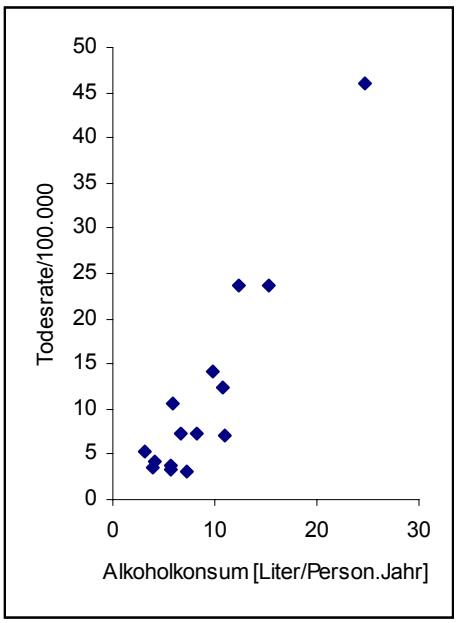
Korrelation und Regression

Untersucht man zwei (oder mehrere) Zufallsvariablen, dann kann man u. U. feststellen, daß zwischen den Zufallsvariablen ein Zusammenhang besteht. Z.B. könnte man erwarten, daß die durchschnittliche Schuhgröße und die Größe von Personen zusammenhängen, d.h. man kann von einer groß gewachsenen Person auch annehmen, daß die Schuhgröße entsprechend groß sein wird. Eine genaue Kenntnis dieses Zusammenhanges ist natürlich sehr vorteilhaft, weil sie gestattet aus der Kenntnis einer Variablen die andere *vorherzusagen.* Der Zusammenhang zwischen 2 Variablen kann natürlich auch negativ sein, z.B. die Zunahme der einen Variablen geht einher mit einer Abnahme der zweiten Variablen. Natürlich können die Variablen auch völlig zusammenhangslos sein; So wird man z.B. zwischen der Populationsdichte von Störchen und der Geburtsrate wahrscheinlich keinen ursächlichen Zusammenhang finden, d.h. aber nicht, daß hohe Populationsdichte mit hoher Storchendichte zufällig koinzidieren können. Zunächst einmal ein Beispiel zur Veranschaulichung:

Tab.1.: Zusammenhang zwischen Alkoholkonsum von Personen über 14 Jahren und Todesrate (Bsp. aus M245 Probability and Statistics, Open University)

Land	Alkoholkonsum [Liter/Jahr.Person]	Todesrate pro 100.000 durch Zirrhose und Alkoholismus
Frankreich	24,7	46,1
Italien	15,2	23,6
West-Deutschland	12,3	23,7
Australien	10,9	7,0
Belgien	10,8	12,3
USA	9,9	14,2
Kanada	8,3	7,4
England und Wales	7,2	3,0
Schweden	6,6	7,2
Japan	5,8	10,6
Niederland	5,7	3,7
Irland	5,6	3,4
Norwegen	4,2	4,3
Finnland	3,9	3,6
Israel	3,1	5,4



Aus dieser Darstellung ist ersichtlich, daß hoher Alkoholkonsum mit erhöhter Sterblichkeitsrate und niedriger Alkoholkonsum mit geringer Sterblichkeitsrate assoziiert ist. Näherungsweise scheint der Zusammenhang linear zu sein.

Nun ist ein quantitative Maß zu finden, wie stark der Zusammenhang zwischen den beiden Variablen ist.

Sind (x_1,y_1), (x_2,y_2),....., (x_i,y_i),....., (x_n,y_n) paarweise Beobachtungen, dann wird durch Subtraktion des Mittelwertes vom jeweiligen Meßwert die Lage im x-y Diagramm so verschoben, daß der Nullpunkt etwa im Zentrum liegt – in Abhängigkeit von der vorliegenden Verteilung, $x' = (x_i - \bar{x})$ und $y' = (y_i - \bar{y})$,

