Hallo,

kurze Anmerkung: Diese Scripte stammen von 1999. Ich kann leider dazu

keine Fragen mehr beantworten! :-(

Euch trotzdem viel Erfolg!

Dorthe

dorthe@luebbert.net

Statistik C

© Dorthe Lübbert, Dorthe.Luebbert@ruhr-uni-bochum.de

1 Regressionsrechnung

1.1 Einfache Regressionsrechnung

Die einfache lineare Regressionsanalyse sucht nach einer linearen Gleichung, die den Zusammenhang zwischen x_i und y_i zum Ausdruck bringt.

Voraussetzung: x_i und y_i sind mindestens intervall-, d.h. metrisch skaliert.

Begriffe:

X: exogene Variable = Einflußfaktor = erklärende Variable = Regressor = unabhängige Variable

Y: endogene Variable = Zielvariable = abhängige Variable = erklärende Variable = Regressand

Schätzgleichung: Gleichung, die exakt die Stichprobe beschreibt

Folgende Probleme lassen sich mit der linearen Einfachregression lösen:

1. Man will wissen, welche Grundrichtung der Beziehung zwischen X und Y besteht.
   Wie groß ist die prop. Veränderung in Y, wenn X_i um eine Einheit erhöht/vermindert wird?
   Bsp.: Pro Jahr zusätzlicher Schulbildung erhöht sich das Einkommen um *b* Einheiten.
2. Man will einen Schätzwert von y für einen X-Wert ermitteln, der außerhalb der Reihe der Beobachtungswerte liegt (→ Extrapolation). Man prognostiziert also.
3. Man will einen Schätzwert von Y wissen, wobei der X-Wert zwischen zwei bekannten X-Werten liegt, selbst aber nicht realisiert ist (→ Interpolation).

1. Bei Zeitreihen wird ein Entwicklungstrend berechnet und als Prognose in die Zukunft fortgeschrieben

Es gibt zwei verschiedene Problemansätze:
Die Frage nach der

a) mathematischen Art der Beziehung zwischen *x* und *y* liefert die *Regressionsgleichung*
b) Stärke der Beziehung liefert den Korrelationskoeffizienten *r* (Bravais-Pearson)

1.2 Das Modell der einfachen linearen Regression

Ein reales Problem kann in die folgende angemessene formale Form übersetzt werden. Zwischen X und Y besteht ein Zusammenhang, der durch die Gleichung $\hat{y}_i = \alpha + \beta x_i + u_1$ zum Ausdruck gebracht werden kann.

Jeder Wert von Y_i läßt sich aus zwei Komponenten zusammengesetzt auffassen:

$\alpha + \beta x_i$: Wert, den yi annehmen würde, falls der Zusammenhang zwischen X und Y streng deterministisch (sprich linear) wäre.

u_i : Wert, um den y_i von seiner deterministischen Komponente $\alpha+\beta x_i$ abweicht (Abweichung zwischen dem realen Wert und der später zu berechnenden Regressionsgerade), u_i ist der Wert der Störgröße u_i. U_i spezifiziert den stochastischen Teil des Zusammenhangs.

U_i läßt sich als Zufallsvariable auffassen, da oft nicht angegeben werden kann, welchen Wert Ui bei vorgegebenem Wert x_i annimmt. Ui läßt sich aber auch als Störvariable auffassen, da die ui die Abweichungen von einer linearen Regressionsfunktion darstellen.
Das nun beschriebene Annahmensystem besteht aus verschiedenen Chrakterisierungen der Störvariablen

1. $E(U_i) = 0$ für alle i
   Die Erwartungwerte der n Störvariablen sind gleich Null
2. $\text{var}(U_i) = \sigma_U{}^2$ für alle i
   Die Varianzen der n Störvariablen sind gleich groß (Homoskedastizität)
3. $\text{cov}(U_i, U_j) = \begin{cases} 0 \text{ für i} \neq \text{j, i, j} = 1....\text{n} \\ \sigma^2_{\text{U}} \text{ für i} = \text{j, i, j} = 1....\text{n} \end{cases}$
   Die n Störvariablen sind unkorreliert, die Kovarianz cov(Ui, Uj) ist für alle Paare der Störvariablen Null, falls i≠j.
4. Ui folgt $N(0;\sigma_U)$ für alle i
   (nur für bestimmte Verfahren wichtig, für Methode der kleinsten Quadarate entbehrlich)

1.3 Die Regressionsgleichung

Die Regressionsgleichung der Stichprobe ergibt sich durch die Gleichung: $y_i = a_i + bx_i + d_i$, wobei d_i die Summe der Schätzfehler, d.h. die Summe der Differenzen zwischen y_i und $a + bx_i$, ist. Der Schätzfehler heißt auch Residuum, die Summe Residuen.
Diese Gleichung zur exakten Beschreibung ist (leider) nicht linear, daher benötigt man als exakte Beschreibung die Gleichung der Regressionsgerade $\hat{y}_i$:
Die Gleichung der Schätzgerade $\hat{y}_i$ lautet: $\hat{y}_i = a_i + bx_i$
Um die beste Regressionsgerade zu bestimmen

a) soll die Summe der Schätzfehler 0 sein, d.h. die einzelnen Fehler sollen sich aufheben, d.h. die Gerade muß durch $\bar{x}$ und $\bar{y}$ laufen

b) die Zahl der Schätzfehler muß minimal sein

1.4 Methode der kleinsten Quadrate für eine einfache Regressionsgleichung

Um die Parameter a und b einer Regressionsgraden so zu bestimmen, daß die Gerade den beobachteten Wertepaaren optimal angepaßt ist, muß die Summe der quadrierten Abweichungen der beobachteten Y_i von den rechnerischen Y_t ein Minimum ergeben. D.h. die Regressionsgerade ist dann optimal berechnet, wenn die Summe der Abweichungsquadrate minimal ist.

