Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Korrelation und Regression

- Einführung
- Streudiagramm
- Kovarianz
- Korrelation
- Regression
- Probleme

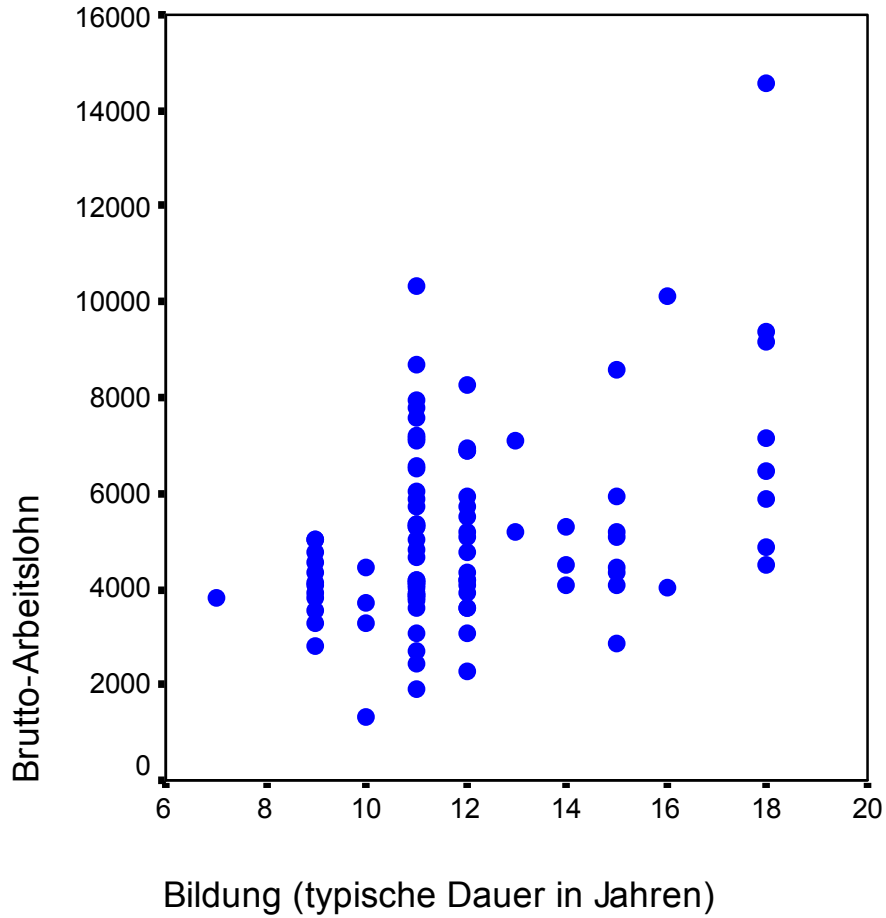
Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Einführung

- Mit Korrelations- und Regressionsanalyse werden Zusammenhänge zwischen zwei metrischen Variablen analysiert.
- Wenn man nur einen Zusammenhang quantifizieren will, aber keine Ursache-Wirkungsbeziehung angenommen werden kann, wird ein Korrelationskoeffizient berechnet.
- Geht man von einer Ursache-Wirkungsbeziehung aus, kann man mit Hilfe der Regressionsanalyse versuchen, die Abhängigkeit des einen Merkmals (Y) vom anderen Merkmal (X) als linearen Zusammenhang durch eine Gleichung auszudrücken

