.

Regression

Multiple Regression und Generalisierte Lineare Modelle

U. Mortensen

Multivariate Methoden
Psychologisches Institut der Universität Mainz
WS 2011/12

Letzte Korrektur 02. 08. 2020

Inhaltsverzeichnis

1 Einführung

Es soll die Frage diskutiert werden, wie der simultane Einfluß mehrerer Variablen $X_1, \cdots, X_p$ auf eine gegebene "abhängige" Variable Y beschrieben werden kann. Allgemein wird man annehmen, dass Y eine Funktion der $X_1, \cdots, X_p$ sein wird, so dass man

$$Y = f(X_1, \cdots, X_p)$$

schreiben kann; die Aufgabe ist dann, die Funktion f zu bestimmen. Daten legen im Allgemeinen nahe, dass Y auch zufällige Komponenten enthält, wobei "zufällig" einfach nur heißen soll, dass nicht alle Unterschiede zwischen den Y-Werten durch entsprechende Unterschiede zwischen den Werten der unabhängigen Variablen $X_1, \cdots, X_p$ erklärt werden können. Man kann diesen Sachverhalt am einfachsten so modellieren, dass man eine zufällige Veränderliche e einführt und zu dem Modell

$$Y = f(X_1, \cdots, X_p) + e \tag{1.1}$$

übergeht. Die Funktion f beschreibt den systematischen Einfluß der Variablen X_j, $j = 1, \ldots, p$ auf die abhängige Variable Y, und e den zufälligen Effekt in einer Y-Messung. Für die i-te Messung hat man dann

$$y_i = f(x_{i1}, \ldots, x_{in}) + e_i, \tag{1.2}$$

und die $x_{i1}, \ldots, x_{in}$ sind die Werte der unabhängigen Variablen bei der i-ten Messung, und e_i ist die zufällige Komponente, auch einfach "Fehler" genannt. Weil Y durch die $X_1, \cdots, X_p$ gewissermaßenb vorhergesagt wird, spricht man von diesen Variablen auch als von den Prädiktoren.

Wenn der Fehler e zufällig verteilt ist, wird es eine Wahrscheinlichkeitsverteilung für e geben, und man nimmt an, dass durch sie ein Erwartungswert $\mathbb{E}(e)$ und eine Varianz $\sigma^2(e)$ definiert ist. Ohne Beshränkung der Allgemeinheit kann man annehmen, dass $\mathbb{E}(e) = 0$ ist. Denn es sei $\mathbb{E}(e) = \eta \neq 0$. Dann gilt für einen individuellen Fehler $e_i = \eta + \varepsilon_i$; ε_i ist einfach die Abweichung von η. Dann kann man aber

$$y_i = f(x_{i1}, \ldots, x_{in}) + \eta + \varepsilon_i$$

schreiben. Da nun η eine Konstante ist, die nicht von einer Messung zur nächsten variiert, kann man sie in den systematischen Teil $f(x_{i1}, \ldots, x_{in})$ integrieren oder "absorbieren". Indem man ε_i nun wieder in e_i umbenennt ist man wieder bei dem Ansatz (1.2) mit $\mathbb{E}(e) = 0$. Bestimmt man nun den Erwartungswert von Y für gegebene $X_1, \cdots, X_p$, so hat man

$$\mathbb{E}(Y|X_1, \cdots, X_p) = f(X_1, \cdots, X_p). \tag{1.3}$$

Die Funktion f bezieht sich somit auf den Erwartungswert der Verteilung der Y-Werte.

Über die genaue Form von f ist üblicherweise kaum etwas bekannt. Aus der Mathematik weiß man, dass sich die meisten Funktionen durch Polynome beliebig genau approximieren lassen. Für den Spezialfall $p = 1$ haben Polynome die Form

$$h(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

n ist der Grad des Polynoms. Liegt eine Funktion $f(x)$ vor, so versucht man, sie durch ein Polynom möglichst niedrigen Grades zu approximieren:

$$f(x) \approx a_0 + a_1 x, \text{ oder} f(x) \approx a_0 + a_1 x + a_2 x^2, \text{ etc.}$$

Für $p = 2$ kann man analog vorgehen, etwa

$$f(x_1, x_2) \approx a_0 + a_1 x_1 + a_2 x_2.$$

Kann man eine Funktion nicht linear approximieren, so kann man nichtlineare Terme hinzufügen:

$$f(x_1, x_2) \approx a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2,$$

oder

$$f(x_1, x_2) \approx a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_1^2 x_2,$$

Das Bemerkenswerte ist, dass man mit einer linearen Approximation häufig sehr weit kommt und die Addition nichtlinearer Terme nicht nötig ist. Die lineare Approximation führt dann zu dem am häufigsten vorkommenden, nämlich dem linearen Regressionsmodell

$$Y = b_0 + b_1 X_1 + \cdots + b_p X_p + e. \tag{1.4}$$

Dieses Modell kann um Terme erweitert werden, die nichtlinear in den Prädiktoren sind, etwa

$$Y = b_0 + b_1 X_1 + \cdots + b_p X_p + b_{p+1} X_1 X_2 + e,$$

bei dem eine Wechselwirkung zwischen den Prädiktoren X_1 und X_2 hinzugefügt wurde. Formal hat man es immer noch mit einem linearen Modell zu tun, – es ist linear in den Koreffizienten $b_0, b_1, \ldots, b_{p+1}$. Benennt man die X_j in Z_j um und setzt man $Z_{p+1} = X_1 X_2$ so hat man wieder das lineare Modell

$$Y = b_0 + b_1 Z_1 + \cdots + b_p Z_p + b_{p+1} Z_{p+1} + e. \tag{1.5}$$

Wenn sich die folgenden Betrachtungen nur auf das Modell (1.4) beziehen, so werden doch Modelle der Art (1.5) implizit gleich mitbehandelt. Natürlich sind die Prädiktoren Z_1 und Z_2 mit Z_{p+1} korreliert; der Fall korrelierender Prädiktoren wird explizit diskutiert werden.

2 Multiple Regression

Für $p = 1$ hat man die einfache Regression, und für den Fall $p > 1$ spricht man von *multipler Regression* von $Y = X_{k+1}$ auf die X_i, $i = 1, \cdots, k$. Die $b_0, b_1, \cdots, b_p$ sind die *Regressionsgewichte* der *Prädiktoren* X_j, und Y ist die *Kriteriumsvariable*.

Beispiel 2.1 Es werde die Merkfähigkeit ($Y = X_0$) als Funktion von p Variablen $X_1, ..., X_p$ betrachtet. Es sei etwa $X_1 =$ Alter, $X_2 =$ intellektuelles Training, $X_3 =$ Ausmaß von Routinetätigkeiten im Beruf sowie im täglichen Leben, $X_4 =$ Tabakgenuß, $X_5 =$ Alkoholgenuß, und es sei

$$Y = b_1X_1 + b_2X_2 + \cdots + b_5X_5 + b_0 + e.$$

Es liegt nahe, die b_j, $j = 1, ..., k$ als "Gewichte" zu betrachten, mit denen die einzelnen Prädiktoren X_j in die Kriteriumsvariable Y (Merkfähigkeit) eingehen. Es wird aber deutlich werden, dass die Schätzungen $\hat{b}_j$ für die b_j im allgemeinen nicht voneinander unabhängig sind und mithin nicht isoliert, d.h. jeweils unabhängig von den anderen Gewichten, interpretiert werden dürfen.

Der Einfluß der Variablen $X_1, \cdots, X_5$ muß nicht notwendig *additiv* auf die Merkfähigkeit wirken. Es kann sein, dass z.B. die Variablen X_4 und X_5 miteinander interagieren: der Effekt von Alkohol kann proportional zum Effekt des Nikotins in die Gleichung eingehen, und damit ist auch der Effekt des Nikotins proportional zum Effekt des Alkohols. Dann geht eine weitere Variable $X_6 = X_4X_5$ in die Gleichung ein, etwa mit dem Gewsicht b_6. Die Gleichung ist jetzt nichtlinear in den Variablen X_5 und X_6, aber immer noch linear in den unbekannten Gewichten $b_0, b_1, \ldots, b_6$.

Im Übrigen kann die Interaktion zwischen mehr als zwei Variablen auftreten; denkbar sind Terme der Form $X_i^p X_j^q X_p^r$ etc., wobei die Exponenten p, q, r ungleich 1 sein können.

□

2.1 Schätzung der Parameter

Es wird angenommen, dass Meßwerte mit Intervallskalenniveau vorliegen; z.B. wurde für jede der Variablen je ein Meßwert von einer Vp erhoben, oder es wurden $k+1$ Variablen an einer Vp, aber zu n verschiedenen Zeitpunkten gemessen. Dann soll also

$$Y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_px_{ip} + e_i, \tag{2.1}$$

$i = 1, ..., n$ geschrieben werden können. Setzt man

$$\hat{y}_i = b_1x_{i1} + \cdots + b_px_{ip} + b_0 \tag{2.2}$$

so ist

$$Y_i = \hat{y}_i + e_i. \tag{2.3}$$

Man die Gewichte $b_0, b_1, \ldots, b_p$ zu einem Vektor $\mathbf{b} = (b_0, b_1, \ldots, b_p)'$ Messwerte x_{ij} zu einer Matrix X zusammenfassen. Die Messwerte x_{ij} können ebenfalls zu einer Matrix X zusammengefasst werden. Weiter sei $\vec{1} = (1, 1, \ldots, 1)'$ ein Vektor mit genau m (die Anzahl der Vpn bzw. Messungen pro Variable) Einsen. Man kann dann die Matrix

$$X_0 = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ & & & \vdots & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tag{2.4}$$

bilden. Mit $\mathbf{e} = (e_1, e_2, \ldots, e_m)'$ erhält man dann die Matrixgleichung

$$\mathbf{y} = X_0\mathbf{b} + \mathbf{e}, \tag{2.5}$$

in der alle Gleichungen (2.1) zusammengefasst sind. Setzt man $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_m)'$, so ergibt sich

$$\hat{\mathbf{y}} = X_0\mathbf{b}, \quad \mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}. \tag{2.6}$$

Die Komponenten des Vektors $\mathbf{b}$ müssen nun geschätzt werden. Dies geschieht durch Anwendung der Methode der Kleinsten Quadrate. Die Komponenten werden so bestimmt, dass die Summe der Fehlerquadrate minimalisiert wird. Es ist ja nach (2.3) $e_i = y_i - \hat{y}_i$, also ist

$$Q(b_0, b_1, \ldots, b_p) = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} (y_i - (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}))^2.$$

In Matrixform erhält man

$$Q(b_0, b_1, \ldots, b_p) = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}). \tag{2.7}$$

Die $b_0, b_1, \ldots, b_p$ werden so bestimmt, dass $Q(b_0, b_1, \ldots, b_p)$ den kleinstmöglichen Wert annimmt. Man erhält das im folgenden Satz zusammengefasste Ergebnis:

Satz 2.1 *Es gelte (2.1) und die Variablen Y, $X_1, ..., X_p$ mögen Intervallskalenniveau haben. Die Kleinste-Quadrate-Schätzungen $\hat{b}_1, ..., \hat{b}_p$ für $b_1, ..., b_p$ ergeben sich als die Lösungen des Systems von n linearen Gleichungen*

$$\begin{aligned} \sum_{i=1}^{n} y_i x_i &= \hat{b}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{b}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{b}_k \sum_{i=1}^{n} x_{1i}x_{ip} \\ \sum_{i=1}^{n} y_i x_{i2} &= \hat{b}_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \hat{b}_2 \sum_{i=1}^{n} x_{i2}^2 + \cdots + \hat{b}_k \sum_{i=1}^{n} x_{i2}x_{ip} \\ &\vdots \\ \sum_{i=1}^{n} y_i x_{ip} &= \hat{b}_1 \sum_{i=1}^{n} x_{i1}x_{ip} + \hat{b}_2 \sum_{i=1}^{n} x_2 x_{ip} + \cdots + \hat{b}_k \sum_{i=1}^{n} x_{ip}^2 \end{aligned} \tag{2.8}$$

Beweis: Anhang (Abschnitt 4), insbesondere Abschnitt 4.1. □

Die additive Konstante $\hat{b}_0$ ergibt sich dann aus (2.2) gemäß

$$\hat{b}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{b}_1 \frac{1}{n}\sum_{i=1}^{n} x_{i1} - \cdots - \hat{b}_p \frac{1}{n}\sum_{i=1}^{n} x_{ip},$$

d.h.

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}_1 - \cdots - \hat{b}_p \bar{x}_p,$$

so dass

$$\bar{y} = \hat{b}_0 + \hat{b}_1 \bar{x}_1 + \hat{b}_2 \bar{x}_2 + \cdots + \hat{b}_p \bar{x}_p + \hat{b}_0. \tag{2.9}$$

Nach (2.2) gilt aber $\hat{y}_i = b_1 x_{i1} + \cdots + b_p x_{ip} + b_0$; summiert man über alle i und teilt man dann durch m, so erhält man

$$\bar{\hat{y}} = b_1 \bar{x}_1 + \cdots + b_p \bar{x}_p + \hat{b}_0. \tag{2.10}$$

Auf der rechten Seite von (2.9) steht aber gerade der Mittelwert der vorhergesagten Werte $\hat{y}_i$, so dass sich wieder

$$\bar{\hat{y}} = \bar{y} \tag{2.11}$$

ergibt.

Im Gleichungssystem (2.8) treten die "Kreuzprodukte"

$$\sum_i y_i x_{ji}, \quad \sum_i x_{si} x_{ji}, \quad \sum_i x_{ij}^2$$

auf, die auf die Varianzen s_j^2 und die Kovarianzen $Kov(Y, X_j)$ und $Kov(X_s, X_j)$ verweisen. Dies legt nahe, das System (2.8) nicht so, wie es dort steht zu lösen, sondern dabei von den Korrelationen r_{yj} und r_{sj} auszugehen. Korrelationen ergeben sich aber immer durch Übergang von den Rohwerten Y_i und X_{ji} zu den entsprechenden standardisierten Werten. Es sei nun

$$y_i = \hat{b}_1 x_{i1} + \hat{\mathbf{b}}_2 x_{i1} + \cdots + \hat{b}_p x_{ip} + e_i$$

und $z_{0i} = (Y_i - \hat{y})/s_y$. Wegen (2.9) findet man sofort

$$z_{0i} = \frac{(y_i - \hat{y})}{s_y} = \hat{b}_1 \frac{1}{s_y}(x_{1i} - \hat{x}_1) + \cdots + \hat{b}_p \frac{1}{s_y}(x_{ki} - \hat{x}_k). \tag{2.12}$$

Um den Übergang von den Differenzen $(X_{ji} - \hat{x}_j)$ zu den $z_{ji} = (X_{ji} - \hat{x}_j)/s_j$ zu erreichen, multipliziert man in (2.12) jede Differenz mit der entsprechenden Streuung. Dividiert man gleichzeitig durch diese Streuung, so bleibt (2.12) jedenfalls richtig:

$$z_{0i} = \hat{b}_1 \frac{s_1}{s_y} \frac{x_{i1} - \hat{x}_1}{s_1} + \cdots + \hat{b}_p \frac{s_k}{s_y} \frac{x_{ip} - \hat{x}_k}{s_k} + \frac{e_i}{s_y}$$

und damit erhält man

$$z_{0i} = \hat{b}_1 \frac{s_1}{s_y} z_{i1} + \cdots + \hat{b}_p \frac{s_k}{s_y} z_{ip} + \epsilon_i, \tag{2.13}$$

wobei $\epsilon_i = e_i/s_y$. Man definiert nun

$$\hat{\beta}_j := \hat{b}_j \frac{s_j}{s_y}, \tag{2.14}$$

wobei die $\hat{\beta}_j$ Schätzungen der $\beta_j = b_j s_j / s_y$ sind. Damit hat man

$$z_{0i} = \hat{\beta}_1 z_{i1} + \hat{\beta}_2 z_{i2} + \cdots + \hat{\beta}_p z_{ip} + \epsilon_i. \tag{2.15}$$

Für die Kleinste-Quadrate-Schätzungen $\hat{\beta}_j$ gilt nun der folgende

Satz 2.2 *Die Schätzungen $\hat{\beta}_j$ in (2.15) ergeben sich aus (2.8) als Lösungen des linearen Gleichungssystems*

$$\begin{aligned} r_{y1} &= \hat{\beta}_1 + \hat{\beta}_2 r_{12} + \cdots + \hat{\beta}_k r_{1p} \\ r_{y2} &= \hat{\beta}_1 r_{21} + \hat{\beta}_2 + \cdots + \hat{\beta}_k r_{2p} \\ &\vdots \\ r_{yp} &= \hat{\beta}_1 r_{p1} + \hat{\beta}_2 r_{p2} + \cdots + \hat{\beta}_p \end{aligned} \tag{2.16}$$

wobei r_{yj} die Korrelation zwischen der Kriteriumsvariablen Y und der j-ten Prädiktorvariablen X_j ist und die $r_{sj} = r_{js}$ die Korrelationen zwischen der s-ten und der j-ten Prädiktorvariablen darstellen.

Beweis: Den Beweis findet man in Abschnitt 4.2, p. 74. □

2.2 Die Matrixschreibweise

Verwendet man die Matrixschreibweise[1], läßt sich die Schätzung der Regressionsgewichte in sehr kompakter Form angeben. Darüber hinaus kann die Frage der Eindeutigkeit der Schätzungen einfacher behandelt werden. Allgemein gilt ja, nach (2.5), $\mathbf{y} = X_0\mathbf{b} + \mathbf{e}$, und dementsprechend muß nach (2.7)

$$Q(\mathbf{b}) = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - X_0\mathbf{b})'(\mathbf{y} - X_0\mathbf{b})$$

minimalisiert werden. $\mathbf{b}$ soll so bestimmt werden, dass $Q(\mathbf{b})$ minimalisiert wird. $\mathbf{b}$ ist durch seine Komponenten bestimmt, also müssen diese Komponenten entsprechend gewählt werden. $\mathbf{b}$ verändert sich, wenn die Komponenten verändert werden. Die Veränderung von $\mathbf{b}$ mit einer Komponente b_j wird durch die partielle Ableitung $\partial\mathbf{b}/\partial b_j$ angegeben. Dies bedeutet, dass

[1] Vergl. das Skriptum *Vektoren und Matrizen*

man die Veränderung aller Komponenten b_p mit b_j betrachtet, die wiederum durch die Ableitungen db_p/db_j gegeben sind. Aber die b_p hängen nicht von b_j ab, also folgt $db_p/db_j = 0$ für alle $k \neq j$, und $db_p/db_j = 1$ für $k = j$. Damit erhält man

$$\frac{\partial \mathbf{b}}{\partial b_j} = (0, \ldots, 0, 1, 0, \ldots, 0)' = \epsilon_j. \tag{2.17}$$

Dies ist die partielle Ableitung von $\mathbf{b}$ nach b_j. Es entsteht also der j-te Einheitsvektor ϵ_j, dessen Komponenten alle gleich Null sind bis auf die j-te, die gleich 1 ist, also $\epsilon_j = (0, \ldots, 0, 1, 0, \ldots, o)'$, und die 1 steht an der j-ten Stelle. Man wird dann auf das folgende Ergebnis geführt:

Satz 2.3 *Die Schätzung für den Vektor* $\boldsymbol{b}$ *ist durch*

$$\hat{\boldsymbol{b}} = (X_0' X_0)^{-1} X_0' \boldsymbol{y} \tag{2.18}$$

gegeben. Standardisiert man die x_{ij}*-Werte, so geht die Matrix* X *in die Matrix* Z *über, und* $\boldsymbol{b}$ *geht in den Vektor* $\vec{\beta}$ *der* β*-Gewichte über. Dabei verschwindet die additive Konstante* b_0*, und (4.8) wird zu*

$$\vec{\hat{\beta}} = (Z'Z)^{-1} Z' \vec{Z}_y, \tag{2.19}$$

wobei $\vec{Z}_y$ *der Vektor der standardisierten* y*-Werte ist.*

Beweis: Den Beweis findet man in Abschnitt 4.3, p. 75.

Dividiert man durch m (und absorbiert man den Faktor $1/m$ in den Vektor $\vec{\hat{\beta}}$, so erhält man wegen $R_{xx} = Z'Z/m$, R_{xx} die Matrix der Korrelationen zwischen den Prädiktorvariablen, und $R_{xy} = Z'Z_y/m$ der Vektor der Korrelationen zwischen den Prädiktoren und der Kriteriumsvariablen,

$$\vec{\hat{\beta}} = R_{xx}^{-1} \vec{R}_{xy}. \tag{2.20}$$

Von diesem Ausdruck läßt sich sofort ablesen, wie die Komponenten des Vektors $\vec{\hat{\beta}}$ von den Korrelationen der Prädiktoren abhängen. Sind diese nämlich unkorreliert, d.h. ist $R_{xx} = I$ die Einheitsmatrix, so ist auch R_{xx}^{-1} eine Einheitsmatrix und man erhält

$$\vec{\hat{\beta}} = \vec{R}_{xy}, \quad R_{xx} = I, \tag{2.21}$$

d.h. die β-Gewichte sind durch die Korrelationen der entsprechenden Prädiktorvariable mit der Kriteriumsvariablen gegeben. Dieses Resultat kann man auch dem Gleichungssystem (2.16) entnehmen, wenn man nämlich dort $r_{ij} = 0$ für $i \neq j$ setzt.

Zum Abschluß soll auf einen geometrischen Aspekt der Schätzung für b bzw. β hingewiesen werden. Nach (2.6) gilt ja $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$, wobei $\hat{\mathbf{y}}$ die Kleinste-Quadrate-Schätzung für $\mathbf{y}$ ist. Dann folgt

Satz 2.4 *Die Vektoren* $\vec{\hat{Y}}$ *und* $\boldsymbol{e}$ *sind orthogonal, d.h. es gilt*

$$\vec{\hat{Y}}\,'\boldsymbol{e} = 0 \tag{2.22}$$

Beweis: Nach (4.8) ist $\hat{\mathbf{b}} = (X_0'X_0)^{-1}X_0'\mathbf{y}$, so dass

$$\hat{\mathbf{y}} = X_0\hat{\mathbf{b}} = X_0(X_0'X_0)^{-1}X_0'\mathbf{y}. \tag{2.23}$$

Weiter ist $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. Dementsprechend ist

$$\hat{\mathbf{y}}'\mathbf{e} = (X_0(X_0'X_0)^{-1}X_0'\mathbf{y})'(\mathbf{y}-\hat{\mathbf{y}}) = \mathbf{y}\,'X_0(X_0'X_0)^{-1}X_0'\mathbf{y} - \vec{y}\,'X_0(X_0'X_0)^{-1}X_0'\hat{\mathbf{y}}$$

Substituiert man hier für $\hat{\mathbf{y}}$ den in (2.23) gegebenen Ausdruck, so ergibt sich die Aussage (2.22). □

Anmerkungen:

1. **Korrelation von Vorhersage und Fehler:** Die Methode der Kleinsten Quadrate bestimmt die Gewichte also derart, dass die Vorhersage $\hat{\mathbf{y}}$ des Kriteriums *nicht* mit den Fehlern $\mathbf{e}$ korreliert.

2. **Projektion:** Nach (2.23) gilt

   $$\hat{\mathbf{y}} = X_0(X_0'X_0)^{-1}X_0'\mathbf{y}.$$

   Die Gleichung beschreibt die Transformation des Vektors $\mathbf{y}$ in den Vektor $\hat{\mathbf{y}}$ mittels der Transformationsmatrix

   $$P_r = X_0(X_0'X_0)^{-1}X_0'. \tag{2.24}$$

   Die Matrix P_r heißt auch *Projektionsmatrix*. P_r ist *idempotent*, d.h. $P_rP_r = P_r^2 = P_r$:

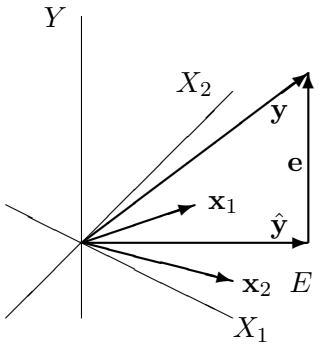
   $$P_r^2 = X_0(X_0'X_0)^{-1}X_0'X_0(X_0'X_0)^{-1}X_0' = X_0(X_0'X_0)^{-1}X_0' = P_r.$$

   Der Name erklärt sich aus der Art der Transformation, die durch P_r vorgenommen wird. Der Vektor $\hat{\mathbf{y}}$ ist ja eine Linearkombination der Vektoren, die als Prädiktoren fungieren:

   $$\hat{\mathbf{y}} = b_0 + b_1\vec{X}_1 + \cdots + b_n\vec{X}_n.$$

   Geometrisch gesehen ist $\hat{\mathbf{y}}$ ein Vektor, der im gleichen Teilraum des V_m liegt wie die $\vec{X}_1, \ldots, \vec{X}_n$ (d.h. $\hat{\mathbf{y}}$ liegt in dem von den $\vec{X}_1, \ldots, \vec{X}_n$ aufgespannten Teilraum $\mathcal{C}(\vec{X}_1, \ldots, \vec{X}_n)$ des V_n). Der Vektor $\vec{Y}$ liegt aber *nicht* in diesem Teilraum: Da $\vec{Y} = \hat{\mathbf{y}} + \mathbf{e}$ und $\mathbf{e}$ senkrecht auf $\hat{\mathbf{y}}$ steht, ist eine zusätzliche Dimension zur Beschreibung von $\vec{Y}$ nötig. Projiziert man nun $\vec{Y}$ auf den $\mathcal{C}(\vec{X}_1, \ldots, \vec{X}_n)$, so erhält man gerade $\hat{\mathbf{y}}$.

Abbildung 1: $\hat{y}$ als Projektion von y auf $C(X)$



$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$

$\hat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 = \vec{\xi}$

Die Projektion ist aber eine Vektortransformation, denn $\vec{Y}$ wird ja in den Vektor $\hat{\mathbf{y}}$ transformiert, und die Transformationsmatrix, die diese Transformation bewirkt, ist nun gerade P_r, – weshalb sie *Projektionsmatrix* heißt. Abbildung 1 soll den Sachverhalt verdeutlichen.

□

Die Ebene E werde durch die Koordinatenachsen X_1 und X_2 definiert und die vorgegebenen Vektoren $\mathbf{x}_1$ und $\mathbf{x}_2$ mögen in dieser Ebene liegen (in anderen Worten: die Achsen X_1 und X_2 sind einfach so gewählt worden, dass sie in der gleichen Ebene wie der durch die Vektoren $\mathbf{x}_1$ und $\mathbf{x}_2$ gebildeten liegen); E enthält dann alle möglichen Linearkombinationen

$$C(X) = \{\vec{\xi}|\vec{\xi} = b\mathbf{x}_1 + c\mathbf{x}_2, \quad \mathbf{x}_1, \mathbf{x}_2 \in E\};$$

Der Vektor $\mathbf{y}$ enthält die von $\mathbf{x}_1$ und $\mathbf{x}_2$ unabhängigen Komponenten e und ist demnach 3-dimensional.

2.3 Eigenschaften der Schätzungen

Es wird zunächst die Annahme gemacht, dass die X_j messfehlerfrei sind. Dann gilt der

Satz 2.5 *Es gelte $Y = Xb + e$ mit $\mathbb{E}(e) = 0$ (Mit $\mathbb{E}$ wird im Folgenden der Erwartungswert bezeichnet). $Kov(e) = \sigma^2 I$, I die Einheitsmatrix, d.h. die Korrelationen zwischen den Fehlern seien alle gleich Null. Dann gilt*
(i) Die die Kleinste-Quadrate-Schätzung $\hat{b} = (X'X)^{-1}X'Y$ für b ist erwartungstreu, d.h. es gilt

$$\mathbb{E}(\hat{b}) = b, \tag{2.25}$$

(ii) Die Kovarianz der Schätzungen $\hat{b}_j$ ist durch

$$Kov(\hat{b}) = \sigma^2(X'X)^{-1} \tag{2.26}$$

gegeben,
(iii) Die Größe

$$s^2 = \frac{1}{n-m}(Y - X\hat{b})'(Y - X\hat{b}) \tag{2.27}$$

ist ein unverzerrter Schätzer für σ^2.

Beweis: s. Anhang. □

Gauß-Markov-Theorem Es gilt der Satz von Gauß-Markov:

Satz 2.6 *Die KQ-Schätzung* $\hat{b}$ *für den unbekannten Parametervektor hat die kleinste Varianz unter allen linearen, verzerrungsfreien (bias-freien) Schätzungen für* b.

