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

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Multiple Regression

Grundlagen

- Oft werden in psychologischen Untersuchungen nicht nur eine sondern **mehrere UVn** betrachtet, die **eine AV** beeinflussen (oder vorhersagen sollen).
- **Beispiele**: Abhängigkeit der Lebenszufriedenheit von sozialem, ökonomischem und Gesundheitsstatus; Beeinflussung sportlicher Leistung durch Trainingszustand und Anwesenheit von Zuschauern.
- Solche Fragestellungen werden auch als **multifaktoriell** bezeichnet
- **Problem**: Die Berechnung vieler paarweiser Korrelationen im multifaktoriellen Fall vernachlässigt mögliche Zusammenhänge zwischen den UVn

Multiple Regression

Grundlagen

Drei **Hauptfragestellungen der Regressionsrechnung**:

1. Gibt es eine statistische Beziehung zwischen mehreren Variablen, die die Vorhersage der AV aus der UV erlaubt?

2. Kann eine möglichst einfache mathematische Regel formuliert werden, die diesen Zusammenhang beschreibt?

3. Wie gut ist diese Regel im Hinblick auf die Vorhersage?

Multiple Regression

Grundgleichung

- Die vorherzusagende Variable (AV, y-Wert) wird als **Kriterium** oder **Response** bezeichnet, die vorhersagenden Variablen (UVn, x-Werte) als **Prädiktoren** oder **erklärende Variablen**.

- Die Vorhersagegleichung der multiplen Regression mit k Prädiktoren wird geschrieben als

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k$$

- Bei standardisierten Daten verwendet man das Symbol β für die k Regressionsparameter (bzw. „-gewichte“)

$$\hat{z}_y = \beta_1 \cdot z_{x_1} + \beta_2 \cdot z_{x_2} + \ldots + \beta_k \cdot z_{x_k}$$

Multiple Regression

Grundgleichung

Gründe für die Annahme einer linearen Gleichung:

- Lineare Zusammenhänge sind **einfach zu verstehen**
- Lineare Zusammenhänge sind mathematisch und statistisch **einfach zu behandeln**
- Lineare Gleichungen haben sich vielfach als **gute Approximationen** für komplexe Beziehungen erwiesen
- Achtung: Auch wenn die Beziehung zwischen zwei ZVn linear „aussieht“, muss es sich nicht zwangsläufig um einen linearen Zusammenhang handeln.

Regression

Methode der kleinsten Quadrate (KQ-Kriterium)

- Zur Minimierung des Vorhersagefehlers wird oft das **Kleinste-Quadrate Kriterium** verwendet (KQ; oder Ordinary Least Squares, OLS)
- Parameter der multiplen Regressionsgleichung werden so gewählt, dass das **Quadrat der Abweichungen** von gemessenem und geschätztem Wert **minimiert** wird
- Für eine Versuchsperson i aus allen n gelte:

$$y_i = \hat{y}_i + e_i \quad \Leftrightarrow \quad e_i = y_i - \hat{y}_i$$

beobachteter Kriteriumswert = vorhergesagter Wert + Messfehler

- Dann soll für alle n Datenwerte erreicht werden, dass

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} e_i^2 \rightarrow \min$$

Minimierung der Quadratsumme des Vorhersagefehlers

Regression

Methode der kleinsten Quadrate (KQ-Kriterium)

- Mithilfe der Allgemeinen Gleichung der einfachen linearen Regression lässt sich für die **Streuung des Vorhersagefehlers QS_e** also schreiben:

$$QS_e = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 \cdot x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik} \right)^2 \rightarrow \min$$

bzw. in der standardisierten Form

$$QS_e = \sum_{i=1}^{n} \left(z_{y_i} - \hat{z}_{y_i} \right)^2 = \sum_{i=1}^{n} \left(z_{y_i} - \beta_1 \cdot z_{x_{i1}} - \beta_2 z_{x_{i2}} - \ldots - \beta_k z_{x_{ik}} \right)^2 \rightarrow \min$$