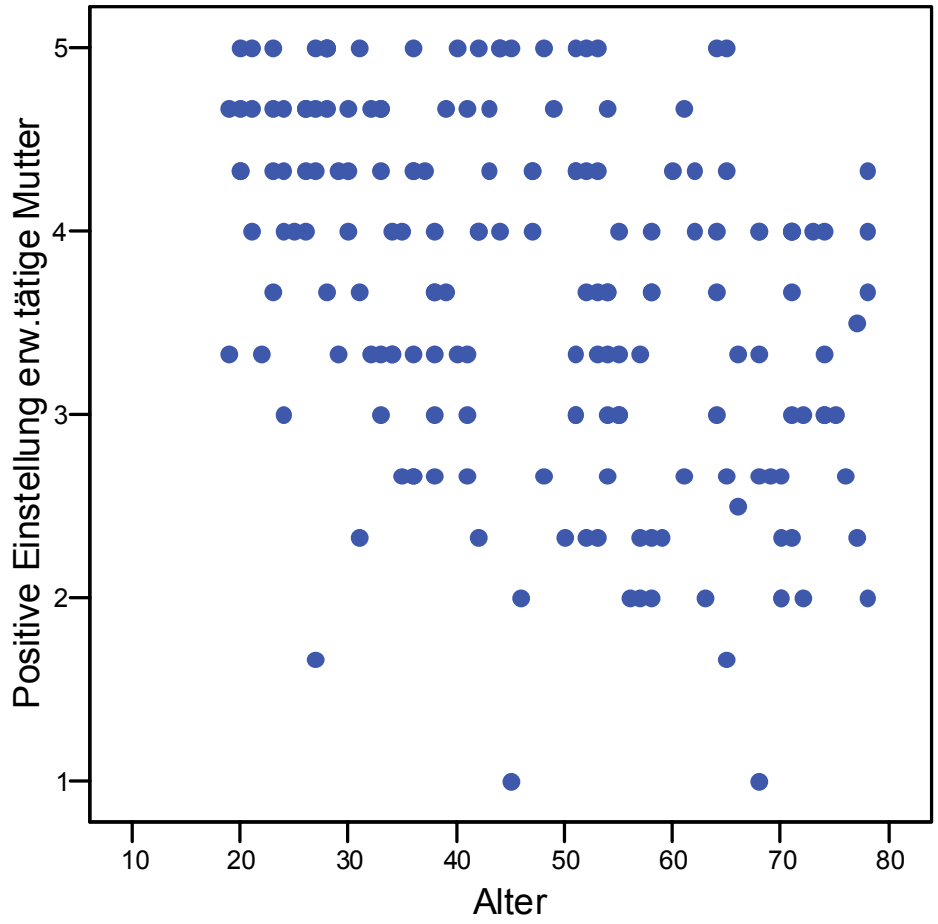
Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Veranschaulichung: Streudiagramm



Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Veranschaulichung: Streudiagramm



Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Kovarianz: „Vorstufe“ der Korrelation

Positiver Zusammenhang:
Hohe Werte in der einen Variablen treten tendenziell gemeinsam mit hohen Werten in der anderen Variablen auf.

Negativer Zusammenhang:
Hohe Werte in der einen Variablen treten tendenziell gemeinsam mit niedrigen Werten in der anderen Variablen auf.

Die „gemeinsame Varianz“ (im Sinne von: „miteinander Variieren“) zweier Variablen.

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

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Kovarianz

Bsp. (fiktiv): Bruttolohn und Bildungsjahre

	X_i	Y_i	$X_i - X_{quer}$	$y_i - y_{quer}$	$X_i - X_{quer} * y_i - y_{quer}$
	2000	9	-1000	-3	3000
	5000	16	2000	4	8000
	4000	16	1000	4	4000
	1500	9	-1500	-3	4500
	2500	10	-500	-2	1000
Summe	15000	60			20500
Ar. Mittel	3000	12	(Gültig für Stichpr.)	Kovar.:	4100

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Kovarianz

Bsp. (fiktiv): Bruttolohn und Körpergroße

	x_i	y_i	$x_i - x_{quer}$	$y_i - y_{quer}$	$x_i - x_{quer} * y_i - y_{quer}$
	2000	1,55	-1000	-0,17	170
	5000	1,65	2000	-0,07	-140
	4000	1,80	1000	0,08	80
	1500	1,75	-1500	0,03	-45
	2500	1,85	-500	0,13	-65
Summe	15000	8,6			0
Ar. Mittel	3000	1,72		Kovar.:	0

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Produkt-Moment-Korrelation, Pearsons Korrelation(skoeffizient)

„Standardisierung“ der Kovarianz:

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

r kann Werte zwischen –1 (perfekter negativer Zusammenhang) und +1 (perfekter positiver Zusammenhang) annehmen.