Beweis: z.B. Seber (1977) □

Der Ausdruck 'minimale Varianz' bedeutet, dass andere Schätzungen als die durch (2.27) gegebene eine mindestens so große Varianz der $\hat{b}_j$ implizieren, – aber dies heißt nicht, dass diese Varianz notwendig auch klein ist. (2.26) zeigt, dass die Varianzen und Kovarianzen der $\hat{b}_j$ von der Matrix $(X'X)^{-1}$ abhängen. Sind die Kovarianzen der Prädiktoren gleich Null, so ist $X'X$ und damit auch $(X'X)^{-1}$ eine Diagonalmatrix.

Es sei $L = \hat{b} - b$ der Vektor der Abweichungen $l_j = \hat{b}_j - b_j$, $j = 1, \ldots, m$. Dann ist

$$\|L\|^2 = (\hat{b} - b)'(\hat{b} - b) \tag{2.28}$$

das Quadrat der Länge $\|L\|$. Je größer der Wert von $\|L\|$, desto größer sind die Abweichungen der Schätzungen von den wahren Werten. Man findet

$$\mathbb{E}(L) = \sigma^2 sp(X'X)^{-1}, \tag{2.29}$$

$$\mathbb{E}(\hat{b}'\hat{b}) = b'b + \sigma^2 sp(X'X)^{-1} \tag{2.30}$$

$$Var(L) = 2\sigma^4 sp(X'X)^{-1}. \tag{2.31}$$

Hier steht *sp* für 'Spur'; die Spur einer Matrix ist die Summe der Diagonalelemente der Matrix.

Die Gleichung (2.29) folgt sofort aus der Gleichung (2.26), da die Diagonalzellen von $Kov(\hat{b})$ gerade die $(\hat{b}_j - b_j)^2$ enthalten ($\hat{b}_j$ und b_j sind die j-te Komponente von $\hat{b}$ bzw. b), und $\|L\|^2 = (\hat{b} - b)'(\hat{b} - b)$ ist dann gerade gleich der Summe der Diagonalzellen von $(X'X)^{-1}$ (also die Spur von $(X'X)^{-1}$), multipliziert mit σ^2. $X'X$ ist symmetrisch und läßt sich in der Form $X'X = V\Lambda V'$ darstellen, wobei V die $(m \times m)$-Matrix der orthonormalen Eigenvektoren von $X'X$ ist und Λ eine $(m \times m)$-Diagonalmatrix ist, in deren Diagonalzellen die zugehörigen Eigenwerte λ_j stehen. Die Gleichung (2.30) zeigt, dass die Länge von $\hat{b}$ stets größer als die Länge von b ist. Dieses Ergebnis ist plausibel, weil wegen des Schätzfehlers die $\hat{b}_k$ von den b_k abweichen. Die Abweichung ist proportional zu σ^2, aber, wie (2.31) zeigt, ist $Var(L)$ auch proportional zu $sp(X'X)^{-1}$.

In den Gleichungen (2.29), (2.30) und (2.31) tritt die Spur $sp(X'X)^{-1}$ auf. Die Spur hängt von den Eigenwerten von $(X'X)^{-1}$ ab, und die wiederum von den Abhängigkeiten zwischen den Prädiktoren, also den Spalten von X. Die Beziehung zwischen der Spur und den Eigenwerten erhellt die Rolle dieser Abhängigkeiten und wird deshalb explizit gemacht.

Die Eigenwerte von $X'X$ und die Eigenschaften von $\hat{b}$: Bekanntlich gilt die Gleichung

$$X'X = V\Lambda V', \tag{2.32}$$

wobei V die orthonormale Matrix der Eigenvektoren von $X'X$ und Λ die Diagonalmatrix der zugehörigen Eigenwerte ist. Für die inverse Matrix $(X'X)^{-1}$ hat man dann

$$(X'X)^{-1} = (V')^{-1}\Lambda^{-1}V^{-1} = V\Lambda^{-1}V',$$

denn wegen der Orthonormalität von V gilt ja $V' = V^{-1}$. Insbesondere kann man dann

$$(X'X)^{-1} = V\Lambda^{-1}V' = \sum_{k=1}^{m} \frac{\mathbf{V}_k\mathbf{V}'_k}{\lambda_k}, \tag{2.33}$$

schreiben, wobei $\mathbf{V}_k$ der k-te Eigenvektor von $X'X$ bzw. $(X'X)^{-1}$ – also die k-te Spalte von V – ist, und $\mathbf{V}_k\mathbf{V}'_k$ ist das dyadische Produkt von $\mathbf{V}_k$ mit sich selbst. Der Wert in der k-ten Diagonalzelle hat die Form

$$\frac{v_{k1}^2}{\lambda_1} + \frac{v_{k2}^2}{\lambda_2} + \cdots + \frac{v_{kn}^2}{\lambda_n},$$

wobei v_{kj} die k-te Komponente des j-ten Eigenvektors ist. Also gilt

$$sp(X'X)^{-1}) = \sum_{k=1}^{n} \left(\frac{v_{k1}^2}{\lambda_1} + \frac{v_{k2}^2}{\lambda_2} + \cdots + \frac{v_{kn}^2}{\lambda_n} \right), \tag{2.34}$$

und für (2.29) folgt

$$\mathbb{E}(L) = \sigma^2 \sum_{k=1}^{n} \left(\frac{v_{k1}^2}{\lambda_1} + \frac{v_{k2}^2}{\lambda_2} + \cdots + \frac{v_{kn}^2}{\lambda_n} \right). \tag{2.35}$$

Ein individueller Summand entspricht dem Erwartungswert $\mathbb{E}[(\hat{b}_k - b_k)^2] = Var(\hat{b}_k)$, so dass man insbesondere

$$Var(\hat{b}_k) = \sigma^2 \left(\frac{v_{k1}^2}{\lambda_1} + \frac{v_{k2}^2}{\lambda_2} + \cdots + \frac{v_{kn}^2}{\lambda_n} \right), \quad k = 1, \ldots, m \tag{2.36}$$

erhält; $v_{k1}, \ldots, v_{kn}$ ist gerade die k-te Zeile der Matrix V der Eigenvektoren von $X'X$. Die Varianz – und das heißt, die Ungenauigkeit der Schätzung – wird also groß, wenn zumindest einige Eigenwerte klein werden. Dies gilt

natürlich auch für die erwartete Länge des Vektors $\hat{b} - b$. Dies ist dann der Fall, wenn Abhängigkeiten zwischen den Prädiktoren bestehen, wie im Folgenden gezeigt wird.

Für die Schätzungen $\hat{\beta}$ gelten die analogen Aussagen:

Satz 2.7 *Es sei $\mathbb{E}(\hat{\beta})$ der Vektor der Erwartungswerte der $\hat{\beta}_i$, d.h.*

$$\mathbb{E}(\hat{\beta}) = (\mathbb{E}(\hat{\beta}_1), \ldots, \mathbb{E}(\hat{\beta}_n))',$$

und $\mathbb{D}(\hat{\beta})$ sei die Matrix der Varianzen und Kovarianzen der Schätzungen, d.h. das (j,k)-te Element von $\mathbb{D}$ sei $\mathbb{E}(\hat{\beta}_j \hat{\beta}_k)$. Dann gelten die Aussagen

$$\mathbb{E}(\hat{\beta}) = \beta \tag{2.37}$$
$$\mathbb{D}(\hat{\beta}) = \sigma^2 R^{-1}, \tag{2.38}$$

wobei R die Matrix der Korrelationen zwischen den Prädiktoren ist. Eine analoge Aussage gilt für die nicht standardisierten Prädiktoren und die nicht standardisierte Kriteriumsvariable.

Eine Ursache für die Korrelationen zwischen den Prädiktoren sind *Multikollinearitäten*. Diese werden durch lineare Abhängigkeiten zwischen den Prädiktoren mit relativ kleinem additiven Fehler erzeugt. Multikollinearitäten stellen ein Problem für die Interpretation der Regressionsgewichte dar, weshalb etwas ausführlicher auf sie eingegangen werden soll.

2.4 Multikollinearität

Die Prädiktoren $X_1, \ldots, X_r$ seien nicht-stochastisch. Von (strenger oder exakter) Multikollinearität wird gesprochen, wenn einer der Spaltenvektoren von X – also eine der Variablen X_p, $1 \leq k \leq r$, als Linearkombination der übrigen Vektoren dargestellt werden kann,

$$X_p = \lambda_1 X_1 + \cdots + \lambda_{k-1} X_{k-1} + \lambda_{k+1} X_{k+1} + \cdots + \lambda_r X_r. \tag{2.39}$$

Der Fall kann sich ergeben, wenn entweder ein Fehler in der Variablenauswahl unterlaufen ist, oder – wahrscheinlicher – wenn einige der unabhängigen Variablen Dummy-Variablen sind, also nur das Vorhandensein eines Faktors bzw. der Stufe eines Faktors anzeigen, wie es bei der Darstellung der Varianzanalyse als Allgemeines Lineares Modell vorkommt, und sich die Werte der Dummy-Variablen zu 1 ergänzen. Die Matrix $X'X$ hat dann keinen vollen Rang, d.h.mindestens einer der Eigenwerte von $X'X$ ist gleich Null, und die inverse Matrix $(X'X)^{-1}$ existiert nicht, so dass keine Lösung wie in (4.8) bzw. (4.9) für die Regressionskoeffizienten gefunden werden kann. Die einfachste Lösung ist, den entsprechenden Prädiktor wegzulassen, da er sowieso keine Information bezüglich der unabhängigen Variablen liefert.

Im allgemeinen Fall ist die strenge Multikollinearität (2.39) unwahrscheinlich. Es kann aber eine "Fast-Multikollinearität" eintreten, wenn etwa

$$X_j = \alpha X_p + \beta + \varepsilon, \quad j \neq k, \ \varepsilon \neq 0 \tag{2.40}$$

gilt und die Varianz $Var(\varepsilon)$ einigermaßen klein ist. Man kann auch zu den standardisierten Werten übergehen und erhält dann

$$Z_j = r_{jk} Z_k + \tilde{\varepsilon}, \quad r_{jk} = \frac{\alpha s_k}{s_j} \tag{2.41}$$

und die entsprechenden β-Gewichte der Prädiktoren betrachten.

Beispiel 2.2 Es wird der Fall nur zweier, aber korrelierter Prädiktoren betrachtet[2]:

$$y_i = b_1 x_{i1} + b_2 x_{i2} + e_i, \tag{2.42}$$

wobei die y_i, x_{i1}, x_{i2} bereits Abweichungen vom jeweiligen Mittelwert seien, so dass $\sum_i x_{i1} = \sum_i x_{i2} = 0$. Zwischen den $x_{i1}-$ und x_{i2}-Werte bestehe die Beziehung

$$x_{i2} = \alpha x_{i1} + v_i \tag{2.43}$$

v_i ein Fehler, und α definiert die Kollinearität. Es gelte

$$\sum_i x_{i1}^2 = \sum_i x_{i2}^2 = 1, \ \sum_i v_i = 0, \ \sum_i v_i x_{i2} = 0,$$

d.h. Prädiktor und Fehler seien unkorreliert. Die Matrizen $X'X$ und $(X'X)^{-1}$ haben nun die Form

$$\mathbf{X}'X = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}, \quad (X'X)^{-1} = \frac{1}{1-\alpha^2} \begin{pmatrix} 1 & -\alpha \\ -\alpha & 1 \end{pmatrix}$$

Aus Gleichung (2.26) folgt

$$Kov(\hat{b}) = \frac{\sigma^2}{1-\alpha^2} \begin{pmatrix} -\alpha & 0 \\ 0 & -\alpha \end{pmatrix}. \tag{2.44}$$

Die Varianz der Schätzungen $\hat{b}_1$ und $\hat{b}_2$ ergeben sich daraus als

$$Var(\hat{b}_1) = Var(\hat{b}_2) = \frac{\sigma^2}{1-\alpha^2}, \tag{2.45}$$

und für die Kovarianz zwischen $\hat{b}_1$ und $\hat{b}_2$ erhält man

$$Kov(\hat{b}_1, \hat{b}_2) = \frac{-\alpha\sigma^2}{1-\alpha^2}. \tag{2.46}$$

[2]Johnstone (1963, p. 161

Die Gleichung (2.45) zeigt, dass für $\alpha \to 1$ die Stichprobenvarianz der Schätzungen $\hat{b}_1$ und $\hat{b}_2$ beliebig groß werden, d.h. je höher die beiden Prädiktoren korrelieren, desto ungenauer werden die Schätzungen der Regressionsgewichte. Die Kovarianz zwischen ihnen ist negativ, wird aber ebenfalls beliebig groß, d.h. wird $\hat{b}_1 > 0$, so folgt $\hat{b}_2 < 0$ und umgekehrt.

Die KQ-Schätzung $\hat{b}$ ist durch

$$\hat{b} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(Xb + e) = (X'X)^{-1}(X'X)b + (X'X)^{-1}e$$

gegeben, d.h.

$$\hat{b} - b = (X'X)^{-1}X'e.$$

Inbesondere erhält man für die Komponenten

$$\begin{pmatrix} \hat{b}_1 - b_1 \\ \hat{b}_2 - b_2 \end{pmatrix} = \frac{1}{1-\alpha^2} \begin{pmatrix} 1 & -\alpha \\ -\alpha & 1 \end{pmatrix} \begin{pmatrix} \sum_i x_{i1}e_i \\ \sum_i x_{i2}e_i \end{pmatrix},$$

d.h.

$$\begin{aligned} \hat{b}_1 - b_1 &= \frac{1}{1-\alpha^2}\left(\sum_i x_{i1}e_i - \alpha \sum_i x_{i2}e_i\right) \\ \hat{b}_2 - b_2 &= \frac{1}{1-\alpha^2}\left(\sum_i x_{i2}e_i - \alpha \sum_i x_{i1}e_i\right) \end{aligned}$$

Berücksichtigt man nun (2.43), so erhält man aus diesen Gleichungen

$$\begin{aligned} \hat{b}_1 - b_1 &= \sum_i x_{i1}e_i - \frac{\alpha}{1-\alpha^2}\sum_i u_i e_i \\ \hat{b}_2 - b_2 &= \frac{\alpha}{1-\alpha^2}\sum_i u_i e_i. \end{aligned}$$

In beiden Gleichungen tritt der Term $\sum_i u_i e_i$ auf und erzeugt deshalb die Kovarianz zwischen $\hat{b}_1$ und $\hat{b}_2$. Ein großer Wert von α erzeugt große und entgegengesetzt wirkende Fehler in den $\hat{b}_j$. Ist $\hat{b}_1$ eine Unterschätzung von b_1, so wird b_2 durch $\hat{b}_2$ überschätzt und umgekehrt.

Das Prinzip überträgt sich auf den Fall von mehr als zwei Prädiktoren. □

Kleine λ_j-Werte sind indikativ für Fast-Multikollinearität. Diese bewirkt also eine Inflationierung der Varianzen der Schätzungen für die Regressionsparameter. Dies bedeutet, dass die t-Werte kleiner und eventuell insignifikant werden.

Es gibt verschiedene Möglichkeiten, die Problematik der Kollinearität anzugehen; Silvey (1969) schlägt die Wahl zusätzlicher, aber speziell gewählter Prädiktoren vor. In vielen Fällen ist ein solches Vorgehen aber kaum möglich, da oft die die Menge der Prädiktoren fest vorgegeben ist. Der im folgenden Abschnitt vorgestellte Ansatz ist daher von größerem praktischen Nutzen.

2.5 Regularisierung und Ridge-Regression

2.5.1 Tychonoff-Regularisierung

Die übliche KQ-Schätzung für den Vektor b hat die Form

$$\hat{\mathbf{b}} = (X'X)^{-1}X'Y.$$

Existieren Multikollinearitäten zwischen den Prädiktoren, so ist im Extremfall der Rang von $X'X$ kleiner als n (X sei eine m,n-Matrix, so dass für $m > n$ der Rang maximal gleich n ist, und dann ist der Rang von $X'X$ gleich n, – dies ist eine notwendige Voraussetzung für die Existenz der Inversen $(X'X)^{-1}$. Nur wenn diese existiert, kann die Schätzung $\hat{\mathbf{b}}$ tatsächlich berechnet werden. Existieren Multikollinearitäten, so können einige der Eigenwerte klein werden und, wie oben beschrieben, große Varianzen und negative Kovarianzen zwischen den Prädiktoren erzeugen. Sind einige der Eigenwerte "klein", entstehen Ungenauigkeiten bei der Berechnung der Inversen. Man sagt, das Problem (hier der Schätzung des Vektors $\vec{b}$) sei schlecht formuliert ("ill posed"). Nun ist der Ausdrück für $\hat{\mathbf{b}}$ die Lösung eines linearen Gleichungssystems, und das schlecht formulierte Probleme können generell bei der Lösung von Gleichungssystemen auftreten, das Problem ist nicht charakteristisch für die Methode der Kleinsten Quadrate. Man kann

$$(X'X)\hat{\mathbf{b}} = A\hat{\mathbf{b}} = X'Y$$

schreiben, mit $A = X'X$; da $X'Y$ ein Vektor ist, hat man damit ein lineares Gleichungssystem mit m Unbekannten, nämlich den Komponenten des Vektors $\hat{\mathbf{b}}$. Bei solchen Gleichungssystemen stellt sich generell die Frage, wie gut die Lösung ist, falls überhaupt eine Lösung existiert. So möchte man im Allgemeinen eine Lösung $\hat{\mathbf{b}}$, die nicht stark von $X'Y$ abhängt. Nimmt man für den Augenblick an, dass die Matrix X der Prädiktoren fest vorgegeben ist (wie bei varianzanalytischen Fragestellungen), so ist damit gemeint, dass $\hat{\mathbf{b}}$ nur wenig von den Messungen Y abhängen sollte. Ist diese relative Unabhängigkeit gegeben, so ist das Problem, eine Lösung für $\hat{\mathbf{b}}$ zu finden, 'gut konditioniert' ('well-conditioned'), andernfalls ist das Problem 'schlecht konditioniert' ('ill-conditioned'). Die Konditioniertheit des Problems, eine Lösung $\hat{\mathbf{b}}$ zu finden, hängt hier im Wesentlichen von der Matrix X ab. Da $\hat{\mathbf{b}} = A^{-1}X'Y = (X'X)^{-1}X'Y$, muß auf jeden Fall die Inverse $(X'X)^{-1}$ existieren. Wie im Abschnitt über Multikollinearität ausgeführt wurde, existiert zwar üblicherweise die Inverse schon wegen der Fehler in den Y-Messungen, aber $X'X$ kann gleichwohl schlecht konditioniert sein, wenn nämlich zumindest einige Eigenwerte klein werden.

Tychonoff-Regularisierung Ein Ansatz, stabile Lösungen für schlecht konditionierte Gleichungssysteme zu finden, geht auf Tychonoff (1943)[3] zurück,

[3] Andrey Nikolayevitch Tychonoff (1906 – 1993), russischer Mathematiker

dessen Arbeit aber erst 1963 in übersetzter Form vorlag; die allgemeine Version erschien in Tychonoff & Arsenin (1977). Im hier gegebenen Zusammenhang hat man das Modell $\mathbf{Y} = X\mathbf{b} + \mathbf{e}$ und man möchte $\mathbf{b}$ so bestimmen, dass $\mathbf{e}'\mathbf{e}$ minimal wird. Da, und $\mathbf{e}'\mathbf{e} = (\mathbf{Y} - X\mathbf{b})'(\mathbf{Y} - X\mathbf{b}) = \|\mathbf{Y} - X\mathbf{b}\|^2$. Ist X schlecht konditioniert, kann man nach Tychonoff vom Ansatz

$$\|X\mathbf{b} - \mathbf{Y}\|^2 + \|\Gamma\mathbf{b}\|^2 \stackrel{!}{=} \min \tag{2.47}$$

ausgehen. Hierin ist Γ eine geeignet gewählte Matrix, die *Tychonoff-Matrix*. $\|\Gamma\mathbf{b}\|^2$ heißt *Regularisierungsterm*, und der Ansatz (2.47) *Tychonoff-Regularisierung*. In der Statistik wird häufig $\Gamma = kI$, I die Einheitsmatrix, und $k \in \mathbb{R}$ gewählt. Für $k = 0$ erhält man dann die übliche Kleinste-Quadrate-Lösung. Die Lösung $\hat{\mathbf{b}}$ ist jedenfalls im allgemeinen Fall durch

$$\hat{\mathbf{b}} = (X'X + \Gamma'\Gamma)^{-1}X'Y \tag{2.48}$$

gegeben.
Beweis: Der Beweis verläuft analog zu der im Anhang, Abschnitt 4.1 gegebenen Herleitung der Schätzung für $\hat{\mathbf{b}}$. □

2.5.2 Ridge-Regression (Hoerl & Kennard, 1970)

Hoerl & Kennard (1970) setzen $\Gamma = hI$, $0 < h \in \mathbb{R}$ und I die Einheitsmatrix und sprechen dann von *Ridge-Regression*; man erhält dann

$$\hat{\mathbf{b}}_h = (X'X + h^2I)^{-1}X'Y. \tag{2.49}$$

Effekt der Regularisierung: Shrinkage Nach (2.36), Seite 14, wird die Varianz der Schätzungen $\hat{\mathbf{b}}$ durch die Reziprokwerte der Eigenwerte von $X'X$ mit bestimmt, und $X'X$ wird in (2.49) durch $X'X + h^2I$ ersetzt. Also muß man zunächst herausfinden, welche Eigenwerte und Eigenvektoren $X'X + h^2I$ hat.

Es wird zunächst gezeigt, dass die folgenden Aussagen gelten:

1. $X'X + h^2I$ hat die gleichen Eigenvektoren wie $X'X$,
2. Sind λ_j die Eigenvektoren von $X'X$, so hat $X'X + h^2I$ die Eigenvektoren $\lambda_j + h^2$, $j = 1, \ldots, p$.

   Beweis: Es sei P_j der j-te Eigenvektor von $X'X$ und λ_j der zugehörige Eigenwert. Die Multiplikation von $X'X + h^2I$ von rechts mit P_j liefert dann

   $$(X'X + h^2I)P_j = X'XP_j + h^2P_j = \lambda_jP_j + h^2P_j = (\lambda_j + h^2)P_j,$$

   und dies bedeutet, dass P_j ein Eigenvektor von $X'X + h^2I$ ist mit zugehörigem Eigenwert $\lambda_j + h^2$. □

Ist M eine symmetrische Matrix mit Eigenvektoren P und Eigenwerten Λ, so gilt $M = P\Lambda P'$ und $M^{-1} = P\Lambda^{-1}P'$. Da $X'X + h^2 I$ symmetrisch ist, folgt dementsprechend (Aussage 2 oben):

$$(X'X + h^2 I)^{-1} = PD^{-1}P', \tag{2.50}$$

mit

$$D^{-1} = \begin{pmatrix} 1/(\lambda_1 + h^2) & 0 & \cdots & 0 \\ 0 & 1/(\lambda_2 + h^2) & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 1/(\lambda_p + h^2) \end{pmatrix}. \tag{2.51}$$

Um die Wirkung der Regularisierung zu verdeutlichen, geht man von der Singularwertzerlegung (SVD) von X aus. Demnach gilt $X = Q\Lambda^{1/2}P'$, wobei Q die Eigenvektoren von XX' und P die Eigenvektoren von $X'X$ sind, und $\Lambda^{1/2}$ sind die Wurzeln aus den Eigenwerten von XX' bzw. $X'X$, $\sqrt{\lambda_j}$. Aus (2.49), Seite 19, folgt dann

$$\hat{\mathbf{b}}_h = PD^{-1}P'P\Lambda^{1/2}Q'Y = PD^{-1}\Lambda^{1/2}Q'y,$$

so dass

$$\hat{\mathbf{b}}_h = \sum_{j=1}^{p} P_j \frac{\sqrt{\lambda_j}}{\lambda_j + h^2} Q_j'Y = \sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + h^2} \frac{P_j Q_j' Y}{\sqrt{\lambda_j}}. \tag{2.52}$$

Weiter ist $\hat{Y} = X\hat{\mathbf{b}} = Q\Lambda^{1/2}P'PD^{-1}\Lambda^{1/2}Q'Y = Q\Lambda D^{-1}Q'Y$, so dass

$$\hat{Y} = \sum_{j=1}^{p} Q_j \frac{\lambda_j}{\lambda_j + h^2} Q_j'Y = \sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + h^2} Q_j Q_j' Y. \tag{2.53}$$

Definition 2.1 *Die Größe*

$$f_j = \frac{\lambda_j}{\lambda_j + h^2} \tag{2.54}$$

heißt Filter.

Offenbar wird f_j klein, je größer h^2 ist.

Man leitet sich nun leicht den Effekt der Regularisierung auf die Schätzung $\hat{\mathbf{b}}$ und die Vorhersage $\hat{Y}$ her: Es sei $h^2 > 0$. Dann folgt

$$\frac{\sqrt{\lambda_j}}{\lambda_j + h^2} = \frac{1}{\sqrt{\lambda_j} + h^2/\sqrt{\lambda_j}} \to \frac{1}{h^2/\sqrt{\lambda_j}} = \frac{\sqrt{\lambda_j}}{h^2} \to 0, \text{ für } \lambda_j \to 0.$$

Analog dazu folgt natürlich $f_j \to 0$ für $\lambda_j \to 0$.

Dies heißt, dass die Summanden in (2.52) und damit die Komponenten von $\hat{\mathbf{b}}_h$ um so kleiner werden, je größer der Wert von h^2 ist. Man spricht von

Shrinkage (Schrumpfung) der Schätzung von $\mathbf{b}$; die Varianz der Schätzungen für $\mathbf{b}$, die für kleine Eigenwerte als Resultat der Korrelationen zwischen den Prädiktoren gewissermaßen aufgebläht wird, wird auf diese Weise reduziert.

Die Regularisierung der Schätzung von b bedeutet demnach eine Reduzierung der Varianz von $\hat{\mathbf{b}}_h$. Man kann nun die Differenz $\hat{\mathbf{b}} - \hat{\mathbf{b}}_h$ betrachten:

$$\begin{aligned}
\hat{\mathbf{b}} - \hat{\mathbf{b}}_h &= (X'X)^{-1}X'Y - (X'X + h^2 I)^{-1}X'Y \\
&= P\Lambda^{-1}P'P\Lambda^{1/2}Q'Y - PD^{-1}P'P\Lambda^{1/2}Q'Y \\
&= P\Lambda^{-1/2}Q'Y - PD^{-1}\Lambda^{1/2}Q'Y \\
&= P(\Lambda^{-1/2} - D^{-1}\Lambda^{1/2})Q'Y \qquad (2.55)
\end{aligned}$$

Für das j-te Element δ_{jj} von $\Lambda^{-1/2} - D^{-1}\Lambda^{1/2}$ erhält man

$$\delta_{jj} = \frac{1}{\sqrt{\lambda_j}} - \frac{\sqrt{\lambda_j}}{\lambda_j + h^2} = \frac{h^2}{\sqrt{\lambda_j}(\lambda_j + h^2)},$$

so dass

$$\hat{\mathbf{b}} - \hat{\mathbf{b}}_h = P\Delta Q'Y, \quad \Delta = \operatorname{diag}\left(\frac{h^2}{\sqrt{\lambda_1}(\lambda_1 + h^2)}, \ldots, \frac{h^2}{\sqrt{\lambda_p}(\lambda_p + h^2)}\right). \qquad (2.56)$$

Für $h^2 = 0$ folgt sofort $\hat{\mathbf{b}} - \hat{\mathbf{b}}_h = 0$, und für $h^2 > 0$ ergibt sich $\hat{\mathbf{b}} - \hat{\mathbf{b}}_h > 0$, und die Differenz wird um so größer ausfallen, je größer[4] h^2.

Die übliche Kleinste-Quadrate-Schätzung geht bei der multiplen Regression von der GLeichung

$$Q(\mathbf{b}) = \sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2 \qquad (2.57)$$

aus; die Komponenten $b_0, b_1, \ldots, b_p$ werden so bestimmt, dass $Q(\mathbf{b})$ ein Minimum wird. Bei der Regularisierung geht dieser Ausdruck in

$$\tilde{Q}(\mathbf{b}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2 + \frac{\lambda}{2}\mathbf{b}'\mathbf{b} \qquad (2.58)$$

über. Diese Gleichung impliziert (vergl. Gleichung (2.49)), dass die b_j kleiner werden, im Zweifel gegen Null streben, es sei denn, die Daten erzwingen einen Wert größer Null. Dieser Sachverhalt illustriert noch einmal den Ausdruck *Shrinkage*, der in diesem Zusammenhang gebraucht wird.

