

Kapitel 8: Regressionsanalyse

1. Lineare Einfachregression	259
a) Arten von Beziehungen	259
b) Die Regressionsgeraden	262
c) Schätzung der Regressionskoeffizienten mit der Methode der kleinsten Quadrate	263
d) Regressionskoeffizienten und Korrelationskoeffizient	268
e) Varianzzerlegung und Bestimmtheitsmaß (Determinationskoeffizient)	269
2. Bemerkungen zur Methode der kleinsten Quadrate	273
a) Eigenschaften der geschätzten Residuen	273
b) Alternativen zur Minimierung der Summe der Quadrate der Abweichungen	275
3. Ergänzungen zur linearen einfachen Regression	279
a) Standardisierte Variablen, gruppierte Daten	279
b) Exkurs zum Regressionsmodell	281
4. Multiple lineare Regression	284
a) Beschreibung des Modells	284
b) Darstellung in Matrixschreibweise	286
c) Multiple Korrelation und multiple Bestimmtheit	288
d) Partielle Regressions- und Korrelationskoeffizienten	289
e) Standardisierte Regressionskoeffizienten, Rekursionsformeln	290
5. Nichtlineare Regression	294

Die Regressionsanalyse beschäftigt sich mit der Schätzung funktionaler Beziehungen zwischen zwei oder mehreren metrisch skalierten Merkmalen. Sie steht im engen Zusammenhang mit der im Kap. 7 behandelten Korrelationsanalyse. Gegenstand der Regressionsanalyse ist die Art der Abhängigkeit zwischen den Variablen, z.B. die Beziehung zwischen dem Jahresumsatz (Y) und den Ausgaben für Werbung (X) einer Unternehmung (Einfachregression) oder zwischen dem Umsatz (Y) den Werbeausgaben (X_1) und den Wareneinkäufen (X_2) (mehrfache [multiple] Regression). Das Ziel kann dabei die Analyse (d.h. das bessere Verständnis des Kausalzusammenhangs) oder die Prognose von bestimmten Größen sein.

1. Lineare Einfachregression

a) Arten von Beziehungen

Bei der einfachen Regression wird die Beziehung zwischen **zwei** Variablen untersucht. Dabei wird davon ausgegangen, dass eine Variable Y von der anderen Variablen X abhängig ist. Daher auch die folgenden synonymen Bezeichnungen:

Variable Y : abhängige-, zu erklärende-, endogene Variable oder Regressand;

Variable X : unabhängige-, erklärende-, einflussausübende-, exogene (besser: vorherbestimmte) Variable, Prädiktor oder Regressor.

Es sind nun verschiedene Begriffspaare zu definieren, nämlich:

- funktionaler und stochastischer Zusammenhang
- einfache und multiple Regression
- lineare und nichtlineare Regression

Def. 8.1: Zusammenhang, Arten von Regressionsfunktionen

- a) Ist Y funktional (deterministisch) abhängig von X , d.h. $y = f(x)$ [Y ist eine Funktion von X], so ist jedem Wert von X ein und nur ein Wert von Y zugeordnet. Bei einer stochastischen Beziehung ist diese Funktion, die **Regressionsfunktion**, von einer **Störgröße** (Restgröße, Residuum) U überlagert (i.d.R. additiv), so dass für die einzelne Beobachtung gilt $y_v = f(x_v) + u_v$. Nach der Art der Regressionsfunktion (d.h. des funktionalen Teils der stochastischen Beziehung) unterscheidet man:
- b) einfache und multiple Regression:
Bei der **einfachen** Regression werden nur zwei Variablen X und Y betrachtet. Von **multipler** Regression spricht man, wenn es eine abhängige Variable Y und mehrere unabhängige Variablen X_1, X_2, \dots, X_p gibt.
- c) lineare und nichtlineare Regression:
Eine Regressionsfunktion ist **linear** (in den Variablen und in den Parametern), wenn gilt

$y_v = a + bx_v$ [a und b heißen Regressionskoeffizienten] (einfache lineare Regression) oder

$y_v = b_0 + b_1x_{1v} + b_2x_{2v} + \dots + b_px_{pv}$ (multiple lineare Regression, p Regressoren),

andernfalls ist sie **nichtlinear** (vgl. Abschn. 5).

Im folgenden soll zunächst nur die einfache lineare Regression betrachtet werden. Die Regressionsfunktion ist dann eine Gerade im x,y-Koordinatensystem des Streudiagramms. Sie stellt den funktionalen (systematischen) Teil des Zusammenhangs zwischen den beiden Variablen X und Y dar. Die Beobachtungen streuen als Punkte, je nach Höhe der Korrelation r_{xy} (vgl. Def. 7.8) mehr oder weniger um die Regressionsgerade.

Bemerkungen zu Def. 8.1:

1. Ist Y eine Funktion von X, also $y = f(x)$ so ist z.B. dem Wert $X = x_1$ **ein** Wert $Y = y_1$ zugeordnet. Ist die Beziehung dagegen stochastisch, so können einem Wert x_1 **mehrere** Werte von Y zugeordnet sein, die auch oberhalb oder unterhalb der Funktion $f(x)$ streuen können, etwa

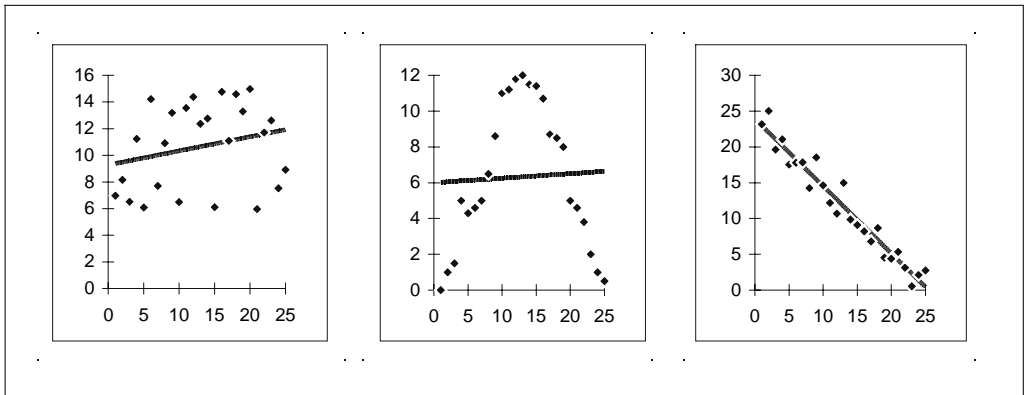
$$Y = y_{11} = f(x_1) + u_1 = y_1 + u_1 < y_1$$
 oder

$$Y = y_{12} = f(x_1) + u_2 = y_1 + u_2 > y_1$$
 je nachdem, welchen Wert U annimmt, z.B. einen positiven ($u_2 > 0$) oder einen negativen Wert ($u_1 < 0$).
2. Die Störgröße U erlaubt es, Zusammenhänge zu betrachten, die nicht exakt einer Funktion folgen, wie dies bei praktisch allen empirischen Daten der Fall ist, sei es wegen der Fehlerhaftigkeit der Messung, sei es weil Y noch von anderen Größen als X determiniert wird. Damit wird es auch möglich, systematische Zusammenhänge bei Beobachtungsdaten festzustellen, d.h. dann, wenn im Unterschied zum Experiment nicht alle Einflußgrößen kontrolliert werden können. Das Residuum U ist Ausdruck aller sonstiger, d.h. anders als X nicht explizit berücksichtigter, nicht kontrollierter oder auch gar nicht bekannter Einflüsse auf Y (weshalb U auch zufällig variieren sollte).
3. Es ist zu unterscheiden zwischen der rein deskriptiven Behandlung der Regression und dem stochastischen Modell, das hier erst später dargestellt wird (Abschn. 3c). Im ersten Fall sind keine Annahmen über

U erforderlich und die Schätzung einer Regressionsfunktion reduziert sich zu einem Problem der Kurvenanpassung, d.h. der Suche nach einer Kurve, die sich einer Punktwolke im Streuungsdiagramm bestmöglich anpaßt.

4. Vor jeder einfachen Regressionsanalyse sollte ein Streuungsdiagramm gezeichnet werden, denn das Streuungsdiagramm (vgl. Bsp. 7.2) erlaubt Rückschlüsse über
 - den jeweiligen Funktionstyp (lineare - oder nichtlineare einfache Regression), vgl. hierzu Abb. 8.1;
 - die Höhe der Korrelation (als Maß der Güte der Anpassung).

Abb. 8.1: Verschiedene Streuungsdiagramme



In Abb. 8.1 sind beispielhaft drei Streuungsdiagramme (mit Regressionsgeraden \hat{y}) gegenübergestellt. Wie leicht zu sehen ist, kann man aus der ersten (linken) Punktwolke auf keinen bzw. einen geringen positiven ($r = +0,2408$) Zusammenhang, aus der zweiten Punktwolke auf einen parabolischen und aus der dritten Punktwolke auf einen beträchtlichen negativen ($r = -0,9727$) linearen Zusammenhang der Variablen X und Y schließen (Ein Streuungsdiagramm mit $r = 0$ ist auch in Abb. 7.1 und 7.3).

5. Die Beschränkung auf lineare Zusammenhänge ist vertretbar, weil
 - viele ökonomische Beziehungen zumindest in guter Näherung durch lineare Funktionen dargestellt werden können,
 - lineare Funktionen relativ einfach zu handhaben und zu interpretieren sind,
 - auch nichtlineare Funktionen häufig linearisiert werden können.

b) Die Regressionsgeraden

Ist ein linearer Zusammenhang zwischen zwei Variablen zu vermuten, so kann dieser durch Schätzung der Regressionskoeffizienten a und b (oder c und d) näher spezifiziert werden.

Def. 8.2: Regressionsgerade

- a) Die lineare Regressionsfunktion (Regressionsgerade) zur Bestimmung von Y (abhängige Variable) durch X (unabhängige Variable) lautet:

$$(8.1) \quad \hat{y}_v = a + bx_v$$

dabei ist \hat{y}_v der Regresswert für die v -te Beobachtung (Einheit) mit $v = 1, 2, \dots, n$ und für die einzelne Beobachtung (x_v, y_v) gilt

$$(8.1a) \quad y_v = \hat{y}_v + u_v = a + bx_v + u_v,$$

d.h. die geschätzte Störgröße u_v für die v -te Beobachtung ist der senkrechte Abstand zwischen y_v und \hat{y}_v im x, y -Koordinatensystem.

- b) Die Größen a und b werden Regressionskoeffizienten genannt, wobei a den Ordinatenabstand und b die Steigung der Regressionsgeraden angibt. Es gilt, die Parameter a und b (mit der Methode der kleinsten Quadrate) sowie s_u^2 (Varianz der Störgröße) zu schätzen.
- c) Der Zusammenhang zwischen abhängiger und unabhängiger Variable ist rein rechnerisch vertauschbar, d.h. neben der Regressionsgeraden nach Gl. 8.1 ist auch

$$(8.2) \quad \hat{x}_v = c + dy_v$$

zu berechnen, wobei dann für x_v gilt

$$(8.2a) \quad x_v = c + dy_v + v_v.$$

Die Störgröße V ist jeweils der waagrechte Abstand zwischen einem Beobachtungspunkt x_v und \hat{x}_v im x, y -Koordinatensystem.

Bemerkungen zu Def. 8.2:

1. Die Regressionskoeffizienten a und b (bzw. c und d) werden mit der Methode der kleinsten Quadrate geschätzt.

2. Die lineare Einfachregression wird hier zunächst nur als bloße Rechentechnik im Rahmen der Deskriptiven Statistik betrachtet. Es ist dann auch zulässig, die Abhängigkeiten durch Vertauschung von X und Y umzukehren, d.h. statt der Regressionsgeraden $\hat{y}_v = a + bx_v$ die Regressionsgerade $\hat{x}_v = c + dy_v$ zu schätzen. Die Formel zur Berechnung von c (bzw. d) ergibt sich aus derjenigen für a (bzw. b) durch Vertauschung von X und Y . Man kann grundsätzlich zwei Regressionsgeraden berechnen, aber meist nur eine der beiden kausal interpretieren.
3. Ob eine Variable abhängig (endogen) oder unabhängig (exogen) ist, liegt für das Modell der Regressionsanalyse nicht in ihrem Inhalt, ihrer sachlichen Interpretation begründet, sondern allein darin, ob die Variable im Rahmen des Modells durch eine Gleichung "erklärt" wird (endogen) oder nicht (exogen). Eine unabhängige (exogene) Variable ist annahmegemäß nicht mit der Störgröße der Gleichung in der sie auftritt korreliert. Auf die Annahmen des zu schätzenden Modells der Regression wird erst im Abschn. 3c eingegangen.

c) Schätzung der Regressionskoeffizienten mit der Methode der kleinsten Quadrate

Die Regressionskoeffizienten sind mit einem eindeutigen und objektiven Verfahren so zu bestimmen, dass sich die Regressionsgerade der Punktwolke im Streudiagramm bestmöglich anpaßt. Alle Kriterien zur Schätzung der Regressionskoeffizienten a und b (bzw. c und d) gehen von den Residuen u_v (bzw. v_v) aus. Ein solches Verfahren ist die Methode der kleinsten Quadrate. Bei ihr werden die Regressionskoeffizienten so bestimmt, dass die Summe der Quadrate der Abweichungen von der Regressionsgeraden ein Minimum annimmt.

Aus Gleichung (8.1a) folgt

$$u_v^2 = (y_v - a - bx_v)^2 \text{ mit } v = 1, 2, \dots, n \text{ und damit:}$$

$$\Sigma u_v^2 = \Sigma (y_v^2 - ay_v - bx_v y_v - ay_v + a^2 + abx_v - bx_v y_v + abx_v + b^2 x_v^2)$$

$$= \Sigma (y_v^2 - 2ay_v - 2bx_v y_v + a^2 + 2abx_v + b^2 x_v^2).$$

Mithin gilt:

$$(8.3) \quad \Sigma u_v^2 = \Sigma y_v^2 - 2a \Sigma y_v - 2b \Sigma x_v y_v + na^2 + 2ab \Sigma x_v + b^2 \Sigma x_v^2$$

Die Summe Σu_v^2 ist bei gegebenen Beobachtungen (also auch gegebenen Werten Σx_v , Σx_v^2 , Σy_v , Σy_v^2 und $\Sigma x_v y_v$) eine Funktion $Q(a, b)$ von a und b und es gilt a und b so zu bestimmen, dass $\Sigma u_v^2 = Q(a, b)$ minimal ist. Hierzu ist lediglich Σu_v^2 gem. Gl. 8.3 partiell

nach a und b zu differenzieren und die beiden Ableitungen sind Null zu setzen, denn man kann leicht zeigen, dass $Q(a,b)$ kein Maximum, sondern nur ein Minimum besitzt [Man kann eine "Regressionsgerade" auch beliebig weit außerhalb der Punktwolke des Streudiagramms legen so dass $\sum u_i^2$ beliebig groß werden kann].

