II. Deskriptive Statistik II

2.1 Lernziele zur Deskriptiven Statistik II

- Schätzung der Überlebensraten nach Kaplan-Meier
- Abhängigkeit zweier Merkmale
- Kontingenztafel
- Regression
- Korrelation

2.2 Schätzung der Überlebensraten nach Kaplan-Meier

In der Medizin werden häufig Merkmale vom Typ einer Überlebenszeit betrachtet. Man versteht darunter Merkmale, die wie eine Überlebenszeit durch ein **Anfangs- und ein Enddatum** charakterisiert sind.

Beide Angaben sind jeweils durch das Eintreten eines Ereignisses gekennzeichnet. Bei den eigentlichen Überlebenszeiten ist das Anfangsdatum z. B. das Datum der Erstdiagnose einer Erkrankung, das Enddatum ist das Todesdatum. Es kann aber auch das Anfangsdatum das Datum einer Operation, und das Enddatum das Datum der Entlassung aus dem Krankenhaus sein. Die Überlebenszeit ist jeweils die Zeitspanne zwischen beiden Daten.

Man spricht von einer zensierten Überlebenszeit, wenn das Endereignis am Stichtag der Auswertung noch nicht eingetreten ist. In diesem Fall steht für die Auswertung nur eine untere Schranke für die noch nicht bekannte tatsächliche Überlebenszeit zur Verfügung.

Unter der **Überlebensrate** *S(t)* versteht man den Anteil der Individuen, deren Überlebenszeit größer als t ist (S für engl.: survival). Es besteht die Aufgabe, diesen Anteil aus den Daten zu schätzen. Den errechneten Schätzwert bezeichnet man üblicherweise mit $\hat{S}(t)$.

Die zensierten Überlebenszeiten bilden ein Problem bei der Berechnung von $\hat{S}(t)$. Ein Verfahren, das es gestattet, die zensierten Überlebenszeiten sinnvoll einzubeziehen, ist das Schätzverfahren von E. Kaplan und P. Meier.

Das Verfahren soll anhand der Beispieldaten aus Tabelle 2.1 beschrieben werden. Die Tabelle enthält in Spalte (2) die Überlebenszeiten von 20 Tieren aus einem Tierversuch. Die Überlebenszeiten sind bereits als Differenz von Anfangs- und Enddatum ausgerechnet und in Tagen angegeben. Da die Tiere im Allgemeinen nicht alle am gleichen Tag in den Versuch aufgenommen werden, müssen die Versuchstage für jedes Tier individuell gezählt werden. Versuchstag 20 für Tier A kann z. B. für Tier B der Versuchstag 5 sein. Zensierte Zeiten sind durch "+" gekennzeichnet. Die Zeiten sind - gleichgültig ob zensiert oder nicht - der Größe nach geordnet:

$$t_0 = 0 < t_1 < t_2 < \ldots < t_n$$

Die verschiedenen Zeiten sind in Spalte (1) durchnummeriert. Zwei Tiere sind zum Beispiel gleichzeitig an Versuchstag 70 eingegangen. Spalte (3) enthält die Anzahl n_i derjenigen Versuchstiere, die den jeweiligen Versuchstag t_i lebend erreichen, man sagt auch, die zum Zeitpunkt t_i im Risiko stehen.

Diese Zahlen $n_1, n_2, n_3 \ldots$ errechnet man sukzessive mit Hilfe der Angaben in Spalte (4). Dort steht die Anzahl d_i der zum Zeitpunkt t_i eingegangenen Tiere. Immer wenn t_i nicht zensiert ist, ist ein Tier eingegangen. Daher ist z. B. $d_3 = 0$, denn $t_3 = 43$ ist eine zensierte Überlebenszeit. Zum Zeitpunkt $t_6 = 70$ sind 2 Versuchstiere eingegangen, d. h. $d_6 = 2$.

Offenbar gilt

$$n_i = n_{i-1} - d_{i-1} \quad (i = 1,2,\ldots),$$

das heißt zum Beispiel für den zweiten Zeitpunkt $t_2 = 40$:

$$n_2 = n_1 - d_1 = 20 - 1 = 19.$$

Nach diesen Vorbereitungen besteht die Grundidee des Kaplan-Meier-Verfahrens darin, zunächst für jeden Zeitpunkt t_i die **bedingten Überlebensraten** q_i auszurechnen: Abhängigkeit zu untersuchen, braucht man andere Methoden.

