Theorie der stochastischen Abhängigkeit

0 Einleitung

Bei der wissenschaftlichen Arbeit ist sehr oft die Lage gegeben, dass nicht nur eine einzige Veränderliche vorliegt; zwei oder mehrere Variable sind zu untersuchen, wobei sie nicht getrennt, sondern gleichzeitig beobachtet werden. Es ist meistens so, dass eine Variable (die sogenannte abhängige Variable) von einer oder sogar mehreren Variablen (den unabhängigen Variablen) abhängig ist. So kann es sein, dass die Kosten pro Pflegetag (abhängige Variable) von der Produktivität (unabhängige Variable) beeinflusst werden. Die betriebswirtschaftliche Theorie zeigt ferner, dass die Einheitskosten nicht nur von der Produktivität, sondern noch von anderen Einflussfaktoren abhängig sind, wie z.B. den Marktpreisen und dem Beschäftigungsgrad.
Generell kann also argumentiert werden, dass eine Variable (z.B. Kosten) als Funktion von einer oder mehreren anderen Variablen (z.B. Produktivität, Marktpreise) betrachtet werden kann. Welches ist die praktische Bedeutung dieser Erkenntnis? Nun, man wird unter anderem daran interessiert sein, für bekannte Werte der unabhängigen Variablen fehlende Werte der abhängigen Variablen zu berechnen (Interpolation, Extrapolation). Ferner wird es möglich sein, die verwendete Theorie mit Hilfe der Statistik zu überprüfen, zu verifizieren.

1 Theorie der stochastischen Abhängigkeit

11 Grundsätzliches

111 Funktionale und stochastische Abhängigkeit

Vorerst soll das Wesen der stochastischen Abhängigkeit kurz erläutert werden. Wir gehen aus vom **Begriff der Funktion.** Nach Dirichlet heisst y eine Funktion von x, d.h. $y = f(x)$, wenn irgendwelchen Werten von x eindeutig y-Werte zugeordnet sind. Die Zuordnung kann u.a. folgendermassen erfolgen:

- Mit **Rechenvorschrift,** und zwar explizit, d.h. $y = f(x)$, oder implizit, d.h. $f(x,y) = 0$.
- Durch eine **Tabelle.**
- Durch eine **grafische Darstellung.**

R.G.D. Allen (Mathematik ..., Berlin 1956, S.29) umschreibt folgendermassen: «Der Begriff der Funktion umschliesst daher die Begriffe der **Beziehung** zwischen den Werten zweier Veränderlicher und der **Abhängigkeit** einer Veränderlichen von der anderen».
Nun hat sich gerade die Betriebswirtschaftslehre sehr oft mit messbaren Grössen zu befassen, wie z.B. mit Kosten, Produktivität, Beschäftigungsgrad, sowie mit den Beziehungen zwischen diesen Grössen. Die Konzeption des funktionalen Zusammenhangs hat daher sehr grosse praktische Bedeutung.
Die funktionale Abhängigkeit ist also durch eine eindeutige Zuordnung charakterisiert, d.h. zu einem gegebenen Wert der unabhängigen Variablen, z.B. x_t, ist der Wert der abhängigen Variablen y_t eindeutig gegeben. Im Gegensatz dazu betrachtet man in der Statistik Wertepaare, bei denen zu jedem x-Wert der Wert y_i nicht eindeutig gegeben ist; zu jedem x_i können zwei oder mehrere y_i gegeben sein. So werden wir folgendem Fall begegnen:

Kosten in Fr.	Produktivität (x_k)	
	100	105
y_{ik}	42.32 56.79	39.44 46.49

Bei diesem Beispiel gehören zu jedem der beiden x-Werte je zwei y-Werte. Man bezeichnet eine derartige Abhängigkeit, also eine mehrdeutige Zuordnung, als **stochastische** (= mutmassliche) **Abhängigkeit.** Das Problem besteht für den Statistiker darin, an Stelle des Punkteschwarms eine mathematische Funktion zu setzen, d.h. die stochastische Abhängigkeit in eine funktionale Abhängigkeit überzuführen.

112 Operationsstufen zur Problemlösung

Zur Lösung des gesamten Problemkreises kann es zweckmässig sein, folgende Operationsstufen zu verwenden:

(1) Festlegung der **Problemlage** bzw. der **Zielsetzung;**

(2) Aufstellung des **Modells:**
- Entwicklung der Hypothese, d. h. Bestimmung der Einflussgrössen bzw. der abhängigen Variablen aus der betriebswirtschaftlichen Theorie (oder Empirie);
- Definition der Einheiten, in denen die verwendeten Variablen gemessen werden;
- Hypothese über die Form der Beziehung;
- Formulierung des Modells in Gleichungsform (= Formulierung des Zusammenhangs) → mathematisches Modell in allgemeiner Form.

(3) **Datenbeschaffung** bzw. Messung der Variablen: Beobachtung, Zählung, Versuch; Primär- oder Sekundärerhebung.

(4) **Schätzverfahren:**
Schätzen der Regressionsparameter; Masszahlen zur Messung der **Stärke der stochastischen Abhängigkeit:** Bestimmtheitsmass und Korrelationskoeffizient.

(5) **Prüfverfahren** (Tests):
- Ursachen der **Regressionsstreuung** → Streuungszerlegung;
- Prüfen der Hypothese, ob überhaupt Abhängigkeit zwischen den Variablen besteht;
- Prüfen der Hypothese über die Form der Abhängigkeit.

(6) **Regressionsgleichung:**
Ersetzung der beobachteten Werte von y_i durch theoretische Werte Y_i.

(7) **Vertrauensgrenzen** der Schätzung. Bestimmung des Fehlers, mit welchem die Schätzwerte behaftet sind (Güte der Schätzung).

(8) **Deutung** der Ergebnisse:
Analyse bzw. Interpretation der Resultate.

113 Systematik

Die nachfolgenden Untersuchungen basieren auf quantitativen Merkmalen. Je nach der **Zahl** der Veränderlichen unterscheidet man:

- Einfache Regression (zwei Variable)
- Mehrfache Regression (drei oder mehr Variable)

Je nach der **Form** der Regressionslinie unterscheidet man lineare oder nichtlineare Regression (Korrelation).

12 Einfache lineare Regression und Korrelation

121 Das Modell

Gegeben sei das **mathematische Modell** einer Funktionalbeziehung

$$\lambda = \alpha + \beta x \tag{1}$$

mit der abhängigen Variablen λ. Das entsprechende **statistische Modell** hat die Form

$$y = \alpha + \beta x + \varepsilon \tag{2}$$

wobei α und β unbekannte Parameter sind, ε aber eine Zufallsvariable ist. In der Praxis sind N Wertepaare (x_i, y_i), $i = 1,2, \ldots, N$, gegeben. Diese Beobachtungen bilden N Beziehungen:

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{3}$$

Nehmen wir nun an, dass für α und β die Schätzwerte a und b vorliegen. Es ist dann

$$y_i = a + b\,x_i + e_i \tag{4}$$

mit e_i als Schätzwert für das unbekannte ε_i. Als Schätzwert für Y_i ergibt sich

$$Y_i = a + b\,x_i \tag{5}$$

In dieser Gleichung (5) erscheinen die Parameter

a = Niveaukonstante,
b = Regressionskoeffizient.

122 Schätzverfahren

122.1 Regressionsparameter

Die Methode der kleinsten Quadrate erlaubt uns, Schätzwerte sowohl für die Niveaukonstante a wie auch für den Regressionskoeffizienten b zu finden. Wir stellen folgende Forderung auf:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - Y_i)^2 = \sum_{i=1}^{N} (y_i - a - b\,x_i)^2 = \text{Minimum}$$

Zur Erfüllung dieser Forderung müssen a und b folgendem Gleichungssystem genügen:

$$\left.\begin{aligned} N\,a + \left\{\sum_{i=1}^{N} x_i\right\} b &= \sum_{i=1}^{N} y_i \\ \left\{\sum_{i=1}^{N} x_i\right\} a + \left\{\sum_{i=1}^{N} x_i^2\right\} b &= \sum_{i=1}^{N} x_i y_i \end{aligned}\right\} \tag{6}$$

Wir erhalten, wenn

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

und

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

für die Niveaukonstante a:

$$a = \bar{y} - b_{yx}\bar{x} \tag{7}$$

und für den Regressionskoeffizienten b:

$$b_{yx} = \frac{S(x_i - \bar{x})\,(y_i - \bar{y})}{S(x_i - \bar{x})^2} \tag{8}$$

