastizität, d.h. die Störgrößen haben eine konstante Varianz.
5. Unabhängigkeit der Störgrößen.
6. Keine lineare Abhängigkeit zwischen den unabhängigen Variablen.
7. Störgrößen sind normalverteilt.

B. 3. 9.
Das Kriterium der Methode der kleinsten Quadratsummen

$$\sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} u_i^* \to Min!$$

ist nicht gerade vorteilhaft als eine Maßzahl für den Grad der Anpassung einer Regressionsfunktion an die gegebenen empirischen Daten. Durch das Kriterium wird zwar eine untere Grenze von Null festgelegt, aber keine obere Grenze.
Daher entsteht die Notwendigkeit der Suche nach einem anderen Kriterium.

B. 3. 10. (***Gesamtvarianz der empirischen Werte und deren Zerlegung***)
Die Varianz der empirischen Werte der erklärenden Variablen y ist:

$$s_y^2 = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}{n-1}.$$

Diese Varianz wird auch als *Gesamtvarianz* bezeichnet.

Nun gilt aber:

$$y_i - \bar{y} = (y_i - y_i^*) - (y_i^* - \bar{y}).$$

Das Quadrieren und über alle i Summieren dieser Beziehung ergibt:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - y_i^*)^2 + 2\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \bar{y}) + \sum_{i=1}^{n}(y_i^* - \bar{y})^2.$$

Es lässt aber nun zeigen, dass

$$\sum_{i=1}^{n} y_i^* = \sum_{i=1}^{n} y_i$$

und

$$\sum_{i=1}^{n} y_i^* u_i = 0$$

und folglich

$$2\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \bar{y}) = 0$$

gilt, so dass:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - y_i^*)^2 + \sum_{i=1}^{n}(y_i^* - \bar{y})^2.$$

Hieraus folgt:

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{n-1} + \frac{\sum_{i=1}^{n}(y_i^* - \bar{y})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n} u_i^{*2}}{n-1} + \frac{\sum_{i=1}^{n}(y_i^* - \bar{y})}{n-1}.$$

Schließlich hat man

$$s_y^2 = s_u^{'2} + s_{y*}^{'2}$$

Hier ist:

$s_u^{'2}$: die sog. „nicht erklärte“ Varianz

$s_{y^*}^{'2}$: Varianz der Regresswerte.

D. 3. 4. (***Einfaches Bestimmtheitsmaß***)
Als *einfaches Bestimmtheitsmaß für die einfache lineare Regression* bezeichnet man:

$$B_{yx} := \frac{\frac{\sum_{i=1}^{n}(y_i^* - \bar{y})^2}{n-1}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \frac{\sum_{i=1}^{n}(y_i^* - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad 0 \le B_{yx} \le 1.$$

B. 3. 11.
Die Formel für das einfache Bestimmtheitsmaß basiert auf den Überlegungen in B. 2. 9. Sie gibt den Anteil der „erklärenden“ Varianz an der Gesamtvarianz an. Je größer der Anteil der Gesamtvarianz, desto besser passt sich die Regressionsfunktion den empirischen Werten an.

B. 3. 12. (***Weitere Formeln für das einfache Bestimmtheitsmaß***)
Es gilt:

$$B_{yx} = \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

bzw.

$$B_{yx} := \frac{\left(n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i\right)^2}{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i\right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i\right)}.$$

D. 3. 5. (***Einfaches Unbestimmtheitsmaß***)
Als *Unbestimmtheitsmaß* bezeichnet man:

$$U_{yx} := \frac{\frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{n-1}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

B. 3. 13.
Offensichtlich gilt

$$B_{yx} + U_{yx} = 1.$$

<u>BS. 3. 1.</u> (*Fortsetzung*)

$$B_{yx} = \frac{(14 \cdot 26907 - 724 \cdot 492)^2}{(14 \cdot 40134 - 724^2)(14 \cdot 18138 - 492^2)} \approx 0.938.$$

Interpretation: Die Produktivität wird zu etwa 93.8% durch das Automatisierungsgrad der Arbeit bestimmt. Der Anteil der „Störfaktoren“ ist etwa 6.2%.