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

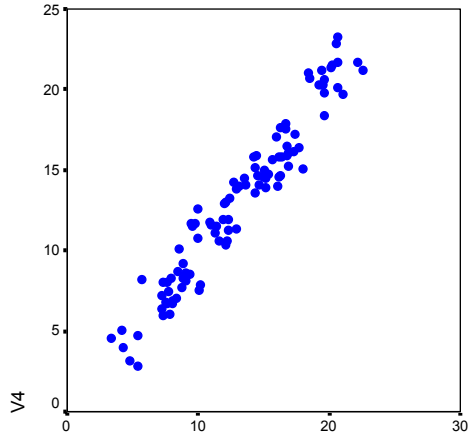
Korrelation

Bsp. (fiktiv): Bruttolohn und Bildungsjahre

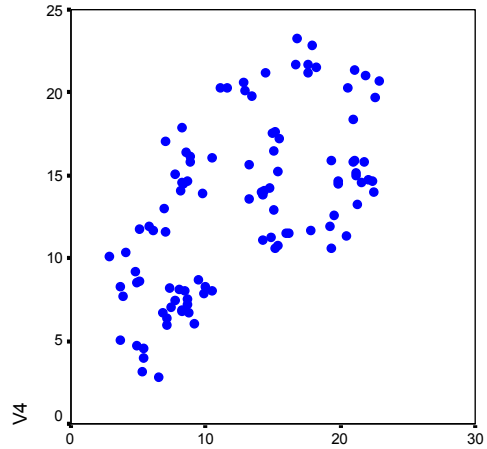
	x_i	y_i	$x_i - x_{quer}$	$y_i - y_{quer}$	$x_i - x_{quer} * y_i - y_{quer}$
	2000	9	-1000	-3	3000
	5000	16	2000	4	8000
	4000	16	1000	4	4000
	1500	9	-1500	-3	4500
	2500	10	-500	-2	1000
Summe	15000	60			20500
Ar. Mittel	3000	12		Kovar.:	4100
				r:	0,96

