Regression bei unvollständigen Daten
Statistik Seminararbeit

Mirko Junge*
http://www.geocities.com/junge_m
mailto:junge_m@web.de

12–14 Mai 1994

Inhaltsverzeichnis

1 Einführung **3**

2 Lineares Regressionsmodell ohne fehlende Daten **3**

3 Homogenitätstest **6**

4 Statistische Methoden bei fehlenden Daten **8**
4.1 Missing–Data–Mechanismen . . . 8
4.2 Nutzung der vollständigen Fälle (complete–case analysis) . . . 9
4.3 Imputationen für fehlende Daten (fill-in methods) . . . 9
4.4 Verfahren auf der Basis von Modellen . . . 10
4.5 Vergleich der Imputationsverfahren . . . 10
4.6 Fehlende Daten in der Y-Matrix . . . 11

5 Regression bei fehlenden Y-Werten **12**
5.1 Analysis of Variance (ANOVA) . . . 12
5.2 Analysis of Covariance (ANCOVA) . . . 13

6 X–Missing **13**
6.1 Der Fall K=2 . . . 14
6.2 Fehlende X–Werte in Matrixschreibweise . . . 16
6.3 Standardverfahren bei unvollstängiger X–Matrix . . . 19

7 Abschluß **19**

*Betreuer: D. Stemann, c/o FernUniversität Hagen, Fachbereich Wirtschaftswissenschaft, Lehrstuhl für Statistik und Ökonometrie, Prof.Dr.Dr. J. Gruber, Feithstraße 140, 58084 Hagen

Ich danke den Mitarbeitern der folgenden Bibliotheken dafür, daß sie mich das benötigte Material kopieren ließen oder mir dies kopierten. Ohne den Literaturschatz, den sie verwalten, wäre diese Arbeit nicht möglich gewesen.

Ärztliche Zentralbibliothek, Fachbereich Medizin der Universität Hamburg
Bibliothek des Fachbereichs Mathematik der Universität Hamburg
Bibliothek der Fachhochschule (Berliner Tor)
Staats– und Universitätsbibliothek der Universität Hamburg
Technische Universitätsbibliothek der Technischen Hochschule Hamburg–Harburg
Universitätsbibliothek Hagen der FernUniversität Hagen
Zahnärztliche Bibliothek der Universität Hamburg

May your quest for knowledge be a happy one

Unersetzbar für die Fertigstellung dieser Unterlagen waren die folgenden Bücher:

Malcom Clark
A plain TEXPrimer
1992, Oxford University Press
ISBN 0–19–853724–7

Donald E. Knuth
The TEXbook
1984, 1986 Addison Wesley / American Mathematical Society
ISBN 0–201–13448–9

typeset on a NEC Ultralite
TEXed on an ATARI ST4
calculations done on a HP48GX
graphics due to Diagramme
printed on a HP DeskJet 500C
acid free paper

1 Einführung

Unvollständige Daten können zum Beispiel durch das nicht vollständige Ausfüllen von Fragebögen auftreten. Antworten können zufällig oder nicht zufällig fehlen. So werden Angaben zum Arbeitsbeginn eher zufällig, Antworten zu Fragen zum Gehalt, Trinkverhalten oder Drogenmißbrauch eher nicht zufällig fehlen.
Auch technische Experimente in der Industrie führen zu unvollständigen Daten. So werden z.B. im Rahmen einer Qualitätssicherung Lebensdauer– und Helligkeitsuntersuchungen von Glühbirnen durchgeführt. Betreibt man eine Glühbirne mit maximaler Helligkeit, so nimmt die Lebensdauer ab. Es ist also nicht möglich, mit einer Glühbirne die maximale Helligkeit und die maximale Lebensdauer zu bestimmen, da jeder Test für sich dem Testobjekt einen irreversiblen Schaden zufügt.
In klinischen Langzeitstudien findet man unvollständige Daten beim sogenannten Drop–out [28, S. 219ff.]: Patienten fallen aus der Studie. Dies kann z.B. durch Verziehen, Unfall oder Tod passieren. Drop-out kann durch organisatorische Maßnahmen minimiert, jedoch nicht ausgeschlossen werden.
Bei der Datenanalyse von Gütern und Geldströmen in der Wirtschaftswissenschaft kommt es vor, daß einige Stromgrößen nur vierteljährlich gemessen werden, andere hingegen monatlich. Ein Beispiel dieser Kategorie ist das Arbeitslosengeld: Es wird wöchentlich bezahlt, die Fehlbeträge des Arbeitsamtes werden jedoch nur monatlich veröffentlicht. Auch hierbei liegen unvollständige Daten vor, wobei allerdings beachtet werden muß, daß die Daten in diesem Fall nicht zufällig fehlen [5].

Im folgenden wird, bevor auf Regression bei fehlenden Daten eingegangen wird, erst das lineare Regressionsmodell vorgestellt, um die Variablenbezeichungen einzuführen. Bevor die Regressionsmodelle in der mathematisch kompakteren Matrixschreibweise präsentiert werden, wird die konventionelle Summendarstellung der Modelle vorgestellt, da diese für die Programmierung auf Computern in Hochsprachen ohne Matrixunterstützung wesentlich einfacher zu implementieren sind. Die hierdurch entstandenen Duplikationen bitte ich zu entschuldigen.

2 Lineares Regressionsmodell ohne fehlende Daten

Mit Hilfe der linearen Regression kann man Zusammenhänge zwischen einer beobachtbaren abhängigen Variablen (Y) und beobachtbaren unabhängigen Variablen ($X_t, \quad t = 1, \ldots, K$) beschreiben, sofern man diese Abhängigkeit durch eine lineare Funktion beschreiben kann. Hierbei ist es unerheblich, ob man die Daten hierfür erst transformieren muß oder nicht. Berücksichtigt man eine Zufallsvariable, die der linearen Funktion additiv überlagert ist, so erhält man eine Regressionsgleichung der Form:

$$Y = \beta_1 X_1 + \ldots + \beta_K X_K + U$$

Zielfunktion der linearen Regression unter Verwendung der Kleinste–Quadrate (KQ) Schätzung ist die Minimierung des Quadrates des Fehlers der Regressionsfunktion. Als Fehler der t–ten Beobachtung wird definiert:

$$\check{u}_t = y_t - \check{y}_t$$

Hierbei ist y_t der beobachtete Wert des Regressanden und $\check{y}$ der durch die Regressionsgleichung mit den beliebigen Parametern $\check{\beta}_t, \quad t = 1, \ldots, K$ prognostizierten Wert des Regressanden. Somit erhält man für das Quadrat des Fehlers:

$$\check{u}_t^2 = (y_t - \check{y}_t)(y_t - \check{y}_t) = (y_t - \check{\beta}_1 x_{t1} - \ldots - \check{\beta}_K x_{tK})^2$$

Die Summe der Quadrate ergibt sich bei N Samples zu:

$$S(\check{\beta}_0, \check{\beta}_1, \ldots, \check{\beta}_K) = \sum_{t=1}^{N} \left(y_t - \check{\beta}_0 - \sum_{k=1}^{K} \check{\beta}_k x_{tk} \right)^2$$

Eine notwendige Bedingung für ein Minimum ist eine Nullstelle in der ersten Ableitung:

$$\frac{\partial \ S(\check{\beta}_0, \check{\beta}_1, \ldots, \check{\beta}_K)}{\partial \check{\beta}_0} = \frac{\partial \ S(\check{\beta}_0, \check{\beta}_1, \ldots, \check{\beta}_K)}{\partial \check{\beta}_1} = \ldots \frac{\partial \ S(\check{\beta}_0, \check{\beta}_1, \ldots, \check{\beta}_K)}{\partial \check{\beta}_K} = 0$$

Man erhält also die Gleichungen:

$$\begin{aligned}
\textstyle\sum_{t=1}^{N} 2(y - \check{\beta}_0 - \check{\beta}_1 x_1 - \ldots - \check{\beta}_k x_k)(-1) &= 0 \\
\textstyle\sum_{t=1}^{N} 2(y - \check{\beta}_0 - \check{\beta}_1 x_1 - \ldots - \check{\beta}_k x_k)(-x_1) &= 0 \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad &= 0 \\
\textstyle\sum_{t=1}^{N} 2(y - \check{\beta}_0 - \check{\beta}_1 x_1 - \ldots - \check{\beta}_k x_k)(-x_K) &= 0
\end{aligned}$$

Durch Umschreiben der Gleichungen erhält man die Normal–Gleichungen.

$$\check{\beta}_0 N + \check{\beta}_1 \sum x_1 + \ldots + \check{\beta}_K \sum x_k = \sum y$$

$$\check{\beta}_0 \sum x_1 + \beta_1 \sum x_1 x_1 + \ldots + \check{\beta}_K \sum x_1 x_K = \sum x_1 y$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad = \quad \vdots$$

$$\check{\beta}_0 \sum x_K + \beta_1 \sum x_K x_1 + \ldots + \check{\beta}_K \sum x_K x_K = \sum x_K y$$

Den Haken ($\check{}$) über den Variablen tauscht man in ein Dach ($\hat{}$), wenn durch die Variablen die Lösung beschrieben wird, die die Summe der Fehler der Regression minimiert. Die allgemeinen Lösungen der Normal–Gleichungen ergeben [11, S. 835]; [26, S. 44]:

$$\hat{\beta}_j = \sum_{k=1}^{K} \frac{\hat{\sigma}_{k0}}{\hat{\sigma}_{jk}}, \quad j \neq 0$$

$$\hat{\beta}_0 = \bar{Y} - \sum_{j=1}^{K} \hat{\beta}_j \bar{X}_j$$

Hierbei sind:

$$\bar{X}_j = \frac{\sum_{i=1}^{N} X_{ij}}{N}, \quad \bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

$$\hat{\sigma}_{j0} = \frac{\sum_{i=1}^{N} (X_{ij} - \bar{X}_j)(Y_i - \bar{Y})}{N-1}, \quad \hat{\sigma}_{jk} = \frac{\sum_{i=1}^{N} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{N-1}$$

Um eine möglichst kompakte Darstellung zu ermöglichen, kann man die Gleichungen auch in Matrixdarstellung aufschreiben. Hierfür empfiehlt es sich die erste Spalte der $\mathbf{X}$–Matrix zu 1 zu setzen:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1K} \\ 1 & x_{22} & \ldots & x_{2K} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{N2} & \ldots & x_{NK} \end{pmatrix}$$

Somit kann man für die Summe der Fehlerquadrate ($\sum_{i=1}^{K}(\check{\beta})$) unter Verwendung von $\check{y} = X\check{\beta}$ und $\check{u} = y - \check{y}$ schreiben:

$$S(\check{\beta}) = \sum_{j=1}^{N} \check{u}_j^2 = \sum_{j=1}^{N} (y_j - \check{\beta}_1 x_{j1} - \ldots - \check{\beta}_K x_{jK})^2 = (y - X\check{\beta})'(y - X\check{\beta}) = \check{u}'\check{u}$$

Ziel ist es, die Funktion $S(\check{\beta})$ zu minimieren. Dies geschieht durch Differenzieren und Nullsetzen der ersten Ableitung:

$$S(\check{\beta}) = (y - X\check{\beta})'(y - X\check{\beta}) = (y' - \check{\beta}'X')(y - X\check{\beta}) = y'y - 2\check{\beta}'X'y + \check{\beta}'X'X\check{\beta}$$

$$\frac{\partial(S(\check{\beta}))}{\partial\check{\beta}} = \frac{\partial(y'y)}{\partial\check{\beta}} - 2\frac{\partial(\check{\beta}'X'y)}{\partial\check{\beta}} + \frac{\partial(\check{\beta}'X'X\check{\beta})}{\partial\check{\beta}} = -2X'y + 2X'X\check{\beta}$$

$$\frac{\partial(S(\check{\beta}))}{\partial\check{\beta}} \,|_{\check{\beta}=\hat{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0 \iff \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