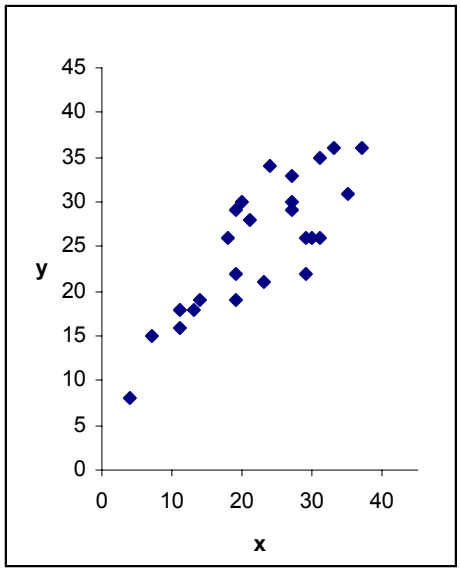
mit den Mittelwerten $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Skalenunabhängigkeit wird weiters erreicht durch Division der Differenz durch die Standardabweichung. Die Standardabweichung der x- und y- Werte ist definiert durch:

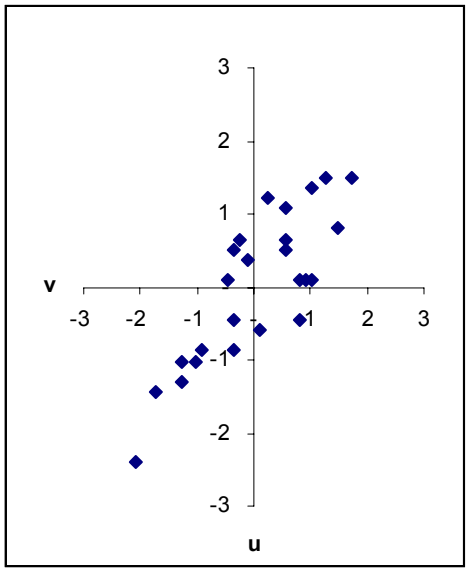
$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{und} \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1}$$

Somit erhalten wir ein Paar standardisierter (vgl. Normalverteilung !) Beobachtungen (u_i,v_i), mit:

$$u_i = \frac{x_i - \bar{x}}{s_x}, \qquad v_i = \frac{y_i - \bar{y}}{s_y} \tag{2}$$



x,y Wertepaare



Standardisierte x,y Werte.
Standardvariable u,v

Aus dem Streudiagramm der standardisierten x und y Werte ist ersichtlich, daß bei positiver Korrelation die meisten Werte im 1. und im 3. Quadranten zu liegen kommen.

Aus der Größe $\sum_{i=1}^{n} u_i v_i$ geht weiters hervor, daß die Punkte im 1. und 3. Quadranten zur Produktsumme die meisten Beiträge liefern, im Vergleich zu den Beiträgen des 2. und 4. Quadranten. Die Produkte der $u_i v_i$ des 1. und 3. Quadranten sind positiv, die des 2. und 4. Quadranten sind negativ, daher wird auch die Gesamtsumme positiv sein. Der umgekehrte Fall tritt auf, wenn die Ausgangsdaten nicht positiv korreliert sind, d.h. wenn eine Zunahme der x-Werte mit einer Abnahme der y-Werte einher geht.
Die bestimmende Größe für einen positiven oder negativen Zusammenhang ist daher: $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, weil die Nenner s_x^2 und s_y^2, nur skalierenden Einfluß haben.

Kovarianz: In Anlehnung an die Definition der Varianz wird daher der Ausdruck

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

als Kovarianz bezeichnet.

Korrelation: Unter Verwendung der standardisierten Werte wird der Ausdruck für r, den **Korrelationskoeffizienten** der Stichprobe:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} u_i v_i$$

$$\text{bzw } r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{oder} \quad r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i{}^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i{}^2 - n\bar{y}^2\right)}} \tag{4}$$

Es kann gezeigt werden, daß: $-1 \le r \le 1$. Für positive Korrelation ist $0 \le r \le 1$, für negative Korrelation ist $-1 \le r \le 0$. Je näher r bei +1 oder –1 liegt, desto besser ist die Korrelation. Für eine Gerade ist r = 1 oder r = -1, je nachdem ob die Gerade positive oder negative Steigung hat.