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

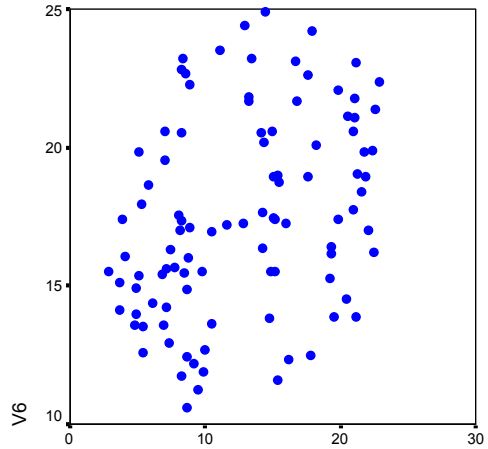
Korrelation



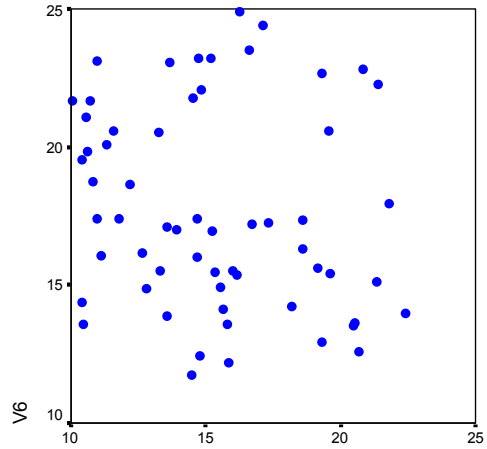
r=0,97



r=0,60



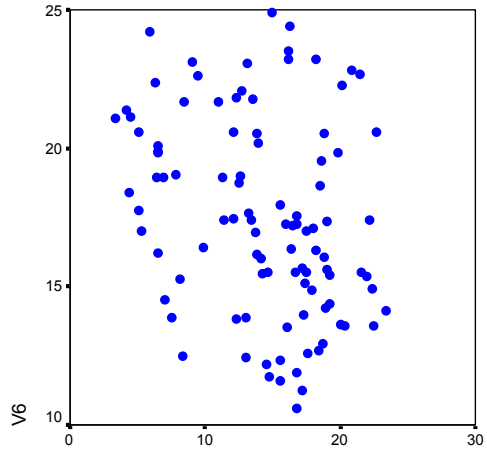
r=0,33



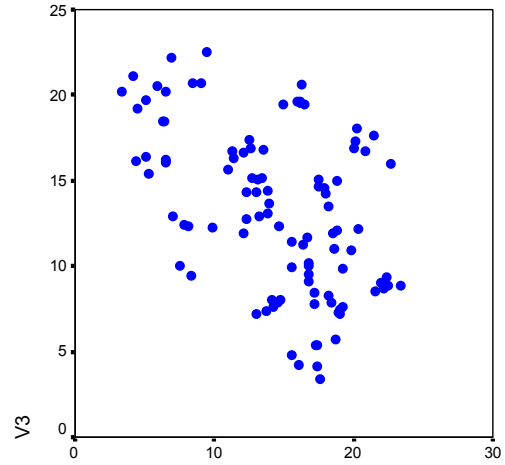
r=0,04

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

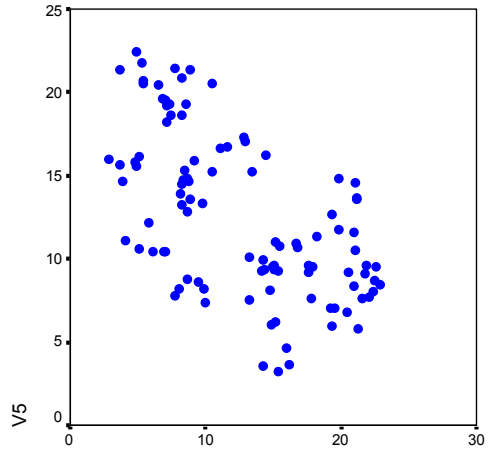
Korrelation



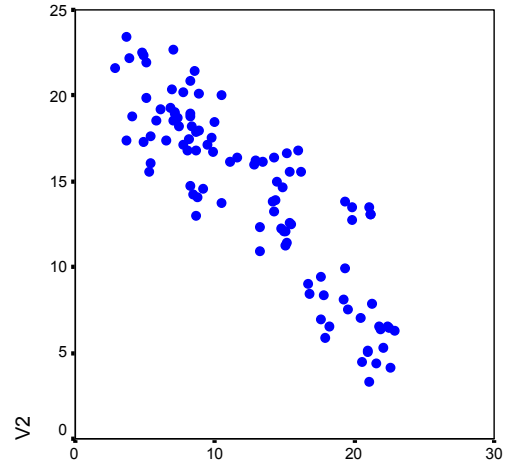
r= –0,26



r= –0,49



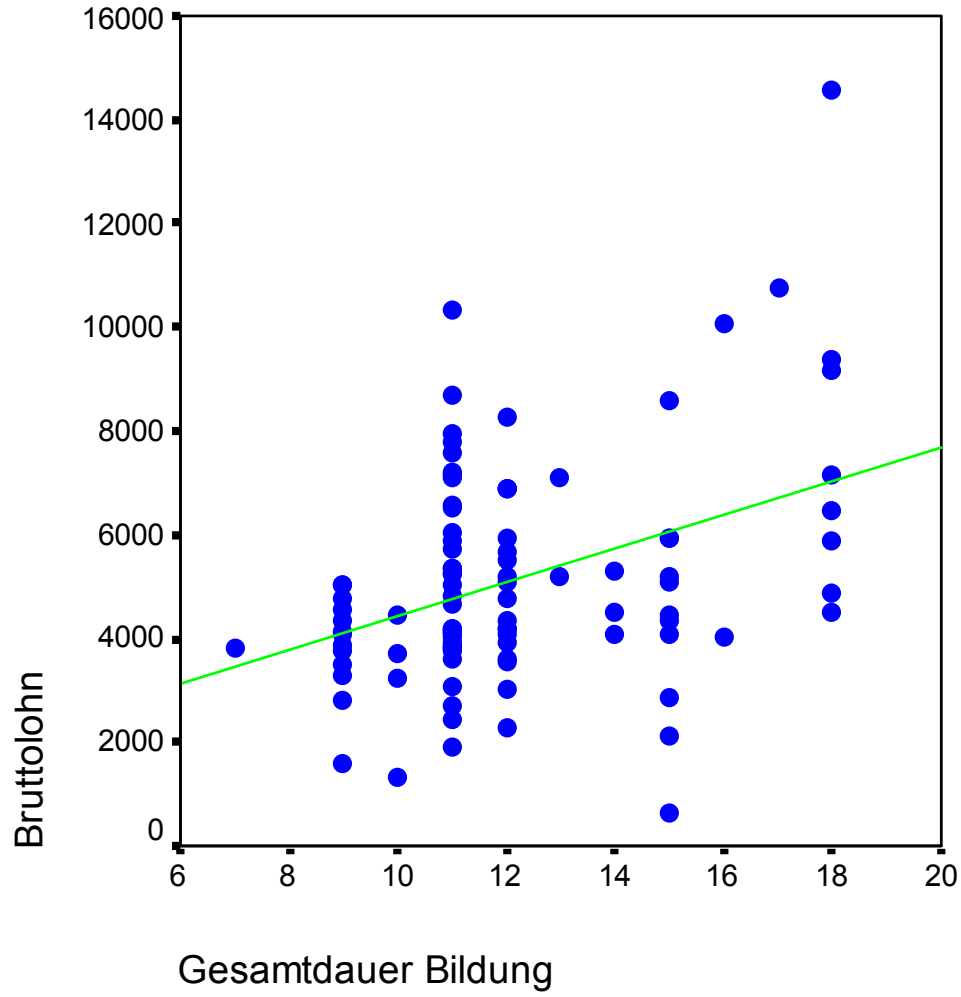
r= –0,60



r= –0,87

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Regression: Die Grundidee visuell



Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Grundidee verbal