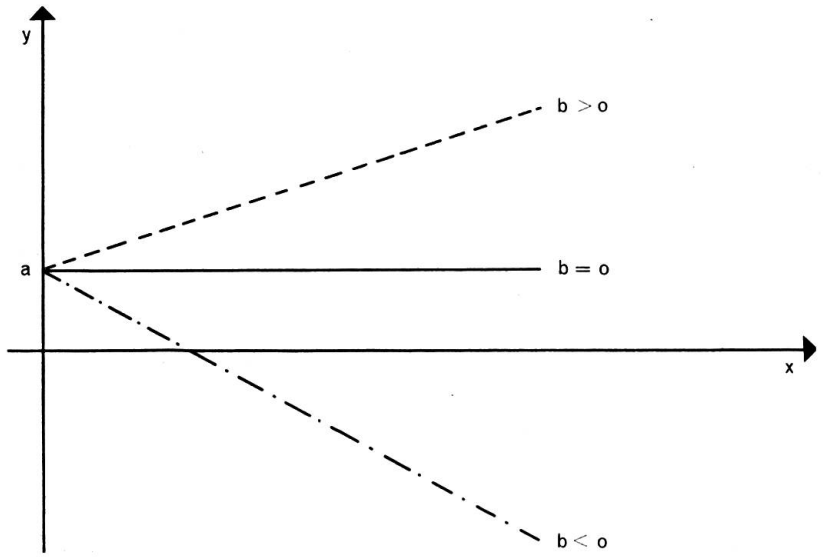
Es ist nun weiter

$$\left.\begin{aligned} S(x_i - \bar{x})\,(y_i - \bar{y}) &= Sx_i y_i - \frac{1}{N}(Sx_i)\,(Sy_i) &&= S_{xy} \\ S(x_i - \bar{x})^2 &= Sx_i^2 - \frac{1}{N}(Sx_i)^2 &&= S_{xx} \\ S(y_i - \bar{y})^2 &= Sy_i^2 - \frac{1}{N}(Sy_i)^2 &&= S_{yy} \end{aligned}\right\} \tag{9}$$

Durch Transformation erhalten wir folgende für die numerische Auswertung geeigneten Ansätze:

$$b_{yx} = \frac{Sx_i y_i - \frac{1}{N}(Sx_i)\,(Sy_i)}{Sx_i^2 - \frac{1}{N}(Sx_i)^2} = \frac{S_{xy}}{S_{xx}} \tag{10}$$

Für a > 0 können beim Regressionskoeffizienten b – grafisch dargestellt – folgende Verhältnisse vorliegen:



Es zeigt somit:

Vorzeichen von b:

Positiv: Regressionslinie steigt
Negativ: Regressionslinie fällt
b = 0: Regressionslinie parallel zur Abszisse

Betrag von b: Winkel der Regressionsgeraden zur Abszisse bzw. Ordinate.

Die **Bedeutung** des Regressionskoeffizienten b: Der in Ansatz (10) dargestellte Richtungsparameter b gibt an, um wieviel die abhängige Veränderliche y variiert (zu- oder abnimmt), wenn die unabhängige Veränderliche x um eine Einheit zunimmt.
Setzen wir nun in (5) für

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

so erhalten wir folgende Regressionsgleichung:

$$Y_i = \bar{y} + b_{yx}(x_i - \bar{x}) \tag{11}$$

Die Regressionsgerade ist im Variationsbereich der unabhängigen Veränderlichen x definiert, d.h. zwischen dem kleinsten und grössten Wert. Sowohl Interpolation wie Extrapolation sind möglich, wobei vor allem bei der Extrapolation sehr oft einige Vorsicht am Platze ist.

Umkehrung des Problems:

Es kann sinnvoll sein, von den y-Werten auf die x-Werte zu schliessen. Der Ansatz für die Regressionsgerade lautet dann

$$X = \bar{x} + b_{xy}(y - \bar{y}) \tag{12}$$

bzw. auch

$$X = a + b_{xy}y \tag{13}$$

Der Regressionskoeffizient b_{xy} ist bestimmt aus

$$b_{xy} = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2} = \frac{S_{xy}}{S_{yy}} \tag{14}$$

Es sind somit bei der stochastischen Abhängigkeit zwei Regressionsgeraden möglich

$$Y = \bar{y} + b_{yx}(x - \bar{x})$$

und

$$X = \bar{x} + b_{xy}(y - \bar{y})$$

122.2 Streuung der Einzelwerte

Die Streuung der Einzelwerte um die Regressionsgerade ist gemäss folgendem Ansatz (Vgl. E.Weber, Grundriss, 1967, S.332) zu bestimmen:

$$s_{yx}^2 = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - Y_i)^2 \tag{15}$$

Durch Transformation ergibt sich

$$s_{yx}^2 = \frac{1}{N-2} \left\{ S_{yy} - \frac{S_{yx}^2}{S_{xx}} \right\} \tag{16}$$

Mit (15) bzw. (16) messen wir das Ausmass der Übereinstimmung zwischen den geschätzten und beobachteten Werten der abhängigen Veränderlichen.

122.3 Bestimmtheitsmass und Korrelationskoeffizient

Während bei der Regressionsanalyse (vgl. 122.1 oben) die Richtung der stochastischen Abhängigkeit festgestellt wird, gibt uns die **Korrelationsanalyse** Aufschluss über den Grad des Zusammenhangs. Eine geeignete Masszahl ist vor allem das Bestimmtheitsmass B. Es ist folgendermassen definiert:

$$B = \frac{[S(x_i - \bar{x})\ (y_i - \bar{y})]^2}{[S(x_i - \bar{x})^2]\ [S(y_i - \bar{y})^2]} = \frac{S_{xy}^2}{S_{xx} S_{yy}} \tag{17}$$

Für die numerische Auswertung lässt sich der Ausdruck nach dem ersten Gleichheitszeichen transformieren. Vor allem lässt sich B auch mit Hilfe der beiden Regressionskoeffizienten errechnen. Es ist

$$B = b_{yx} b_{xy} \tag{18}$$

Deutung des Bestimmtheitsmasses:

Das Bestimmtheitsmass B gibt den Anteil der Streuung der abhängigen Veränderlichen y an, der sich aus der Variabilität der unabhängigen Veränderlichen x erklären lässt.

Die **Eigenschaften** des Bestimmtheitsmasses:

(1) B ist positiv, d.h. $B > 0$;
(2) Variation: $0 \leq B \leq 1$;
(3) Symmetrie in x und y, d.h. es ist

$$B = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{S_{yx}^2}{S_{yy} S_{xx}}$$

Aus dem Bestimmtheitsmass lässt sich durch einfache Operation der Korrelationskoeffizient gewinnen, der somit in enger Beziehung zu dieser Masszahl steht:

$$r = \sqrt{B} \tag{19}$$

Eigenschaften des Korrelationskoeffizienten:

(1) Vorzeichen: $r \lessgtr 0$
(2) Variation: $-1 \leq r \leq +1$
(3) Symmetrie in bezug auf die Variablen x und y:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{s_{yx}}{s_y s_x}$$

Für praktische Auswertungen hat das Bestimmtheitsmass grössere Bedeutung; ihm ist vor dem Korrelationskoeffizienten der Vorrang zu geben.

123 Prüfen von Hypothesen

123.1 Grundsätzliches

Ein **Prüfverfahren** (Test) wird benützt, damit entschieden werden kann, ob eine aufgestellte Hypothese angenommen oder verworfen werden soll, bzw. nach A. Linder (Statistische Methoden ..., S.43): «..., ob die Angaben einer Stichprobe mit einer Hypothese verträglich sind oder ihr widersprechen.»

Unter einer **Hypothese** versteht man nach Merriam-Webster «a tentative theory or supposition provisionally adopted to explain certain facts and to guide in the investigation of others».
In bezug auf die Arten von Hypothesen unterscheidet man:

- Nullhypothese;
- Alternativhypothese (Gegenhypothese).

Von grosser Bedeutung bei der Regressionsrechnung ist der Test, ob der Richtungsparameter b_1 wesentlich oder nur zufällig von einem zweiten Regressionskoeffizienten b_2 verschieden ist. Wir formulieren folgende Hypothese (einseitiger Test):

(1) Nullhypothese $H_0 : b_1 = b_2 = 0.62$
(2) Alternativhypothese $H_A : b_1 > b_2$
[bzw. $H_A : b_1 < b_2$]

Der zweiseitige Test prüft, ob zwischen b_1 und b_2 keine Differenz vorliegt bzw. die Alternativhypothese der gegenseitigen Abweichung, somit z.B.:

(1) Nullhypothese $H_0 : b_1 = b_2 = 0.62$
(2) Alternativhypothese $H_A : b_1 \neq b_2$

Fehler bei Prüfverfahren:

Die Gültigkeit von Hypothesen muss überprüft werden. Die Resultate lassen erkennen, ob die Hypothese angenommen oder aber verworfen werden muss. Dabei können zwei Typen von Fehlern unterlaufen:

Typ I: Verwerfen einer Hypothese, die richtig ist (= Risiko 1. Art);
Typ II: Annahme einer Hypothese, die falsch ist (= Risiko 2. Art).