[4] $d\delta_{jj}/d(h^2) = \lambda_j\sqrt{\lambda_j}/(\sqrt{\lambda_j}(\lambda_j + h^2) > 0$, also wächst δ_{jj} mit dem Wert von h^2.

2.5.3 Das Lasso (Tibshirani, 1996)

Das Acronym Lasso steht für Least Absolute Shrinkage and Selection Operator. Der Lasso-Fall wurde von Tibshirani (1996) eingeführt und diskutiert.

Eine Verallgemeinerung von (2.58) ist

$$L = \frac{1}{2}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2 + \frac{\lambda}{2}\sum_{j=1}^{p}|b_j|^q \tag{2.59}$$

wobei L für 'Verlust' (Loss) steht, – je größer L, desto größer der Fehler. Für $q = 2$ erhält man wieder die Ridge-Regression in der Form (2.58).

Lasso: Ein Spezialfall – das Lasso – ist $q = 1$:

$$L = \frac{1}{2}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2 + \frac{\lambda}{2}\sum_{j=1}^{p}|b_j|. \tag{2.60}$$

Für hinreichend großen Wert von λ werden einige der Komponenten von $\mathbf{b}$ gegen Null gedrängt, so dass man ein sparsames Modell für die Regression erhält, bei dem die Gefahr des *overfitting* minimalisiert ist.

Für die Schätzungen von $\mathbf{b}$ über die Lasso-Regularisierung gibt es keine geschlossene Form mehr. Für vorgegebene λ-Werte müssen die Schätzungen auf numerischem Wege, d.h. iterativ gefunden werden. Die Mimimalisierung von L ist äquivalent der Minimalisierung von

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij})^2$$

unter der Nebenbedingung

$$\sum_{j=1}^{p}|b_j| \leq \lambda_0,$$

λ_0 eine geeignet gewählte Konstante. Wird λ_0 hinreichend klein gewählt, werden einige der b_j exakt gleich Null. Damit ermöglicht das Lasso-Kriterium eine Art von Prädiktorauswahl. Man muß den geeigneten Wert von λ_0 adaptiv – also durch *trial and error* – finden, bis der Fehler minimalisiert wird. Sind $\hat{\mathbf{b}}_j$ die normalen Kleinste-Quadrate-Schätzungen (die also ohne Regularisierung berechnet werden) und ist

$$t_0 = \sum_{j=1}^{p}|\hat{b}_j|,$$

so werden für $\lambda_0 = t_0/2$ die $\hat{b}_j$-Werte bis zu 50 % geschrumpft.

Ridge-Regression und das Lasso sind offenbar Formen der Auswahl von Prädiktoren, wie sie auch als *subset selection* bekannt sind. Die im Folgenden besprochene PCA-Regression gehört ebenfalls dazu.

2.5.4 PCA- bzw. SVD-Regression

Allgemein ist eine Regularisierung angebracht, wenn X Multikollinearitäten enthält.

Zwei Vektoren $\mathbf{x}, \mathbf{y}$ heißen *kollinear*, wenn sie parallel sind, d.h. wenn etwa $\mathbf{y} = \lambda\mathbf{x}$, $\lambda \in \mathbb{R}$ gilt. Für das Skalarprodukt gilt dann $|\mathbf{x}'\mathbf{y}|/\|\mathbf{x}\|\|\mathbf{y}\| = 1$, da nun $\cos\theta = 1$, denn der Winkel θ zwischen den Vektoren ist gleich Null. Repräsentieren diese Vektoren Messungen oder Schätzungen von Parametern, so wird man strenge Kollinearität kaum finden, da sie stets mit Mess- bzw. Schätzfehlern behaftet sind, so dass etwa $\mathbf{x} = \mathbf{x}_0 + \xi_x$, $\mathbf{y} = \mathbf{y}_0 + \xi_y$ gilt, wobei $\mathbf{x}_0, \mathbf{y}_0$ die "wahren" Vektoren sind. Die Fehler ξ_x, ξ_y implizieren dann $|\mathbf{x}'\mathbf{y}|/\|\mathbf{x}\|\|\mathbf{y}\| < 1$, d.h. $\mathbf{x}$ und $\mathbf{y}$ sind gewissermaßen numerisch linear unabhängig. Diese Betrachtungen übertragen sich auf den Fall, dass im fehlerfreien Fall einige Prädiktorvariablen als Linearkombinationen anderer Prädiktorvariablen dartellbar sind; man spricht dann von *Multikollinearitäten*. Sie bedeuten ebenfalls hohe Korrelationen zwischen den entsprechenden Prädiktoren. Natürlich stellt sich die Frage, wie hoch eine Korrelation sein muß, damit sie als "hoch" betrachtet werden kann. Diese Frage hat keine einfache Antwort, denn eben wegen der stets existierenden Fehler in den Variablen wird man stets Korrelationen zwischen den Variablen beobachten, – auch wenn die wahre Korrelation gleich Null ist, und Stichprobenfehler können auch in diesem Fall relativ hohe Korrelationen erzeugen, oder die Korrelationen können relativ klein ausfallen, obwohl eine Multikollinearität vorliegt. Welches Maß auch immer man für Multikollinearität in den Daten definiert, in bestimmten Wertebereichen wird es nicht eindeutig für oder gegen die Existenz von linearen Abhängigkeiten verweisen.

Vernachlässigt man für einen Moment Mess- und Schätzfehler, so gilt in jedem Fall die Aussage, dass die Existenz von linearen Abhängigkeiten bedeutet, dass die $(m \times p)$-Matrix X der Prädiktoren einen Rang kleiner als $\min(m, p)$ hat, und dies bedeutet, dass die Anzahl der Eigenwerte λ_j von $X'X$ (oder $Z'Z$) kleiner als p ist, d.h. für $r < p$ Eigenwerte gilt $\lambda_j > 0$ und für die restlichen $p - r$ gilt $\lambda_k = 0$, $k > r$. Die Mess- und Schätzfehler bedeuten dann zwar, dass alle $\lambda_j > 0$, aber zumindest die λ_k für $k > r$ werden klein sein, da sie eben nur die hoffentlich kleinen Fehler reflektieren. Gleichung (2.36), Seite 14 zeigt die Auswirkungen auf die Varianz der Schätzungen der Regressionskoeffizienten. Ein Maß für Kollinearität ist der *Konditionsindex*

$$KI_j = \sqrt{\frac{\lambda_j}{\min_k \lambda_k}}, \quad j = 1, \ldots, p \tag{2.61}$$

Ein Wert $KI_j > 30$ gilt als multikollinearitätsverdächtig.

Die Singularwertzerlegung (SVD[5]): Offenbar ist es optimal, linear unabhängige Prädiktoren zu verwenden, die also nur wenig, also nur fehlerbe-

[5]Singular Value Decomposition

dingt, miteinander korrelieren. Es kann ein mühsamer Prozess sein, derartige Prädiktoren zu finden. Es kann eleganter sein, aus gegebenen Prädiktoren linear unabhängige, insbesondere orthogonale Prädiktoren zu berechnen. Für die Matrix X stets die Singularwertzerlegung (SVD – Singular Value Decomposition)

$$X = Q\Sigma P', \quad \Sigma = \Lambda^{1/2}, \quad \mathrm{rg}(X) = r \leq \min(m, p) \tag{2.62}$$

wobei Q eine (m, r)-dimensionale Matrix mit orthonormalen Spaltenvektoren und P ein (r, r)-dimensionale Matrix mit ebenfalls orthonormalen Spaltenvektoren ist, und $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ ist eine (r, r)-Diagonalmatrix mit $\sigma_i = \lambda_i^{1/2}$, λ_i die von Null verschiedenen Eigenwerte von XX' bzw. $X'X$ (diese beiden Matrizen haben identische von Null verschiedene Eigenwerte). (2.62) impliziert $XX' = Q\Lambda Q'$, d.h. Q enthält die Eigenvektoren von XX', die zu von Null verschiedenen Eigenwerten korrespondieren, und $X'X = P\Lambda P'$, d.h. P enthält die Eigenvektoren von $X'X$, die zu von Null verschiedenen Eigenwerten korrespondieren [6].

PCA-Regression: Ersetzt man in $\mathbf{y} = X\mathbf{b} + \mathbf{e}$ die Matrix X durch $Q\Sigma P'$, so erhält man

$$\mathbf{y} = X\mathbf{b} + \mathbf{e} = Q\Sigma P'\mathbf{b} + \mathbf{e} \tag{2.63}$$

Setzt man $\mathbf{u} = \Sigma P'\mathbf{b}$, so hat man damit einen neuen Parametervektor für das Modell

$$\mathbf{y} = Q\mathbf{u} + \mathbf{e} \tag{2.64}$$

definiert. Hier ist die Matrix Q die Matrix der Prädiktoren; die Spalten $\mathbf{q}_1, \ldots, \mathbf{q}_p$ sind orthogonal und damit linear unabhängig. Die rechte Seite von (2.64) ist als *PCA-Regression* bekannt.

Wendet man hierauf die Methode der Kleinsten Quadrate zur Schätzung von $\mathbf{u}$ an, so ergibt sich die Schätzung

$$\hat{\mathbf{u}} = (Q'Q)^{-1}Q'\mathbf{y} = Q'\mathbf{y} \tag{2.65}$$

da ja $Q'Q = I$ und damit $(Q'Q)^{-1} = I$. Wegen (2.64) hat man

$$\hat{\mathbf{u}} = Q'(Q\mathbf{u} + \mathbf{e}) = \mathbf{u} + Q'\mathbf{e}. \tag{2.66}$$

$Q'\mathbf{e}$ ist der neue Fehlervektor. Man hat dann $\mathbf{u} = \hat{\mathbf{u}} - Q'\mathbf{e}$ und mithin

$$\mathbb{E}(Q'\mathbf{e}) = Q'\mathbb{E}(\mathbf{e}) = \mathbb{E}(\hat{\mathbf{u}}) - \mathbb{E}(\mathbf{u}) = \mathbb{E}(\hat{\mathbf{u}}) - \mathbf{u} = 0 \tag{2.67}$$

wegen $\mathbb{E}(\mathbf{e}) = 0$; $\mathbf{u}$ ist daher eine unverzerrte (biasfreie) Schätzung. Für die Varianz von $\hat{\mathbf{u}}$ um $\mathbf{u}$ erhält man

$$\mathbb{E}(\hat{\mathbf{u}} - \mathbf{u})^2 = \mathbb{E}(\mathbf{u} + Q'\mathbf{e} - \mathbf{u})^2 = \mathbb{E}(\mathbf{e}'QQ'\mathbf{e}) = \mathbb{E}(\mathbf{e}'\mathbf{e}) = \sigma^2 \tag{2.68}$$

[6] Eine Herleitung und weitere Diskussion der Singularwertzerlegung findet man u.a. in Mortensen, Kurze Einführung in die Vektor- und Matrixrechnung für die Multivariate Statistik (2016)

d.h. die Varianz der Schätzungen $\hat{\mathbf{u}}$ ist gleich der Fehlervarianz.

Es sei $m > p$ – dies ist der Normalfall, so dass $\text{rg}(X) \leq p$. Betrachtet man nur $p - r$ Eigenwerte als "substantiell", so kann an X durch $\hat{X} = Q_r \Lambda_r^{1/2} P_r'$ approximiert werden. Nach (2.52) (p. 20) gilt

$$\hat{\mathbf{b}}_h = \sum_{j=1}^{m} f_j \frac{P_j Q_j' Y}{\lambda_j}. \tag{2.69}$$

Man kann dann eine Schätzung $\hat{\mathbf{b}}_r$ definieren, indem man den Filter in (2.54) eingeführten Filter f_j umdefiniert gemäß

$$f_j = \begin{cases} 1, & \text{für } j \leq r \\ \\ 0, & \text{für } j > r, \end{cases} \tag{2.70}$$

und Die Schätzung $\hat{\mathbf{b}}_r$ ist als *truncated SVD-estimate* (TSVD)-Schätzung bekannt. Die TSVD-Schätzung ist auf den ersten Blick attraktiv, weil kein besonderer Parameter h eingeführt werden muß; Betrachtungen, die auf der Analogie der KQ-Schätzung mit der Signalfilterung beruhen, zeigen aber, dass die regularisierten Schätzungen $\hat{\mathbf{b}}_h$ robuster als die TSVD-Schätzungen $\hat{\mathbf{b}}_r$ sind.

Alternativer Ansatz (Knüsel (2008)): Man geht wieder von (2.63) aus, multipliziert dieses mal aber $\mathbf{y} = Q\Sigma P'\mathbf{b} + \mathbf{e}$ von links mit Q':

$$Q'\mathbf{y} = Q'Q\Sigma P'\mathbf{b} + Q'\mathbf{e}. \tag{2.71}$$

Mit $\tilde{\mathbf{y}} = Q'\mathbf{y}$, $\tilde{\mathbf{b}} = P'\mathbf{b}$ und $\tilde{\mathbf{e}} = Q'\mathbf{e}$ erhält man, wegen $Q'Q = I_p$, die $p \times p$-Einheitsmatrix,

$$\tilde{\mathbf{y}} = \Sigma\tilde{\mathbf{b}} + \tilde{\mathbf{e}}. \tag{2.72}$$

Diese Gleichung heißt auch die *kanonische Form* des linearen Modells $\mathbf{y} = X\mathbf{b} + \mathbf{e}$. Für die i-te Komponente von $\tilde{\mathbf{y}}$ hat man nun

$$\tilde{y}_i = \begin{cases} \sigma_i \tilde{b}_i + \tilde{e}_i, & i = 1, \ldots, r,\ r \leq p \\ \tilde{e}_i, & i = r+1,\ r < p \end{cases} \tag{2.73}$$

Man findet sofort

$$\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \tilde{e}'QQ'\tilde{e} = \mathbf{e}'\mathbf{e}. \tag{2.74}$$

Der Vektor $\tilde{\mathbf{b}}$ soll nun nach der Methode der Kleinsten Quadrate (KQ-Methode) geschätzt werden. Demnach soll $\tilde{\mathbf{b}}$ so bestimmt werden, dass $\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \sum_i \tilde{e}_i^2 = \sum_i e_i^2$ minimal wird. Diese Bedingung ist erfüllt, wenn die $\tilde{b}_i$ so bestimmt werden, dass die $\tilde{e}_i = 0$, $i = 1, \ldots, r$. Die $\tilde{b}_{r+1}, \ldots \tilde{b}_p$ imFalle $r < p$ können beliebig gewählt werden, da sie sich nicht auf die Quadratsumme auswirken. Damit hat man

$$\tilde{b}_i = \begin{cases} \tilde{y}_i/\sigma_i, & i = 1, \ldots, r \\ \text{beliebig für}, & i = r+1, \ldots, p \end{cases} \tag{2.75}$$

Für die minimale Quadratsumme hat man dann

$$Q_{\min} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \tilde{e}_i^2 = \sum_{i=r+1}^{n} \tilde{e}_i^2 = \sum_{i=r+1}^{n} \tilde{y}_i^2. \tag{2.76}$$

(Das Q in $Q_{\min}$ steht hier für Quadratsumme, nicht für die Matrix Q). Aus $\tilde{\mathbf{b}} = P'\mathbf{b}$ folgt nun

$$\mathbf{b} = P\tilde{\mathbf{b}}, \tag{2.77}$$

und die die Komponenten $\tilde{b}_i$, $i = 1, \ldots, r$ sind eindeutig festgelegt, d.h. sie sind *identifizierbar*. Die $\tilde{b}_i$, $i = r+1, \ldots, p$ sind beliebig wählbar, d.h. $Q_{\min}$ ist eindeutig bestimmt, der Vektor $\tilde{\mathbf{b}}$ nicht. Damit ist die Gesamtmenge der Lösungsvektoren $\mathbf{b}$ ein $(p-r)$-dimensionaler Raum.

Multikollinearität: Dieser Fall ist gegeben, wenn $r < \min(m, p)$. Aus $X = Q\Sigma p'$ folgt dann $XP = Q\Sigma$ und umgekehrt, d.h. es gilt

$$X = Q\Sigma P' \Longleftrightarrow XP = Q\Sigma \tag{2.78}$$

Setzt man Q als $(m \times m)$-Matrix an, d.h. schreibt man auch die nicht zu Eigenwerten ungleich Null korrespondierenden Eigenvektoren von XX' in Q ein, und schreibt man analog P als $(p \times p)$-Matrix an, und, korrespondierend dazu Σ als $(m \times m)$-Matrix von Nullen, in der nur die erst r Diagonalzellen ungleich Null sind, so erhält man

$$X\mathbf{p}_{r+1} = \cdots X\mathbf{p}_p = \vec{0}. \tag{2.79}$$

$\mathbf{z}_j = X\mathbf{p}_j = \vec{0}$ impliziert, dass die Spaltenvektoren von X linear abhängig sind, denn $\mathbf{z}_j$ ist ja eine Linearkombination dieser Spaltenvektoren, mit dem Koeffizientenvektor $\mathbf{p}_j$. Daraus folgt, dass (mindestens) einer der Spaltenvektoren von X (d.h. eine Variable) aus X eliminiert werden kann. So sei $\mathbf{x}_2 = \mathbf{x}_1$. Dann ist

$$b_1\mathbf{x}_1 + b_2\mathbf{x}_2 = b_1\mathbf{x}_1 + b_2 a\mathbf{x}_1 = (b_1 + ab_2)\mathbf{x}_1.$$

Dann ist $\tilde{b}_1 = b_1 + ab_2$ eindeutig festgelegt, $\tilde{b}_2$ ist aber frei wählbar (nicht identifizierbar).

Da beliebige Werte für die $\tilde{b}_i$ für $i = r+1, \ldots, p$ gewählt werden können, kann man insbesondere $\tilde{b}_i = 0$ für $r+1 \leq i \leq p$ setzen. Die Frage ist dann, wie sich diese Wahl auf die Bestimmung der b_i auswirkt. Man hat nun

$$\sum_{i=1}^{r} \tilde{b}_i^2 = \tilde{\mathbf{b}}'\tilde{\mathbf{b}} \stackrel{!}{=} \min \tag{2.80}$$

Es ist

$$\tilde{\mathbf{b}}'\tilde{\mathbf{b}} = \mathbf{b}'PP'\mathbf{b} = \|\mathbf{b}\|^2,$$

d.h. die KQ-Schätzungen bewirken nun, dass $\mathbf{b}$ von minimaler Länge ist. Man spricht von *Schätzungen mit minimaler Länge*.

2.5.5 Variablen- bzw. Prädiktorselektion

Will man eine durch die Messungen Y repräsentierte Variablen durch andere Variable - also durch "Prädiktoren" - vorhersagen oder "erklären", so muß man sich entscheiden, welche Variablen als Prädiktor besonders geeignet sind. Die Vorhersage oder Erklärung soll so gut wie möglich anhand einer kleinstmöglichen Anzahl von Prädiktoren geschehen. Wird man nicht durch irgendeine Theorie auf spezielle Prädiktoren geführt, so muß man sich oft zwischen verschiedenen möglichen Prädiktorvariablen entscheiden.

Eine erste mögliche Strategie ist, einfach alle zur Verfügung stehenden Prädiktoren einzubinden. Das führt oft zu guten Vorhersagen, hat aber den Nachteil, dass viele Variablen erhoben werden müssen und zwischen den Prädiktoren im allgemeinen Abhängigkeiten bestehen werden. Dadurch wird die Vorhersage redundant, und die Regressionskoeffizienten sind nicht leicht interpretierbar. Man kann dann versuchen, *rückwärts* zu arbeiten und sukzessive Prädiktoren wegzulassen,wobei man jedesmal nachsieht, ob die Vorhersage durch Elimination eines Prädiktors wesentlich schlechter wird oder nicht. Diese Methode ist als *Backward Elimination Procedure* bekannt. Man kann auch umgekehrt vorgehen und mit einem Prädiktor beginnen. Dazu wählt man zunächst diejenige Variable aus, die die beste Vorhersage von Y gestattet. Man nimmt dann so lange weitere Prädiktoren hinzu, bis eine zufriedenstellende Vorhersage von Y erreicht wurde. Dies ist die *Forward Selection Procedure*. Eine Variante dieser Strategie ist die *Stepwise Regression Procedure*. Hier wird nach jeder Hinzunahme einer Variable das Modell der Multiplen Regression an die Daten angepaßt und alle bisher verwendeten Prädiktoren neu bewertet; eine Prädiktorvariable, die zunächst eine gute Vorhersage von Y zu gestatten schien, kann in Kombination mit anderen Prädiktoren eine relativ untergeordnete Rolle spielen und sogar überflüssig werden. Details dieser Verfahren findet man, zusammen mit numerischen Beispielen, z.B. in Draper und Smith (1966). Hier, wie etwa auch in Seber (1977), wird auch die *Residuenanalyse* , d.h. der Analyse der "Fehler" (= Residuen), die bei der Vorhersage von Y durch einen gegebenen Satz von Prädiktoren gemacht werden, eingehend diskutiert.

2.5.6 Bayes-Ansätze und Regularisierung

Die Schätzung von Parametern, etwa bei einem Regressionsproblem, mit der Methode der Kleinsten Quadrate setzt nicht die Annahme einer speziellen Verteilung voraus. Die Maximum-Likelihood-Methode dagegen geht von einer speziellen Verteilung aus. Für den Fall, dass die Normalverteilung angenommen werden kann, ergibt sich der gleiche Lösungsansatz wie bei der Methode der Kleinsten Quadrate. Der Bayessche Ansatz setzt nicht nur eine Annahme über die Verteilung der Messwerte voraus, sondern erfordert die Annahme einer a priori-Verteilung. Die Gauß-Verteilung hat in diesem

Fall eine Reihe interessanter Eigenschaften: sie gehört zu den Konjugierten a-priori-Verteilungen, d.h. ist die a priori-Verteilung eine Gauß-Verteilung, so ist die a posteriori-Verteilung ebenfalls eine Gauß-Verteilung. Darüber hinaus entspricht sie der Annahme maximaler Entropie, d.h. maximalen Unwissens über die wahren Werte der Parameter.

Es sei $\mathbf{b}$ der Vektor der zu schätzenden Parameter für eine multiple Regression. Die konjugierte a priori-Gauß-Verteilung ist dann

$$\phi(\mathbf{b}) = \mathcal{N}(\mathbf{b}|\mathbf{m}_0, \mathbf{S}_0). \tag{2.81}$$

$\mathbf{m}_0$ der Vektor der Erwartungswerte für die Komponenten von $\mathbf{b}$, und $\mathbf{S}_0$ die dazu korrespondierende Varianz-Kovarianz-Matrix. Die a posteriori-Verteilung ist dann durch

$$p(\mathbf{b}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{b})\phi(\mathbf{b}) \tag{2.82}$$

mit

$$p(\mathbf{b}|\mathbf{x}) = \mathcal{N}(\mathbf{b}|\mathbf{m}_N, \mathbf{S}_N), \tag{2.83}$$

wobei

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{b}\Phi'\mathbf{y}) && (2.84)\\ \mathbf{S}_0^{-1} &= \mathbf{S}_0^{-1} + \mathbf{b}\Phi'\Phi && (2.85) \end{aligned}$$

Φ ist eine Matrix von Basisfunktionen, d.h. es wird vom allgemeinen Fall

$$y(\mathbf{x}, \mathbf{b}) = b_0 + \sum_{j=1}^{p-1} b_j \phi_j(\mathbf{x}) \tag{2.86}$$

ausgegangen, und $\Phi = [\phi_1(\mathbf{x}_i), \ldots, \phi_{p-1}]$, $i = 1, \ldots, N$, N die Anzahl der Fälle in der Stichprobe. Wählt man für die a Priori-Verteilung insbesondere $p(\mathbf{b}|\alpha) = \mathcal{N}(\mathbf{b}|0, \alpha I)$, I die Einheitsmatrix, so erhält man für die log-posteriori-Verteilung

$$\log p(\mathbf{b}|y) = -\frac{b}{2}\sum_{i=1}^{N}(y_i - \mathbf{b}'\phi(\mathbf{x}_i))^2 - \frac{\alpha}{2}\mathbf{b}'\mathbf{b} + \text{const} \tag{2.87}$$

Wählt man einen "normalen" KQ-Ansatz und addiert einen Regularisierungsterm $-\lambda\mathbf{b}'\mathbf{b}$, so ist dieser Ansatz demnach äquivalent einem Bayesschen Ansatz mit der Gauß-Verteilung als a priori-Verteilung.

Eine Methode, die Probleme der Fast-Multikollinearität zu überwinden, ist die *Ridge-Regression*. Dabei werden iterativ die Diagonalelemente von $X'X$ solange erhöht, bis es zu einer Stabilisierung von $(X'X)^{-1}$ kommt; diese Erhöhung der Diagonalelemente impliziert eine Vergrößerung der Eigenwerte λ_j. Alternativ kann man bei der Ridge-Regression so vorgehen, dass man *penalisierte Schätzungen* für $\vec{\beta}$ berechnet. Dazu wird nicht nur, wie üblich,

$\| \mathbf{y} - X\vec{\beta} \|^2$ minimalisiert, sondern der "penalisierende" Term $-\lambda \| \vec{\beta} \|^2$ addiert, d.h. es wird

$$Q(\beta) = \frac{1}{2\sigma^2}(\mathbf{y} - X\vec{\beta})'(\mathbf{y} - X\vec{\beta}) + \lambda\vec{\beta}'\vec{\beta} \tag{2.88}$$

minimalisiert. Das Verfahren der Ridge-Regression ist allerdings kritisiert worden, so dass die anfängliche Euphorie über das Verfahren verflogen ist. Ein einfacher und eleganter Ausweg ist die im folgenden Abschnitt besprochene PCA-Regression, bei der zu unabhängigen Prädiktoren übergegangen wird.

2.5.7 Diskussion der verschiedenen Shrinkage-Methoden

Ridge-Regression und PCA-Regression führen beide auf "geschrumpfte" (shrinked) Schätzungen der Regressionsgewichte. Während bei der Ridge-Regression das Ausmaß der Schrumpfung von der Größe der Eigenwerte abhängt und deshalb eine schrittweise zunehmende Schrumpfung vom ersten Gewicht an bedeutet, bleibt die Schrumpfung bei der PCA-Regression gleich Null für die ersten r Eigenwerte, um vom r-ten an total zu werden: die korrespondierenden Gewichte werden gleich Nul gesetzt. Ridge-Regression schrumpft alle Richtungen (d.h. Richtungen der Eigenvektoren von $X'X$), wobei aber Richtungen mit geringer Varianz *mehr*. Die PCA-Regression läßt alle Richtungen mit großer Varianz ungeschrumpft, und verwirft alle übrigen. Frank & Friedman (1993) haben die verschiedenen Shrinkage-Verfahren untersucht und kommen zu der Folgerung, dass die Ridge-Regression den übrigen Verfahren vorzuziehen sind.