- Wir suchen eine Gerade, die den Zusammenhang zwischen den beiden Merkmalen „möglichst gut“ beschreibt.
- „Möglichst gut“ heißt: Die gesuchte Gerade soll so gewählt werden, dass die Abstände zwischen der Geraden und den empirisch beobachteten Werten minimiert werden.

(Aus bestimmten Gründen werden die quadrierten Abstände minimiert.)

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Grundidee formal

Wir schätzen eine Gleichung

$\hat{y}_i = a + bx_i$, so dass gilt

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = Minimum$$

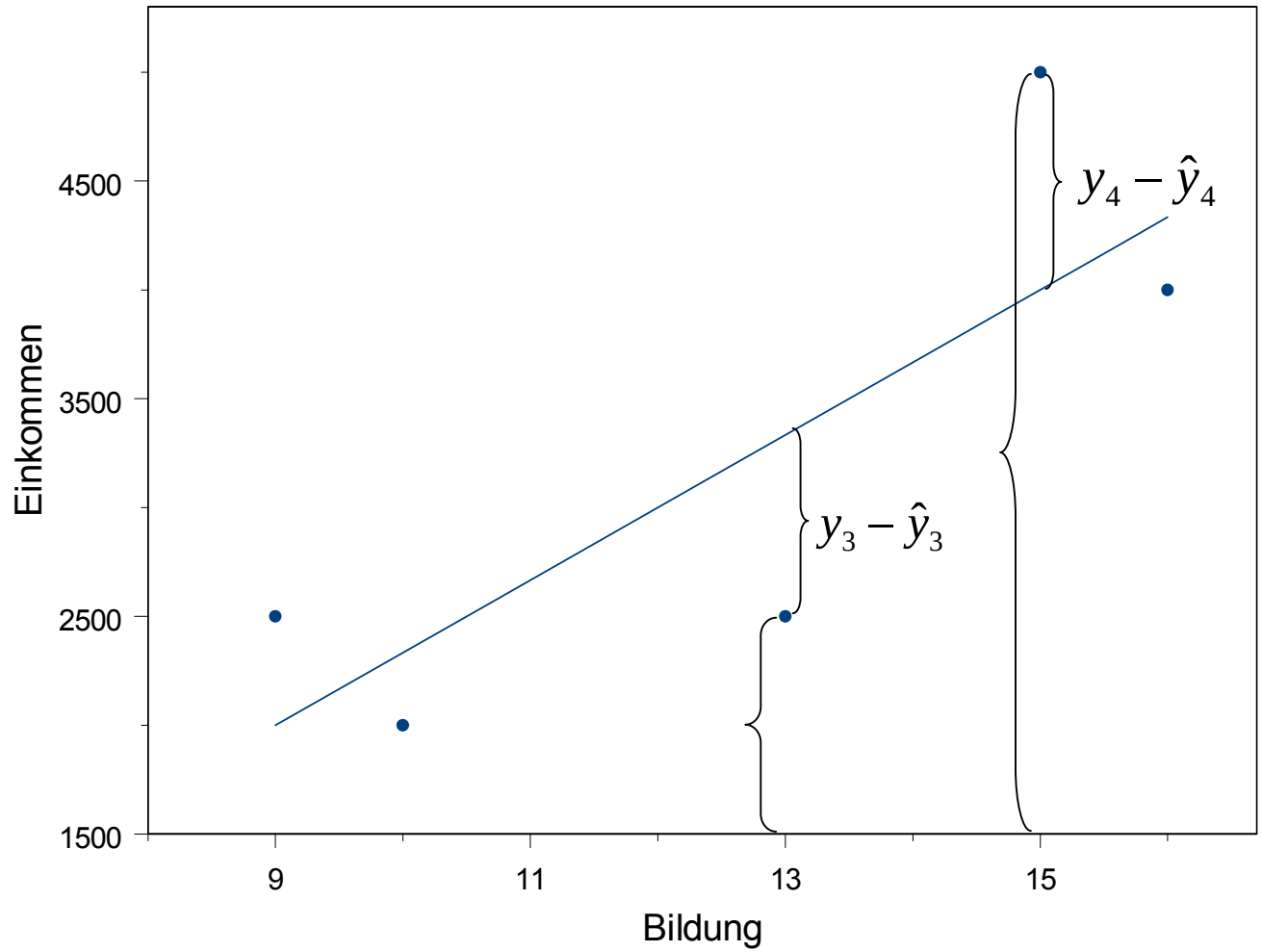
Es werden also die quadrierten Abweichungen der Geraden von den beobachten Werten minimiert („Methode der Kleinsten Quadrate“; englisch: Ordinary Least Squares = OLS).