$$z(a,b) = \sum_{i=1}^{n} d_i{}^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 \quad \rightarrow \text{Minimum}$$

Durch partielle Ableitung und Nullsetzen dieser Ableitungen ergeben sich die Normalgleichungen zur Bestimmung der Koeffizienten einer linearen Kleinste-Quadrate-Reressionsfunktions. Löst man das System der Normalgleichungen nach a und b auf, erhält man die Regressionskoeffizienten *a* und *b:*
Für eine **einfache Regressionsgleichung** ergeben sich die Regressionskoeffizienten:

$$b = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i{}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \bar{y} - b\bar{x}$$

1.5 Bedeutung der Regressionsfunktionsbestandteile

Eine univariate Regressionsfunktion hat die Funktion $y_i + a + bx_i$
Dabei bedeuten:

x_i Werte auf der X-Achse

y_i Werte auf der Regressionsfunktion
Man nennt diese Werte auch zu erwartende oder theoretische Werte, weil diese Y-Werte in Abhängigkeit von Veränderungen der Variablen X zu erwarten wären, wenn die Regressionslinie den Zusammenhang zwischen X und Y korrekt wiederspiegelt. Insoweit kommt in der Regressionsfunktion selbst eine Hypothese über den vermuteten Zusammenhang zwischen X und Y zum Ausdruck.
a Ordinatenabschnitt der linearen Funktion
b Steigung (= Tangens des Steigungswinkels) der Funktion
Die Koeffzienten a und b spezifizieren den deterministischen Teil des Zusammenhangs und stellen die wahren Parameter für die gesamte Population her.

1.6 Beziehung zwischen Regressionskoeffizient $r^2 = R^2$ und $\mathrm{var}(\hat{y})$ und $\mathrm{var}(y)$

Zwischen $r^2 = R^2$ und $\mathrm{var}(\hat{y})$ und $\mathrm{var}(y)$ bestehen verschiedene Beziehungen:

$R^2 = 0$ wenn beide Merkmalswerte unkorreliert sind

$R^2 = 1$ wenn das Streuungsdiagramm auf einer Geraden mit positiver oder negativer Steigung liegt

Je größer R^2, desto stärker werden die empirischen Y-Werte durch die theoretischen y-Werte bestimmt/determiniert.

2 Bivariate Regressionsrechnung

Die bivariate Regressionsrechnung will die Beziehung der drei Merkmal X_1, X_2 und Y klären. Y ist die Variable, die erklärt werden soll, hängt also statstistisch von X_1 und X_2 ab.
Gesucht ist eine Gleichung für die Geraden durch diesen dreidimensionalen Raum. Diese Gleichung ermöglicht es wie bei der einfachen Regressionsrechnung, die Tendenz der Abhängigkeit zwischen Y und X_1 und X_2 soll durch eine lineare Funktion der Art $\hat{y}_i = b_o + b_1 x_{1i} + b_2 b_{2i} + d_i$ bestimmt werden, wobei:

b_o: Regressionskonstante

b_1, b_2: (partielle) Regressionskoeffizienten

d_i: Differenz zwischen y_i und $b_o + b_1 x_{1i} + b_2 b_{2i}$

Um die Regressionskoeffizienten zu bestimmen, wendet man die ***Methode der kleinsten Quadrate*** an:

2.1 Methode der kleinsten Quadrate für die bivariate Regressionsgleichung

Durch Nullsetzen der partiellen Ableitungen erhält man ein System von Normalgleichungen, die ein lineares Gleichungssystem mit drei Unbekannten. Löst man dieses System auf, ergeben sich die folgenden die Regressionskoeffizienten: [→ *Schwarze*, S. 159]

$$b_o = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{b}_2$$

$$b_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2} \cdot \frac{\sqrt{\mathrm{var}(y)}}{\sqrt{\mathrm{var}(x_1)}}$$

$$b_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2} \cdot \frac{\sqrt{\mathrm{var}(y)}}{\sqrt{\mathrm{var}(x_2)}}$$

r_{y1} Korrelationskoeffizient für Y und X_1

r_{y2} Korrelationskoeffizeint für Y und X_2

r_{12} Korrelationskoeffizient für X_1 und X_2

2.2 Beispiel für eine bivariate Regressionsgleichung

Bei einer statistischen Erhebung in den USA wurden von 20 Bauernhöfen die bewirtschaftete Fläche X_2 (in 10 acres; 1 acre=040467 ha), die Anzahl der unterhaltenen Milchkühe X_2 und das erzielte Jahreseinkommen Y (in 10 Dollar) ermittelt. Die Daten stammen aus den späten 20er Jahren. Man erhielt folgendes Ergebnis:

X_1	18	0	14	6	1	9	6	12	7	2	17	15	7	0	12	16	2	6	12	15
X_2	6	22	18	8	12	10	17	11	16	23	7	12	24	16	9	11	22	11	16	8

y	96	83	126	61	59	90	82	88	86	76	102	108	96	70	80	113	76	74	98	80

X_1: Fläche
X_2: Milchkühe
Y: Jahreseinkommen
Der vermutete Zusammenhang zwischen Fläche und Anzahl der Kühe auf der einen Seite, und dem Jahreseinkommen auf der anderen Seite soll durch die Regressionsgleichung zum Ausdruck gebracht werden.
Ich berechne die entsprechenden Korrelationskoeffizienten (siehe univariate Regressionsrechnung):

$r_{yi} = 0{,}708231402$

$r_{y2} = 0{,}007803807$

$r_{12} = -0{,}61508498$

Für b_1 ergibt sich $b_1 = \frac{0{,}708231402 - 0{,}007803807 \cdot -0{,}61508498}{1 - -0{,}61508498^2} \cdot \frac{\sqrt{287{,}1157895}}{\sqrt{31{,}9447}} =$

usw.
entsprechend der oben angegebenen Formel berechnet man die anderen Elemente.

