

Biometrie

**Für Studierende der
Agrarbiologie und
der Agrarwissenschaften (B. Sc.)**

an der Universität Hohenheim

4./6. Semester

Prof. Dr. Hans-Peter Piepho

Institut für Kulturpflanzenwissenschaften (340)
FG Biostatistik
Universität Hohenheim
e-mail: piepho@uni-hohenheim.de

Hohenheim, im März 2016

Copyright: Hans-Peter Piepho
Nur für den internen Gebrauch, Vervielfältigung nur mit Genehmigung des Autors

A 4002

Glossar griechischer Buchstaben

Buchstabe	Umschrift	häufige Verwendung
α	Alpha	Fehler 1. Art/Achsenabschnitt
β	Beta	Fehler 2. Art/Steigung
χ	Chi	Chi-Quadrat-Verteilung
δ	Delta	Lack-of-fit Effekt
γ	Gamma	Effekt im linearen Modell
η	Eta	Linearer Prädiktor, Erwartungswert, systematischer Teil des linearen Modells
λ	Lambda	Kontrast, schätzbare Funktion
μ	Mü	Mittelwert, Erwartungswert
π	Pi	Kreiskonstante/Binomialwahrscheinlichkeit
θ	Theta	Parameter
ρ	Rho	Korrelation
σ	Sigma	Standardabweichung
τ	Tau	Behandlungseffekt

Glossar lateinischer Buchstaben und Abkürzungen

Buchstabe	häufige Verwendung
c_i	1. Koeffizienten für linearen Kontrast 2. Studentisiertes Residuum
k	Koeffizientenvektor für Schätzbare Funktion
n	Stichprobenumfang
x	Einflussvariable
y	Zielvariable
r	Korrelation/ Zahl der Wiederholungen pro Behandlung
a	Achsenabschnitt (Schätzung)
b	Steigung (Schätzung)
s	Standardabweichung
ase	asymptotischer Standardfehler
$s.e./s.f.$	Standardfehler
\log	Natürlicher Logarithmus (Basis e) = \ln

Inhaltsverzeichnis

Seite

6.	Korrelation und Regression	1
6.1	Die Pearsonsche Produkt-Moment Korrelation	3
6.2	Regression	10
6.2.1	Streuungszerlegung	16
6.2.2	t-Tests und Vertrauensintervalle	20
6.2.3	Inverse Regression	25
6.3	Vergleich von Korrelation und Regression	30
6.4	Nichtlineare Regression durch Transformation der Variablen	32
6.5	Korrelation bei nichtlinearen Zusammenhängen	46
6.6	Test auf Linearität	50
6.7	Residuen, Modellvoraussetzungen und Ausreißer	55
6.7.1	Was sind Residuen?	55
6.7.2	Residuen-Plots	56
6.8	Lineare Modelle in Matrizenschreibweise	67
6.8.1	Normalengleichungen und ihre Lösung	72
6.8.2	Vertrauensintervalle und Tests für Linearkombinationen der Parameter	81
6.9	Vergleich von geschachtelten Modellen mittels F-Test	93
6.10	Multiple lineare Regression	97
6.10.1	Sequentieller Modellaufbau – Varianzanalyse	101
6.10.2	Erweiterung auf mehr als zwei Variablen	104
6.10.3	Das multiple Bestimmtheitsmaß	105
6.10.4	Multikollinearität	106
6.10.5	Variablenselektion	110
6.11	Polynomregression	126
6.12	„Eigentliche“ nichtlineare Regression	132
6.12.1	Einige "eigentliche" (intrinsisch) nichtlineare Modelle	133
6.12.2	Schätzen der Parameter	137
6.12.3	Startwerte	146
6.12.4	Schließende Statistik	152
6.12.5	Numerische Probleme - "ill-conditioning"	154
*6.12.6	Ein weiteres Beispiel	156
6.13	Lineare Kontraste	163
	Anhang: Einige Grundlagen der Matrizenrechnung	167
7.	Transformationen zur Erzielung der Voraussetzungen	173
7.1	Beispiel einer einfachen Varianzanalyse	173
7.2	Studentisierte Residuen im allgemeinen linearen Modell	182
7.3	Einige gängige Transformationen	187
7.4	Generalisierte lineare Modelle	188
8.	Versuchsanlagen	189
8.1	Wiederholung	191
8.2	Randomisation	194
8.3	Blockbildung	197
8.4	Randomisierte vollständige Blockanlage	197
8.4.1	Randomisation	198
8.5	Lateinisches Quadrat	201
8.5.1	Randomisation	204
8.5.2	Cross-over Design	206

