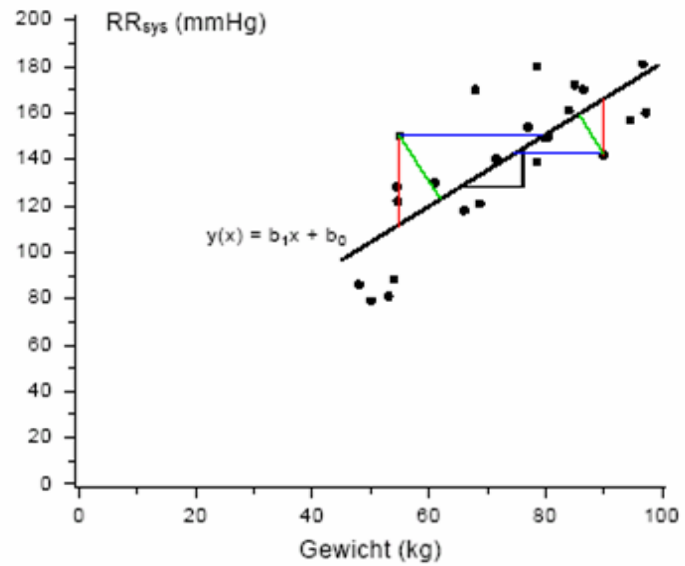
Einfache lineare Regression

- Linearer Zusammenhang zwischen zwei metrischen Größen wird als Gerade visualisiert
- Finde Gerade $Y = \alpha + \beta \cdot X$
- β : Steigung der Geraden, d.h. erhöht sich X um eine Einheit, so erhöht sich Y um β Einheiten.
- α : Achsenabschnitt, d.h. Wert von Y für $X = 0$

Bestimmung der Regressionsgerade

Welche Gerade ist die „Beste“?

- Sie sollte etwa in der „Mitte“ der Punktwolke liegen
- Abweichungen der Wertepaare (x_i, y_i) (Punkte) von der Geraden sollten möglichst „klein“ (minimal) sein



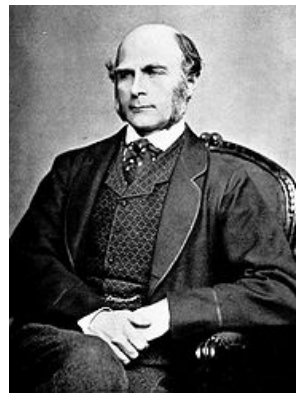
Methode der kleinsten Quadrate

- Y ist Zielgröße und X Einflussgröße
- Y soll mit Hilfe von X erklärt oder prognostiziert werden
- Lineares Modell $Y = \alpha + \beta X + \varepsilon$
- Minimierung der Abstände in Y-Richtung
- Wähle $\hat{\alpha}$ und $\hat{\beta}$ so, dass $\sum_{i=1}^{n} \left(y_i - (\hat{\alpha} + \hat{\beta} x_i) \right)^2$ minimal wird

Geschichte

Idee der KQ-Schätzung von Gauss (1795) im Alter von 18 Jahren

Veröffentlichung von Legendre
Idee der Regression von Galton (1886)



Lineare Einfachregression und Kleinste-Quadrate-Schätzer

Seien $(x_1, y_1), \ldots, (x_n, y_n)$ Beobachtungen der Merkmale X und Y, dann heißt

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

lineare Einfachregression, wobei α den Achsenabschnitt, β die Steigung und ε den Fehler bezeichnet.

Die Kleinste-Quadrate-Schätzer für $\hat{\alpha}$ und $\hat{\beta}$ sind gegeben durch

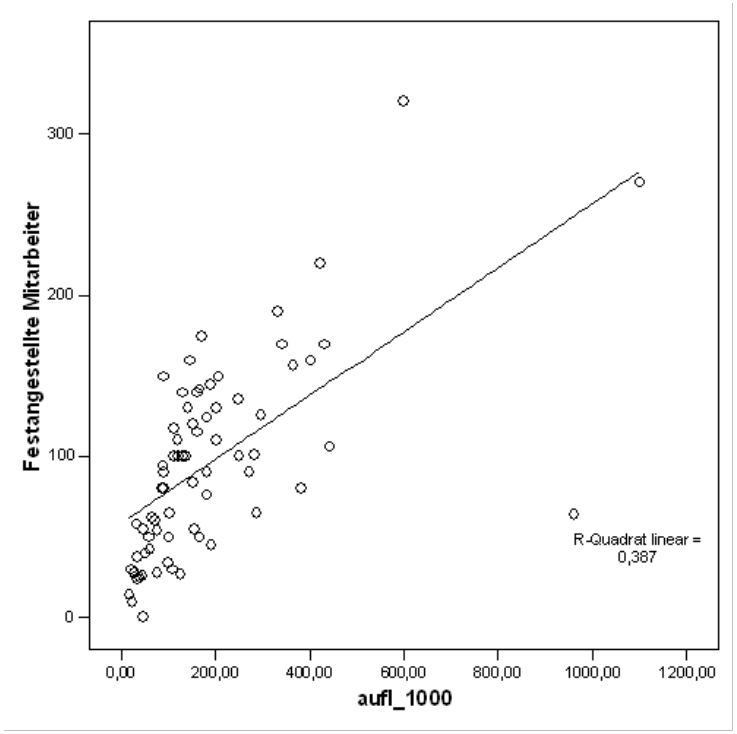
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{S_{xy}}{S_x^2}.$$

Die Residuen berechnen sich durch

$$\varepsilon_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n,$$

mit $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$.