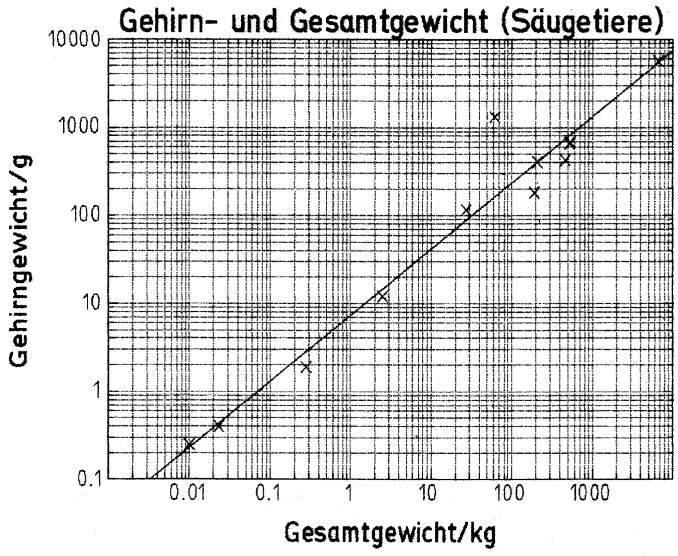
Somit erhält man:

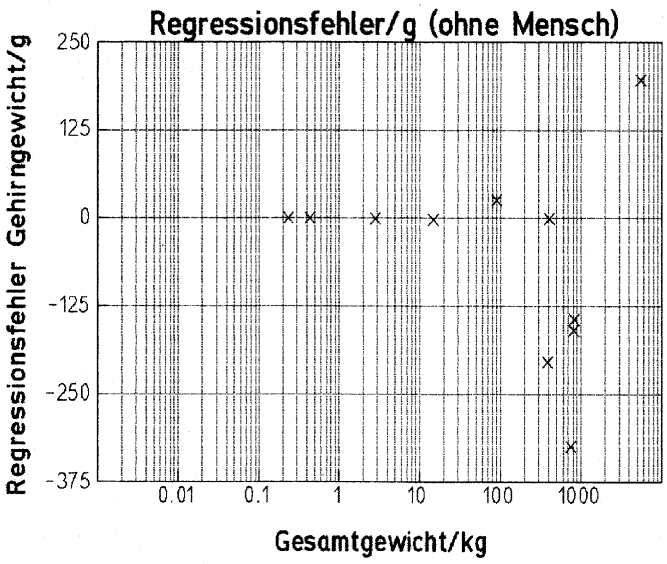
$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Im folgenden ist ein Beispiel der linearen Regression anhand von Daten aus der Biologie dargestellt. Für die Untersuchung von Schlafeigenschaften bei Säugetieren interessiert unter anderem das Gehirngewicht (Y in [g]). Dies soll unter Benutzung der vorhandenen Daten für den Asiatischen Elephanten geschätzt werden, für den keine Daten zur Verfügung standen. Die Daten enthält die Tabelle:

Säugetier	Gesamt-gewicht (G) [kg]	Gehirn-gewicht (B) [g]	$\log_{10}(G)$ X	$\log_{10}(B)$ Y	$\hat{y}$	$(Y-\hat{y})^2$	$10^{\hat{y}}$ [g]	Regressions-Fehler $B-10^{\hat{y}}$ [g]
Elephant(Afr.)	6654.0	5712.0	3.8231	3.7568	3.7416	$2.310 \cdot 10^{-4}$	5515.8	196.2
Elephant(Asia)	2547.0	x	3.4060	x	3.4284	x	2681.6	x
Giraffe	529.0	680.0	2.7235	2.8325	2.9157	$6.922 \cdot 10^{-3}$	823.63	-143.6
Pferd	521.0	655.0	2.7168	2.8162	2.9108	$8.949 \cdot 10^{-3}$	814.25	-159.3
Kuh	465.0	423.0	2.6675	2.6263	2.8737	$6.121 \cdot 10^{-2}$	747.60	-324.6
Gorilla	207.0	406.0	2.3160	2.6085	2.6097	$1.440 \cdot 10^{-6}$	407.09	-1.088
Schwein	192.0	180.0	2.2833	2.2553	2.5852	$1.088 \cdot 10^{-1}$	384.73	-204.7
Mensch	62.0	1320.0	1.7924	3.1206	2.2165	$8.174 \cdot 10^{-1}$	164.60	1155.4
Ziege	27.66	115.0	1.4419	2.0607	1.9532	$1.156 \cdot 10^{-2}$	89.780	25.22
Hase	2.50	12.10	0.3979	1.0828	1.1692	$7.465 \cdot 10^{-3}$	14.762	-2.662
Ratte	0.280	1.90	-0.5528	0.2787	0.4551	$3.112 \cdot 10^{-2}$	2.8514	-0.9514
Maus	0.023	0.40	-1.6383	-0.3979	-0.3601	$1.429 \cdot 10^{-3}$	0.4364	-0.0364
Lt.br.Bat	0.010	0.25	-2.0000	-0.6021	-0.6318	$8.821 \cdot 10^{-4}$	0.2334	0.0166
Summe						$1.056 \cdot 10^{0}$		539.89

Graphisch dargestellt erhält man:





Wie man aus der graphischen Darstellung von B über G sieht, eignet sich eine logarithmische Darstellung für eine lineare Interpolation. Man setzt also:

$$\mathbf{X} = \begin{pmatrix} 1 & 3.8231 \\ 1 & 2.7135 \\ \vdots & \vdots \\ 1 & -2.0000 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 3.7568 \\ 2.8325 \\ \vdots \\ -0.6021 \end{pmatrix}$$

Einsetzen in die Formel $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ ergibt:

$$\hat{\beta} = \begin{pmatrix} 0.8703 \\ 0.7511 \end{pmatrix}$$

Somit ergibt sich $\hat{y} = x\hat{\beta}$, oder als Funktionsgleichung geschrieben $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$.
Benutzt man die lineare Regressionsgleichung zur Interpolation des Wertes für den Asiatischen Elephanten mit 2547kg Gesamtgewicht ($\log(G) = 3.4060$) erhält man

$$\hat{y} = \begin{pmatrix} 1 & 3.4060 \end{pmatrix} \begin{pmatrix} 0.8703 \\ 0.7511 \end{pmatrix} = 3.4284$$

Die lineare Regression liefert also als Schätzung für das Gehirngewicht des Asiatischen Elephanten ein Gewicht von $10^{3.4284}g = 2681g$.

3 Homogenitätstest

Durch Homogenitätstest muß man klären, ob signifikante Schichtungseffekte im Datenbestand vorliegen, die zu einer Verzerrung führen [Toutenburg(1992), 199]. Mittels des χ^2–Tests kann man u.a. eine Hypothese über die gemeinsame Verteilung zweier Merkmale testen. Von besonderer Bedeutung ist hierbei die Überprüfung auf Unabhängigkeit zweier Merkmale (χ^2–Unabhängigkeitstest) [13, Chp. 12.3.3]; [30, S. 555ff.]:
Als Nullhypothese wird hierbei die Unabhängigkeit der Merkmale formuliert. Hieraus folgt, daß bei Nichtablehnung der Nullhypothese eine Abhängigkeit der Merkmale statistisch nicht nachgewiesen werden kann und bei Ablehnung der Nullhypothese eine statistische Abhängigkeit, bei einer Irrtumswahrscheinlichkeit von α (Signifikanzniveau) der Merkmale vorliegt.
Die Testgröße χ^2_* berechnet sich aus absoluten Häufigkeiten nach

$$\chi^2_* = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(h_{oij} - h_{eij})^2}{h_{eij}}$$

Hierbei bezeichnet die h_{oij} die Häufigkeit der beobachteten Ausprägungen eines Merkmals x_i und y_j; h_{eij} sind die bei Unabhängigkeit der Merkmale zu erwartenden Häufigkeiten. Formal geschrieben sind $h_{oij} = h(x_i, y_j)$, $h_{eij} = \frac{1}{N}h(x_i)h(y_j)$, mit $h(x_i) = \sum_{j=1}^{s} h(x_i, y_j)$, $h(y_j) = \sum_{i=1}^{r} h(x_i, y_j)$ und $N = \sum_{i=1}^{r}\sum_{j=1}^{s} h(x_i, y_j)$. Sind erwartete Häufigkeiten h_{eij} bei einem Unabhängigkeitstest kleiner als 5, so sind Spalten oder Zeilen so zusammenzufassen, daß jedes $h_{eij} \geq 5$ wird [Gruber(1993), 12.3]. Einige Autoren meinen, daß man Zeilen oder Spalten nur zusammenfassen muß, wenn ein $h_{eij} < 1$ [30, S. 550].
Die Annahmegrenze ergibt sich aus der χ^2–Verteilung mit $\nu = (r-1)(s-1)$ Freiheitsgraden und einem Signifikanzniveau α.

$$\chi^2_\nu(x) = \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})}\int_0^x t^{\frac{\nu-1}{2}}e^{-\frac{1}{2}t}dt$$

Dieses Integral ist nicht elementar lösbar. Numerische Näherungen kann man Tabellen entnehmen [19, S. 37f.]; [12, S. 460] oder von Taschenrechnern/Computern approximieren lassen [17].
Ist $\chi^2 \leq \chi^2_*$, so wird die Nullhypothese abgelehnt. Es liegt also mit einer Irrtumswahrscheinlichkeit von α (Signifikanzniveau) eine statistische Abhängigkeit der Merkmale vor.

Als Beispiel sei eine Untersuchung angegeben, bei der die Zahl der pro Tag gerauchten Zigaretten und die Todesrate durch Lungen– oder Bronchialkrebs untersucht wurden [20], wie in [31, S. 429] zitiert]. (Daß Zigarettenrauch auch canceröse Entartungen in anderen Geweben als den Untersuchten hervorruft, wurde außer acht gelassen, kann jedoch nicht vernachlässigt werden [10, S. 811]; [22, S. 664].

Zahl der Totesfälle / Personenjahre Beob.		
Alter	Nichtraucher	Raucher (1–2 Pack./Tag)
35–44	0/ 35200	4/ 40600
45–54	0/ 15100	10/ 12800
55–64	25/214000	245/103000
65–74	49/171000	194/ 50000
75+	4/ 8490	7/ 1270
Total:	78/443790	460/207670

Die Nullhypothese sei, daß die Todesfälle unabhängig vom Rauchen sind. Damit $h_{eij} \geq 5$, also jede der erwarteten Häufigkeiten größer gleich 5, müssen Zeilen zusammengefaßt werden:

Zahl der Totesfälle / Personenjahre Beob.		
Alter	Nichtraucher	Raucher (1–2 Pack./Tag)
35–64	25/264300	259/156400
65+	53/179490	201/ 51270
Total:	78/443790	460/207670

Somit erhält man für die erwartete Häufigkeiten h_{eij} Werte:

Erwartete Häufigkeiten (h_{eij})		
Alter	Nichtraucher	Raucher (1–2 Pack./Tag)
35–64	34.51	221.4
65+	36.83	217.2

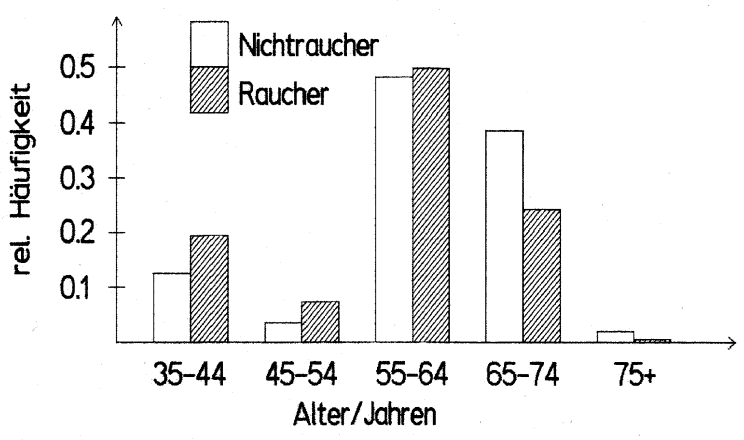
Somit ergibt sich $\chi^2_* = 15.74$. Mit einem Signifikanzniveau $\alpha = 0.05$ und der Zahl der Freiheitsgrade $\nu = (2-1)(2-1) = 1$ ist $\chi^2 = 3.84$. Also:

$$\chi^2 \leq \chi^2_*$$

Die Nullhypothese wird abgelehnt: Mit einer Irrtumswahrscheinlichkeit von 5% ist die Zahl der Todesfälle abhängig von den Rauchgewohnheiten.
Setzt man nur $h_{eij} \geq 1$ voraus [Wonnacott(1990), 550], so müssen nur die ersten beiden Zeilen zusammengefaßt werden. Man erhält $\chi^2_* = 18.89$. Bei einem Signifikanzniveau $\alpha = 0.05$ und $\nu = (4-1)(2-1) = 3$ ist $\chi^2 = 7.81$. Auch in diesem Fall ist $\chi^2 \leq \chi^2_*$, und die Nullhypothese wird abgelehnt.
Ein Problem könnte sich jedoch daraus ergeben, daß nicht jede der Altersklassen mit der gleichen Häufigkeit in der Untersuchung vertreten ist. In dem folgenden Graphen ist die relative Anzahl der Personenjahre, die für die einzelnen Altersklassen beobachtet wurden, dargestellt.