2.3 Korrelationskoeffizient nach Bravais-Pearson

Die Korrelationsrechnung dient dazu, die Stärke des Zusammenhangs zwischen zwei Untersuchungsvariablen in einer einzigen statistischen Maßzahl zum Ausdruck zu bringen. *r* ist eine dimensionslose Größe
Voraussetzung für die Anwendung des Korrelationskoeffizienten von Bravais-Pearson sind mindestens *intervallskalierte Daten.*

$$r = \frac{\operatorname{cov}(x, y)}{+\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{1}{n}\sum_{1=1}^{n}(y_i - \bar{y})^2}} \qquad -1 \le r \le 1$$

$\operatorname{cov}(x, y)$	Kovarianz zwischen X und Y
$\operatorname{var}(x)$	Varianz von X
$\operatorname{var}(y)$	Varianz von Y
$\bar{x}$	arithmetisches Mittel von X
$\bar{y}$	arithmetisches Mittel von Y
n	Anzahl von (yi, xi);Anzahl der statistischen Einheiten

2.3.1 Interpretation von r

Der Korrelationskoeffizient von Bravais-Pearson nimmt nur Werte zwischen -1 und +1 an.Wertebereich von -1 bis +1:
r=-1 maximaler reziproker Zusammenhang, d.h. mit sehr hoher Wahrscheinlichkeit nehmen die Y-Werte tendenziell ab, wenn die Werte der Variablen X zunehmen
r=0 kein Zusammenhang zwischen X und Y
r=+1 maximaler gleichgerichteter Zusammenhang, d.h. mit sehr hoher Wahrscheinlichkeit nehmen die Werte der Variablen Y tendenziell zu, wenn die X-Werte zunehmen.

2.3.2 Anmerkungen zum Korrelationskoeffizienten *r*

- in der Praxis taucht ein Wert für r größer 0,5 nur selten auf, man betrachtet ein *r* zwischen 0,3 und 0,5 als ein Indiz für einen starken Zusammenhang
- je größer die Zahl der Merkmalsträger, desto aussagekräftiger ist *r*
- die Treffsicherheit von *Prognosen* ist um um so höher, je größer *r* ist, d.h. je stärker der Zusammenhang zwischen zwei Variablen X und Y ist und je größer N ist.
- die Interpretation des Korrelationskoeffizienten muß immer auf dem Hintergrund einer *linearen Regressionsfunktion* erfolgen. Wäre in einem konkreten Fall eine nichtlineare Funktion

angemessen, dann könnte sich beispielsweise ein r-Wert nahe bei 0 ergeben, weil gleichwohl eine lineare Funktion unterstellt wird.
- Die Prüfung, ob eine nichtlineare Funktion zugrunde gelegt werden muß, kann z.B. graphisch oder durch eine Clusteranalyse erfolgen.

2.4 Determinationskoeffizient

r^2 heißt Determinationskoeffizient oder Bestimmtheitsmaß. r^2 gibt an, welcher Anteil der Streung von Y durch die Regressionsgerade „bestimmt" oder „erklärt" werden kann. Anders ausgedrückt: Der Determinationskoeffizient gibt an, wie groß der Anteil der Varianz der Untersuchungsvariablen ist, der sich auf die Variation der einen exogenen Variablen zurückführen läßt.
Der Determinationskoeffizient hat seinen Namen daher, daß er denjenigen Anteil an der Varianz der Y-Werte angibt, der durch die Variation der X-Werte determiniert wird. Dies geht auf das Prinzip der Varianzzerlegung zurück.

2.4.1 Prinzip der Varianzzerlegung

In jedem konkreten Anwendungsbeispiel kann man davon ausgehen, daß die Y-Werte streuen. Diese Streuung kann mit der *Varianz* (quadrierte Standardabweichung) gemessen werden. Die Aufgabe der Regressionsrechnung kann man auch so erklären, daß man fordert, eine Variable (X) zu finden, die die interessierende abhängige Variable (Y) beeinflußt und in diesem Sinne „statistisch erklärt".
„Erklären" bedeutet hier, daß die Veränderungen der Variablen statistisch zurückgeführt werden auf Veränderungen der Variable X. Das bedeutet aber weiterhin, daß ein mehr oder weniger großer Teil der Varianz von Y dadurch statistische erklärt wird, daß die Variation der Variablen X als statistischer Erklkärungsgrund angenommen wird.
Formal sieht das folgendermaßen aus:
Die Variation der Abhängigen Y (Var(y)) läßt sich in zwei Teile zerlegen:
- 1. Teil: $\mathrm{var}(y_t)$**:** Variation der zu erwartenden (theoretischen) Y-Werte, die auf der Regressionsgeraden liegen
- 2. Teil: $\mathrm{var}(y_r)$ Reststreuung, d.h. Variation der Y-Werte um die Regressionsgerade herum

  y_r =Restwerte= $y - y_t$)