$$q_i = \frac{n_i - d_i}{n_i} \quad (i = 1,2,\ldots)$$

das heißt zum Beispiel für den zweiten Zeitpunkt t_2=40:

$$q_i = \frac{n_2 - d_2}{n_2} = \frac{19 - 1}{19} = 0.9474$$

Das ist der Anteil derer, die den Zeitpunkt t_2 überleben, von all denen, die ihn erreichen. Die q_i werden in Spalte (5) berechnet. Die geschätzte Überlebensrate $\hat{S}(t)$ erhält man durch Aufmultiplizieren aller q_i. Dies ist in Spalte (6) notiert:

$$\hat{S}(t) = \prod_{t_i \le t} q_i$$

das heißt zum Beispiel für den zweiten Zeitpunkt t_2=40:

$$\hat{S}(t_2) = \hat{S}(40) = q_1 \cdot q_2 = 0.95 \cdot 0.9474 = 0.9$$

Beispiel 2.1
Tabelle 2.1 enthält aus einem Tierversuch 20 Überlebenszeiten in Tagen. Die Zeiten sind bereits aufsteigend sortiert. An den mit (+) gekennzeichneten Zeitpunkten endet die Beobachtungszeit, ohne dass das betrachtete Ereignis (hier Tod des Versuchstiers) eingetreten ist. Solche am Stichtag der Auswertung noch anhaltenden Überlebenszeiten nennt man ***zensiert****.*

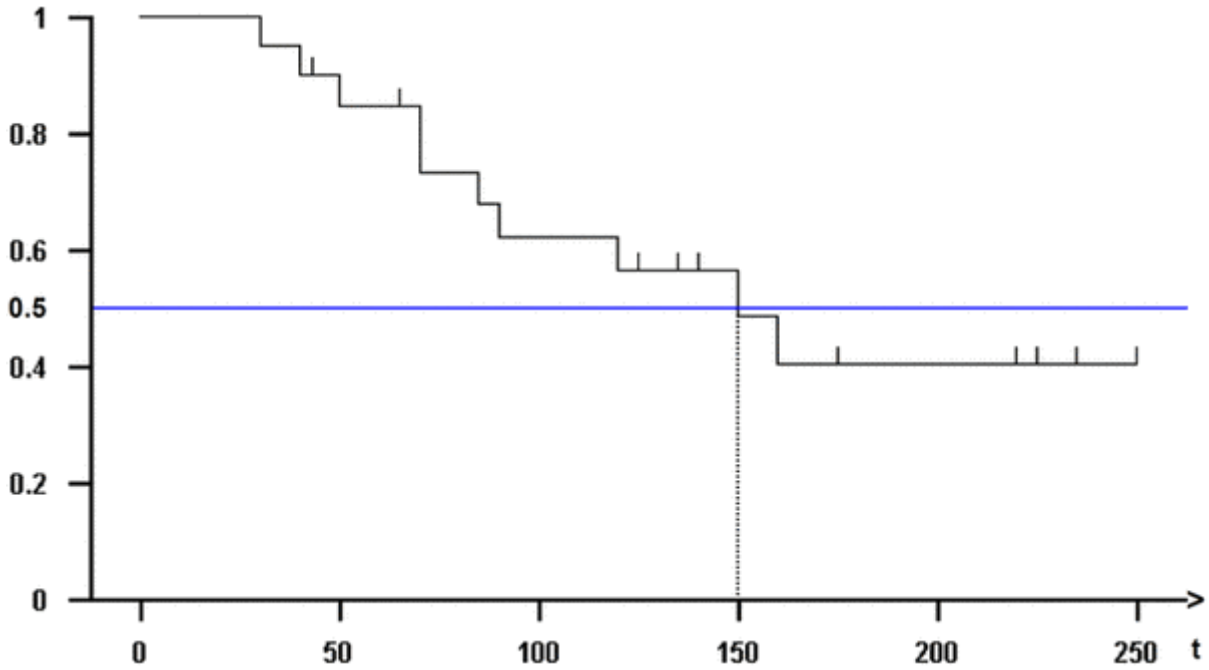
Tabelle 2.1: Rechenschema zum Kaplan-Meier-Schätzer

(1) **i**	**(2)** **Tage** t_i	**(3)** **im Risiko** n_i	**(4)** **Ereignisse** d_i	**(5)** **Anteil Überlebender** $q_i=(n_i-d_i)/n_i$	**(6)** **kumulative Überlebensrate** $q_1 \cdot q_2 \cdot \ldots \cdot q_i$
0	0	20	0	20/20=1	1.0000
1	30	20	1	19/20=0.9500	0.9500
2	40	19	1	18/19=0.9474	0.9000
3	43^+	18	0	18/18=1	0.9000
4	50	17	1	16/17=0.9412	0.8471
5	65^+	16	0	16/16=1	0.8471
6	70	15	2	13/15=0.8667	0.7341
7	85	13	1	12/13=0.9231	0.6776
8	90	12	1	11/12=0.9167	0.6212
9	120	11	1	10/11=0.9091	0.5647
10	125^+	10	0	10/10=1	0.5647
11	135^+	9	0	9/9=1	0.5647
12	140^+	8	0	8/8=1	0.5647
13	150	7	1	6/7=0.8571	0.4840
14	160	6	1	5/6=0.8333	0.4034
15	175^+	5	0	5/5=1	0.4034
16	220^+	4	0	4/4=1	0.4034
17	225^+	3	0	3/3=1	0.4034
18	235^+	2	0	2/2=1	0.4034
19	250^+	1	0	1/1=1	0.4034