Relative Verteilung der Beobachtungen		
Alter	Nichtraucher	Raucher (1–2 Pack./Tag)
35–44	0.0793	0.1955
45–54	0.0340	0.0616
55–64	0.4822	0.4960
65–74	0.3853	0.2408
75+	0.0191	0.0061

Graphisch dargestellt ergibt sich die Tabelle zu:



Aus der relativen Verteilung erkennt man, daß die Raucher–Gruppe in den Altersgruppen $35-44$ und $45-54$ relativ mehr Personenjahre auf sich vereinigt, wohingegen sie in den anderen Altersgruppen relativ weniger Personenjahre besitzt. Hieraus wird klar, daß ein Vergleich der akkummulierten Sterberaten von $78/443790 \approx 0.1758/1000$ für Nichtraucher bzw. $460/207670 \approx 2.215/1000$ für Raucher kein unverzerrten Schätzer für die Auswirkungen des Rauchens ist. Um die Daten dennoch vergleichbar zu machen, führt Woolson eine Standardisierung und Anpassung der Häufigkeiten ein [31, S. 428ff.]. Die standardisierten Häufigkeiten werden mit T bezeichnet, der Index 0 bezeichne die Nichtraucher und der Index 1 die Raucher:

$$T_0 = \sum_{i=1}^{N} \frac{n_{i0} + n_{i1}}{N_\Sigma} r_{i0}, \quad T_1 = \sum_{i=1}^{N} \frac{n_{i0} + n_{i1}}{N_\Sigma} r_{i1}$$

Wobei n_{ij} für die Zahl aller Beobachtungen angibt, die gemacht wurden, um die Häufigkeit h_{oij} zu erhalten; $r_{ij} = h_{oij}/n_{ij}$ bezeichnet die relative Häufigkeit und $N_\Sigma = \sum_{l=1}^{N}(n_{l0} + n_{l1})$. T_0 ist also die Todesrate der Standardpopulation.
Für die Differenz der beiden angepaßten Häufigkeiten erhält man:

$$T_0 - T_1 = \sum_{i=1}^{N} \frac{n_{i0} + n_{i1}}{N_\Sigma}(r_{i0} - r_{i1})$$

Unter der Annahme, daß r_{i0} und r_{i1} mit $p_{0j} = p_{1j}$ unabhängig binomial verteilt sind, und daß $(n_{i0} + n_{i1})/N_\Sigma$ nicht zufällig sind, wie im Falle sehr großer Populationen, so ist die Varianz von $T_0 - T_1$ [Woolson(1987), 435]:

$$\hat{\sigma}^2_{(T_0 - T_1)} = \sum_{i=1}^{N} \left(\frac{n_{0i} + n_{1i}}{N_\Sigma} \right)^2 \frac{\bar{r}_i(1 - \bar{r}_i)(n_{i0} + n_{i1})}{n_{i0} n_{i1}}, \quad \bar{r}_i = \frac{h_{oi0} + h_{oi1}}{n_{i0} + n_{i1}}$$

Die Testgröße $\chi^2_\star$ errechnet sich dann zu:

$$\chi^2_\star = \frac{(T_0 - T_1)^2}{\hat{\sigma}^2_{(T_0-T_1)}}$$

Die Testgröße $\chi^2_\star$ folgt einer χ^2–Verteilung mit einem Freiheitsgrad ($\nu = 1$).
Wendet man die obigen Formeln auf das Raucherbeispiel an, so erhält man bei der Berechnung folgende Zwischenwerte: (Es wurde nicht mit der Sterbehäufigkeit (h_{oij}, sondern mit der Überlebenshäufigkeit $h'_{oij} = n_{ij} - h_{oij}$ gerechnet.)

Zwischenwerte der Berechnung							
Alter	n_{0j}	n_{1j}	h'_{0j}	r_{0j}	h'_{1j}	r_{1j}	$\bar{r}_i(1-\bar{r}_i)$
35–44	35200	40600	35200	1.0000	40596	0.9999015	$5.276766 \cdot 10^{-5}$
45–54	15100	12800	15100	1.0000	12790	0.9992188	$3.582945 \cdot 10^{-4}$
55–64	214000	103000	213950	0.9997664	102755	0.9976214	$9.297334 \cdot 10^{-4}$
65–74	171000	50000	170951	0.9997135	49806	0.99612	$1.098339 \cdot 10^{-3}$
75+	8490	1270	8486	0.9995289	1263	0.9944882	$1.125779 \cdot 10^{-3}$

Mit Hilfe dieser Zwischenwerte berechnet man leicht:

$$T_0 = 0.999682, \quad T_1 = 0.997299, \quad T_0 - T_1 = 0.002383$$

$$\hat{\sigma}^2_{(T_0-T_1)} = 6.7443 \cdot 10^{-9}$$

Somit ergibt sich:

$$\chi^2_\star = 832.9$$

Mit einem Signifikanzniveau $\alpha = 0.05$ und einem Freiheitsgrad $\nu = 1$ erhält man $\chi^2 = 3.8415$. Somit ist

$$\chi^2 \leq \chi^2_\star$$

Die Nullhypothese wird auch hier abgelehnt: Mit einer Irrtumswahrscheinlichkeit von 5% ist die Zahl der Todesfälle abhängig von den Rauchgewohnheiten.

4 Statistische Methoden bei fehlenden Daten

4.1 Missing–Data–Mechanismen

Eine ignorierbare Nichtresponse liegt stets dann vor, wenn die fehlenden Daten zufällig fehlen, also die beobachteten Werte eine zufällige Stichprobe der Gesamtstichprobe sind.
Um die Begriffe der Missing–Data–Mechanismen näher zu erläutern, gehen wir von einer bivariaten Stichprobe (X, Y) aus, wobei angenommen wird, daß X vollständig beobachtet wurde, während Y fehlende Werte aufweist. Durch Umordnung kann man ohne Beschränkung der Allgemeinheit ein sogenanntes monotones Pattern erzeugen. Hierfür werden die n vollständigen Samples zuerst geschrieben, und die unvollständigen $N - n$ Samples am Schluß.
Man kann somit schreiben:

$$X = \begin{pmatrix} X_c \\ X_* \end{pmatrix}, \quad y = \begin{pmatrix} y_c \\ y_* \end{pmatrix}$$

Hierbei ist X_c (c: für complete) der Teil der Matrix X, der die vollständigen Daten enthält ($n \times K$ Matrix), und X_* ist der Teil von X, der unvollständige Datensätze enthält. Entsprechendes gilt für den y–Vektor.
Die Wahrscheinlichkeit der Response von Y kann auf verschiedene Weisen von X und Y abhängen [27, S. 201]:

(i) sie hängt von Y und X ab

(ii) sie hängt von X, aber nicht von Y ab

(iii) sie ist unabhängig sowohl von X als auch von Y

Im Fall (i) sind die Daten weder MAR (missing at random) noch OAR (observed at random). Somit ist der Missing–Data–Mechanismus nicht ignorierbar.
Den Fall (i) kann man beim Bestimmen von chemischen Konzentrationen vorfinden. Sind bei einigen Datensätze die Konzentrationen kleiner als die mit dem entsprechenden Gerät meßbaren, so fehlen die Daten nicht MAR, da nur kleine Konzentrationen fehlen.
Im Fall (ii) sind die fehlenden Daten MAR. Hieraus folgt, daß die beobachteten y–Werte y_{obs} nicht notwendiger Weise eine zufällige Stichprobe von y bilden. Innerhalb der durch die X–Werte definierten Klassen bilden sie jedoch zufällige Stichproben.

Als Beispiel für diesen Fall sei eine Untersuchung über die Zartheit von tiefgefrorenem Truthahnfleisch angeführt: Das Fleisch wird vor dem Verarbeiten und Tieffrieren auf Zartheit untersucht und entsprechend einer Punkteskala markiert (X). Von Interesse ist jedoch die Zartheit (Y) nach dem Auftauen der Truthähne. Während der Lagerung werden die als besonders zart markierten Zubereitungen gestohlen, so daß für diese Tiere die Zartheit nach dem Auftauen (Y) nicht mehr bestimmt werden kann. Die Response von Y ist MAR, da der Prozeß, der den Verlust verursacht hat, eine Funktion von X ist, jedoch nicht von Y, obwohl X und Y wahrscheinlich voneinander abhängen [29, S. 221].
Im Fall (iii) sind die fehlenden Daten MAR und die beobachteten Daten sind OAR, so daß die fehlenden Daten MCAR (missing completely at random) sind. Somit bilden die Daten y_{obs} eine zufällige Stichprobe von $y = (y_{obs}, y_{mis})$.
In den Fällen (ii) und (iii) ist der Missing–Data–Mechanismus bei Verfahren auf der Basis der Likelyhoodfunktion ignorierbar, im Fall (iii) auch bei Verfahren auf der Basis der Stichprobe [27, S. 201].

4.2 Nutzung der vollständigen Fälle (complete–case analysis)

In der complete–case Analysis streicht man alle unvollständig beobachteten Zeilen der Datenmatrix. Hierdurch erreicht man das gleiche, wie durch Auffüllen der fehlenden Werte mit neutralen Werten [14, S. 82]. Dies setzt voraus, daß der Anteil der unvollständigen Datensätze im Verhältnis zu den vollständigen vernachlässigbar ist und, daß auch keine Blockbildung im Missing-Pattern vorliegt. Durch Homogenitätstests muß man klären, ob durch das Weglassen von einzelnen Datenzeilen signifikante Schichtungseffekte vorliegen, die zu einer Verzerrung (Selectivity Bias) führen.
Daß diese Prüfung sinnvoll ist, zeigt sich z.B. in einer Untersuchung an vollständigen Daten, bei denen durch Zufall einige Datenwerte gelöscht wurden. In fast allen Fällen lieferte die complete–case Analysis ein Ergebnis, das besser mit den Werten der vollständigen Datenreihe übereinstimmte. Nur in den Fällen, in denen es viele unvollständige Datensätze gab, oder die fehlenden Werte MAR waren, lieferten die Auffüllmethoden bessere Ergebnisse. Eine vollständige Übersicht der Versuchsläufe ist in [14] angegeben.