- Die Minimierung der Regressionsparameter erfolgt über partielle Differenzierung nach jedem einzelnen der b- bzw. β-Gewichte

Regression

Normalgleichungen der multiplen Regression

- Die partielle Differenzierung der nichtstandardisierten Gleichung mit k Prädiktoren führt immer auf ein **System von $k+1$ Normalgleichungen**, das wie folgt aufgebaut ist:

$$\sum_{i=1}^{n} y = \sum_{i=1}^{n} b_0 + b_1 \sum_{i=1}^{n} x_1 + b_2 \sum_{i=1}^{n} x_2 + \ldots + b_k \sum_{i=1}^{n} x_k$$

$$\sum_{i=1}^{n} yx_1 = b_0 \sum_{i=1}^{n} x_1 + b_1 \sum_{i=1}^{n} x_1^2 + b_2 \sum_{i=1}^{n} x_1 x_2 + \ldots + b_k \sum_{i=1}^{n} x_1 x_k$$

$$\sum_{i=1}^{n} yx_2 = b_0 \sum_{i=1}^{n} x_2 + b_1 \sum_{i=1}^{n} x_1 x_2 + b_2 \sum_{i=1}^{n} x_2^2 + \ldots + b_k \sum_{i=1}^{n} x_2 x_k$$

$$\ldots$$

$$\sum_{i=1}^{n} yx_k = b_0 \sum_{i=1}^{n} x_k + b_1 \sum_{i=1}^{n} x_1 x_k + b_2 \sum_{i=1}^{n} x_2 x_k + \ldots + b_k \sum_{i=1}^{n} x_k^2$$

Regression

Normalgleichungen der multiplen Regression

- In der standardisierten Form ergibt sich ein **System von k Normalgleichungen**:

$$\sum_{i=1}^{n} z_{x_1} z_y = \beta_1 \sum_{i=1}^{n} z_{x_1}^2 + \beta_2 \sum_{i=1}^{n} z_{x_1} z_{x_2} + \ldots + \beta_k \sum_{i=1}^{n} z_{x_1} z_{x_k}$$

$$\sum_{i=1}^{n} z_{x_2} z_y = \beta_1 \sum_{i=1}^{n} z_{x_1} z_{x_2} + \beta_2 \sum_{i=1}^{n} z_{x_2}^2 + \ldots + \beta_k \sum_{i=1}^{n} z_{x_2} z_{x_k}$$

$$\ldots$$

$$\sum_{i=1}^{n} z_{x_k} z_y = \beta_1 \sum_{i=1}^{n} z_{x_1} z_{x_k} + \beta_2 \sum_{i=1}^{n} z_{x_2} z_{x_k} + \ldots + \beta_k \sum_{i=1}^{n} z_{x_k}^2$$

Regression

Multiple Regression - Zusammenfassung

- Die partielle Differenzierung einer multiplen Regressionsgleichung mit k Prädiktoren führt immer auf ein **System von $k+1$ (bzw. k) Normalgleichungen**
- Prinzip: Die summierte Ausgangsgleichung wird nacheinander mit jedem Prädiktor $x_0 \ldots x_k$ (bzw. $z_1 \ldots z_k$) multipliziert
- Die Normalgleichungen liefern dann für $k+1$ (bzw. k) unbekannte Regressionsparameter genau so viele Gleichungen.
- Dieses Gleichungssystem kann nun durch Substitution oder Diagonalisierung für die Parameter gelöst werden