a und b sind die aus der Stichprobe berechneten Schätzer für die unbekannten Parameter der Grundgesamtheit, meist als α und β bezeichnet.

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Grundidee graphisch

Bitte beachten: Y- und X-Achse müssen bis 0 verlängert gedacht werden



Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Regressionsgleichung

Im Beispiel (vorherige Seite!)

$$\hat{y}_i = -1000 + 333,\overline{3}\,x_i$$

bzw.

$$y_i = -1000 + 333,\overline{3}\,x_i + e_i$$

Die Berechnung von a und b überlassen wir der einschlägigen Statistik-Software.

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Multiple Regression

$$\hat{Y} = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2$$

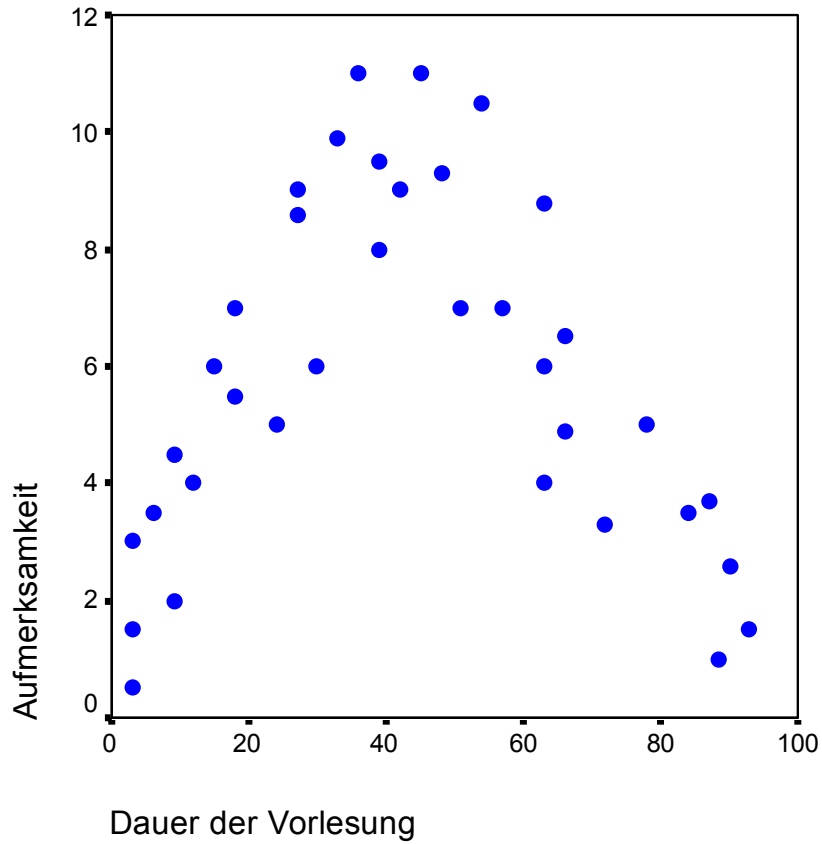
$$\hat{Y} = -750 + 232 \cdot X_1 + 63,5 \cdot X_2$$