8.6	Anlagen mit unvollständigen oder ungleich großen Blöcken	207
8.7	Auswertung einer Blockanlage	209
8.7.1	Varianzanalyse einer Blockanlage	209
8.7.2	Mittelwertvergleiche in einer Blockanlage	213
8.8	Auswertung eines Lateinischen Quadrats	222
8.8.1	Varianzanalyse eines Lateinischen Quadrats	222
8.8.2	Mittelwertvergleiche in einem Lateinischen Quadrat	226
8.9	Regression in einer Blockanlage	231
9.	Zweistufige Stichproben	241
9.1	Modell und Auswertung	241
9.2	Optimale Allokation	246
10.	Zweifaktorielle Varianzanalyse – Wechselwirkung	250
10.1	Wechselwirkungen (Interaktionen)	250
10.2	Modellierung	251
10.3	Varianzanalyse bei balancierten Daten	264
10.4	Mittelwertvergleiche bei balancierten Daten	257
10.5	Varianzanalyse bei unbalancierten Daten	267
10.6	Mittelwertvergleiche bei unbalancierten Daten	269
10.7	Zweifaktorielle Versuche in anderen Versuchsanlagen (Blockanlage etc.)	277
11.	Elementare nichtparametrische Verfahren	279
11.1	Vergleich zweier unabhängiger Stichproben	281
11.1.1	Median-Test	282
11.1.2	Mann-Whitney-Test	282
11.2	Kruskal-Wallis-Test (H-Test) für mehr als zwei unverbundene Stichproben	287
11.3	Vergleich zweier verbundener Stichproben	289
11.3.1	Vorzeichen-Test	290
11.3.2	Vorzeichen-Rangtest	293
11.4	Friedman-Test für mehr als zwei verbundene Stichproben	295
11.5	Vor- und Nachteile nichtparametrischer Verfahren	298
12.	Kovarianzanalyse	300
12.1	Modellbildung	302
12.2	Varianzanalyse	304
12.3	Mittelwertvergleiche	307
12.4	Ein soziologisches Beispiel	311
12.4.1	Modellierung	313
12.4.2	Varianzanalyse und Modellselektion	314
12.4.3	Schätzen des selektierten Modells	315
12.4.4	Welcher Faktor hat den größeren Einfluß?	317
13.	Messwiederholungen	319
14.	Einführung in multivariate Verfahren	331
14.1	Zwei einführende Beispiele	331
14.2	Distanzmaße und Ähnlichkeiten	333
14.2.1	Euklidische Distanz	333
14.2.2	Binäre Daten (0-1)	338
14.2.3	Gemischte Daten	340

14.3	Clusteranalyse	342
14.3.1	Average Linkage	342
14.3.2	Single Linkage	344
14.3.3	Complete Linkage	344
14.3.4	Anwendung auf die Margarine-Daten	345
14.4	Hauptkomponentenanalyse	350
14.4.1	Erster Überblick	350
* 14.4.2	Mathematische Details zur Berechnung der Hauptkomponenten	366
* 14.4.3	Biplots	377
Anhang A:	Tabellen wichtiger Verteilungen	383
I.	Standardnormalverteilung - $P(Z > z)$	384
II.	t-Verteilung (zweiseitig)	385
II(b).	t-Verteilung (einseitig)	386
III.	Standardnormalverteilung - Quantile	387
IV.	Chi-Quadrat-Verteilung	388
V.	F-Verteilung	389
VI.	F-Verteilung	390
VII.	Studentisierte Variationsbreiten	391
VIII.	Chi-Quadrat-Verteilung	392
Anhang B:	Was ist eigentlich SAS?	393
Anhang C:	Die Methode von Lagrange	393
Anhang D:	Was besagt der p -Wert?	394
Anhang E:	Freiheitsgrade	399
Anhang F:	Fehlerfortpflanzung (Delta-Methode)	400