Das ergibt die folgenden zwei Gleichungen:

$$1) \frac{dQ}{da} = -2\sum y_v + 2na + 2b\sum x_v \stackrel{!}{=} 0$$

$$2) \frac{dQ}{db} = -2\sum x_v y_v + 2a\sum x_v + 2b\sum x_v^2 \stackrel{!}{=} 0 \quad ,$$

die umgeformt das System der **Normalgleichungen** darstellen:

(8.4a) $an + b\sum x_v = \sum y_v$ 1. Normalgleichung
(8.4b) $a\sum x_v + b\sum x_v^2 = \sum x_v y_v$ 2. Normalgleichung

Wird dieses Normalgleichungssystem nach a und b aufgelöst so erhält man als Schätzwerte für die Regressionskoeffizienten a und b:

(8.5a) $a = \frac{\sum x_v^2 \sum y_v - \sum x_v \sum x_v y_v}{n\sum x_v^2 - (\sum x_v)^2}$	(8.5b) $b = \frac{n\sum x_v y_v - \sum x_v \sum y_v}{n\sum x_v^2 - (\sum x_v)^2}$
---	---

Mit $n\sum x_v^2 - (\sum x_v)^2$ steht die n^2 -fache Varianz der Variablen X im Nenner von a und b gem. Gl. 8.5 und die Steigung b ist das Verhältnis von Kovarianz und Varianz:

(8.6a) $b = \frac{\sum (x_v - \bar{x})(y_v - \bar{y})}{\sum (x_v - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \quad .$
--

Wie man leicht sieht, gilt aufgrund der ersten Normalgleichung:

(8.6b) $a = \bar{y} - b\bar{x}$

d.h. die Regressionsgerade $\hat{y}_v = a + bx_v$ verläuft durch den Schwerpunkt $P(x,y)$ des Streudiagramms.

Man kann also a und b bestimmen

1. aufgrund der beiden Normalgleichungen und
2. indem man zunächst b nach Gl. 8.6a und dann a gem. Gl. 8.6b berechnet.

Regressionsgerade $\hat{x} = c + dy$

Man erhält die entsprechenden Formeln zur Bestimmung von c und d, indem man in den Normalgleichungen (Gl. 8.4) bzw. in den Formeln (Gl. 8.6) für a und b einfach x und y vertauscht:

$$(8.4a^*) \quad cn + d\Sigma y_v = \Sigma x_v$$

$$(8.4b^*) \quad c\Sigma y_v + d\Sigma y_v^2 = \Sigma x_v y_v$$

$$(8.6a^*) \quad d = \frac{S_{xy}}{S_y^2}$$

$$(8.6b^*) \quad c = \bar{x} - d\bar{y}$$

denn aufgrund der ersten Normalgleichung (Gl. 8.4a*) verläuft auch die Funktion $\hat{x} = c + dy$ durch den Schwerpunkt [die beiden Regressionsgeraden schneiden sich also in diesem Punkt].

Die Zusammenhänge werden an einem einfachen Zahlenbeispiel demonstriert und anschließend einige Eigenschaften der Regressionsgeraden interpretiert.

Beispiel 8.1:

Gegeben seien die folgenden Daten für n = 4 Personen (Es ist klar, dass man in der Praxis nicht bei n = 4 Werten eine Regressionsgerade schätzen würde. Das Beispiel soll aber möglichst einfach und überschaubar sein):

x_v	2	3	7	8
y_v	4	5	10	5

Für die folgenden Berechnungen empfiehlt es sich, eine Arbeitstabelle aufzustellen:

x_v	y_v	x_v^2	y_v^2	$x_v y_v$	$(x_v - \bar{x})^2$	$(y_v - \bar{y})^2$	$(x_v - \bar{x})(y_v - \bar{y})$
2	4	4	16	8	$(2-5)^2=9$	$(4-6)^2=4$	$(-3)(-2)=6$
3	5	9	25	15	$(3-5)^2=4$	$(5-6)^2=1$	$(-2)(-1)=2$
7	10	49	100	70	$2^2=4$	$4^2=16$	8
8	5	64	25	40	$3^2=9$	$(-1)^2=1$	-3
Σ	20	24	126	166	26	22	13

Es ist nützlich, zunächst die Parameter der Randverteilungen und der zweidimensionalen Verteilung gem. Kap. 7 zu berechnen:

$$\text{Mittelwerte: } \bar{x} = \Sigma x_v / n = 20/4 = 5$$

$$\bar{y} = \Sigma y_v / n = 24/4 = 6$$

$$\text{Varianzen: } s_x^2 = [\Sigma (x_v - \bar{x})^2] / n = 26/4 = 6,5$$

$$s_y^2 = [\Sigma (y_v - \bar{y})^2] / n = 22/4 = 5,5$$

$$\text{Kovarianz: } s_{xy} = [\Sigma (x_v - \bar{x})(y_v - \bar{y})] / n = 13/4 = 3,25$$

Berechnung der Regressionsgeraden $\hat{y} = a + b \cdot x$ und $\hat{x} = c + d \cdot y$:

Die Normalgleichungen lauten

für a und b:

für c und d:

$an + b\Sigma x_v = \Sigma y_v$	$cn + d\Sigma y_v = \Sigma x_v$
$a\Sigma x_v + b\Sigma x_v^2 = \Sigma x_v y_v$	$c\Sigma y_v + d\Sigma y_v^2 = \Sigma x_v y_v$
$4a + 20b = 24$	$4c + 24d = 20$
$20a + 126b = 133$	$24c + 166d = 133$

Daraus folgt: $a = 3,5$ und $b = 1/2$ sowie

$$c = 32/22 = 1,4545 \text{ und } d = 13/22 = 0,5909$$

Ferner gilt: $b = s_{xy} / s_x^2 = 13/26 = 1/2$ und $d = s_{xy} / s_y^2 = 13/22 = 0,5909$.

Die Regressionsgeraden schneiden sich im Schwerpunkt, d.h.:

$$a = \bar{y} - b \cdot \bar{x} = 3,5 \text{ und } c = \bar{x} - d \cdot \bar{y} = 32/22 = 1,4545.$$

Für die beiden Regressionsgeraden erhält man also:

$$\hat{y}_v = 3,5 + 0,5x_v \quad \text{und} \quad \hat{x}_v = 32/22 + (13/22)y_v.$$

Die Berechnung der Funktion $Q(a,b)$

$$Q(a,b) = \Sigma u_v^2 = \Sigma y_v^2 - 2a\Sigma y_v - 2b\Sigma x_v y_v + na^2 + 2ab\Sigma x_v + b^2\Sigma x_v^2$$

für dieses Beispiel führt zu

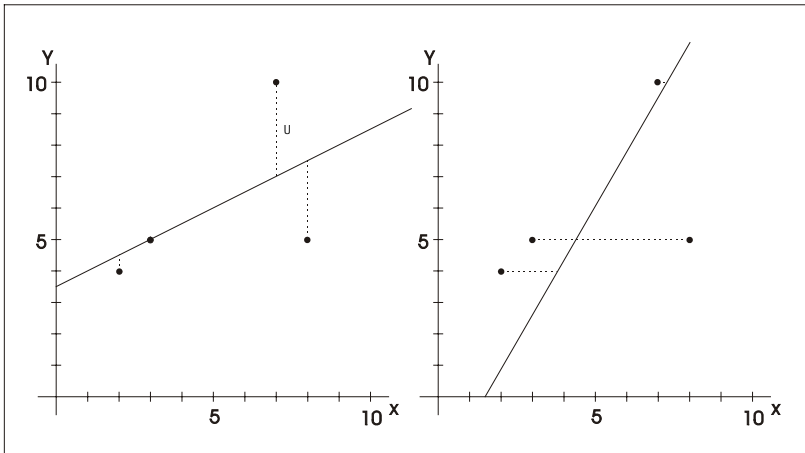
$$Q = \Sigma u_v^2 = 166 - 48a - 266b + 4a^2 + 40ab + 126b^2,$$

einer Funktion mit einem Minimum von $\Sigma u_v^2 = 15,5$. Die folgende Tabelle gibt einige Funktionswerte von $Q(a,b)$ in der Nähe des Minimums an:

	b=0,3	b=0,4	b=0,5	b=0,6	b=0,7
a=3,3	22,30	17,72	15,66	16,12	19,10
a=3,4	21,38	17,20	15,54	16,40	19,78
a=3,5	20,54	16,76	15,50	16,76	20,54
a=3,6	19,78	16,40	15,54	17,20	21,38
a=3,7	19,10	16,12	15,66	17,72	22,30

Mit beispielsweise $b = 0,5$ gilt $Q(a, b=1/2) = 64,5 - 28a + 4a^2$, was eine Parabel im a, Q -Koordinatensystem ist. Man sieht, dass $Q = \sum u_v^2$ in der Tat an der Stelle $a = 3,5$ mit $\sum u_v^2 = 15,5$ ein Minimum hat.

Abb. 8.2: Streuungsdiagramm und Regressionsgeraden für Bsp. 8.1



Wie man sieht, schneiden sich die beiden Regressionsgeraden im Punkt $S (\bar{x}=5, \bar{y}=6)$, dem Schwerpunkt.

Mit diesem Demonstrationsbeispiel lassen sich auch einige Eigenschaften der Methode der kleinsten Quadrate zeigen. Man kann z.B. die in Abb. 8.2 eingezeichneten (senkrechten) Residuen u_v und die waagrechten Residuen v_v bestimmen und die Residuen u_v der Regressionsgerade $\hat{y}_v = a + bx_v = 3,5 + 1/2x_v$ vergleichen mit den Abweichungen $u_{v,*}$ um eine alternative Gerade, etwa um $\hat{y}_v^* = 3,4 + 0,6x_v$:

Regressionsgerade $\hat{y}_v = 3,5 + 0,5x_v$					alternative Gerade $\hat{y}_v^* = 3,4 + 0,6x_v$		
x_v	y_v	\hat{y}_v	u_v	u_v^2	\hat{y}_v^*	u_v^*	$(u_v^*)^2$
2	4	4,5	-0,5	0,25	4,6	-0,6	0,36
3	5	5	0	0	5,2	-0,2	0,04
7	10	7	3	9	7,6	2,4	5,76
8	5	7,5	-2,5	6,25	8,2	-3,2	10,24
Σ	20	24	0	15,5	25,6	-1,6	16,4

Wie man sieht gilt: $\Sigma u_v = 0$ und $\Sigma y_v = \Sigma \hat{y}_v = 24$ (beides infolge der ersten Normalgleichung), während dies für eine andere Gerade nicht gelten muss. Abgesehen davon, dass die Summe der Abweichungsquadrate $\Sigma (u_v^*)^2$ mit 16,4 größer ist als $\Sigma u_v^2 = 15,5$ ist auch u_v mit x nicht korreliert, wohl aber u_v^* mit x (denn die Kovarianz zwischen u_v^* und x beträgt $(\Sigma x u_v^*)/n - \bar{u}^* \bar{x} = -2,65 - 5(-0,4) = -0,65$).

Für die Störgröße v in der Regression von x (abhängig) auf y (unabhängig) gilt:

$$v_v = x_v - \hat{x}_v = x_v - [32/22 + (13/22)y_v].$$

y_v	x_v	\hat{x}_v	v_v	v_v^2
4	2	3,818	-1,818	3,3058
5	3	4,409	-1,409	1,9855
10	7	7,364	-0,364	0,1322
5	8	4,409	+3,591	12,8946
Σ	24	20	0	18,3182

d) Regressionskoeffizienten und Korrelationskoeffizient

Nach Def. 7.8 gilt für den Korrelationskoeffizienten

$$(8.7) \quad r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y} = \begin{cases} +\sqrt{bd} & \text{wenn } b, d > 0 \\ -\sqrt{bd} & \text{wenn } b, d < 0 \end{cases}$$

Weil das Vorzeichen von b und d allein durch die Kovarianz s_{xy} bestimmt wird, haben b und d stets das gleiche Vorzeichen. Der (lineare) Korrelationskoeffizient r_{xy} ist also das geometrische Mittel der Steigungen der beiden Regressionsgeraden.

Im Unterschied zu b und d ist r invariant gegenüber linearen Transformationen von X und Y (vgl. Bem. Nr. 2 zu Def. 7.8). Mit den Lineartransformationen

$$x^* = p_1 + q_1 \cdot x \text{ und } y^* = p_2 + q_2 \cdot y$$

erhält man für die Koeffizienten der Regressionsfunktion

$$\begin{array}{llll} y^* = a^* + b^* x^* & a^* = p_2 + q_2 a - p_1 b^* & \text{und} & b^* = b(q_2/q_1) \\ x^* = c^* + d^* y^* & c^* = p_1 + q_1 c - p_2 d^* & \text{und} & d^* = d(q_1/q_2). \end{array}$$

Die Regressionskoeffizienten a, b, c und d werden also von Maßstabsänderungen in den Skalen von X und/oder Y berührt, der Korrelationskoeffizient dagegen nicht.

e) Varianzzerlegung und Bestimmtheitsmaß (Determinationskoeffizient)

Die Regressionsfunktion $\hat{y} = a + bx$ bedeutet, dass die Streuung von Y wegen der linearen Abhängigkeit von X zum Teil durch die Streuung von X erklärt werden kann. Die Varianz von Y ist zu zerlegen in einen auf \hat{y} (und damit auf X) zurückgehenden Teil und in eine Residualvarianz:

Man kann dies zeigen, indem man den Abstand (die Abweichung) eines

Datenpunkts vom Mittel zerlegt in zwei Abstände u_v und $\hat{y}_v - \bar{y}$:

$$(y_v - \bar{y}) = (y_v - \hat{y}_v) + (\hat{y}_v - \bar{y}) \quad \text{wobei} \quad u_v = (y_v - \hat{y}_v).$$

Berücksichtigt man, dass gilt (vgl. Gl. 8.11ff.) $\Sigma y = \Sigma \hat{y}$, $\Sigma u_v = 0$ und $\Sigma \hat{y}_v u_v = \Sigma x_v u_v = 0$, so ergibt sich

$$\begin{array}{ccccc} (y_v - \bar{y}) & = & (\hat{y}_v - \bar{y}) & + & (u_v - \bar{u}) \\ \text{T} & & \text{E} & & \text{R} \end{array}$$

Die beiden Differenzen E und R auf der rechten Seite dieser Identität deuten auf zwei Variationsquellen hin:

E: Die Differenz von dem durch die Regression geschätzten Wert (dem Regresswert) \hat{y}_v und dem arithmetischen Mittel \bar{y} ist verantwortlich für die durch die Regression "erklärte" Streuung.

R: Die Abweichung des beobachteten Wertes y_v von dem geschätzten Wert \hat{y}_v (d.h. das Residuum) kann als "durch die Regression nicht erklärte Abweichung" bezeichnet werden.