Lineare Regression

Wie ist nun durch die Datenpunkte eine Gerade zu legen, die „am besten" paßt? Um eine Gerade der Form $y = \hat{\alpha} + \hat{\beta}x$ zu finden, benötigen wir die „Maximum Likelihood Schätzer" (MLS) für α und β, d.h. wir müssen Werte für die Parameter α und β finden, so daß die Summe der Abweichungsquadrate $\sum_{i=1}^{n}(y_i - f(x_i))^2$ ein Minimum wird. Diese Prinzip heißt „Methode der kleinsten Fehlerquadrate".

Die MLS $\hat{\alpha}$ und $\hat{\beta}$ (Bezeichnung: „Alpha Dach und Beta Dach") können durch Lösung der Ableitung von R nach α und β gefunden werden:

$$R_{\min} = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \tag{5}$$

$$\frac{dR}{d\alpha} = 0, \qquad \frac{dR}{d\beta} = 0 \tag{6}$$

D.h. $\hat{\alpha}$ und $\hat{\beta}$ erfüllen die Gleichungen:

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \tag{7}$$

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = 0 \tag{8}$$

Daher ist:

$$\sum_{i=1}^{n} y_i - \hat{\alpha}\sum_{i=1}^{n} 1 - \hat{\beta}\sum_{i=1}^{n} x_i = 0 \;\Rightarrow\; n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{9}$$

und

$$\hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{10}$$

Verwenden wir nun die Definition de Mittelwertes, dann können wir $n\bar{x} = \sum_{i=1}^{n} x_i$ berechnen. Durch Multiplikation von (9) mit $\bar{x}$, kann diese Gleichung umgeschrieben werden:

$$\hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta} n\bar{x}^2 = \bar{x}\sum_{i=1}^{n} y_i \tag{11}$$

Durch Subtraktion der Gleichung (11) von Gleichung (10) wird $\hat{a}$ eliminiert und man erhält:

$$\begin{aligned}\hat{\beta}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) &= x_1 y_1 + x_2 y_2 + ... + x_n y_n - \bar{x}(y_1 + y_2 + ... + y_n) \\ &= (x_1 - \bar{x})y_1 + (x_2 - \bar{x})y_2 + ... + (x_n - \bar{x})y_n \\ &= \sum_{i=1}^{n}(x_i - \bar{x})y_i\end{aligned} \tag{12}$$

Nun ist aus der Berechnung der Standardabweichung bekannt, daß:

$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$, daher ist der MLS für $\hat{\beta}$ definiert durch:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

und nach Gleichung (7) ist dann $\hat{\alpha}$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Überprüfung des Modells

Das lineare Modell ermöglicht eine Reihe von Fehlinterpretationen, die eine Überprüfung der Annahme erfordern. Ob die Steigung β von einer vorgegebenen Steigung β_0 abweicht, kann mit (a) ANOVA oder (b) mittels t-Test überprüft werden.

