

Partielle Korrelation

Ziel:

Bestimmung der Korrelation zweier Merkmale unter „konstant halten“ eines dritten Merkmals

Beispiel:

Korrelation der Zahl der freien und festen Mitarbeiter unter konstanter Auflage

Idee:

Herausrechnen des Einflusses des dritten Merkmals durch lineare Regression



Partieller Korrelationskoeffizient (Definition)

Es soll der lineare Zusammenhang zwischen X und Y bei festen Z bestimmt werden. Betrachte lineare Regressionen

$$\begin{aligned}X &= \alpha_1 + \beta_1 Z + \varepsilon_1 \\Y &= \alpha_2 + \beta_2 Z + \varepsilon_2.\end{aligned}$$

Aus den Daten (x_i, y_i, z_i) werden die Parameter nach der KQ-Methode geschätzt. Man erhält die bereinigten Variablen X^{BZ} und Y^{BZ} als Residuen der Regressionen:

$$\begin{aligned}X^{BZ} &= X - \hat{\alpha}_1 - \hat{\beta}_1 Z \\Y^{BZ} &= Y - \hat{\alpha}_2 - \hat{\beta}_2 Z\end{aligned}$$

Dann heißt die Maßzahl

$$r_{XY|Z} = r_{X^{BZ} Y^{BZ}}$$

partieller Korrelationskoeffizient zwischen X und Y unter Z .

Berechnung der partiellen Korrelation

Es gilt:

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r_{XZ}^2} \sqrt{1 - r_{YZ}^2}}$$



Beispiel: Korrelation der Anzahl freier Mitarbeiter mit der Anzahl fest angestellter Mitarbeiter

Korrelationen

		Festangestellte Mitarbeiter	Freie Mitarbeiter
Festangestellte Mitarbeiter	Korrelation nach Pearson	1	,490**
	Signifikanz (2-seitig)		,000
	N	68	57
Freie Mitarbeiter	Korrelation nach Pearson	,490**	1
	Signifikanz (2-seitig)	,000	
	N	57	57

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

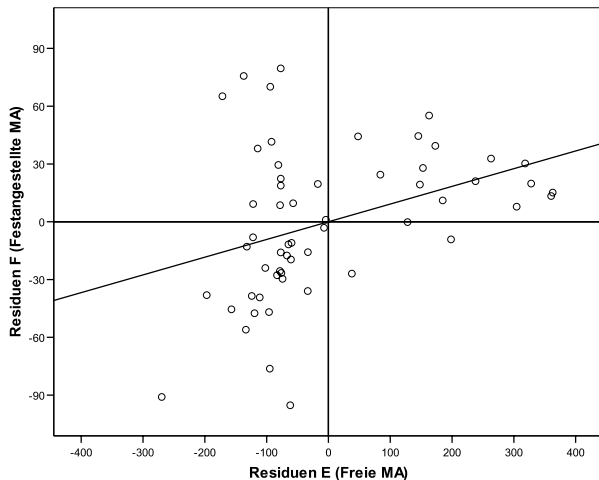
Einfache Korrelation

Korrelationen

Kontrollvariablen			Festangestellte Mitarbeiter	Freie Mitarbeiter
auf_1000	Festangestellte Mitarbeiter	Korrelation	1,000	,366
		Signifikanz (zweiseitig)	.	,006
		Freiheitsgrade	0	54
	Freie Mitarbeiter	Korrelation	,366	1,000
		Signifikanz (zweiseitig)	,006	.
		Freiheitsgrade	54	0

Nach Auflage bereinigte Korrelation

Freie und fest angestellte Mitarbeiter in der Zeitungsstudie (bereinigt nach der Größe der Zeitung)



Multiple Regressionsmodell

Gegeben sind die Zielgröße Y und die Einflussgrößen X_k

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_p \cdot x_p + \varepsilon$$

Das Modell kann aus den entsprechenden Daten mit Hilfe der KQ-Methode geschätzt werden. Analog zum linearen Modell ist das Bestimmtheitsmaß r^2 ein zentrales Kriterium für die Modellanpassung.

Die Parameter b_k haben folgende Interpretation:
Steigt das Merkmal X_k um eine Einheit und werden die anderen Einflussgrößen festgehalten, so steigt Y im Durchschnitt um b_k Einheiten.

Beispiel: Festangestellte und Freie Mitarbeiter

FAM: Anzahl festangestellter Mitarbeiter

FM: Anzahl freier Mitarbeiter

AT: Auflage in Tausend

$$FAM = a + b_1 \cdot FM + b_2 \cdot AT + \varepsilon$$

$$FAM = 31 + 0.092 \cdot FM + 0.32 \cdot AT + \varepsilon$$

$$FAM = 67 + 0.17 \cdot FM + f$$

Der Zusammenhang zwischen *FAM* und *FM* wird bei Berücksichtigung von *AT* geringer.

Zusammenfassung multiples Regressionsmodell

Das multiple Regressionsmodell ist nützlich, um Zusammenhänge zwischen Merkmalen zu analysieren.

Es ermöglicht:

- Quantifizierung des Zusammenhangs
- Herausrechnen von Störgrößen
- Auswahl von relevanten Einflussgrößen



Erweiterungen des Modells beinhalten:

- Nichtlineare Zusammenhänge
- Einbeziehung von nominalen Merkmalen als Einflussgrößen (z.B. Geschlecht, Nationalität, etc.)
- Binäre Zielgrößen (krank/gesund)



Das nichtlineare Regressionsmodell

Zusammenhang zwischen X und y : β kann Vektor sein

$$Y = f(X, \beta) + \varepsilon$$

KQ- Schätzer aus Daten Y_i, X_i :

$$\hat{\beta} := \arg \min_{\beta} \sum_{i=1}^n (y_i - f(x_i, \beta))^2$$

Berechnung oft nicht mit geschlossenen Formeln, aber numerisch möglich, z.B. mit Paket nls in R.

Beispiel: Wachstum mit Obergrenze β_1

$$y = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 * t)} + \varepsilon$$