Durch Quadrieren der Abstandsgleichung und anschließender Summation erhält man:

$$\begin{array}{rcccc} \Sigma(y_v - \bar{y})^2 & = & \Sigma(y_v - \hat{y}_v)^2 & + & 2\Sigma(y_v - \hat{y}_v)(\hat{y}_v - \bar{y}) & + & \Sigma(\hat{y}_v - \bar{y})^2 \\ \Sigma T^2 & = & \Sigma R^2 & + & 2\Sigma RE & + & \Sigma E^2 \end{array}$$

Wie leicht zu sehen ist, fällt der mittlere Ausdruck auf der rechten Seite weg, denn wegen $\sum u_v = \sum u_v \cdot x_v = 0$ gilt

$$\sum (y_v - \hat{y}_v)(\hat{y}_v - \bar{y}) = \sum u_v(\hat{y}_v - \bar{y}) = \sum u_v(a + bx_v - \bar{y}) = 0.$$

Somit gilt auch (Varianzzerlegung)

(8.8)	$\frac{1}{n}\sum (y_v - \bar{y})^2$	=	$\frac{1}{n}\sum (\hat{y}_v - \bar{y})^2$	+	$\frac{1}{n}\sum (y_v - \hat{y}_v)^2$
	totale Varianz		=		erklärte Varianz + Residualvarianz
	s_y^2		=		$s_{\hat{y}}^2 + s_u^2$

Diese Gleichung legt den Gedanken nahe, die erklärte Varianz $s_{\hat{y}}^2$ und die Residualvarianz $s_u^2 = n^{-1}\sum (u_v - \bar{u})^2 = n^{-1}\sum (u_v - 0)^2 = (\sum u_v^2)/n$ durch die Gesamtvarianz zu dividieren. Die Anteile sind nach Def. 7.10 das **Bestimmtheitsmaß** B_{yx} (Bestimmtheit von Y durch X) und das **Unbestimmtheitsmaß**¹ $U_{yx} = 1 - B_{yx}$. Offenbar gilt $0 \leq B_{yx} \leq 1$ und entsprechend $0 \leq U_{yx} \leq 1$, weil B und U Varianzanteile darstellten.

Speziell für die einfache lineare Regression gilt für das Bestimmtheits- und Unbestimmtheitsmaß:

1. Symmetrie: $B_{yx} = B_{xy}$ mit $B_{xy} = \frac{s_x^2}{s_y^2}$
2. Das Bestimmtheitsmaß B_{yx} ist das Quadrat des Korrelationskoeffizienten r_{xy} ($B_{yx} = r_{xy}^2$).

Anders als beim linearen Korrelationskoeffizient r kann man mit dem Bestimmtheitsmaß r^2 nicht zwischen positiver und negativer Korrelation differenzieren.

Wegen $\hat{y}_v = a + bx_v$ und $\bar{y} = a + b\bar{x}$ erhält man für die erklärte Varianz $s_{\hat{y}}^2$ auch $s_{\hat{y}}^2 = b^2 s_x^2$ und weil $b = s_{xy}/s_x^2$

(8.9)	$B_{yx} = \frac{s_{xy}^2}{s_x^2 s_y^2} = b \cdot d = r_{xy}^2.$
-------	---

¹ Die Größe U (Unbestimmtheitsmaß) und u (Störgröße) sollte nicht verwechselt werden.

Entsprechend ist die mit X erklärte Varianz $s_x^2 = d^2 s_y^2$ und damit

$$B_{xy} = \frac{d^2 s_y^2}{s_x^2} = \frac{s_{xy}^2}{s_x^2 s_y^2} = b \cdot d = r_{xy}^2 = B_{yx}$$

Für die Grenzfälle $r_{xy}^2 = 1$ und $r_{xy}^2 = 0$ gilt:

- $r_{xy}^2 = B_{yx} = 1$ ($U_{yx} = 0$)
 $u_v = 0$ und $v_v = 0$ für alle $v = 1, 2, \dots, n$, alle n Beobachtungen liegen genau auf den Regressionsgeraden, die in diesem Fall "zusammenfallen" und lauten $\hat{y} = a + bx$ und $\hat{x} = (-a/b) + (1/b)y$
- $r_{xy}^2 = B_{yx} = 0$ ($U_{yx} = 1$)
für jedes v ist der Regresswert \hat{y}_v identisch mit dem Mittelwert \bar{y} (die y Regressionsgerade verläuft also parallel zur x-Achse und hat die Steigung $b = 0$ [sie lautet $\hat{y} = \bar{y}$]) und für jedes v ist der Regresswert \hat{x}_v identisch mit dem Mittelwert \bar{x} . Die Regressionsgerade \hat{x} lautet $\hat{x} = \bar{x}$ und steht senkrecht auf der Regressionsgeraden $\hat{y} = \bar{y}$.

Man könnte vermuten, dass der Winkel α zwischen den Regressionsgeraden mit dem Korrelationskoeffizienten zusammenhängt, da $\alpha = 0^\circ$ bedeutet $r = +1$ oder $r = -1$ und $\alpha = 90^\circ$ impliziert $r = \pm 0$. Der Zusammenhang ist jedoch nicht ganz so einfach. Man kann leicht zeigen, dass gilt

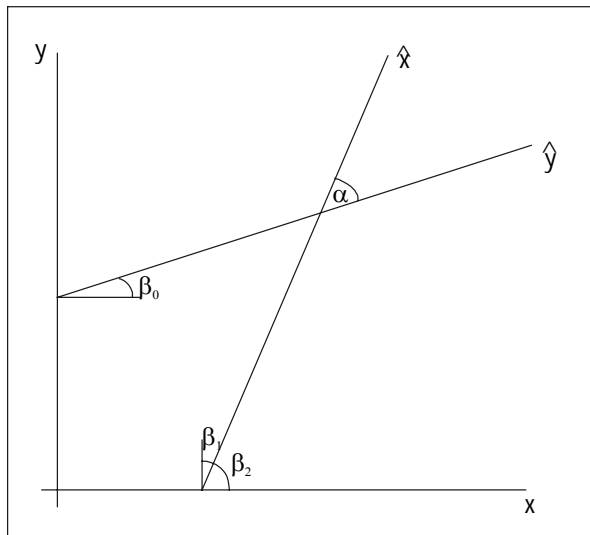
$$(8.10) \quad \tan(\alpha) = \frac{1-bd}{b+d} = \frac{s_x s_y (1-r^2)}{r(s_x^2 + s_y^2)}$$

Abb. 8.3: Zur Geometrie der Regressionsgeraden

und dass die Steigung $\tan(\beta_2)$ der Regressionsgeraden \hat{x} im x,y-Koordinatensystem betragsmäßig stets größer ist als die Steigung b der Regressionsgeraden \hat{y} , denn (vgl. Abb. 8.3) es gilt:

$$\begin{aligned} \tan(\beta_0) &= b = s_{xy}/s_x^2; \\ \tan(\beta_1) &= d = s_{xy}/s_y^2; \\ \tan(\beta_2) &= \cot(\beta_1) = 1/d, \end{aligned}$$

so dass sich die Behauptung über die Steigungen aus der Schwarzen Ungleichung ergibt.



Beispiel 8.2:

Man verifiziere Gl. 8.7 bis 8.9 für das Bsp. 8.1!

Lösung 8.2:

Für das Bsp. 8.1 erhält man: $r = 13/\sqrt{22 \cdot 26} = 0,54356$, da die Kovarianz $13/5$ und die Varianzen $s_x^2 = 22/5$ und $s_y^2 = 26/5$ betragen. Die Steigungen sind $b = 0,5$ und $d = 13/22$, so dass $r = \sqrt{13/44} = 0,54356$. Ferner erhält man $\Sigma u_v^2 = 15,5$ und $\Sigma (y_v - \bar{y})^2 = 22$, so dass gilt $U_{yx} = 15,5/22 = 0,7045 = 1 - r_{yx}^2$ (es gilt $r_{yx}^2 = 1 - U_{yx} = 1 - 0,7045 = 0,2954 = (0,54356)^2$) und $\Sigma v_v^2 = 18,318$ und $\Sigma (x_v - \bar{x})^2 = 26$, so dass gilt $U_{xy} = 18,318/26 = 0,7045 = 1 - r_{xy}^2$. Man sieht, dass die Bestimmtheit hier 29,54% beträgt und dass $B_{yx} = B_{xy}$ das Quadrat des Korrelationskoeffizienten ist.

Beispiel 8.3:

Es sei X der Intelligenzquotient (IQ) des Vaters und Y der des Sohnes. Psychologen fanden heraus, dass der IQ (praktisch in allen Generationen) mit Mittelwert 100 und Standardabweichung 16,4 symmetrisch verteilt ist, und dass die Korrelation zwischen dem IQ des Vaters und des Sohnes $r_{xy} = +0,5$ ist:

- Man bestimme die Regressionsgerade $\hat{y} = a + bx$.
- Welcher IQ ist für den Sohn zu erwarten, wenn der Vater einen IQ von 75 (d.h. leichte Debilität) und welcher, wenn der Vater einen IQ von 130 (überragende Intelligenz) hat?
- Kann man aufgrund der Ergebnisse schließen, dass durch die Vererbung ein unaufhaltsamer Trend zum Mittelmaß besteht, so dass es nach einigen Generationen nur noch Personen mit einem IQ von 100 gibt?
- Wie ändert sich die Situation hinsichtlich der unter c) gegebenen Interpretation, wenn die Korrelation nicht $r_{xy} = +0,5$, sondern $r_{xy} = +0,25$ beträgt.

Lösung 8.3:

- Aus $r_{xy} = 0,5$ folgt wegen $s_x = s_y = 16,4$ für die Kovarianz $s_{xy} = r_{xy}s_x s_y = \frac{1}{2} \cdot (16,4)^2 = \frac{1}{2} s_x^2$ und damit $b = s_{xy}/s_x^2 = \frac{1}{2}$. Aus $\bar{y} = a + b\bar{x}$ folgt dann wegen $\bar{x} = \bar{y} = 100$ und $b = \frac{1}{2}$ für a der Wert: $a = 50$.
Die Regressionsgerade lautet also $\hat{y} = 50 + \frac{1}{2}x$ (Sohn in Abhängigkeit vom Vater)

- b) Die Fragen nach dem zu erwartenden IQ des Sohnes bei gegebenem IQ des Vaters laufen auf eine deterministische (d.h. funktionale) Interpretation der Regressionsfunktion hinaus. Man kann mit der Regressionsfunktion leicht die folgenden Werte nachrechnen:

Vater (x)	Sohn (\hat{y})	\hat{y} ist
75 (unter 100)	$50 + \frac{1}{2} \cdot 75 = 87,5$	größer als x
130 (über 100)	$50 + \frac{1}{2} \cdot 130 = 115$	kleiner als x

Man erkennt, dass grundsätzlich gilt: $\hat{y} > x$ wenn $x < 100$ (der Sohn ist intelligenter als der Vater, wenn dieser unterdurchschnittlich intelligent ist) und umgekehrt $\hat{y} < x$ wenn $x > 100$.

- c) Das legt die (verfehlte) Interpretation, auf die hier bezug genommen wird nahe. Sie ist verfehlt, weil die Regressionsfunktion deterministisch interpretiert wird, wozu bei einer Bestimmtheit von nur 25% ($r^2 = 0,25$) kein Anlaß besteht. Zu einem ähnlichen (Fehl-) Schluß gelangten amerikanische Wirtschaftsforscher bei einer langfristigen Analyse von Unternehmensgewinnen). Man bezeichnet den dargestellten Zusammenhang auch als regression to the mean.

- d) Die Regressionsgerade lautet jetzt: $\hat{y} = 75 + \frac{1}{4} x$ und die Regression zum Mittel (in der falschen Interpretation also der Trend zum Mittelmaß) ist noch schneller: denn ist $x = 75$, so ist (deterministisch interpretiert) $\hat{y} = 93,75$ statt 87,5 und bei $x = 130$ ist $\hat{y} = 107,5$ statt 115. Wegen $s_x = s_y$ ist generell bei diesem Beispiel die Regressionsfunktion

$$\hat{y} = 100(1 - r) + rx$$

und für die Abweichung vom Mittelwert gilt

$$(\hat{y} - 100) = r(x - 100).$$

Sie ist also bei gegebenen x umso kleiner (und damit die Geschwindigkeit der regression to mean umso größer) je kleiner r ist.

2. Bemerkungen zur Methode der kleinsten Quadrate

a) Eigenschaften der geschätzten Residuen

Aus den beiden Normalgleichungen (Gl. 8.4a und 8.4b) für die Bestimmung der Regressionsgerade $\hat{y} = a + bx$

$an + b\sum x_v = \sum y_v$	1. Normalgleichung
$a\sum x_v + b\sum x_v^2 = \sum x_v y_v$	2. Normalgleichung

ergeben sich die folgenden Eigenschaften der geschätzten Störgröße u (für v in der Regression $\hat{x} = c + dy$ gelten die folgenden Ausführungen analog), die ihrerseits gewisse Folgerungen erlauben:

- aus der ersten Normalgleichung folgt:

1. Die Summe der Residuen (und somit auch der Mittelwert \bar{u}) ist Null

$$(8.11) \quad \Sigma u_v = n\bar{u} = 0$$

oder äquivalent

2. Die geschätzte Regressionsgerade verläuft durch den Schwerpunkt
3. Die Regresswerte \hat{y} und die beobachteten y -Werte sind in der Summe und damit auch im Mittel gleich

$$(8.12) \quad \Sigma \hat{y}_v = \Sigma y_v \quad \text{und} \quad \bar{\hat{y}} = \bar{y}.$$

- aus der zweiten Normalgleichung folgt:

4. Multipliziert man $y_v = a + bx_v + u_v$ mit x_v und summiert man über alle n Beobachtungen so erhält man $\Sigma x_v y_v = a\Sigma x_v + b\Sigma x_v^2 + \Sigma x_v u_v$, so dass wegen der zweiten Normalgleichung U und X nicht miteinander korreliert sind:

$$(8.13) \quad \Sigma x_v u_v = 0 \quad \text{und} \quad s_{ux} = r_{ux} = 0.$$

5. Multipliziert man $\hat{y}_v = a + bx_v$ mit u_v und summiert man über alle n Beobachtungen, so erhält man $\Sigma \hat{y}_v u_v = a\Sigma u_v + b\Sigma x_v u_v$, so dass wegen $\Sigma x_v u_v = \Sigma u_v = 0$ gilt:

$$(8.14) \quad \Sigma \hat{y}_v u_v = r_{\hat{y}u} = 0.$$

6. Die entsprechende Betrachtung mit $y_v = a + bx_v + u_v$ führt zu $\Sigma y_v u_v = a\Sigma u_v + b\Sigma x_v u_v + \Sigma u_v^2$ und wegen $\Sigma x_v u_v = \Sigma u_v = 0$ zu

$$(8.15) \quad \Sigma y_v u_v = \Sigma u_{yx}^2 \quad \text{und} \quad s_u^2 = \frac{\Sigma u_v^2}{n} = s_{uy}.$$

Hieraus folgt $r_{yu} = s_u/s_y$ und damit auch

$$(8.16) \quad (r_{yu})^2 = \frac{s_u^2}{s_y^2} = 1 - (r_{xy})^2 = U_{xy}$$

das Unbestimmtheitsmaß (Unbestimmtheit von Y durch X) als Bestimmtheit von Y durch U (und umgekehrt).