Es besteht also folgender Zusammenhang:

$$\mathrm{var}(y) = \mathrm{var}(y_t) + \mathrm{var}(y_r)$$

Rechnet man diese Varianzen aus, stellt man fest, daß der prozentuale Anteil von $\mathrm{var}(y_t)$ an der Gesamtvarianz var(y) mit dem numerischen Wert des Determinationskoeffizienten übereinstimmt.
- Je höher der Wert des Determinationskoeffizienten ist (d.h. je stärker der Zusammenhang zwischen *x* und *y*), desto kleiner ist die Restreuung, weil sich die Punkte je in diesem Fall sehr eng um die Regressionsgerade herum streuen, desto höher ist der Anteil von var(y_t) an der Gesamtstreuung.

2.5 Rangkorrelationskoeffizient nach Spearman

$$r_{sp} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

$$d_i = Rg(x_i) - Rg(y_i)$$

$$-1 \le r_{SP} \le 1$$

Voraussetzung: ordinalskalierte Daten
Der Rangkorrelationskoeffizient beruht nicht auf den direkten Merkmalsausprägungen x_i bzw y_i, sondern auf den zugeordneten Rangnummern $Rg(x_i)$. Der Rangkorrelationskoeffizient von Spearman ist der auf diese Rangnummern $Rg(x_i)$ angewandte Bravais-Pearson-Korrelationskoeffizient, aus diesem Grunde ist auch der Wertebereich für r_{SP} mit dem von *r* identisch! Anders ausgedrückt ergibt sich r_{sp} aus r, wenn man dort die X- und Y-Werte durch deren Rangplätze ersetzt. Nach einigen Umformungen ergibt sich die obige Formel.

Vorgehensweise: Die Daten müssen der Größe nach sortiert sein, erst danach werden die Ränge vergeben. Haben mehrere Merkmalsträger den gleichen Rang inne, erhalten sie den gleichen (gemittelten) Rangplatz, die Rangplätze davor und danach bleiben entsprechend leer.
Entsprechend der Formel subtrahiere ich den jeweiligen Rang y_i von x_i, quadriere das Ergebnis und addiere alle Ergebnisse für x_i, i=1…n usw.

3 Multiple lineare Regression

Die multiple Regressionsanalyse ist ein Instrument zur Untersuchung des funktionalen Zusammenhangs zwischen einem quantitativem Merkmal x mit Ausprägungen y und Merkmalen x_1 …x_i

Die multiple Regressionsrechnung hat die Aufgabe, den Zusammenhang zwischen mehr als zwei Variablen zu beschreiben und damit zu prognostischen Aussagen für eine als abhängig angesehene Variable Y zu gelangen, von der unterstellt wird, daß sie nicht nur von X_1, sondern auch von X_2 (und eventuell weiteren Variablen) abhängt.

X ist also die unabhängige, erklärende Variablengruppe

Y ist die Variablengruppe der abhängigen Variablen (bei der univariaten multiplen Regressionsrechnung enthält Y nur eine Variable)

3.1 Schätzfunktion

Entsprechend dem Regressionsmodell der einfachen linearen Regression (vgl. „**Das Modell der einfachen linearen Regression**“, S. 2) lautet die Modellfunktion (also die exakte Beschreibung der Stichprobe) für das multiple Funktion:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + u_i, \quad \text{i} = 1\ldots\text{n},\ \beta_j \neq 0 \text{ für alle j}$$

Wiederum wendet man die „Methode der der kleinsten Quadrate“ an und erhalten die Regressionsgleichung, die hier **Regressionshyperebene** genannt wird.

$$y_i = b_0 + \sum_{j=1}^{k} b_j k_{ij} + d_i$$

b_o Regressionskonstante

$b_j,\ \text{j} = 1\ldots\text{k}$ (partielle) Regressionskoeffizienten

$$\hat{\text{y}}_1 = b_o + \sum_{j=1}^{k} b_j x_{ji} \qquad \text{i} = 1\ldots\text{n},\ \text{b}_\text{j} \neq 0,\ \text{j} = 1\ldots\text{k} \qquad \text{Regressionshyperebene}$$

Die Regressionskoeffizienten lauten ausgeschrieben (Matrixrechnung)

$$\underline{b} = (\underline{x}'\,\underline{x})^{-1}\,\underline{x'y}$$

$$\underline{b} = \begin{pmatrix} \mathbf{b_0} \\ . \\ . \\ \mathbf{b}_k \end{pmatrix} \qquad \underline{\text{x}} = \begin{pmatrix} \mathbf{1}\ \text{x}_{11} \ \ldots\ \text{x}_{\text{k1}} \\ . \ . \qquad . \\ . \ . \qquad . \\ \mathbf{1}\ \text{x}_{1\text{n}}\ \ldots\ \text{x}_{\text{kn}} \end{pmatrix} \qquad \underline{\text{y}} = \begin{pmatrix} \mathbf{y_1} \\ . \\ . \\ y_n \end{pmatrix}$$

$\underline{b}$ Vektor der Regressionskoeffizienten

$\underline{\text{x}}$ Matrix der Werte der exogenen Merkmale Xj, $\text{j} = !\ldots\text{k}$. Die Werte der Scheinvariablen X_0, die zur Regressionskonstanten b_o gehört, sind alle 1.