Beispiel: Zahl der Mitarbeiter in Abhängigkeit von der Auflage



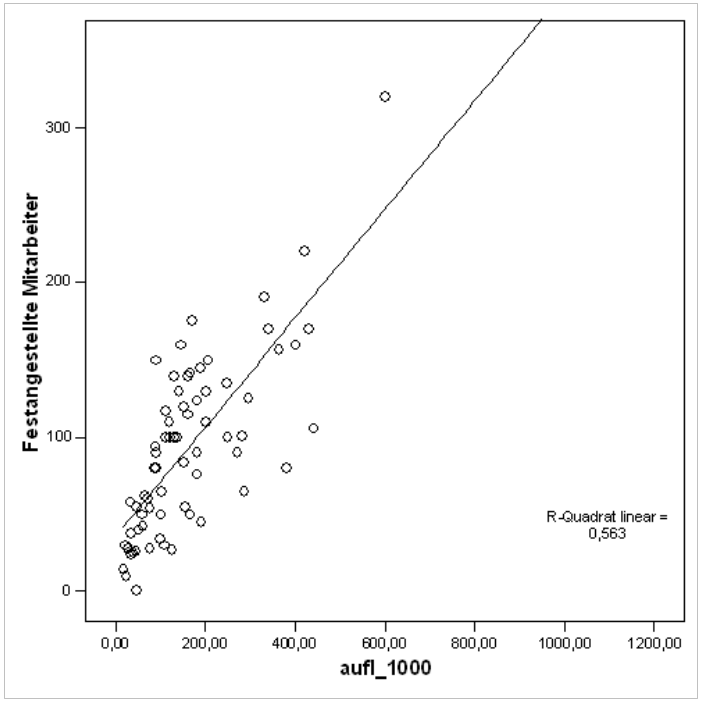
Interpretation:
Mit einer Auflagensteigerung von 1000 ist durchschnittlich die Einstellung von 0.199 Mitarbeitern verbunden.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	Korrelationen		
		B	Standardfehler	Beta			Nullter Ordnung	Partiell	Teil
1	(Konstante)	58,193	8,043		7,235	,000			
	aufl_1000	,199	,030	,622	6,549	,000	,622	,622	,622

a. Abhängige Variable: Festangestellte Mitarbeiter

Regression ohne 2 Extremwerte



Beachte:
Jetzt werden 0.352 Mitarbeiter bei einer Auflagensteigerung von 1000 eingestellt.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	Korrelationen		
		B	Standardfehler	Beta			Nullter Ordnung	Partiell	Teil
1	(Konstante)	36,078	7,740		4,661	,000			
	aufl_1000	,352	,038	,750	9,220	,000	,750	,750	,750

a. Abhängige Variable: Festangestellte Mitarbeiter

Standardabweichung des Störterms

Die geschätzte Abweichung der y-Werte von der Geraden ergibt sich zu:

$$\begin{aligned} s_\varepsilon &= \sqrt{\frac{1}{n-2}\sum \varepsilon_i^2} \\ \varepsilon_i &= y_i - \hat{y}_i \end{aligned}$$

Wichtiges intuitives Maß zur Modellanpassung

Streuungs- und Quadratsummenzerlegung

Ziel: Erklärung der Streuung von Y durch X:

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

Streuung von Y	=	Erklärte Streuung	+	Rest
SST	=	SSM	+	SSE
Quadratsumme Gesamt (Total)	=	Quadratsumme Regression (Model)	=	Quadratsumme Residuen (Error)

Das Bestimmtheitsmaß R^2

Anteil der durch die Regression (d.h. durch X) erklärten Varianz

$$\begin{aligned} R^2 &= \frac{SSM}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \\ &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned}$$

Es gilt: Bestimmtheitsmaß = Quadrat der Korrelation zwischen X und Y

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = r^2$$

Nachweis von $R^2 = r_{XY}^2$

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i) = \hat{\alpha}+\hat{\beta}\bar{x} = (\bar{y}-\hat{\beta}\bar{x})+\hat{\beta}\bar{x} = \bar{y}$$

Daraus folgt:

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2 = \sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i-\hat{\alpha}+\hat{\beta}\bar{x})^2 = \hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

somit für R^2:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \\ &= \frac{s_{XY}^2\cdot s_X^2}{(s_X^2)^2\cdot s_Y^2} = \left(\frac{s_{XY}}{s_X s_Y}\right)^2 = r_{XY}^2 \end{aligned}$$

Umkehrregression I

Vertauscht man die Rollen von X und Y, so erhält man die Umkehrregression.