Aus Tabelle 2.1 kann man ablesen, dass der empirische Median der Überlebenszeiten $\tilde{x} = 150$ Tage beträgt.

Abbildung 2.1 zeigt die **geschätzte Überlebensrate** $\hat{S}(t)$ in Abhängigkeit von der Überlebenszeit als Treppenfunktion. Es ist üblich, die **Zensierungszeitpunkte** durch einen **senkrechten Strich** zu markieren. Den empirischen Median $\tilde{x} = 150$ Tage kann man am Schnittpunkt der waagerechten Linie mit der Treppenfunktion ablesen.

Abbildung 2.1: Kaplan-Meier-Plot für zensierte Überlebenszeiten



Die bisherigen Auswertungsmethoden beschränkten sich auf die Betrachtung **eines** Merkmals. Will man **gleichzeitig mehrere** Merkmale in die Auswertung einbeziehen, um deren Abhängigkeit zu untersuchen, braucht man andere Methoden.

2.3 Kontingenztafel

Zur Untersuchung der Abhängigkeit zweier qualitativer Merkmale eignet sich die Kontingenztafel, in der die gemeinsame Häufigkeitsverteilung zweier Merkmale tabellarisch dargestellt wird. Sind A und B zwei qualitative Merkmale mit den Ausprägungen $A_1, A_2, ..., A_k$ bzw. $B_1, B_2, ..., B_l$, dann ist die zugehörige Kontingenztafel ein rechteckiges Zahlenschema, das in der i-ten Zeile und der j-ten Spalte die absolute Häufigkeit n_{ij} enthält, mit der die Ausprägungskombination A_iB_j bei den n Beobachtungseinheiten einer Stichprobe beobachtet wurde.

Ergänzt wird das Schema um je zwei Zeilen und zwei Spalten, die die Ausprägung der Merkmale und die **Spaltensummen** bzw. die **Zeilensummen** enthalten.

Tabelle 2.2: Allgemeine Kontingenztafel

B / A	B_1	B_2	…	B_j	...	B_ℓ	Zeilensumme
A_1	n_{11}	n_{12}	…	n_{1j}	…	$n_{1\ell}$	$n_{1.}$
A_2	n_{21}	n_{22}	…	n_{2j}	…	$n_{2\ell}$	$n_{2.}$
⋮	⋮	⋮	…	⋮	…	⋮	⋮
A_i	n_{i1}	n_{i2}	…	n_{ij}	…	$n_{i\ell}$	$n_{i.}$

$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	$n_{k\ell}$	$n_{k.}$
Spaltensumme	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.j}$	$\dots$	$n_{.\ell}$	$n=n_{..}$

Tabelle 2.2 kann man die allgemein üblichen Bezeichnungen entnehmen. Mit den Zeilen- bzw. den Spaltensummen erhält man wieder die absoluten Häufigkeiten des Merkmals *A* bzw. die des Merkmals *B*. In analoger Weise stellt man Kontingenztafeln für diskrete und klassierte stetige Merkmale auf. Häufig werden zusätzlich zu den **absoluten** Häufigkeiten auch die **relativen** Häufigkeiten in die Kontingenztafel eingetragen. Je nach Fragestellung interessiert man sich für die

- **Gesamtprozente**, das sind die relativen Häufigkeiten bezogen auf den Stichprobenumfang *n*
- **Zeilenprozente**, das sind die relativen Häufigkeiten bezogen auf die jeweiligen Zeilensummen. Sie müssen sich in jeder Zeile zu 100 % addieren
- **Spaltenprozente**, das sind die relativen Häufigkeiten bezogen auf die jeweilige Spaltensumme. Sie addieren sich in jeder Spalte zu 100 %.

Beispiel 2.2
Tabelle 2.3 enthält von 20 Patienten einer klinischen Studie die Daten zu den Merkmalen Therapie (TAD/TAD, TAD/HAM), Therapieergebnis (CR=complete Remission, PR=Partial Remission, NR= Non responder, ED=Early Death), Geschlecht und Alter.