<u>D. 3. 6.</u> (*Mehrfaches bzw. multiples Unbestimmtheitsmaß*)
Als *mehrfaches* bzw. *multiples Bestimmtheitsmaß* bezeichnet man:

$$B_{y;1,2,\ldots,m} = \frac{\sum_{i=1}^{n} (y_i^* - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \qquad 0 \le B_{y;1,2,\ldots,m} \le 1.$$

<u>B. 3. 14.</u>
In Matrizenschreibweise lässt sich das mehrfache Bestimmtheitsmaß folgendermaßen darstellen:

$$B_{y;1,2,\ldots,m} = \frac{a_{(1)}^T s_{xy}}{s_y^2}, \qquad 0 \le B_{y;1,2,\ldots,m} \le 1.$$

Hier sind:

$$a_{(1)} = \begin{pmatrix} a_1 \\ . \\ . \\ . \\ a_m \end{pmatrix}, \qquad s_{xy} = \begin{pmatrix} s_{1y} \\ . \\ . \\ . \\ s_{my} \end{pmatrix}.$$

BS. 3. 4. (***Fortsetzung der Beispiele BS. 3. 2. und BS. 3. 3.***)

y_i	x_{i1}	x_{i2}	x_{i3}	$(y_i - \bar{y})^2$	$(x_{i1} - \bar{x}_{i1}) \cdot (y_i - \bar{y})$	$(x_{i2} - \bar{x}_{i2}) \cdot (y_i - \bar{y})$	$(x_{i3} - \bar{x}_{i3}) \cdot (y_i - \bar{y})$
20	32	33	127	229.306121	298.530611	90.8571426	-176.306119
24	30	31	120	124.163264	241.959183	89.1428568	-51.7346916
28	36	41	116	51.0204076	112.244897	-14.2857142	-4.59183562
30	40	39	117	26.4489792	60.2448974	0.0000000	-8.44897862
31	41	46	106	17.163265	44.3877546	-28.9999997	38.7653054
33	47	43	128	4.59183655	10.1020406	-8.5714284	-27.0918346
34	56	34	109	1.30612235	-4.89795902	5.7142855	7.26530538
37	54	38	114	3.44897975	4.24489808	-1.8571429	-2.52040862
38	60	42	115	8.16326555	23.6734698	8.5714287	-1.02040862
40	55	35	121	23.5918372	15.9591839	-19.4285716	27.4081634
41	61	39	110	34.306123	54.3877556	0.0000000	-31.3775526
43	67	44	111	61.7346946	120.102042	39.2857145	-34.2346956
45	69	40	108	97.1632662	170.387756	9.8571429	-72.5204106
48	76	41	113	165.306124	312.244899	25.7142858	-30.3061246
492	724	546	1615	847.714286	1463.571430	196.000000	-366.714286

$\bar{y} = 35.1428571,\qquad \bar{x}_1 = 51.7142857,\qquad \bar{x}_2 = 39.0000000,\qquad \bar{x}_3 = 115.3571429$

$$s_y^2 = \frac{847.714286}{13} = 65.2087912$$

$$s_{x_1y} = \frac{1463.57143}{13} = 112.582418$$

$$s_{x_2y} = \frac{196.00000}{13} = 15.07692308$$

$$s_{x_3y} = \frac{-366.714286}{13} = -28.2087912$$

BS. 3. 2.:

$$a_{(1)} = \begin{pmatrix} 0.5259 \\ 0.1591 \end{pmatrix}, \qquad s_{xy} = \begin{pmatrix} 112.582418 \\ 18.6373626 \end{pmatrix},$$

$$B_{y;12} = \frac{\begin{pmatrix} 0.5259 & 0.1591 \end{pmatrix} \cdot \begin{pmatrix} 112.582418 \\ 18.6373626 \end{pmatrix}}{65.2087912} = 0.953434297$$