Matrixalgebraische Berechnung

der multiplen Regression

- Wir haben gesehen, dass die Normalgleichungen der multiplen Regression für standardisierte Daten lauteten:

$$\sum_{i=1}^{n} z_{x_1} z_y = \beta_1 \sum_{i=1}^{n} z_{x_1}^2 + \beta_2 \sum_{i=1}^{n} z_{x_1} z_{x_2} + \ldots + \beta_k \sum_{i=1}^{n} z_{x_1} z_{x_k}$$

$$\sum_{i=1}^{n} z_{x_2} z_y = \beta_1 \sum_{i=1}^{n} z_{x_1} z_{x_2} + \beta_2 \sum_{i=1}^{n} z_{x_2}^2 + \ldots + \beta_k \sum_{i=1}^{n} z_{x_2} z_{x_k}$$

$$\ldots$$

$$\sum_{i=1}^{n} z_{x_k} z_y = \beta_1 \sum_{i=1}^{n} z_{x_1} z_{x_k} + \beta_2 \sum_{i=1}^{n} z_{x_2} z_{x_k} + \ldots + \beta_k \sum_{i=1}^{n} z_{x_k}^2$$

- Weiterhin ist die Korrelation zweier Variablen x_p und x_q:

$$r_{x_p x_q} = \frac{1}{n} \sum_{i=1}^{n} z_{i,x_p} z_{i,x_q}$$

Matrixalgebraische Berechnung

der multiplen Regression

- Damit reduziert sich das Normalgleichungssystem zu:

$$
\begin{aligned}
r_{x_1 y} &= \beta_1 + \beta_2 r_{x_1 x_2} + \beta_3 r_{x_1 x_3} + \ldots + \beta_k r_{x_1 x_k} \\
r_{x_2 y} &= \beta_1 r_{x_1 x_2} + \beta_2 + \beta_3 r_{x_2 x_3} + \ldots + \beta_k r_{x_2 x_k} \\
r_{x_3 y} &= \beta_1 r_{x_1 x_3} + \beta_2 r_{x_2 x_3} + \beta_3 + \ldots + \beta_k r_{x_3 x_k} \\
&\ldots \\
r_{x_k y} &= \beta_1 r_{x_1 x_k} + \beta_2 r_{x_2 x_k} + \beta_3 r_{x_3 x_k} + \ldots + \beta_k
\end{aligned}
$$

- In Matrixnotation ist dies:

$$R_{xx} \times \vec{\beta} = \overrightarrow{r_{xy}} \quad \text{mit} \quad R_{xx} = \frac{1}{n} \cdot Z^T Z$$

Matrixalgebraische Berechnung

der multiplen Regression

- In Matrixnotation ist dies:

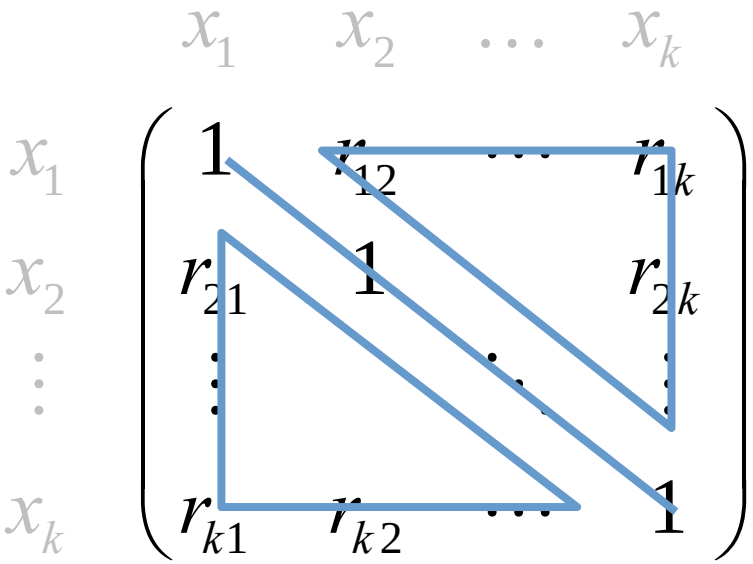
$$R_{xx} \times \vec{\beta} = \overrightarrow{r_{xy}} \quad \text{mit} \quad R_{xx} = \frac{1}{n} \cdot Z^T Z$$

- wobei: $R_{xx} = k \times k$ Matrix der Prädiktorinterkorrelationen

Exkurs: Die Korrelationsmatrix R

Aufbau und Bedeutung

- Die Korrelationsmatrix R stellt die Korrelationen zwischen k Variablen in Matrixschreibweise dar.
- Sie ist quadratisch und enthält $k \times k$ Korrelationen