Tabelle 2.3: Therapie, Therapieergebnis, Geschlecht und Alter für 20 Patienten

Lfd. Nr.	Therapie	Therapieergebnis	Geschlecht	Alter
1	TAD/TAD	PR	WEIBLICH	19
2	TAD/HAM	ED	MÄNNLICH	55
3	TAD/TAD	NR	WEIBLICH	48
4	TAD/TAD	CR	WEIBLICH	49
5	TAD/HAM	PR	MÄNNLICH	32
6	TAD/HAM	CR	WEIBLICH	22
7	TAD/TAD	CR	WEIBLICH	43
8	TAD/TAD	CR	MÄNNLICH	44
9	TAD/HAM	CR	WEIBLICH	24
10	TAD/TAD	CR	WEIBLICH	36
11	TAD/HAM	ED	MÄNNLICH	38
12	TAD/TAD	CR	MÄNNLICH	55
13	TAD/TAD	CR	WEIBLICH	28
14	TAD/HAM	CR	MÄNNLICH	48
15	TAD/HAM	NR	WEIBLICH	35
16	TAD/TAD	CR	WEIBLICH	43
17	TAD/HAM	CR	WEIBLICH	37
18	TAD/HAM	CR	WEIBLICH	49

19	TAD/TAD	CR	WEIBLICH	36
20	TAD/HAM	ED	WEIBLICH	29

Mit den Angaben aus Tabelle 2.3 erhält man für die beiden qualitativen Merkmale „***Therapie***“ und „***Therapieergebnis***“ die folgende Kontingenztafel.

Tabelle 2.4: Therapie und Therapieergebnis

Therapie	**Ergebnis**				**Zeilensumme**
	CR	**PR**	**NR**	**ED**	
TAD/TAD	8	1	1	0	10
Zeilenprozent	80	10	10	0	100
TAD/HAM	5	1	1	3	10
Zeilenprozent	50	10	10	30	100
Spaltensumme	13	2	2	3	20
Zeilenprozent	65	10	10	15	100

Tabelle 2.5 enthält die Daten für die Merkmale „***Therapie***“ und „***Therapieergebnis***“ von allen 140 Patienten.

Tabelle 2.5: Therapie und Therapieergebnis bei 140 Patienten einer klinischen Studie

Therapie	**Ergebnis**				**Zeilensumme**
	CR	**PR**	**NR**	**ED**	
TAD/TAD	48	5	13	7	73
Zeilenprozent	65.75	6.85	17.81	9.59	100
TAD/HAM	47	3	12	5	67
Zeilenprozent	70.15	4.48	17.91	7.46	100
Spaltensumme	95	8	25	12	140
Zeilenprozent	67.86	5.71	17.86	8.57	100

Bei gleich guten Therapien würde man gleiche Zeilenprozente in beiden Therapiearmen erwarten. Bei der vorliegenden Tabelle müsste man untersuchen, ob die Abweichungen von der Gleichheit so groß sind, dass sie vernünftigerweise nicht mehr durch den Zufall erklärt werden können. Diese Untersuchung ist Gegenstand des Chi-Quadrat-Tests.