BS. 3. 3.:

$$a_{(1)} = \begin{pmatrix} 0.52123 \\ 0.15092 \\ -0.02389 \end{pmatrix}, \quad s_{xy} = \begin{pmatrix} 112.582418 \\ 18.6373626 \\ -28.2087912 \end{pmatrix},$$

$$B_{y;123} = \frac{\begin{pmatrix} 0.5213 & 0.15092 & -0.02389 \end{pmatrix} \cdot \begin{pmatrix} 112.582418 \\ 15.07692308 \\ -28.2087912 \end{pmatrix}}{65.2087912} = 0.945248801$$

B. 3. 15.

Oftmals, vor allem bei kleinem Stichprobenumfang n, wird ein *korrigiertes Bestimmtheitsmaß* ermittelt, da die Anzahl der erklärenden Variablen die Anzahl der Freiheitsgrade wesentlich vermindert. Dadurch wird die Bestimmtheit überschätzt.

D. 3. 7 (*Korrigiertes Bestimmtheitsmaß*)

Als *korrigiertes Bestimmtheitsmaß* bezeichnet man

$$B^*_{y;1\ldots m} = 1 - \frac{s_u^2}{s_y^2}$$

$$= 1 - U_{y;1\ldots m} \frac{n-1}{n-m-1}$$

$$= 1 - (1 - B_{y;1\ldots m}) \frac{n-1}{n-m-1}$$

BS. 3. 5. (*BS. 3. 4. fortgesetzt*)

BS. 3. 2.:

$$B^*_{y;12} = 1 - (1 - 0.953434297) \frac{14-1}{14-2-1} = 0.944967805.$$

BS. 3. 3.:

$$B^*_{y;123} = 1 - (1 - 0.945248801) \frac{14-1}{14-3-1} = 0.928823441.$$

<u>D. 3. 8</u> (*Varianz* bzw. *der Standardfehler der Residuen*)
Als *Varianz* bzw. *Standardfehler der Residuen* (auch als *Standardfehler der Regressionsschätzung* genannt) bezeichnet man:

$$s_u^2 = \frac{\sum_{i=1}^{n} u_i^{*2}}{n-(m+1)}$$

$$= \frac{1}{n-(m+1)} u^{*T} u^*$$

<u>D. 3. 9</u> (*Varianz* bzw. *der Standardfehler der Regressionsparameter*)
Als *Varianz* bzw. *Standardfehler der Parameter* bezeichnet man:

$$s_{a_i} = s_u \sqrt{x^{(ii)}}, \quad i = 0,1,...,n.$$

Dabei ist x^{ii} das $i-$te Diagonalelement der Matrix $(X^T X)^{-1}$.

<u>D. 3. 10.</u> (*Relative Standardfehler der Regressionsparameter*)

$$s'_{a_i} = \frac{s_{a_i}}{a_i}, \quad i = 0,1,...,n.$$

<u>B. 3. 16.</u>
Je größer diese relativen Standardfehler der Regressionsparameter sind, desto weniger zuverlässig sind die geschätzte Regressionsfunktion und die darauf bauenden Prognosewerte.

BS. 3. 6.

Zu BS. 3. 1.:

y_i	x_i	y_i^*	$u_i^* = y_i - y_i^*$
20	32	24.4276	-4.4276
24	30	23.3406	0.6594
28	36	26.6016	1.3984
30	40	28.7756	1.2244
31	41	29.3186	1.6814
33	47	32.5806	0.4194
34	56	37.4716	-3.4716
37	54	36.3846	0.6154
38	60	39.6456	-1.6456
40	55	36.9276	3.0724
41	61	40.1896	0.8104
43	67	43.4496	-0.4496
45	69	44.5376	0.4624
48	76	48.3416	-0.3416
492	724	491.9924	0.0076