2.6 Der multiple Korrelationskoeffizient und Signifikanztests

Es sei r die Korrelation zwischen den zwei Variablen X und Y. Für einen "gewöhnlichen" Korrelationskoeffizienten r_{xy} gilt bekanntlich die Beziehung $r^2 = 1 - s_e^2/s_y^2$; $r^2 = D$, D der Determinationskoeffizient. Man kann diese Beziehung auch als *Definition* von r interpretieren. Diese Auffassung liegt der folgenden Definition zugrunde:

Definition 2.2 *Es werde das Modell (**??**) betrachtet und es seien $\hat{\beta}_j = \hat{\boldsymbol{b}}_j s_j/s_y$ die Kleinste-Quadrate-Schätzungen für die β_j bzw. für die b_j, s_e^2 sei die zu diesen Schätzungen korrespondierende Fehlervarianz (s_e der Standardschätzfehler) und s_y^2 sei die Varianz der gemessenen Y-Werte. Dann heißt die durch*

$$R^2_{0.12\ldots k} = 1 - \frac{s_e^2}{s_y^2} \tag{2.89}$$

bestimmte Größe $R_{0.12..k} = \sqrt{1 - s_e^2/s_y^2}$ der multiple Korrelationskoeffizient*.*

Man zeigt nun leicht die Gültigkeit des folgenden Satzes:

Satz 2.8 *Es sei $R_{0.12..k}$ der in (2.89) eingeführte multiple Korrelationskoeffizient, und es gelte*

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{1i} + \cdots + \hat{b}_k x_{ki},$$

Es sei $s^2(\hat{y})$ die Varianz der vorhergesagten Werte $\hat{Y}_i$, und s_y^2 sei die Varianz der gemessenen Y-Werte. Dann gilt

$$\begin{aligned} R^2_{0.12\ldots k} &= \frac{s^2(\hat{y})}{s_y^2} && (2.90)\\ &= r^2(Y, \hat{Y}) && (2.91)\\ &= \hat{\beta}_1 r_{y1} + \cdots + \hat{\beta}_k r_{yk} = \vec{\hat{\beta}}' \vec{R}_{xy}. && (2.92) \end{aligned}$$

Beweis: Den Beweis des Satzes findet man in Abschnitt 4.4, p. 76. □

R^2 ist eine Schätzung des Anteils an der Gesamtvarianz der Kriteriumsvariablen, der durch die Prädiktoren erklärt werden kann. Damit ist R^2 ein Maß für die Effektivität der multiplen Regression. Die Signifikanz von R^2 kann mit dem F-Test

$$F^{k-1}_{N-k} = \frac{R^2(N-k)}{(1-R^2)(k-1)} \tag{2.93}$$

getestet werden. Der Standardfehler für die Schätzung $\hat{z}_{0i}$ ist durch

$$s_{\hat{z}} = \sqrt{1-R^2} \tag{2.94}$$

gegeben.

Es läßt sich zeigen, dass die Schätzungen $\hat{\beta}_i$ normalverteilt sind mit dem Erwartungswert β_i und der Varianz $a_{ii}\sigma^2$, wobei a_{ii} das i-te Element in der Hauptdiagonalen von $(Z'Z)^{-1}$ ist, und σ^2 ist die Fehlervarianz. Man kann dann beliebige Hypothesen über β_i mit dem t-Test überprüfen:

$$t = \frac{\hat{\beta}_i - \beta_i}{\sqrt{a_{ii}}} \frac{1}{\sqrt{\sum_{i=1}^k e_i^2/(n-k)}}, \quad df = n-k \tag{2.95}$$

Um zu prüfen, ob β_i signifikant von Null abweicht, muß man hierin nur $\beta_i = 0$ setzen. Das $(1-\alpha)$-Konfidenzintervall für $\hat{\beta}_i$ ist dann

$$I_{1-\alpha} = \hat{\beta}_i \pm t_{\alpha/2}\sqrt{a_{ii}}\sqrt{\frac{\sum_j e_j^2}{n-k}}. \tag{2.96}$$

Die Frage ist, wie sich Multikollinearitäten auf die t-Werte auswirken.[7] Andererseits kann man den F-Test für die Nullhypothese $\vec{\beta} = \vec{0}$ betrachten. Der F-Test kann in der Form

$$\hat{F} = \frac{\vec{\hat{\beta}}'(X'X)\vec{\hat{\beta}}/k}{s^2}, \tag{2.97}$$

wobei s^2 eine Schätzung für σ^2 ist. Es sei $\vec{\pi} = P'\vec{\beta}$. Dann erhält man für den Erwartungswert von $\hat{F}$

$$\begin{aligned} \mathbb{E}(\hat{F}) &= \frac{\vec{\beta}'(X'X)^{-1}\vec{\beta}}{s^2} = \frac{\vec{\beta}'P\Lambda^{-1}P'\vec{\beta}/k}{s^2} \\ &= \frac{\vec{\pi}'\Lambda^{-1}\vec{\pi}}{s^2} = \frac{1}{ks^2}\sum_{j=1}^{r}\frac{\pi_i^2}{\lambda_j}. \end{aligned} \tag{2.98}$$

Hieraus folgt, dass kleine λ_j-Werte den $\hat{F}$-Wert *inflationieren*.

2.7 Illustration: zwei Prädiktoren

In diesem Fall lassen sich die Schätzungen für die $\hat{b}_j$ *bzw.* für die $\hat{\beta}_j$ leicht angeben. Nach (2.16) gilt nun

$$\begin{aligned} r_{y1} &= \hat{\beta}_1 + \hat{\beta}_2 r_{12} \\ r_{y2} &= \hat{\beta}_1 r_{21} + \hat{\beta}_2 \end{aligned} \tag{2.99}$$

wobei natürlich $r_{12} = r_{21}$ erfüllt ist. Die zweite Gleichung kann nach $\hat{\beta}_2$ aufgelöst werden:

$$\hat{\beta}_2 = r_{y2} - \hat{\beta}_1 r_{12}; \tag{2.100}$$

hieraus wird sofort ersichtlich, dass für den Fall $r_{12} \neq 0$ $\hat{\beta}_1$ und $\hat{\beta}_2$ nicht unabhängig voneinander sind. Setzt man nun (2.100) in die erste Gleichung von (2.99) ein, so erhält man

$$r_{y1} = (r_{y2} - \hat{\beta}_1 r_{12})r_{12} + \hat{\beta}_1 = r_{y2}r_{12} + \hat{\beta}_1(1 - r_{12}^2)$$

Löst man diesen Ausdruck nach $\hat{\beta}_1$ auf und setzt ihn dann in (2.100) ein, so folgen die Gleichungen

$$\hat{\beta}_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \tag{2.101}$$

$$\hat{\beta}_2 = \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}. \tag{2.102}$$

Nach (2.14) gilt $\hat{\beta}_j = \hat{\mathbf{b}}_j s_j/s_y$, also folgt $\hat{b}_j = \hat{\beta}_j s_y/s_j$, d.h. aber

$$\hat{b}_1 = \left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)\frac{s_y}{s_1} \tag{2.103}$$

[7] Die folgende Darstellung ist an Bierens (2007) orientiert.

$$\hat{b}_2 = \left(\frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2} \right) \frac{s_y}{s_2}, \tag{2.104}$$

und

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}_1 - \hat{b}_2 \bar{x}_2, \tag{2.105}$$

vergl. (2.9) und (2.11).

2.8 Beta-Gewichte und Partialkorrelationen

Die Ausdrücke für $\hat{\beta}_1$ und $\hat{\beta}_2$ bzw. $\hat{b}_1$ und $\hat{\mathbf{b}}_2$ legen nahe, dass die Beta- bzw. Regressionsgewichte mit den Partialkorrelationen zwischen den Variablen Y, X_1 und X_2 verwandt sind. Die Beziehung zwischen β-Gewichten und Partialkorrelationen soll im Folgenden verdeutlicht werden. Insbesondere gilt

$$\hat{\beta}_1 = r_{y1.2} \sqrt{\frac{1 - r_{y2}^2}{1 - r_{12}^2}} \tag{2.106}$$

Für $\hat{\beta}_2$ etc. gelten analoge Aussagen.

Beweis: Es seien wieder r_{y1} und r_{y2} die Korrelationen zwischen der abhängigen Variablen Y mit den Prädiktorvariablen X_1 und X_2, und r_{12} sei die Korrelation zwischen den Variablen X_1 und X_2. Dann ist die partielle Korrelation $r_{y1.2}$ (also die Korrelation zwischen Y und X_1, nachdem der Effekt der Variablen X_2 herauspartialisiert wurde)

$$r_{y1.2} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}}. \tag{2.107}$$

Andererseits ist

$$\hat{\beta}_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2},$$

d.h.

$$\hat{\beta}_1 (1 - r_{12}^2) = r_{y1} - r_{y2} r_{12}. \tag{2.108}$$

Multipliziert man $r_{y1.2}$ in (2.107) mit $\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}$, so folgt

$$r_{y1.2} \sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)} = r_{y1} - r_{y2} r_{12}. \tag{2.109}$$

Durch Gleichsetzen der beiden rechten Seiten von (2.108) und (2.109) gibt sich dann

$$\hat{\beta}_1 (1 - r_{12}^2) = r_{y1.2} \sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)},$$

d.h.

$$\hat{\beta}_1 = r_{y1.2} \frac{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}}{1 - r_{12}^2},$$

woraus sofort (2.106) folgt. □

Offenbar gilt also

$$\hat{\beta}_1 = r_{y1.2} \text{ genau dann, wenn } 1 - r_{y2}^2 = 1 - r_{12}^2. \tag{2.110}$$

Aus Symmetriegründen folgt dann ebenfalls

$$\hat{\beta}_2 = r_{y2.1} \text{ genau dann, wenn } 1 - r_{y1}^2 = 1 - r_{12}^2. \tag{2.111}$$

$\hat{\beta}_1 = r_{y1.2}$ also dann, wenn $|r_{y2}| = |r_{12}|$, wenn also die Korrelationen zwischen der Kriteriumsvariablen Y und und dem Prädiktor X_2 einerseits und zwischen den Prädiktoren X_1 und X_2 andererseits betragsmäßig identisch sind. Dieser Fall wird in strenger Form selten erfüllt sein. Wenn außerdem $\hat{\beta}_2 = r_{y2.1}$ gelten soll, folgt, dass die Regressionsgewichte genau dann gleich denPartialkorrelationen sind, wenn

$$|r_{y1}| = |r_{y2}| = |r_{12}|, \tag{2.112}$$

wenn also Y, X_1 und X_2 betragsmäßig gleich korrelieren. Diese Bedingung wird selten in strenger Form erfüllt sein; hat man Grund, anzunehmen, dass (2.112) zumindest in guter Näherung erfüllt ist, so kann das die Interpretation der Regressionsgewichte erleichtern, – sind sind dann approximativ als Partialkorrelationen deutbar.

Die Beziehung (2.106) erlaubt sofort, die Partialkorrelation $r_{y1.2}$ zu berechnen, was wiederum für die Interpretation der Zusammenhänge nützlich sein kann:

$$r_{y1.2} = \hat{\beta}_1 \sqrt{\frac{1 - r_{12}^2}{1 - r_{y2}^2}}, \tag{2.113}$$

d.h. wenn (2.112) nicht erfüllt ist, kann man sich zumindest die Partialkorrelationen ausrechnen, was für die Interpretation hilfreich sein kann.

Beispiel 2.3 Es werde die Beziehung zwischen der Merkfähigkeit (X_1), dem Alter (X_2) und dem intellektuellen Training (X_3) betrachtet; gegeben seien die Korrelationen $r_{12} = -.5$, $r_{13} = .8$ und $r_{23} = -.6$. Hier soll die Beziehung

$$X_1 = b_0 + b_2 X_2 + b_3 X_3 + e$$

betrachtet werden. Es ist also $Y = X_1$, $s_y = s_1$. Es sei $s_1 = 15$, $s_2 = 17$, $s_3 = 11$. Es werden zunächst die Beta-Gewichte berechnet. Es ist

$$\begin{aligned} \hat{\beta}_2 &= \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} = \frac{-.5 - .8 \times -.6}{1 - .6^2} = -.031, \\ \hat{b}_2 &= \hat{\beta}_2 \times \frac{s_1}{s_2} = -.031 \times \frac{15}{17} = -.027. \end{aligned}$$

und nach (2.113) gilt

$$r_{12.3} = \hat{\beta}_2 \sqrt{\frac{1 - r_{23}^2}{1 - r_{13}^2}} = -.031 \times \sqrt{\frac{1 - .6^2}{1 - .8^2}} = -.041.$$

Weiter ist

$$\hat{\beta}_3 = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} = \frac{.8 - (-.5)(-.6)}{(1 - .6^2)} = .781,\ \hat{b}_3 = \hat{\beta}_3 \frac{s_1}{s_3} = 1.065,$$

$$r_{13.2} = \hat{\beta}_3 \sqrt{\frac{1 - r_{23}^2}{1 - r_{12}^2}} = .722,$$

d.h. die Korrelation zwischen den Merkmalen Merkfähigkeit und intellektuelles Training nach Auspartialisierung des Alters beträgt $\approx .72$, so dass nur $\approx 50\%$ der Variation der Merkfähigkeit auf Unterschiede im intellektuellen Training zurückgeführt werden können. □

2.9 Suppressorvariable

Die multiple Korrelation $R_{0.12\ldots k}$ ist durch (2.92), d.h. durch

$$R_{0.12\ldots k} = \sqrt{\beta_1 r_{y1} + \cdots + \beta_k r_{yk}}$$

ausgedrückt worden; dabei sind die r_{yj}, $1 \leq j \leq k$ die Korrelationen zwischen den Prädiktoren x_j und dem Kriterium y. Nun kann die multiple Korrelation durch Hinzunahme eines Prädiktors *erhöht* werden, auch wenn dieser Prädiktor gar nicht mit dem Kriterium korreliert. Wir beschränken uns auf den Fall der Vorhersage einer Variablen y durch zwei Prädiktoren x_1 und x_2, da in diesem Fall die Verhältnisse noch übersichtlich sind. Dann ist jedenfalls

$$R_{0.12} = \sqrt{\beta_1 r_{y1} + \beta_2 r_{y2}},$$

und es gelte insbesondere $r_{y2} = 0$, so dass $R_{0.12} = \sqrt{\beta_1 r_{y1}}$, und nach (2.101) und (2.102) folgt

$$\beta_1 = \frac{r_{y1}}{1 - r_{12}^2}, \quad \beta_2 = \frac{r_{y1}r_{12}}{1 - r_{12}^2}.$$

Es soll nun $R_{0.12} = \sqrt{\beta_1 r_{y1}} > r_{y1}$ gelten, d.h. die Vorhersage von y durch x_1 *und* x_2 soll besser sein als die Vorhersage von y aufgrund des Prädiktors x_1 allein, *obwohl* y und x_2 nicht miteinander korrelieren. Dann folgt sofort

$$R_{0.12}^2 = \beta_1 r_{y1} > r_{y1}^2 \Rightarrow \beta_1 > r_{y1},$$

d.h. aber

$$\frac{r_{y1}}{1 - r_{12}^2} > r_{y1},$$

und dies ist der Fall, wenn $r_{12} \neq 0$ ist. Für $r_{12} = 0$ ist $\beta_1 = r_{y1}$ und $r_{0.12} = r_{y1}$; dies ist natürlich klar, denn wenn x_2 weder mit dem Kriterium y noch mit dem Prädiktor x_1 korreliert, kann die Vorhersage von y durch Hinzunahme von x_2 als Prädiktor nicht verbessert werden. Also ist $r_{12} \neq 0$ eine *notwendige* Bedingung für die Verbesserung der Vorhersage durch Hinzunahme von x_2, *wenn* $r_{y2} = 0$ gilt.

Es sei nun $r_{y1} > 0$. Dann ist $\beta_1 = r_{y1}/(1 - r_{12}^2) > 0$; das Vorzeichen von β_2 hängt aber, wegen $\beta_2 = r_{y1}r_{12}/(1-r_{12}^2)$, noch vom Vorzeichen von r_{12} ab. Für $r_{12} > 0$ ist ß$_2 > 0$, für $r_{12} < 0$ folgt ß$_2 < 0$. Ist dagegen $r_{y1} < 0$, so folgt ß$_1 < 0$. Ist $r_{12} > 0$, so folgt jetzt $\beta_2 < 0$, und für $r_{12} < 0$ folgt ß$_2 > 0$. Für $r_{y1} > 0$ ist das Vorzeichen von β_2 also immer dem von r_{12} entgegengesetzt.

Betrachtet man nun die partielle Korrelation $r_{y1\cdot 2}$, so findet man

$$r_{y1.2} = \frac{r_{y1}}{1 - r_{12}^2} = \beta_1.$$

Wegen $1 - r_{12}^2 < 1$ ist also $r_{y1.2} = \beta_1 > r_{y1}$.

> **Zusammenfassung:** Partialisiert man also aus x_1 die nicht mit y korrelierende Variablen x_2 heraus, so erhöht sich die Korrelation von x_1 mit y. Dementsprechend kann man sagen, dass x_1 nicht nur die Größe A mißt, die auch mit y erfaßt wird, sondern eine davon unabhängig variierende Größe B, die insbesondere durch x_2 repräsentiert wird und auf diese Weise den Zusammenhang zwischen y und x_1 reduziert. Die in x_1 mit gemessene Größe B erzeugt also eine Art "Rauschen", das die Vorhersage von y durch x_1 stört. Indem man B unabhängig von x_1 durch Messung der Variablen x_2 erfaßt und aus x_1 herauspartialisiert, wird dieses Rauschen gewissermaßen unterdrückt; dieser Sachverhalt motiviert die Bezeichnung *Suppressorvariable* für x_2: Die für die Vorhersage von y-Werten irrelevante Variabilität der x_1-Werte wird durch Hinzunahme der x_2-Variablen unterdrückt.

Analoge Betrachtungen gelten für mehr als zwei Prädiktorvariable; Variable, die mit der eigentlich interessierenden (= Kriteriums-) Variable "nichts" zu tun haben, weil sie nur wenig oder gar nicht mit ihr korrelieren, können gleichwohl helfen, die Vorhersage der Kriteriumsvariable zu verbessern. Allerdings werden die Beziehungen zwischen den Variablen schnell sehr unübersichtlich, wenn die Anzahl der Prädiktorvariablen steigt. Für ein qualitatives Verständnis der Suppressionseffekte ist der hier betrachtete Fall hinreichend.

2.10 Fehler in den Variablen

Es sei X die Matrix der Messungen der Prädiktorvariablen, die allerdings nicht messfehlerfrei seien. Es gelte demnach

$$X = \tilde{X} + V, \tag{2.114}$$

wobei V die Matrix der Messfehler sei und $\tilde{X}$ die "wahren" Werte der Prädiktoren enthalte. Für die Kriteriumsvariable Y gelte also

$$\mathbf{y} = \tilde{X}\vec{b} + \mathbf{e}. \tag{2.115}$$

Aus (2.114) folgt $\tilde{X} = X - V$, mithin erhält man

$$\mathbf{y} = (X - V)\vec{b} + \mathbf{e} = X\vec{b} + (\mathbf{e} - V\vec{b}). \tag{2.116}$$

Für die Kleinste-Quadrate-Schätzung für $\vec{b}$ erhält man

$$\hat{\mathbf{b}} = \vec{b} + (X'X)^{-1}X'(\mathbf{e} - V\vec{b}). \tag{2.117}$$

Nun werden die Schätzungen $\hat{\mathbf{b}}$ stets von den wahren Werten $\vec{b}$ abweichen, und in bezug auf (2.117) ist die Frage also, ob $X'(\mathbf{e} - V\vec{b}) \to \vec{0}$ strebt oder nicht. Es ist

$$\frac{1}{n}X'(\mathbf{e} - V\vec{b}) = \frac{1}{n}X'\mathbf{e} - \frac{1}{n}X'V\vec{b}, \tag{2.118}$$

n der Stichprobenumfang. Nimmt man an, dass die Elemente von $\mathbf{e}$ unkorreliert sind so kann man

$$\frac{1}{n}X'\mathbf{e} \to \vec{0} \tag{2.119}$$

annehmen. Aber unter Berücksichtigung von (2.114) hat man $X'V\vec{b} = (V' + \tilde{X}')V\vec{b}$ und mithin

$$\frac{1}{n}X'V\vec{b} = \frac{1}{n}V'V\vec{b} + \frac{1}{n}\tilde{X}'V\vec{b}. \tag{2.120}$$

Selbst wenn $\tilde{X}'V\vec{b}/n \to 0$, bleibt noch der Term $V'V\vec{b}/n$, der die Varianzen und Kovarianzen der Messfehler V enthält, so dass $\vec{b}$ i.a. nicht biasfrei geschätzt wird.

2.11 Kreuzvalidierung

Ein zweites Problem bei der Interpretation der $\hat{\mathbf{b}}_j$ und $\hat{\beta}_j$ ergibt sich daraus, dass sie eben Kleinste-Quadrate-Schätzungen sind. Diese Schätzungen optimieren die Vorhersage von y *anhand der vorliegenden* Meßwerte x_{ij}, d.h. $\sum_{ij} e_{ij}^2$ wird für die gegebene Stichprobe von Meßwerten minimalisiert. Ist diese Stichprobe nicht repräsentativ, so sind auch die Schätzungen der $\hat{\mathbf{b}}_j$ und $\hat{\beta}_j$ nicht repräsentativ. Das klingt fast trivial, — aber bei der Diskussion von Meßdaten wird dieser Sachverhalt oft übersehen. Man denke etwa an die Diagnostik: Anhand eines Tests soll ein Merkmal optimal erfaßt werden. Dazu werden Untertests $T_1, \cdots, T_k$ entwickelt, und die Scores in diesen Tests sollen als Prädiktoren für das Merkmal dienen. Dazu werden anhand einer Eichstichprobe die Gewichte $\hat{\beta}_j$ bestimmt, so dass $\hat{y} = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$ eine optimale Vorhersage von y ist, wobei die x_j die Scores (Punktwerte) in den Untertests T_j kennzeichnen. Wendet man nun den Test auf eine beliebige Person an, so ist der für sie auf diese Weise bestimmte y-Wert nun dann eine gute Schätzung, wenn die Person der Eichstichprobe entspricht, denn nur dann sind die Gewichte $\hat{\beta}_1, \cdots, \hat{\beta}_k$ vernünftige Gewichte für sie. Gehört sie aber einer Population an, die der Eichstichprobe nicht gut entspricht, so sind die $\hat{\beta}_j$ suboptimal.

Das hier angesprochene Problem ist ziemlich grundlegender Natur. Eine notwendige Bedingung für die Interpretierbarkeit der $\hat{\beta}_j$ ist deswegen die *Kreuzvalidierung*. Dazu wird die Untersuchung an einer zweiten Stichprobe wiederholt, wobei sowohl die $y - wie$ die x_{ij}-Werte gemessen werden. Die y-Werte werden nun anhand der $\hat{\mathbf{b}}_j - bzw.$ $\hat{\beta}_j$-Werte, die anhand der Daten aus der *ersten* Stichprobe gewonnen wurden, vorausgesagt; sie mögen mit $\hat{y}_1$ bezeichnet werden. Es zeigt sich, dass die Korrelation $r(y, \hat{y}_1)$ im allgemeinen geringer ist, als wenn man die y-Werte aufgrund von Regressionsgewichten voraussagt, die anhand der gleichen Stichprobe gewonnen wurden. Insbesondere kann sich zeigen, dass bestimmte Prädiktoren, die in der einen Stichprobe ein Regressionsgewicht $\neq 0$ erhielten und damit für die Vorhersage von Belang waren, für eine valide Vorhersage kaum geeignet, also überflüssig sind. Dies gilt insbesondere für Prädiktoren, die sich als Kombination anderer Prädiktoren ergeben, etwa $z = x_i x_j$, oder $z = x_j^2$, etc. Dies sind nichtlineare Terme, — die Regression bleibt natürlich auch bei Verwendung solcher Terme linear, denn die Linearität bezieht sich auf die Gewichte $\hat{b}_j$ und $\hat{\beta}_j$, in diesen Gewichten sind die Gleichungen stets linear. Nichtlineare Glieder der Form $x_i x_j$ oder x_j^2 etc "überleben" die Kreuzvalidierung häufig nicht; dies ist ein weiterer Grund für die Vorherrschaft der auch in den Prädiktoren linearen Regression. Die Instabilität der Schätzungen der Gewichte für solche nichtlinearen Terme ergibt sich daraus, dass ihre Meßfehler an die Variablen x_i, x_j etc. gekoppelt sind und multiplikativ in die Regressionsgleichungen eingehen.

2.12 Anwendung: Das Brunswick-Modell

Brunswick (1956)[8] schlug ein allgemeines Modell der Urteilsbildung vor, das auf einer Anwendung der multiplen Regression beruht. Die betrachtete Situation hat die folgende Struktur:

1. Es sollen Urteile über die Ausprägungen Y einer bestimmten Variablen getroffen werden. Diese Variable ist das "Kriterium" oder auch die "distale Variable".

2. Die Urteilsbildung erfolgt auf der Basis "proximaler" Variablen x_i, $i = 1, ..., n$, oder "Cues".[9]

3. Y kann anhand der x_i durch multiple Regression vorhergesagt werden; es soll gelten

$$y_0 = \beta_{01} x_1 + \beta_{02} x_2 + \cdots + \beta_{0n} x_n + e_0 \tag{2.121}$$

[8] Brunswick, E.: Perception and the representative design of psychological experiments. University of California Press, Berkeley 1956

[9] cue = Stichwort, Hinweis, Symptom

wobei y_0, x_i und z_0 bereits *standardisierte* Werte sind; e_0 ist der Fehler. y_0 ist die objektive Vorhersage von Y, d.h. die β_{0i} sind aus objektiven Messungen gewonnen worden. Statt der linearen Regression kann im Prinzip auch ein nichtlinearer Zusammenhang bestehen.

4. Eine Person beurteilt Y ebenfalls aufgrund der x_i, gewichtet diese Prädiktoren aber unter Umständen nicht optimal. Ihre Vorhersagen von Y entsprechen der Gleichung

$$y_s = \beta_{s1}x_1 + \beta_{s2}x_2 + \cdots + \beta_{sn}x_n + e_s \tag{2.122}$$

Der Index s steht für *subjektiv*.

Ist der lineare Ansatz korrekt, so ist die Vorhersage von Y durch (2.121) optimal (im Sinne der Methode der Kleinsten Quadrate). Die Vorhersage durch y_s ist genau dann suboptimal, wenn *nicht* für alle i $\beta_{si} = \beta_{0i}$ gilt. Die Vorhersageleistung der Person kann durch einen Korrelationskoeffizienten ausgedrückt werden:

$$r_a = r(y_o, y_s) \tag{2.123}$$

Dabei steht der Index a für *achievement*, d.h. für die Leistung. Die Korrelationen r_{xy}, d.h. die Korrelationen zwischen den "Cues" und der Variablen Y, heißen auch *cue validities*, auf deutsch vielleicht *Symptomgültigkeiten*. Sie sind natürlich eng mit den β_{xi} verwandt. Die Korrelationen $r_{xy;s}$ sind die *cue dependencies*; sie reflektieren die Abhängigkeit des Urteils y_s von den x_i und stehen zu den β_{si} in Beziehung.

Bevor einige inhaltliche Betrachtungen über das Modell angestellt werden, soll eine zuerst von Tucker (1964) abgeleitete Formel für r_a angegeben werden:

Satz 2.9 *Es sei*

$$\begin{array}{rcl} \hat{y}_0 &=& \beta_{01}x_1 + \beta_{02}x_2 + \cdots + \beta_{0n}x_n \\ \hat{y}_s &=& \beta_{s1}x_1 + \beta_{s2}x_2 + \cdots + \beta_{sn}x_n \end{array} \tag{2.124}$$

so dass

$$\begin{array}{rcl} y_0 &=& \hat{y}_0 + z_0 \\ y_s &=& \hat{y}_s + z_s. \end{array}$$

Dann gilt

$$r_a = Kov(y_0, y_s) = \underbrace{Kov(\hat{y}_0, \hat{y}_s)}_{G} + Kov(e_0, e_s). \tag{2.125}$$

Beweis: Allgemein gilt: Sind a, b, c und d irgendwelche Variablen, dann folgt

$$Kov(a+b, c+d) = Kov(a,c) + Kov(a,d) + Kov(b,c) + Kov(b,d) \tag{2.126}$$

Mit $a = \hat{y}_0$, $b = z_0$, $c = \hat{y}_s$, $d = z_s$ folgt sofort (2.125), wenn man bedenkt, dass in diesem Falle ja $Kov(a, d) = Kov(b, c) = 0$ ist: Fehler und Vorhersagen sind bei Kleinste-Quadrate-Schätzungen ja unabhängig voneinander. □

Es sei $G = r(\hat{y}_0, \hat{y}_s)$. Aufgrund von (2.90) ist bekannt, dass $s^2(\hat{y_0}) = R_0^2 s_y^2$ und $s^2(\hat{y}_s) = R_s^2 s_y^2$, wobei aber wegen der vorausgesetzten Standardisierung $s_y = 1$. Dann ist

$$Kov(\hat{y}_0, \hat{y}_s) = GR_0R_s.$$

Weiter sei $r(e_0, e_s) = C$ die Korrelation zwischen den Fehlervariablen. Es folgt

$$r(e_0, e_s) = Kov(e_0, e_s)/(s_{e_0} s_{e_s}).$$

Aus (2.89) folgt aber

$$s^2(e_0) = \sqrt{1 - R_0^2}, \quad s^2(e_s) = \sqrt{1 - R_s^2}$$

Eingesetzt in (2.125) ergibt sich dann

Satz 2.10 *Es gilt*

$$r_a = GR_0R_s + C\sqrt{(1 - R_0^2)(1 - R_s^2)} \tag{2.127}$$

Hammond und Summers (1972) haben Satz 2.10 zum Ausgangspunkt einiger allgemeiner Betrachtungen über den Erwerb von Wissen einerseits und über die Anwendung ("Kontrolle") dieses Wissens andererseits gemacht. Während r_a die "Urteilsleistung" mißt, d.h. eben die Kovariation zwischen dem Merkmal Y und den Aussagen über dieses Merkmal, repräsentiert G das "Wissen" über die Aufgabe. In der Tat bedeutet ein hoher Wert von G, dass eine Person weiß, welche "Cues" sie zu berücksichtigen und wie sie sie zu gewichten hat. R_o repräsentiert die Ungewißheit auf der objektiven Seite, und R_s repräsentiert die Ungewißheit auf der Seite der urteilenden Person. R_0 definiert die obere Grenze der Urteilsgenauigkeit, wenn die Prädiktoren, Symptome etc. x_i zur Grundlage des Urteils gemacht werden. R_s definiert dagegen die "kognitive Kontrolle". Selbst wenn das Wissen bezüglich des Zusammenhanges zwischen Y und den x_i perfekt ist, kann r_a einen suboptimalen Wert annehmen, wenn nämlich die Varianz des Fehlers z_s groß ist. Umgekehrt kann die Person ihr Wissen optimal nutzen, das Kriterium Y aber nur unvollständig voraussagen, wenn die objektive Struktur nur eine unvollständige Voraussage zuläßt.