- Die **Hauptdiagonale** enthält die Korrelationen der Variablen mit sich selbst (r_{xx} = 1)
- Die untere und obere Dreiecksmatrix sind **symmetrisch**

Matrixalgebraische Berechnung

der multiplen Regression

- In Matrixnotation ist dies:

$$R_{xx} \times \vec{\beta} = \vec{r_{xy}} \quad \text{mit} \quad R_{xx} = \frac{1}{n} \cdot Z^T Z$$

- wobei: $R_{xx} = k \times k$ Matrix der Prädiktorinterkorrelationen

$\vec{r}_{xy} = k \times 1$ Vektor der Kriteriumskorrelationen

$\vec{\beta} = k \times 1$ Vektor der Regressionsgewichte

$Z = n \times k$ Vektor der z-standardisierten Daten

- Lösung: Inverse Interkorrelationsmatrix vormultiplizieren

$$R_{xx}^{-1} R_{xx} \times \vec{\beta} = R_{xx}^{-1} \vec{r_{xy}} \quad \Leftrightarrow \quad \vec{\beta} = R_{xx}^{-1} \vec{r_{xy}}$$

Matrixalgebraische Berechnung

Rückrechnung der unstandardisierten Parameter

- Wurden die ***β*-Parameter** für die z-standardisierten Daten matrixalgebraisch bestimmt, kann die Berechnung der unstandardisierten ***b*-Parameter** vorgenommen werden über

$$b_i = \beta_i \frac{SD_y}{SD_{x_i}} \qquad \text{mit } i = 1, 2, ..., k$$

- Die Konstante b_0 wird dann berechnet als

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - ... - b_k\bar{x}_k$$

Bortz, S. 183 – 189

Matrixalgebraische Berechnung

Spezialfall: Nur ein Prädiktor

Bei nur einem Prädiktor vereinfacht sich die Berechnung der Regressionsgewichte erheblich.

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

1. Steigung: $b_1 = r_{xy} \cdot \frac{s_y}{s_x}$ oder $b_1 = \frac{\mathrm{cov}(x,y)}{s_x^2}$

2. y-Achsenabschnitt: $b_0 = \bar{y} - b_1 \cdot \bar{x}$

Interpretation der Lösung

b- und *β*-Gewichte

- Die Größe eines ***b*-Gewichtes** gibt an, um wieviele Einheiten sich der Wert des unstandardisierten Kriteriums verändert, wenn der Betrag des unstandardisierten Prädiktors um 1 steigt.
- Die Größe des ***β*-Gewichtes** gibt dasselbe für die standardisierten Variablen an
- Das ***b*-Gewicht** beantwortet die Frage: „Ich möchte einen der Prädiktoren um 1 erhöhen. Welchen sollte ich wählen, damit das Kriterium maximal steigt?“
- Das ***β*-Gewicht** beantwortet die Frage: „Mit welchem Prädiktor erhöhe ich das Kriterium am effizientesten?“
- Das b-Gewicht liefert also eine **absolute**, das β-Gewicht eine **relative Information**.

Regression

Interpretation der Lösung

Vorsicht bei der Interpretation der Regressionsgleichung

- Bei der **Korrelationsrechnung** bedeutet ein Zusammenhang niemals **Kausalität**, lediglich **Assoziation**
- Bei der **Regressionsrechnung** gilt zunächst dasselbe
- Die Kausalitätsvermutung wird (wenn überhaupt) schon bei der Aufstellung der Regressionsgleichung getroffen, nicht erst bei der Interpretation der Ergebnisse.
- Um tatsächlich Kausalität festzustellen, müssen weitere Randbedingungen vorliegen (i.e. zeitliche Antezedenz von Ursache vor Wirkung, Generalisierbarkeit etc.).

Relevante Excel Funktionen

- Multiple Regression
  - MMULT()
  - MTRANS()
  - MINV()