$$u^* = y - y^* = \begin{pmatrix} -4.4276 \\ 0.6594 \\ 1.3984 \\ 1.2244 \\ 1.6814 \\ 0.4194 \\ -3.4716 \\ 0.6154 \\ -1.6456 \\ 3.0724 \\ 0.8104 \\ -0.4496 \\ 0.4624 \\ -0.3416 \end{pmatrix}$$

$$s_u^2 = \frac{1}{n-(m+1)} u^{*T} u^* = \frac{52.2638878}{12} = 4.355323983\,; \qquad s_u \approx 2.086941298$$

$$X^T X = \begin{pmatrix} 14 & 724 \\ 724 & 40134 \end{pmatrix}; \qquad (X^T X)^{-1} = \begin{pmatrix} 1.06456 & -0.01920 \\ -0.01920 & 0.00037 \end{pmatrix},$$

$$s_{a_0}^2 = 4.355323983 \cdot 52.88929 \approx 4.6365, \qquad s_{a_0} \approx 2.1533,$$

$$s'_{a_0} = \frac{2.1533}{7.0356} \approx 0.3061, \text{ also ca. } 30.61\%$$

$$s_{a_1}^2 = 4.355323983 \cdot 0.00037 \approx 0.00161, \qquad s_{a_1} \approx 0.0401,$$

$$s'_{a_0} = \frac{0.0401}{0.5435} \approx 0.07378, \text{ also ca. } 7.38\%.$$

<u>Zu BS. 3. 4</u>.:

$$s_u^2 = \frac{1}{n-(m+1)} u^{*T} u^* = \frac{46.756936}{14-(3+1)} = 4.65211605; \quad s_u \approx 2.1623$$

$$s_{a_0}^2 = 4.65211605 \cdot 52.88929 \approx 246.0711 \qquad s_{a_0} \approx 15.6859,$$

$$s'_{a_0} = \frac{15.6859}{5.05729} \approx 3.1016,$$

$$s_{a_1}^2 = 4.65211605 \cdot 0.00052 \approx 0.00242 \qquad s_{a_1} \approx 0.04919,$$

$$s'_{a_1} = \frac{0.04919}{0.52123} \approx 0.09437,$$

$$s_{a_2}^2 = 4.65211605 \cdot 0.00489 \approx 0.02275 \qquad s_{a_2} \approx 0.1508,$$

$$s'_{a_2} = \frac{0.1508}{0.15092} \approx 0.9992,$$

$$s_{a_3}^2 = 4.65211605 \cdot 0.00242 \approx 0.01126 \qquad s_{a_3} \approx 0.1061,$$

$$s'_{a_3} = \left| \frac{0.1061}{-0.02389} \right| \approx 4.4414.$$

Während für die einfache Regressionsfunktion die Standardfehler der Regressionsparameter akzeptabel sind, trifft das bei der multiplen Regressionsfunktion nur für den Standardfehler von a_1 zu.

<u>B. 2. 17.</u> (*Konfidenzintervall der Regressionsparameter*)
Sei

$$Y = A_0 + A_1 x_1 + \ldots + A_m x_m$$

die Regressionsgleichung der Grundgesamtheit.

Das *Konfidenzintervall* für den Regressionsparameter $A_i,\ i = 1, 2, \ldots, m,$ mit dem Signifikanzniveau α wird folgendermaßen bestimmt:

$$\left[a_i - t_{n-m-1} \cdot s_{a_i},\ a_i - t_{n-m-1} \cdot s_{a_i}\right],\ i = 0, 1, \ldots, m.$$

<u>BS. 3. 7.</u>
Sei $\alpha = 0.05$.

<u>Zu BS. 3. 1</u>.:

$$A_0 \in \left[7.0356 - 2.179 \cdot 2.1533,\ 7.0356 + 2.179 \cdot 2.1533\ \right] \approx \left[2.3436,\ 11.7276\right],$$

d.h. mit einer Wahrscheinlichkeit von 95% wird A_0 dem obigen Intervall angehören.