4.3 Imputationen für fehlende Daten (fill-in methods)

Unter Imputation versteht man das Auffüllen der unvollständigen Teilmatrix X_*. Da der fehlende Wert unbekannt ist, muß immer mit einer Abweichung von Imputation zum Originalwert gerechnet werden. Dies kann für die durchgeführte Analyse so starke Auswirkungen haben, daß das Ergebnis zum größten Teil nur von der Wahl der Imputationen und nicht mehr von den wirklich beobachteten Daten abhängt [27, S. 199]. Die Möglichkeit von so gravierenden Auswirkungen ist deshalb beim Auffüllen der Matrix X_* zu beachten und durch systematisches Verändern der Imputationen zu testen.
Man unterscheidet verschiedene Arten von Imputationen [27, S. 199]; [29, S. 224f]:

- Hot–Deck Imputation
- Cold–Deck Imputation
- Mean Imputation
- Regression (Correlation) Imputation
- Multiple Imputation

Unter Hot–Deck Imputation versteht man das Einsetzen von realisierten Werten der betreffenden Variablen. Es treten in der aufgefüllten X–Matrix also nur numerische Werte auf, die auch tatsächlich beobachtet werden könnten. Ein solches Imputationsverfahren ist zum Beispiel dann sinnvoll, wenn der Wertebereich für die aufzufüllende Variable nicht stetig oder beschränkt ist.
Als Cold–Deck Imputation bezeichnet man das Einsetzen eines konstanten Wertes aus einer externen Quelle für die fehlende Variable. Dies kann zum Beispiel eine Konstante der Population sein, wie das mittlere Alter der weiblichen Population. Es handelt sich hierbei also um eine Imputation, die die fehlenden Werte mit Konstanten auffüllt, die keine direkte Verbindung zu den beobachteten Werten hat, also auch andere Fehler und Erfassungsgrundlagen besitzt.
Mean–Imputation bezeichnet das Einsetzen des Stichproben– bzw. Spaltenmittelwertes in der fehlenden Variablen. Es wird der Mittelwert der betreffenden Spalte über alle vorhandenen Werte berechnet, also nicht nur über die von X_c. Bei univariaten Verteilungen ist der Stichprobenmittelwert der beste Schätzer für den Mittelwert [9]. Es wird also mittelwertsneutral aufgefüllt.
Bei der Regressions Imputation wird die Korrelationsstruktur innerhalb der X_c–Matrix ausgenutzt. Es wird nicht nur die Korrelation innerhalb der betreffenden Spalte der X_c–Matrix benutzt! Die fehlenden Datenwerte werden mittels der klassischen Regression vorhergesagt. Hierbei ist zu beachten, daß die Regressionkoeffizienten nur aus der X_c–Matrix gewonnen werden, also der Trend in den vollständigen Daten auf die unvollständigen Daten übertragen wird [29, S. 224]. Damit dies sinnvoll ist, muß sichergestellt sein, daß die vollständigen Daten repräsentativ für die gesamte Datenmenge sind. Es ist in jedem Falle ein Homogenitätstest (siehe oben) durchzuführen. Weiterhin muß sichergestellt sein, daß zwischen den Spalten der X–Matrix keine zu große Korrelation besteht, da in diesem Fall die klassische Regression nicht angewandt werden darf [4, S. 173ff.]; [15, S. 204].

Die Daten in der folgenden Tabelle stellen jeweils das Gewicht und die Größe von männlichen Jugendlichen im Alter von 9 und 18 Jahren dar [29, S. 52]. Die fehlenden Daten für die achtzehnjährigen werden jeweils mit den angegebenen Methoden aufgefüllt.

Sample Nummer	Gewicht 9 Jahre [kg]	Länge 9 Jahre [cm]	Gewicht 18 Jahre [kg]	Länge 18 Jahre [cm]
1	41.5	139.4	110.2	179.0
2	31.0	144.3	79.4	195.1
3	30.1	136.5	76.3	183.7
4	34.1	135.4	74.5	178.7
5	24.5	128.9	55.7	171.5
6	29.8	136.0	68.2	181.8
7	26.0	128.5	78.2	172.5
8	30.1	133.2	66.5	174.6
9	37.9	145.6	70.5	190.4
10	27.0	132.4	57.3	173.8
11	25.9	133.7	50.3	172.6
12	31.1	138.3	70.8	185.2
13	34.6	134.6	73.7	178.4
14	34.6	139.0	75.2	177.6
15	43.1	146.0	83.1	183.5
16	33.2	133.2	74.3	178.1
17	30.7	133.3	72.2	177.0
18	31.6	130.3	88.6	172.9
19	33.4	144.5	75.9	188.4
20	29.4	125.4	64.9	169.4
μ_c	31.98	135.93	73.92	179.21
σ_c	4.73	5.72	12.30	6.66

					Hot–Imputation			Cold–Imputation		Mean–Imputation		Regression–Imputation	
					Gewicht [kg]	Länge [cm]	Sample No.	Gewicht [kg]	Länge [cm]	Gewicht [kg]	Länge [cm]	Gewicht [kg]	Länge [cm]
1*	30.2	135.8	65.6	·	x	183.7	(3)	x	175	x	179.21	x	179.31
2*	31.1	139.9	66.4	·	x	185.2	(12)	x	175	x	179.21	x	184.20
3*	27.6	136.8	·	182.4	57.3	x	(10)	70	x	73.95	x	63.41	x
4*	32.3	140.6	·	185.8	88.6	x	(18)	70	x	73.95	x	72.54	x
5*	29.0	138.6	·	·	64.9	169.4	(20)	70	175	73.95	179.21	65.22	183.75
6*	31.4	140.0	·	·	88.6	172.0	(18)	70	175	73.95	179.21	69.88	184.55

Für die Schätzwerte der Regressionskoeffizienten erhält man folgende Werte:

$$\hat{\beta}_{g18} = \begin{pmatrix} 12.98 \\ -0.6627 \\ 1.315 \\ 0.1186 \end{pmatrix}, \quad \hat{\beta}_{l18} = \begin{pmatrix} 47.99 \\ 2.684 \\ -1.910 \\ 1.111 \end{pmatrix}, \quad \hat{\beta}_{g18,l18} = \begin{pmatrix} 71.86 & 21.50 \\ 2.243 & -0.3967 \\ -0.5173 & 1.254 \end{pmatrix}$$

Hierbei bezeichnet $\hat{\beta}_{g18}$ den Schätzwert für den Regressionskoeffizienten für die Schätzung des Gewichtes bei 18 jährigen in Abhängigkeit des Gewichts als neunjähriger, der Länge als neunjähriger und der Länge als achtzehnjähriger. Dementsprechend bezeichnet $\hat{\beta}_{g18,l18}$ die Schätzung des Regressionskoeffizienten bei achtzehnjährigen für Gewicht und Länge, wobei Gewicht und Länge als neunjähriger die unabhängigen Daten sind.
In der Multiplen Imputation wird durch wiederholte Imputation und Auswertung jedes vervollständigten Datensatzes eine Variabilität in der Zielgröße erreicht. Aus diesem Variabilitätsbereich wird die endgültige Zielgröße z.B. durch Mittelwertsbildung berechnet [23].

4.4 Verfahren auf der Basis von Modellen

Die Grundidee besteht in der Faktorisierung der Likelihood–Funktion nach der Beobachtungs– und Fehlendstruktur, so daß iterative Verfahren, beginnend mit den vollständigen Daten, eine schrittweise Maximierung der gesamten Likelihoodfunktion ermöglichen [18].

4.5 Vergleich der Imputationsverfahren

Afifi und Elashoff haben die verschiedenen Methoden, die man bei einfacher Regression benutzen kann, näher untersucht und festgestellt, daß keine der bekannten Methoden über alle Datensätze immer die besten Ergebnisse erzielt [1]. Allgemein gefaßt fanden sie heraus, daß die Mean Imputation für Datenreihen am besten ist, die sehr gering miteinander korreliert sind. Die complete–case Analysis soll am besten für Daten mit geringer Korrelation sein und die Regression Imputation für stark korrelierte Daten.
Für die multiple Regression stellte Haitovsky folgende Bauernregeln auf [14, S. 80]:

- Die klassische Methode, also die complete–case Analysis, sollte gewählt werden, wenn der Anteil der unvollständigen Daten an der Gesamtdatenmenge gering und nicht zu sehr über die multivariaten Beobachtungswerte verteilt ist. Weiterhin sollte keiner der paarweisen Korrelationskoeffizienten zu groß sein.

- Die Regressionsimputation sollte benutzt werden, wenn eine Variable mit fehlenden Werten stark mit einer Variablen mit nur wenigen fehlenden Datenwerten korreliert ist.

4.6 Fehlende Daten in der Y-Matrix

Ein praktisches Beispiel, das zu einer unvollständigen Y–Matrix führt, ist das folgende [21, S. 523ff.]:
Man habe eine Anzahl von p verschiedenen psychologischen Tests, die entworfen wurden, um das Abschneiden von Studenten in ihrem ersten Studienjahr vorherzusagen. Die Zahl der Studienanfänger, die zu Beginn des Semesters an den Tests teilnehmen sei N_1. Am Ende des Schuljahres sei eine Erfolgsskala der Studenten verfügbar. Aus diesen Daten kann man die Mittelwerte der einzelnen Tests p und die Covarianzmatrix der Tests bestimmen.
Die Sinnhaftigkeit eines der Tests p kann mit Hilfe der Korrelation zwischen Erfolgsskala und der durch den Test gemachten Vorhersage beurteilt werden. Wenn der nächste Jahrgang von N_2 Studienanfängern den Test absolviert, stehen N_2 zusätzliche X–Werte zur Verfügung. Betrachtet man diese Werte als Stichproben einer $p+1$–variaten Verteilung (p psychologische Tests und die Erfolgsskala), so handelt es sich um unvollständige Daten, da das Abschneiden der Studenten am Ende des Semesters noch nicht bekannt ist. Es fehlen also Y–Werte.
Die zweite Gruppe von Stichproben stellt eine Basis für die Schätzung der Kovarianzen zwischen den Testergebnissen dar. Da die Kovarianzen in den Gleichungen für die Regressionskoeffizienten auftreten, stellt sich die Frage, ob man nicht die N_1 und N_2 Samples zusammennehmen könnte, um eine bessere Schätzung der Regressionskoeffizienten zu erhalten, als dies nur mit den N_1 Samplen möglich wäre.