Die folgende Übersicht zeigt die vier Möglichkeiten:

Hypothese	Entscheid	
	Annahme	Ablehnung
Hypothese ist richtig	Richtiger Entscheid	Fehlertyp I
Hypothese ist falsch	Fehlertyp II	Richtiger Entscheid

Unser Entscheid ist also richtig, wenn wir

(1) H_0 annehmen, wenn H_0 richtig ist;
(2) H_0 ablehnen, wenn H_0 falsch ist.

Hingegen begehen wir einen Fehler von **Typ I,** wenn wir H_0 ablehnen, wenn diese Hypothese richtig ist, bzw. einen Fehler vom **Typ II,** wenn wir H_0 annehmen, wenn H_0 falsch ist. Schwerwiegender ist meistens Fehlertyp II; man versucht ihn deshalb zu minimalisieren, indem der Stichprobenumfang erhöht wird. Dieses Vorgehen kann mit dem Hinweis darauf begründet werden, dass mehr Informationen bessere Entscheide ermöglichen.
Für die Praxis ist es nützlich, auch bei den Prüfverfahren einen bestimmten **Arbeitsablauf** einzuhalten:

(1) Problemlage
(2) Formulierung der Teilhypothese
(3) Prüfverteilung
(4) Sicherheitsgrad (Entscheidungswahrscheinlichkeit)
(5) Sicherheitsschwelle
(6) Bereitstellung der Prüf-Rechenformel; Berechnung des Prüfwertes
(7) Entscheid (Vergleich Prüfwert mit Sicherheitsschwelle); →Annahme oder Ablehnung der Hypothese (bzw. Alternativhypothese).
(8) Schlussfolgerungen.

123.2 Varianzanalyse

Die Quadratsumme der Abweichungen der Einzelwerte y_i von ihrem Mittelwert $\bar{y}$ kann folgendermassen zerlegt werden:

$$S(y_i - \bar{y})^2 = S(Y_i - \bar{y})^2 + S(y_i - Y_i)^2 \tag{20}$$

Nun ist aber

$$S(y_i - \bar{y})^2 = Sy_i^2 - \frac{1}{N}(Sy_i)^2 = S_{yy}$$

$$S(Y_i - \bar{y})^2 = b_{yx}\left\{Sx_i y_i - \frac{1}{N}(Sx_i)(Sy_i)\right\} = b_{yx}S_{xy}$$

$$S(y_i - Y_i)^2 = S_{yy} - b_{yx}S_{xy}$$

Wir erhalten folgendes Schema der Streuungszerlegung:

Streuung	Summe der Quadrate	FG	Durchschnittsquadrate
Auf der Regression	$b_{yx}S_{xy}$	1	$s_1^2 = b_{yx}S_{xy}$
Um die Regression	$S_{yy} - b_{yx}S_{xy}$	$N - 2$	$s_2^2 = \frac{1}{N-2}\{S_{yy} - b_{yx}S_{xy}\}$
Insgesamt	S_{yy}	$N - 1$	

(21)

Es ist ferner

$$s_2^2 = s_{yx}^2 = \frac{1}{N-2}\{S_{yy} - b_{yx}S_{xy}\}$$

Es kann gezeigt werden, dass

$$s_{yx}^2 = \hat{\sigma}_y^2$$

d.h. s_{yx}^2 ist ein zuverlässiger Schätzwert der unbekannten Streuung $\hat{\sigma}_y^2$.

Prüfverfahren: F-Test

$$F = \frac{DQ\ (\text{auf der Regression})}{DQ\ (\text{um die Regression})} = \frac{s_1^2}{s_2^2} \tag{22}$$

Freiheitsgrade: $n_1^* = 1;\quad n_2^* = N - 2$

Entscheidungskriterien:

$F \geq F_P$: Regression gesichert für Irrtumswahrscheinlichkeit von P %
$F < F_P$: Weitere Berechnungen nicht sinnvoll

(23)

Ist die Regression gesichert, werden von Null verschieden sein:

- Regressionskoeffizient;
- Bestimmtheitsmass (Korrelationskoeffizient).

Folge: Regressionsgerade kann bestimmt werden.

123.3 Prüfen der Regressionsparameter

Wir prüfen, ob die Abweichungen der beiden Regressionsparameter a und b einer Stichprobe von den entsprechenden Werten der Grundgesamtheit wesentlich oder nur zufällig sind. Als Prüfstreuung dient uns Ansatz (16) oben, d.h.

$$s_{yx}^2 = \frac{1}{N-2}\left\{S_{yy} - \frac{S_{yx}^2}{S_{xx}}\right\}$$

Als Prüfverteilung muss die t-Verteilung in Betracht gezogen werden.

(1) Prüfen der Niveaukonstanten a:

Wir verwenden folgenden t-Test:

$$t = \frac{a - \alpha}{s_{yx}} \sqrt{N} \tag{24}$$

mit: $n^* = (N-2)$ Freiheitsgraden.

Wir setzen nun $\alpha = 0$. Es ist dann folgende Hypothese zu prüfen:

$$H_0 : a = 0$$

Der Prüfansatz wird dann zu

$$t = \frac{a}{s_{yx}} \sqrt{N}, \text{ bzw. } t^2 = \frac{a^2 N}{s_{yx}^2}.$$

(2) Prüfen des Regressionskoeffizienten b:

Auch hier kommt ein t-Test zur Anwendung

$$t = \frac{b_{yx} - \beta}{s_{yx}} \sqrt{S_{xx}} \tag{25}$$

mit: $n^* = (N-2)$ Freiheitsgraden.

Wir prüfen folgende Hypothese:

$$H_0 : b_{yx} = \beta$$

Diese Hypothese wird angenommen, wenn

$$\frac{|b_{yx} - \beta|}{s_{yx}} \sqrt{S_{xx}} < t_P$$

andernfalls wird sie abgelehnt.

Nun ist aber β meistens nicht bekannt. Es ist daher sinnvoll, den Richtungsparameter b_{yx} mit Null zu vergleichen. Die Frage lautet: Ist b wesentlich oder nur zufällig von Null verschieden? Dies ist bei der Regressionsanalyse die eigentliche Kardinalfrage. Sollte nämlich $b = 0$ sein, so hat die in Betracht gezogene unabhängige Veränderliche **keinen Einfluss** auf die abhängige Variable.
Der Prüfansatz hat folgende Form:

$$t = \frac{b_{yx}}{s_{yx}} \sqrt{S_{xx}} \tag{26}$$

mit: $n^* = (N-2)$ Freiheitsgraden.

Es ist folgende Hypothese zu prüfen:

$$H_0 : b_{yx} = 0$$

Diese Hypothese wird abgelehnt, wenn

$$\frac{|b_{yx}|}{s_{yx}} \sqrt{S_{xx}} > t_P$$

Der Regressionskoeffizient ist in diesem Fall wesentlich von Null verschieden; die Hypothese eines bestehenden Einflusses muss in der Folge akzeptiert, bzw. die Nullhypothese muss abgelehnt werden.

(3) Prüfen der Differenz zwischen zwei Regressionskoeffizienten

Es ist folgende Hypothese zu prüfen:

$$H_0 : b_1 = b_2$$

d.h. wir prüfen, ob die Differenz zwischen den beiden Regressionskoeffizienten Null ist. Für das Prüfverfahren vgl. E.Weber, Grundriss (6. Auflage 1967), S.338f.

123.4 Prüfen des Bestimmtheitsmasses und des Korrelationskoeffizienten

(1) Prüfen des Bestimmtheitsmasses

Es ist naheliegend, die Abweichung des Bestimmtheitsmasses von Null zu prüfen; denn für $B = 0$ muss eine Abhängigkeit zwischen den betrachteten Variablen negiert werden.

Wir stellen also folgende Hypothesen auf:

$$H_0: B = 0$$
$$H_A: B > 0$$

Nach A. Linder (Methoden, S. 182) lässt sich das Prüfverfahren aus der Streuungszerlegung gewinnen. Der Ansatz (21) kann zu diesem Zweck auch folgendermassen formuliert werden:

Streuung	Summe der Quadrate	FG	Durchschnittsquadrate
Regression	BS_{yy}	1	$s_1^2 = BS_{yy}$
Einzelwerte um Regression	$(1 - B)\, S_{yy}$	$N - 2$	$s_2^2 = \frac{1}{N-2}(1 - B)\, S_{yy}$
Insgesamt	S_{yy}	$N - 1$	.

