Mathematische und statistische Methoden I

Sprechstunde
jederzeit nach
Vereinbarung und
nach der Vorlesung

Wallstr. 3, 6. Stock,
Raum 06-206

Dr. Malte Persike


persike@uni-mainz.de


lordsofthebortz.de


twitter.com/methodenlehre


tinyurl.com/gplusmethodenlehre

WiSe 2011/2012

Fachbereich Sozialwissenschaften
Psychologisches Institut
Johannes Gutenberg Universität Mainz

Kennwerte

Test der Gewichte gegen Null

Kennwerte der multiplen Regression

1. Der multiple Korrelationskoeffizient R

- Definition: Der **multiple Korrelationskoeffizient R** repräsentiert die Korrelation zwischen dem Kriterium y und **allen** Prädiktoren $x_1 \ldots x_k$
- Dabei berücksichtigt R etwaige Interkorrelationen zwischen den Prädiktoren (und entfernt sie)
- Der multiple Korrelationskoeffizient R ist definiert als

$$R_{y \cdot x_1 x_2 \ldots x_k} = \sqrt{\sum_{j=1}^{k} \beta_j r_{x_j y}}$$

- Er ist mathematisch äquivalent zur Korrelation zwischen den gemessenen y-Werten und den vorhergesagten y^{dach}-Werten, also

$$R_{y \cdot x_1 x_2 \ldots x_k} = r_{y\hat{y}}$$

Kennwerte

Test der Gewichte gegen Null

Kennwerte der multiplen Regression

2. Der multiple Determinationskoeffizient R^2

- Definition: Der **multiple Determinationskoeffizient R^2** repräsentiert die Varianzaufklärung, die **alle** Prädiktoren $x_1 \ldots x_k$ am Kriterium y leisten
- Der multiple Determinationskoeffizient R^2 ist definiert als

$$R^2 = \frac{\text{Erklärte Streuung}}{\text{Gesamt-Streuung}} = 1 - \frac{\text{Fehlerstreuung}}{\text{Gesamt-Streuung}}$$

- Rechnerisch:

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = 1 - \frac{Var(e)}{Var(y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2}$$

Kennwerte der multiplen Regression

2. R und R^2

- R und R² sind tatsächlich direkt ineinander transformierbar

$$R = \sqrt{\sum_{j=1}^{k} \beta_j r_{x_j y}} \quad \Leftrightarrow \quad \sum_{j=1}^{k} \beta_j r_{x_j y} = R^2$$

- Für die Bewertung des R können wieder die Daumenregeln nach Cohen (1988) verwendet werden:

$R < \pm 0.10$	$\rightarrow$	keine Korrelation
$R < \pm 0.30$	$\rightarrow$	kleine Korrelation
$R < \pm 0.50$	$\rightarrow$	mittlere Korrelation
$R \geq \pm 0.50$	$\rightarrow$	hohe Korrelation

Kennwerte

Test der Gewichte gegen Null

Kennwerte der multiplen Regression

2. R und R^2

- Dies führt aber auf ein Problem bei der Bewertung des R^2, denn die Quadratur der Daumenregeln liefert

$R^2 < \pm 0.01$	$\rightarrow$	keine Korrelation
$R^2 < \pm 0.10$	$\rightarrow$	kleine Korrelation
$R^2 < \pm 0.25$	$\rightarrow$	mittlere Korrelation
$R^2 \geq \pm 0.25$	$\rightarrow$	hohe Korrelation

- In der Praxis bedeuten 25% aufgeklärte Varianz, dass 75% der Streuung in der AV nicht durch die Regressionsgleichung, d.h. die Prädiktoren erklärt wird

Kennwerte

Test der Gewichte gegen Null

Kennwerte der multiplen Regression

2. R und R^2

- Daher hat Cohen alternative **Daumenregeln für die Bewertung des R^2** vorgeschlagen (beruhend auf seiner Definition der Effektstärke)

$R^2 < \pm 0.20$	$\rightarrow$	keine Varianzaufklärung
$R^2 < \pm 0.50$	$\rightarrow$	kleine Varianzaufklärung
$R^2 < \pm 0.80$	$\rightarrow$	mittlere Varianzaufklärung
$R^2 \geq \pm 0.80$	$\rightarrow$	hohe Varianzaufklärung

- Diese Regeln sind recht streng, insbesondere in der Feldforschung, wo 20-30% Varianzaufklärung bereits als gutes Ergebnis gewertet werden

Kennwerte

Test der Gewichte gegen Null

Kennwerte der multiplen Regression

3. Abhängigkeit

a) Sind die Prädiktoren **unabhängig**, so sind die ß-Gewichte gleich den Kriteriumskorrelationen und die aufgeklärte Varianz ist die Summe der Quadrate der ß-Gewichte

- **Erklärung**: Bei perfekt unabhängigen Prädiktoren ist die Prädiktorinterkorrelationsmatrix R_{xx} gleich der Identitätsmatrix I.