7. Aus Gl. 8.14 folgt mit $y_v = \hat{y}_v + u_v$ die Varianzzerlegung $s_y^2 = s_{\hat{y}}^2 + s_u^2$ sowie $s_{y\hat{y}} = s_{\hat{y}}^2$, so dass auch gilt

$$r_{y\hat{y}} = \frac{s_{y\hat{y}}}{s_{\hat{y}}s_y} = \frac{s_{\hat{y}}^2}{s_{\hat{y}}s_y} = \frac{s_{\hat{y}}}{s_y} = \sqrt{\frac{s_{\hat{y}}^2}{s_y^2}} = \sqrt{B_{xy}} = r_{xy} \text{ also}$$

$$(8.17) \quad r_{y\hat{y}} = r_{xy}.$$

Die Korrelation zwischen x und y ist also auch als Korrelation zwischen y und dem Regresswert \hat{y} zu interpretieren, eine Interpretation, die sich auf die multiple Regression übertragen läßt.

8. Die Regressionsgerade $\hat{y} = a + bx$ verläuft durch die zwei Punkte $P_1(\bar{x}, \bar{y})$, d.h. den Schwerpunkt und $P_2(\bar{x}_w, \bar{y}_w)$, wobei \bar{x}_w und \bar{y}_w gewogene Mittel der x_v - bzw. y_v -Werte sind mit den Gewichten $w_v = x_v / \sum x_v$.

b) Alternativen zur Minimierung der Summe der Quadrate der Abweichungen

Man kann sich alternative Schätzungen der Regressionskoeffizienten a und b der Regressionsfunktion $\hat{y} = a + bx$ (die folgende Betrachtung gilt übrigens ganz entsprechend auch für nichtlineare und multiple Regression) vorstellen, von denen jedoch nicht alle sinnvoll und eindeutig sind. Dargestellt werden im folgenden sechs Alternativen zur Methode der kleinsten Quadrate:

- 1.) Fordert man z.B. den Ausgleich positiver und negativer Abweichungen von der Regressionsfunktion, d.h. dass die Summe der Residuen den Wert Null annimmt, so erhält man nur eine Gleichung zur Bestimmung von a und b, nämlich die erste Normalgleichung.

Die Forderung $\sum u_v = 0$ führt zur Gleichung $a + b\bar{x} = \bar{y}$, die von allen Geraden erfüllt wird, die durch den Schwerpunkt gehen. Eine von Ihnen ist die Regressionsgerade nach der Methode der kleinsten Quadrate. Das Kriterium $\sum u_v = 0$ ist also in $\sum u_v^2 = \text{Min}$ impliziert, führt aber allein nicht zu einer eindeutigen Lösung.

- 2.) Die Minimierung der Abweichungen $\sum u_v$ scheitert daran, dass die Funktion $Q = \sum u_v = \sum y_v - na - b\sum x_v$ keine Extremwerte besitzt (bei ge-

gebenem $a = a_0$ oder $b = b_0$ ist die Funktion $Q(a_0, b)$ bzw. $Q(a, b_0)$ eine Gerade). Wie Beispiel 8.4 (Abb. 8.4) zeigt, ist die Lösung nicht eindeutig.

- 3.) Auch das Kriterium der Minimierung der absoluten Abweichungen

$$\sum |u_v| = \sum |y_v - a - bx_v| = \text{Min!}$$

führt nicht in jedem Fall zu einem eindeutigen und befriedigenden Ergebnis (Bsp. 8.4) und bereitet zudem beträchtliche Schwierigkeiten bei der Bestimmung der Regressionskoeffizienten.

- 4.) Ein sinnvolles und zugleich eindeutiges Kriterium ist jedoch die **orthogonale Regression**, d.h. die Minimierung der orthogonalen (statt senkrechten) Abstände der Datenpunkte zur (orthogonalen) Regressionsgeraden.

Mit der orthogonalen Regressionsgeraden $\hat{y}_v^o = A + Bx_v$ (statt $\hat{y}_v = a + bx_v$ für die Regression nach der Methode der kleinsten Quadrate) erhält man für die quadrierten orthogonalen Abstände $d_v^2 = u_v^2/(1+B^2)$. Zur Schätzung von B ist B aus der quadratischen Gleichung $B^2s_{xy} + B(s_x^2 - s_y^2) - s_{xy} = 0$ zu bestimmen (wobei nur eine Lösung erkennbar sinnvoll ist) und für A gilt $A = \bar{y} - B\bar{x}$ (Schwerpunktbedingung). Beispiel 8.5 demonstriert den Rechengang. Die Steigung B liegt zwischen den beiden Steigungen b und d , wobei das Verhältnis der Varianzen s_y^2/s_x^2 von Bedeutung ist. Außerdem ist die orthogonale Regression symmetrisch, d.h. die Geraden \hat{y}^o und \hat{x}^o "fallen zusammen".

- 5.) Eine Regressionsgerade, die durch beiden Punkte $P_1(\bar{x}_1, \bar{y}_1)$ und $P_2(\bar{x}_2, \bar{y}_2)$ "läuft" hat die Parameter $a = (\bar{x}_2\bar{y}_1 - \bar{x}_1\bar{y}_2)/(\bar{x}_2 - \bar{x}_1)$ und $b = (\bar{y}_2 - \bar{y}_1)/(\bar{x}_2 - \bar{x}_1)$. Man kann also z.B. eine Regressionsgerade durch die Punkte $P_1(\bar{x}_1, \bar{y}_1)$ und $P_2(\bar{x}_2, \bar{y}_2)$ wobei \bar{x}_1 der Mittelwert der ersten (mit kleineren x -Werten) Hälfte der Daten und \bar{x}_2 der Mittelwert zweiten Hälfte der Daten ist (Regressionsgerade nach Wald). Oder man teilt die Daten in drei gleich umfangreiche Teilgesamtheiten nach Maßgabe zunehmender x -Werte und bestimmt eine Regressionsgerade, die durch die Punkte $P_1(Z_{x1}, Z_{y1})$ und $P_3(Z_{x3}, Z_{y3})$ läuft, wobei Z_x und Z_y die Mediane (Zentralwerte) sind. Dies sind Vorschläge, eine Regressionsgerade zu bestimmen, die "robuster" ist gegenüber Ausreißern als die mit der Methode der kleinsten Quadrate errechnete Regressionsgerade.
- 6.) Natürlich kann man sich viele weitere Bedingungen vorstellen, die zu einer eindeutigen und sinnvollen Schätzung einer Geraden, die sich den Daten anpaßt, führen würden. Vorgeschlagen wird insbesondere auch eine Gewichtung der quadrierten Abweichungen also

$$\Sigma w_v(u_v)^2 = \Sigma w_v(y_v - a - bx_v)^2 = \text{Min}$$

so dass die Regressionsfunktion nach der Methode der kleinsten Quadrate "nur" der Spezialfall gleicher Gewichte $w_1 = w_2 = \dots = w_n$ aller Beobachtungen ist.

Beispiel 8.4 (vgl. Abb. 8.4):

a) Gegeben seien die folgenden Daten (*Abb. 8.4 oben*)

(x,y): (2,5), (4,7) und (6,7)

Die Geraden $\hat{y} = 2 + x$ und $\hat{y} = 14 - 2x$ führen jeweils zu $\Sigma u_v = 1$ obgleich sie sich den Daten offensichtlich sehr unterschiedlich gut anpassen. Man erhält übrigens

für $\hat{y} = 2 + x$, $\Sigma |u_v| = 3$ und $\Sigma u_v^2 = 3$ dagegen

für $\hat{y} = 14 - 2x$ die Werte $\Sigma |u_v| = 11$ und $\Sigma u_v^2 = 51$,

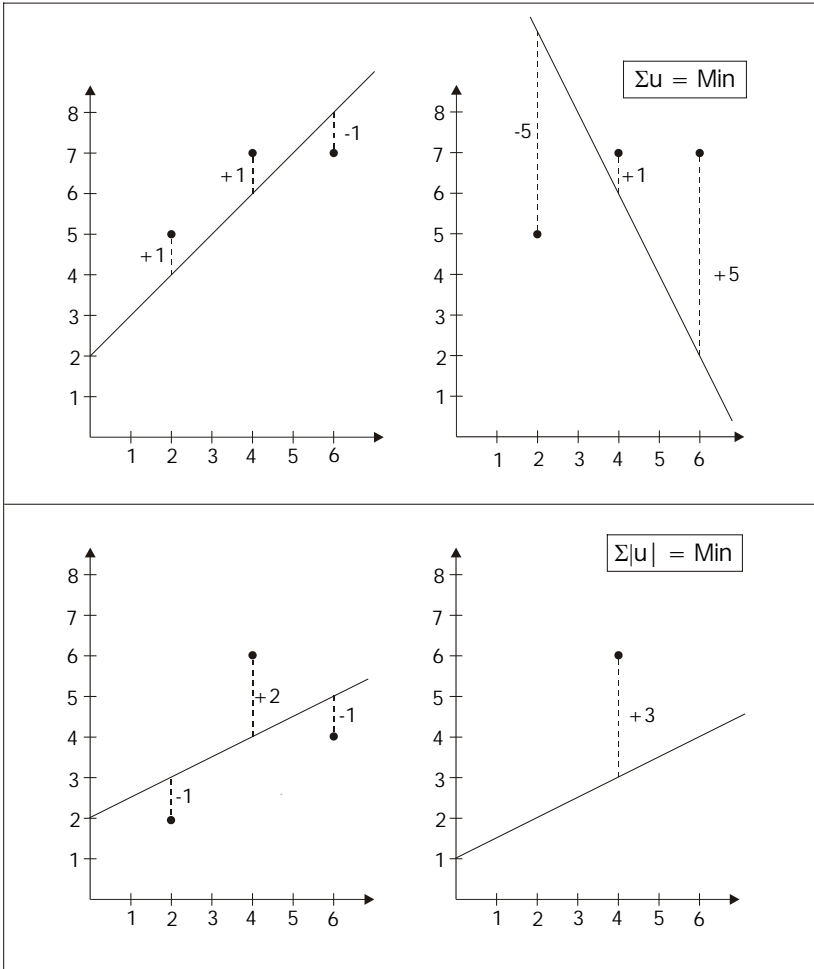
die offensichtlich schlechter sind.

b) Gegeben seien die folgenden Daten (*Abb. 8.4 unten*)

(x,y): (2,2), (4,6) und (6,4).

Die Gerade $\hat{y} = 1 + 1/2x$ (*Abb. 8.4 rechts unten*) ist, gemessen am Kriterium $\Sigma |u_v| = \text{Min}$ besser ($\Sigma |u_v| = 3$) als die Gerade $\hat{y} = 2 + 1/2x$ (*Abb. 8.4 links unten*, $\Sigma |u_v| = 4$), nicht aber nach dem Kriterium der kleinsten Quadrate, denn man erhält $\Sigma u_v^2 = 9$ für die rechte und $\Sigma u_v^2 = 6$ für die linke Gerade.

Abb. 8.4: Alternativen zur Minimierung der Summe der Quadrate der (senkrechten) Abweichungen (vgl. Bsp. 8.4)



Beispiel 8.5:

Man bestimme die kleinste-Quadrate- und die orthogonale Regression sowie die orthogonalen (d_v) und die senkrechten (u_v) Residuen für die beiden Datensätze des Beispiels 8.4!

Lösung 8.5:

a.) Für die Daten $(x;y)$: (2,5), (4,7) und (6,7) erhält man mit der Methode der kleinsten Quadrate: $\hat{y} = 13/3 + 1/2x$ und $\hat{x} = -11/2 + 3/2y$.

Wegen $s_x^2 = 8/3$ $\bar{x} = 4$ $s_y^2 = 8/9$ $\bar{y} = 19/3$ und $s_{xy} = 4/3$ erhält man für die orthogonale Regression $\hat{y}^0 = 4,19259 + 0,53518x_v$ und somit die folgenden Residuen:

x	y	y	u	u ²	y ⁰	(u ⁰) ²
2	5	16/3 = 5,3	1/3	1/9	5,2629	0,06915
4	7	19/3 = 5,3	-2/3	4/9	6,3333	0,44444
6	7	22/3 = 7,3	1/3	1/9	7,4037	0,16766
Σ			2/3			0,67657

Man beachte, dass $\Sigma(u^0)^2 = 0,67657$ zwar größer ist als $\Sigma u^2 = 0,667$ (Σu^2 wird ja minimiert), nicht aber die Summe der quadrierten **orthogonalen** Abstände für die gilt: $\Sigma d_v^2 = \Sigma u_v^2 / (1+b^2) = (2/3) / (5/4) = 0,533$ bei der Regressionsgeraden (kleinste Quadrate) bzw. $\Sigma d_v^2 = \Sigma (u^0_v)^2 / (1+B^2) = 0,52593$ (dieses Σd_v^2 wird ja auch minimiert).

b) Für die Daten von Teil b des Beispiels 8.3 erhält man für die Regressionsfunktion $\hat{y} = 2 + 1/2x$ (kleinste Quadrate, die in Abb. 8.4 links unten eingezeichnete Gerade ist die Regressionsgerade) $\Sigma u_v^2 = 6$ und $\Sigma d_v^2 = 4,8$. Für die lineare orthogonale Regression erhält man $\hat{y}^0 = x$ (also die 45° Linie weil $s_x^2 = s_y^2$!) und $\Sigma u_v^2 = 8$ aber für die minimierte Größe $d_v^2 = 1/2 \cdot 8 = 4$.

3. Ergänzungen zur linearen einfachen Regression

a) Standardisierte Variablen, gruppierte Daten

Def. 8.3: Zentrierung, Standardisierung

Die Transformation $x_v^z = x_v - \bar{x}$ und entsprechend $y_v^z = y_v - \bar{y}$ heißt Zentrierung und die Transformation $x_v^s = (x_v - \bar{x}) / s_x = x_v^z / s_x$ (y_v^s ist analog definiert) stellt eine Standardisierung dar. Zum Begriff der Standardisierung vgl. Kap. 2. In einem anderen Sinne wird der Begriff in Kap. 9 gebraucht).