$\underline{\text{y}}$ Vektor der Werte von Y

$\underline{\text{x}}'$ transponierte Matrix von $\underline{\text{x}}$

$\underline{x}$ muß nichtsingulä r sein, damit die Inverse $(\underline{\text{x}}', \underline{x})^{-1}$ existiert.

4 Bedeutung der Regressionskoeffizienten

b_0 gibt an, welcher Wert für die deterministische Komponente der Untersuchungsvariablen zu erwarten ist, falls sämtliche exogenen Variablen den Wert Null realisieren.

Geometrisch gibt b_0 die Schnittebene der Hyperebene mit der Y-Achse an.

→Abbildung vgl. Tiede S. 171

Die übrigen Koeffizienten b_j, *j=1...k* geben (bei Beachtung der jeweiligen Vorzeichen) den (positiven oder negativen) Beitrag an, um den sich der Schätzwert für die deterministische Komponente $\hat{y}_i$ des Wertes y_i, falls der Wert x_{ji} der j-ten exogenen Variablen um eine Einheit erhöht.
b_{ij} besitzen jeweils eine Dimension: b_j wird in Einheiten von Y pro Einheit von X_j gemessen. Da die Xj recht unterschiedliche Maßstäbe haben können, kann man meistens durch einen Vergleich der Regressionskoeffizienten keinen Hinweis auf die Bedeutung der Variation der Einflußgrößen für die Variation der Untersuchungsvariablen geben.

4.1 Multiple und partielle Koeffizienten

4.1.1 Multipler Determinationskoeffizient

Der multiple Determinationskoeffizient ist so konstruiert wie der Determinationskoeffizient des einfachen Modells, die Varianzzerlegungsformel gilt auch für den multiplen Fall, d.h. di eStreuung der Untersuchungsvariablen läßt sich zerlegen:

a) in die Streuung der geschätzten deterministischen Komponente, den statistisch erklärten Teil

b) in die Streuung der stochastischen Komponente, den durch die Variaton nicht erklärten Teil

Der **multiple Determinationskoeffizient** (*multiples Bestimmtheitsmaß*) gibt an, wie groß der Anteil der Varianz der geschätzten deterministischen Komponente an der gesamten Varianz der Untersuchungsvariablen ist. Die Aussagekraft des multiplen Determinationskoeffizienten wird häufig überschätzt. Der multiple Determinationskoeffizient ist ein globales Maß und für die Beschreibung der spezifischen Einwirkungen der einzelnen exogenen Variablen auf die Untersuchungsvariable ungeeignet.
Die Formel lautet:

$$r_{Y.12..k^2} = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)} = 1 - \frac{\mathrm{var}(d)}{\mathrm{var}(y)} = \sum_{j=1}^{k} b_j^{*2} + 2\sum_{j=1}^{k-1}\sum_{l=j+1}^{k} b_j * b_l * r_{jl} = \frac{[\mathrm{cov}(y,\hat{y})]^2}{\mathrm{var}(y)\,\mathrm{var}(\hat{y})}$$

$$b_j^* = b_j \frac{\sqrt{\mathrm{var}(x_j)}}{\sqrt{\mathrm{var}(y)}} \qquad j = 1...k$$

r_{jl} Korrelationskoeffizient für X_j und X_{j+1}, $j = 1...k-1$

4.1.2 Multipler Korrelationskoeffizient

Die Wurzel des multiplen Determinationskoeffizienten wird als **multipler Korrelationskoeffizient** bezeichnet: $+\sqrt{r_{Y.12...k^2}}$

- Das Vorzeichen ist bedeutungslos, da es in Hinblick auf eine exogene Variable positiv und eine andere negativ sein kann.

Der Korrelationskoeffizient gibt Antwort auf die folgende
Frage: Wie verbessern sich Prognosen der interessierenden Variable Y, wenn sie nicht allein abhängig von X, sondern auch zugleich als abhängig von der Drittavariablen Z angesehen wird?
Antwort: In diesem Fall versucht man, zusätzliche Informationen zur Verbesserung von Vorhersagen zu nutzen und verwendet in diesem Zusammenhang den sogenannten multiplen Korrelationskoeffizienten

4.1.3 Multipler Regressionskoeffizient

4.2 Partielle Regressionsrechnung

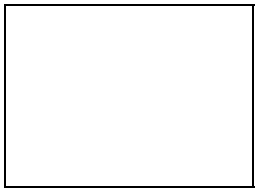
Die partielle Regressionsrechnung hat die Aufgabe, den Zusammenhang zwischen zwei interessierenden Variablen X und Y um den eventuellen Einfluß einer dritten Variable Z (oder weiterer Variablen) zu „bereinigen“. Wenn beispielsweise X mit Y korrelliert, aber sowohl X als auch Y ihrererseits mit Z korrelieren, dann ist die hohe Korrelation zwischen X und Y eine matematisch notwendige Folge des gemeinsamen Einflusses von Z. Das „Heraus-Partialisieren“ von Z zeigt dann den Zusammenhang zwischen X und Y, der übrig bleibt, wenn der gemeinsame Einfluß von Z eliminiert wird.

4.2.1 Partieller Determinationskoeffizient

$$r_{[yj]}^{\ 2} = \frac{\mathrm{cov}^2(d_{[y]}, d_{[j]})}{\mathrm{var}(d_{[y]}\,\mathrm{var}(d_{[j]})} = j = 1....k$$ [→ Tiede. S.181]

Der partielle Determinationskoeffizient ist analog zum Determinationskoeffizient konstruiert. Er gibt den Varianzanteil der (nach Erklärung durch die übrigen exogenen Variablen verbliebenen) Restvarianz der Untersuchungsvariablen Y an, der von der (um den Einflllluß der übrigen exogenen Variablen bereinigten) Varianz der exogenen Variablen X_j herrührt.