Bei kontrollierten Experimenten, wie klinischen Studien in der Pharmakologie oder technischen Laboruntersuchungen, wird die X-Matrix durch gezielte Versuchsplanung festgelegt und die Systemantwort Y beobachtet. Die fehlenden Werte werden also eher in der Systemantwort Y, als im Versuchsplan auftreten. Somit ist es, selbst wenn für die Daten MCAR-Annahme gilt, vorteilhafter, mit einem aufgefüllten Y-Vektor die Standardanalyse balanzierter Modelle durchzuführen, als mit einem kleineren complete–case Datensatz zu arbeiten [Toutenburg(1992), 203]. Falls der Versuchsplan z.B. vollständig gekreuzt ist, würde die Beschränkung auf den complete–case Datensatz zu Schwierigkeiten bei der Interpretation führen.
Beschränkt man sich auf den bivariaten Fall und setzt eine unabhängige Normalverteilung der Werte von $\mathbf{X}$ und $\mathbf{Y}$, so kann man für die Momente der Verteilung folgendes ableiten [16, S. 84ff.]; [2, S. 200ff.]:
Seien x und y unabhängig voneinander normalverteilt und seien μ_x und μ_y die Mittelwerte, σ_x^2 und σ_y^2 die Varianzen und $\rho = \sigma_{xy}/(\sigma_x\sigma_y)$ der Korrelationskoeffizient, so kann man die Dichtfunktion als $n(x, y \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho)$ schreiben. Sowohl abhängige wie unabhängige Variable (x, y) werden in n Fällen beobachtet, nur die abhängige Variable (y) in $N - n$ Fällen.
Die zweidimensionale Verteilung von x und y kann man als Produkt der Grenzdichte von x und der bedingten Dichte von $y \mid x$ schreiben [12, S. 62]; [16, S. 84ff.]:

$$n(x, y \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho) = n(x \mid E(X), \sigma_x^2) \cdot n(y \mid E(Y \mid x), \sigma_{y|x}^2) = n(x \mid \mu_x, \sigma_x^2) \cdot n(y \mid \nu + \beta_{yx}x, \sigma_{y|x}^2)$$

Hierbei sind

$$E(Y \mid x) = \mu_y - \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \quad \sigma_{y|x}^2 = \sigma_y^2(1 - \rho^2) \tag{2}$$

woraus folgt:

$$\nu = \mu_y - \beta_{yx}\mu_x, \quad \beta_{yx} = \frac{\rho\sigma_y}{\sigma_x} \tag{3}$$

Somit kann die Likelihood-Funktion wie folgt geschrieben werden:

$$\prod_{i=1}^{n} n(x_i, y_i \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho) \prod_{i=n+1}^{N} n(x_i \mid \mu_x, \sigma_x^2) = \prod_{i=1}^{N} n(x_i \mid \mu_x, \sigma_x^2) \prod_{i=1}^{n} n(y_i \mid \nu + \beta_{yx}x_i, \sigma_{y|x}^2) \tag{1}$$

Die Maximum–Likelihood–Schätzung (ML–Schätzung) von μ_x, σ_x^2, ν, β_{yx} und $\sigma_{y|x}^2$ sind die Werte, die Formel (1) maximieren.
Ein Maximum in Bezug auf μ_x und σ_x^2 erhält man durch die Maximierung von $\prod_{i=1}^{N} n(x_i \mid \mu_x, \sigma_x^2)$. Hieraus erhält man die bekannten ML–Schätzer einer univariaten Normalverteilung von N Beobachtungen:

$$\hat{\mu}_x = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \hat{\sigma}_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})$$

Um (1) in Bezug auf ν, β_{yx} und $\sigma_{y|x}^2$ zu maximieren, muß man $\prod_{i=1}^{n} n(y_i \mid \nu + \beta_{yx}x_i, \sigma_{y|x}^2)$ maximieren. Man erhält die Regressionsparameter:

$$\hat{\nu} = \bar{y}^* - \hat{\beta}_{yx}\bar{x}^*$$

$$\hat{\beta}_{yx} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}^*)(x_i - \bar{x}^*)}{\sum_{i=1}^{n}(x_i - \bar{x}^*)^2}$$

$$\hat{\sigma}_{y|x}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}^*)^2 - \hat{\beta}_{yx}^2\sum_{i=1}^{n}(x_i - \bar{x}^*)^2}{n}$$

Wobei $\bar{x}^* = \frac{1}{n}\sum_{i=1}^n x_i$ und $\bar{y}^* = \frac{1}{n}\sum_{i=1}^n y_i$. Die ML–Schätzer für μ_y, σ_y^2 und ρ erhält man durch Auflösen der Gleichungen (2, 3)

$$\sigma_{y|x}^2 = \sigma_y^2(1-\rho^2), \quad \nu = \mu_y - \beta_{yx}\mu_x, \quad \beta_{yx} = \frac{\rho\sigma_y}{\sigma_x}$$

indem man $\nu = \hat{\nu}$, $\beta_{yx} = \hat{\beta}_{yx}$ und $\sigma_{y|x}^2 = \hat{\sigma}_{y|x}^2$ setzt. Also:

$$\hat{\mu}_y = \bar{y}^* + \hat{\beta}_{yx}(\bar{x} - \bar{x}^*)$$

$$\hat{\sigma}_y^2 = \hat{\sigma}_{y|x}^2 + \hat{\rho}^2\hat{\sigma}_y^2 = \hat{\sigma}_{y|x}^2 + \hat{\beta}_{yx}^2\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y}^*)^2 + \hat{\beta}_{yx}^2\left(\hat{\sigma}_x^2 - \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}^*)^2\right)$$

$$\hat{\rho} = \frac{\hat{\beta}_{yx}\hat{\sigma}_x}{\hat{\sigma}_y}$$

5 Regression bei fehlenden Y-Werten

5.1 Analysis of Variance (ANOVA)

Yates schlug für das Auffüllen der $N-n$ nicht beobachteten Y-Werte folgende Methode vor [32]: Man unterteilt den Datensatz in einen vollständigen, bei dem keine Y-Werte fehlen (Index c: complete), und einen unvollständigen (fehlende Werte durch Asterix $(*)$ gekennzeichnet) und sortiert — ohne Beschränkung der Allgemeinheit — um:

$$\begin{pmatrix} y_{obs} \\ y_{mis} \end{pmatrix} = \begin{pmatrix} X_c \\ X_* \end{pmatrix}\beta + \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix}$$

Man schätzt β aus dem vollständigen Submodell gemäß $b_c = (X_c'X_c)^{-1}X_c'y_{obs}$. Der fehlende Teil des Y-Vektors besteht aus $N-n$ Elementen, die entsprechend der klassischen Vorhersage $\hat{y}_{mis} = X_* b_c$ geschätzt werden. Für die Quadratsumme der Residuen kann man jetzt schreiben:

$$S(\beta) = \sum_{i=1}^N \epsilon_i^2 = \epsilon'\epsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \left(\begin{pmatrix} \mathbf{y}_{obs} \\ \hat{\mathbf{y}}_{mis} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_c \\ \mathbf{X}_* \end{pmatrix}\beta\right)' \left(\begin{pmatrix} \mathbf{y}_{obs} \\ \hat{\mathbf{y}}_{mis} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_c \\ \mathbf{X}_* \end{pmatrix}\beta\right)$$

Als Summe geschrieben erhält man:

$$S(\beta) = \sum_{i=1}^n (y_i - x_i'\beta)^2 + \sum_{i=n+1}^N (\hat{y}_i - x_i'\beta)^2$$

Der erste Summand $(\sum_{i=1}^n (y_i - x_i'\beta)^2)$ wird minimal für $\beta = b_c$. Der zweite Summand $(\sum_{i=n+1}^N (\hat{y}_i - x_i'\beta)^2)$ wird für $\beta = b_c$ Null, da $\hat{y}_{mis} = X_* b_c$. Somit ist b_c der KQ-Schätzer im nach Yates aufgefüllten Modell.
Die Schätzung für σ^2 ohne fehlende Daten lautet

$$\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Hierbei ist K die Anzahl der Spalten der $\mathbf{X}$–Matrix. Bei fehlenden Y–Werten berechnet sich der Schätzer zu

$$\hat{\sigma}_{mis}^2 = \frac{1}{n-K}\sum_{t=1}^n (y_i - \hat{y}_i)^2$$

Die Auffüllmethode nach Yates liefert

$$\hat{\sigma}_{Yates}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=n+1}^N (\hat{y}_i - \hat{y}_i)^2}{N-K} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{N-K}$$

Somit ist $\hat{\sigma}_{Yates}^2 < \hat{\sigma}_{mis}^2$, d.h. die Varianz ist verzerrt und wird bei der Methode nach Yates unterschätzt. Durch Multiplikation mit $(N-K)/(n-K)$ muß korrigiert werden:

$$\hat{\sigma}_{mis}^2 = \hat{\sigma}_{Yates}^2 \frac{N-K}{n-K} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-K}$$

Wendet man diese Formeln auf die Daten der Säugetiere aus obiger Tabelle an, so erhält man:

$$\hat{\sigma}_{Yates}^2 = \frac{1.056}{13-2} = 0.0960, \quad \hat{\sigma}_{mis}^2 = \frac{1.056}{12-2} = 0.1056$$

Wie bereits erwartet, wird mit der Methode nach Yates die Varianz unterschätzt.

5.2 Analysis of Covariance (ANCOVA)

Um den Korrekturfaktor für die Varianz in Yates´ ANOVA zu eleminieren, schlug Bartlett die folgende Methode vor, die als Bartlett´s ANCOVA bezeichnet wird [3, S. 147ff.]:

- jeder fehlende Wert wird durch eine beliebige Schätzung (guess) aufgefüllt: $y_{mis} \longrightarrow \hat{y}_{mis}$.
- es wird eine $N \times (N-n)$ –Indikationsmatrix $\mathbf{Z}$ als Kovariable eingeführt, die n–Nullzeilenvektoren für vollständige Beobachtungen und $N-n$ Vektoren e_i' für unvollständige.

Die Z–Matrix hat somit folgendes Aussehen:

$$Z = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Über die Kovariable wird ein zusätzlicher Parameter γ, ein $N-n \times 1$–Vektor eingeführt und mitgeschätzt:

$$\begin{pmatrix} y_{obs} \\ \hat{y}_{mis} \end{pmatrix} = X\beta + Z\gamma + \epsilon = (X, Z) \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \epsilon$$

Für die KQ–Schätzung von $\begin{pmatrix} \beta \\ \gamma \end{pmatrix}$ minimiert man:

$$S(\beta, \gamma) = \sum_{i=1}^{n} (y_i - x_i'\beta - 0'\gamma)^2 + \sum_{i=n+1}^{N} (\hat{y}_i - x_i'\beta - e_i'\gamma)^2$$

Der erste Summand wird für $\hat{\beta} = b_c$ minimal, der zweite wird für $\hat{\gamma} = \hat{y}_{mis} - X_* b_c$ Null. Somit ist $\begin{pmatrix} b_c \\ \hat{y}_{mis} - X_* b_c \end{pmatrix}$ KQ–Schätzung von $\begin{pmatrix} \beta \\ \gamma \end{pmatrix}$.
Wählt man entsprechend der Yates Methode $\hat{y}_{mis} = X_* b_c$, so wird $\hat{\gamma} = 0$. Die Einführung der Variable γ, an deren Wert man gar nicht interessiert ist, hat den Vorteil, daß die Zahl der Freiheitsgrade bei der Schätzung von σ^2 gleich N minus der Anzahl der geschätzten Parameter, also $N - K - (N-n) = n - K$, ist. Das heißt, wir erhalten einen unverzerrten Schätzer für σ^2:

$$\hat{\sigma}^2 = \hat{\sigma}^2_{mis} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - K}$$

6 X–Missing

Im Folgenden wird davon ausgegangen, daß ein Datensatz von N Beobachtungen von $K+1$ Variablen vorliegt und zwar in der Form, daß

$$f(Y \mid X_1, \dots, X_K) \sim N(\beta_0 + \beta_1 X_1 + \dots + \beta_K X_K, \sigma_\epsilon^2)$$

Um die klassische Regressionsfunktion (siehe oben) so zu erweitern, daß sie auch in den Fällen gilt, in denen in den unabhängigen Variablen Datenwerte fehlen, führt man folgende Bezeichnungen ein:
Sei W_{ij} eine Indikatorfunktion, so daß $W_{ij} = 1$ falls X_{ij} beobachtet wurde und $W_{ij} = 0$ falls X_{ij} nicht beobachtet wurde. Man geht davon aus, daß die fehlenden Daten zufällig, also unabhängig voneinander fehlen. Hieraus folgt, daß die gemeinsame Verteilung irgendeiner Menge von Ws das Produkt der Wahrscheinlichkeit der einzelnen Indikatorfunktionen ist. Die Zahl der Daten für die X_i beobachtet wurde, kann man wie folgt schreiben:

$$N_j = \sum_{i=1}^{N} W_{ij}$$

Die Zahl der Datensätze, in denen sowohl X_j, als auch X_k beobachtet wurden, ergibt sich zu:

$$N_{jk} = \sum_{i=1}^{N} W_{ij} W_{ik}$$

Der Mittelwert von X_j, der auch den beobachteten Werten von X_j basiert ergibt sich zu

$$\bar{X}_{j(j)} = \frac{1}{N_j}\sum_{i=1}^{N} W_{ij}X_{ij} = \frac{\sum_{i=1}^{N} W_{ij}X_{ij}}{\sum_{i=1}^{N} W_{ij}}$$