(a) Überprüfung mit ANOVA

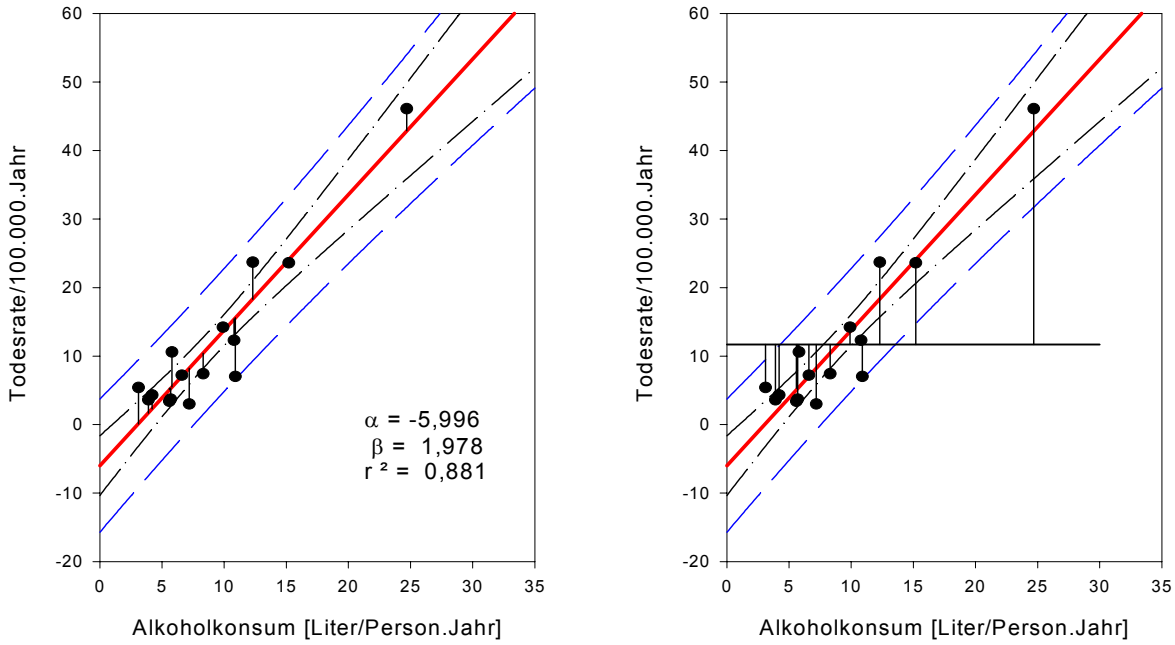
In der unten stehenden Abbildung sind die Abweichungen der einzelnen Datenpunkte (y_i) von der linearen Regressionslinie (yi –f(x_i)) und die Abweichungen der Datenpunkte (y_i) vom Mittelwert $(\bar{y}_i - f(x_i))$ dargestellt. Die Gesamtvarianz ergibt sich aus den Abweichungen der Meßpunkte vom Mittelwert, die aus der Regression resultierende Varianz ergibt sich aus den Abweichungen vom Funktionswert der linearen Regression. Ist die Regression „gut“, d.h. die Gerade liegt nahe den Datenpunkten, dann ist diese Varianz klein, im Extremfall wird sie Null.

Das Modell für den linearen Zusammenhang ist: $y = \hat{\alpha} + \hat{\beta}x + \text{Restfehler}$, bzw in dem angeführten Beispiel:

Todesrate (y) = Todesrate ohne Alkoholkonsum ($\hat{\alpha}$) + ($\hat{\beta}$)*Alkoholkonsum (x) + Restfehler.

Die Regressionsgleichung lautet: **y =– 5,996 + 1,978.x**

Der Durchschnitt der Todesrate (y) beträgt 11,7



Abweichung der Meßwerte y_i von der linearen Regression $f_{(xi)}$ (links), und Abweichung der Meßwerte y_i vom Mittelwert $\bar{y}$ (rechts). (Konfidenzintervalle: schwarz strich-punktiert; „Prediction“ Intervalle: blau strichliert)

Zur Überprüfung des linearen Modells sind die Gesamtvarianz und die Fehlervarianz (Restfehler), das ist der Anteil der Varianz, der der linearen Beziehung zwischen x und y zuzuordnen ist, zu bestimmen. Die Fehlervarianz beträgt in diesem Beispiel Fall 226,1.

Land	X: Alkoholkonsum [Liter/Jahr.Person]	Y: Todesrate pro 100.000 durch Zirrhose und Alkoholismus	y-f(y)	$(y-f(y))^2$
Frankreich	24,7	46,1	3,24	10,51
Italien	15,2	23,6	-0,47	0,22
West-Deutschland	12,3	23,7	5,37	28,81
Australien	10,9	7,0	-8,56	73,33
Belgien	10,8	12,3	-3,07	9,40
USA	9,9	14,2	0,61	0,38
Kanada	8,3	7,4	-3,02	9,13
England und Wales	7,2	3,0	-5,25	27,51
Schweden	6,6	7,2	0,14	0,02
Japan	5,8	10,6	5,12	26,25
Niederland	5,7	3,7	-1,58	2,49
Irland	5,6	3,4	-1,68	2,82
Norwegen	4,2	4,3	1,99	3,95
Finnland	3,9	3,6	1,88	3,54
Israel	3,1	5,4	5,26	27,71
				Σ = **226,08**