$$\vec{\beta} = I \times r_{xy} \Leftrightarrow \vec{\beta} = r_{xy}$$

- Damit gilt für den multiplen Korrelationskoeffizienten R

$$R_{y \cdot x_1 x_2 \ldots x_k} = \sqrt{\sum_{j=1}^{k} r_{x_j y}^2}$$

- Und R^2 ist einfach die Summe der quadrierten Kriteriumskorrelationen

$$R^2_{y \cdot x_1 x_2 \ldots x_k} = \sum_{j=1}^{k} r_{x_j y}^2$$

Kennwerte der multiplen Regression

3. Abhängigkeit

a) Sind die Prädiktoren **unabhängig**, so sind die *ß*-Gewichte gleich den Kriteriumskorrelationen und die aufgeklärte Varianz ist die Summe der Quadrate der *ß*-Gewichte

b) Sind die Prädiktoren **abhängig** (interkorreliert), so sind 3 Fälle zu unterscheiden:

1. Der Prädiktor klärt zumindest Teile der Varianz am Kriterium auf, die andere Prädiktoren nicht aufklären: er ist **nützlich**.
2. Der Prädiktor enthält (nur) Information, die auch andere Prädiktoren enthalten: er ist **redundant**
3. Der Prädiktor unterdrückt irrelevante Varianz in anderen Prädiktoren: er ist ein **Suppressor**

Bortz, S. 349

Kennwerte

Test der Gewichte gegen Null

Kennwerte der multiplen Regression

3a. Nützlichkeit

- **Nützlichkeit** = Der Beitrag, den eine Variable zur Varianzaufklärung des Kriteriums leistet, der von den anderen Variablen nicht geleistet wird

- Die Nützlichkeit einer Variablen x_j berechnet sich als

$$U_j = R^2_{y,x_{1,2,\ldots,k+j}} - R^2_{y,x_{1,2,\ldots,k-j}}$$

U_j ist also der Betrag, um den R^2 wächst, wenn die Variable x_j in die multiple Regressionsgleichung aufgenommen wird.

Kennwerte der multiplen Regression

3b. Redundanz

- **Redundanz** = die vielen Variablen messen Aspekte gemeinsam, so dass man prinzipiell weniger Prädiktoren benötigte → **unerwünschter Aspekt**
- Die Variable x_j ist redundant zur Vorhersage von Variable y wenn gilt

$$\left| \beta_{x_j} \cdot r_{x_j y} \right| < r^2_{x_j y}$$

- Prädiktoren enthalten empirisch nahezu immer gemeinsame Varianzanteile und sind somit „teilweise redundant“. Echte Redundanz liegt erst gemäß obiger Definition vor.
- **Multikollinearität**: Die Kovarianz eines Prädiktors mit dem Kriterium ist in den anderen Prädiktoren (fast) vollständig enthalten → extremer Fall von Redundanz, der **unbedingt zu vermeiden** ist.

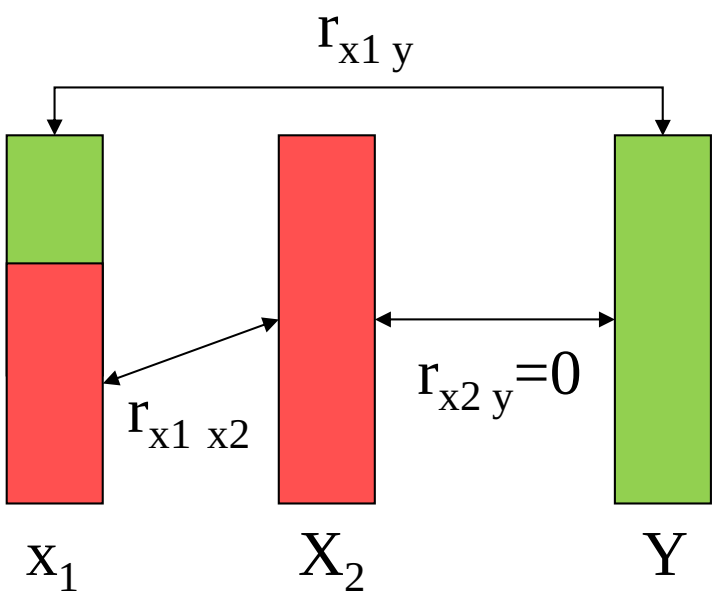
Bortz, S. 352 – 354

Kennwerte

Test der Gewichte gegen Null

Kennwerte der multiplen Regression

3c. Suppression



x_2 „bindet“ irrelevante Prädiktorinformation

x_2 hängt nicht mit y zusammen, trotzdem erhöht sie R^2

Kennwerte der multiplen Regression

3c. Suppression

- **Defintion:** Eine Variable x_j ist ein **Suppressor**, wenn gilt:

$$U_{x_j} > r^2_{x_j y}$$

- Die Zunahme der erklärten Varianz durch Aufnahme der Variable ist also größer als die einzelne Varianzaufklärung.