Goldberg (1970)[10] hat das Modell auf Klinische Urteile angewendet und fand, dass das Wissen, so wie es durch G ausgedrückt werden kann, oft

[10]Goldberg, L.R. (1970) Man versus model of man: a rationale, plus some evidence, for a method of improving on clinical inferences. (1970) Psychological Bulletin, 73, 422-432

sehr gut ist, die Urteile aber trotzdem suboptimal bleiben, da die Urteiler inkonsistent mit dem Wissen umgehen, also einen Mangel an "kognitiver Kontrolle" zeigen.

Das Modell ist ebenfalls zur Diskussion von interpersonellen Konflikten angesetzt worden (Brehmer, Azuma, Hammond, Kostron und Varonos (1970)). Es zeigte sich, dass die kognitiven Systeme als Resultat der sozialen Interaktion konvergierten, d.h. G wuchs, sich aber die kognitive Kontrolle *verringerte*, d.h. R_s wurde geringer.

Weiter ist R_0 der multiple Korrelationskoeffizient für die "objektive" Vorhersage von Y anhand der x_i, und R_s ist der multiple Korrelationskoeffizient für die entsprechende subjektive Vorhersage.

2.13 Spezialfall: Zeitreihen

In den bisherigen Beispielen sind die Variablen y, $x_1, ..., x_k$ anscheinend von der Zeit unabhängige Größen. Wir wissen aber, dass viele psychologisch interessante Variablen in der Zeit variieren. So ist z.B. bekannt, dass das Ausmaß an Depressivität mit dem Tagesverlauf schwanken kann (Tagesschwankungen), und dass möglicherweise noch weitere Periodizitäten feststellbar sind. Auf solche *Zeitreihen* wird in einem späteren Kapitel noch ausführlich eingegangen, hier soll nur gezeigt werden, wie der Ansatz der multiplen Regression dazu dienen kann, die Struktur der zeitlichen Variation von Meßwerten zu erfassen.

Es sei X_t etwa das Ausmaß an Depressivität zum Zeitpunkt t, X_{t-1} sei das Ausmaß an Depressivität zum Zeitpunkt $t-1$ etc. Die X_t, X_{t-1}, X_{t-2}, ... bilden dann eine *Zeitreihe*; Zeitreihen sind spezielle *stochastische* oder *Zufallsprozesse*. Für die Werte X_t können drei Modelle betrachtet werden:

Modell 1:

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_k X_{t-p} + e_t. \tag{2.128}$$

Formal entspricht dieser Ansatz der multiple Regression, denn X_t übernimt die Rolle der Kriteriumsvariablen Y, und X_{t-1}, X_{t-2}, $\cdots$, X_{t-p} fungieren als Prädiktoren; a_1, $\cdots$, a_p sind Regressionsgwichte. e_t ist eine zufällig zum Zeitpunkt t auf die Variable einwirkende Größe (ein "random shock"). Der *Prozeß* X_t ist also dadurch charakterisiert, dass seine Werte durch Regression bis auf einen Wert e_t durch $k \geq 1$ vorangegangene Werte bestimmt werden. Man spricht deshalb auch von einem *autoregressiven Prozeß k-ter Ordnung*, oder kurz von einem $AR(k)$-Prozeß.

Modell 2: Ein alternatives Modell ist durch

$$X_t = b_1 e_{t-1} + b_2 e_{t-2} + \cdots + b_r e_{t-q} + e_t \tag{2.129}$$

gegeben. Wieder ist X_t durch einen Regressionsansatz definiert, allerdings sind hier die Prädiktoren nicht die vergangenen Werte von X_t selbst, sondern

die vergangenen r zufälligen "Stöße" *random shocks*, die gewissermaßen von außen auf den Organismus einwirken. X_t wird durch ein gleitendes Mittel der vergangenen r Werte e_t, e_{t-1}, $\cdots$, e_{t-q} dargestellt, und man spricht auch von einem *Moving-Average-Prozeß der Ordnung* r, abgekürzt durch $MA(r)$-Prozeß.

Modell 3: Der Verlauf einer Variablen X kann gelegentlich auch durch eine Kombination von AR- und MA-Prozessen beschrieben werden:

$$X_t = a_1 X_{t-1} + \cdots + a_p X_{t-p} + b_1 e_{t-1} + \cdots + b_r e_{t-1} + e_t \tag{2.130}$$

Dies ist dann ein ARMA(p, q)-Prozeß. Die Koeffizienten a_i und b_j repräsentieren das "System", das die X_t generiert; aus ihnen lassen sich u.U. Schlüsse auf die Funktionsweise des Systems ziehen. □

Die Schätzung der Parameter a_i, $i = 1, ..., p$ bzw. b_j, $j = 1, ..., q$ geschieht im Prinzip in der gleichen Weise wie bei einer multiplen Regression, d.h. gemäß Satz 2.2, (2.16), wobei die Parameter β_j durch die a_i, bzw. b_j zu ersetzen sind, und die Korrelationen durch die *Autokorrelationen mit einer Verzögerung* k

$$r(t, t-k) = \frac{Kov(x_t, X_{t-k})}{s(x_t)s(x_{t-k})}. \tag{2.131}$$

Auf weitere Details soll hier nicht eingegangen werden, da zusätzliche Eigenschaften von Zeitreihen diskutiert werden müssen, die einen Exkurs in die Theorie der stochastischen Prozesse voraussetzen.

2.14 Abschließende Bemerkungen

Die multiple Korrelation ist hier keinesweges vollständig beschrieben worden. Detailliertere Informationen über die Regression im allgemeinen und die multiple Regression im besonderen findet man in speziell über dieses Gebiet geschriebenen Lehrbüchern, wie etwa Christensen (1987), Draper und Smith (1966) oder Seber (1977).

3 Generalisierte lineare Modelle

3.1 Binäre Kriteriumsvariable

Es gelte die Beziehung

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + e_i, \quad i = 1, \ldots, m \tag{3.1}$$

mit kontinuierlich variierenden Prädiktorwerten x_{ij} impliziert, dass die abhängige Variable y_i ebenfalls kontinuierlich variiert. Oft hat man aber eine abhängige Variable, die *nicht* kontinuierlich variiert, sondern eher eine Kategorie anzeigt: ein Bewerber für eine bestimmte berufliche Position wird seine Aufgaben entweder meistern oder nicht, eine Operation wird erfolgreich verlaufen

oder nicht, ein Straftäter wird rückfällig oder nicht, etc. Man möchte das in Frage stehende Ereignis anhand der Prädiktoren vorhersagen, möglichst anhand eines Ausdrucks, wie er auf der rechten Seite von (3.1) steht. Statt einer kontinuierlich variierenden Kriteriumsvaraiablen hat man aber nun eine *Indikatorvariable*:

$$Y_i = \begin{cases} 1, & y_i > y_0 \\ 0, & y, \leq y_0 \end{cases} , \tag{3.2}$$

wobei y_i einfach der Wert der rechten Seite von (3.1) und y_0 ein Schwellenwert ist, den man so bestimmen muß, dass die Entscheidungen möglichst fehlerfrei sind. Die Frage ist nun, wie man eine solche Festlegung für y_0 findet.

Diese Frage kann man mit einiger Ausführlichkeit diskutieren. Bei medizinischen Diagnosen kann es charakteristische Symptome oder Symptomausprägungen geben, von denen man weiß, dass sie einen bestimmten Krankheitszustand widerspiegeln, – d.h. man hat irgendwie Erfahrungen sammeln können derart, dass ein solcher Schluß von der Symptomausprägung (in Bezug auf (3.1) müßte man von der Ausprägung einer Kombination von Symptomen sprechen) auf den Zustand eines Patienten mit an Sicherheit grenzender Wahrscheinlichkeit möglich ist. In anderen Fällen sind solche "kritischen" Ausprägungen nicht gegeben, man denke an die Möglichkeit des Rückfalls eines Straftäters, oder an die Möglichkeit eines erneuten Auftretens einer Panikattacke eines Patienten, der sich seiner Angstzustände wegen einer Therapie unterzogen hat. Es liegt demnach nahe, die rechte Seite von (3.1) mit der Wahrscheinlichkeit einer korrekten Klassifizierung oder Diagnose in Verbingung zu bringen. Ein solcher Ansatz würde den Fall, dass der Schwellenwert y_0 anhand irgendwelcher Erfahrungen oder gar theoretischer Betrachtungen bestimmt wird, als Spezialfall einschließen. Generell hat man dann

$$P(Y_i = 1|\mathbf{x}) = \phi(\mathbf{x}), \tag{3.3}$$

wobei $\phi \in (0, 1)$ eine geeignet gewählte Funktion ist. Geht man davon aus, dass die inverse Funktion ϕ^{-1} für die in Frage kommenden $\mathbf{x}$ existiert, so ist

$$g(P(Y_i = 1|\mathbf{x})) = \phi^{-1}[P(Y_i = 1|\mathbf{x})] = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} \tag{3.4}$$

die *Link-Funktion* (d.h. $g = \phi^{-1}$ ist die Link-Funktion, deren Wert durch die rechte Seite von (3.8) gegeben ist)[11]. Beispiele für Link-Funktionen werden im Zusammenhang mit den entsprechenden Wahrscheinlichkeitsmodellen gegeben.

Für die Wahl der Funktion ϕ hat man verschiedene Möglichkeiten:

1. Man geht von der Annahme aus, dass das Ereignis dann eintritt, wenn eine zufällig variierende Größe ξ einen Schwellenwert überschreitet. ξ wird nicht direkt beobachtet und ist insofern eine latente Variable.

[11] Link - engl. 'Verbindung'

kann z.B. die Verkalkung der Herzkranzgefäße repäsentieren. Ist sie größer als ein kritischer Wert, kommt es zum Herzinfarkt ($Y_i = 1$, – das ist sicherlich ein vereinfachtes Modell, aber es dient auch nur zur Illustration). Dementsprechend nimmt man eine Verteilungsfunktion F für ξ an: $F_\xi(x) = P(\xi \leq x)$. Dann hat man

$$P(Y_i = 1|\mathbf{x}) = 1 - F_\xi(y_0|\mathbf{x}_i), \quad \mathbf{x}_i = (x_{i1}, \ldots, x_{ip})' \tag{3.5}$$

Die Abhängigkeit von $\mathbf{x}$ läßt sich z.B. modellieren, indem man den Erwartungswert $\mu = \mathbb{E}(\xi)$ von ξ als Funktion der rechten Seite von (3.1) auffasst, so dass man

$$\mu_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} \tag{3.6}$$

hat; hier taucht der Fehlerterm e_i nicht mehr auf. Der Grund dafür liegt darin, dass die stochastischen Aspekte in (3.1) eben nur über die zufälligen Größen e_i eingebracht werden, während sie im Ansatz (3.5) in der Verteilung P_ξ ausgedrückt werden. Man entscheidet jetzt für $Y_i = 1$, wenn $P(Y_i = 1|\mathbf{x}) > P_0$ ist, und P_0 hängt von der Größe des Risikos ab, das man einzugehen bereit ist. Mit P_0 wird implizit auch der Schwellenwert y_0 definiert. Es wird nun deutlich, dass man jetzt noch eine Beziehung zwischen dem Risiko und der der Wahrscheinlichkeit P_0 finden muß, sofern es um Entscheidungen geht. In bestimmten Anwendungen ist aber die Wahl einer solchen kritischen Größe nicht nötig, etwa dann, wenn man nur daran interessiert ist, zu erfahren, ob eine bestimmte Therapie die Wahrscheinlichkeit einer Heilung erhöht, oder die Wahrscheinlichkeit eines Rückfalls erniedrigt, etc.

2. Für den Fall, dass in Abhängigkeit vom Wert der Symptome oder allgemein Prädiktoren eine Entscheidung getroffen werden soll, kann man vom Satz von Bayes ausgehen. Demnach gilt ja

$$P(Y_i = 1|\mathbf{x}) = P(\mathbf{x}|Y_i = 1)\frac{P(Y_i = 1)}{P(\mathbf{x})}. \tag{3.7}$$

Hierin ist $P(Y_i = 1|\mathbf{x})$ die Posteriori-Wahrscheinlichkeit für $Y_i = 1$, gegeben $\mathbf{x}$, und $P(Y_i = 1)$ ist die Priori-Wahrscheinlichkeit für das Ereignis, das durch $Y_i = 1$ indiziert wird. $P(\mathbf{x}|Y_i = 1)$ ist die Likelihood der Daten oder Symptome $\mathbf{x}$, und $P(\mathbf{x})$ ist die Wahrscheinlichkeit, dass man die Symptome oder Prädiktorwerte $\mathbf{x}$ überhaupt beobachtet. Es zeigt sich, dass man unter bestimmten Bedingungen zu einer Lösung geführt wird, die dem Ansatz 3.5 äquivalent ist.

In jedem Fall wird die rechte Seite von (3.1) (ohne e_i), also

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip},$$

mit einer Wahrscheinlichkeitsfunktion verbunden, etwa $P_i = P(Y_i = 1|\mathbf{x})$ Die Umkehrfunktion g

$$g(P_i) = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} \tag{3.8}$$

heißt *Link-Funktion*. Die Klasse der durch (3.8) definierten Modelle ist die der *Generalisierten Linearen Modelle*. Der Fall, dass $g(P_i) = y_i$, ist ein Spezialfall.

3.2 Das logistische Modell

Das logistische Modell ist durch

$$P(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp(\lambda(\mathbf{x}))} \tag{3.9}$$

definiert, wobei oft

$$\lambda(\mathbf{x}) = \mathbf{b}'\mathbf{x} = b_0 + b_1 x_1 + \cdots + b_p x_p, \tag{3.10}$$

mit $\mathbf{x} = (1, x_1, x_2, \ldots, x_p)'$ angesetzt wird.

Es ist $P(Y = 0|\mathbf{x}) = 1 - P(Y = 1|\mathbf{x})$. Dementsprechend hat man

$$P(Y = 0|\mathbf{x}) = 1 - \frac{1}{1 + \exp(\lambda(\mathbf{x}))} = \frac{1 + \exp(\lambda(\mathbf{x})) - 1}{1 + \exp(\lambda(\mathbf{x}))} = \frac{\exp(\lambda(\mathbf{x}))}{1 + \exp(\lambda(\mathbf{x}))},$$

und die Multiplikation von Zähler und Nenner mit $\exp(-\lambda(\mathbf{x}))$ liefert

$$P(Y = 0|\mathbf{x}) = \frac{1}{1 + \exp(-\lambda(\mathbf{x}))}. \tag{3.11}$$

Gilt (3.10), so wird das Vorzeichen von λ bei der Schätzung der Parameter mitbestimmt, so dass man keine großen Fallunterscheidungen abhängig davon, ob man $P(Y = 1|\mathbf{x})$ oder $P(Y = 0|\mathbf{x})$ betrachtet, machen muß.

Zur Vereinfachung werde $p(\mathbf{x}) = P(Y = 1|\mathbf{x})$ geschrieben. Löst man (3.9) nach $\lambda(\mathbf{x})$ auf, erhält man

$$\log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = \lambda(\mathbf{x}), \tag{3.12}$$

d.h. die Link-Funktion ist durch

$$g = \log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) \tag{3.13}$$

definiert.

Definition 3.1 *Die Größe* $\log(p(x)/1-p(x))$ *heißt* Logit*, und die in (3.13) gegebene Transformation* g *der Wahrscheinlichkeit* $p(x)$ *heißt* Logit-Transformation*.*

Anmerkung: Der Ausdruck 'logistisch' geht auf das französische Wort 'logis' für Wohnung zurück: 1836 bekam der belgische Mathematiker Pierre Verhulst (1804 – 1849) von der französischen Regierung den Auftrag, das Wachstum der Bevölkerung von Paris und damit den notwendigen Bau von Wohnungen, Straßen, Kanalisation etc. abzuschätzen. Verhulst nahm an, dass das Wachstum zunächst exponentiell sein würde, sich dann aber abschwächen und gegen Null streben würde, dass also von einem gewissen Zeitpunkt an die Größe der Bevölkerung konstant bleiben würde: Schließlich muß eine Stadt mit Wasser, Lebensmitteln etc versorgt werden, und diese Güter kommen in erster Linie aus der Umgebung, die die Bevölkerung der Stadt "tragen" (s. unten) muß. Die implizite Annahme ist hier, dass die "Tragfähigkeit" der städtischen Umgebung konstant bleibt. Verhulst nahm das Modell

$$\frac{dN(t)}{dt} = rN(t)(1 - N(t)/K), \quad r > 0, \quad K > 0 \tag{3.14}$$

an. Dies ist eine Differentialgleichung: der Ausdruck $dN(t)/dt$ ist ein Maß für die Veränderung der Anzahl $N(t)$ der Bewohner, die hier als stetige und differenzierbare Funktion der Zeit angenommen wird. Diese Annahme ist vernünftig, denn Unstetigkeiten der Art

$$f(t) = \begin{cases} g(t), & t \leq t_0 \\ g(t) + a, a > 0, & t > t_0 \end{cases}$$

bedeuten plötzlich auftretende Ereignisse wie ein Atomschlag ($a < 0$) oder ein jählings einsetzender Zuzug von Flüchtlingen ($a > 0$), aber derartige Ereignisse haben einerseits eine sehr kleine Wahrscheinlichkeit, verkomplizieren andererseits aber die Betrachtungen enorm. Diese Argumentation überträgt sich auf die Möglichkeit von "Knicken" in der Funktion $N(t)$, an denen die Funktion $N(t)$ dann nicht differenzierbar ist.

Ist $N(t)/K$ klein im Vergleich zu 1, so ist $dN(t)/dt \approx rN(t)$, woraus $N(t) = k\exp(rt)$ folgt, k eine Konstante, d.h. N wächst in guter Näherung exponentiell. Für $N(t) \to K$ folgt aber $1 - N(t)/K \to 0$ und damit $dN(t)/dT \to 0$, d.h. die Größe der Bevölkerung verändert sich schließlich nicht mehr und bleibt konstant, wenn sie den Wert K erreicht hat; K gilt demnach *Trägerkonstante*: sie gibt an, wieviele menschen durch die Umgebung der Stadt "getragen" werden können. Das logistische Modell (3.9) kann u. a. aus einer Verteilungsfunktion hergeleitet werden, die eine zu (3.14) analoge Form hat, – dies ist die logistische Verteilung (vergl. Abschnitt 3.2.1).

3.2.1 Herleitung des Modells aus der logistischen Verteilung

Es werde angenommen, dass das in Frage stehende Ereignis, also $Y = 1$, eintritt, wenn eine nicht direkt beobachtbare ("latente") Variable ξ einen kritischen Wert überschreitet: man löst eine Aufgabe, wenn die kognitiven Fähigkeiten, zusammengefasst in der zufälligen Veränderlichen ξ, gerade einen bestimmten Wert überschreitet, ein Unfall geschieht, wenn die Aufmerksamkeit ξ einen bestimmten Wert unterschreitet, eine Krankheit bricht aus, wenn eine Belastung, repräsentiert durch die Variable xi, einen kritischen Wert überschreitet, etc. Die unabhängigen Variablen $\mathbf{x} = (x_1, \ldots, x_p)'$ bestimmen die Parameter der Verteilung von ξ, etwa den Erwartungswert $\mu = \mathbb{E}(\xi)$, so dass $\mu = \mu(x_1, \ldots, x_p)$. Je größer μ, desto größer soll die Wahrscheinlichkeit sein, dass ξ einen kritischen Wert überschreitet (je kleiner μ, desto größer die Wahrscheinlichkeit, dass ein kritischer Wert unterschritten wird).

Es ist eine Standardannahme, im Rahmen solcher Annahmen zu postzulieren, dass xi normalverteilt ist. Eine mathematisch einfacher zu handhabende, aber im Übrigen sehr ähnliche Verteilung ist die logistische Verteilung, die durch die Differentialgleichung

$$f(x) = \frac{dF(x)}{dx} = kF(x)(1 - F(x)). \tag{3.15}$$

charakterisiert ist; man sieht die Ähnlichkeit zum Verhulstschen Ansatz (3.14). Es läßt sich zeigen, dass die Lösung der Differentialgleichung auf die Verteilungsfunktion

$$P(\xi \leq S) = \frac{1}{1 + \exp\left(-\frac{(S-\mu)}{\sigma}\frac{\pi}{\sqrt{3}}\right)}, \tag{3.16}$$

oder

$$P(\xi \leq S) = \frac{1}{1 + \exp\left(-1.8138(S - \mu)/\sigma\right)}, \quad 1.8138 \approx \pi/\sqrt{3} \tag{3.17}$$

führt. μ und σ sind die Parameter der Verteilung; es gilt

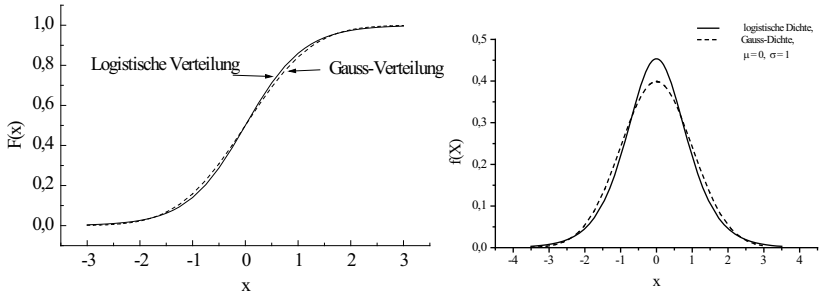
$$\mathbb{E}(\xi) = \mu, \quad Var(\xi) = \sigma^2, \tag{3.18}$$

wobei hier $\pi = 3.14159\ldots$. Für $\mu = 0$ und $\sigma^2 = 1$ erhält man die standardisierte logistische Verteilung. Die Abbildung 2 zeigt die Verläufe von Gauß- und logistischer Verteilung. Man sieht, dass die Unterschiede zwischen den Verteilungen für die meisten praktischen Zwecke vernachlässigbar sind.

Das Modell (3.9) ergibt sich aus (3.17) durch *Reparametrisierung*. Setzt man $a = 1.8138$, so ist ja in (3.17)

$$-1.8138(S - \mu)/\sigma = \alpha S - \beta\mu,$$

Abbildung 2: Verläufe der standardisierten Gauß- und der logistischen Verteilung



mit $\alpha = -a/\sigma$, $\beta = a/\sigma$. Setzt man $\mu = b_0 + b_1 x_1 + \cdots + b_p x_p$, so erhält man

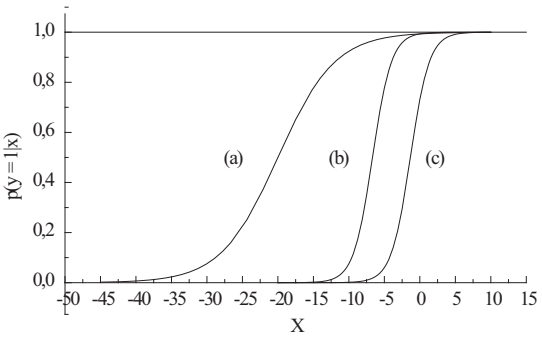
$$\alpha S - \beta\mu = \alpha S - \beta(b_0 + b_1 + \cdots + b_p x_p)$$

und führt man die Umbenennungen

$$b_0 \to \alpha S - \beta b_0, \quad b_j \to \beta b_j$$

durch, so hat man bereits das logistische Modell (3.9).

Abbildung 3: Logistische Funktionen für verschiedene Parameterwerte



3.2.2 Herleitung des Modells über den Satz von Bayes

Nach dem Satz von Bayes gilt

$$P(Y = 1|\mathbf{x}) = P(\mathbf{x}|Y = 1)\frac{P(Y = 1)}{P(\mathbf{x})}. \tag{3.19}$$

$P(Y = 1|\mathbf{x})$ ist die Posteriori-Wahrscheinlichkeit, $P(\mathbf{x}|Y = 1)$ ist die Likelihood von $\mathbf{x}$, gegeben $Y = 1$, und $P(Y = 1)$ ist die Priori-Wahrscheinlichkeit. Wegen des Satzes der Totalen Wahrscheinlichkeit gilt

$$P(\mathbf{x}) = P(\mathbf{x}|Y_i = 1)P(Y_i = 1) + P(\mathbf{x}|Y_i = 0)P(Y_i = 0).$$

Setzt man zur Abkürzung $\pi_i = P(Y_i = 1)$ und berücksichtigt man, dass $P(Y_i = 0) = 1 - P(Y_i = 1) = 1 - \pi_i$ ist, so erhält man

$$P(\mathbf{x}) = P(\mathbf{x}|Y_i = 1)\pi_i + P(\mathbf{x}|Y_i = 0)(1 - \pi_i). \tag{3.20}$$

Setzt man diesen Ausdruck in (3.19) ein, so erhält man

$$P(Y_i = 1|\mathbf{x}) = \frac{P(\mathbf{x}|Y_i = 1)\pi_i}{P(\mathbf{x}|Y_i = 1)\pi_i + P(\mathbf{x}|Y_i = 0)(1 - \pi_i)}. \tag{3.21}$$

Dividiert man den Zähler und den Nenner auf der rechten Seite durch $P(\mathbf{x}|Y_i = 1)\pi_i$, so bekommt man den Ausdruck

$$P(Y_i = 1|\mathbf{x}) = \frac{1}{1 + \frac{P(\mathbf{x}|Y_i=0)(1-\pi_i)}{P(\mathbf{x}|Y_i=1)\pi_i}}. \tag{3.22}$$

Nun ist aber

$$\lambda(\mathbf{x}) = \frac{P(\mathbf{x}|Y_i = 1)}{P(\mathbf{x}|Y_i = 0)} \frac{\pi_i}{(1 - \pi_i)} \tag{3.23}$$

wobei $P(\mathbf{x}|Y_i = 1)/P(\mathbf{x}|Y_i = 0)$ bekanntlich der Likelihood-Quotient ist. Andererseits ist

$$\frac{P(\mathbf{x}|Y_i = 0)}{P(\mathbf{x}|Y_i = 1)} \frac{(1 - \pi_i)}{\pi_i} = \frac{1}{\lambda(\mathbf{x})}, \tag{3.24}$$

und offenbar ist $1/\lambda(\mathbf{x}) = \exp(-\log \lambda(\mathbf{x}))$, so dass (3.22) in der Form

$$P(Y_i = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\log \lambda(\mathbf{x}))} \tag{3.25}$$

geschrieben werden kann. Damit hat man eine Beziehung zur logistischen Funktion hergestellt. Welche Form sie annimmt, hängt von der Wahl der Wahrscheinlichkeitsfunktion $P(\mathbf{x}|Y_i)$ ab.