Folgerungen:

1. Regressionsgerade bei zentrierten Variablen x_v^z und y_v^z :

Die Gleichung $y_v = a + bx_v + u_v$ läßt sich wegen $\bar{y} = a + b\bar{x}$ umformen zu

$$(8.18) \quad y_v^z = bx_v^z + u_v,$$

d.h. es verschwindet das Absolutglied a bei Zentrierung und b ist gem. Gl. 8.6a zu bestimmen, in diesem Fall also $b = \Sigma x_v^z y_v^z / \Sigma (x_v^z)^2$. Das Absolutglied a verschwindet und die Steigung b bleibt unverändert.

Das Verschwinden des Absolutglieds durch Verwendung zentrierter Variablen (d.h. durch Verschiebung des Koordinatensystems) ist nicht zu verwechseln mit der "**homogenen Regression**", bei der eine Regressionsgerade gesucht wird, die durch den Ursprung des (ursprünglichen) Koordinatensystems verläuft.

Es wird dann $\Sigma(u_v^*)^2 = \Sigma(y_v - b_H x_v)^2$ statt $\Sigma(u_v)^2 = \Sigma(y_v - a - b x_v)^2$ minimiert und das führt zur ersten Normalgleichung $\Sigma x_v y_v = b_H \Sigma x_v^2$. Die Steigung bleibt nicht unverändert.

Im Beispiel 8.1 erhält man die homogenen Regressionsgeraden $\hat{y}_v = 1,055x_v$ (denn $b_H = \Sigma x_v y_v / \Sigma x_v^2 = 133/126 = 1,055$ während $b = 1/2$) und $\hat{x}_v = 0,8012y_v$ (denn $d_H = \Sigma x_v y_v / \Sigma y_v^2 = 133/166$).

2. Regressionsgerade bei standardisierten Variablen x_v^S und y_v^S :

Die arithmetischen Mittel \bar{x}^S und \bar{y}^S der standardisierten Variablen sind Null und ihre Varianzen sind Eins. Bei Verwendung standardisierter Variablen folgt aus der ersten Normalgleichung der Regressionsfunktion

$$\hat{y}^S = a_S + b_S x^S \quad \bar{x}^S = \bar{y}^S = 0 \quad \text{dass } a_S = 0.$$

Und für die zweite Normalgleichung erhält man

$$a_S \Sigma x_v^S + b_S \Sigma (x_v^S)^2 = \Sigma x_v^S y_v^S = nr_{xy}, \quad \text{woraus unter Berücksichtigung von } a_S = 0 \text{ und } \frac{1}{n} \Sigma (x_v^S)^2 = \frac{1}{n} \Sigma [(x_v - \bar{x})/s_x]^2 = s_x^2/s_x^2 = 1 \text{ folgt}$$

$$(8.19) \quad b_S = r_{xy}.$$

In der Regressionsfunktion $\hat{y}_v^S = a_S + b_S x_v^S$ mit den standardisierten Variablen x_v^S und y_v^S ist also der Ordinatenabschnitt $a_S = 0$ und die Steigung b_S gleich dem Korrelationskoeffizienten. Die beiden durch den Ursprung verlaufenden Regressionsgeraden lauten

$$\hat{y}_v^S = r_{xy} \cdot x_v^S \quad \text{und} \quad \hat{x}_v^S = r_{xy} \cdot y_v^S.$$

Gl. 8.19 liefert auch den Grund dafür, dass die in Bsp. 8.3 Teil c und d dargestellte regression to the mean umso schneller erfolgt, je geringer r ist, denn bei standardisierten Variablen ist die Regressionsfunktion $\hat{y}^S = b_S x^S = r_{xy} x^S$, so dass die

Standardabweichung der Regresswerte \hat{y}^S genau r_{xy} beträgt (die Standardabweichung von x^S ist wegen der Standardisierung natürlich 1). Mithin streut \hat{y}^S um den Mittelwert 0 umso weniger, je kleiner betragsmäßig r_{xy} ist.

Bei **gruppierten Daten** erhält man mit der Notation von Def. 7.2

$$(8.20a) \quad a = \frac{\sum x_i^2 n_i \cdot \sum y_j n_j - \sum x_i n_i \cdot \sum x_i y_j n_{ij}}{n \sum (x_i^2 n_i) - (\sum x_i n_i)^2}$$

$$(8.20b) \quad b = \frac{n \sum x_i y_j n_{ij} - \sum x_i n_i \cdot \sum y_j n_j}{n \sum (x_i^2 n_i) - (\sum x_i n_i)^2} = \frac{n^2 s_{xy}}{n^2 s_x^2}$$

für die Koeffizienten a und b der Regressionsgerade \hat{y} und entsprechend c und d für die Regressionsgerade \hat{x} .

Bei gruppierten Daten kann man auch die in Kapitel 7 dargestellten Regressionslinien bestimmen. Die bedingten Mittelwerte

$$\bar{y}|x_i = \frac{\sum_j y_j n_{ij}}{n_i}$$

müssen nicht notwendig auf einer Geraden und schon gar nicht auf der Regressionsgeraden \hat{y} liegen (und entsprechend müssen die Werte $\bar{x}|y_j$ nicht identisch sein mit \hat{x}).

Die Regressions**linien** (vgl. Def. 7.6) sind i.d.R. von den Regressions**geraden** verschieden.

b) Exkurs zum Regressionsmodell

Es wurde gelegentlich darauf hingewiesen, dass die rein deskriptive Behandlung der Regression, die allein in diesem Rahmen dargestellt werden soll, von dem stochastisch fundierten Modell der Regression zu unterscheiden ist. So ist es z.B. im Kontext dieses Modells nicht mehr möglich, die Abhängigkeiten zwischen X und Y einfach zu vertauschen.

Bei der modellmäßigen Erfassung des Zusammenhangs der beiden Variablen X und Y wird davon ausgegangen, dass für die wahre Regression (in der Grundgesamtheit) gilt:

$$(8.21) \quad Y = \alpha + \beta X + U.$$

Hierin ist U (und deshalb auch Y) eine Zufallsvariable, während x nichtzufällig schwankt² (die Varianz σ_u^2 in der Grundgesamtheit ist nicht Null). Die Parameter α und β sind die Regressionskoeffizienten der "wahren" Regressionsfunktion der Grundgesamtheit, die es anhand der n beobachteten Wertepaare (x_v, y_v) einer Stichprobe zu schätzen gilt. Dabei sind die Werte y_v als Realisationen einer Zufallsvariable Y aufzufassen.

Es werden folgende Annahmen über das Modell und die nicht zu beobachtende Zufallsvariable U gemacht:

1. Die x_v - Werte ($v=1, \dots, n$) sind fest vorgegeben, d.h. sie sind frei von Meßfehlern und somit keine Zufallsvariablen.
2. Die y_v - Werte sind wegen der Überlagerung durch eine Störgröße U_v , deren Realisation u_v genannt sei: $y_v = \alpha + \beta x_v + u_v$. Bekannt sind aber weder α und β noch u_v . Vielmehr sind α , β und u_v mit den Beobachtungen x_v , y_v zu schätzen. Die (Stichproben-) Schätzwerte heißen $a = \hat{\alpha}$ und $b = \hat{\beta}$; für u_v (wie bisher immer bezeichnet) wäre \hat{u}_v eigentlich die bessere Bezeichnung.
3. Die zufälligen Störvariablen U_v sollen folgende Eigenschaften haben:
 - a) $E(U_v) = 0$ ($v=1, \dots, n$)
 - b) $\text{Var}(U_v) = \sigma_u^2 = \sigma$
 - c) die Größen U_1, \dots, U_n sind stochastisch unabhängig und
 - d) unabhängig identisch normalverteilt mit $E(U) = 0$ und $\sigma_u = \sigma$.

Bemerkungen zu Voraussetzung Nr. 3:

zu a)

Diese Forderung ist nicht damit zu verwechseln, dass die mit der Methode der kleinsten Quadrate **geschätzte** Störgröße u im Mittel Null ist. Man beachte, dass $E(U)$ eine Aussage über einen Erwartungswert ist, der sich auf eine Wahrscheinlichkeitsverteilung bezieht. Es ist ja auch ein Stichprobenmittel von $\bar{x} = 0$ nicht gleichbedeutend damit, dass in der Grundgesamtheit $\mu = 0$ ist. Aus a) folgt auch, dass x und U nicht korreliert sind, was ebenfalls nicht damit zu verwechseln ist, dass der Regressor mit der **geschätzten** Störgröße nicht korreliert (zweite Normalgleichung).

zu b)

Diese Eigenschaft wird auch Homoskedastizität (Homoskedastie) genannt und sie besagt, dass die U_v , unabhängig von dem jeweiligen x -Wert, alle dieselbe Varianz σ_u^2 haben.

² Abweichend von der bisherigen Notation (X = Variable, x = konkreter Wert einer Variable) soll hier zwischen großen und kleinen Buchstaben in einer Weise unterschieden werden, wie das in der Induktiven Statistik üblich ist: Y = Zufallsvariable, x = nicht zufällige Variable oder Realisation der Zufallsvariable X .

zu c)

Dies bedeutet v.a. bei Zeitreihen, dass z.B. einem großen (kleinen) Wert von U_t nicht notwendig ein großer (kleiner) Wert U_{t+1} folgen darf, d.h. dass die Störvariablen nicht autokorreliert sein dürfen.

zu d)

Wenn es nicht von Interesse ist, eine Intervallschätzung oder einen Test für α , β , σ oder die Korrelation durchzuführen, ist diese Annahme nicht erforderlich. Es wird oft übersehen, dass für eine rein deskriptive Behandlung der Regression alle obigen Annahmen über U nicht erforderlich sind.

Ist 3a nicht erfüllt, so sind a und b keine erwartungstreue Schätzer für α und β . Von 3b und 3c wird Gebrauch gemacht, bei der Bestimmung der Varianzen und Kovarianzen der Stichprobenverteilungen von a und b und wäre 3d nicht gegeben, so wären die Schätzer a und b (für α und β) bzw. s_u^2 für σ_u^2 nicht t - bzw. χ^2 -verteilt.

Weitere Bemerkungen und Zusammenfassung:

- Man kann auch unterscheiden das "klassische" Modell der "Regression" $Y = \alpha + \beta X + U$ (nicht-stochastischer Regressor x) vom Modell der Korrelation $Y = \alpha + \beta X + U$ (stochastischer Regressor X).
- Es ist, wie gesagt, im stochastischen Modell nicht mehr möglich, die Abhängigkeiten zwischen X und Y einfach zu vertauschen, wie dies bei rein deskriptiver Betrachtung möglich ist, denn: gilt z.B. bei einem stochastischen Regressor X in der Regression $Y = \alpha_0 + \alpha_1 X + U$ die Annahme $E(UX) = 0$, so kann für die Umkehrfunktion $X = \beta_0 + \beta_1 Y + V$ wegen $\beta_0 = -\alpha_0/\alpha_1$ und $\beta_1 = 1/\alpha_1$ sowie $V = -U/\alpha_1$ nicht auch zugleich $E(VY) = 0$ gelten. Vielmehr ist $E(VY) = -\alpha_1^{-2}E(UY) = -\alpha_1^{-2}E(U^2) = -\sigma_u^2/\alpha_1^2 < 0$.
- Die aufgrund der Stichprobe zu schätzenden Parameter sind bei einfacher Regression: α , β und (was häufig vergessen wird) $\sigma^2 = \sigma_u^2$ (die Varianz der Störgröße) und die Schätzwerte sind a (für α), b (für β) und $\hat{\sigma}^2 = \Sigma u^2/(n-2)$ für σ^2 der Varianzschätzer, was nicht zu verwechseln ist mit der Stichprobenvarianz $s^2 = \Sigma u^2/n$.

Übersicht 8.1: Größen und Parameter des Modells der einfachen (multiplen) Regression

Größen	beobachtbar	nichtbeobachtbar	Schätzwert
nichtzufällig ^(a)	$\mathbf{x}^{(b)}$ (x_1, \dots, x_p)	$\alpha, \beta, \sigma_u^2$ $(\beta_0, \beta_1, \dots, \beta_p, \sigma_u^2)$	a, b, σ_u^2 ^(c) $(b_0, b_1, \dots, b_p, \sigma_u^2)$
zufällig	\mathbf{y}	Störgröße U	$\hat{\mathbf{y}}$ und $\hat{\mathbf{u}}$ bzw. \mathbf{u}

(a) nichtzufällige Größen, bzw. zu schätzende Konstanten im Modell; zufällige Größen sind dagegen Y und U , wobei y_v eine Realisation von Y ist;

- (b) mit $\mathbf{x}, (\mathbf{x}_1, \mathbf{x}_2, \dots), \mathbf{y}, \mathbf{U}, \mathbf{u} = \hat{\mathbf{u}}$ und $\hat{\mathbf{y}}$ sind jeweils Vektoren (Spaltenvektoren mit n Zeilen) gemeint, also $\mathbf{x}' = [x_1 \ x_2 \ \dots \ x_n]$ oder $\mathbf{x}_1' = [x_{11} \ x_{12} \ \dots \ x_{1n}]$.
- (c) Der Schätzer der Varianz der Störgröße $\sigma_u^2 = \Sigma u^2 / (n-p)$ (bei $p-1$ Regressoren, also z.B. $p=2$ bei einfacher Regression) ist erwartungstreu, also $E(\hat{\sigma}_u^2) = \sigma_u^2 = \sigma^2$. Die oben betrachtete Größe $s^2 = \Sigma u^2 / n$ ist aber nicht erwartungstreu.

4. Multiple lineare Regression

a) Beschreibung des Modells

Multiple Regression (Mehrfachregression) heißt Y auf den Einfluß nicht nur einer, sondern mehrerer erklärender Variablen X_1, \dots, X_p zurückzuführen. Im linearen Fall bedeutet dies, das Modell:

$$(8.24) \quad Y_v = \beta_0 + \beta_1 x_{1v} + \beta_2 x_{2v} + \dots + \beta_p x_{pv} + U_v \quad (v=1, \dots, n)$$

mit Stichprobendaten durch die Gleichung

$$(8.25) \quad y_v = b_0 + b_1 x_{1v} + b_2 x_{2v} + \dots + b_p x_{pv} + u_v \quad (v=1, \dots, n)$$

zu schätzen, wobei

$$(8.26) \quad \hat{y}_v = b_0 + b_1 x_{1v} + b_2 x_{2v} + \dots + b_p x_{pv}$$

die (geschätzte) Regressionsfunktion ist.

Das Absolutglied b_0 ist für die Interpretation irrelevant und kann beseitigt werden durch Zentrierung (deviation scores) oder Standardisierung. Ein Beispiel für eine zweifache Regression ist die bereits erwähnte Abhängigkeit des Jahresumsatzes (Y) von den Ausgaben für Werbung (X_1) und den Wareneinkäufen (X_2). Das Streudiagramm ist dann eine dreidimensionale Darstellung (Y, X_1, X_2) und der Ausdruck Punkt-"wolke" wäre jetzt (im Unterschied zur einfachen [bivariaten] Regression) erstmals korrekt. Bei zwei Regressoren ist die Funktion $y = y(x_1, x_2)$ eine Ebene. Der Regresswert \hat{y} ist ein Punkt auf der Ebene. Die Beobachtungen (Datenpunkte) liegen auf, ober- oder unterhalb dieser Ebene. Die senkrechten [in Richtung der y -Achse] Abstände von dieser Ebene stellen die Störgröße dar.