Das Verhältnis der beiden Durschnittsquadrate ergibt einen F-Test:

$$F = \frac{s_1^2}{s_2^2} = \frac{BS_{yy}(N-2)}{(1-B)S_{yy}} = \frac{B(N-2)}{(1-B)} \tag{27}$$

Mit: $n_1^* = 1$; $n_2^* = (N-2)$ FG und dem **Entscheidungskriterium:**

$F \geq F_P$: B ist wesentlich verschieden von Null; Einfluss besteht.

$F < F_P$: B ist bloss zufällig von Null verschieden.

Das Bestimmtheitsmass B kann auch mit Tafelwerten (Sicherheitspunkte) überprüft werden (vgl. Linder, Methoden, S. 183 bzw. S. 469–471). Verwendet wird folgender Ansatz:

$$B_P = \frac{F_P}{(N-2) + F_P} \tag{28}$$

Ein Bestimmtheitsmass, das kleiner ist als (28), weicht nur zufällig von Null ab. Es gilt also das Entscheidungskriterium:

$B \geq B_P$: B ist signifikant von Null verschieden.

(2) Prüfen des Korrelationskoeffizienten

Auch für das Prüfen des Korrelationskoeffizienten kann ein t-Test angesetzt werden. Der Ansatz lautet

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

mit: $n^* = (N-2)$ Freiheitsgraden.

Als Entscheidungskriterium gilt:

$t \geq t_P$: Der Korrelationskoeffizient ist signifikant von Null verschieden.

Es ist selbstverständlich möglich, mit Hilfe von Sicherheitspunkten (Tabellenwerten) das Prüfverfahren durchzuführen.

123.5 Linearität der Regression

Wir haben weiter oben (vgl. Ziff. 112) dargelegt, dass bei der Aufstellung des Modells eine Hypothese über die **Form der Beziehung** zwischen den betrachteten Variablen zu formulieren ist. Meistens wird man die Einzelwerte in einer Grafik zur Darstellung bringen und auf Grund dieser Zeichnung einen Vorentscheid fällen. Es ist aber angebracht, dieses Vorgehen durch ein Prüfverfahren abzustützen. So

wird man z.B. prüfen, ob die Hypothese der Linearität angenommen werden kann oder aber verworfen werden muss. Geeignet zur Lösung dieses Problems ist die Varianzanalyse bzw. ein F-Test. E.Weber (vgl. Grundriss, 1967, S.339ff.) bemerkt dazu: «Oftmals ist aus der Anordnung der Wertepaare (x,y) ersichtlich, ob lineare Regression angenommen werden kann. In Zweifelsfällen muss eine Prüfung auf Linearität vorgenommen werden. Diese Prüfung beruht darauf, die Summe der Quadrate der Abweichungen der Mittel der Beobachtungswerte von den entsprechenden Regressionswerten mit der Quadratsumme der Abweichungen der Beobachtungswerte von den jeweiligen Spaltenmitteln zu vergleichen.»
Das führt zu folgender **Anordnung:**

Unabhängige Veränderliche	x_1	...	x_k	...	x_p	Total
Einzelwerte y_{ik}	y_{11}	...	y_{1k}	...	y_{1p}	
			$\vdots$			
Allgemein:			y_{ik}			
Umfang	n_1	...	n_k	...	n_p	$n = \sum_{k=1}^{p} n_k$

Nach E.Weber (Grundriss, S.340ff.) führen wir die Varianzanalyse in zwei Stufen durch:

1. Stufe: Aufteilung der SQ (insgesamt) in SQ (zwischen den Spalten) und SQ (innerhalb der Spalten).
2. Stufe: Zerlegung der SQ (zwischen den Spalten).

Als Ergebnis der zwei Streuungszerlegungen erhalten wir einen F-Test zur Beurteilung, ob Linearität angenommen werden darf oder nicht.
In den nachstehenden Schemata der Streuungszerlegung sind die für numerische Auswertung benötigten Ansätze enthalten. Für die Theorie verweisen wir auf A.Linder (Methoden, S.156ff.) bzw. E.Weber (Grundriss, S.340/341).
Wir nehmen vorerst eine einfache Streuungszerlegung mit einer Aufteilung der SQ (insgesamt) in SQ (zwischen den Spalten) und SQ (innerhalb der Spalten) vor, d.h. also
SQ (insgesamt) = SQ (zwischen Spalten) + SQ (innerhalb).

Streuungszerlegung 1.Stufe:

Streuung	SQ	FG	DQ
Zwischen den Spalten	$\sum_k \frac{1}{n_k}\left(\sum_i y_{ik}\right)^2 - \frac{1}{n}\left(\sum_{k\,i} y_{ik}\right)^2 = B$	$p-1$	.
Innerhalb der Spalten	$A - B = C$	$n-p$	$s^2_{III} = \frac{C}{n-p}$
Insgesamt	$S_{yy} = A$	$n-1$	.

(29)

Es ist (vgl. oben)

$$S_{yy} = \sum_{k\,i} y^2_{ik} - \frac{1}{n}\left(\sum_{k\,i} y_{ik}\right)^2$$

Das SQ (zwischen den Spalten) lässt sich weiter aufteilen in

$$SQ\ (\text{zwischen}) = \sum_k [n_k (\bar{y}_{\bullet k} - Y_k)^2] + \sum_k [n_k (Y_k - \bar{y}_{\bullet\bullet})^2]$$

Streuungszerlegung 2. Stufe:

Streuung	SQ	FG	DQ
Auf der Regression	$(S_{xy})^2 / S_{xx} = D$	1	$s_I^2 = D$
Mittelwerte um Regression	$B - D = E$	$p - 2$	$s_{II}^2 = \frac{E}{p-2}$
Zwischen den Spalten	$S_k \frac{1}{n_k} \left(S y_{ik}\right)_i^2 - \frac{1}{n} \left(SS y_{ik}\right)_{k\,i}^2 = B$	$p - 1$	.

(30)

Aus (29) und (30) gewinnen wir folgenden F-Test:

$$F = \frac{s_{II}^2}{s_{III}^2}, \quad \text{mit:} \quad \left.\begin{matrix} n_1^* = (p-2) \\ n_2^* = (n-p) \end{matrix}\right\} \text{Freiheitsgraden}$$

Voraussetzung: $s_{II}^2 > s_{III}^2$.

Als **Entscheidungskriterium** verwenden wir:

$F < F_P$: Die DQ s_{II}^2 und s_{III}^2 sind nur zufällig voneinander verschieden: Lineare Regression ist zulässig. Der Test sagt nur aus, dass lineare Regression zulässig sei. Es ist aber möglich, dass eine andere Form den Daten besser entspricht.
Ist hingegen $F > F_P$, so darf keinesfalls lineare Regression angenommen werden.

124 Vertrauensgrenzen der Schätzung

124.1 Grundsätzliches

Für die Schätzwerte müssen zufällige Schwankungen in Rechnung gestellt werden. Es wird unsere Aufgabe sein, in den folgenden zwei Abschnitten Überlegungen in bezug auf die Unsicherheit der Schätzwerte anzustellen, insbesondere für

- den Regressionskoeffizienten b, bzw.
- für die Regressionswerte Y_i.

Dabei wird es darum gehen, eine geeignete Konzeption von Streuungsbereichen zu finden.

124.2 Vertrauensgrenzen für die Regressionsparameter

Wie bereits weiter oben angedeutet, nimmt der Regressionskoeffizient b im Rahmen der Regressionsanalyse eine ausserordentlich wichtige Stellung ein. Es ist daher gerechtfertigt, Überlegungen anzustellen in bezug auf die Unsicherheit des aus einer Stichprobe gewonnenen Schätzwertes.
Wir haben oben in Ziff. 123.3 folgenden t-Test angesetzt

$$t = \frac{b_{yx}}{s_{yx}} \sqrt{S_{xx}}$$

Aus diesem Ansatz können folgende Vertrauensgrenzen bestimmt werden (für die theoretischen Grundlagen vgl. K.W. Smillie, Introduction, S. 10/11):

$$b \pm t_P \frac{s_{yx}}{\sqrt{S_{xx}}} \tag{31}$$

Es sind $n^* = N-2$ Freiheitsgrade zu berücksichtigen.

Es ist ferner in Ansatz (31):

$$s_{yx}^2 = \frac{1}{N-2} \left\{ S_{yy} - \frac{S_{yx}^2}{S_{xx}} \right\}$$

bzw.

$$S_{xx} = Sx_i^2 - \frac{1}{N} (Sx_i)^2$$

Selbstverständlich ist es auch möglich, Vertrauensgrenzen für die Niveaukonstante a in Rechnung zu stellen:

$$a \pm t_P \frac{s_{yx}}{\sqrt{N}} \tag{32}$$

Sowohl für Ansatz (31) als auch für (32) sind $n^* = N - 2$ Freiheitsgrade in Rechnung zu stellen.