2.4 Regression und Korrelation

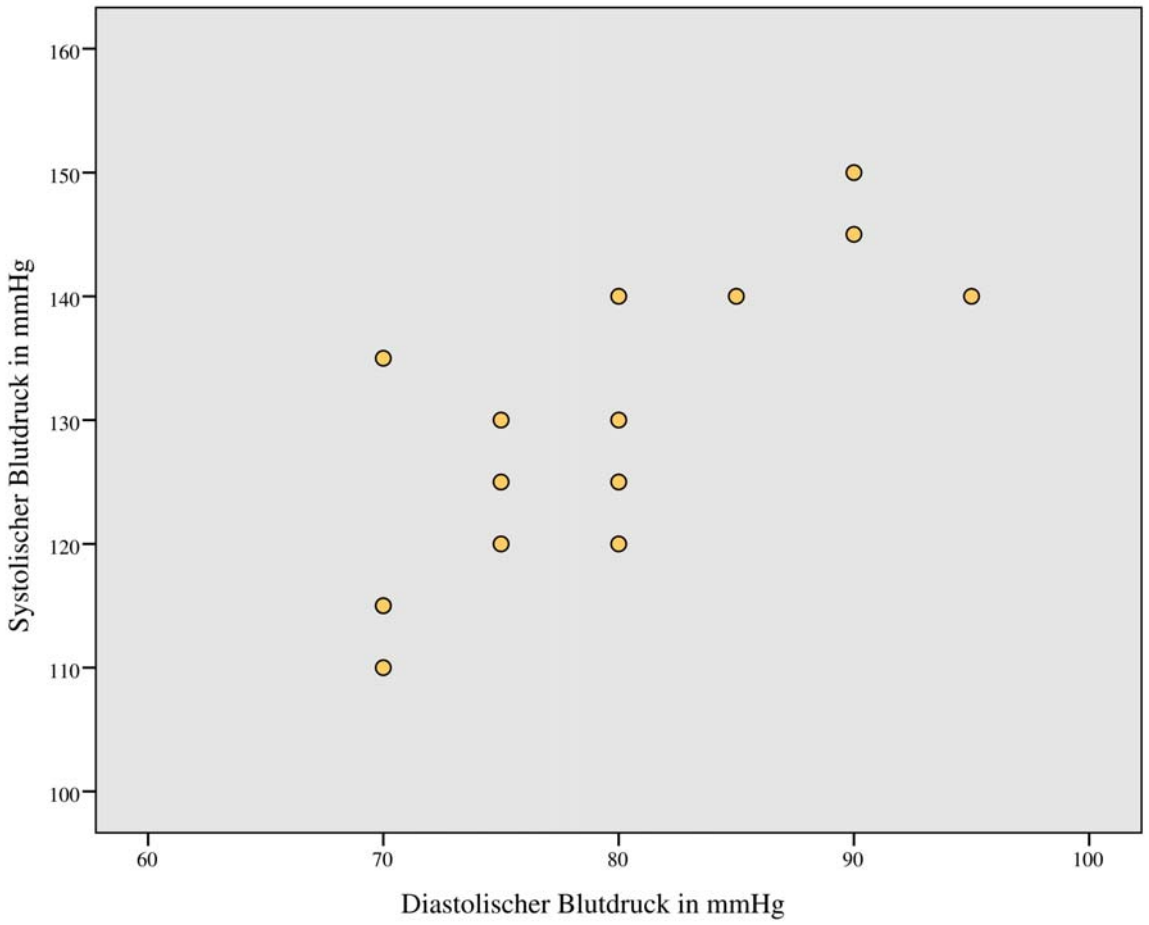
An n Beobachtungseinheiten werden **zwei stetige** Merkmale X und Y beobachtet, die nicht klassiert werden. Es ist ratsam, die Untersuchung der gemeinsamen Verteilung zweier stetiger Merkmale mit der Zeichnung einer **Punktwolke (Scatterplot)** zu beginnen, denn die Punktwolke liefert auf einen Blick Informationen, die für das weitere Vorgehen wichtig sind. Dazu trägt man das Merkmal X an der x-Achse, das Merkmal Y an der y-Achse ab und zeichnet das an der i-ten Beobachtungseinheit festgestellte Wertepaar (x_i,y_i) als Punkt in das Koordinatensystem ein (i=1,2,...,n). Jede Beobachtungseinheit liefert also genau einen Punkt für die Punktwolke.

Beispiel 2.3
Tabelle 2.6 enthält von 15 Patienten die Angaben zum diastolischen und zum systolischen Blutdruck, die in Abbildung 2.2 als Punktwolke dargestellt sind. Der diastolische Blutdruck (RR_{dias}*) ist an der x-Achse, der systolische (*RR_{sys}*) an der y-Achse abgetragen.*

Tabelle 2.6: Diastolischer und systolischer Blutdruck von 15 Patienten

Lfd. Nr.	RRdias	RRsys
1	80	120
2	70	115
3	80	125
4	70	110
5	70	115
6	80	130
7	85	140
8	75	120
9	75	125
10	90	150
11	80	140
12	70	135
13	95	140
14	75	130
15	90	145

Abbildung 2.2: Punktwolke diastolischer und systolischer Blutdruck



Zur Untersuchung der **Abhängigkeit** von zwei oder mehr stetigen Merkmalen dient die Regressionsrechnung. Hier wird nur der Fall der **linearen** Regression für **zwei** Merkmale betrachtet.

$X(=RR_{dias})$ und $Y(=RR_{sys})$ seien die beiden stetigen Merkmale und es soll ***Y* in Abhängigkeit von *X*** untersucht werden. Oft ist aus dem inhaltlichen Zusammenhang nicht unmittelbar klar, ob man Y in Abhängigkeit von X, oder XY in Abhängigkeit von untersuchen soll. Wenn man Y in Abhängigkeit von X untersucht, spricht man von der "Regression von Y auf X", wenn man X in Abhängigkeit von Y untersucht, spricht man von der "Regression von X auf Y".