Es werde nun angenommen, dass $\mathbf{x}$ multivariat normalverteilt ist, d.h.

$$f(\mathbf{x}) = A \exp(-(\mathbf{x} - \vec{\mu})'\Sigma^{-1}(\mathbf{x} - \vec{\mu})), \tag{3.26}$$

wobei A eine Normierungskonstante ist und $\vec{\mu} = (\mu_1, \ldots, \mu_p)$ der Vektor der Erwartungswerte der X_j. Es gibt zwei Möglichkeiten: $\mathbf{x}$ kommt asus einer Population mit dem Erwartungwert $\vec{\mu} = \vec{\mu}_0$, oder aus einer Poulation mit dem Erwartungswert $\vec{\mu} = \vec{\mu}_1$. $\vec{\mu}_0$ definiere etwa die Population der "Unfähigen",

$\vec{\mu}_1$ die der "Fähigen", oder "Gesunde" versus "Kranke", etc. Der Einfachheit halber werde angenommen, dass die Varianz-Kovarianzmarix Σ fürbeide Populationen identisch ist. Dann ist

$$\begin{aligned}\frac{P(\mathbf{x}|Y_i=0)}{P(\mathbf{x}|Y_i=1)} &= \frac{\exp(-(\mathbf{x}-\vec{\mu}_0)'\Sigma^{-1}(\mathbf{x}-\vec{\mu}_0))}{\exp(-(\mathbf{x}-\vec{\mu}_1)'\Sigma^{-1}(\mathbf{x}-\vec{\mu}_1))} \\ &= \exp((\mathbf{x}-\vec{\mu}_1)'\Sigma^{-1}(\mathbf{x}-\vec{\mu}_1)) - (\mathbf{x}-\vec{\mu}_0)'\Sigma^{-1}(\mathbf{x}-\vec{\mu}_0)),\end{aligned}$$

so dass

$$\log\left(\frac{P(\mathbf{x}|Y_i=0)}{P(\mathbf{x}|Y_i=1)}\right) = (\mathbf{x}-\vec{\mu}_1)'\Sigma^{-1}(\mathbf{x}-\vec{\mu}_1)) - (\mathbf{x}-\vec{\mu}_0)'\Sigma^{-1}(\mathbf{x}-\vec{\mu}_0).$$

Multipliziert man die Terme aus, so ergibt sich für die rechte Seite

$$2\mathbf{x}'\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0) + (\vec{\mu}_1-\vec{\mu}_0)'\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0),$$

wobei $\Sigma^{-1}(\vec{\nu}_1-\vec{\mu}_0)$ ein Vektor und $(\vec{\mu}_1-\vec{\mu}_0)'\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0)$ ein Skalar ist. Setzt man $\mathbf{b} = 2\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0)$ und $b_0 = (\vec{\mu}_1-\vec{\mu}_0)'\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0) + \log(\pi/(1-\pi)$, so erhält man wieder

$$P(Y=1|\mathbf{x}) = \frac{1}{1+\exp(b_0+b_1x_1+\cdots+b_px_p)},$$

d.h. das logistische Modell.

Der Bayessche Ansatz zur Herleitung des logistischen Modells läßt u. U. die Interpretation der Entscheidung zwischen zwei Populationen zu. Für $P(Y=1|\mathbf{x}) > P_0$ entscheidet man z.B., dass eine Person aus der Population der "Fähigen", der "Willigen", der "Kranken" etc ist, andernfalls gehört sie zur jeweiligen Komplementärpopulation.

3.2.3 Regularisierte logistische Regression

Die Parameter für einen logistischen Regressionsansatz werden üblicherweise mit der Maximum-Likelihood-Methode (ML-Methode) geschätzt. Für N Beobachtungen hat man die Likelihood

$$l(\mathbf{b}) = \sum_{i=1}^{N} \log p_{g_i}(\mathbf{x}_i;\mathbf{b}), \tag{3.27}$$

mit $p_k(\mathbf{x}_i;\mathbf{b}) = P(\mathcal{C}_k|\mathbf{x}_i,\mathbf{b})$, $\mathbf{x}_i$ der Vektor der Prädiktorwerte für den i-ten Fall, $\mathbf{b}$ der Parametervektor. $g_i = 1$, wenn der i-te Fall zu cC_1 gehört, $g_i = 0$ sonst. Es sei $y_i = 1$, wenn der i-Fall zu $\mathcal{C}_1$ gehört, und $y_i = 0$, wenn $g_i = 1$, und $y_i = 0$ für $g_i = 0$. Weiter werde $p_1(\mathbf{x},\mathbf{b}) = p(\mathbf{x},\mathbf{b})$ gesetzt und $p_2(\mathbf{x},\mathbf{b}) = 1 - p_1(\mathbf{x},\mathbf{b})$. Dann hat man

$$\begin{aligned} l(\mathbf{b}) &= \sum_{i=1}^{N} \left[y_i \log p(\mathbf{x}_i;\mathbf{b}) + (1-y_i)\log(1-p(\mathbf{x},\mathbf{b}))\right] \\ &= \sum_{i=1}^{N} \left[y_i\mathbf{b}'\mathbf{x}_i - \log(1+\exp(\mathbf{b}'\mathbf{x}_i)\right] \end{aligned} \tag{3.28}$$

Die Likelihodd wird maximiert, wenn die partiellen Ableitungen von $l(\mathbf{b})$ gleich Null gesetzt werden:

$$\frac{\partial \mathbf{b}}{\partial b_k} = \sum_{i=1}^{N} \mathbf{x}_i (y_i - p(\mathbf{x}_i, \mathbf{b}) = 0. \tag{3.29}$$

Die Gleichungen sind nichtlinear in den Unbekannten b_k.

Man kann zu einer penalisierten Version der Schätzung übergehen, wobei man insbesondere das Lasso (Abschnitt 2.5.3) benutzt; dann wird das Maximum

$$\max_{b_0, \mathbf{b}} \left[\sum_{i=1}^{N} \left(y_i (b_0 + \mathbf{b} * \mathbf{x}_i) - \log(1 + \exp(b_0 + \mathbf{b}' \mathbf{x}_i)) \right) - \lambda \sum_{j=1}^{p} |b_j| \right] \tag{3.30}$$

bestimmt.

Zur Berechnung kann das R-Packet glmet (Friedman et al, 2001) verwendet werden.

3.2.4 Kategoriale Prädiktorvariablen

Es ist hier nicht notwendig, dass die x_k Intervallskalenniveau haben. Es ist möglich, dass die x_k kategoriale Variablen sind. Der Unterschied zu den später zu betrachtenden *loglinearen Modellen* besteht dann eigentlich nur noch darin, dass eine (oder mehrere) Variable als Antwortvariable ausgezeichnet werden, die Betrachtung in dieser Hinsicht also asymmetrisch ist.

Im Falle kategorialer erklärender Variablen entsprechen die b_p in (**??**) den Haupteffekten in einer Varianzanalyse. Dies führt zur *Effektkodierung* für die x_k: es gelte

$$x_k = \begin{cases} 1, & \text{es liegt Kateg. } k \text{ vor} \\ -1, & \text{es liegt Kateg. } j \neq k \text{ vor} \\ 0, & \text{sonst} \end{cases} \tag{3.31}$$

für $j = 1, \cdots, r-1$; es muß nur bis zur $(r-1)$-ten Kategorie betrachtet werden, da die letzte bei Bestimmung dieser Kategorien festliegt. Es gilt also

$$b_p = -\sum_{j=1}^{k-1} b_j, \quad \text{oder} \quad \sum_{j=1}^{k} b_j = 0 \tag{3.32}$$

Dementsprechend gilt dann für eine gegebene Link-Funktion g

$$g(p(x)) = b_0 + b_j, \quad j = 1, \cdots, k-1 \tag{3.33}$$

und

$$g(p(x)) = b_0 - b_1 - \cdots - b_{k-1}, \quad j = k \tag{3.34}$$

Bei varianzanalytischen Designs ist es üblich, mehr als nur eine unabhängige Variable (Faktoren) zu betrachten. Es seien etwa die Faktoren A und B gegeben. Sie können durch die Vektoren X^A und X^B repäsentiert werden, deren Komponenten durch

$$x_i^A, i = 1, \cdots, I-1; \quad x_j^B, j = 1, \cdots, J-1 \tag{3.35}$$

gegeben sind und die wie in (3.31) als Dummy-Variablen definiert sind. $x_i^A = x_j^B = 1$ heißt dann, dass die Bedingung $A_i \cap B_j$ (A_i und B_j) gegeben ist. Die beiden Vektoren können zu einem einzelnen Merkmalsvektor $\vec{X}$ zusammengefaßt werden:

$$\vec{X} = (x_1^A, \cdots, x_{I-1}^A, x_1^B, \cdots, x_{J-1}^B, x_1^A x_2^B, \cdots, x_1^A x_{J-1}^B, \cdots, x_{I-1}^A x_{J-1}^B)' \tag{3.36}$$

Die Komponenten $x_1^A x_1^B$ etc. repräsentieren die möglichen Wechselwirkungskomponenten. Dann ergibt sich das Modell

$$g(p(\vec{X})) = b_0 + \vec{X}'\vec{b} \tag{3.37}$$

wobei $\vec{b}$ ein Vektor ist, dessen Komponenten die Haupt- und Wechselwirkungseffekte repräsentieren. Für g kann wieder die Logit-Transformation gewählt werden.

3.2.5 Interpretation der Parameter

Aus dem logistischen Modell

$$P(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp(b_0 + b_1 x_1 + \cdots + b_p x_p)}$$

ergibt sich der Wettquotient (odds ratio)

$$\frac{P(Y = 1|\mathbf{x})}{P(Y = 0|\mathbf{x})} = \frac{P(Y = 1|\mathbf{x})}{1 - P(Y = 1|\mathbf{x})} = \exp(b_0 + b_1 x_1 + \cdots + b_p x_p). \tag{3.38}$$

Schreibt man zur Vereinfaschung $p(\mathbf{x})$ für $P(Y = 1|\mathbf{x})$, so erhält man

$$\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = e^{b_0} e^{b_1 x_1} \cdots e^{b_p x_p}. \tag{3.39}$$

Angenommen, man verändert nun x_j um eine Einheit, sodass x_j in $x_j + 1$ übergeht. Dann hat man etwa für $j = 1$

$$\frac{p(\tilde{\mathbf{x}})}{1 - p(\tilde{\mathbf{x}})} = e^{b_0} e^{b_1(x_1+1)} \cdots e^{b_p x_p} = e^{b_0} e^{b_1 x_1} e^{b_1} \cdots e^{b_p x_p}, \tag{3.40}$$

wobei $\tilde{\mathbf{x}} = (x_1+1, x_2, \ldots, x_p)'$. Die Wettquotienten für $\mathbf{x}$ und $\tilde{\mathbf{x}}$ unterscheiden sich demnach um den Faktor $\exp(b_1)$;

$$\frac{p(\tilde{\mathbf{x}})}{1 - p(\tilde{\mathbf{x}})} = \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} e^{b_1} \tag{3.41}$$

Für die Logits gilt

$$\text{Logit}(\tilde{\mathbf{x}}) = \text{Logit}(\mathbf{x}) + b_1 \tag{3.42}$$

Ein Regressionskoeffizient ist also nicht so einfach zu interpretieren wie bei der multiplen Regression (einfach: bei unabhängigen Prädiktoren!). b_j definiert den Faktor e^{b_j}, um den sich die Wettchance verändert, oder die additive Konstante, um die sich das Logit verändert. Die Veränderung der Wahrscheinlichkeit $P(Y = 1|\mathbf{x})$ mit der Veränderung eines Prädiktors ist keineswegs direkt aus b_j ablesbar. Man hat etwa

$$P(Y = 1|\tilde{\mathbf{x}}) = \frac{1}{1 + \exp(b_0 + b_1 x_1 + b_p x_p)\exp(b_1)}. \tag{3.43}$$

Wie stark sich der Faktor $\exp(b_1)$ oder allgemein $\exp(b_j)$ auswirkt, hängt von den Werten der Komponenten von $\mathbf{x}$ ab.

Odds Ratio und relatives Risiko Für die Interpretation ist die Wettchance bzw. der Odds Ratio (3.41) wichtig. Zuerst seien die Definitionen

Tabelle 1: Odds, Odds Ratio und Relatives Risiko

	Risiko R		
Effekt E	1	0	Σ
1	a (n_{11})	b (n_{12})	$a+b$
0	c (n_{21})	d (n_{22})	$c+d$
Σ	$a+c$	$b+d$	N

zusammengefasst:

1. **Odds** (Wettchancen) Diese Größen sind schon eingeführt worden und werden nur wegen der Systematik noch einmal mit der Nomenklatur der Tabelle 1 definiert:

$$\Omega_1 = \frac{P(E = 1|R = 1)}{P(E = 0|R = 1)}, \quad \Omega_2 = \frac{P(E = 1)|R = 0)}{P(E = 0|R = 0)} \tag{3.44}$$

   Hier ist $P(E = 1|R = 1) = P(Y = 1|\mathbf{x}_1)$, $P(E = 0|R = 1) = P(Y = 0|\mathbf{x}_1)$, $P(E = 1|R = 0) = P(Y = 1|\mathbf{x}_0)$, $P(E = 0|R = 0) = P(Y = 0|\mathbf{x}_0)$. $\mathbf{x}_0$ und $\mathbf{x}_1$ repräsentieren zwei verschiedene Vektoren von Bedingungen, die zu "Risiko nicht vorhanden" $R = 0$ und "Risiko vorhanden" $R = 1$ korrespondieren.

   Aus der Definition der bedingten Wahrscheinlichkeit folgen sofort die Schätzungen

$$\hat{\Omega} = \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}, \quad \hat{\Omega}_2 = \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}. \tag{3.45}$$

2. **Relatives Risiko** Das relative Risiko ist das Verhältnis etwa der Wahrscheinlichkeit, dass man erkrankt ($E = 1$), wenn man unter bestimmten Bedingung lebt (etwa wenn man raucht, $R = 1$) zur Wahrscheinlichkeit, dass man erkrankt, wenn diese Bedingung nicht gegeben ist (wenn man also nicht raucht):

$$RR = \frac{P(E = 1|R = 1)}{P(E = 1|R = 0)} \tag{3.46}$$

Es ist $P(E = 1|R = 1) = a/(a + c)$ die bedingte Wahrscheinlichkeit (dh eine Schätzung davon) eines Effekts, gegeben den Risikofaktor. Diese Größe entspricht dem Risiko, einen Effekt zu zeigen, wenn das Risiko gegeben ist. $P(E = 1|R = 0) = b/(b + d)$ ist eine Schätzung der bedingten Wahrscheinlichkeit eines Effekts, wenn kein Risikofaktor gegeben ist; diese Größe ist das Risiko, wenn der Risikofaktor nicht vorliegt. Dann ist das *relative Risiko* durch den Quotienten

$$\widehat{RR} = \frac{\hat{P}(E = 1|R = 1)}{\hat{P}(E = 1|R = 0)} = \frac{a/(a + c)}{b/(b + d)} \tag{3.47}$$

definiert. Offenbar gilt $0 \leq RR < \infty$.

3. **Odds Ratio** Der Odds-Ratio ist das Verhältnis der Odds:

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{P(E = 1|R = 1)}{P(E = 0|R = 1)} \frac{P(E = 0)|R = 0)}{P(E = 1|R = 0)} \tag{3.48}$$

Aus der Tabelle 1 erhält man die Schätzung

$$\hat{\Theta} = \frac{a/(a + c)}{c/(a + c)} \frac{d/(b + d)}{b/(b + d)} = \frac{a}{c} \cdot \frac{d}{b}. \tag{3.49}$$

Beispiel 3.1 Aspirin lindert nicht nur den Kopfschmerz, sondern verringert auch das Risiko, einen Herzinfarkt zu erleiden. In einer längeren Studie hat man 11034 Ärzten ein Placebo und 11037 anderen Ärzten täglich eine Aspirintablette gegeben[12]. Dabei handelte es sich um einen Blindversuch: keiner der beteiligten Ärzte wußte, ob er ein Placebo oder eine Aspirintablette schluckte. Die Daten sind hier zu einer 2×2-Tabelle zusammengefaßt worden. Für das relative Risiko eines Herzinfarkts erhält man

$$\widehat{RR}_{HI} = \frac{p(B_1|A_1)}{p(B_1|A_2)} = \frac{104/11037}{189/11034} = \frac{104}{189} \cdot \frac{11034}{11037} = .5501$$

Der Anteil der Personen, die einen Herzinfarkt bekamen *und* Aspirin genommen haben ist nur etwas mehr als halb so groß wie der Anteil derjenigen

[12]Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study (1988), N. Engl. J. Med. 318, 262-264

Tabelle 2: Aspirin und die Wahrscheinlichkeit von Herzinfarkten

	Herzinfarkt		
Medikament	B_1 $(y = 1)$	B_2 $(y = 0)$	Σ
Aspirin $(A_1; x = 1)$	104	10933	11037
Placebo $(A_2; x = 0)$	189	10845	11034
Σ	293	21778	22071

Personen, die einen Herzinfarkt bekamen *und* kein Aspirin genommen haben. Man kann auch das relative "Risiko", *keinen* Herzinfarkt zu bekommen, berechnen:

$$\widehat{RR}_{\neg HI} = \frac{p(B_2|A_1)}{p(B_2|A_2)} = \frac{10933}{10845} \cdot \frac{11034}{11037} = 1.00784$$

Man sieht, dass die beiden Arten von Risiken nicht komplementär sind. Der Wert von $R_{\neg HI}$ spiegelt die Tatsache wider, dass insgesamt die Wahrscheinlichkeit, einen Herzinfarkt zu bekommen, relativ klein ist ($p(HI) = 293/22071 = .01328$); der Effekt des Aspirins geht, betrachtet man die (Teil-)Population der Personen ohne Herzinfarkt, gewissermaßen verloren.

Die Odds, einen Herzinfarkt zu bekommen, wenn man Aspirin nimmt, sind

$$\widehat{\Omega}_1 = \frac{p(B_1|A_1)}{p(B_2|A_1)} = \frac{104/11037}{10933/11037} = \frac{104}{10933} = .00951$$

Die entsprechenden Odds für die Bedingung, ein Placebo verabreicht bekommen zu haben, sind

$$\widehat{\Omega}_2 = \frac{p(B_1|A_2)}{p(B_2|A_2)} = \frac{189/11034}{10845/11034} = \frac{189}{10845} = .01743$$

Offenbar sind die Odds für einen Herzinfarkt unter der Bedingung, Aspirin bekommen zu haben, geringer als die unter der Bedingung, nur ein Placebo verabreicht bekommen zu haben. Man enthält für das Kreuzproduktverhältnis

$$\widehat{\Theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{104 \cdot 10845}{189 \cdot 10933} = \frac{.00951}{.01743} = .54584$$

Die Odds für einen Herzinfarkt unter der Bedingung, Aspirin eingenommen zu haben, sind fast auf die Hälfte reduziert relativ zu der Bedingung, nur ein Placebo verabreicht bekommen zu haben. □

3.3 Dichotome Prädiktoren und Beispiele

Es werden zunächst einige Begriffe eingeführt. A_1 und A_2 mögen zwei verschiedene Bedingungen repräsentieren, A_1 etwa stehe für "es wird ein Placebo verabreicht", und A_2 stehe für "es wird ein Medikament verabreicht".

B_1 und B_2 bezeichne verschiedene Resultate, wie B_1 "der Patient erleidet einen Herzinfarkt", B_2 "der Patient erleidet keinen Herzinfarkt".

Definition 3.2 *Der Quotient*

$$R = \frac{p(B_j|A_1)}{p(B_j|A_2)} \geq 0 \tag{3.50}$$

heißt relatives Risiko.

Im Falle der stochastischen Unabhängigkeit von unabhängiger und abhängiger Variable gilt $p(B_j|A_1) = p(B_j|A_2) = p(B_j)$; in diesem Fall ist das relative Risiko $R = 1$.

Definition 3.3 *Die Quotienten*

$$\Omega_1 = \frac{p(B_1|A_1)}{p(B_2|A_1)}, \quad \Omega_2 = \frac{p(B_1|A_2)}{p(B_2|A_2)} \tag{3.51}$$

heißen Odds *oder* Wettchancen.

Man kann z.B. wetten, dass B_1 bzw. B_2 eintritt, wenn entweder A_1 oder A_2 zutrifft. Die Chancen, zu gewinnen, sind dann durch Ω_1 bzw. Ω_2 gegeben.

Die Odds lassen sich leicht aus den Häufigkeiten n_{ij} ausrechnen. Denn es ist ja $P(A_i|B_j) = n_{ij}/n_{i+}$, und deshalb hat man sofort

$$\Omega_1 = \frac{n_{11}/n_{1+}}{n_{12}n_{1+}} = \frac{n_{11}}{n_{12}}, \quad \Omega_2 = \frac{n_{21}/n_{2+}}{n_{22}/n_{2+}} = \frac{n_{21}}{n_{22}} \tag{3.52}$$

Definition 3.4 *Das Verhältnis*

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{p(B_1|A_1)p(B_2|A_2)}{p(B_2|A_1)p(B_1|A_2)} \geq 1 \tag{3.53}$$

heißt Kreuzproduktverhältnis*, oder auch* odds ratio.

Wegen (3.52) hat man sofort

$$\Theta = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{3.54}$$

Θ ist ein Maß für die Stärke der Assoziation zwischen der unabhängigen und der abhängigen Variablen. Sind diese beiden Variablen stochastisch unabhängig, so ergibt sich $\Theta = 1$, wie man leicht überprüft. Die Assoziation zwischen den Variablen ist um so größer, je mehr Θ von 1 abweicht. Für $\Theta \to 0$ ist die Beziehung zwischen den Variablen *negativ*, d.h. $x = 1$ impliziert $y = 0$ und $x = 0$ bedeutet $y = 1$, und für $\Theta \to \infty$ ist sie *positiv*; für größer werdendes Θ gehen $\pi_{1|2} = p(B_1|A_2)$ und $\pi_{2|1} = p(B_2|A_1)$ gegen Null.

Der Prädiktor x repräsentiere nun eine dichotome unabhängige Variable, wenn x also nur die Werte 0 oder 1 annehmen kann. Die Kontingenztabelle

Tabelle 3: Logistische Regression für einen dichotomen Prädiktor

	Abhängige Variable (y)	
Unabh. Variable (x)	$y = 1$	$y = 0$
$x = 1$	$e^{A+B}/(1 + e^{A+B})$	$1/(1 + e^{A+B})$
$x = 0$	$e^{B}/(1 + e^{B})$	$1/(1 + e^{B})$

ist dann eine einfache 2×2-Tabelle: Die Maximum-Likelihood-Schätzungen für die Parameter werden in Abschnitt 3.5.4 hergeleitet.

Die ursprünglich für deskriptive Zwecke eingeführten Größen *Odds* und *Kreuzproduktverhältnis* erweisen sich nun gerade als die ML-Schätzungen für die unbekannten Parameter A und B (vergl. Abschnitt 3.5.4, Seite 72)

$$\frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{\Omega_1}{\Omega_2} = \Theta = \exp(\hat{A}), \quad \text{bzw. } \log\Theta = \hat{A} \tag{3.55}$$

und (vergl. (3.52))

$$\frac{n_{21}}{n_{22}} = \Omega_2 = \exp(\hat{B}), \quad \text{bzw. } \log\Omega_2 = \hat{B}. \tag{3.56}$$

Beispiel 3.2 Aspirin lindert nicht nur den Kopfschmerz, sondern verringert auch das Risiko, einen Herzinfarkt zu erleiden. In einer längeren Studie hat man 11034 Ärzten ein Placebo und 11037 anderen Ärzten täglich eine Aspirintablette gegeben[13]. Dabei handelte es sich um einen doppelten Blindversuch: keiner der beteiligten Ärzte wußte, ob er ein Placebo oder eine Aspirintablette schluckte. Die Daten sind hier zu einer 2×2-Tabelle zusammengefaßt worden. Für das relative Risiko eines Herzinfarkts erhält man

$$R_{HI} = \frac{p(B_1|A_1)}{p(B_1|A_2)} = \frac{104/11037}{189/11034} = \frac{104}{189} \times \frac{11034}{11037} = .5501$$

Der Anteil der Personen, die einen Herzinfarkt bekamen *und* Aspirin genommen haben ist nur etwas mehr als halb so groß wie der Anteil derjenigen Personen, die einen Herzinfarkt bekamen *und* kein Aspirin genommen haben. Man kann auch das relative "Risiko", *keinen* Herzinfarkt zu bekommen, berechnen:

$$R_{\neg HI} = \frac{p(B_2|A_1)}{p(B_2|A_2)} = \frac{10933}{10845} \times \frac{11034}{11037} = 1.00784$$

[13] Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study (1988), N. Engl. J. Med. 318, 262-264

Tabelle 4: Aspirin und die Wahrscheinlichkeit von Herzinfarkten

	Herzinfarkt		
Medikament	B_1 $(y = 1)$	B_2 $(y = 0)$	Σ
Aspirin $(A_1; x = 1)$	104	10933	11037
Placebo $(A_2; x = 0)$	189	10845	11034
Σ	293	21778	22071

Man sieht, dass die beiden Arten von Risiken nicht komplementär sind. Der Wert von $R_{\neg HI}$ spiegelt die Tatsache wider, dass insgesamt die Wahrscheinlichkeit, einen Herzinfarkt zu bekommen, relativ klein ist ($p(HI) = 293/22071 = .01328$); der Effekt des Aspirins geht, betrachtet man die (Teil-)Population der Personen ohne Herzinfarkt, gewissermaßen verloren.

Die Odds, einen Herzinfarkt zu bekommen, wenn man Aspirin nimmt, sind

$$\Omega_1 = \frac{p(B_1|A_1)}{p(B_2|A_1)} = \frac{104/11037}{10933/11037} = \frac{104}{10933} = .00951$$

Die entsprechenden Odds für die Bedingung, ein Placebo verabreicht bekommen zu haben, sind

$$\Omega_2 = \frac{p(B_1|A_2)}{p(B_2|A_2)} = \frac{189/11034}{10845/11034} = \frac{189}{10845} = .01743$$

Offenbar sind die Odds für einen Herzinfarkt unter der Bedingung, Aspirin bekommen zu haben, geringer als die unter der Bedingung, nur ein Placebo verabreicht bekommen zu haben. Man enthält für das Kreuzproduktverhältnis

$$\Theta = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{104 \times 10845}{189 \times 10933} = \frac{.00951}{.01743} = .54584$$

Die Odds für einen Herzinfarkt unter der Bedingung, Aspirin eingenommen zu haben, sind fast auf die Hälfte reduziert relativ zu der Bedingung, nur ein Placebo verabreicht bekommen zu haben. □

Beispiel 3.3 (Fortsetzung des Beispiels 3.2) In Beispiel 3.2 werde

$$x = \begin{cases} 0, & \text{Placebo} \\ 1, & \text{Aspirin} \end{cases} \tag{3.57}$$

definiert. Für die Odds hat man mit der Abkürzung HI für "Herzinfarkt"

$$\Omega_1 = \frac{P(\text{HI; ja}|\text{Aspirin})}{P(\text{HI; nein}|\text{Aspirin})} = e^{\hat{A}+\hat{B}}, \quad \Omega_2 = \frac{P(\text{HI; ja}|\text{Placebo})}{P(\text{HI; nein}|\text{Placebo})} = e^{\hat{B}},$$

und für das Kreuzproduktverhältnis erhält man

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\exp(\hat{A} + \hat{B})}{\exp(\hat{B})} = e^{\hat{A}};$$

hier reflektiert Θ den Zusammenhang zwischen der Medikamenteneinnahme (Placebo oder Aspirin) und der Erkrankung (Herzinfarkt oder nicht). Der Befund $\Theta = \exp(\hat{A})$ oder $\log \Theta = \hat{A}$ zeigt, dass tatsächlich θ und das Steigmaß A der Regressionsgeraden $Ax + B$ direkt aufeinander bezogen sind. Für $A = 0$ hängt die Erkrankung nicht von der Medikamention ab, und $\log \Theta = 0$ genau dann, wenn $\Theta = 1$, d.h. wenn $\Omega_1 = \Omega_2$. Aus (3.113) erhält man die Schätzung

$$\hat{A} = \log \Theta = \log 1.833 = .60595$$

und aus (3.114)

$$\hat{B} = \log \frac{n_{11}}{n_{12}} = .01743$$

□

Beispiel 3.4 Die Gefahr einer Infektion bei Kaiserschnittgeburten hängt von verschiedenen Faktoren ab, u.a. davon, ob der Kaiserschnitt geplant oder nicht geplant war, und welche Risikofaktoren wie Übergewicht, Diabetes etc vorhanden sind oder nicht. Die Tabelle 5 enthält die Daten einer Studie am Klinikum Großhadern in München[14]: Die Wahrscheinlichkeit einer Infektion kann mit dem Logit-Ansatz modelliert werden. Die Kodierung wird in Tabelle 6 angegeben. Dann gilt

$$\log \frac{P(\text{Infektion})}{P(\text{keine Infektion})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3. \tag{3.58}$$

Die Interpretation der Schätzungen ist direkt: Antibiotika verringern die log odds einer Infektion, und wegen der strengen Monotonie folgt weiter, dass sie das relative Risiko einer Infektion verringern. Risikofaktoren und ein nichtgeplanter Kaiserschnitt erhöhen die log odds und damit das relative Risiko einer Infektion.