Def. 8.4: multiple lineare Regression

- a) Das Gleichungssystem (Gleichungen) 8.26 beschreibt eine multiple lineare Regression mit p Regressoren und für die v -te Beobachtung ($v=1, 2, \dots, n$) gilt:

$$(8.26a) \quad \hat{y}_v = b_0 + b_1 x_{1v} + b_2 x_{2v} + \dots + b_p x_{pv} = \sum_{i=0}^p b_i x_{iv} \quad \text{mit } x_{0v} = 1$$

oder in anderer Schreibweise, z.B. bei drei Regressoren:

$$(8.26b) \quad \hat{y} = b_{y0.123} + b_{y1.23}x_1 + b_{y2.13}x_2 + b_{y3.12}x_3$$

- b) Die Koeffizienten b_i oder $b_{y_i,jk}$ heißen **partielle Regressionskoeffizienten** (als Maß für den isolierten Einfluß des Regressors X_i "bei Konstanz" von X_j , X_k usw.), da die partielle Ableitung der Regressionsfunktion $b_i = \beta y / \beta x_i$ ist, bzw. (synonym) multiple Regressionskoeffizienten.
- c) Zur Beschreibung des Modells werden auch die folgenden Größen herangezogen: **multiple Korrelation** und **multiple Bestimmtheit** sowie die **partielle Korrelation** (vgl. Bem. Nr. 3 und Teile c und d dieses Abschnitts).
- d) Unter **Orthogonalität** versteht man, dass (bei stochastischen Regressoren) die Regressoren X_1, \dots, X_p stochastisch unabhängig sind. Wenn (im anderen Extremfall) zwischen den unabhängigen Variablen X_1, \dots, X_p lineare Abhängigkeiten bestehen spricht man von offener **Kollinearität** (Multikollinearität). In diesem Fall ist eine Kleinste-Quadrate-Schätzung der Regressionskoeffizienten nicht möglich.

Bemerkungen zu Def. 8.4:

1. Gl. 8.26 stellt im $p+1$ dimensionalen Raum eine p - dimensionale Hyperebene dar. Im Fall einer zweidimensionalen Regression [zwei Regressoren X_1 und X_2] ist dies, wie gesagt, eine Ebene im anschaulichen Sinne (im dreidimensionalen $[Y, X_1, X_2]$ Raum).
2. Es sollen die gleichen Modellannahmen wie bisher gelten und zusätzlich Kollinearität (Multikollinearität) möglichst ausgeschlossen sein, d.h. im Idealfall Orthogonalität herrschen. In der Praxis ist weder Orthogonalität noch offene Kollinearität üblich, sondern eine **verdeckte Kollinearität**, d.h. eine mehr oder weniger starke Korrelation zwischen den unabhängigen Variablen X_1, \dots, X_p . Bei offener Kollinearität ist eine Kleinste-Quadrate-Schätzung der Regressionskoeffizienten nicht möglich und bei starker verdeckter Kollinearität ist sie nur mit großem (und unbekanntem) Fehler möglich, also sehr unzuverlässig.

3. Während die Begriffe partielle und multiple Regressionskoeffizienten synonym sind, gilt dies keineswegs für die partiellen und multiplen Korrelationskoeffizienten. Der
- **multiple Korrelationskoeffizient**, etwa $R_{y,123}$ ist die einfache Korrelation zwischen y und $\hat{y} = f_3(x_1, x_2, x_3)$; entsprechend wäre z.B. $R_{y,12}$ die Korrelation zwischen y und $\hat{y} = f_2(x_1, x_2)$ (es gilt übrigens $R_{y,12} \mid R_{y,123}$);
 - **partielle Korrelationskoeffizient**, etwa $r_{y,1,23}$ ist die Korrelation zwischen Y und X_1 wobei der Einfluß von X_2 und X_3 in einem noch beschriebenen Sinne "ausgeschaltet" ist.
4. Das Kriterium für die beste Anpassung der Regressionsfunktion (Regressions- [hyper] -ebene) ist wieder die Minimierung der Quadrate der Abweichungen u_v .

Auf die Regressionsfunktion wird somit wieder die Methode der Kleinsten Quadrate angewendet, d.h. es sind die Werte b_0, b_1, \dots, b_p zu bestimmen, die die Funktion $Q(b_0, b_1, \dots, b_p)$ minimieren:

$$Q(b_0, b_1, \dots, b_p) = \sum u_v^2 = \sum (y_v - \hat{y}_v)^2 = \text{Min}$$

mit $Q(b_0, b_1, \dots, b_p) = \sum u_v^2 = \sum (y_v - b_0 - b_1 x_{1v} - \dots - b_p x_{pv})^2$.

Q ist partiell nach b_0, b_1, \dots, b_p zu differenzieren und die Ableitungen sind Null zu setzen, womit man $p+1$ Normalgleichungen erhält. Im Falle von zwei Regressoren lautet das System der drei Normalgleichungen (jeweils summiert über $v = 1, 2, \dots, n$):

$$\begin{array}{l} 1. \quad b_0 n \quad + \quad b_1 \sum x_{1v} \quad + \quad b_2 \sum x_{2v} \quad = \quad \sum y_v \\ 2. \quad b_0 \sum x_{1v} + b_1 \sum x_{1v}^2 + b_2 \sum x_{1v} x_{2v} = \sum x_{1v} y_v \\ 3. \quad b_0 \sum x_{2v} + b_1 \sum x_{1v} x_{2v} + b_2 \sum x_{2v}^2 = \sum x_{2v} y_v. \end{array}$$

Man beachte, dass die Schätzwerte b_0, b_1, b_2 linear in y (und damit auch in u) sind, also lineare Schätzer für $\beta_0, \beta_1, \beta_2$ sind.

b) Darstellung in Matrixschreibweise

Erheblich einfacher und übersichtlicher sind die Schätzgleichungen in Matrixschreibweise darzustellen. Dazu vereinbart man die Datenmatrix \mathbf{X} , den Datenvektor \mathbf{y} , den Vektor \mathbf{u} der geschätzten Residuen und den Vektor \mathbf{b} der zu schätzenden Regressionskoeffizienten wie folgt:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdot & x_{p1} \\ 1 & x_{12} & x_{22} & \cdot & x_{p2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{1n} & x_{2n} & \cdot & x_{pn} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ b_p \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_n \end{bmatrix}$$

Die Spaltenvektoren \mathbf{y} und \mathbf{u} haben jeweils n Zeilen (n Beobachtungen) \mathbf{X} hat n Zeilen und $p+1$ Spalten und die Elemente x_{jv} ($v = 1, 2, \dots, n$ und $j = 0, 1, \dots, p$). Der dummy - Regressor X_0 (mit $x_{01} = \dots = x_{0n} = 1$) wird erforderlich zur Schätzung des Absolutglieds. Das Gleichungssystem 8.26 stellt sich dann in Matrixschreibweise wie folgt dar:

$$\boxed{(8.27) \quad \mathbf{y} = \mathbf{Xb} + \mathbf{u}.}$$

Das Matrizenprodukt $\mathbf{X'X}$ ist die quadratische Matrix der n -fachen Anfangsmomente bzw. Anfangsproduktmomente:

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \dots & x_{1n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix} = \begin{bmatrix} n & \Sigma x_1 & \dots & \Sigma x_p \\ \Sigma x_1 & \Sigma x_1^2 & \dots & \Sigma x_1 x_p \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \Sigma x_p & \Sigma x_1 x_p & \dots & \Sigma x_p^2 \end{bmatrix}$$

und man erkennt leicht, dass die Normalgleichungen in Matrixschreibweise lauten:

$$\boxed{(8.28) \quad \mathbf{X'Xb} = \mathbf{X'y} \text{ (Normalgleichungen).}}$$

Die zu minimierende Residualquadratsumme ist:

$$\Sigma u_v^2 = \mathbf{u'u} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = \mathbf{y'y} - \mathbf{b'X'y} - \mathbf{y'Xb} + \mathbf{b'X'Xb}$$

Berücksichtigt man, dass $\mathbf{b'X'y}$ ein Skalar ist und deshalb $(\mathbf{b'X'y})' = \mathbf{y'Xb}$ ist, gilt weiter:

$$\boxed{(8.29) \quad \mathbf{u'u} = \mathbf{y'y} - 2\mathbf{y'Xb} + \mathbf{b'X'Xb}.}$$

Durch partielle Differentiation von Gl. 8.29 nach dem Vektor \mathbf{b} erhält man das Normalgleichungssystem der multiplen Regression, d.h. Gl. 8.28, in ausführlicher Schreibweise ohne den Summationsindex v

$$\begin{array}{rcl}
 b_0 n & + & b_1 \Sigma x_1 + \dots + b_m \Sigma x_m = \Sigma y \\
 b_0 \Sigma x_1 & + & b_1 \Sigma x_1^2 + \dots + b_m \Sigma x_1 x_m = \Sigma x_1 y \\
 \cdot & \cdot & \cdot \\
 b_0 \Sigma x_p & + & b_1 \Sigma x_1 x_p + \dots + b_m \Sigma x_m^2 = \Sigma x_m y.
 \end{array}$$

Werden beide Seiten der Gl. 8.28 mit der Inversen $(\mathbf{X}'\mathbf{X})^{-1}$ prämultipliziert, so gelangt man zum Schätz-Vektor der Regressionskoeffizienten:

$$(8.30) \quad \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} .$$

Voraussetzung für die Berechnung von \mathbf{b} nach Gl. 8.30 ist natürlich, dass die Inverse der Matrix $\mathbf{X}'\mathbf{X}$ überhaupt existiert, was z.B. nicht der Fall ist bei offener Multikollinearität.

Folgerungen aus den Normalgleichungen:

- 1) Aus Gl. 8.29 folgt übrigens wegen $\mathbf{y}' = \mathbf{b}'\mathbf{X}' + \mathbf{u}'$ und deshalb $\mathbf{y}'\mathbf{X}\mathbf{b} = \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$

$$\mathbf{u}'\mathbf{u} = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{y}'\mathbf{y}$$

und damit die folgende Beziehung:

$$(8.31) \quad \mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{u}'\mathbf{u} \quad (\text{Varianzzerlegung}) .$$

- 2) Aus Gl. 8.28 folgt wegen $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{u}$, dass das Produkt $\mathbf{X}'\mathbf{u}$ einen Nullvektor ergeben muss also

$$(8.32) \quad \mathbf{X}'\mathbf{u} = \mathbf{0}$$

gilt (oder ausführlich $\Sigma u = \Sigma x_1 u = \Sigma x_2 u = \dots = \Sigma x_p u = 0$).

- 3) Aus Gl. 8.31 und 8.32 folgt generell $s_{\hat{y}}^2 = s_{y\hat{y}}$, so dass ganz allgemein im Regressionsmodell gilt:

$$(8.17a) \quad r_{y\hat{y}} = R_{y,ij\dots} = \frac{s_{\hat{y}}}{s_y} = \sqrt{B_{y,ij\dots}}$$

c) Multiple Korrelation und multiple Bestimmtheit

Die Güte des Gesamtzusammenhangs wird beschrieben mit dem multiplen Korrelationskoeffizient $R_y = R_{y,12}$ [oder $r_{y,12}$] (bei zwei Regressoren X_1 und X_2 , was im folgenden als Beispiel dienen mag), bzw. mit dem multiplen Bestimmtheitsmaß R_y^2 . Anders als bei der einfachen Regression macht

es hier (wie auch bei nichtlinearer Regression) keinen Sinn, zwischen positiver und negativer Korrelation zu unterscheiden, so dass R nie negativ sein kann.

Die Bezeichnung des multiplen Korrelationskoeffizienten mit $R_{y,12}$ macht deutlich, dass in der Regressionsfunktion Y durch X_1 und X_2 "erklärt" wird.

Def. 8.5: multiple Korrelation und Bestimmtheit

- a) Der multiple Korrelationskoeffizient $R_{y,12}$ ist die einfache Korrelation zwischen y und $\hat{y} = b_{y1.2} \cdot x_1 + b_{y2.1} \cdot x_2$ also zwischen dem beobachteten Wert y und dem Regresswert \hat{y}

$$(8.33) \quad R_{y,12} = r_{y\hat{y}}$$

- b) Die multiple Bestimmtheit $B_{y,12}$ ist analog zum einfachen Bestimmtheitsmaß der Anteil der durch die multiple Regression erklärten Varianz der abhängigen Variable Y an der Gesamtvarianz von Y also:

$$(8.34) \quad B_{y,12} = \frac{s_{\hat{y}}^2}{s_y^2} \quad (\text{mit } \hat{y} \text{ als Funktion von } x_1 \text{ und } x_2)$$

und analog bei mehr als zwei Regressoren; es ist nach Gl. 8.17a gleich $R_{y,12}^2$.

- c) Die Größe $U_{y,12} = 1 - B_{y,12} = s_u^2/s_y^2$, wobei s_u^2 die Varianz der Störgröße U ist, ist entsprechend das multiple Unbestimmtheitsmaß.

Auch für das multiple Bestimmtheitsmaß liegt der Wertebereich zwischen 0 (keine Erklärung) und 1 (alle Beobachtungen "liegen auf" der Regressionshyperebene, perfekte Erklärung von Y durch X_1, X_2, \dots, X_p). Den multiplen Korrelationskoeffizienten $R_{y,12..p}$ zwischen Y und X_1, \dots, X_p erhält man auch aus der positiven Wurzel des multiplen Bestimmtheitsmaßes.

d) Partielle Regressions- und Korrelationskoeffizienten

Der partielle Regressionskoeffizient erster Ordnung $b_{y1.2}$ ist der isolierte (partielle) Einfluss von X_1 auf Y in dem Sinne, dass der Einfluß

- von X_2 auf Y in Höhe von $b_{y2}x_2$ aus der einfachen Regression $y = b_{y2}x_2 + u$ und
- von X_2 auf X_1 in Höhe von $b_{12}x_2$ aus der einfachen Regression

$$x_1 = b_{12}x_2 + v$$

ausgeschaltet ist (wobei hier zentrierte Werte [deshalb kein Absolutglied] vorausgesetzt sind).

Dann ist $b_{y1,2}$ der Regressionskoeffizient in der einfachen Regression von $u = y - b_{y2}x_2$ (Regressand) auf $v = x_1 - b_{12}x_2$ (Regressor) und der partielle Korrelationskoeffizient $r_{y1,2}$ ist der (einfache) Korrelationskoeffizient zwischen u und v .