Zu a_1 :

$$A_1 \in \left[0.5435 - 2.179 \cdot 0.0401,\ 0.5435 + 2.179 \cdot 0.0401\ \right] \approx \left[0.4561,\ 0.6308\right],$$

d.h. mit einer Wahrscheinlichkeit von 95% wird A_1 dem obigen Intervall angehören.

<u>Zu BS. 3. 4</u>.:
Nur für A_1:

$$A_1 \in \left[0.52123 - 2.228 \cdot 0.04919,\ 0.52123 + 2.228 \cdot 0.04919\ \right] \approx \left[0.4116,\ 0.6212\right].$$

<u>B. 3. 18.</u> (*Test der Positivität von* a_1)
Wir beschränken uns auf den Fall, dass die Standardabweichung des Störfaktors in der Grundgesamtheit σ_u nicht bekannt ist.
Dann sind folgende Schritte durchzuführen:

Schritt 1:
Formuliere der Nullhypothese H_0 und der Alternativhypothese H_1.

Schritt 2:
Wende die $t-$Verteilung an.

Schritt 3:
Bestimme die Annahme und die Ablehnungsbereiche auf Grund des Freiheitsgrades $df = n - m - 1$ und α. Die kritische Grenze sei mit $t_{\alpha, n-m-1}$ bezeichnet.

Schritt 4:
Berechne

$$s_{stat} = \frac{a_1 - A_1}{s_{a_1}}.$$

Schritt 5:
Entscheide über die Annahme oder der Ablehnung der Nullhypothese durch Verglich von $t_{\alpha, n-m-1}$ und s_{stat}.

<u>BS. 3. 8.</u> (*BS. 3. 1.*)

1.

$$H_0 : A_1 = 0, \quad H_1 : A_1 > 0.$$

2.
Da σ_u nicht bekannt ist, wird die $t-$Verteilung benutzt.

3.
Der Test ist wegen $H_1 : A_1 > 0$ rechtsseitig. $df = 14 - 2 = 12$. Damit liegt die Ablehnungsregion rechts von 1.782.

4.

$$t_{stat} = \frac{a_1 - A_1}{s_{a_1}} = \frac{0.5435}{0.0401} \approx 13.554.$$

5.
Wegen $13.5536 > 1.782$ wird die Nullhypothese abgelehnt. Dies bedeutet, dass mit zunehmendem Automatisierungsgrad der Arbeit das Niveau der Arbeitsproduktivität steigt.

<u>B. 3. 19.</u> (*Multiple nichtlineare Regression*)
Wir beschränken und auf den Fall der sog. Cobb-Douglas-Produktionsfunktion:

$$y^* = \gamma A^{\alpha} K^{\beta}, \ \alpha, \beta > 0 \quad (A: \text{ Arbeit; } K: \text{ Kapital})$$

Die Funktion wird nun linearisiert:

$$\lg y^* = \lg(\gamma A^{\alpha} K^{\beta})$$

$$\lg y^* = \lg \gamma + \alpha \cdot \lg A + \beta \cdot K.$$

Mit

$$Y^* := \lg y^*, \quad a_0 := \gamma, \quad a_1 := \alpha, \quad a_2 := \beta$$

erhält man:

$$Y^* = a_0 + A \cdot a_1 + K \cdot a_2 .$$

Die Normalgleichungen (siehe B. 3. 5.)

$$\begin{cases} n \cdot a_9 \quad + a_1 \sum_{i=1}^{n} x_{i1} \quad + a_2 \sum_{i=1}^{n} x_{i2} \quad = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_{i1} + a_1 \sum_{i=1}^{n} x_{i1}^2 \quad + a_2 \sum_{i=1}^{n} x_{i1} x_{i2} = \sum_{i=1}^{n} x_{i1} y_i \\ a_0 \sum_{i=1}^{n} x_{i2} + a_1 \sum_{i=1}^{n} x_{i1} x_{i2} + a_2 \sum_{i=1}^{n} x_{i2}^2 \quad = \sum_{i=1}^{n} x_{i2} y_i \end{cases}$$

liefern die Regressionsparameter α, β und γ. Dabei wird

$$x_{i1} := A_i, \ x_{i2} := K_i$$

gesetzt.