124.3 Vertrauensgrenzen für die Regressionswerte Y_i

Die Streuung des zum Wert x gehörenden Regressionswertes beträgt (vgl. A. Linder, Methoden, S.160; K.W. Smillie, Introduction, S.11), dargestellt in unserer Symbolik,

$$s_Y^2 \sim s_{yx}^2 \left\{ \frac{1}{N} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right\} \tag{33}$$

Der Ausdruck in der Klammer von (33)

$$\frac{(x_i - \bar{x})^2}{S_{xx}}$$

zeigt, dass s_Y^2 mit zunehmender Entfernung der x_i von $\bar{x}$ grösser wird. Mit Hilfe von (33) können wir nun die Vertrauensgrenzen definieren

$$a + bx_i \pm t_P s_Y \tag{34.1}$$

bzw.

$$Y_i \pm t_P s_Y \tag{34.2}$$

Es sind in diesem Ansatz für die Bestimmung des Tafelwertes t_P insgesamt $n^* = (N-2)$ Freiheitsgrade zu berücksichtigen.
In der Praxis liegt sehr oft der Fall vor, dass für einen Einzelwert von x der dazugehörige Y-Wert prognostiziert werden muss. Für die Streuung gilt dann:

$$s_{Y(E)}^2 \sim s_{yx}^2 \left[1 + \frac{1}{N} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \tag{35}$$

Für den Wert t_P bei der Bestimmung der Vertrauensgrenzen sind auch hier $n^* = N-2$ Freiheitsgrade vorzumerken.

125 Zwei Regressionsgerade

Es ist oft die Sachlage gegeben, dass zwei Regressionsgerade miteinander zu vergleichen sind. Das Vorgehen zur Lösung dieses Problems gestaltet sich folgendermassen:

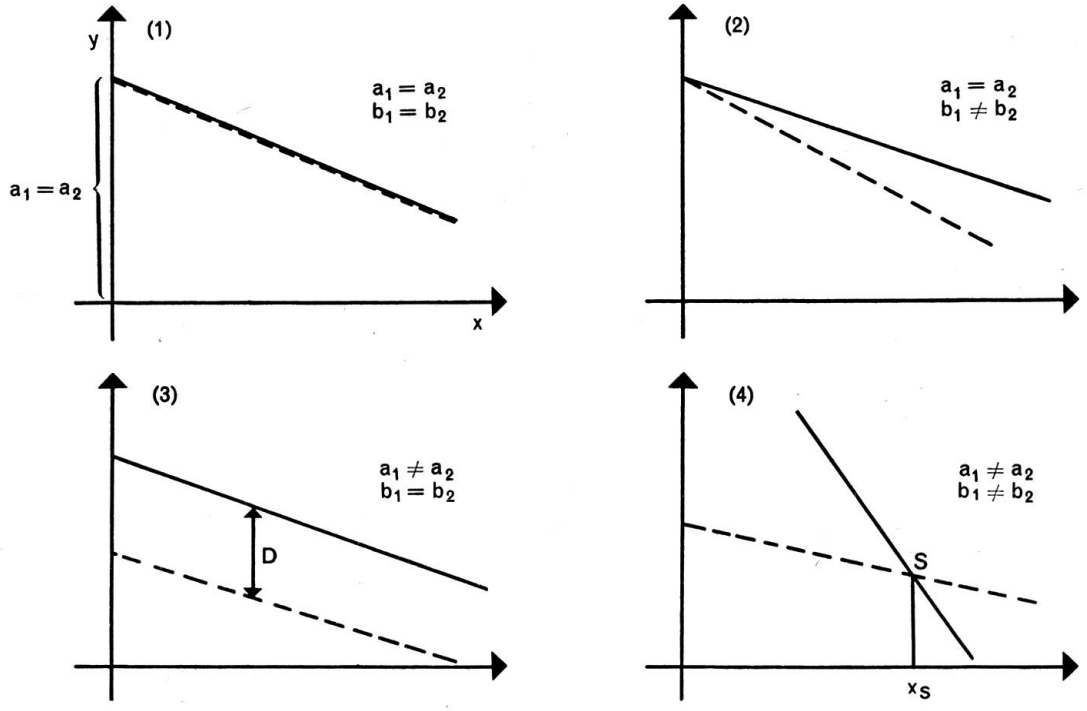
(1) Man wird vorerst für beide Regressionsgerade die **Linearität** prüfen;
(2) Es wird zu prüfen sein, ob die Regressionskoeffizienten b_1 bzw. b_2 überhaupt **von Null verschieden** sind;
(3) Sind lineare Abhängigkeiten als zulässig erkannt bzw. die Richtungsparameter verschieden von Null, so wird man schliesslich die **Parallelität** prüfen.

In bezug auf Niveaukonstanten bzw. Regressionskoeffizienten sind nun folgende Fälle denkmöglich:

Regressions-koeffizienten	Niveaukonstanten	
	$a_1 = a_2$	$a_1 \neq a_2$
Gleich	$b_1 = b_2$	$b_1 = b_2$
Verschieden	$b_1 \neq b_2$	$b_1 \neq b_2$

Sollte der Fall vorliegen, wo die Niveaukonstanten verschieden, die Regressionskoeffizienten aber in Vorzeichen und Betrag gleich sind, so kann es nützlich sein, die Distanz D zwischen beiden Regressionsgeraden zu berechnen und in die Überlegungen einzubeziehen.
Bei Nichtparallelität wird es interessant sein, den Schnittpunkt beider Geraden zu bestimmen.

Aus unserem soeben aufgeführten Schema lassen sich folgende Grafiken gewinnen:



Die beiden Probleme

- Linearität (vgl. Ziff. 123.5)
- Verschiedenheit von Null (vgl. 123.3)

haben wir oben behandelt. Für die in der Praxis hauptsächlich vorkommenden Fälle (3) und (4) – verschiedene Niveaukonstanten – ist es gegeben, Varianzanalyse und Regressionsanalyse zu kombinieren; es resultiert die sogenannte Kovarianzanalyse, auch etwa Mitstreuungszerlegung genannt.

A. Linder (Methoden, S. 220) bemerkt dazu: «Die Mitstreuungszerlegung wird vor allem benützt, um festzustellen, wie eine oder mehrere unabhängige Veränderliche sich auf die Unterschiede zwischen den Durchschnitten einer abhängigen Veränderlichen auswirken. Sie kann aber auch dazu dienen, die Unterschiede zwischen mehreren Regressionskoeffizienten zu beurteilen.»

Als Grundlage für die Berechnungen seien N Beobachtungen mit Werteverbindungen (x,y) gegeben. Wir nehmen eine Ausgliederung in p Gruppen (1, 2, ... , k, ... , p) vor. Innerhalb der Teilgesamtheiten sind die entsprechenden Werteverbindungen vorhanden. Aus diesen können Summenwerte und Durchschnitte gebildet werden. Für theoretische Erörterungen verweisen wir u. a. auf die Monographien von A. Linder und E. Weber, bzw. auf das Werk von B. Ostle.

Für die numerische Auswertung gelangen wir zu folgendem Schema der Streuungszerlegung (36):

Streuung	SQ	FG	DQ
1. Gerade	$S_{y_1y_1} - b_1 S_{x_1y_1}$		.
.....			
k-te Gerade	$S_{y_ky_k} - b_k S_{x_ky_k}$		.
.....			
p-te Gerade	$S_{y_py_p} - b_p S_{x_py_p}$		.
p Regressionsgerade	$\sum_{k=1}^{p} S_{y_ky_k} - \sum_{k=1}^{p} b_k S_{x_ky_k} = E$	$N - 2p$	$s_1^2 = \frac{E}{N-2p}$
Nichtparallelität	$C - E = D$	$p - 1$	$s_2^2 = \frac{D}{p-1}$
Innerhalb der Gruppen	$\sum_{k=1}^{p} S_{y_ky_k} - b^* \left(\sum_{k=1}^{p} S_{x_ky_k} \right) = C$	$N - p - 1$	$s_3^2 = \frac{C}{N-p-1}$
Zwischen den Gruppen	$A - C = B$	$p - 1$	$s_4^2 = \frac{B}{p-1}$
Insgesamt	$S_{yy} - b S_{xy} = A$	$N - 2$	.

Es ist ferner [vgl. SQ (innerhalb Gruppen)]

$$b^* = \frac{S^*_{xy}}{S^*_{xx}} = \frac{\sum_k SS_{x_k y_k}}{\sum_k SS_{x_k x_k}}$$

Vgl. zu diesem Schema auch: B. Ostle, Statistics in Research, 1964, S. 203.