Empfohlene Lehrbücher

- Backhaus, K., Erichson, B., Plinke, W., Weiber, R. 2000. Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. Springer, Berlin.
- Bortz J., Lienert, G. A. 1998. Kurzgefasste Statistik für die klinische Forschung. Springer, Berlin.
- Dean, A., Voss, D. T. 1998. Design and analysis of experiments. Springer, Berlin.
- Draper, N., Smith, H. 1998. Applied regression analysis. 3rd ed. Wiley, New York.
- Linder, A., Berchtold, W. 1982. Statistische Methoden II und III. Birkhäuser, Basel.
- Mead, R., Curnow, R. N., Hasted, A. M. 1993. Statistical methods in agriculture and experimental biology. 2nd edition. Chapman & Hall, London.
- Mead, R., Gilmour, S. G., Mead, A. 2012. Statistical principles for the experimental design of experiments. Cambridge University Press, Cambridge.
- Munzert, M. 1992. Einführung in das pflanzenbauliche Versuchswesen. Parey, Berlin.
- Munzert, M. 2015. Landwirtschaftliche und gartenbauliche Versuche mit SAS. Springer, Berlin.
- Petersen, R. G. 1994. Agricultural field experiments. Marcel Dekker, New York.
- Schabenberger, O., Pierce, F. J. 2000. Contemporary statistical models. CRC Press, Boca Raton.
- Snedecor, G. W., Cochran, W.G. 1967. Statistical methods. Iowa State University Press, Ames.
- Steel, R. G. D., Torrie, J. H. 1980. Principles and procedures of statistics. 2nd edition. McGraw Hill, New York.
- Thomas, E. 2006. Feldversuchswesen. Ulmer, Stuttgart.

Vorbemerkung

Die Vorlesung "Biometrie" wird angeboten für Studierende der Agrarbiologie und der Agrarwissenschaften (B. Sc.). Die Vorlesung und somit auch das Skript schließen nahtlos an die Vorlesung "Statistik" im 1. Semester für Studierende in den Bachelor-Studiengängen Agrarwissenschaften, der Agrarbiologie und Nachwachsende Rohstoffe.

Da sich der Inhalt dieser Vorlesung inhaltlich auf den der Vorlesung "Statistik" bezieht, wurde die Nummerierung der Kapitel von der Vorlesung "Statistik" ausgehend weitergeführt. Daher beginnt das Skript mit Kapitel 6. Außerdem überlappt das Ende des Skriptes zur Vorlesung "Statistik" mit dem Anfang des Skriptes zur Vorlesung "Biometrie", weil hier zu Beginn vor allem auf Methoden der linearen Regression Bezug genommen wird, die am Ende der Vorlesung "Statistik" behandelt werden. Die Überlappung der Skripte dient der besseren Lesbarkeit.

Im vorliegenden Skript wird an verschiedenen Stellen auf Kapitel des Skriptes "Statistik" Bezug genommen. Dies erleichtert es den Studierenden, Bezüge zu der von ihnen absolvierten Vorlesung herzustellen. Ergänzende Materialien zur Vorlesung Biometrie sind unter <http://ilias.uni-hohenheim.de/> erhältlich.

Übungen

Begleitend zur Vorlesung werden Übungen angeboten. Zweck der Übungen ist, dass die Studierenden selbständig die gestellten Aufgaben lösen. Wenn Fragen auftreten, stehen Tutoren für die Beantwortung zur Verfügung. Das Bearbeiten der Übungsaufgaben stellt eine wesentliche Vorbereitung auf die Klausur dar. In den Übungen werden keine Aufgaben an der Tafel vorgerechnet, es geht ums selber Rechnen. Das Vorrechnen von Beispielen passiert in der Vorlesung.