Jetzt können wir eine Varianzanalyse durchführen, in dem wir eine Zerlegung der Varianzen vornehmen. Variationsquellen sind hier die lineare Regression und der Restfehler, der aus der Abweichung von der Regressionsgeraden resultiert. Beide Variationen zusammen ergeben die totale Variation. Die Summe der Abweichungsquadrate (SQ) läßt sich demnach folgendermaßen zerlegen:

$$SQ_{tot} = SQ_{linReg} + SQ_{Restfehler}$$

$$SQ_{tot} = \sum_i (y_i - \bar{y})^2$$

$$SQ_{linreg} = \sum_i (f(y_i) - \bar{y})^2$$

$$SQ_{Restfehler} = \sum_i (f(y_i) - y_i)^2$$

SQ....Quadratsumme

FG.....Freiheitsgrad

MQ....Mittlere Quadratsumme, SQ/FG

Variationsquelle	FG	SQ	MQ	F	P
Lineare Regression*	1 (für lin.Reg)	1680,2	1680,2	96,6	< 0.001
Restfehler	13 (n-2)	226,1	17,4		
Total	14 (n-1)	1906,3	136,2		

* Der Freiheitsgrad ist hier die Anzahl der freien Parameter –1. Es gibt 2 freie Parameter: α und β

Der p-Wert für $F_{1,13}$ ist extrem klein; daraus können wir schließen, daß die Veränderungen von X und Y signifikant miteinander verknüpft sind. Die Steigungskonstante $\hat{\beta}$ ist daher verschieden von 0.

<u>Überprüfung mittels t-Test</u>

Mit diesem Test kann überprüft werden, ob die Steigung von einer vorgegebenen Steigung abweicht.

$$t = \frac{\left|\hat{\beta} - \beta_0\right|}{\sqrt{\frac{MQ_{Restfehler}}{SQ_x}}}$$

$MQ_{Restfehler} = 17,4$

SQ_x = Summe der quadrierten Abweichungen von X : $\sum_i (x_i - \bar{x})^2$

In diesem Fall ist $SQ_x = 429,5$

Für die Überprüfung, ob überhaupt eine Steigung bzw. Korrelation vorhanden ist, d.h. für $\beta_0 = 0$ erhalten wir: $t = 1,978/\sqrt{(17,4/429,5)} = 9,83$. Für ein $\alpha = 0,01$ und einen Freiheitsgrad von (n-2) = 13 liegt der kritische Wert bei 3,0 und damit weit unter der Prüfgröße von 9,83. Die Hypothese H_0: $\beta = \beta_0$ wird daher verworfen.

Konfidenzintervalle

Konfidenzintervalle können durch Umformen der Gleichung für den t-Wert angegeben werden:

$$\hat{\beta} \pm t_{\nu,\alpha} \cdot \sqrt{\frac{MQ_{Restfehler}}{SQ_x}}$$

Das Konfidenzintervall des Wertes $f_{(xi)}$ wird wie folgt berechnet:

$$KI_{f(x_i)} = f(x_i) \pm t_{\nu,\alpha} \sqrt{MQ_{\mathrm{Re}\,stfehler} \cdot \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SQ_x}\right)}$$

Mit zunehmender Entfernung von $\bar{x}$ wächst der Ausdruck in der Wurzel durch die Zunahme der Differenz $(x_i - \bar{x})$. Das Konfidenzintervall wird daher zwangsläufig größer, je weiter entfernt die Datenpunkte von $\bar{x}$ sind. Diese Aufweitung ist natürlich sinnvoll, weil für die Randbereiche, in denen keine oder nur mehr wenige Meßwerte mehr vorhanden sind, zuverlässige Informationen fehlen. Aus der Verbindung der Konfidenzinteralle für alle x_i entstehen die Konfidenzlinien oder der Konfidenzbereich, innerhalb dessen $(1-\alpha)$% aller Beobachtungswerte zu erwarten sind. Mithilfe des Konfidenzintervalles kann auch überprüft werden, ob der Achsenabschnitt α signifikant von Null verschieden ist. Liegt 0 außerhalb dieses Intervalls, dann kann angenommen werden, daß α nicht Null ist.