4.2.2 Partieller Korrelationskoeffizient



Der partielle Korrelationskoeffizient entsteht aus der Wurzel des partiellen Determinationskoeffizienten. Er beantwortet die
Frage: Wie stellt sich der Zusammenhang zwischen X und Y dar, wenn der etwaige gemeinsame Einfluß einer dritten Variable Z „ausgeschaltet" wird?
Antwort: Der Einfluß der Variablen Z wird auspartialisiert, sein Einfluß „ausgeschaltet". Man berechnet den *partiellen Korrelationskoeffizienten* zwischen X und Y - unter Ausschaltung von Z.

5 Faktorenanalyse

In der empirischen Realität gibt es meist Zusammenhänge nicht nur zwischen zwei Untersuchungsvariablen, sondern sehr viele, möglicherweise alle Untersuchungsvariablen hängen zusammen. Die Faktorenanalyse versucht, die hohe Dimensionalität des Untersuchungsraums zu reduzieren. Die Faktorenanalyse ist also ein datenreduzierendes Verfahren. Sie ermöglicht es ohne entscheidenden Informationsverlust, viele wechselseitig hoch korrelierende Variablen durch wenige voneinander unabhängige Faktoren zu ersetzen.

Graphische Interpretation:
Jeder Merkmalsträger kann als ein Punkt in einem hochdimensionierten Achsenkreuz präsentiert werden, dessen Dimensionalität von der Anzahl der Untersuchungsvariablen bestimmt wird. Die Faktorenanalyse beantwortet somit die Frage, ob es eine deutlich geringere Anzahl von Faktoren gibt, die die Zusammenhänge zwischen allen Untersuchungsvariablen weitgehend zu erklären in der Lage sind. [*Voss, S. 164*].

Faktorenanalyse als heuristisches hypothesengenerierendes Verfahren
Die Faktorenanalyse ist ein heuristisches hypothesengenerierendes Verfahren. D.h. es muß ein Ordnungssystem erstellt werden, das mit den thoeretischen Kontexten der unterstellten Variablen am besten zu vereinbaren ist. Dann werden Hypothesen über Strukturen formuliert, die den untersuchten Merkmalen zugrunde liegen.
Voraussetzung für die Faktorenanalyse ist also, daß Korrelationen zwischen einzelnen Merkmalen oder Merkmalsgruppen bestehen. Je höher die Korrelation zwischen den Beobachtungsmerkmalen, umso besser lassen sich die resultierenden Faktoren erklären. Dabei versucht die Faktorenanalyse die einfachste Struktur zu finden, die die Ausgangsdaten möglichst genau wiedergibt und erklärt. Die Forderungen einer möglichst guten Abbildung der Beobachtungsdaten einerseits und der möglichst geringen Zahl von Faktoren andererseits stehen in Konkurrenz zueinander. Das führt dazu, daß Ergebnisse der Faktorenanalyse von subjektiven Aspekten abhängen können. Das gilt insbesondere für die Anzahl der gewählten Faktoren und deren Interpretation. Durch verschiedene Bedingungen, die an die Faktoren gestellt werden, resultieren eine Menge verschiedener Verfahren.

Ein wesentlicher Grundgedanke der Faktoranalyse besteht darin, den Merkmalsträgern (Beobachtungen pro Untersuchungseinheit) Faktorwerte (f) und den Variablen Ladungskoeffizienten (Korellationskoeffizienten vor ihrer Standardisierung) so zuzuordnen, daß aus dieser Gleichung Z Schätzwerte errechnet werden können. Aus diesen wird eine Korrelationskoeffizientenmatrix R erstellt, die möglichst gut mit der empirischen Korrelationskoeffizientenmatrix R übereinstimmen soll.

5.1 *Rechenweg*

Äußerlich sieht dieses Modell aus wie ein System von m multiplen Regressionsmodellen. Der entscheidende Unterschied besteht darin, daß die Einflußgrößen bei den Regressionsmodellen vorgegeben und mit der eigentlichen Variablen zusammen explizit gemessen werden, wohingegen die Faktoren hypothetische Konstrukte sind, die aus der standardisierten Datenmatrix Z herausgerechnet (extrahiert) werden sollen.

Ausgangspunkt einer Faktorenanalyse ist eine **empirische (quantitative) Datenmatrix Y**. Die Datenmatrix enthält die Merkamslwerte der interessierenden Merkmal Y_i, die am i-ten Objekt beobachtet wurden.

Aus der empiritschen Datenmatrix Y wird die **standardisierte Datenmatrix** $\hat{Z}_{ij}$ berechnet. (Dazu standardisiert man die Matrix Y so, daß der Mittelwert jeder Spalte Null und die empirische Varianz jeder Spalte Eins ist).

Die Faktorenanalyse geht nun davon aus, daß sich die korrelierten, beobachteteten Merkmale als Linearkombination von unbekannten nichtbeobachteten Faktoren darstellen lassen. Jedes Element der standardisierten Datenmatrix läßt sich als Linearkombination von Realisationen der unbekannten Faktoren beschreiben.
Das heißt: Die Matrix yst ist darstellbar als Produkt zweier Matrizen, $\hat{Z}_{ij}$ ergibt sich durch Multiplikation der Ladungsmatrix mit der Matrix der Faktorenwerte.

$$\hat{Z}_{ij} = \sum_{k-1}^{1} a_{jk} \cdot i_k$$

a_{jk} = unbekannte Faktorenmatrix

i_k = bekannte Ladungsmatrix (erklärter Varianzanteil)

Zur Lösung dieser Gleichung sucht man (bzw. SPSS) Werte für die geschätzte Faktorenmatrix, bis die obige Gleichung erfüllt werden kann (Das ist deshalb so schwierig und aufwendig, weil es sich um Matritzen handelt!).