Klausur

Termin: Ende des SS

Es muss eine Mindestpunktzahl (50 %) erreicht werden. In der Klausur dürfen Taschenrechner, das Skript, Ihre handschriftlichen Aufzeichnungen sowie Bücher verwendet werden.

Eine **Nachholklausur** wird **einmal** während des jeweils folgenden Semesters angeboten. Wird auch diese nicht bestanden, kann im darauffolgenden Semester an der für die jeweilige Vorlesung am Semester-Ende stattfindenden Klausur teilgenommen werden. **Während eines Semesters wird somit nur eine Klausur angeboten. Es wird keine mündliche Prüfung angeboten.**

Ich weise ausdrücklich darauf hin, dass das primäre Lernen mit Altklausuren nicht zu empfehlen ist. Die Fragen ändern sich von Jahr zu Jahr. Konzentrieren Sie sich auf den Stoff der Vorlesung und die Übungen. In der Klausur geht es vor allem darum, das Verständnis der verschiedenen Methoden und die Fähigkeit zur richtigen Methodenwahl abzufragen, nicht um den alleinigen Nachweis, dass reine Rechentechnik beherrscht wird. Selbstverständlich müssen Sie in der Klausur zwar auch rechnen, aber nicht nur.

Begleitskripte in R und SAS

Aus Studiengebühren finanziert gibt es Begleitskripte zur Vorlesung, welche die Verrechnung vieler der in der Vorlesung behandelten Beispiele in den Statistik-Paketen R und SAS erläutern (Siehe <http://ilias.uni-hohenheim.de/>).

Kontaktzeit und Workload

Die Kontaktzeit für ein Modul an der Universität Hohenheim beträgt 56 Stunden. Die sogenannte Workload eines Moduls beträgt 150 bis 180 Stunden. Dies bedeutet, dass Sie für jede Stunde Vorlesung oder Übung jeweils etwa zwei Stunden selbständig nacharbeiten müssen. Dies können sie anhand des Skriptes, ihrer Aufzeichnungen und mit den Übungsaufgaben tun. Darüber hinaus ist ihnen begleitend dringend die selbstständige Lektüre von Lehrbüchern empfohlen (siehe Liste mit empfohlenen Büchern).

6. Korrelation und Regression

Bisher (vorangegangene Vorlesung Statistik) wurden bereits verschiedene Verfahren zur Erfassung des Zusammenhangs zweier Variablen behandelt, ohne dass diese entsprechend bezeichnet wurden. Der Zusammenhang zweier Variablen x und y kann symbolisch wie folgt dargestellt werden:

$x \longrightarrow y$ (x bewirkt y)

$x \longleftrightarrow y$ (x und y hängen wechselseitig voneinander ab)

$\begin{array}{c} \nearrow x \\ z \searrow \\ \downarrow y \end{array}$ (x und y hängen direkt von z ab)

Die Wahl des Verfahrens zur Ermittlung des Zusammenhanges hängt vom Skalenniveau der Variablen ab. Sind beide Variablen metrisch skaliert, kommen die Korrelation und die Regression zur Auswertung in Frage. Diese Verfahren stehen im Zentrum dieses Kapitels. Im folgenden wird eine kleine Übersicht gegeben über bisher behandelte Verfahren und deren Abhängigkeit vom Skalenniveau der Variablen.

Tab. 6.1: Abhängigkeit der Wahl des Verfahrens zur Ermittlung eines Zusammenhangs.

Skalenniveau		Verfahren (Beispiele)
x	y	
metrisch	metrisch	Korrelation, Regression
kategorial	metrisch	paarweiser t-Test, Varianzanalyse
kategorial	kategorial	χ^2 -Test

Beispiel: Es soll der Zusammenhang ermittelt werden zwischen Sortenwahl (kategorial) und Ertrag (metrisch). Hierzu wird ein Sortenversuch mit fünf Sorten durchgeführt. Auswertung: Varianzanalyse mit anschließendem multiplen t-Test (Kap. 3 und 4).