Aus der Streuungszerlegung (36) resultiert ein F-Test, der Antwort auf die Frage gibt, ob die berechneten Regressionskoeffizienten wesentlich oder nur zufällig voneinander abweichen. Als Prüfstreuung verwenden wir s_1^2.

Formulierung der **Hypothese:**

$H_0: b_1 = b_2 = \ldots = b_p = b$

H_A: Mindestens zwei Regressionskoeffizienten sind signifikant voneinander verschieden.

Der F-Test lautet

$$\text{I.} \qquad F = \frac{s_2^2}{s_1^2}; \quad \text{mit:} \quad \left.\begin{aligned} n_1^* &= p-1 \\ n_2^* &= N-2p \end{aligned}\right\} \text{ Freiheitsgraden} \qquad (37)$$

Die errechnete Verhältniszahl vergleichen wir mit dem Tabellenwert. Bei

$F < F_P$: Die Regressionskoeffizienten sind nur zufällig voneinander verschieden.

Sind mindestens zwei Regressionskoeffizienten signifikant voneinander verschieden, so kann man prüfen, zwischen welchen Parametern Unterschiede bestehen (t-Test).

Unter der Voraussetzung der Parallelität der Regressionskoeffizienten kann man testen, ob die Abstände D_k wesentlich oder nur zufällig voneinander verschieden sind. Dazu verwenden wir die untersten drei Zeilen der Streuungszerlegung von (36) und erhalten folgenden F-Test:

$$\text{II.} \qquad F = \frac{s_4^2}{s_3^2}; \quad \text{mit:} \quad \left.\begin{aligned} n_1^* &= p-1 \\ n_2^* &= N-p-1 \end{aligned}\right\} \text{ Freiheitsgraden} \qquad (38)$$

Das Entscheidungskriterium:

$F \geq F_P$: Mindestens zwei der Distanzen sind signifikant voneinander verschieden.

Man kann auch hier prüfen, welche Distanzen voneinander verschieden sind. Ist $F < F_P$, so weichen die Distanzen nur zufällig voneinander ab; die Regression durch den Schwerpunkt kann die stochastische Abhängigkeit generell angeben. Wir hätten dann die Verhältnisse, die in Grafik (1) dargestellt sind. Es liegen also vor:

- Linearität;
- Verschiedenheit der beiden b von Null;
- Parallelität;
- Distanzen nur zufällig voneinander verschieden.

In der Praxis dürfte jedoch dieser Spezialfall nur selten gegeben sein.

13 Einfache nichtlineare Regression

131 Grundsätzliches

Im Abschnitt 12 haben wir das einfachste Modell behandelt; es betrifft die lineare Regression und Korrelation. Als Ergebnis des Linearitätstests muss aber in vielen Fällen die Hypothese einer nichtlinearen stochastischen Abhängigkeit angenommen werden. Sehr oft weisen schon theoretische Untersuchungen auf eine nichtlineare Beziehung hin.

Zur Lösung der vorliegenden Probleme drängen sich folgende Methoden auf:

- Transformation auf den linearen Fall;
- Mehrfache lineare Regression;
- Orthogonale Polynome.

Es wird fallweise zu prüfen sein, welches Verfahren geeignet ist, die Daten auszuwerten.

132 Transformation auf einfache lineare Regression

Theoretische oder empirische Untersuchungen lassen es oft als zweckmässig erscheinen, nichtlineare Regression anzunehmen. Der Übergang zu Logarithmen muss sich in vielen Fällen – besonders bei betriebswirtschaftlichen und volkswirtschaftlichen Problemen – als naheliegendste und einfachste Lösungsmöglichkeit aufdrängen. Folgende Ansätze können einer derartigen Transformation unterzogen werden:

$$Y = a\,b^x \tag{39}$$

Die Transformation ergibt

$$\log Y = \log a + x(\log b)$$

Die Bestimmung der Parameter erfolgt in der oben dargestellten Art und Weise (vgl. Ziff. 12).

$$Y = a\,x^b \tag{40}$$

Das Ergebnis der Transformation

$$\log Y = \log a + b(\log x)$$

Auch hier stellen sich keine neuen Probleme. Die Erkenntnisse aus der linearen Regression finden analoge Anwendung.

133 Mehrfache lineare Regression

In bestimmten Fällen führt die Methode der mehrfachen linearen Regression zum verfolgten Ziel, eine nichtlineare Beziehung statistischen Verfahren zu unterwerfen (vgl. dazu A. Linder, Methoden, S. 207 ff.).
Eine Regressionsgleichung von der Form

$$Y = a + bx + cx^2$$

kann in die Regressionsgleichung

$$Y = a + b_1x_1 + b_2x_2$$

oder auch

$$Y = \bar{y} + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2)$$

übergeführt werden, mit $x_1 = x$ und $x_2 = x^2$. Für die Berechnung der Regressionskoeffizienten b_1 und b_2 sind die entsprechenden Ansätze in Abschnitt 14 (mehrfache lineare Regression) massgebend.
Man wird im Anschluss an die Berechnungen zur Auffindung der Regressionsparameter verschiedene Hypothesen prüfen, so z. B. ob die Regressionsgleichung mit den beobachteten Werten gut übereinstimmt. Ferner könnte man die Frage stellen, ob nicht das lineare Glied weggelassen werden könnte, d. h. man wird für b_1 die Hypothese prüfen, ob

$$H_0: \beta_1 = 0$$

Man wird eine Streuungszerlegung mit einem F-Test ansetzen, um zu prüfen, ob sich der Wert von b_1 mit der Hypothese verträgt, dass $\beta_1 = 0$. Ist dies der Fall, wird man die eine der unabhängigen Variablen weglassen können.

134 Regression mit orthogonalen Polynomen

Bisher haben wir bei einfacher linearer Regression folgenden Ansatz benützt

$$Y = a_0 + a_1x$$

Durch Erweiterung erhalten wir ein allgemeines Polynom p-ten Grades

$$Y = a_0 + a_1x + a_2x^2 + \ldots + a_px^p \tag{41}$$

Wir stellen die Forderung F auf

$$F = S(y_i - Y_i)^2 = \text{Minimum}.$$

Die Bestimmungsgleichungen lassen sich mit Hilfe der Determinantenmethode lösen.

Die Berechnungen können vereinfacht werden, wenn wir an Stelle von (41) folgenden Ansatz verwenden:

$$Y = A_0 + A_1\varphi_1 + \ldots + A_p\varphi_p \tag{42}$$

In dieser Formel sind

$\varphi_k (k = 1, 2, \ldots, p)$ = Orthogonale Polynome;

$A_k (k = 0, 1, \ldots, p)$ = Konstanten, definiert durch folgende Ansätze:

$$A_0 = SY/N, \text{ bzw.} \tag{43}$$

$$A_k = \frac{SY\varphi_k}{S(\varphi_k)^2}, \quad k = 1, 2, \ldots, p \tag{44}$$

Für eingehendere Darstellungen vgl. u.a.:

Draper, N.R.: Applied Regression Analysis, 1966, S.150ff.
Fisher, R.A.: Statistical Methods, 13th ed. 1963, insbes. Kap.27 und 28 (S.147–156).
Grossen, H.: Regression mit orthogonalen Polynomen, Bern 1948.
Linder, A.: Statistische Methoden, 1964, S.210ff.
Mather, K.: Statistical Analysis in Biology, London 1964 (S.129ff.).
Weber, E.: Grundriss, 6. Aufl. 1967, S.352ff.

Für Tafelwerte stehen zur Verfügung:

Fisher/Yates: Statistical Tables, 6th ed. 1963, S.33ff. bzw. 98ff.

Die Methode der orthogonalen Polynome ist vor allem gut geeignet zur Behandlung von Zeitreihen (äquidistante Werte von x).

14 Mehrfache lineare Regression und Korrelation

141 Allgemeines

In den vorhergehenden Abschnitten haben wir zwei Veränderliche in ihrer Abhängigkeit untersucht. Es gibt aber Problemstellungen, wo drei und mehr Variable gleichzeitig analysiert werden müssen. So kann es vorkommen, dass die Kosten pro Pflegetag nicht nur von der Produktivität abhängig sind, sondern auch von den Marktpreisen der Produktionsfaktoren. Eine abhängige Variable steht dann mehreren Einflussfaktoren, mehreren unabhängigen Variablen gegenüber. Solche Probleme lassen sich nur mit Hilfe der mehrfachen linearen Regression und Korrelation lösen. Dabei bildet die Theorie der einfachen linearen Regression und Korrelation die Basis, auf dem die mehrfache Regression aufgebaut werden kann.