Geht man von

$$\log(P(\text{Infektion})/P(\text{keine Infektion})) \text{ zu } P(\text{Infektion})/P(\text{keine Infektion})$$

über, so erhält man

$$\frac{P(\text{Infektion})}{P(\text{keine Infektion})} = \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_3 x_3).$$

[14] In: Fahrmeier,L., Tutz, G. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer-Verlag New York Berlin Heidelberg 1994, p.2 und p. 29

Tabelle 5: Infektionen beim Kaiserschnitt

	geplant		nicht geplant	
	Infektion		Infektion	
	ja	nein	ja	nein
Antibiotika				
mit Risikofakt.	1	17	11	87
ohne Risikofakt.	0	2	0	0
keine Antibiot.				
mit Risikofakt.	28	30	23	3
ohne Risikofakt.	8	32	0	9

Tabelle 6: Kodierung der Variablen

Kaiserschnitt	geplant:	$x_1 = 0$	nicht geplant:	$x_1 = 1$
Risikofaktor	vorhanden:	$x_2 = 1$	nicht vorhanden:	$x_2 = 0$
Antibiotika	verabreicht:	$x_3 = 1$	nicht verabreicht:	$x_3 = 0$

Die Schätzungen für die β_j, $j = 0, \ldots, 3$ sind

$$\hat{\beta}_0 = -1.89, \quad \hat{\beta}_1 = 1.07, \quad \hat{\beta}_2 = 2.03, \quad \hat{\beta}_3 = -3.25$$

Man findet für $f_i = \exp(\hat{\beta}_i x_i)$, $i = 0, 1, 2, 3$

$$f_0 = 0.151072, \quad f_1 = 2.91538, \quad f_2 = 7.61409, \quad f_3 = 0.0387742$$

Hieraus kann abgelesen werden, dass ein nicht geplanter Kaiserschnitt das relative Risiko einer Infektion um den Faktor $f_1 = \exp(1.07) = 2.92$ erhöht. Die Existenz von Risikofaktoren erhöht das relative Risiko für eine Infektion um den Faktor $f_2 = \exp(\hat{\beta}_2) = 7.6$ im Vergleich zu einer Situation ohne Risikofaktor.

Verabreichte Antibiotika reduzieren das Risiko um den Faktor $f_3 = 0.0387742$. Setzt man $x_3 = 0$, d.h. betrachtet man den Fall, dass keine Antibiotika gegeben werden, so ist $p(\text{Infektion})/(1 - p(\text{Infektion})) = 3.353$. Die Gabe von Antibiotika senkt das Risiko um den Faktor $f_3 \approx 4/100 = 1/25$,

d.h. man erhält $p(\text{Infektion})/(1-p(\text{Infektion})) = 3.353 \times .04 = .14$. Das Risiko beträgt jetzt nur noch den 25-ten Teil des ursprünglichen Risikos. □

3.4 Modelle für Anzahlen (counted data)

3.4.1 Poisson-Regression

Gelegentlich wird die Anzahl bestimmter Ereignisse erhoben, ohne dass die Gesamtzahl von Beobachtungen von vornherein festgelegt werden kann. Beispiele hierfür sind die Anzahl von Unfällen in einem Stadtbereich innerhalb eines bestimmten Zeitintervalls, die Anzahl der Selbstmorde in einer Stadt innerhalb eines Jahres, oder die Anzahl suizidaler Gedanken (suicidal ideations) von Patienten mit *posttraumatic stress disorder* (PTSD) (Boffa et a. (2017)). Der formale Punkt bei derartigen Folgen ist, dass eine Folge von Ereignissen nicht durch das Eintreten eines bestimmten zufälligen Ereignisses begrenzt wird, wie etwa eine Folge von Roulette-Versuchen mit dem Resultat, dass eine bestimmte Farbe (etwa schwarz) durch das Auftreten einer bestimmten anderen Farbe (rot) beendet wird, oder eine Anzahl von nicht erfolgreichen Bewerbungen durch eine erfolgreiche Bewerbung beendet wird.

Die Binomialverteilung

$$P(K = k|np) = \binom{n}{k}(p^k(1-p)^{n-k}, \quad \binom{n}{k} = \frac{k!}{n!(n-k)!}, \tag{3.59}$$

mit

$$\mathbb{E}(K) = np, \quad Var(K = np(1-p) \tag{3.60}$$

ist ein Beispiel für eine Verteilung, bei der die Gesamtzahl n der als unabhängig vorausgesetzten Versuche von vornherein festliegt. Man könnte an die geometrische Verteilung

$$P(K = k|p) = p(1-p)^{k-1}, \tag{3.61}$$

mit

$$\mathbb{E}(K) = \frac{1}{p}, \; Var(K) = np(1-p) \tag{3.62}$$

denken, bei K die Gesamtzahl der unabhängigen Versuche ist; $P(K = k|p)$ ist die Wahrscheinlichkeit, dass $k-1$ "Misserfolge" dem ersten "Erfolg" vorangehen. Bei den oben genannten Beispielen gibt es aber kein Ereignis, dass die Folge jeweils betrachteten Ereignisse stoppt.

Für die Binomialverteilung ist aber schon von dem französischen Mathematiker und Physiker S. D. Poisson[15] eine Approximation hergeleitet worden, die die in einer computerlosen, also schrecklichen Zeit die aufwändige Berechnung der Binomialkoeffizienten $\binom{n}{k}$ überlüssig machte: Für unabhängige Versuche mit binärem Ausgang und dem Grenzwert $pn \to \lambda < \infty$

[15] Siméon Denis Poisson (1781 – 1840)

für (i) $n \to \infty$ und $p \to 0$ gilt dann

$$P(K = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \tag{3.63}$$

mit

$$\mathbb{E}(K) = Var(K) = \lambda. \tag{3.64}$$

Ein bemerkenswerter Aspekt dieser Verteilung ist, dass Erwartungswert und Varianz identisch sind. Hängt nun die Verteilung von kontrollierbren Bedingungen ab, so liegt nahe, den Parameter λ als Funktion dieser Bedingungen aufzufassen, wobei die Bedingungen durch die unabhängigen Variablen $x_1, \ldots, x_q$ repräsentiert werden. Setzt man $\mathbf{b} = (b_0 + b_1, \cdots, b_N)'$ und $\mathbf{x} = (1, x_1, \ldots, x_q)$ so kann man

$$\lambda_i = b_0 + b_1 x_{i1} + \cdots + b_q x_{iq} = \mathbf{b}'\mathbf{x} \tag{3.65}$$

setzen. Dies ist das *lineare Poisson-Modell*. Alternativ kann man den Logarithmus von λ als lineare Funktion der Prädiktoren ansetzen:

$$\log \lambda_i = b_0 + b_1 x_{i1} + \cdots + b_q x_{iq} = \mathbf{b}'\mathbf{x} \tag{3.66}$$

so dass

$$\lambda_i = e^{b_0 + b_1 x_{i1} + \cdots + b_q x_{iq}} = e^{b_0} e^{b_1 x_{i1}} \cdots e^{b_q x_{iq}}. \tag{3.67}$$

Natürlich können auch Wechselwirkungen zwischen Variablen untersucht werden, etwa die zwischen x_1 und x_2, so dass man

$$\lambda = b_0 + b_1 x_1 + \cdots + b_q x_q + b_{q+1} x_1 x_2 \tag{3.68}$$

bzw.

$$\log \lambda = b_0 + b_1 x_1 + \cdots + b_q x_q + b_{q+1} x_1 x_2 \tag{3.69}$$

bekommt.

Beispiel 3.5 Es wird eine biomedizinische Studie zur immunaktivierenden Fähigkeit zweier Stoffe untersucht. Stoff 1 ist TNF (Tumor Nekrosefaktor), Stoff 2 ist IFN (Interferon). Dazu wurde die Anzahl von differenzierenden Zellen in Abhängigkeit von der Dosis für TNF und der für INF betrachtet, vergl. Tabelle 7. Getested wurde das log-lineare Modell

$$\lambda_i = \exp(b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_1 x_{i3}), \tag{3.70}$$

wobei x_1 für TNF und x_2 für INF steht. $x_1 x_2$ repräsentiert eine Interaktion, bzw. einen "synergistischen" Effekt. Die folgenden Parameter wurden bestimmt:

$$\hat{\mathbf{b}}_0 = 3.436, \quad \hat{b}_1 = .016, \quad \hat{\mathbf{b}}_2 = .009, \quad \hat{b}_3 = -.001.$$

Der sehr kleine Wert für b_3 legt nahe, dass keine synergistischen Effekte existieren. □

Tabelle 7: Häufigkeiten von differenzierenden Zellen

Anz. diff. Zellen	Dosis TNF	Dosis INF
11	0	0
18	0	4
20	0	20
39	0	100
22	0	1
38	1	4
52	1	20
69	6	100
31	10	4
68	10	4
69	10	20
128	10	10
102	100	0
171	100	0
180	100	20
193	100	100

3.4.2 Überdispersion und die Negative Binomialregression

Selbst wenn die Forderung der Unabhängigkeit der beobachteten Ereignisse erfüllt ist und die Poisson-Verteilung als vernünftige Wahl zur Charakterisierung der Daten betrachten kann, widerspricht die Beziehung (3.64) oft den Daten: die geschätzte Varianz $\hat{s}^2$ ist größer als die Schätzung $\bar{k}$ des Erwartungswerts λ. Man spricht dann von Überdispersion (over dispersion).

Dieser Befund bedeutet noch nicht, dass die Poisson-Verteilung grundsätzlich die falsche Verteilung ist. Denn eine implizite Annahme bei der Poisson-Regression ist, dass der Parameter λ für alle Fälle identisch ist, – d.h. dass die individuellen Werte nur unwesentlich voneinander abweichen. Das muß aber nicht so sein. In der untersuchten Population können größere als nur zufällige Abweichungen von einem Wert λ existieren, d.h. die Population kann heterogen in Bezug auf die individuellen λ-Werte sein.

Ein erster Ansatz, das Überdispersionsproblem zu lösen, besteht darin, einen Überdispersionsparameter ϕ einzuführen:

$$Var(y_i|\mathbf{x}_i) = \phi\lambda_i), \quad \left\{ \begin{array}{l} \lambda_i = \mathbf{x}_i\mathbf{b}, \\ \lambda_i = \exp(\mathbf{x}_i'\mathbf{b}) \end{array} \right. , \; i = 1, \ldots, m \tag{3.71}$$

mit

$$\mathbb{E}(y_i|\mathbf{x}_I) = \lambda_i, \quad Var(y_i|\mathbf{x}_i). \tag{3.72}$$

Für die Dichtefunktion erhält man

$$f(y_i|\mathbf{b}) = \exp(-\lambda_i)\frac{\lambda_i^{y_i}}{y_i!}, \quad \mathbb{E}(y_I|\mathbf{x}_i) = \lambda_i \tag{3.73}$$

Bei der Maximum-Likelihood-Schätzung der Parameter wird auf diese Dichte zurückgegriffen.

Die negative Binomialregression Der Ausdruck ist insofern ein wenig irreführend, weil die Ausgangsverteilung ja die Poisson-Verteilung ist und sich die negative Binomialverteilung erst durch eine Modifikation der Poisson-Verteilung ergibt. In der Tat kann K (Anzahl der SIs etc) die Werte $K = 0, 1, 3, \ldots$ annehmen, d.h. im Prinzip kann K beliebig große Werte annehmen. Die Annahme dabei ist, dass die einzelnen Beobachtungen stochastisch unabhängig sind. Die Poisson-Verteilung kann aus der Binomialverteilung hergeleitet werden:

$$\mathcal{B}(K, n, p) = P(K = k|p, n) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.74}$$

mit

$$\mathbb{E}(K) = np, \quad Var(K) = np(1-p) = npq, \; q = 1-p \tag{3.75}$$

Hier liegt also die Gesamtzahl n der Beobachtungen fest. Die Berechung von $\binom{n}{k}$ wird für große Werte von n sehr länglich und ohne Computer extrem zeitintensiv, so dass der französische Mathematiker und Physiker S. D. Poisson[16] eine Approximation suchte und fand (die Herleitung ist zu länglich, um sie hier zu reproduzieren), unter den Annahmen (i) die Beobachtungen sind stochastisch unabhängig, (ii) p ist "klein" und n ist "groß" und

$$pn \to \lambda \text{ für } p \to 0, \; n \to \infty \tag{3.76}$$

folgt

$$\mathcal{B}(K, n, p) \to P(K = k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$$

mit $\mathbb{E}(K) = Var(K) = \lambda$. (3.76) legt die Vermutung nahe, die Approximation $P(K = k|\lambda)$ für $\mathcal{B}(K, n, p)$ gelte nur für wirklich kleine Werte von p, also etwa $p = .1$ oder $p = .2$. Der Punkt ist, dass man eigentlich nur $\lambda < \infty$ annimmt, d.h. p soll "klein" im Vergleich zu n sein, und der Wert von n kann *beliebig* groß sein. $np \to \lambda$ heißt ja nur, dass für hinreichend großen Wert von n die Beziehung $p \approx \lambda/n \leq 1$ gelten soll, und diese Beziehung kann auch für $p = .99$ gelten, man muß nur einen entsprechenden Wert von λ wählen. Praktisch bedeutet die Approximation, dass man zuläßt, dass die zufällige Veränderliche K beliebig groß werden kann, das in Frage stehende Ereignis also nie eintritt. Abgesehen davon kann man die Poisson-Verteilung

[16] Siméon Denis Poisson (1781 – 1840)

auch einfach annehmen, ohne sie als Approximation an die Binomialverteilung anzusehen. In der Tat sind viele Häufigkeiten, z.B. die von Unfällen, zumindest in erster Näherung Poisson-verteilt.

Man kann eine Reihe von Gründen für die Beobachtung $Var(K) > \mathbb{E}(K)$ (bzw. für die entsprechenden Schätzungen dieser Parameter) anführen, auf die hier nicht ausführlich eingegangen werden kann. (**??**) ist eine ad-hoc-Annahme mit α als einem sogenannten Jesus-Parameter: er hilft, zumindest manchmal, aber man weiß nicht was er eigentlich bedeutet. Die implizite Annahme im Possion-Modell ist nun, dass λ für alle Personen, die befragt werden, denselben Wert hat. Diese Annahme ist sicher nicht berechtigt, so dass man für die i-te Person einen Parameter λ_i ansetzen müsste. Das kann man machen, indem man z.B.

$$\lambda_i = \lambda\varepsilon_i, \quad i = 1, \ldots, m \tag{3.77}$$

setzt. λ wird sozusagen individuell modifiziert, ε_i wird auch als *Heterogenitätsterm* bezeichnet, weil er eben die Heterogenität der Personen bezüglich des Parameters λ reflektiert. Die Verteilung der Häufigkeiten (SI, suicide attempts, etc) für die i-te Person ist dann

$$P(K = k_i|\lambda, \varepsilon_i) = e^{-\lambda_i}\frac{\lambda_i^{k_i}}{k_i!} = e^{-\lambda\varepsilon_i}\frac{(\lambda\varepsilon_i)^{k_i}}{k_i!}, \quad k_i = 0, 1, 2, \ldots \tag{3.78}$$

Es ist aber klar, dass man die individuellen ε_i und damit die λ_i nicht schätzen kann, – sie schätzen zu wollen, hieße, für jede Person einen "freien Parameter" (das ist ein Parameter, dessen Wert man nicht kennt und der eben aus den Daten geschätzt werden muß) schätzen zu müssen, plus etwaiger weiterer freier Parameter, – man hat dann mehr freie Parameter als Datenpunkte). Ein Ausweg aus dieser Problemlage ist nun, eine Wahrscheinlichkeitsverteilung für die λ_i anzunehmen. Ebenso, wie man annimmt, dass der Intelligenzquotient in der Bevölkerung normalverteilt ist, kann man annehmen, dass die λ-Werte in der Bevölkerung irgendwie verteilt sind. Man könnte etwa postulieren, dass auch die λ-Werte in der Population normalverteilt sind. Diese Annahme (die in Psychologen-, Biologen- und Medizinerkreisen stets locker von der Hand geht) erweist sich aber zumindest aus mathematischer Sicht als nicht sehr klug, wie die folgenden Betrachtungen zeigen.

Denn man ist ja an der Verteilung der k_i-Werte interessiert, die von exogenen Variablen, die als unabhängige Variablen in die Verteilung eingehen (das sind die Prädiktoren $X_1, \ldots, X_n$) abhängen, und deren Effekt will man ja erkunden. Da die Wahl einer Person (ein Feuerwehrmann in der Untersuchung von Boffa et al) stichprobentheoretisch ein zufälliges Ereignis ist, ist auch der zugehörige λ_i-Wert zufällig. Am Ende fasst man aber die beobachteten Häufigkeiten zusammen, d.h. man betrachtet zufällige Variable, deren Verteilung sich zusammensetzt (i) aus einer Verteilung für $K = k_i$ und (ii) einer Verteilung für den Parameter λ. Um zu sehen, wie diese Zusammensetzung aufgebaut ist, muß man beachten, dass die Verteilung $P(K = k_i|\lambda, \varepsilon_i)$

in (3.78) eine bedingte Wahrscheinlichkeit ist: $K = k_i$ hat eine bestimmte Wahrscheinlichkeit *unter der Bedingung*, dass die i-te Person den Parameter λ_i hat. Allgemein gilt für eine bedingte Wahrscheinlichkeit

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A|B)P(B), \tag{3.79}$$

d.h. die Wahrscheinlichkeit, dass die zufälligen Ereignisse A und B gemeinsam auftreten, ist gleich dem Produkt $P(A|B)P(B)$. Demnach ist die Wahrscheinlichkeit, dass ein Wert $K = k_i$ auftritt *und* die i-te Person den Parameter λ_i hat, unter der Bedingung, dass die unabhängigen Prädiktorvariablen durch den Vektor $\mathbf{x}_i$ gegeben sind, durch[17]

$$P(K = k_i|\mathbf{x}_i \cap \lambda_i) = f(K = k_i|\mathbf{x}_i, \lambda_i)g(\lambda_i) \tag{3.80}$$

gegeben. Hier ist f die Wahrscheinlichkeit für $K = k_i$, unter der Bedingung, dass der individuelle Parameter durch λ_i gegeben ist und die Bedingungen durch $\mathbf{x}_i$ gegeben sind, – f entspricht der Wahrscheinlichkeit $P(A|B)$ – , und g ist die Warscheinlichkeit, dass die i-te Person den Parameter λ_i hat. $g(\lambda_i)$ entspricht der Wahrscheinlichkeit $P(B)$.

Nun gilt es, zwei Aspekte zu betrachten: (1) an den individuellen λ_i-Werten ist man eigentlich gar nicht interessiert, man möchte wissen, was die Bedingung $\mathbf{x}_i$ sozusagen im Mittel bewirkt. (2) Da die Population aus einzelnen Personen besteht, hat man es streng genommen mit einer diskreten Verteilung der λ_i- Werte zu tun. Dies ist ein mathematisch sehr unangenehmer Sachverhalt, denn man muß nun, um die Wahrscheinlichkeit zu berechnen, dass λ_i in einem bestimmten Intervall liegt, über die Wahrscheinlichkeiten für die einzelnen λ_i-Werte summieren, – das ist ein sehr aufwändiger und vielfach praktisch gar nicht möglicher Prozess. Deswegen macht man von der Möglichkeit Gebrauch, die diskrete Verteilung durch eine stetige Verteilung zu approximieren, wie schon bei der Verteilung der IQs (Normalverteilung), die ja letztlich auch eine diskrete Menge bilden. Wie in der Wahrscheinlichkeitstheorie gezeigt wird (darauf kann ich hier icht im Detail eingehen) ist die Approximation durch eine stetige Verteilung nachgerade beliebig genau, vereinfacht aber die Mathematik beträchtlich. Deshalb geht man für g auch zu einer stetigen Verteilung über. Man hat im diskreten Fall den Ausdruck (Satz von der Totalen Wahrscheinlichkeit → Statistik I oder II), wenn A unter den exhaustiven und disjunkten Bedingungen B_1, $B_2, \ldots B_p$ auftreten kann,

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_p)P(B_p) = \sum_{j=1}^{p} P(AB_j)P(B_j). \tag{3.81}$$

(vergl. (3.79)). Wir sind interessiert an $f(K = k_i|\mathbf{x}_i)$; diese Wahrscheinlichkeit entspricht $P(A)$ in der Gleichung (3.81). Der Punkt bei dieser Gleichung

[17] Das Zeichen ∩ steht für "und".

ist, dass man die einzelnen Bedingungen B_j gewissermaßen los wird, indem man über sie summiert. Wenn man statt der diskreten Verteilung zu stetigen Verteilungen übergeht, entspricht der Summe ein Integral (auf die Details des Übergangs gehe ich nicht ein, die kann man ersteinmal glauben, da sie zur Standardmathematik gehören). Demnach hat man für $f(K = k_i|\mathbf{x}_i)$ dann

$$f(K = k_i|\mathbf{x}_i) = \int_0^\infty f(K = k_i|\mathbf{x}_i, \lambda)g(\lambda)d\lambda) \tag{3.82}$$

Hier erscheint λ nicht mehr mit dem Index i, weil nun λ alle Werte aus dem Intervall $[0, \infty)$ annehmen kann, und $f(K = k_i|\mathbf{x}_i, \lambda)$ ist durch (3.78) gegeben. Natürlich ist die Wahrscheinlichkeit, dass ein λ-Wert unendlich ist, gleich Null, aber die Frage nach den möglichen Werten der Wahrscheinlichkeit für λ wird durch die Annahme einer bestimmten Verteilung g automatisch mitgeklärt.

Die Frage ist nun, *welche* Verteilung man annehmen soll. Wie oben angedeutet könnte man die Standardannahme der Normalverteilung machen, aber dafür gibt es einerseits keinen logisch zwingenden Grund, und andererseits erschwert sie das Leben, wie der Versuch, das Integral in (3.82) in möglichst geschlossener Form zu erhalten, zeigt. Es gibt Verteilungen, die in Abhängigkeit von den Parameterwerten eine Form annehmen können, die der einer Normalverteilung sehr nahe kommen, die aber auch stark von der Normalverteilung abweichen können. Man ist also flexibler, wenn man eine andere als die Normalverteilung wählt.

Bewährt hat sich im gegebenen Zusammenhang die *Gamma-Verteilung*. In der Psychologie wird sie oft verwendet, um die Verteilung von Reaktionszeiten zu beschreiben: Reaktionszeiten sind oft die Summe von Teilzeiten (im einfachsten Fall: erst ein Signal verarbeiten, auf das die Vp reagieren soll, dann den neuronalen Befehl an den Finger geben, der die jeweilige Taste drücken soll, dann die Ausführung der Bewegung (= Reaktion); jede dieser Zeiten ist oft exponentialverteilt, und die Summe voneinander unabhängiger Zeiten ist dann Gamma-verteilt). Bei der Gamma-Verteilung sind kleinere Werte der jeweiligen zufälligen Veränderlichen wahrscheinlicher als größere, die Verteilung ist also asymmetrisch (links-steil). Man muß ein derartiges Modell für die Gamma-Verteilung aber nicht annehmen, um diese Verteilung zu wählen, man kommt auch auf anderen Wegen auf die Gamma-Verteilung. Im hier gegebenen Zusammenhang ist die Entscheidung für die Gamma-Verteilung eher eine ad-hoc-Annahme, die wegen ihrer Form und ihrer mathematischen Eigenschaften naheliegt: sie ist wohl die einzige Verteilung, die einen geschlossenen Ausdruck für das Integral (3.82) zuläßt. Die Gamma-Verteilung ist durch

$$g(\lambda) = \frac{\theta^p}{\Gamma(p)}\lambda^{p-1}e^{-\theta\lambda},\ \lambda > 0,\ g(\lambda) = 0,\ \lambda \leq g(\lambda) = 0. \tag{3.83}$$

definiert. λ ist hier die zufällige Veränderliche, und p und θ sind die Pa-

rameter der Verteilung. Hier wird ein weiterer Vorteil des Ansatzes (3.82) deutlich: man muß nicht nur nicht eine diskrete Summe für individuelle, unbekannte und daher zu schätzende λ_i-Werte bestimmen, sondern hat nur zwei freie Parameter, nämlich p und θ. $\Gamma(p)$ ist die Gamma-Funktion

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx, \tag{3.84}$$

der Quotient $\frac{\theta^p}{\Gamma(p)}$ in (3.83) ist ein Normalisierungsfaktor (das Integral über $g(\lambda)$ muß gleich 1 sein, der Faktor hat die gleiche Bedeutung wie der Faktor $1/\sigma\sqrt{2\pi}$ bei der Normalverteilung). Der Erwartungswert und die Varianz der Gamma-Verteilung sind durch

$$\mathbb{E}(\lambda) = \frac{p}{\theta}, \quad Var(\lambda) = \frac{p}{\theta^2} \tag{3.85}$$

gegeben. Setzt man nun den Ausdruck für g in das Integral (3.82) ein, so findet man (Details werden hier ausgelassen, da länglich und nicht weiter instruktiv)

$$f(k_i|\mathbf{x}_i) = \frac{\Gamma(k_i + \theta)}{k_i! \Gamma(\theta)} \left(\frac{\theta}{\theta + \lambda_i}\right)^\theta \left(\frac{\lambda_i}{\theta + \lambda_i}\right)^{k_i}, \tag{3.86}$$

wobei $\mathbb{E}(\lambda) = 1$ gesetzt wurde; nach (3.85) gilt allgemein $\mathbb{E}(\lambda) = p/\theta$, so dass $\mathbb{E}(\lambda) = 1$ den Wert von p festlegt, nämlich $p = \theta$. Damit hat man $f(k_i|\mathbf{x}_i)$ bestimmt: die Verteilung ist die negative Binomialverteilung. Der Erwartungswert für k_i, die Häufigkeit von Reaktionen (SIs etc) ergibt sich genmäß

$$\mathbb{E}(k_i|\mathbf{x}_i) = e^{\mathbf{x}_i' \mathbf{b}}. \tag{3.87}$$

Es handelt sich um einen bedingten Erwartungswert, also um die Erwartung unter der Bedingung, dass $\mathbf{x}_i$ vorliegt). Die bedingte Varianz ist

$$Var(k_i|\mathbf{x}_i) = \lambda_i(1 + \lambda_i/\theta) > \mathbb{E}(k_i|\mathbf{x}_i), \tag{3.88}$$

d.h. der Überdispersion der eigentlich Poisson-verteilten Häufigkeiten wird durch die Betrachtung der bedingten Verteilungen Rechnung getragen, die sich wiederum aus der Heterogenität der individuellen Parameterwerte ergibt.

Das Modell wird oft *reparametrisiert*: es wird $\alpha = 1/\theta$ gesetzt. (3.88) wird dann zu

$$f(k_i|\mathbf{x}_i) = \frac{\Gamma(k_i + \alpha^{-1}}{k_i! \Gamma(\alpha^{-1}} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i}\right)^{k_i} \tag{3.89}$$

Für die Varianz erhält man dann aus (3.88) den Ausdruck

$$Var(k_i|\mathbf{x}_i) = \lambda_i + \alpha\lambda_i^2. \tag{3.90}$$

Dieses Modell wird unter dem Namen *NB2-Modell* geführt (für den Fall, dass dieses Modell in Artikeln zu Ihrem Thema genannt wird).