Zur Veranschaulichung trägt man die Daten als Punktwolke in ein Koordinatensystem ein (Abbildung 2.2). Bei der linearen Regression von Y auf X geht man davon aus, dass zwischen den beiden Merkmalen ein linearer Zusammenhang der Form

$$Y = \beta_0 + \beta_1 X$$

besteht. Die Abweichung der tatsächlich festgestellten Wertepaare von der durch die Gleichung beschriebenen Geraden führt man auf den Einfluss nicht erfasster Störgrößen zurück. Es stellt sich die Aufgabe, β_0 und β_1 vernünftig aus den Daten zu schätzen.

Dieses Problem wurde mathematisch von C. F. Gauß gelöst. Man erhält für β_1 bzw. β_0 die **Schätzwerte** b_1 bzw. b_0, die aus den Daten mit Hilfe der Formeln

$$b_1 = \frac{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

bzw.

$$b_0 = \bar{y} - b_1 \bar{x}$$

berechnet werden.

Die Gerade

$$y = b_0 + b_1 x$$

heißt (empirische) **Regressionsgerade** der Regression von *Y* auf *X*. Der Anstieg der Regressionsgeraden b_1, heißt (empirischer) **Regressionskoeffizient**. Außerdem hat sich für den Zähler des Regressionskoeffizienten

$$s_{xy} = \frac{1}{n-1} S_{xy} \quad \text{mit} \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

die Bezeichnung "(empirische) **Kovarianz** von *X* und *Y*" eingebürgert. Sie wird analog zur empirischen Varianz

$$s_x^2 = \frac{1}{n-1} S_{xx} \quad \text{mit} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}$$

Wer will, kann sich durch Ausmultiplizieren der Quadrate davon überzeugen, dass gilt

$$S_{xy} = \sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}.$$

Mit Hilfe von Varianz und Kovarianz lässt sich die Formel für den Regressionskoeffizienten zu

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

vereinfachen.

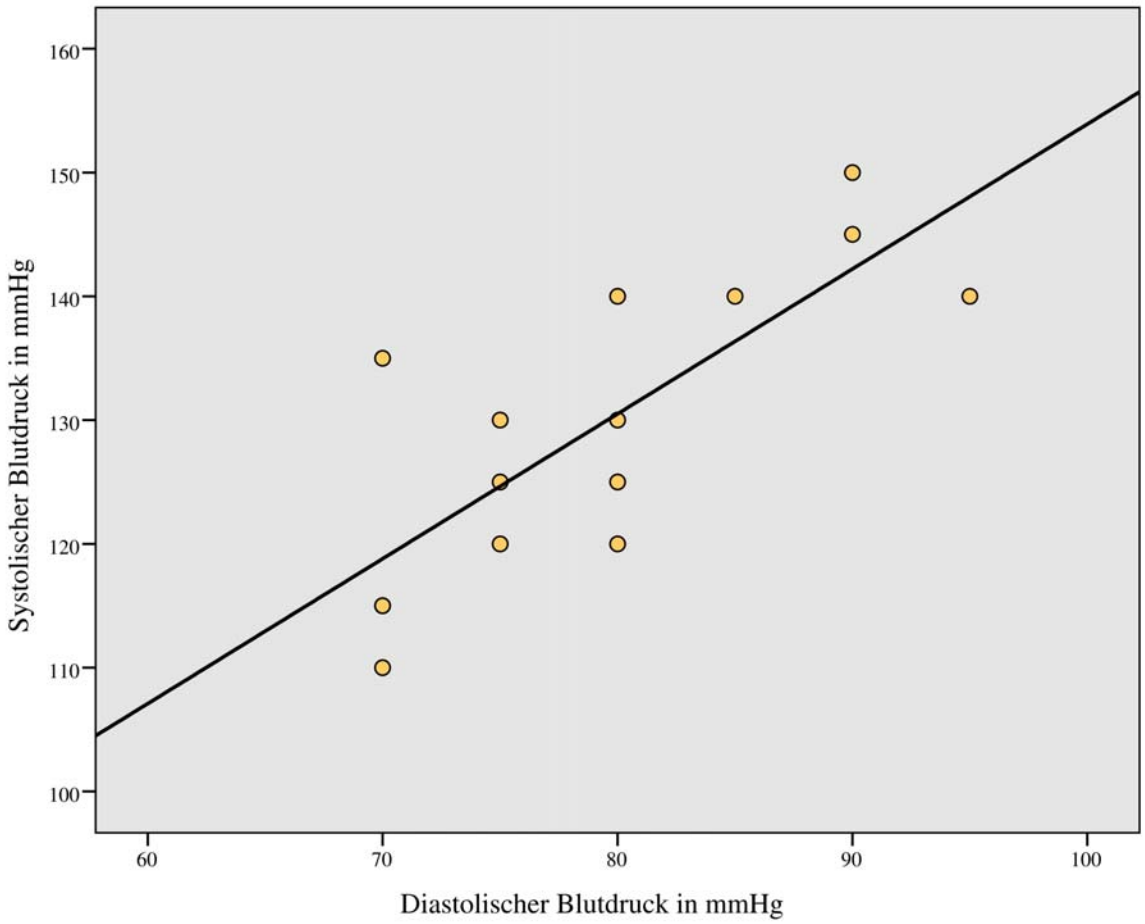
Für die Beispieldaten aus Tabelle 2.6 sind alle für die Regression wichtigen Kenngrößen in Tabelle 2.7 zusammengetragen.