142 Das Modell

Wie in Abschnitt 121 oben nehmen wir eine lineare Beziehung zwischen einer abhängigen und p unabhängigen Variablen an:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \tag{45}$$

Es sollen ferner n Werteverbindungen vorhanden sein:

$$(x_{1i}, x_{2i}, \ldots, x_{pi}, y_i), \quad i = 1, 2, \ldots, n,$$

für p unabhängige Variable $x_1, x_2, \ldots, x_p$ und einer abhängigen Veränderlichen y. Die Beobachtungen genügen, analog zu Ziff.121, folgenden n Beziehungen:

$$y_i = \alpha + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i \tag{46}$$

Für $\alpha, \beta_1, \beta_2, \ldots, \beta_p$ sollen nun Schätzwerte vorliegen, so dass (46) wird:

$$y_i = a + b_1 x_{1i} + \ldots + b_p x_{pi} + e_i \tag{47}$$

Als Schätzwert für Y_i ergibt sich damit für zwei unabhängige Variable:

$$Y_i = a + b_1 x_{1i} + b_2 x_{2i} \tag{48}$$

Der Ansatz (48) enthält die Niveaukonstante a und die beiden Regressionskoeffizienten b_1 und b_2, die zu bestimmen sind.

Für die Bedingungen und Annahmen in bezug auf das Modell vgl. K.W.Smillie, Introduction, S.41 oben. Erwähnt sei lediglich die Beschränkung (vgl. Smillie, S.41, Ziff.3.2.5): «The number of sets of observations is greater than the number of regression coefficients to be estimated, i.e. $n > p + 1$, and no one independent variable is a linear combination of any of the remaining independent variables».
Wie bei der einfachen linearen Regression sind Schätzwerte aufzusuchen, Hypothesen zu prüfen und Vertrauensgrenzen zu bestimmen.

143 Schätzverfahren

143.1 Regressionsparameter

Vorerst sind die mehrfachen (partiellen) Regressionskoeffizienten zu bestimmen. Gegeben seien zwei unabhängige Veränderliche und eine abhängige Variable:

(1) Unabhängige Veränderliche

$$x_1: x_{11}, x_{12}, \ldots, x_{1N}$$

$$x_2: x_{21}, x_{22}, \ldots, x_{2N}$$

(2) Abhängige Veränderliche

$$y: y_1, y_2, \ldots, y_N$$

Es bestehen also folgende Werteverbindungen:

$$(x_{1i}, x_{2i}; y_i); \quad i = 1, 2, \ldots, N$$

Es gilt die Forderung

$$S(y_i - a - b_1 x_{1i} - b_2 x_{2i})^2 = \text{Minimum}.$$

Wir verzichten auf Beweise und geben als Resultat:

Niveaukonstante a:

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 \tag{49}$$

oder auch

$$a = \frac{Sy_i - b_1 Sx_{1i} - b_2 Sx_{2i}}{N} \tag{50}$$

Regressionskoeffizienten b_1 und b_2:

$$b_1 = \frac{1}{\triangle} \begin{vmatrix} S_{x_1 y} & S_{x_1 x_2} \\ S_{x_2 y} & S_{x_2 x_2} \end{vmatrix}$$

$$b_2 = \frac{1}{\triangle} \begin{vmatrix} S_{x_1 x_1} & S_{x_1 y} \\ S_{x_1 x_2} & S_{x_2 y} \end{vmatrix} \tag{51}$$

wobei für den Ausdruck $\triangle$ gilt

$$\triangle = \begin{vmatrix} S_{x_1 x_1} & S_{x_1 x_2} \\ S_{x_1 x_2} & S_{x_2 x_2} \end{vmatrix}$$

Die mehrfachen (partiellen) Regressionskoeffizienten messen die stochastische Abhängigkeit je zweier Veränderlicher unter Konstantsetzung der dritten Variablen, deren Einfluss dadurch bei der Analyse eliminiert wird. Man könnte deshalb auch setzen:

$$b_1 = b_{yx_1 \cdot x_2}$$

bzw.

$$b_2 = b_{yx_2 \cdot x_1}$$

Die Methode der Multiplikatoren:
(vgl. dazu E.Weber, Grundriss, S.345ff.)

Wir bezeichnen die Multiplikatoren mit dem Buchstaben c und verwenden – bei drei unabhängigen Variablen – folgendes Gleichungssystem:

$$\begin{array}{l|c|c|c|} c_{\bullet 1} S_{x_1x_1} + c_{\bullet 2} S_{x_1x_2} + c_{\bullet 3} S_{x_1x_3} = & 1 & 0 & 0 \\ c_{\bullet 1} S_{x_2x_1} + c_{\bullet 2} S_{x_2x_2} + c_{\bullet 3} S_{x_2x_3} = & 0 & 1 & 0 \\ c_{\bullet 1} S_{x_3x_1} + c_{\bullet 2} S_{x_3x_2} + c_{\bullet 3} S_{x_3x_3} = & 0 & 0 & 1 \end{array} \qquad (52)$$

Die Auflösung des Systems liefert die Matrix der Multiplikatoren, d. h. man erhält die Regressionskoeffizienten aus den Gleichungen

$$\begin{aligned} b_1 &= c_{11} S_{x_1y} + c_{12} S_{x_2y} + c_{13} S_{x_3y} \\ b_2 &= c_{12} S_{x_1y} + c_{22} S_{x_2y} + c_{23} S_{x_3y} \\ b_3 &= c_{13} S_{x_1y} + c_{23} S_{x_2y} + c_{33} S_{x_3y} \end{aligned} \qquad (53)$$

Für die Auswertung vgl. E.Weber, Grundriss, S.346f.: Schema für die Berechnung der partiellen Regressionskoeffizienten.
Die Methode der Multiplikatoren ist sehr gut geeignet, um zu einem bestimmten System (z.B. zwei unabhängige und eine abhängige Variable) **eine oder mehrere Variablen hinzuzufügen** (vgl. dazu die Arbeit von W.G.Cochran: The omission or addition of an independent variable in multiple linear regression; in Suppl. to the Journal of the Royal Statistical Society, Vol. V, 1938, S.171 ff.). Vor allem sind die Ansätze von Cochran geeignet für die Erstellung eines Elektronenrechner-Programmes.

143.2 Die Bestimmtheit

(1) Totales Bestimmtheitsmass

Für zwei unabhängige und eine abhängige Variable hat der Ansatz für das Bestimmtheitsmass (totale Bestimmtheit) folgende Form:

$$B_T = \frac{1}{S_{yy}} \{ b_1 S_{x_1y} + b_2 S_{x_2y} \} \qquad (54)$$

Die Deutung des Bestimmtheitsmasses kann selbstverständlich auch aus der Streuungszerlegung heraus erfolgen.

(2) Partielle Bestimmtheitsmasse

Setzen wir für

$$B_1 = B_{yx_1 \cdot x_2}$$

und

$$B_2 = B_{yx_2 \cdot x_1}$$

dann erhalten wir für die partiellen Bestimmtheitsmasse erster Ordnung:

$$\left.\begin{aligned} B_1 &= \frac{B_{yx_1} - 2\sqrt{B_{yx_1} B_{yx_2} B_{x_1x_2}} + B_{yx_2} B_{x_1x_2}}{(1 - B_{yx_2})(1 - B_{x_1x_2})} \\ B_2 &= \frac{B_{yx_2} - 2\sqrt{B_{yx_2} B_{yx_1} B_{x_1x_2}} + B_{yx_1} B_{x_1x_2}}{(1 - B_{yx_1})(1 - B_{x_1x_2})} \end{aligned}\right\} \qquad (55)$$

Diese Bestimmtheitsmasse nach Ansatz (55) lassen sich auch darstellen in den Bestimmtheitsmassen nullter Ordnung.

143.3 Korrelationskoeffizient

Sowohl der totale wie der partielle Korrelationskoeffizient lassen sich aus der allgemeinen Formel

$$r = \sqrt{B}$$

herleiten. Auch hier lassen sich die partiellen Korrelationskoeffizienten erster Ordnung auf Koeffizienten nullter Ordnung zurückführen.

144 Prüfen von Hypothesen

144.1 Streuungszerlegung

Es wird vorerst generell zu entscheiden sein, ob eine mehrfache lineare Regression statthaft bzw. sinnvoll ist. Wir zerlegen zu diesem Zweck wiederum die Quadratsummen und gelangen zu folgendem Schema der Streuungszerlegung:
(zwei unabhängige, eine abhängige Variable)

Streuung	SQ	FG	DQ
Auf Regression	$b_1 S_{x_1 y} + b_2 S_{x_2 y} = C$	2	$s_1^2 = \frac{1}{2} C$
Um Regression	$A - C = B$	$N - 3$	$s_2^2 = \frac{B}{N-3}$
Insgesamt	$S_{yy} = Sy_i^2 - \frac{1}{N}\left(Sy_i\right)^2 = A$	$N - 1$	.