- **Vereinfachung**: Bei nur zwei Prädiktoren x_1 und x_2 ist x_2 ein Supressor, wenn gilt:

$$\left| r_{x_1 z.x_2} \right| > \left| r_{x_1 z} \right| \cdot \sqrt{\frac{1 - r^2_{x_1 x_2}}{1 - r^2_{x_2 z}}}$$

Grundlagen

Linearisierbare Formen

Polynome

Nichtlineare Regression

Grundlagen

- Bei einer Reihe psychologischer Fragestellungen ergeben sich **nichtlineare Zusammenhänge** zwischen UV & AV.
- Beispiele: Reaktionszeit, Blutalkohol und psychomotorische Leistungen, Fehlerraten in Leistungstests bei verschiedenen Aufgabenschwierigkeiten
- Solche nichtlinearen Zusammenhänge lassen sich in **zwei Klassen** einteilen:

1. Zusammenhänge, die sich durch eine einfache (nichtlineare) Transformationen in lineare Zusammenhänge überführen lassen
2. Zusammenhänge, für die eine nichtlineare Regressionsgleichung gelöst werden muss.

Grundlagen

Linearisierbare Formen

Polynome

Nichtlineare Regression

Linearisierbare und polynomiale Formen

Fall 1: Linearisierende Transformation, z.B.

$$\hat{y} = b_0 \cdot x^{b_1} \xrightarrow{\ln(\bullet)} \ln(\hat{y}) = \ln(b_0) + b_1 \cdot \ln(x)$$

(hier nicht behandelt)

Fall 2: Nicht (einfach) linearisierbar

$$\hat{y} = b_0 + b_1 \cdot x + b_2 \cdot x^2$$

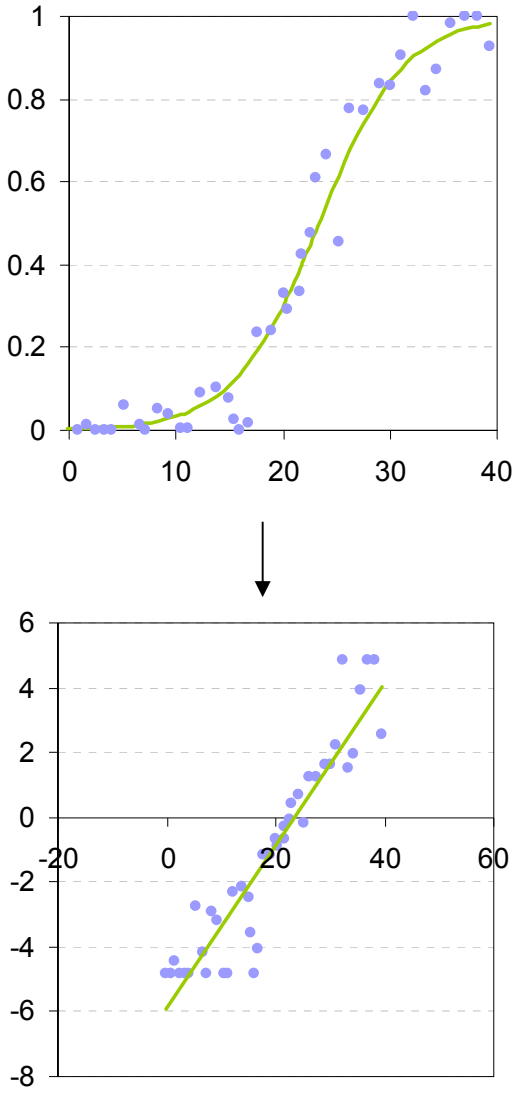
Grundlagen

Linearisierbare Formen

Polynome

Nichtlineare Regression

Beispiel: Logistische Regression



- Gemessene Daten verlaufen **ogivenförmig** und variieren zwischen 0 und 1
- Umformung der y-Werte durch Logarithmieren bewirkt eine **Linearisierung** der Daten
- Mithilfe dieser neuen y-Werte kann eine lineare Regression bestimmt werden, um die Parameter b_0 und b_1 zu errechnen

Grundlagen
Linearisierbare Formen
Polynome

Polynomiale Regression

Grundlagen und Durchführung

- Häufig können Merkmalszusammenhänge durch **Polynome 2. oder 3. Ordnung** gut beschrieben werden, d.h.

$$\hat{y} = b_0 + b_1 \cdot x + b_2 \cdot x^2$$

oder

$$\hat{y} = b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$$

- Dies ist formal eine **lineare multiple Regression**, allerdings nicht mit mehreren Prädiktoren, sondern mit einem Prädiktor sowie Transformationen seiner selbst.

Grundlagen

Linearisierbare Formen

Polynome

Polynomiale Regression

Grundlagen und Durchführung

- Eine solche **polynomiale Regression** wird berechnet, indem einfach die transformierten Prädiktorterme x^2, x^3 usw. bestimmt werden
- Dann wird auf diesen eine übliche **lineare multiple Regression** durchgeführt
- Die Einträge der Korrelationsmatrix sind dabei dann die Korrelationen des Prädiktors mit sich selbst in den transformierten Formen
- Es können alle von **Kennwerte und Gütemaße** der multiplen Regression bestimmt werden.
- Die polyn. Regression ist auch über die **KQ-Methode** (inkl. Normalgleichungen) herzuleiten. Dies führt auf dasselbe Ergebnis wie der hier verfolgte Ansatz.