Tabelle 2.7: Regressions- und Korrelationsrechnung

RR_{dias}		RR_{sys}	
(1) $\sum x_i$	1185	(1) $\sum y_i$	1940
(2) $\bar{x} = \sum x_i / n$	79.00	(2) $\bar{y} = \sum y_i / n$	129.33
(3) $\sum x_i^2$	94525	(3) $\sum y_i^2$	252950
(4) $(\sum x_i)^2 / n$	93615	(4) $(\sum y_i)^2 / n$	250907
(5) $S_{xx} = (3) - (4)$	910	(5) $S_{yy} = (3) - (4)$	2043
(6) $s_x^2 = S_{xx} / (n-1)$	65.00	(6) $s_y^2 = S_{yy} / (n-1)$	145.93
(7) $s_x = \sqrt{s_x^2}$	8.0623	(7) $s_y = \sqrt{s_y^2}$	12.0811
(8) $b_1 = S_{xy} / S_{xx} = (13) / (5)$	1.1703	(8) $a_1 = S_{xy} / S_{yy} = (13) / (5)$	0.5212
(9) $b_0 = \bar{y} - b_1 \cdot \bar{x}$	36.8773	(9) $a_0 = \bar{x} - a_1 \cdot \bar{y}$	11.5905
(10) $y = b_0 + b_1 \cdot x$	y=36.88+1.17 x	(10) $x = a_0 + a_1 \cdot y$	x=11.59+0.52y
(11) $\sum x_i \cdot y_i$		154325	
(12) $(\sum x_i) \cdot (\sum y_i) / n$		153260	
(13) $S_{xy} = (11) - (12)$		1065	
(14) $r = S_{xy} / \sqrt{S_{xx} \cdot S_{yy}}$		0.781	

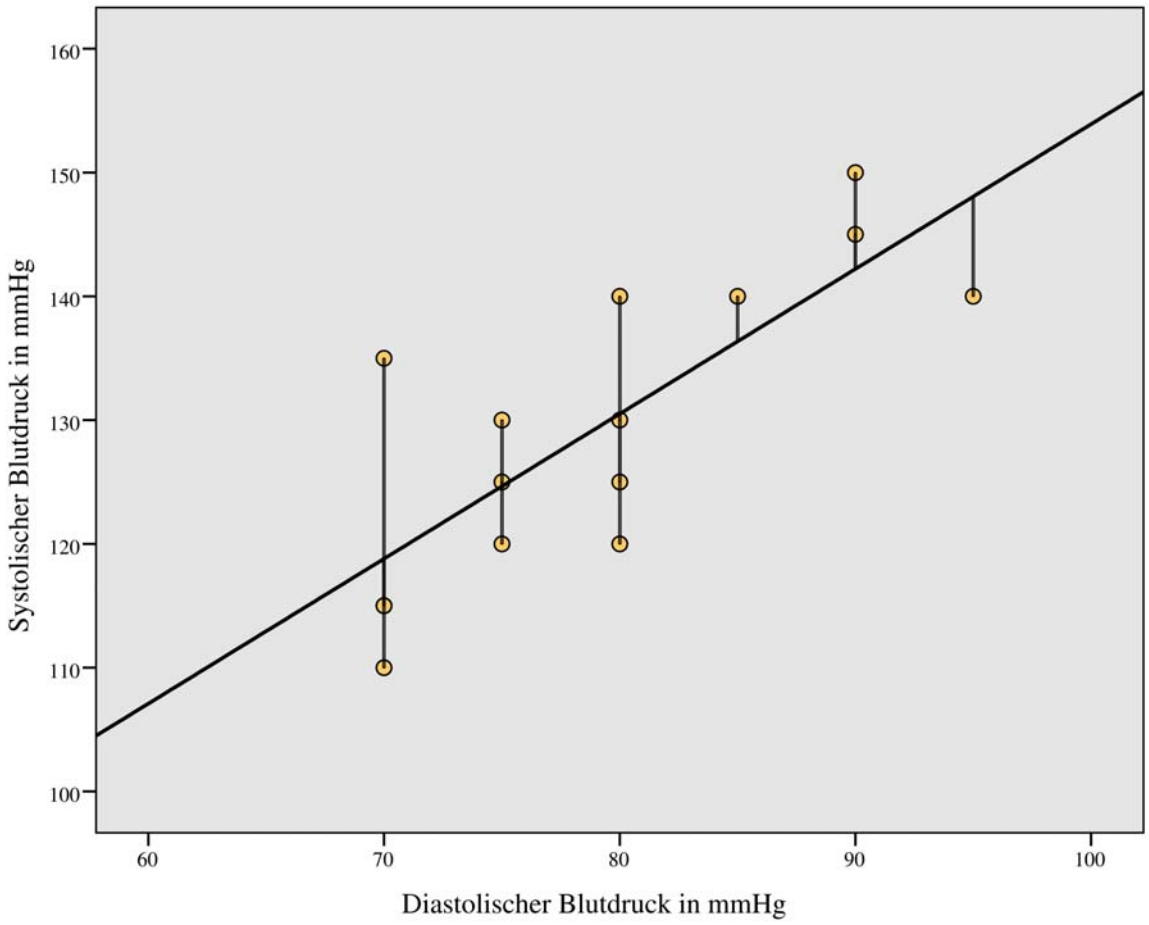
Abbildung 2.3 zeigt noch einmal die Punktwolke mit der berechneten Regressionsgerade.