Man beachte, dass $b_{y1,2}$ und $b_{1y,2}$ verschiedene Größen sind, nicht aber $r_{y1,2}$ und $r_{1y,2}$. Es gilt ferner:

$$(8.35) \quad r_{ij,k} = \sqrt{b_{ij,k} b_{ji,k}}$$

analog zur einfachen (bivariaten) Regression.

Bei der einfachen Regressionsanalyse gab es nur einen Korrelationskoeffizienten r . Er steht auf der gleichen Ebene wie der multiple Korrelationskoeffizient $R_{y,12,\dots}$, da auch bei einfacher Korrelation r als Korrelation zwischen y und \hat{y} dargestellt werden kann.

e) Standardisierte Regressionskoeffizienten, Rekursionsformeln

1. Standardisierte Regressionskoeffizienten

Man beachte:

- Bei Kollinearität misst der Regressionskoeffizient $b_{ij,k}$ nicht den isolierten Einfluß von X_j (Regressor, "Ursache") auf X_i (Regressand, "Wirkung"), sondern einen "gemischten Einfluß von X_j **und** X_k ", ohne dass man diese Einflüsse trennen könnte.
- Aussagen dergestalt, welche Einflußgröße die (kausal) wichtigere sei sind außerdem auch deshalb nur mit großer Vorsicht zu machen, weil die Regressionskoeffizienten maßstabsabhängig sind.

Es kann sehr wohl X_k ein bedeutsamerer (größerer) Einfluß auf Y sein, als X_i - auch wenn $b_{yi,k} > b_{yk,i}$ gilt - einfach deshalb, weil die Varianz von X_i kleiner ist als die von X_k . Statt der mit Gl. 8.30 zunächst errechneten nichtstandardisierten Regressionskoeffizienten $b_{yi,k}$ sind die standardisierten Regressionskoeffizienten $c_{yi,k}$ (oft auch $\beta_{yi,k}$ genannt, was dann aber mit den "wahren" Koeffizienten der Gl. 8.24 verwechselt werden könnte) zu berechnen nach der Formel:

$$(8.36) \quad c_{y_i.k} = b_{y_i.k} \frac{s_i}{s_y},$$

wobei s^2 die entsprechenden Varianzen sind.

Man erhält den Vektor \mathbf{c} der standardisierten Regressionskoeffizienten aus dem Normalgleichungssystem

$$(8.28a) \quad \mathbf{Rc} = \mathbf{r}$$

mit der Matrix \mathbf{R} der Korrelationskoeffizienten (anstelle von $\mathbf{X}'\mathbf{X}$ in Gl. 8.28 und dem Vektor \mathbf{r} der (einfachen) Korrelationen zwischen den Regressoren und Y (anstelle von $\mathbf{X}'\mathbf{y}$). Bei zwei Regressoren gilt z.B.:

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} c_{y1.2} \\ c_{y2.1} \end{bmatrix} = \begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix}$$

2. Rekursionsformeln

Der partielle Regressionskoeffizient erster Ordnung $b_{ij.k}$ läßt sich rekursiv aus den einfachen Regressionskoeffizienten (und entsprechend auch die partielle Korrelation erster Ordnung aus den einfachen Korrelationen) wie folgt bestimmen:

$$(8.37) \quad b_{ij.k} = \frac{b_{ij} - b_{ik}b_{ki}}{1 - b_{ik}b_{ki}},$$

$$(8.38) \quad r_{ij.k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}.$$

Man kann die partiellen Regressionskoeffizienten auch auf Varianzen und Kovarianzen zurückführen. Dann gilt:

$$(8.37a) \quad b_{ij.k} = \frac{s_{ij}s_k^2 - s_{jk}s_{ik}}{s_j^2s_k^2 - s_{jk}^2},$$

$$(8.37b) \quad b_{ij.k} = \frac{s_i}{s_j} \cdot \frac{r_{ij} - r_{jk}r_{ik}}{1 - r_{jk}^2}.$$

Der zweite Bruch in Gl. 8.37b stellt den standardisierten Regressionskoeffizienten $c_{ij.k}$ (gem. Gl. 8.36) dar. Wie man sieht sind bei unkorrelierten Regressoren ($r_{jk} = 0$) die partiellen Regressionskoeffizienten $b_{ij.k}$ identisch mit den einfachen b_{ij} .

Für die Interpretation der partiellen Regressionskoeffizienten mag auch der folgende Zusammenhang von Interesse sein:

$$(8.37c) \quad b_{ij} = b_{ij.k} + b_{ik.j} \cdot b_{jk}$$

wonach der einfache Regressionskoeffizient b_{ij} die Summe eines direkten Einflusses von X_j auf X_i (gemessen als $b_{ij.k}$) und eines indirekten Einflusses von X_k über X_j auf X_i (mit $b_{ik.j} \cdot b_{jk}$) ist. Deshalb mißt auch b_{ij} nicht, wie man meinen möchte den "echten" Einfluß von x_j auf x_i , sondern stets einen direkten **und** indirekten Einfluß von x_j auf x_i . Der indirekte Einfluß kann auch dominieren, wie das bei einer Scheinkorrelation der Fall ist. Eine entsprechende Beziehung gilt auch für die Korrelation.

Entsprechende Formeln existieren auch für die partiellen Regressions- und Korrelationskoeffizienten zweiter Ordnung, die rekursiv aus den Koeffizienten erster Ordnung zu bestimmen sind.

Für die rekursive Bestimmung der multiplen Bestimmtheit gilt:

$$(8.39) \quad R_{i,jk}^2 = \frac{r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}r_{jk}}{1 - r_{jk}^2}$$

Aus rekursiven Beziehungen ergibt sich übrigens auch:

$$(8.40) \quad 1 - R_{y,123} = (1 - R_{y,12})(1 - r_{y3,12}),$$

so dass $R_{y,123} \geq R_{y,12}$ oder allgemein die multiple Bestimmtheit bei p Regressoren nicht kleiner sein kann als bei $p-1$ Regressoren³. Aus Gl. 8.40 folgt übrigens auch die folgende Interpretation der partiellen Bestimmtheit im Sinne der proportionalen Fehlerreduktion ([PRE] vgl. Def. 7.14):

$$(8.40a) \quad r_{y3,12}^2 = \frac{R_{y,123}^2 - R_{y,12}^2}{1 - R_{y,12}^2}$$

wonach $r_{y3,12}^2$ die Zunahme der Bestimmtheit von Y durch den Regressor X_3 (oder die Abnahme der Unbestimmtheit!) ins Verhältnis zur Unbestimmtheit mit den beiden Regressoren X_1 und X_2 setzt. Es gilt $R_{y,123}^2 > R_{y,12}^2$ es sei denn $r_{y3,12} = 0$.

Zusammenhänge dieser Art veranlassen auch manche Anwender mit verschiedenen Verfahren der stufenweisen Regression (stepwise regression) durch Hinzufügen oder Weglassen von Regressoren zu einer Regressionsgleichung mit möglichst großer Bestimmtheit und wenig Regressoren zu gelangen. Es kann hier nicht näher begründet werden, warum ein solches Vorgehen mit Skepsis zu betrachten ist. Der Hauptgrund hierfür ist die Kollinearität. Denn nur bei Orthogonalität (d.h. Abwesenheit von Kollinearität) ist die multiple Bestimmtheit eine Summe der einfachen Bestimmtesten:

³ Dem Umstand wird oft durch Berechnung einer korrigierten Bestimmtheit $R_{y,123}^{-2}$ Rechnung getragen: $R_{y,123}^{-2} = 1 - (n-1)(1-R^2)/(n-p)$ wobei $R_{y,123}^{-2} \mid R$

$$(8.40b) \quad R_{y,123}^2 = r_{y1}^2 + r_{y2}^2 + r_{y3}^2.$$

Aus Gl. 8.28a folgt auch Gl. 8.37b, wonach $c_{ij,k} = (r_{ij} - r_{jk}r_{ik}) / (1 - r_{jk}^2)$ und dass bei Orthogonalität ($r_{jk} = 0$) gilt $c_{ij,k} = r_{ij}$. Ferner wird erkennbar, dass die multiple Bestimmtheit nur bei Orthogonalität mit Gl. 8.40b als Summe partieller Bestimmtheiten darstellbar ist. Wegen der Standardisierung ($s_y = s_1 \dots = s_p = 1$) gilt dagegen z.B. bei zwei korrelierten Regressoren:

$$(8.41) \quad R_{y,12}^2 = c_{y1,2}^2 + c_{y2,1}^2 + 2c_{y1,2}c_{y2,1}r_{12}.$$

Zur Summe der partiellen Bestimmtheiten tritt also bei Kollinearität noch das doppelte Produkt der c-Koeffizienten mit der positiven oder negativen Korrelation r_{12} hinzu, das bei Orthogonalität wegfällt.

Beispiel 8.6:

Der Bierverkauf einer Brauerei hänge wie folgt von der erklärenden Variable "Ausgaben für Fernsehwerbung" X_1 und einer zweiten erklärenden Variable "Ausgaben für Zeitungswerbung" X_2 ab. Es ergibt sich die folgende (natürlich fiktive) Datenkonstellation:

v	Y_v	X_{1v}	X_{2v}
1	0,8	1,5	1,0
2	1,0	4,0	1,5
3	1,4	4,5	3,0
4	1,6	7,0	5,0
5	1,7	8,5	4,5
6	2,1	8,5	5,5
7	2,1	11,5	7,0
8	2,5	12,0	8,5

Man bestimme die lineare Zweifachregression $\hat{y} = b_0 + b_1x_1 + b_2x_2$ und zeige, ob und in welchem Maße die Hinzunahme eines zweiten Regressors (X_2) zur Erklärung von Y beiträgt. Hilfsangaben:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0,725175874 & -0,171375851 & 0,1403529 \\ -0,171375851 & 0,173451736 & -0,238957444 \\ 0,1403529 & -0,238957444 & 0,350478607 \end{bmatrix} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 13,2 \\ 109,15 \\ 69,65 \end{bmatrix}$$

Lösung 8.6:

Man erhält

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \begin{bmatrix} 0,642226963 \\ 0,026709722 \\ 0,181288202 \end{bmatrix}$$

also: $\hat{y} = 0,642226963 + 0,026709722x_1 + 0,181288202x_2$. Das (unkorrigierte) multiple Bestimmtheitsmaß beträgt $R_{y,12}^2 = 0,957045$. Es wird also 95,7% der Varianz der abgesetzten Biermenge Y durch die Zweifachregressionsfunktion erklärt.

Für die einfache Regression von Y auf X_1 erhält man $\hat{y} = 0,569682 + 0,1503126x_1$ und für die Bestimmtheit $R_{y,1}^2 = 0,916971$. Das Beispiel enthält eine beachtliche Kollinearität, denn X_1 und X_2 korrelieren mit $0,969171$. Deshalb ist auch die multiple Bestimmtheit $R_{y,12}^2$ sehr viel kleiner als die Summe der einfachen Bestimmtheiten $R_{y,1}^2 = r_{y1}^2 = 0,916971$ und $R_{y,2}^2 = r_{y2}^2 = 0,955287$. Der Erklärungsbeitrag von X_2 ist nach Gl. 8.40a die partielle Bestimmtheit und er beläuft sich auf nur 48%:

$r_{y2,1}^2 = (R_{y,12}^2 - R_{y,1}^2) / (1 - R_{y,1}^2) = 0,040074 / 0,083029 = 0,48262$, so dass die partielle Korrelation nur $0,6947$ beträgt. Man kann den partiellen Korrelationskoeffizienten auch mit $r_{y2,1} = (r_{y2,1} - r_{y1} r_{12}) / \sqrt{(1 - r_{y1}^2)(1 - r_{12}^2)}$ rekursiv berechnen.

5. Nichtlineare Regression

Oft lassen theoretische Erkenntnisse oder die Betrachtung des Streudiagramms darauf schließen, dass ein nichtlinearer Zusammenhang zwischen den Variablen besteht und somit eine nichtlineare Regressionsfunktion bestimmt werden muss, worauf hier nur kurz eingegangen werden kann.

Eine Regressionsfunktion ist **nichtlinear**

- **in den Variablen**, wenn Regressoren wie etwa x , x^2 , x^{-1} , e^x , x_1x_2 , $\sin(x)$ oder $\log(x)$ auftreten
- **in den Regressionskoeffizienten (Parametern)**, wenn in der Gleichung Koeffizienten auftreten wie b^2 oder b^{-1} usw.
- in den Variablen **und** in den Parametern, wenn beides zutrifft, wie etwa bei der logistischen Funktion als Regressionsfunktion $\hat{y}_i = k / [1 + \exp(a - bx_i)]$. Man kann dann auch von **intrinsischer** Nichtlinearität sprechen.

Ist eine Funktion allein nichtlinear in den Variablen oder in den Parametern, nicht aber in beiden, so kann sie meist linearisiert werden durch eine Variablen- bzw. Parametersubstitution oder durch eine Variablentransformation. Übersicht 8.2 gibt hierzu einige Hinweise. Wenn sie nichtlinear in den Variablen **und** in den Parametern ist, so sind andere Methoden notwendig (z.B. eine Näherung durch eine Taylorreihenentwicklung, die dann meist nach einigen Gliedern abgebrochen wird).

1. Variablensubstitution

Ist z.B. eine parabolische (Polynom zweiten Grades) Regression zu schätzen ($\hat{y}_v = b_0 + b_1x_v + b_2x_v^2$) so kann man x durch x_1 und x^2 durch x_2 substituieren, also neue Variablen definieren (*redefining variables*). Das führt zu den folgenden Normalgleichungen:

$$b_0n + b_1\Sigma x_1 + b_2\Sigma x_2 = \Sigma y$$

$$b_0\Sigma x_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1 x_2 = \Sigma x_1 y$$

$$b_0\Sigma x_2 + b_1\Sigma x_1 x_2 + b_2\Sigma x_2^2 = \Sigma x_2 y$$

die äquivalent sind dem Gleichungssystem:

$$b_0n + b_1\Sigma x + b_2\Sigma x^2 = \Sigma y$$

$$b_0\Sigma x + b_1\Sigma x^2 + b_2\Sigma x^3 = \Sigma xy$$

$$b_0\Sigma x^2 + b_1\Sigma x^3 + b_2\Sigma x^4 = \Sigma x^2 y .$$

Die Parabel $y = b_0 + b_1x + b_2x^2$ ist zwar nicht linear in x , wohl aber linear in den Variablen x_1 und x_2 und zu schätzen mit einer multiplen Regression mit den Regressoren $x_1 = x$ und $x_2 = x^2$.

Bei Nichtlinearität in den Variablen ist eine Linearisierung durch Variablensubstitution möglich. Das bedeutet, dass die Formeln der linearen Regression zur Bestimmung der Regressionskoeffizienten bei Verwendung der transformierten Variablen angewandt werden können und (im Unterschied zu vielen Fällen der Variablentransformation) auch die üblichen Annahmen über die (additive) Störgröße gültig bleiben.