Die Regressionskoeffizienten (oder Regressionsgewichte) b_1 und b_2 geben jeweils den um die andere Variable „bereinigten“ Einfluss wieder; man sagt, der Einfluss der anderen Variablen wurde „auspartialisiert“.

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Probleme bei Korrelation und Regression

Nur lineare Zusammenhänge werden erfasst



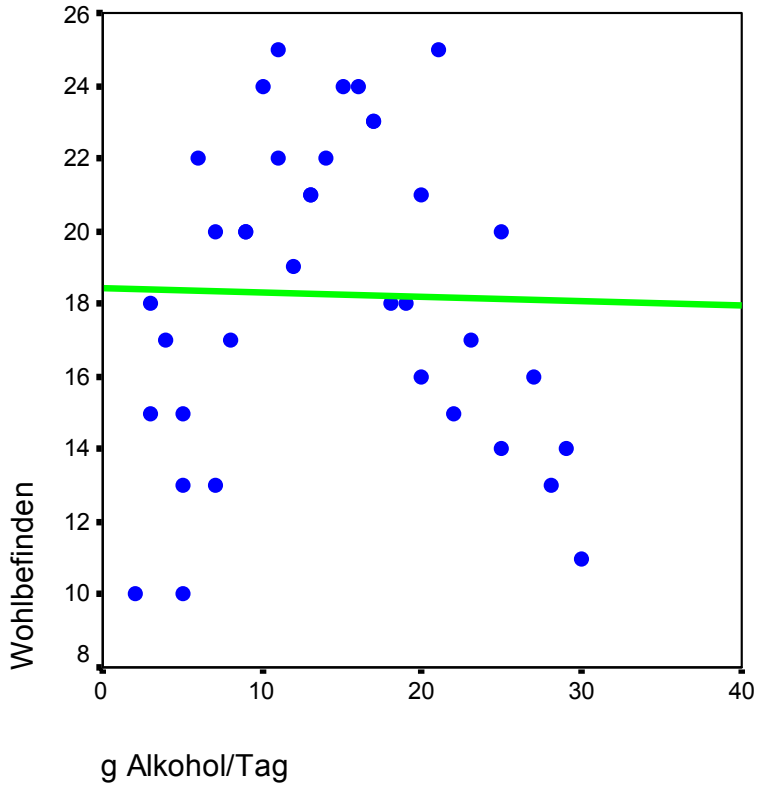
Korrelation: -0,05, d.h. praktisch gleich null.

Das Beispiel ist fiktiv, Sie sind natürlich ständig aufmerksam! ☺

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Probleme bei Korrelation und Regression