Abbildung 2.3: Punktwolke und Regressionsgerade $y = 36.88 + 1.17x$



Man kann mathematisch zeigen, dass die so berechnete Regressionsgerade die eindeutig bestimmte Gerade ist, die die **Summe der Abstandsquadrate** der Punkte von der Geraden **minimiert**. Hierbei werden die Abstände parallel zur y-Achse gemessen.

Abbildung 2.4: Schema einer linearen Regression - Methode der kleinsten Quadrate



Nach der Durchführung der Rechnung stellt sich die Frage, wie "gut" die ermittelte Gerade zu den Punkten passt oder - etwas spezifischer – wie viel von der Streuung der *Y*-Werte durch ihre angenommene Abhängigkeit, der Korrelation, von den *X*-Werten erklärt wird.

Eine Maßzahl hierfür ist der (empirische) Korrelationskoeffizient *r*, der durch

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

oder mit den eingeführten Abkürzungen vereinfacht

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

erklärt ist.

Man kann zeigen, dass immer $-1 \leq r \leq +1$ gilt.

Die Grenzfälle r=+1 und r=-1 treten auf, wenn schon alle gemessenen Punkte (x_i,y_i) auf einer Geraden liegen, wobei die Gerade für r=+1 steigt und für r=-1 fällt. Für r=0 verläuft die Gerade parallel zur x-Achse.

r^2, das **Quadrat** des Korrelationskoeffizienten, heißt **Bestimmtheitsmaß**. r^2 lässt sich interpretieren als Anteil der durch die Regression erklärten Streuung der Y-Werte. Hat man z. B. r=0.7 erhalten, dann ist r^2=0.49, d. h., 49 % der Streuung der Y-Werte werden durch die lineare Abhängigkeit von X erklärt. Damit ist r bzw. r^2 das gesuchte Maß. Man darf sich aber nicht zu dem Trugschluss verleiten lassen, dass ein r^2 nahe bei 1 einen linearen Zusammenhang "beweist". Es wird nur ausgesagt, dass ein angenommener linearer Zusammenhang einen großen Anteil der Streuung der Y-Werte erklärt.

Die bisherigen Rechnungen gelten für die Regression von Y auf X, bei der Y das abhängige und X das unabhängige Merkmal ist. Durch Vertauschung der Rollen von X und Y kommt man zur Regression von X auf Y, bei der X das abhängige und Y das unabhängige Merkmal ist. Die Gleichung der Regressionsgeraden sei

$$x = \alpha_0 + \alpha_1 y\,.$$

Ganz analog zu den Rechnungen oben erhält man für α_1 bzw. α_0 die Schätzwerte

$$a_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{S_{xy}}{S_{yy}},$$

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

Die Kenngrößen für die Regression von X auf Y sind ebenfalls in Tabelle 2.7 aufgeführt. Trägt man beide Regressionsgeraden in das gleiche Koordinatensystem ein, erkennt man, dass sich die beiden Geraden im Punkt $(\bar{x}, \bar{y})$ - dem sogenannten Schwerpunkt - schneiden. Der Korrelationskoeffizient r ist symmetrisch in X und Y. Daher erhält man für beide Regressionen das gleiche r.

Für r^2=1 sind beide Regressionsgeraden identisch.

Abbildung 2.5: Punktwolke und Regressionsgeraden y=36.88+1.17x **und** x=11.59+0.52y