Daten $(X_i, Y_i), \quad i = 1, \ldots, n$

Regression:	$Y = \alpha + \beta X$	$\beta = \frac{s_{XY}}{s_X^2}$
Umkehrregression:	$X = \gamma + \delta Y$	$\delta = \frac{s_{XY}}{s_Y^2}$

Im XY-Koordinatensystem hat die Gerade der Umkehrregression die Darstellung

$$Y = -\frac{\gamma}{\delta} + \frac{1}{\delta} X$$

Umkehrregression II

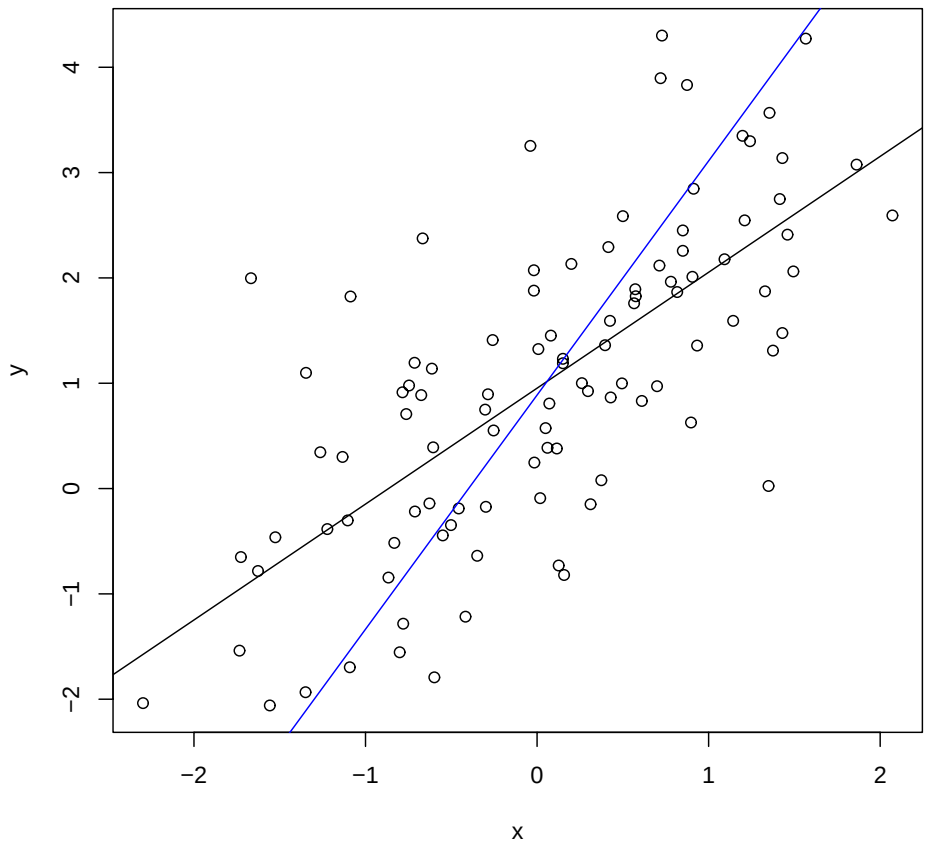
Es gilt:

$$\beta \cdot \delta = \frac{S_{XY}^2}{S_X^2 S_Y^2} = r^2 \leq 1$$

$\Rightarrow |\beta| \leq \frac{1}{|\delta|}$
Gerade der Umkehrregression steiler

und
$\Rightarrow \beta \cdot \delta \geq 0$
β und δ haben gleiches Vorzeichen

Beispiel: Umkehrregression



Orthogonale Regression

Falls man die orthogonalen Abstände zur Gerade minimiert, erhält man eine Gerade zwischen Regression und Umkehrregression. Löse Minimierungsproblem in α, β

$$(\alpha_{ORR}, \beta_{ORR}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \underbrace{\frac{1}{1+\beta_i^2}(y_i - \alpha - \beta x_i)^2}_{\text{Orthogonaler Abstand}}$$

$$\begin{aligned} \hat{\beta}_{ORR} &= \frac{1}{2S_{XY}} \left[(S_Y^2 - S_X^2) + \sqrt{4S_{XY}^2 + (S_Y^2 - S_X^2)^2} \right] \\ \hat{\alpha}_{ORR} &= \bar{y} - \hat{\beta}_{ORR} \cdot \bar{x} \end{aligned}$$

Wichtige Eigenschaften der linearen Regression

- Asymmetrie: Regressionsgerade von Y auf X verschieden von Regressionsgerade von X auf Y
- Die Regressionsgerade geht durch $(\bar{x}, \bar{y})$
- Interpretation der Steigung b steht im Mittelpunkt der Interpretation
- R^2-Wert gibt den Varianz-Erklärungsanteil wieder
- R^2 ist Quadrat der Korrelation
- s_ε gibt durchschnittliche Abweichung der Werte von der Regressionsgeraden an