3.5 Die Schätzung der Parameter

3.5.1 Grundsätzliches zur Maximum-Likelihood-Methode

Es wird angenommen, dass die zufällige Veränderliche X die die Wahrscheinlichkeitsdichte $f(X)$ hat; f kann die Normalverteilung sein, oder die Binomialverteilung, falls $X = 0, 1, \ldots, n$, etc. f ist durch Parameter spezifiziert; im Falle der Normalverteilung sind dies der Erwartungswert $\mu = \mathbb{E}(X)$ und die Varianz $\sigma^2 = Var(X)$, im Falle der Binomialverteilung ist es der Parameter p. Man beobachtet eine Stichprobe von Werten $x_1, x_2, \ldots, x_n$ von X und fragt nach der Wahrscheinlichkeit, mit der diese Stichprobe gewonnen wird; diese Wahrscheinlichkeit ist die *Likelihood* L der Stichprobe. Hat man die x_i unabhängig voneinander bestimmt, so ist L durch das Produkt der Wahrscheinlichkeiten für die einzelnen x_i gegeben, also

$$L(x_1, \ldots, x_n) = L(\mathbf{x}) = \prod_{i=1}^{n} P(x_i), \quad \mathbf{x} = (x_1, \ldots, x_n)'. \tag{3.91}$$

Natürlich hängen die Werte $p(x_i)$ von den Parametern der Verteilung der x_i-Werte ab, so dass man L auch als bedingte Wahrscheinlichikeit auffassen kann:

$$L(\mathbf{x}|\pi_1, \ldots, \pi_k) = \prod_{i=1}^{n} P(x_i|\pi_1, \ldots, \pi_k); \tag{3.92}$$

hierbei sind die $\pi_1, \ldots, \pi_k$ die Parameter der Verteilung. Bei der Normalverteilung hat masn etwa $\pi_1 = \mu$, $\pi_2 = \sigma^2$. Im allgemeinen beobachtet man nun solche Werte $\mathbf{x}$, die am wahrscheinlichsten sind, - dies folgt aus dem Begriff der Wahrscheinlichkeit. Also werden die $\pi_1, \ldots, \pi_k$ Werte haben derart, dass die Likelihood $L(\mathbf{x}|\pi_1, \ldots, \pi_k)$ maximal ist. Damit hat man die Methode der Parameterschätzung auch schon gefunden: man maximiert L als Funktion der Parameter $\pi_1, \ldots, \pi_k$.

Beispiel 3.6 X sei binomialverteilt, so dass

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.93}$$

gilt. $\pi_1 = p$ ist der Parameter der Verteilung. Es sei $x_1, \ldots, x_n$ die beobachtete Folge von Resultaten, wobei $x_i = 1$ oder $x_i = 0$. Da die x_i unabhängig sind, ist die Likelihood-Funktion durch

$$L(x_1, \ldots, x_n|p) = \prod_{i=1}^{n} P(X = x_i|p) \tag{3.94}$$

gegeben. Nun ist $P(X = 1|p) = p$, $P(X = 0|p) = 1 - p$. Es seien gerade k Werte der x_i gleich 1, und $n - k$ Werte seien gleich 0. Damit ergibt sich

$$L(x_1, \ldots, x_n|p) = p^k (1-p)^{n-k}. \tag{3.95}$$

Gesucht ist der p-Wert, für den L maximal wird. Also bildet man dL/dp und bestimmt den Wert $\hat{p}$, für den diese Ableitung gleich Null wird. Man findet

$$\frac{dL}{dp} = kp^{k-1}(1-p)^{n-k} - (n-k)p^k(1-p)^{n-k-1},$$

so dass

$$k\hat{p}^{k-1}(1-\hat{p})^{n-k} = (n-k)\hat{p}^k(1-\hat{p})^{n-k-1}.$$

Löst man diese Gleichung nach $\hat{p}$ auf, so erhält man

$$\hat{p} = \frac{k}{n}. \tag{3.96}$$

Die relative Häufigkeit des Ereignisses $x_i = 1$ ist also gerade die Maximum-Likelihood-Schätzung $\hat{p}$ für p. □

Beispiel 3.7 Es wird die Anzahl der Unfälle in einer Fabrik betrachtet. Für einen festen Zeitraum der Dauer Δt ist die Verteilung dieser Anzahl in guter Näherung durch

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{3.97}$$

gegeben; dabei wird die Annahme zugrundegelegt, dass die Unfälle unabhängig voneinander geschehen und die Wahrscheinlichkeit eines Unfalls relativ klein ist. Man findet, dass $\mathbb{E}(X) = \mu = \lambda$ und $Var(X) = \sigma^2 = \lambda$. In m aufeinander folgenden Abschnitten der Länge Δt findet man die Anzahlen $x_1 = n_1$, $x_2 = n_2, \ldots, x_m = n$. Man muß die Funktion

$$L(n_1, \ldots, n_m|\lambda) = \prod_{k=1}^{m}\left(e^{\lambda}\frac{\lambda^{n_k}}{n_k!}\right). \tag{3.98}$$

Die Differentiation nach λ ist ein wenig umständlich durchzuführen. Man kann nun von der Tatsache Gebrauch machen, dass der Logarithmus von L eine monotone Funktion von L ist; wird L maximal, so wird auch $\log L$ maximal. Man findet dann

$$\log L(n_1, \ldots, n_m|\lambda) = \sum_{k=1}^{m}(\lambda + n_k \log\lambda - \log n_k!).$$

Man findet dann

$$\frac{dL}{d\lambda} = \sum_{k=1}^{m}(1 + \frac{n_k}{\lambda}) = m + \frac{1}{\lambda}\sum_{k=1}^{m} n_k.$$

Für $\lambda = \hat{\lambda}$ wird dieser Ausdruck gleich Null, und man erhält

$$\hat{\lambda} = \frac{1}{m}\sum_{k=1}^{m} n_k. \tag{3.99}$$

□

3.5.2 Anwendung: logistische Regression

Obwohl x im Prinzip stetig variieren kann, wird man im allgemeinen die Wirkung von nur endlich vielen Werten $x_1, \cdots, x_k, \cdots, x_r$ untersuchen. Die abhängige Variable y nehme aber nur die zwei Werte 1 oder 0 an. Man hat dann eine Datentabelle von der Form der Tabelle 8. Es werde nun angenom-

Tabelle 8: Logistische Regression, r Prädiktorwerte

	Abhängige Variable		
Unabh. Variable	$y = 1$	$y = 0$	Σ
x_1	n_{11}	n_{12}	n_{1+}
x_2	n_{21}	n_{22}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	n_{k+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	n_{r+}
Σ	n_{+1}	n_{+2}	$n_{++} = n$

men, dass

$$p(x_k) = p(y = 1|x_k) = \frac{\exp(Ax_k + B)}{1 + \exp(Ax_k + B)}$$

gilt. Die Parameter A und B sind unbekannt und müssen aus den Daten, d.h. aus den n_{ks}, $s = 1, 2$ geschätzt werden. Dazu wird die Maximum-Likelihood-Methode angesetzt. Zur Vereinfachung werde

$$p_k = p(x_k), \quad n_k = n_{k1}, \quad N_k = n_{k+}$$

gesetzt. Sicherlich ist dann

$$n_{k2} = N_k - n_k$$

Die Likelihood-Funktion ist dann durch

$$L = \prod_{k=1}^{r} p_k^{n_k} (1 - p_k)^{N_k - n_k} \tag{3.100}$$

gegeben, und daraus erhält man sofort die Log-Likelihood-Funktion

$$\log L = \sum_k (n_k \log p_k + (N_k - n_k)(1 - p_k)) \tag{3.101}$$

Um die Schätzungen für A und B zu gewinnen, müssen die partiellen Ableitungen von $\log L$ bezüglich A und B hergeleitet werden. Dazu wird der folgende Satz bewiesen:

Satz 3.1 *Es gelte*

$$P(y=1|x_k) = p_k = \frac{\exp(\phi_k(x_k))}{1+\exp(\phi_k(x_k))}, \quad \phi_k(x_k) = Ax_k + B \tag{3.102}$$

Weiter sei entweder $\theta = A$ oder $\theta = B$; ϕ_k kann dann jeweils als Funktion von θ aufgefaßt werden. Dann gilt

$$\frac{\partial \log L}{\partial \theta} = \sum_k (n_k - N_k p_k) \frac{\partial \phi_k}{\partial \theta}. \tag{3.103}$$

Beweis: Aus (3.101) erhält man sofort

$$\frac{\partial \log L}{\partial \theta} = \sum_k \left(n_k \frac{1}{p_k} \frac{\partial p_k}{\partial \theta} - (N_k - n_k) \frac{1}{1-p_k} \frac{\partial p_k}{\partial \theta} \right) = \sum_k \left(\frac{n_k}{p_k} - \frac{N_k - n_k}{1-p_k} \right) \frac{\partial p_k}{\partial \theta}$$

Nun ist

$$\frac{\partial p_k}{\partial \theta} = \frac{dp_k}{d\theta} \frac{\partial \phi_k}{\partial \phi_k},$$

und

$$\frac{dp_k}{d\phi_k} = \frac{e^{\phi_k}(1+e^{\phi_k}) - e^{2\phi_k}}{(1+\exp(\phi_k))^2} = \frac{e^{\phi_k}}{(1+\exp(\phi_k))^2} = \frac{p_k}{1+\exp(\phi_k)}$$

Deshalb erhält man

$$\begin{aligned} \frac{\partial \log L}{\partial \theta} &= \sum_k \frac{n_k - n_k p_k - N_k p_k + n_k p_k}{p_k(1-p_k)} \frac{\partial p_k}{\partial \theta} = \\ &= \sum_k \frac{n_k - N_k p_k}{p_k(1-p_k)} \frac{p_k}{(1+\exp(\phi_k))} \frac{\partial \phi_k}{\partial \theta} = \\ &= \sum_k \frac{n_k - N_k p_k}{(1-p_k)(1+\exp(\phi_k))} \frac{\partial \phi_k}{\partial \theta} \end{aligned}$$

Aber es ist

$$(1-p_k)(1+e^{\phi_k}) = \frac{1}{1+\exp(\phi_k)}(1+e^{\phi_k}) = 1,$$

und damit folgt (3.103). □

Die ML-Schätzungen für A und B erhält man aus (3.103), indem man die Gleichungen

$$\left. \frac{\partial \log L}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0 \tag{3.104}$$

bildet.

3.5.3 Der allgemeine Fall

Nach Satz 3.1, Gleichung (3.103) gilt für die Likelihood-Funktion

$$\frac{\partial \log L}{\partial \theta} = \sum_k (n_k - N_k p_k) \frac{\partial \phi_k}{\partial \theta},$$

Da $\theta = A$ oder $\theta = b$ hat man zwei Gleichungen; gleich Null gesetzt erhält man

$$\sum_k n_k \frac{\partial \phi_k}{\partial \theta} = \sum_k N_k p_k \frac{\partial \phi_k}{\partial \theta} \tag{3.105}$$

Wegen

$$\frac{\partial \phi_k}{\partial \theta} = \begin{cases} x_k, & \theta = A \\ 1, & \theta = B \end{cases} \tag{3.106}$$

erhält man für (3.105) in ausführlicher Schreibweise

$$\sum_k n_k x_k = \sum_k N_k \hat{p}_k x_k \tag{3.107}$$

$$\sum_k n_k = \sum_k N_k \hat{p}_k \tag{3.108}$$

Hier ist $\hat{p}_k$ statt p_k geschrieben worden, weil die Gleichungen für die Schätzungen $\hat{A}$ und $\hat{B}$ gelten. Für $\hat{\pi}_k$ setzt man nun den in (3.102) gegebenen Ausdruck ein, der dann nach $\hat{A}$ und $\hat{B}$ aufgelöst werden muß. Da die Wahrscheinlichkeiten $\hat{\pi}_k$ in nichtlinearer Weise von $\hat{A}$ und $\hat{B}$ abhängen, kann man die Lösungen für $\hat{A}$ und $\hat{B}$ im allgemeinen Fall – wenn also x nicht mehr auf nur zwei Werte, 0 und 1, beschränkt ist – nicht mehr in geschlossener Form hinschreiben. Die Gleichungen müssen numerisch gelöst werden; hierzu kann z.B. die Newton-Raphson-Methode herangezogen werden. Computerprogramme zur logistischen Regression enthalten diesen oder einen ähnlichen Algorithmus.

3.5.4 Spezialfall: dichotome Prädiktoren

Nach (3.103) muß zunächst $\partial \phi_k / \partial \theta$ bestimmt werden. Für $\theta = A$ ist wegen $\phi_k = A x_k + B$

$$\frac{\partial \phi_k}{\partial \theta} = \frac{\partial \phi_k}{\partial A} = x_k$$

und da x_k nur die Werte 0 oder 1 annehmen kann ist

$$\frac{\partial \phi_k}{\partial A} = \begin{cases} 0, & x_k = 0 \\ 1, & x_k = 1 \end{cases}$$

Für $\theta = B$ findet man

$$\frac{\partial \phi_k}{\partial B} = 1$$

für beide x-Werte. Dementsprechend liefert (3.103) die Gleichungen

$$\left.\frac{\partial \log L}{\partial A}\right|_{A=\hat{A}} = n_1 - \frac{N_1 \exp(\hat{A}+\hat{B})}{1+\exp(\hat{A}+\hat{B})} = 0 \tag{3.109}$$

$$\left.\frac{\partial \log L}{\partial B}\right|_{B=\hat{B}} = n_1 - \frac{N_1 \exp(\hat{A}+\hat{B})}{1+\exp(\hat{A}+\hat{B})} + n_2 - \frac{N_2 \exp(\hat{B})}{1+\exp(\hat{B})} = 0 \tag{3.110}$$

Die Gleichung (3.110) vereinfacht sich aber wegen (3.109) sofort zu

$$\left.\frac{\partial \log L}{\partial B}\right|_{B=\hat{B}} = n_2 - \frac{N_2 \exp(\hat{B})}{1+\exp(\hat{B}))} = 0$$

woraus sofort

$$\frac{n_2}{N_2 - n_2} = \exp(\hat{B}) \tag{3.111}$$

folgt.

Aus (3.109) folgt

$$n_1(1+\exp(\hat{A}+\hat{B})) = N_1 \exp(\hat{A}+\hat{B})$$

so dass

$$\frac{n_1}{N_1 - n_1} = \exp(\hat{A}+\hat{B}),$$

und wegen (3.111) hat man

$$\frac{n_1}{N_1 - n_1} = \frac{n_2}{N_2 - n_2}\exp(\hat{A})$$

so dass

$$\frac{n_1}{N_1 - n_1}\frac{N_2 - n_2}{n_2} = \exp(\hat{A}) \tag{3.112}$$

Nun war $n_k = n_{k1}$ und $N_k - n_k = n_{k2}$. Dementsprechend ist (vergl. (3.54))

$$\frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{\Omega_1}{\Omega_2} = \Theta = \exp(\hat{A}), \quad \text{bzw. } \log\Theta = \hat{A} \tag{3.113}$$

und (vergl. (3.52))

$$\frac{n_{21}}{n_{22}} = \Omega_2 = \exp(\hat{B}), \quad \text{bzw. } \log\Omega_2 = \hat{B} \tag{3.114}$$

4 Anhang: Beweise:

4.1 Satz 2.1

Beweis: Gemäß der Methode der kleinsten Quadrate muß die Funktion

$$Q(b_0, b_1, ..., b_p) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2$$

der $k+1$ Variablen $b_0, b_1, ..., b_p$ minimiert werden. Allgemein soll gelten

$$\frac{\partial Q(b_0, b_1, ..., b_p)}{\partial b_j} = 2 \sum_{i=1}^{n} e_i \, \frac{\partial e_i}{\partial b_j}|_{b_j = \hat{b}_j} = 0$$

Für $j = 0$ insbesondere erhält man

$$\frac{\partial e_i}{\partial b_0} = \frac{\partial (y_i - b_0 - b_1 X_{1i} - \cdots - b_p X_{ki})}{\partial b_0} = -1$$

so dass

$$\frac{\partial Q}{\partial b_0}|_{b_0 = \hat{b}_0} = -2 \sum_{i=1}^{n} e_i = 0$$

Hieraus folgt sofort

$$\sum_{i=1}^{n} e_i = 0$$

d.h. der mittlere Fehler $\bar{e} = \sum_{i=1}^{n} e_i / n = 0$ ist gleich Null, da ja $-2 \neq 0$.

Für $j > 0$ ergibt sich

$$\frac{\partial e_i}{\partial b_j}|_{b_j = \hat{b}_j} = \frac{\partial Q(y_i - b_1 X_{1i} - \cdots - b_p X_{ki})}{\partial b_j} = -X_{ji}.$$

Mithin ist

$$\frac{\partial Q}{\partial b_j}|_{b_j = \hat{b}_j} = -2 (\sum_{i=1}^{n} e_i X_{ji})|_{b_j = \hat{b}_j} = 0, \tag{4.1}$$

für $j = 1, ..., k$. Substituiert man hierin $e_i = Y_i - \hat{b}_1 X_{1i} - \hat{\mathbf{b}}_2 X_{2i} - \cdots - \hat{b}_p X_{ki}$, so erhält man das Gleichungssystem (2.8).

□

4.2 Satz 2.2

Beweis: Man betrachtet wieder die Funktion

$$Q(\beta_1, ..., \beta_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (z_{0i} - \beta_1 \, z_{1i} - \cdots - \beta_k \, z_{ki})^2$$

Die Schätzungen $\hat{\beta}_j$ ergeben sich als Lösungen der k Gleichungen

$$\frac{\partial Q}{\partial \beta_j}|_{\beta_j=\hat{\beta}_j} = 2\sum_{i=1}^{n} \epsilon_i \frac{\partial \epsilon_i}{\beta_j}|_{\beta_j=\hat{\beta}_j} = 0$$

Es ist aber

$$\frac{\partial \epsilon_i}{\partial \beta_j} = -z_{ji},$$

und in die Gleichungen eingesetzt ergibt sich das Gleichungssystem

$$\begin{aligned}
\sum_{i=1}^{n} z_{0i}\, z_{1i} &= \hat{\beta}_1 \sum_{i=1}^{n} z_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} z_{1i}\, z_{2i} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} z_{1i}\, z_{ki} \\
\sum_{i=1}^{n} z_{0i}\, z_{2i} &= \hat{\beta}_1 \sum_{i=1}^{n} z_{2i} z_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} z_{2i}^2 + \cdots + \hat{\beta}_k \sum_{i=1}^{n} z_{1i}\, z_{ki} \\
&\vdots \\
\sum_{i=1}^{n} z_{0i}\, z_{ki} &= \hat{\beta}_1 \sum_{i=1}^{n} z_{1i}\, z_{ki} + \hat{\beta}_2 \sum_{i=1}^{n} z_{1i}\, z_{ki} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} z_{ki}^2 \qquad (4.2)
\end{aligned}$$

Es ist aber $\sum_i z_{0i} z_{ji}/n = r_{0j} = r_{yj}$, $\sum_i z_{si} z_{ji}/n = r_{sj}$ und $\sum_i z_{ji}^2 = 1$. Dividiert man das Gleichungssystem (4.2) durch n, so erhält man die Behauptung des Satzes.

□

4.3 Satz 2.3

Es ist

$$\begin{aligned}
(\mathbf{y} - X_0\vec{b})'(\mathbf{y} - X_0\vec{b}) &= \vec{y}'\mathbf{y} - \mathbf{y}'X_0\vec{b} - (X_0\vec{b})'\mathbf{y} + (X_0\vec{b})X_0\vec{b} \\
&= \mathbf{y}'\mathbf{y} - \mathbf{y}'X_0\vec{b} - \vec{b}'X_0'\mathbf{y} + \vec{b}'X_0'X_0\vec{b}.
\end{aligned}$$

Nun ist aber $X_0\vec{b}$ ein Vektor, so dass $\mathbf{y}'X_0\vec{b} = \vec{b}'X_0'\mathbf{y}$ ein Skalarprodukt ist; folglich ist

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - X_0\vec{b})'(\mathbf{y} - X_0\vec{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'X_0\vec{b} + \vec{b}'X_0'X_0\vec{b} \qquad (4.3)$$

Die partielle Ableitung nach b_j liefert dann (Anwendung der Produktregel)

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial b_j} = -2\vec{y}'X_0\vec{\epsilon}_j + \vec{\epsilon}_j{}'X_0'X_0\vec{b} + \vec{b}'X_0'X_0\vec{\epsilon}_j, \quad j = 1, \ldots, k \qquad (4.4)$$

Fasst man diese Gleichungen zusammen, so bilden die $\vec{\epsilon}_j$ die Einheitsmatrix I und man erhält die Matrixgleichung

$$-2X_0'\mathbf{y} + 2X_0'X_0\hat{\mathbf{b}} = 0, \qquad (4.5)$$

wobei $\hat{\mathbf{b}}$ der Vektor ist, für den der Ausdruck links gleich Null ist. Dieser Vektor minimalisiert das Skalarprodukt $\mathbf{e}'\mathbf{e}$. Es folgt $\mathbf{y}'X_0 = X_0'X_0\hat{\mathbf{b}}$. Läßt sich die Inverse $(X_0'X_0)^{-1}$ bilden, so folgt sofort die Gleichung

$$\hat{\mathbf{b}} = (X_0'X_0)^{-1}X_0'\mathbf{y}. \tag{4.6}$$

Standardisiert man die x_{ij}-Werte, so geht die Matrix X in die Matrix Z über, und $\vec{b}$ geht in den Vektor $\vec{\beta}$ der β-Gewichte über. Dabei verschwindet die additive Konstante b_0, und (4.8) wird zu

$$\vec{\hat{\beta}} = (Z'Z)^{-1}Z'\vec{Z}_y, \tag{4.7}$$

wobei $\vec{Z}_y$ der Vektor der standardisierten y-Werte ist. □ yyy

$$\hat{\mathbf{b}} = (X_0'X_0)^{-1}X_0'\mathbf{y}. \tag{4.8}$$

Standardisiert man die x_{ij}-Werte, so geht die Matrix X in die Matrix Z über, und $\vec{b}$ geht in den Vektor $\vec{\beta}$ der β-Gewichte über. Dabei verschwindet die additive Konstante b_0, und (4.8) wird zu

$$\vec{\hat{\beta}} = (Z'Z)^{-1}Z'\vec{Z}_y, \tag{4.9}$$

wobei $\vec{Z}_y$ der Vektor der standardisierten y-Werte ist.

4.4 Satz 2.8

Beweis: Es sei $y_i = Y_i - \bar{y}$, $\hat{y}_i = \hat{Y}_i - \bar{y}$ (vergl. (2.11)) und $e_i - \bar{e} = e_i$ wegen $\bar{e} = 0$. Es sei zunächst angemerkt, dass aus $y_i = \hat{y}_i + e_i$ wieder

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$$

folgt, denn $\sum_i y_i^2 = \sum_i (\hat{y}_i + e_i)^2 = \sum_i \hat{y}_i^2 + \sum_i e_i^2 + 2\sum_i \hat{y}_i e_i$, und $\sum_i \hat{y}_i e_i = \sum_i \sum_j \hat{\mathbf{b}}_j x_{ji} e_i = \sum_j \hat{\mathbf{b}}_j \sum_i x_{ji} e_i = 0$, da ja $\sum_i x_{ji} e_i = 0$ gemäß (4.1). Dann gilt

$$s^2(y) = s^2(\hat{y}) + s^2(e) \tag{4.10}$$

und folglich

$$\frac{s^2(\hat{y})}{s^2(y)} = 1 - \frac{s^2(e)}{s^2(y)}. \tag{4.11}$$

Nach (2.89) ist aber $R^2_{0.12\cdots k} = 1 - s^2(e)/s^2(y)$; mithin gilt (2.90).

Für die Korrelation $r(y, \hat{y})$ gilt

$$r^2(y, \hat{y}) = \frac{\sum_i y_i \hat{y}_i)^2}{\sum_i y_i^2 \sum_i \hat{y}_i^2}$$

Nun ist $\sum_i y_i\hat{y}_i = \sum_i(\hat{y}_i + e_i)\hat{y}_i = \sum_i \hat{y}_i^2 + \sum_i \hat{y}_i e_i$, und da $\sum_i \hat{y}_i e_i = 0$ ist $\sum_i y_i\hat{y}_i = \sum_i \hat{y}_i^2$. Dann folgt aber sofort

$$r^2(y,\hat{y}) = \frac{(\sum_i \hat{y}_i^2)^2}{\sum_i y_i^2 \sum_i \hat{y}_i^2} = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = \frac{s^2(\hat{y})}{s^2(y)} = R^2_{0.12\cdots k},$$

d.h. aber $r(y,\hat{y}) = R_{0.12\ldots k}$.

Um (**??**) einzusehen betrachtet man $\sum_i e_i^2$; es ist $\sum_i e_i^2 = \sum_i e_i e_i = \sum_i e_i(y_i - \hat{y}_i) = \sum_i e_i y_i$, denn $\sum_i e_i\hat{y}_i = 0$, wie oben gezeigt wurde. Dann ist

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

$$= \sum_{i=1}^n y_i^2 - \sum_{i=1}^n y_i((\hat{b}_1 x_{1i} + \cdots + \hat{b}_p x_{ki}) = \tag{4.12}$$

$$\sum_{i=1}^n y_i^2 - \hat{b}_1 \sum_{i=1}^n y_i x_{1i} - \cdots - \hat{b}_p \sum_{i=1}^n y_i x_{ki}, \tag{4.13}$$

mithin $\hat{b}_1 \sum_i y_i x_{1i} + \cdots + \hat{b}_p \sum_i y_i x_{ki} = \sum_i y_i^2 - \sum_i e_i^2$ und also

$$\frac{\hat{b}_1 \sum_i y_i x_{1i} + \cdots + \hat{b}_k \sum_i y_i x_{ki}}{\sum_i y_i^2} =$$

$$\frac{\sum_i y_i^2 - \sum_i e_i^2}{\sum_i y_i^2} = 1 - \frac{s^2(e)}{s^2(y)} = R^2_{0.12\ldots k}. \tag{4.14}$$

Sei $s_y = s(y)$, so folgt wegen (2.14) $\hat{\beta}_j = \hat{\mathbf{b}}_j s_j / s_y$

$$\frac{\hat{b}_j \sum_i y_i x_{ji}}{s_y^2} = \frac{\hat{b}_j}{s_y} \frac{\sum_i y_i x_{ji}}{s_y} = \frac{\hat{b}\, s_j}{s_y} \frac{\sum_i y_i x_{ji}}{s_x s_y}$$

und mithin

$$R^2_{0.12\ldots k} = \hat{\beta}_1 r_{y1} + \hat{\beta}_2 r_{y2} + \cdots + \hat{\beta}_k r_{yk}$$

□

Literatur

[1] Bierens, H. J. (2007) Multicollinearity. Pennsylvania State University

[2] Christensen, R.: Plane Answers to Complex Questions. The Theory of Linear Models. Springer-Verlag, 1987

[3] Fahrmeir, L., Kneib, T., Lang, S: Regression – Modelle, Methoden, Anwendungen. Springer-Verlag, Heidelberg 2007

[4] Fox, J.: Regression Diagnostics. Newbury Park 1991

[5] Friedman, J. (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29 (5), 1189 – 1232

[6] Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning – Data mining, inference, and prediction. Springer Science+Business Media, LLC 2009

[7] Hoerl, A. E., Kennard, R. W. (1970) Ridge Regression: Biased Estimation for Nonorthognal Problems. *Technometrics*, 12(1), 55–67

[8] Johnston, J.: Econometric Methods. McGraw-Hill 1963

[9] Knüsel, L. (2008) Singularwert-Zerlegung und Methode der kleinsten Quadrate. Technical Report Number 031, 2008; Department of Statistics, University of Munich, www.stat.uni-muenchen.de

[10] Mandel, J. (1982): Use of the singular value decomposition in regression analysis. *The American Statistician*, 36 (1), 15 – 24

[11] Rozeboom, W.W. (1979) Ridge Regression: Bonanza or Beguilement? *Psychological Bulletin*, 82, 242-249

[12] Seber, G.A.F.: Linear Regression Analysis. John Wiley and Sons, New York 1977

[13] Silvey, S.D. (1969) Multicollinearity and imprecise estimation. *J. Royal Society of Statistics, Series B*, 31, 539 – 552

[14] Tibshirani, R. (1996) Regression Shrinkage via the Lasso. *Journal of the Royal Statistical Society*, 58(1), 267–288

[15] Tychonov, A.N., Arsenin, V. Y.: Solution of ill-posed problems. Washington, Winston & Sons 1977

Index