Der Mittelwert von X_j, der auf den Werten von X_j basiert, für die sowohl X_j, als auch X_k beobachtet wurden, errechnet sich zu

$$\bar{X}_{j(jk)} = \frac{1}{N_{jk}}\sum_{i=1}^{N} W_{ij}W_{ik}X_{ij} = \frac{\sum_{i=1}^{N} W_{ij}W_{ik}X_{ij}}{\sum_{i=1}^{N} W_{ij}W_{ik}}$$

Der Mittelwert von Y für die Daten, bei denen X_j beobachtet wurde, ist

$$\bar{Y}_{(j)} = \frac{1}{N_j}\sum_{i=1}^{N} W_{ij}Y_{i} = \frac{\sum_{i=1}^{N} W_{ij}Y_{i}}{\sum_{i=1}^{N} W_{ij}}$$

Durch Analogschluß mit dem linearen Regressionsmodell ohne fehlende Daten erhält man:

$$\hat{\hat{\sigma}}_{jk} = \frac{\sum_{i=1}^{N} W_{ij}W_{ik}(X_{ij} - \bar{X}_{j(jk)})(X_{ik} - \bar{X}_{k(jk)})}{N_{jk} - 1}, \quad j,k \neq 0$$

$$\hat{\hat{\sigma}}_{j0} = \frac{\sum_{i=1}^{N} W_{ij}(X_{ij} - \bar{X}_{j(j)})(Y_{i} - \bar{Y}_{(j)})}{N_j - 1}$$

Setzt man für $\hat{\sigma}_{jk}$ $\hat{\hat{\sigma}}_{jk}$ und für $\hat{\sigma}_{j0}$ $\hat{\hat{\sigma}}_{j0}$, so kann man für die Regressionskoeffizienten schreiben:

$$\hat{\hat{\beta}}_j = \sum_{k=1}^{K} \frac{\hat{\hat{\sigma}}_{k0}}{\hat{\hat{\sigma}}_{jk}}, \quad j \neq 0$$

$$\hat{\hat{\beta}}_0 = \bar{Y} - \sum_{j=1}^{K} \hat{\hat{\beta}}_j \bar{X}_{j(j)}$$

Hieraus errechnet sich die Kovarianzmatrix zu [11, S. 837ff.]:

$$Cov(\hat{\hat{\beta}}_j, \hat{\hat{\beta}}_k) = \sigma_\epsilon^2 \sum_{l=1}^{K}\sum_{m=1}^{K} \frac{\hat{\hat{\sigma}}_{lm}}{\hat{\hat{\sigma}}_{jl}\hat{\hat{\sigma}}_{km}} \frac{N_{lm} - 1}{(N_l - 1)(N_m - 1)}, \quad j,k \neq 0$$

Für die Schätzung von $Cov(\hat{\hat{\beta}}_j, \hat{\hat{\beta}}_k)$ braucht man eine Schätzung für σ_ϵ^2, die man durch Analogschluß zum Fall ohne fehlende Daten erhält:

$$\hat{\hat{\sigma}^2}_\epsilon = \frac{N-1}{(N-1)-K}(\hat{\hat{\sigma}}_{00} - \sum_{j=1}^{K} \hat{\hat{\beta}}_j \hat{\hat{\sigma}}_{j0}), \quad \hat{\hat{\sigma}}_{00} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1}$$

6.1 Der Fall K=2

Beschränkt man sich auf den Fall $K = 2$, so erhält man für den Fall ohne fehlende Daten:

$$V(\hat{\beta}_1) = \frac{\sigma_\epsilon^2 \hat{\sigma}_{22}}{(N_{12}-1)\hat{D}}, \quad V(\hat{\beta}_2) = \frac{\sigma_\epsilon^2 \hat{\sigma}_{11}}{(N_{12}-1)\hat{D}}, \quad Cov(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma_\epsilon^2 \hat{\sigma}_{12}}{(N_{12}-1)\hat{D}}$$

Wobei $\hat{D} = \hat{\sigma}_{11}\hat{\sigma}_{22} - \hat{\sigma}_{12}^2$. Es wurde $(N_{12} - 1)$ und nicht $(N - 1)$ benutzt, da die Schätzung nur auf vollständigen Daten beruht.
Bei fehlenden Daten in der X–Matrix erhält man für $K = 2$ die folgenden Gleichungen:

$$\hat{\hat{\beta}}_1 = \frac{\hat{\hat{\sigma}}_{22}\hat{\hat{\sigma}}_{10} - \hat{\hat{\sigma}}_{12}\hat{\hat{\sigma}}_{20}}{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{22} - \hat{\hat{\sigma}}_{12}\hat{\hat{\sigma}}_{12}} = \frac{\hat{\hat{\sigma}}_{22}\hat{\hat{\sigma}}_{10} - \hat{\hat{\sigma}}_{12}\hat{\hat{\sigma}}_{20}}{\hat{\hat{D}}}$$

$$\hat{\hat{\beta}}_2 = \frac{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{20} - \hat{\hat{\sigma}}_{12}\hat{\hat{\sigma}}_{10}}{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{22} - \hat{\hat{\sigma}}_{12}\hat{\hat{\sigma}}_{12}} = \frac{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{20} - \hat{\hat{\sigma}}_{12}\hat{\hat{\sigma}}_{10}}{\hat{\hat{D}}}$$

Hierbei ist $\hat{\hat{D}} = \hat{\sigma}_{11}\hat{\sigma}_{22} - \hat{\sigma}^2{}_{12}$. Die Varianzen und Covarianzen erhält man aus obiger Formel für die Covarianz:

$$V(\hat{\hat{\beta}}_1) = \frac{\sigma_\epsilon^2 \hat{\hat{\sigma}}_{22}}{\hat{\hat{D}}^2(N-1)} \left(\frac{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{22}}{P_1} + \hat{\sigma}^2{}_{12}(\frac{1}{P_2} - \frac{2P_{12}}{P_1 P_2}) \right)$$

$$V(\hat{\hat{\beta}}_2) = \frac{\sigma_\epsilon^2 \hat{\hat{\sigma}}_{11}}{\hat{\hat{D}}^2(N-1)} \left(\frac{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{22}}{P_2} + \hat{\sigma}^2{}_{12}(\frac{1}{P_1} - \frac{2P_{12}}{P_1 P_2}) \right)$$

$$Cov(\hat{\hat{\beta}}_1, \hat{\hat{\beta}}_2) = -\frac{\sigma_\epsilon^2 \hat{\hat{\sigma}}_{12}}{\hat{\hat{D}}^2(N-1)} \left(\frac{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{22}}{P_1 P_2} - \hat{\sigma}^2{}_{12}\frac{P_{12}}{P_1 P_2} \right)$$

Hierbei sind

$$P_1 = \frac{N_1 - 1}{N-1}, \quad P_2 = \frac{N_2 - 1}{N-1}, \quad P_{12} = \frac{N_{12} - 1}{N-1} = P_1 + P_2 - 1$$

Es ist zu beachten, daß $P_1 + P_2 > 1$, da wenigstens eins der X für jeden jeweiligen Datenwert beobachtet wurde.
Für große Datensätze führt man als Effektivität für $\hat{\hat{\beta}}_i$ im Verhältnis zu $\hat{\beta}_i$ folgenden Quotienten ein:

$$\mathit{Eff}(\hat{\hat{\beta}}_i) = \frac{V(\hat{\beta}_i)}{V(\hat{\hat{\beta}}_i)}$$

Im Falle $K = 2$ erhält man durch einfaches Einsetzen:

$$\mathit{Eff}(\hat{\hat{\beta}}_1) = \frac{1}{P_{12}} \left(\frac{P_1 P_2}{P_2 + \frac{r^2}{1-r^2}(1 - P_{12})} \right)$$

$$\mathit{Eff}(\hat{\hat{\beta}}_2) = \frac{1}{P_{12}} \left(\frac{P_1 P_2}{P_1 + \frac{r^2}{1-r^2}(1 - P_{12})} \right)$$

Die Variable r wird als Populationskoeffizient der Korrelation zwischen X_1 und X_2 bezeichnet und ist definiert durch:

$$r = \frac{\hat{\sigma}_{12}^2}{\hat{\sigma}_{11}\hat{\sigma}_{22}}$$

Wie man leicht sieht, werden die Effektivitäten von $\hat{\hat{\beta}}_1$ und $\hat{\hat{\beta}}_2$ am größten, wenn $r = 0$. In diesem Falle gilt:

$$\mathit{Eff}(\hat{\hat{\beta}}_1) = \frac{P_1}{P_{12}}, \quad \mathit{Eff}(\hat{\hat{\beta}}_2) = \frac{P_2}{P_{12}}$$

Mit steigendem r werden die Effektivitäten kleiner.
Von besonderem Interesse ist der Fall, in dem $\mathit{Eff}(\hat{\hat{\beta}}_i) \geq 1$. Setzt man diese Bedingung in die Gleichung für $\mathit{Eff}(\hat{\hat{\beta}}_i)$ ein, so erhält man:

$$r^2 \leq \frac{P_2(1 - P_2)}{P_{12}(1 - P_{12}) + P_2(1 - P_2)}$$

$$r^2 \leq \frac{P_1(1 - P_1)}{P_{12}(1 - P_{12}) + P_1(1 - P_1)}$$

Zu beachten ist, daß P_1 und P_2 nicht gleichzeitig Null sein dürfen, da sonst der Nenner in obigen Ausdrücken Null wird.
Für den Fall $P_1 = P_2 = P$ erhält man:

$$r^2 \leq \frac{P}{5P - 2}, \quad (0.5 < P < 1)$$

Hieraus liest man ab, daß immer wenn $r^2 < 1/3$ und $P_1 = P_2 = P$, $\mathit{Eff}(\hat{\hat{\beta}}_i) > 1$. Selbstverständlich gibt es auch in diesem Fall Werte von r^2 die Größer als $1/3$ sind, bei denen immer noch $\mathit{Eff}(\hat{\hat{\beta}}_i) > 1$ gilt, sofern $P < 1$.