(56)

Der Testansatz lautet

$$F = \frac{DQ\ (\text{auf Regression})}{DQ\ (\text{um Regression})} = \frac{s_1^2}{s_2^2} \tag{57}$$

mit: $n_1^* = 2$; $n_2^* = (N-3)\ FG$

Geprüft wird die Hypothese

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Entscheidungskriterium:

$F \geq F_P$: Mehrfache lineare Regression ist zulässig.

144.2 Prüfen der partiellen Regressionskoeffizienten

Es bestehen dafür zwei grundsätzliche Möglichkeiten:

(1) Prüfen der Abweichung des Regressionskoeffizienten b_1 von einem Wert β_1;
(2) Prüfen, ob ein partieller Regressionskoeffizient wesentlich oder nur zufällig von Null abweicht.

In beiden Fällen wird ein t-Test als Prüfverfahren dienen, wobei die Multiplikatoren im Prüfansatz Verwendung finden.
Bei K.W. Smillie hat dieser Ansatz folgende Form (vgl. Introduction, S. 48, Ansatz 3.4.3):

$$t = \frac{b_j - \beta_j}{s\sqrt{c_{jj}}} = \frac{b_j - \beta_j}{s_{b_j}}$$

Die Zahl der Freiheitsgrade beträgt $n^* = N-p-1$.

Der Vergleich von t mit dem diesbezüglichen Tabellenwert zeigt, ob die formulierte Hypothese anzunehmen oder zu verwerfen ist.

144.3 Prüfen der Bestimmtheiten

(1) Mehrfache (totale) Bestimmtheit

Aus Ansatz (54) ergibt sich folgende Verallgemeinerung für das Bestimmtheitsmass:

$$B_T = \frac{1}{S_{yy}}\left\{b_1 S_{x_1 y} + b_2 S_{x_2 y} + \ldots + b_p S_{x_p y}\right\}$$

Als Prüfverfahren dient ein F-Test, der aus der Streuungszerlegung abgeleitet wird (vgl. A. Linder, Methoden)

$$F = \frac{B(N-p-1)}{(1-B)\,p} \tag{58}$$

mit: $n_1^* = p$ und $n_2^* = (N-p-1)\ FG$

Der F-Wert wird mit dem benötigten Tafelwert verglichen. Wenn

$F > F_P$: B ist wesentlich von Null verschieden.

Andernfalls ist B nur zufällig von Null verschieden.

(2) Partielle Bestimmtheiten

Für partielle Bestimmtheiten (p-1)ter Ordnung ist für das Prüfen der Nullhypothese ebenfalls ein F-Test anzusetzen:

$$F = \frac{B\,(N-2p)}{(1-B)\,p} \tag{59}$$

mit: $n_1^* = p$ und $n_2^* = (N-2p)$ FG.

Aus Tafeln mit F-Verteilung ist der Tabellenwert zu entnehmen und mit dem errechneten F-Wert zu vergleichen.

145 Vertrauensgrenzen

Für die berechneten Regressionswerte kann man Vertrauensgrenzen bestimmen (vgl. dazu A. Linder, Methoden, S. 196). Analog zu Ansatz (34.2) oben gilt auch hier:

$$Y \pm t_P s_Y$$

Die Streuung s_Y^2 des Regressionswertes Y kann gemäss folgendem Ansatz berechnet werden:

$$\begin{aligned} s_Y^2 \sim s_2^2 \Big\{ \frac{1}{N} &+ c_{11}(x_1-\bar{x}_1)^2 + 2c_{12}(x_1-\bar{x}_2)(x_2-\bar{x}_2) + \ldots + 2c_{1p}(x_1-\bar{x}_1)(x_p-\bar{x}_p) \\ &+ c_{22}(x_2-\bar{x}_2)^2 \quad + \ldots + 2c_{2p}(x_2-\bar{x}_2)(x_p-\bar{x}_p) \\ &\quad \ddots \quad + \ldots + 2c_{3p}(x_3-\bar{x}_3)(x_p-\bar{x}_p) \\ &\quad\quad \ddots \quad\quad\quad \vdots \\ &\quad\quad\quad \ddots + c_{pp}(x_p-\bar{x}_p)^2 \Big\} \end{aligned} \tag{60}$$

Vgl. zu (60) A. Linder, Statistische Methoden, 1964, S. 196.

Wie in Ansatz (35) kann man an Stelle der Vertrauensgrenzen für Y diejenigen für einen Einzelwert ausrechnen; in der Klammer des Ausdruckes (60) ist in diesem Fall der Wert 1 hinzuzufügen.

mit: $n^* = (N-p-1)$ Freiheitsgraden

c_{ik} = Multiplikatoren

s_2^2 = DQ (um Regression)

Wie bereits weiter oben erläutert, wird auch hier s_Y^2 und damit die Ungenauigkeit um so grösser, je mehr wir uns von den Durchschnitten der unabhängigen Variablen ($\bar{x}_1, \bar{x}_2, \ldots$) entfernen.

146 Unechte Variable

Es kann in der Praxis erforderlich sein, qualitative Variable in die Untersuchungen einzubeziehen. So ist bei betriebswirtschaftlichen Spitalanalysen die Sachlage gegeben, den Einfluss von Spezialabteilungen auf die Kosten pro Pflegetag abzuklären. Wir haben dann z. B. drei Einflussgrössen:

x_1 = Produktivität;

x_2 = Durchschnittliche Personalkosten;

x_3 = Spezialabteilungen mit der Setzung:

0 Keine Spezialabteilung

1 Es bestehen Spezialabteilungen.

Die statistische Basistabelle sieht dann folgendermassen aus:

Spital	Produktivität x_1	Personalkosten (Fr.) x_2	Spezialabteilung x_3	Betriebskosten pro Pflegetag (Fr.) y
A	129	9438	0	32.67
B	83	10621	1	59.95
C	111	9995	0	46.77
.				
.				
.				

Für Einzelheiten vgl. die Monographie von K.W. Smillie, S.69ff. (dummy variables). Diese Arbeit enthält auch ein durchgerechnetes Beispiel.

15 Mehrfache nichtlineare Regression und Korrelation

151 Allgemeines

Im Abschnitt 14 haben wir bei den Ausführungen über die Mehrfachkorrelation immer lineare Abhängigkeit vorausgesetzt. Wie dies die Autoren Ezekiel-Fox (vgl. Methods, 3rd ed. 1959, S.204ff.) andeuten, kann die Annahme linearer Beziehungen bei verschiedenen Untersuchungen zu begrenzten Ergebnissen führen. Es ist daher gegeben, auch bei mehrfacher Regression mit nichtlinearen Ansätzen zu arbeiten.

152 Transformation

In Abschnitt 132 haben wir gezeigt, dass durch geeignete Transformationen die nichtlineare in lineare Regression übergeführt werden kann. Selbstverständlich ist dieses Verfahren auch bei mehrfacher nichtlinearer Regression anwendbar. Man wird also für eine bestimmte Variable den Übergang zu Logarithmen vornehmen.

153 Mehrfache nichtlineare Regression

Analog zu Abschnitt 133 können wir mit Hilfe einer Substitution das Problem zu lösen versuchen. Es sei z. B.

$$Y = \bar{y} + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + b_3(x_2 - \bar{x}_2)^2$$

Die Abhängigkeit in bezug auf die zweite Variable sei also nichtlinear. Wir setzen nun

$$x_2^2 = x_3$$

Das führt uns zu folgender Gleichung:

$$Y = \bar{y} + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + b_3(x_3 - \bar{x}_3)$$

Die Verwendung von Elektronenrechnern erspart die mühselige Rechenarbeit auf Tischrechnern.

16 Elektronenrechner – Programme

Heute stehen zur Berechnung der notwendigen Hilfszahlen, Parameter, Regressionswerte usw. Elektronenrechner zur Verfügung, die in einem Bruchteil des Zeitaufwandes, der früher für «Handarbeit» benötigt wurde, alle Ergebnisse bereitstellen. Es gibt in der Tat eine Reihe vorzüglicher Programme, die vor allem bei mehrfacher Regression gute Dienste leisten. Ein grosser Teil der untenstehenden Resultate wurde jedoch zu Kontrollzwecken zweifach errechnet, einmal auf dem Tischrechner, zum andern Mal auf dem Computer. Wir wollen uns hier nicht weiter mit Programmierung befassen; die diesbezüglichen Probleme werden in der Spezialliteratur eingehend dargestellt.