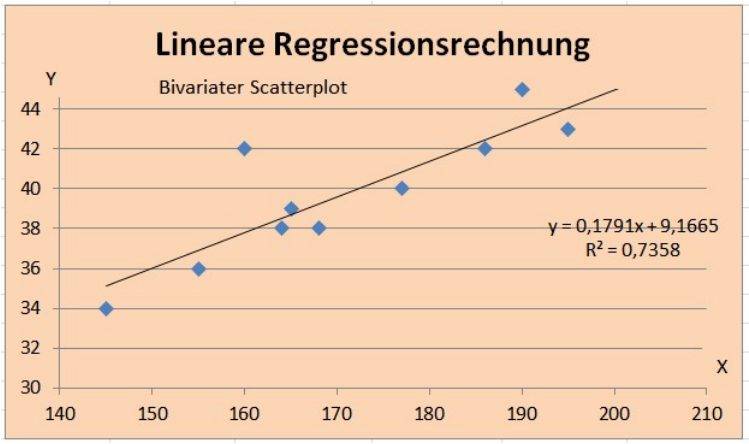
Fragestellung der Regressionsanalyse

Die Regressionsanalyse stellt ein Modell dar, um die Auswirkungen einer oder mehrerer unabhängigen Variablen (hier: Regressoren) auf eine abhängige Variable (hier: Regressand) zu messen oder/und vorherzusagen. Sie ist somit eine Weiterentwicklung der Korrelationsforschung, die sich lediglich auf die Ermittlung von Zusammenhangsmaßen von Variablen beschränkt. Mit der Regression wird versucht, ein oder mehrere Merkmale (UV) zur Erklärung eines anderen Merkmals (AV) heranzuziehen.

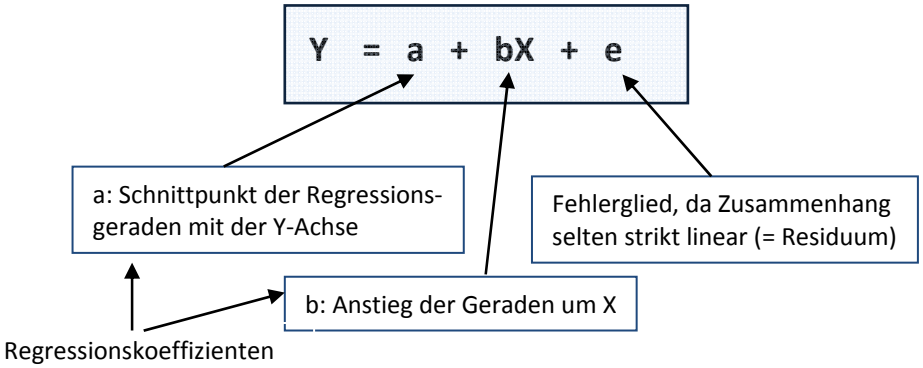
Funktionsprinzip im bivariaten Fall (Beispiel 1)

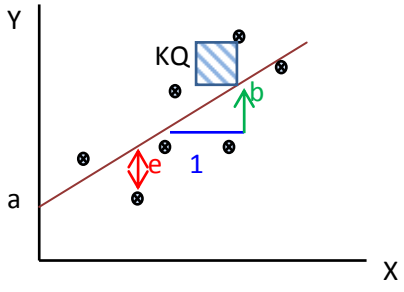
Körpergröße X = UV	Schuhgröße Y = AV
165	39
177	40
168	38
190	45
164	38
160	42
155	36
145	34
195	43
186	42



Annahme eines funktionalen (linearen) Zusammenhangs ***Y = f(X) + e***
Formale Darstellung (Geradengleichung): **Y = a + bX + e**

Grundstruktur der *einfachen, bivariaten* Regression





Lösung der Geradengleichung mit der ***Methode der kleinsten Fehlerquadrate***, d.h. es wird zu jedem Datenpunkt der Wert $\hat{a}$ und $\hat{b}$ gesucht, der –unter Einbeziehung von $e_1 - e_s$ - die Funktion ***F(a, b)*** minimiert.

$$\hat{b} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\overline{X^2} - \bar{X}^2} = \frac{S_{xy}}{{S_x}^2}$$

$$\hat{a} = \bar{Y} - \hat{b} \cdot \bar{X}$$

Nach manueller Berechnung (am Beispiel 1) mit Hilfstabelle oder mit Taschenrechner:

$\hat{a}$ = 9,166488479

$\hat{b}$ = 0,179082177

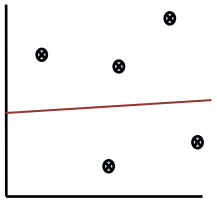
Da jetzt die Werte für ***a*** und ***b*** in die Geradengleichung eingesetzt werden können:

$\hat{Y} = 0{,}17908 \cdot X + 9{,}1665$ (Gegenprobe durch Einsetzen von bekanntem X/Y)

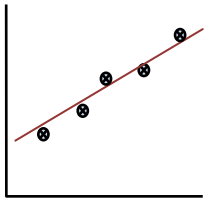
wird eine Prognose von ***Y*** für beliebige Ausprägungen von ***X*** möglich. Jedoch: Modellierter Zusammenhang ist ***nur auf den gegebenen Stichprobenbereich beschränkt***!

Beurteilung der Anpassungsgüte

Mit der o.g. Geradengleichung ist zwar ein funktioneller Zusammenhang hergestellt, aber noch keine Aussage über die Güte der Anpassung getroffen. Es sind folgende Konstellationen möglich:



Niedrige Erklärungskraft



Hohe Erklärungskraft

Um die Erklärungskraft des Modells zu quantifizieren, werden die geschätzten Residuen $(\hat{e}_i = y_i - \hat{y}_i)$ analysiert und über das Bestimmtheitsmaß R^2 ausgedrückt:

$$R^2 = \frac{SQ_{Regression}}{SQ_{Total}} = 1 - \frac{SQ_{Residual}}{SQ_{Total}}$$

wobei

$SQ_{Residual} = \sum(Y_i - \widehat{Y_i})^2$ … Abweichung von Original-Punktwolke zum vorhergesagten Wert auf der Regressionslinie - **Nicht durch Regression erklärter Anteil**

$SQ_{Regression} = \sum(\widehat{Y_i} - \bar{Y})^2$ … **Durch die Regression erklärter Anteil** an der Gesamtvariabilität

$SQ_{Total} = \sum(Y_i - \bar{Y})^2 = SQ_{Regression} + SQ_{Residual}$ … **Totale Variabilität** der Y-Reihe

Im bivariaten Fall gilt: $R^2 = r^2$ und $r = \sqrt{R^2}$

r ist der Korrelationskoeffizient, also ein Maß für die Stärke des linearen Zusammenhangs zwischen X und Y.

Interpretation des Bestimmtheitsmaßes

R^2 = 0: Untere Schranke, Geraden verläuft parallel zur X-Achse, keine Erklärungskraft
R^2 = 1: Obere Schranke, alle Punkte liegen auf der Regressionslinie, 100% Erklärungskraft
R^2 = x: Mit dem Modell können $x \cdot 100$ Prozent der Varianz erklärt werden
Am Beispiel 1: R^2 = 0,735797976 (r = 0,857786673) → 74% Erklärungskraft