5.2 Interpretation

5.3 Begriffe aus der Faktorenanalyse

Ladungsmatrix

Die **Ladungsmatrix** heißt auch Matrix der Faktorladungen. Die Koeffizienten der Ladungsmatrix beschreiben die Ladungen des k-ten nichtbeobachteten Faktors bezüglich des j-ten beobachteten Merkmals. Eine Faktorladung a_{ij} entspricht der Korrelation zwischen einer Variablen *i* mit einem Faktor *j*. Die Ladungsmatrix beschreibt den Zusammenhang zwischen Merkmalen und Faktoren.

Matrix der Faktorenwerte

beschreibt die n beobachteten Objekte bezüglich der Faktore. Die Faktorenmatrix beschreibt den Zusammenhang zwischen Faktoren und Objekten.

Die standardisierte Datenmatrix beschreibt den Zusammenhang zwischen Merkmalen und Objekten.
Die Faktorenmatrix beschreibt den Zusammenhang zwischen Faktoren und Objekten.
Die Ladungsmatrix beschreibt den Zusammenhang zwischen Merkmalen und Faktoren.

Ladungsmuster

Ein wesentlicher Grundgedanke der Faktorenanalyse besteht darin, den Merkmalsträgern (Beobachtungen pro Untersuchungseinheit) Faktorwerte (f) und den Variablen Ladungskoeffizienten (Korrelationskoeffizienten, wenn diese vorher standardisiert worden sind) so zuzuorden, daß Z Schätzwerte errechnet werden können, und aus diesen eine Korrelationskoeffizientenmatrix R-Schätzwert, die möglichst gut mit der epirischen Korrelationskoeffizientenmatrix R übereinstimmen soll. Die Korrelation zwischen Z und einem Faktor F beruht im Wesentlichen auf der transponierten Matrix von Faktorwerten, denn die Zielfunktion lautet: $Z = F \cdot A' = fm_2 ai_2 .... fm_n ai_q$. Aus Z wird $R = A \cdot A'$ abgeleitet (Korrelationskoeffizientenmatrix). Diese Umformung verdeutlicht, wie wichtig die Ladungsmuster für die Bestimmung der Faktorwerte sind.

Ladungskoeffizient (=Faktorladung)

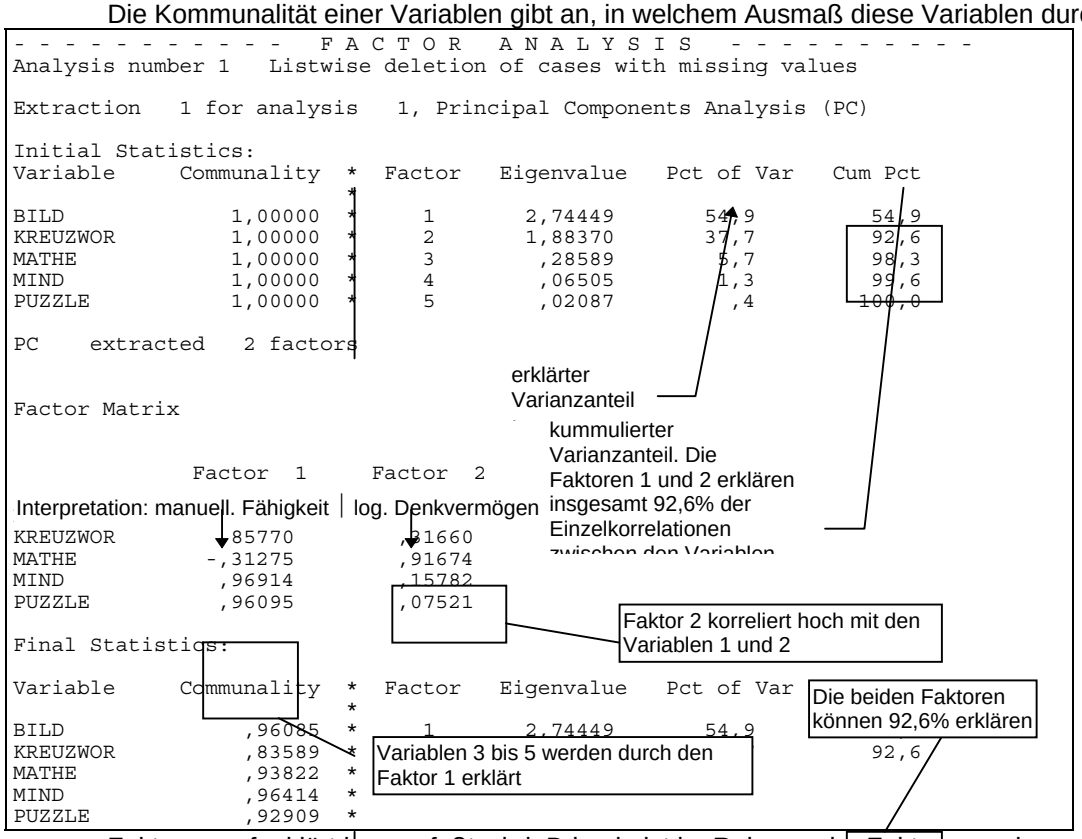
Eine Faktorladung a_{ij} entspricht der Korrelation zwischen einer Variablen *i* mit einem Faktor *j*

a_{ij}^2 als erklärter Varianzanteil entspricht dem Determinationskoeffizienten

In der Faktorenanalyse entscheidet sich die Bedeutung der Faktoren aufgrund der Faktorladung. Sie ist letztlich ausschlaggebend für die Wertigkeit des Faktors.
Die Koeffizienten a_{ij} der gemeinsamen Faktoren und die Koeffizienten d_j der spezifischen Faktoren werden als Faktorladungen bezeichnet.