In den meisten Fällen interessiert nicht primär die Effizienz der Schätzer für die Regressionskoeffizienten, sondern die Effizienz der Schätzung für den Output Y. Für den Fall ohne fehlende Daten erhält man:

$$\hat{Y} = \bar{Y} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2, \quad x_j = X_j - \bar{X}_j$$

$$V(\hat{Y}) = V(\bar{Y}) + x_1^2 V(\hat{\beta}_1) + x_2^2 V(\hat{\beta}_2) + 2x_1 x_2 Cov(\hat{\beta}_1, \hat{\beta}_2), \quad da \;\; Cov(\bar{Y}, \hat{\beta}_j) = 0$$

Einsetzen der entsprechenden Definitionsgleichungen liefert:

$$V(\hat{Y}) = \frac{\sigma_\epsilon^2}{N_{12}-1}\left(1 + \frac{x_1^2\hat{\sigma}_{22}}{\hat{D}} + \frac{x_2^2\hat{\sigma}_{11}}{\hat{D}} - \frac{x_1x_2\hat{\sigma}_{12}}{\hat{D}}\right)$$

Im Fall mit fehlenden Daten sei $\hat{\hat{Y}}$ die Vorhersage für den Schätzer von Y. Man erhält:

$$\hat{\hat{Y}} = \bar{Y} + \hat{\hat{\beta}}_1 x_{1(1)} + \hat{\hat{\beta}}_2 x_{2(2)}, \quad x_{j(j)} = X_j - \bar{X}_{j(j)}$$

Nach [11, S. 839] ergibt sich für die Varianz:

$$V(\hat{\hat{Y}}) = V(\bar{Y}) + x_{1(1)}^2 V(\hat{\hat{\beta}}_1) + x_{2(2)}^2 V(\hat{\hat{\beta}}_2) + 2x_{1(1)}x_{2(2)}Cov(\hat{\hat{\beta}}_1, \hat{\hat{\beta}}_2), \quad da \lim_{N_j\to\infty}\left(Cov(\bar{Y}, \hat{\hat{\beta}}_j)\right) = 0$$

Einsetzen der Definitionsgleichungen liefert:

$$V(\hat{\hat{Y}}) = \frac{\sigma_\epsilon}{N-1}\left(1 + \frac{x_{1(1)}^2\hat{\hat{\sigma}}_{22}}{\hat{D}^2}\left(\frac{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{22}}{P_1} + \sigma_{12}^2\left(\frac{1}{P_2} - \frac{2P_{12}}{P_1P_2}\right)\right) + \frac{x_{2(2)}^2\hat{\hat{\sigma}}_{11}}{\hat{D}^2}\left(\frac{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{22}}{P_2} + \sigma_{12}^2\left(\frac{1}{P_1} - \frac{2P_{12}}{P_1P_2}\right)\right) - \right.$$

$$\left. \frac{2x_{1(1)}x_{2(2)}\hat{\hat{\sigma}}_{12}}{\hat{D}^2}\left(\frac{\hat{\hat{\sigma}}_{11}\hat{\hat{\sigma}}_{22}}{P_1P_2} - \sigma_{12}^2\frac{P_{12}}{P_1P_2}\right)\right)$$

Die relative Effektivität von $\hat{\hat{Y}}$ zu $\hat{Y}$ sei definiert durch:

$$\mathit{Eff}(\hat{\hat{Y}}) = \frac{V(\hat{Y})}{V(\hat{\hat{Y}})}$$

Nach Einsetzen der entsprechenden Gleichungen erhält man:

$$\mathit{Eff}(\hat{\hat{Y}}) = \frac{P_1P_2((1-r^2) + Z_1^2 + Z_2^2 - 2rZ_1Z_2)}{P_{12}\left(P_1P_2(1-r^2) + Z_1^2\left(P_2 + (\frac{r^2}{1-r^2})(1-P_{12})\right) + Z_2^2\left(P_1 + (\frac{r^2}{1-r^2})(1-P_{12})\right) - 2rZ_1Z_2\frac{(1-P_{12}r^2)}{1-r^2}\right)}$$

Mit $Z_i^2 = x_i^2/\sigma_{ii}$. Die Effizienz wird maximiert für $Z_1 = Z_2 = 0$, also wenn die Schätzung durch Regression benutzt wird, um Y bei $\bar{X}_1$ und $\bar{X}_2$ zu schätzen. In diesem Punkt ist $\mathit{Eff}(\hat{\hat{Y}}) = 1/P_{12}$.
Wenn $Z_1 = Z_2$ und beide im Verhältnis zu $(1-r^2)$ groß sind, so ist

$$\mathit{Eff}(\hat{\hat{Y}}) = \frac{2P_1P_2(1+r)}{P_{12}(P_{12}(1+2r)+1)}$$

Dieser Ausdruck ist größer als 1, wenn

$$r \geq -\frac{1}{2} - \frac{1}{2}\left(\frac{P_{12} - P_1P_2}{P_{12}^2 - P_1P_2}\right)$$

Damit der Nenner nicht Null wird, dürfen wieder P_1 und P_2 nicht beide gleich 1 sein. $\mathit{Eff}(\hat{\hat{Y}}) \geq 1$ gilt immer dann, wenn $r \geq -0.5$.

Für den Fall $K = 3$ hat Edgett [8] einen maximum Likelihoodschätzer für eine trivariate normalverteilte Population mit fehlenden Daten in einer Variablen abgeleitet. Auf eine ausführliche Darstellung wird hier aus Platzgründen verzichtet.

6.2 Fehlende X–Werte in Matrixschreibweise

Man führt folgende Bezeichnung ein:

$$\begin{pmatrix} y_c \\ y_* \end{pmatrix} = \begin{pmatrix} X_c \\ X_* \end{pmatrix}\beta + \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix}, \quad \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix} \sim (0, \sigma^2 I)$$

Das Submodell ohne fehlende Daten ist $y_c = X_c\beta + \epsilon_c$, wobei $Rang(X_c) = K$. Das Submodell mit fehlenden Daten ist $y_* = X_*\beta + \epsilon_*$. Hierbei ist zu beachten, daß y_* vollständig beobachtet ist, X_* jedoch unvollständig, aber es gibt in X_* noch Beobachtungen ($X_* \neq X_{mis}$, wobei in X_{mis} keine Beobachtungen vorliegen (mis: missing)).
Bei diesem aus zwei Submodellen zusammengesetzten Modell handelt es sich um ein mixed Modell, (auch Aitken Modell genannt) [25, S. T736]; [27, S. 207]. Wesentliche Voraussetzung für die folgende Lösung ist die Unkorreliertheit der beiden Zufallsprozesse, also:

$$E(\epsilon_c\epsilon_*') = 0$$

Somit gilt:

$$E\begin{pmatrix}\epsilon_c\\ \epsilon_*\end{pmatrix}(\epsilon_c,\epsilon_*)' = \sigma^2\begin{pmatrix}W & 0\\ 0 & I\end{pmatrix}$$

Folgender Satz liefert die Lösung für das mixed Modell [5, S. 319]; [27, S. 142]:
Im Mixed Modell

$$\begin{pmatrix}y_c\\ y_*\end{pmatrix} = \begin{pmatrix}X_c\\ X_*\end{pmatrix}\beta + \begin{pmatrix}\epsilon_c\\ \epsilon_*\end{pmatrix}, \quad \begin{pmatrix}\epsilon_c\\ \epsilon_*\end{pmatrix} \sim (0, \sigma^2 I)$$

hat die beste lineare erwartungstreue Schätzung von β die Gestalt

$$\hat{\hat{\beta}}(X_*) = (X_c'X_c + X_*'X_*)^{-1}(X_c'y_c + X_*'y_*) = b_c + S_c^{-1}X_*'(I_n + X_*S_c^{-1}X_*')^{-1}(y_* - X_*b_c)$$

mit

$$V(\hat{\hat{\beta}}(X_*)) = \sigma^2(S_c + S_*)^{-1}$$

Hierbei sind

$$b_c = (X_c'X_c)^{-1}X_c'y_c, \quad S_c = X_c'X_c, \quad S_* = X_*'X_*$$

Ein Beweis findet sich in [24]. Es sei angemerkt, daß b_c der Schätzer für das vollständige Modell ist. Für den Fall, daß σ^2 nicht bekannt, so kann man es schätzen:

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{N}(y_t - \bar{y})^2}{N-K} = \frac{\begin{pmatrix}y_c\\ y_*\end{pmatrix}'\begin{pmatrix}y_c\\ y_*\end{pmatrix} - \left(\begin{pmatrix}E\end{pmatrix}'\begin{pmatrix}y_c\\ y_*\end{pmatrix}\right)^2}{N-K}$$

Wobei E ein N–Zeilen Einsvektor ist.

Die relative Effizienz eines Schätzers zu einem anderen kann man im Gegensatz zu obiger Definition von *Eff* auch über den mittleren quadratischen Fehler definieren [25, S. T736]:

$$MSE(\hat{\beta}) = E((\hat{\beta}-\beta)(\hat{\beta}-\beta)') = Cov(\hat{\beta}) + (bias(\hat{\beta}))(bias(\hat{\beta}))'$$

Seien $\hat{\beta}_1$ und $\hat{\beta}_2$ zwei gegebene Schätzer von β, dann ist $\hat{\beta}_2$ ein besserer Schätzer, dann und nur dann, wenn

$$MSE(\hat{\beta}_1) - MSE(\hat{\beta}_2) \geq 0$$

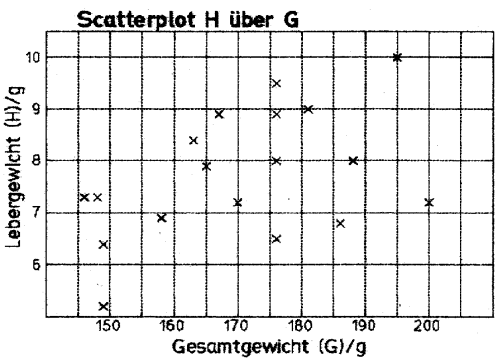
Ein Effizienzkriterium (*eff*)kann man daraus durch Division mit $MSE(\hat{\beta}_2)$ erhalten: $\hat{\beta}_2$ ist ein besserer Schätzer, dann und nur dann, wenn:

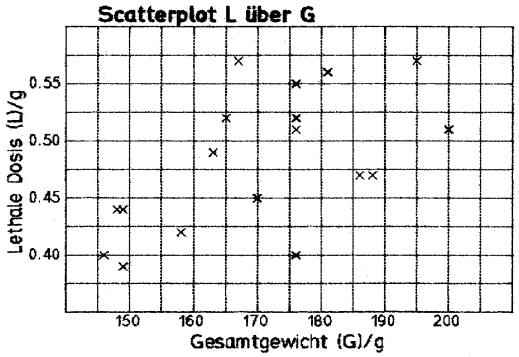
$$eff(\hat{\beta}_1, \hat{\beta}_2) = \frac{MSE(\hat{\beta}_1)}{MSE(\hat{\beta}_2)} \geq 1$$

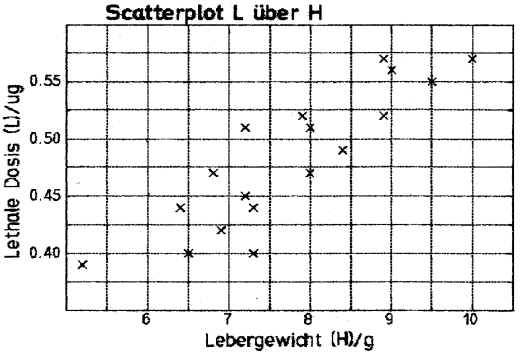
Durch das Einführen von Wichtungen in den unvollständigen Daten kann man das Mixed–Modell zum Weighted–Mixed–Modell erweitern [25, S. T736f.].

Als Beispiel sei eine pharmakologische Untersuchung angeführt, bei der es darum geht, die Dosis eines Pharmakons festzustellen, die ein Versuchstier nicht mehr überlebt. Normalerweise werden diese Dosen auf das Gesamtgewicht des Versuchtstiers bezogen angegeben. Dies ist nicht immer sinnvoll, da die meisten pharmakologischen Substanzen in der Leber abgebaut werden. Die Effizienz des Abbaus ist also eher eine Funktion des Lebergewichtes. Somit müßten die Dosen auf das Lebergewicht bezogen angegeben werden. Ganz ohne Einfluß ist das Körpergewicht bei der Bestimmung der tötlichen Dosis jedoch nicht, da bei größerem Körpergewicht mehr Flüssigkeit zum Verdünnen der applizierten Substanz vorhanden ist, die Konzentration im Körper also geringer ist. Die Rohdaten sind in der folgenden Tabelle zusamengefaßt. Die fehlenden Daten sind aufgrund eines experimentellen Fehlers zustandegekommen: Die Leber wurde ohne Wiegen weiterverarbeitet.