Nur lineare Zusammenhänge werden erfasst



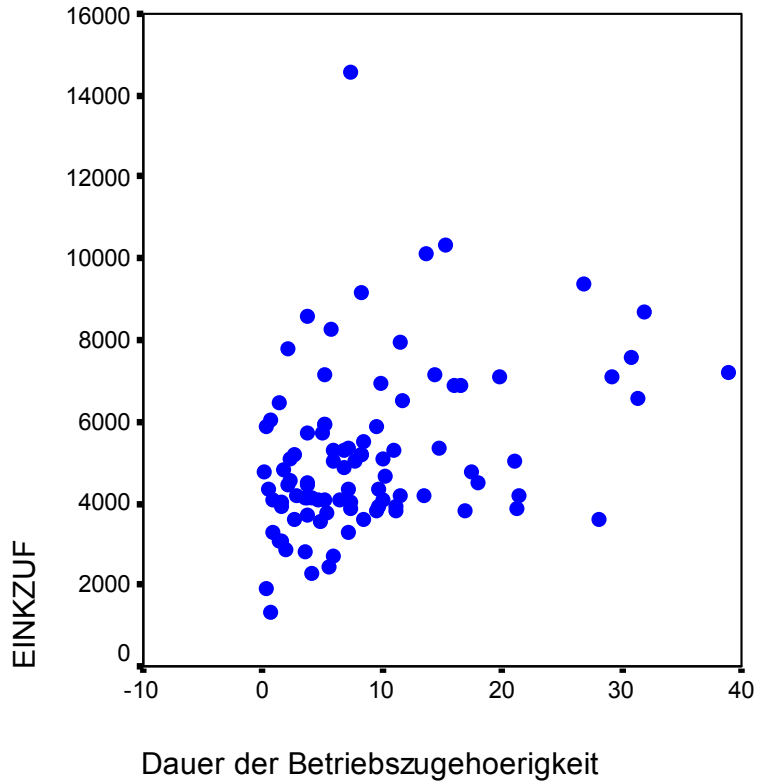
Die Gerade ist quasi horizontal – was nicht dem „eigentlichen“ Zusammenhang entspricht.

In der Regressions-analyse kann man sich aber durch „Tricks“ an die Kurve annähern

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Probleme bei Korrelation und Regression

Einzelne Fälle können starken Einfluss ausüben (nicht zuletzt wegen Multiplikation)



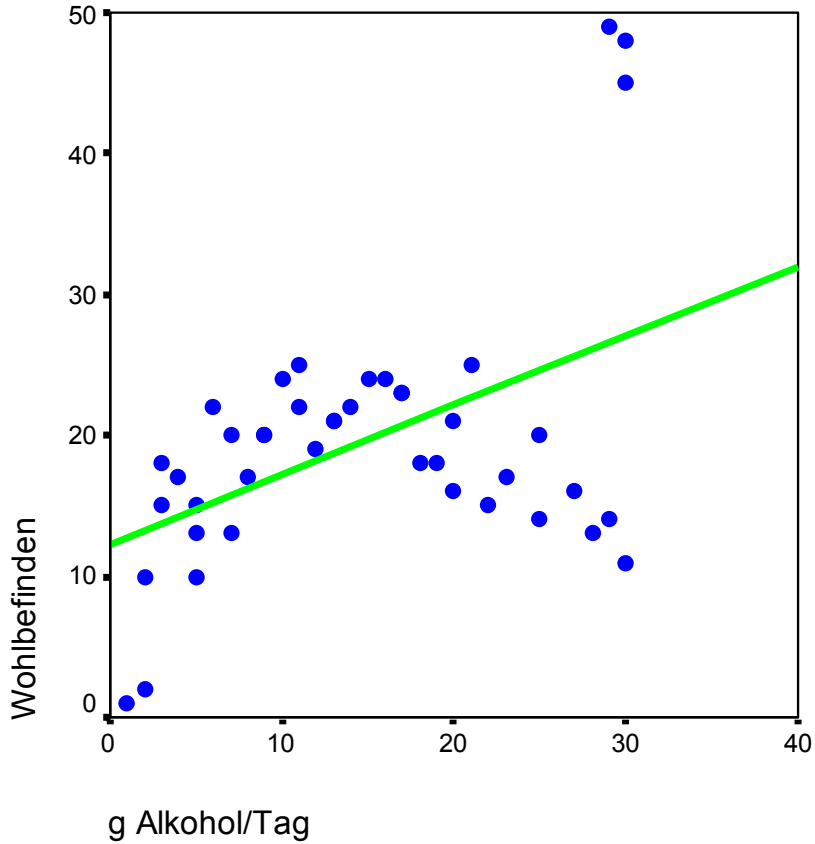
Korrelation über alle Fälle: r=0,35.

Korrelation ohne Einkommen über 14.000: r=0,39.

Einführung
Streudiagramm
Kovarianz
Korrelation
Regression
Probleme

Probleme bei Korrelation und Regression

Einzelne Fälle können starken Einfluss ausüben (nicht zuletzt wegen Quadrierung)



Die gleichen Daten wie vorhin plus einige Extremwerte (links unten, rechts oben) erzeugen einen deutlich steigende Gerade