Kommunalität einer Variablen

Die Kommunalität einer Variablen gibt an, in welchem Ausmaß diese Variablen durch die



Faktoren aufgeklärt bzw. erfaßt wird. D.h. sie ist im Rahmen der Faktorenanalyse ein Maß für den Grad des Zusammenhangs einer Variablen mit allen anderen Variablen, statistisch gesehen erklärt die Kommunalität den Anteil der gemeinsamen Varianz. Jede Variable hat eine spezifische Kommunalität.

Die Varianz einer standardisierten Variablen ist immer 1. Die Kommunalität muß folglich kleiner 1 sein, sollte aber möglichst gegen 1 tendieren: Weicht sie stark von 1 ab, kann man annehmen, daß die Faktoren schlecht gewählt worden sind.

Eigenwert

Eigenwerte spielen bei der Faktorenanalyse die quasi entscheidende Rolle: Sie werden vor der Faktorrotation berechnet und dienen zumeist als Kriterium für die Entscheidung, ob Faktoren im faktorenanalytischem Modell beibehalten oder weggelassen werden.

Der Eigenwert λ_j eines Faktors j gibt an, wieviel von der Gesamtvarianz aller Variablen durch diesen Faktor erfaßt wird. Ist ein Eigenwert kleiner als 1, erklärt er also weniger als die Varianz einer einzigen Variablen, wird der entsprechende Faktor für unbedeutend erklärt.

Die Eigenwertbestimmung der Faktoren dient also dazu, unwichtige Faktoren zu eleminieren.

5.4 Faktorenanalyse mit SPSS

1. Dateneingabe

Die Variablen werden definiert <Data><Define Variable> und die Werte eingegeben

2. Faktorenanalyse durchführen

Durch <Statistics><Data Reduction> <Factor> erhält man das Auswahlmenü für die Faktorenanalyse. Man wählt diejenigen Variablen aus, die in die Berechnung einfließen sollen, stellt ggf Optionen ein und bestätigt mit <OK>

3. Ausgabe

SPSS wirft folgende Output-Datei aus:

6 Clusteranalyse

Wenn es Zusammenhänge zwischen einer größeren Zahl von Untersuchungsvariablen gibt, werden sich die Merkmalsträger in einem hochdimensionalen Achsenkreuz in bestimmter Weise „klumpen“. Dies „Klumpen“ (engl.: Cluster) zu isolieren und auf der Grundlage der eventuellen Isolationserfolge

dann zu inhaltlichen Interpretationen der beobachteten Zusammenhänge zu gelangen, ist Aufgabe der *Cluster-Analyse*.

Die Clusteranalyse teilt also viele, multivariate und durch einen festen Satz von Merkmalen beschriebene Untersuchungsobjekte nach Maßgabe ihrer Ähnlichkeit in homogene Gruppen oder Cluster ein, die allerdings extern möglichst gut voneinander seperierbar sein sollen. Die Ähnlichkeit bzw. Unähnlichkeit hängt von den Merkmalen der Objekte ab, diese müssen durch sorgfältige inhaltliche Überlegungen begründet werden.

Methodisch gesehen mißt die Clusteranalyse Abstände zwischen Merkmalsträger. Wertepaare, die in geringem Abstand vorkommen, werden in gemeinsame Klumpen aufgenommen. Das verwendete Distanzmaß ist die euklidische Distanz, die nach dem Satz des Pythagoras berechnet wird. Voraussetzung: Die Variablen müssen unabhängig sein, bei Korrelation kommt es zu Problemen mit dem Distanzmaß.

6.1.1 Unterschied Varianzanalyse - F-Test

Beim F-Test Test werden die Varianzen zweier Grundgesamtheiten bewertet. Stammen sie aus einer GG? Gibt es einen signifikanten Unterschied?

Varianzanalyse: Stammen zwei oder mehrere Mittelwerte aus der gleichen Grundgesamtheit?

7 Begriffe Statistik C

Chi-Quadrat-Verteilung
Chi-Quadrat-Test
Dichtefunktion

- Exogene Variablen

Faktorenanalyse
Faktorladung
Freiheitsgerade
F-Verteilung
Gamma-Funktion
Gamma-Verteilung
Interaktion
Interaktion bei der zweifaktoriellen Varianzanalyse
Konfidenzbereich
Ladung
Ladungskoeffizient
Linearitätshypothese

- Methode der kleinsten Quadrate

Momentenmethode

- Multiple Regressionsrechnung
- Multiples Regressionsmodell
- Rangkorrelationskoeffizient
- Rangkorrelationskoeffizienten von Spearman
- Regressionskoeffizienten
- Regressionskoeffizient, partieller

Streuung der Standardabweichung
Student-Verteilung/t-Verteilung
Varianzanalyse zweifacher Klassifikation
Varianzquotiententest
Wechselwirkungen
Zufallsvariable