Übersicht 8.2: Einige leicht linearisierbare Funktionen

Die Übersicht ist ähnlich aufgebaut, wie Übers. 9.5, in der zugleich Trendmodelle behandelt werden (vgl. auch Kap. 11). Allerdings ist hier x an die Stelle von t in Übers. 9.5 gesetzt worden. Der Parameter k ist wie in Übers. 9.5 ein Sättigungsniveau.

	Funktion f(x) y=f(x)	linearisiert: y*=a*+b*x*			
		y*	a*	b*	x*
1	$(a+bx)^\alpha$				
1a	$\alpha=1$: Gerade $y = a+bx$	y	a	b	x
1b	$\alpha=-1$: $\frac{1}{a+bx}$	$\frac{1}{y}$	a	b	x
1c	$\alpha=1/2$: $\sqrt{a+bx}$	y^2	a	b	x
1d	Potenzfunktion bx^α	ln(y)	ln(b)	α	ln(x)
2	Parabel $a+bx+cx^2$ (1)				
3	$a \cdot \exp(bx^\alpha)$				
3a	$\alpha=1$: ae^{bx} (2)	ln(y)	ln(a)	b	x
3b	$\alpha=-1$: $ae^{b/x}$	ln(y)	ln(a)	b	$\frac{1}{x}$
4	$k+be^{cx}$ (3)				
5	$k+\frac{b}{c+x}$ (Hyperbel)				
5a	$k+b/x$ (c=0)	y	k	b	x^{-1}
5b	$b/(c+x)$ (k=0)	y^{-1}	c/b	b^{-1}	x
6	$\frac{k(x+a)}{x+b}$ (b>a)				
6a	$kx/(x+b)$ (a=0)	y^{-1}	k^{-1}	b/k	x^{-1}
6b	$x/(cx+b)$ (a=0, k=1)	y^{-1}	c	b	x^{-1}
6c	$k(x+a)/x$ (b=0)	y	k	ak	x^{-1}
7	$\ln(y)=K-a/(b+x)$ (4)				
7a	$\ln(y)=K-a/x$ (b=0) (5)	ln(y)	K	-a	x^{-1}

- (1) Linearisierbar als multiple lineare Regression mit Variablensubstitution $z = x^2$; man erhält dann: $a+bx+cz$. Entsprechend ist die im Beispiel 9.13 genannte logarithmische Parabel zu linearisieren mit $y^* = \ln(y)$ und $z = x^2$, also $y^* = a+bx+cz$.
- (2) Oder in allgemeinerer Form $y = ar^x$, mit $r=e^b$ und damit $b = \ln(r)$.
- (3) Oder $y=k+br^x$ mit $r=e^c$ (k: Sättigungsniveau); das ist die "modifizierte Exponentialfunktion". Mit $k=0$ erhält man den Fall 3a. Im Falle $k \neq 0$ ist aber die Schätzung schwierig. Mit einem Versuchswert für k kann man die lineare Funktion $\ln(y-k) = \ln(b) + cx$ schätzen.
- (4) In Übers. 9.5 unter Nr. 9 (Johnson-Funktion); $k=e^K$ ist das Sättigungsniveau.
- (5) Dies ist äquivalent mit Nr. 3b, denn $y = k \cdot e^{-a/x}$ ($k=e^K$) hat die gleiche Form wie $y = ae^{b/x}$. Die Funktion steigt S-förmig mit einem Wendepunkt bei $x=a/2$ auf ein Sätti-

gungsniveau von k . Sie ist im Unterschied zu der logistischen Funktion jedoch nicht punktsymmetrisch um $x=a/2$ weil bei diesem Wert von x die Ordinate $y=k-2$ und nicht $y=k/2$ ist.

2. Variablentransformation

Viele v.a. in den Parametern nichtlineare Funktionen lassen sich durch geeignete Transformationen, z.B. durch Logarithmieren in eine lineare Form bringen. Dadurch wird die nicht nur komplizierte, sondern häufig auch nicht eindeutige Lösung eines nichtlinearen Normalgleichungssystems umgangen.

Eine direkte Schätzung der Regressionsfunktion

$$(8.42) \quad y = b \cdot x^\alpha + u$$

mit der Methode der kleinsten Quadrate führt zu den beiden (in den zu schätzenden Parametern α und b) nichtlinearen Normalgleichungen

$$(1) \quad \Sigma y x^\alpha = b \Sigma x^{\alpha+2} \quad \text{und} \\ (2) \quad 2b\alpha \Sigma y x^{\alpha-1} = (\alpha+2)b^2 \Sigma x^{\alpha+1}.$$

Oder im Modell

$$(8.43) \quad y = e^{-bx} + u$$

wäre der eine Parameter b durch die Normalgleichung

$$\Sigma x_i \cdot \exp(-2bx_i) = \Sigma y_i \cdot x_i \cdot \exp(-bx_i)$$

zu schätzen.

Offensichtlich kann die wegen ihrer konstanten Elastizität in der Ökonomie (als Modell der Kosten- oder Nachfragefunktion) sehr beliebte Potenzfunktion $y_i = b \cdot x_i^\alpha$ (gem Gl. 8.42) als nichtlineare Funktion durch Logarithmieren linearisiert werden, denn aus

$$(8.42a) \quad y_i = b \cdot x_i^\alpha \quad \text{folgt}$$

$$(8.42b) \quad \ln(y_i) = \ln(b) + \alpha \cdot \ln(x_i).$$

Bei Variablentransformationen dieser Art ist jedoch auch die Störgröße in der Regression zu beachten. So ist z.B. Gl. 8.42b nur dann die linearisierte Schätzfunktion, wenn man das Modell

$$(8.42^*) \quad y_i = b \cdot x_i^\alpha \cdot \exp(u_i)$$

zugrundelegt, in dem die Störgröße (der **Störfaktor**) e^u multiplikativ auftritt (also umso größere Werte annimmt, je größer y ist). Entsprechendes

gilt bei Potenzfunktionen mit zwei Regressoren x_1 und x_2 , etwa der Cobb-Douglas-Funktion $y = b x_1^\alpha x_2^\beta e^u$. Dagegen ließe sich das Modell

$$(8.42) \quad y_i = b \cdot x_i^\alpha + u_i$$

nicht so einfach durch Logarithmieren linearisieren, weil die Störgröße U hier nicht in einer Form vorkommt, bei der sie nach Transformation als additives Störglied erscheint.

Auch bei Variablensubstitutionen ist die Störgröße zu beachten. Im Beispiel 8.7 wird auf dieses Problem (einer ungerechtfertigten deterministischen Interpretation der Regressionsgleichung, d.h. Vernachlässigung der Störgröße) eingegangen.

3. Intrinsische Nichtlinearität

Nicht linearisierbar durch Variablensubstitution oder -transformation sind Funktionen, die nichtlinear in den Variablen **und** in den Parametern sind. Das gilt z.B. für die in Übers. 9.5 unter Nr. 9 genannte logistische Funktion $y_i = k/[1+\exp(a-bx_i)]$.

Eine Linearisierung wäre z.B. dann möglich, wenn die Sättigungsgrenze k bekannt wäre (oder wenn man hierfür einen brauchbaren "Versuchswert" ansetzen könnte). Man erhält dann die Schätzgleichung $\ln[(k-y_i)/y_i] = a - bx_i$. Die logistische Funktion ist linear in den Variablen x und $\ln[(k-y)/y]$ (den sog. "Logits") und in den zu schätzenden Parametern a und b . Ist ein Parameter (wie etwa k) dagegen nicht als bekannt vorauszusetzen, so sind Funktionen dieser Art nicht einfach durch Transformationen (z.B. Logarithmieren) oder Variablensubstitution zu linearisieren.

Def. 8.6: Linearität einer Regression

Eine Regressionsfunktion ist linear (oder linearisiert), wenn die abhängige Variable y (bzw. die transformierte oder substituierte Variable y) im Mittel eine lineare Funktion von z_1, z_2, \dots ist, wobei die Größen z_1, z_2 usw. **bekannt** Funktionen der Regressoren x_1, x_2 sind.

Erläuterung zu Def. 8.6:

So ist z.B. die Funktion

(8.43a) $y = e^{-bx+u} = e^{-bx}e^u$ oder äquivalent $\ln(y_i) = -bx_i + u_i$ linear in $\ln(y)$ und x . Linearisierbar ist auch die Funktion

(8.43b) $y = \exp(-\frac{b}{x} + u) = e^{-b/x}e^u$ oder äquivalent $\ln(y_i) = -b(1/x_i) + u_i$, weil $1/x$ eine **bekannt** Funktion von x ist.

Dagegen ist die unter Nr. 8 in Übers. 9.5 genannte Funktion

$y = \exp(k + br^x + u)$ mit $r = e^c$ (Gompertz-Funktion)

nichtlinear, weil r^x in $\ln(y) = k + br^x + u_i$ **keine bekannte** Funktion von x ist solange r nicht bekannt ist.

Beispiel 8.7:

- a) Ist die nichtlineare Funktion (vgl. Nr. 6a in Übers. 8.2)

$$(8.44) \quad y_i = \frac{kx_i}{x_i + b}$$

zu linearisieren, indem man x und y durch die reziproken Variablen substituiert, also mit

$$(1) \quad \frac{1}{y_i} = \frac{1}{k} + \frac{b}{k} \frac{1}{x_i} \quad \text{oder ist}$$

$$(2) \quad \frac{x_i}{y_i} = \frac{b}{k} + \frac{1}{k} x_i \quad \text{die "richtige" Schätzgleichung?}$$

- b) Man schätze die Funktion 8.44 mit den beiden unter a) genannten Schätzgleichungen für folgende Daten:

x	1	2	3	4	5	6	7
y	2,7	4,4	5,3	6,6	6,5	6,3	7,7

- c) Kann man von den gleichen Eigenschaften der Störgröße wie bei linearer Regression ausgehen?

Lösung 8.7:

- a) Die Gleichungen 1 und 2 sind nur dann äquivalente Umformungen von Gl. 8.44, wenn man die Störgröße nicht beachtet, wenn man also den Fehler begeht, die Regressionsfunktion deterministisch zu interpretieren. Die Parameter $1/k$ und b/k mit der Methode der kleinsten Quadrate zu schätzen setzt voraus, dass gilt:

$$(1) \quad \frac{1}{y_i} = \frac{1}{k} + \frac{b}{k} \frac{1}{x_i} + u_{1i}$$

bzw. dass man rücktransformiert von der Gleichung

$$(8.44a) \quad y_i = \frac{kx_i}{x_i + b + w_{1i}}$$

mit $w_{1i} = kx_i u_{1i}$ als Störgröße ausgehen kann. Entsprechend läuft der Ansatz (2) darauf hinaus, dass gilt

$$(2) \quad \frac{x_i}{y_i} = \frac{b}{k} + \frac{1}{k} x_i + u_{2i} \text{ und damit r\"ucktransformiert}$$

$$(8.44b) \quad y_i = \frac{kx_i}{x_i + b + w_{2i}} \text{ mit } w_{2i} = ku_{2i} \text{ als St\"orgr\"o\ss}e.$$

In keinem der beiden F\"alle ist damit aber direkt das Modell

$$(8.44c) \quad y_i = \frac{kx_i}{x_i + b} + u_i \text{ mit } u_i \text{ als St\"orgr\"o\ss}e$$

gesch\"atzt.

Es handelt sich also (und das wird oft \"ubersehen) bei allen drei Gleichungen (8.44a bis 8.44c) um verschiedene Modelle, die nur bei einer deterministischen Interpretation der Regression als \"aquivalent angesehen werden k\"onnen.

- b) Die Daten stammen aus einem Lehrbuch von G. Ross⁴, der die Gl. 8.44 mit verschiedenen Methoden gesch\"atzt hatte. So ergab z.B. eine direkte Sch\"atzung von Gl. 8.44c mit nichtlinearen Normalgleichungen: $k = 9,956$ und $b = 2,549$.

Mit dem Ansatz (1) erh\"alt man die (zur\"ucktransformierten) Parametersch\"atzer $k = 10,32046$ und $b = 2,79428$ bei einer Bestimmtheit von $r^2 = 0,98718$ und mit dem Ansatz (2) $k = 9,94663$ und $b = 2,57574$ bei $r^2 = 0,95197$.

- c) Anhand des Ansatzes 1 (Gl. 8.44a) soll gezeigt werden, dass die \"ublichen Eigenschaften einer St\"orgr\"o\ss}e nicht gelten f\"ur die Abweichung zwischen y und dem inversen Sch\"atzwert f\"ur y^{-1} nach der folgenden Sch\"atzgleichung

$$(1) \quad y_i^{-1} = k^{-1} + (b/k)x_i^{-1} + u_{1i}$$

Sch\"atzt man Gl.(1) so erh\"alt man das Ergebnis $1/k = 0,096895$ und $b/k = 0,270751$ woraus sich die oben angegebenen Werte k und b zur\"uckrechnen lassen. Die

folgende Tabelle gibt die mit der Sch\"atzgleichung errechneten Regresswerte $\left(\frac{1}{\hat{y}}\right) =$

$\frac{1}{\hat{y}_i}$ an sowie die sich daraus ergebenden Sch\"atzwerte \hat{y} f\"ur y (mit $\hat{y} = 1/(\hat{y}^{-1})$). F\"ur die St\"orgr\"o\ss}e

$$u_i = \frac{1}{y_i} - \left(\frac{1}{\hat{y}}\right)$$

⁴ Ross, Garvin J.S., Nonlinear Estimation, New York usw. (Springer Verlag) 1990.

gelten natürlich die in den Gl. 8.11ff genannten Eigenschaften (so ist etwa $\sum u_i = 0$), nicht aber für die Störgröße $w_i = y_i - \hat{y}_i$ (so ist $\sum w_i$ nicht 0 sondern 0,0174):

x_i (1)	y_i (2)	y_i^{-1} (3)	$\hat{1/y}_i$ (4)	u_i (5)	\hat{y}_i (6)	w_i (7)
1	2,7	0,37037	0,36765	0,00272	2,720006	-0,020006
2	4,4	0,22727	0,23227	-0,00500	4,305323	0,094677
3	5,3	0,18868	0,18715	0,00500	5,343439	-0,043439
4	6,6	0,15152	0,16458	-0,01307	6,075973	0,524027
5	6,5	0,15385	0,15105	0,00280	6,620534	-0,120534
6	6,3	0,15873	0,14202	0,01671	7,041256	-0,741256
7	7,7	0,12987	0,13557	-0,00570	7,376062	0,323938

Zwischen den Spalten bestehen folgende Beziehungen: (3) ist der reziproke Wert von (2) und (6) der reziproke Wert von (4). Ferner gilt (5) = (3)-(4) sowie (7)=(2)-(6).