Ratte Nummer	Gesamt-gewicht (G) [g]	Leber-gewicht (H) [g]	Lethal-Dose (L) [μg]	
(1)	176.0	6.5	0.40	
(2)	176.0	9.5	0.55	
(3)	176.0	8.9	0.52	
(4)	200.0	7.2	0.51	
(5)	167.0	8.9	0.57	
(6)	188.0	8.0	0.47	
(7)	195.0	10.0	0.57	
(8)	176.0	8.0	0.51	
(9)	165.0	7.9	0.52	
(10)	158.0	6.9	0.42	
(11)	148.0	7.3	0.44	
(12)	149.0	5.2	0.39	
(13)	163.0	8.4	0.49	
(14)	170.0	7.2	0.45	
(15)	186.0	6.8	0.47	
(16)	146.0	7.3	0.40	
(17)	181.0	9.0	0.56	
(18)	149.0	6.4	0.44	
				lin. Reg. Lebergew. ($\hat{H}$) [g]
(19)	190.0	x	0.53	8.554
(20)	179.0	x	0.50	8.050
(21)	147.0	x	0.47	7.624
(22)	153.0	x	0.44	7.057
(23)	157.0	x	0.42	6.679
(24)	160.0	x	0.53	8.665
(25)	188.0	x	0.49	7.836







Für die Korrelationskoeffizienten erhält man:

$$corr(G,H) = 0.47, \quad corr(G,L) = 0.57, \quad corr(H,L) = 0.87$$

Aufgrund der Scatterplot und der Korrelationskoeffizienten kann man festhalten, daß die maximale Dosis mehr vom Lebergewicht, als von Körpergewicht abhängt. Weiterhin ist die Abhängigkeit von Körpergewicht zu Lebergewicht eher gering.
Für den complete–case Fall errechnet man:

$$\hat{\beta}(X_c) = \begin{pmatrix} 4.842 \cdot 10^{-2} \\ 8.318 \cdot 10^{-4} \\ 3.770 \cdot 10^{-2} \end{pmatrix}, \quad V(\hat{\beta}(X_c)) = \begin{pmatrix} 5.349 \cdot 10^{-3} & -2.781 \cdot 10^{-5} & -7.269 \cdot 10^{-5} \\ -2.781 \cdot 10^{-5} & 2.274 \cdot 10^{-7} & -1.416 \cdot 10^{-6} \\ -7.269 \cdot 10^{-5} & -1.416 \cdot 10^{-6} & 4.056 \cdot 10^{-5} \end{pmatrix}, \quad V(\hat{Y}) = 8.071 \cdot 10^{-4}$$

Für die Analyse mit fehlenden Daten muß man die fehlenden Daten durch eine geeignete Imputation auffüllen. In der Tabelle mit den Rohdaten sind bereits die durch lineare Regression geschätzten Werte angegeben. Mit ihnen erhält man:

$$\hat{\hat{\beta}}(X_*) = \begin{pmatrix} 5.682 \cdot 10^{-2} \\ 6.501 \cdot 10^{-4} \\ 4.058 \cdot 10^{-2} \end{pmatrix}, \quad V(\hat{\hat{\beta}}(X_*)) = \begin{pmatrix} 1.474 \cdot 10^{-2} & -7.039 \cdot 10^{-5} & -3.437 \cdot 10^{-4} \\ -7.039 \cdot 10^{-5} & 6.063 \cdot 10^{-7} & -4.192 \cdot 10^{-6} \\ -3.437 \cdot 10^{-4} & -4.192 \cdot 10^{-6} & 1.361 \cdot 10^{-4} \end{pmatrix}, \quad V(\hat{\hat{Y}}) = 6.038 \cdot 10^{-4}$$

Die Effektivität *Eff* berechnet sich zu

$$Eff(\hat{\hat{Y}}) = \frac{V(\hat{Y})}{V(\hat{\hat{Y}})} = \frac{8.071 \cdot 10^{-4}}{6.038 \cdot 10^{-4}} = 1.337$$

Die Schätzung unter Berücksichtigung der fehlenden Daten ist also unter dem gewählten Effektivitätskriterium (Verhältnis der Varianzen) Effektiver, als das ohne Berücksichtigung der fehlenden Daten.

6.3 Standardverfahren bei unvollstängiger X–Matrix

Zum Auffüllen der fehlenden Werte der X–Matrix können selbstverständlich die im Abschnitt 4.3 (Imputationen für fehlende Daten) gemachten Imputationsarten benutzt werden. Es muß nur für den jeweiligen Datensatz, bzw. jede Datenspalte die Sinnhaftigkeit der jeweiligen Imputationsart geprüft werden.
Man kann sich auch auf eine complete–case Analysis beschränken, also in der Regressionsberechnung nur die Datensätze ohne fehlende Samples berücksichtigen. Voraussetzung hierfür ist allerdings, daß der Prozentsatz der vollständigen Daten im Verhältnis zu den Vollständigen nicht zu groß ist, und daß die Daten MAR sind, also keine Schichtungseffekte auftreten. Das Regressionsmodell reduziert sich dann zu:

$$\hat{\beta} = S_c^{-1} X_c' y_c = (X_c' X_c)^{-1} X_c' y_c, \quad V(\hat{\beta}_c) = \sigma^2 S_c^{-1} = \sigma^2 (X_c' X_c)^{-1}$$

7 Abschluß

Die in dieser Arbeit zusammengestellten Methoden zur Behandlung von Datensätzen mit fehlenden Daten wurden durch nicht immer optimalen Bespiele veranschaulicht. Dies liegt zum einen daran, daß die meisten Methoden ohne Beispiele veröffentlicht wurden, zum anderen, daß wenn Beispiele angeführt wurden, diese so unvollständig dargestellt wurden, daß man den Rechengang nicht verfolgen kann. Um das geschilderte dennoch mit anschaulichen Daten zu unterlegen, wurde Daten aus anderen Veröffentlichungen auf die darzustellenden Methoden übertragen. (Nennenswehrte Ausnahmen sind Weisberg [29] und Woolson [31]. Dies konnte nicht immer ohne einen Verlust an Klarheit geschehen. Dies bitte ich zu entschuldigen.
Zum Abschluß eine Warnung: Trauen Sie niemals einem Coputerprogramm, wie man mit den fehlenden Daten umzugehen hat. Der Programmierer des Programms konnte beim Erstellen die Einzelheiten Ihres Problems gar nicht kennen und die Default–Analysen des Programms sind mit Sicherheit für Ihre Daten nicht adäquat.

Literatur

[1] Missing observations in multivariate statistics (I: Review of the literature)
Afifi, A.A.; Elashoff, R.M. (1966)
Journal of the American Statistical Association; Volume 61; Page 595–604

[2] Maximum Likelihood Estimates for a Multivariate Normal Distribution when some Observations are missing
Anderson, T.W. (Columbia University) (1957)
Journal of the American Statistical Association; Volume 52; Page 200–204

[3] Some Examples of Statistical Methods of Research in Agriculture and Applied Biology
Bartlett, M.S. (ICI) (1937)
Journal of the Royal Statistical Society (B); Volume 4 No. 2; Page 137–183 (discussion on paper page 170–183)

[4] Alternative Methods of Regression
Birkes, David; Dodge, Yadolah (1993)
John Wiley & Sons; ISBN 0–471–56881–3

[5] The Use of Incomplete Observations in Multiple Regression Analysis (A Generalized Least Squares Approach)
Dagenais, Marcel G. (Universite de Montreal) (1973)
Journal of Econometrics; Volume 1; Page 317–328

[6] Maximum Likelihood from Incomplete Data via the EM Algorithm
Dempster, A.P.; Laird, N.M.; Rubin, D.B. (Harvard University) (1977)
Journal of the Royal Statistical Society (B); Volume 39 No. 1; Page 1–38 (discussion on paper page 22–38)

[7] Analysis of Experiments with Missing Data
Dodge, Yadolah (University of Neuchatel) (1985)
John Wiley & Sons; ISBN 0–471–88736–6

[8] Multiple Regression with Missing Observations Among the Independent Variables
Edgett, George L. (Queen´s University) (1956)
Journal of the American Statistical Association; Volume 51; Page 122–131

[9] Stein´s Paradox in Statistics
Efron, Bradley (Stanford University; Morris, Carl (Rand Corp.) (1979)
Scientific American; Volume 236 No. 5; Page 119–127, 148

[10] Pharmakologie und Toxikologie
Forth, Wolfgang; Henschler, Dietrich; Rummel, Walter; Starke, Klaus (1992)
BI–Wissenschaftsverlag (6. Auflage); ISBN 3–411–15026–2

[11] Linear Regression Analysis with Missing Observations Among the Independent Variables
Glasser, M. (Harvard School of Public Health) (1964)
Journal of the American Statistical Association; Volume 51; Page 834–844

[12] Formeln und Tabellen der angewandten mathematischen Statistik
Graf, Henning, Stange, Wilrich (1987)
Springer–Verlag; ISBN 3–540–16901–6

[13] Statistik Grundkurs (Kurs 0055/0056)
Gruber, Josef; Schwarze, Jochen; Fuskova, Lidmila; Kunitz, Harald (1984/1993)
FernUniversität Hagen

[14] Missing Data in Regression Analysis
Haitovsky, Yoel (Technion, Israel Institute of Technology) (1968)
Journal of the Royal Statistical Society (B); Volume 30 No. 1; Page 67–82

[15] Introduction to Mathematical Statistics
Hoel, Paul G. (UCLA) (1984)
John Wiley & Sons (5th Edition); ISBN 0–471–80530–0

[16] Distributions in Statistics. Continuous Multivariate Distributions
Johnson, N.L.; Kotz, S. (1972)
John Wiley & Sons; ISBN 0–471–44370–0

[17] Junge, Mirko (1989,1994)
A Statistic Suite for the HP28/HP48

[18] Statistical Analysis with Missing Data
Little, Roderick J.A. (UCLA); Rubin, Donald B. (Harvard University) (1987)
John Wiley & Sons; ISBN 0–471–80254–9

[19] New Cambridge Elementary Statistical Tables
Lindley, D.V.; Scott, W.F. (1984)
Cambridge University Press; ISBN 0–521–26922–9

[20] Standardization of risk ratios
Miettinen, O.S. (1972)
American Journal of Epidemology; Volume 96

[21] Estimation of Parameters from Incomplete Multivariate Samples
Nicholson, George E. (University of North Carolina) (1957)
Journal of the American Statistical Association; Volume 52; Page 523–526

[22] Allgemeine und spezielle Pathologie
Riede, Ursus–Nikolaus; Schaefer, Hans–Eckart (1993)
Georg Thieme Verlag (3. Auflage); ISBN 3–13–683303–9

[23] Multiple Imputation for Non–Response in Surveys
Rubin, Donald B. (1987)
John Wiley & Sons; ISBN 0–471–08705–x

[24] Less sensitive Tests by introducing stochastic linear hypotheses
Schaffrin, B. (The Ohio State University) (1987)
Proceedings of the Second International Tampere Conference in Statistics; Page 647–664

[25] Weighted Mixed Regression
Schaffrin, B. (The Ohio State University); Toutenburg, Helge (Universität Dortmund) (1990)
Zeitschrift angewante Mathematische Mechanik (ZAMM); Volume 70; Page T735–T738

[26] Ökonometrie
Schneeweiß, Hans (1990)
Physica Verlag (4. Auflage); ISBN 3–7908–0486–x

[27] Lineare Modelle
Toutenburg, Helge (Universität München) (1992)
Physica Verlag; ISBN 3–7908–0641–2

[28] Datenverlust bei klinischen Studien
Walther, Winfried (Akademie für zahnärztliche Fortbildung); Toutenburg, Helge (Universität Regensburg, Lehrstuhl für Statistik) (1991)
Deutsche Zahnärztliche Zeitung; Jahrgang 46; Seiten 219–222

[29] Applied Linear Regression
Weisberg, Sanford (University of Minesota) (1980)
John Wiley & Sons; ISBN 0–471–04419–9

[30] Introductory Statistics
Wonnacott, Thomas H.; Wonnacott, Ronald J. (University of Western Ontario) (1990)
John Wiley & Sons (5th Edition); ISBN 0–471–51733–X

[31] Statistical Methods for the Analysis of Biomedical Data
Woolson, R.F. (1987)
John Wiley & Sons; ISBN 0–471–80615–3

[32] The analysis of replicated experiments when the field results are incomplete
Yates, F. (1933)
Emp. Journal of Experimental Agriculture; Volume 1; Page 129–142