BS. 3. 9.
Die nachfolgende Tabelle zeigt 7 verschiedene Kombinationen von Arbeit und Kapital und die entsprechende Produktionsmenge.
Die Abhängigkeit der Produktion von den Faktoren Arbeit und Kapital soll in der Form einer Regressionsfunktion des Cobb-Douglas-Typs

$$y^* = \gamma A^\alpha K^\beta, \ \alpha, \beta, \gamma > 0$$

dargestellt werden.

Arbeit	Kapital	Produktion
20	30	1814
25	35	2142
30	42	2526
35	39	2590
40	48	3029
45	50	3242
50	46	3243

Lösung:

Arbeitstabelle

A_i	K_i	y_i	x_{i1}	x_{i2}	$\lg y_i$	x_{i1}^2	x_{i2}^2	$x_{i1} \cdot \lg y_i$	$x_{i2} \cdot \lg y_i$	$x_{i1} \cdot x_{i2}$
20	30	1814	1.3010	1.4771	3.2586	1.6927	2.1819	4.2396	4.8134	1.9218
25	35	2142	1.3979	1.5441	3.3308	1.9542	2.3841	4.6563	5.1430	2.1585
30	42	2526	1.4771	1.6232	3.4024	2.1819	2.6349	5.0258	5.5230	2.3977
35	39	2590	1.5441	1.5911	3.4133	2.3841	2.5315	5.2704	5.4308	2.4567
40	48	3029	1.6021	1.6812	3.4813	2.5666	2.8266	5.5773	5.8529	2.6934
45	50	3242	1.6532	1.6990	3.5108	2.7331	2.8865	5.8041	5.9648	2.8088
50	46	3243	1.6990	1.6628	3.5109	2.8865	2.7648	5.9650	5.8379	2.8250
245	290	18586	10.6744	11.2785	23.9082	16.3992	18.2103	36.5384	38.5657	17.2619

$$\begin{cases} 7a_0 & +10.6744a_1 + 11.2785a_2 = 23.9082 \\ 10.6744a_0 & +16.3992a_1 + 17.2619a_2 = 36.5384 \\ 11.2785a_0 & +17.2619a_1 + 18.2103a_2 = 38.5657 \end{cases}$$

Das Gleichungssystem hat die Lösung

$$a_0 = 1.9990, \quad a_1 = 0.4028, \quad a_2 = 0.4979 .$$

Damit erhalten wir

$$\gamma = 99.77, \quad \alpha = 0.4028, \quad \beta = 0.4979 .$$

Die gesuchte Regressionsfunktion lautet:

$$y = 99.77 A^{0.4028} K^{0.4979}$$

A_i	K_i	y_i	y_i^{*2}	$(y_i^* - \bar{y})^2$	$(y_i - \bar{y})^2$
20	30	1814	1813.48316	708391.05	707521.311
25	35	2142	2142.3069	263000.72	263315.595
30	42	2526	2524.64451	17029.8183	16677.8783
35	39	2590	2589.05722	4367.31216	4243.59221
40	48	3029	3029.70085	140293.688	139769.161
45	50	3242	3242.13532	344560.152	344401.303
50	46	3243	3245.13086	348085.84	345576.017
		18586	18586.45882	1825728.58	1821504.86

$$B_{y12} = \frac{1821504.86}{1825728.58} \approx 0.9977 .$$

Die Produktion wird zu etwa 99.77% durch die Faktoren Arbeit und Kapital bestimmt.

(Letzte Aktualisierung: 26.08.19)