Beispiel: Es soll der Zusammenhang ermittelt werden zwischen vorherrschendem Bodentyp und Besitzform landwirtschaftlicher Betriebe in einer Region. Hierzu wird eine Zufallsstichprobe von Betrieben befragt. Die Betriebe werden bezüglich Besitzform und Bodentyp in Klassen eingeteilt. Auswertung: χ^2 -Test (Kap. 5).

Beispiel: In einer Erhebung wurden für einen Zeitraum von 26 Jahren die erzielten Ernten (y) den Regenmengen von April bis Juni (x) gegenübergestellt (Jacob, A. und Rüter H. 1961 Der Vegetationsversuch. VDLUFA, Berlin; Daten aus: Scheinert, R.: Pflanzenbau 5, 236-239, 1928/29). Beide Merkmale sind metrisch.

Jahr	x	y	x^2	y^2	xy
1900	177	26,2	31329	686,44	4637,4
1901	96	25,0	9216	625,00	2400,0
1902	144	32,6	20736	1062,76	4694,4
1903	105	26,6	11025	707,56	2793,0
1904	111	19,6	12321	384,16	2175,6
1905	135	20,4	18225	416,16	2754,0
1906	209	29,2	43681	852,64	6102,8
1907	161	33,8	25921	1142,44	5441,8
1908	246	26,6	60516	707,56	6543,6
1909	108	22,6	11664	510,76	2440,8
1910	137	24,2	18769	585,64	3315,4
1911	71	16,6	5041	275,56	1178,6
1912	119	29,8	14161	888,04	3546,2
1913	108	19,4	11664	376,36	2095,2
1914	132	30,6	17424	936,36	4039,2
1915	89	16,4	7921	268,96	1459,6
1916	147	30,4	21609	924,16	4468,8
1917	98	19,2	9604	368,64	1881,6
1918	106	20,2	11236	408,04	2141,2
1919	123	25,6	15129	655,36	3148,8
1920	156	31,0	24336	961,00	4836,0
1921	191	35,8	36481	1281,64	6837,8
1922	162	31,6	26244	998,56	5119,2
1923	235	33,6	55225	1128,96	7896,0
1924	147	30,4	21609	924,16	4468,8
1925	110	30,2	12100	912,04	3322,0
Summe	3623	687,6	553187	18988,96	99737,8

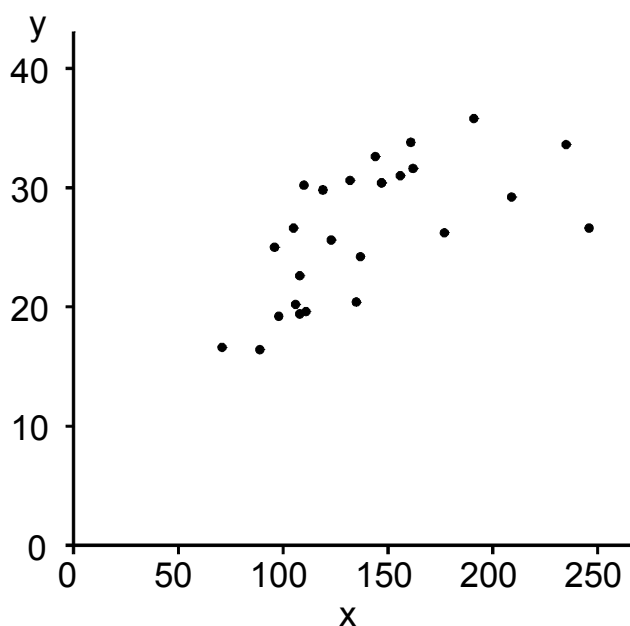


Abb. 6.1: Plot der Erträge (y ; dt/ha) gegen die in den Monaten April bis Juni gefallene Regenmenge (x ; mm).

Diese Daten können graphisch dargestellt werden, indem die Messwerte (x_i = Regenmenge; y_i = Ertrag) als Koordinaten von Punkten in einem kartesischen Koordinatensystem aufgefaßt werden. Hierdurch wird jedes Jahr bzw. jedes Messwertpaar (x_i, y_i) als Punkt repräsentiert (siehe Abb. 6.1). Offenbar besteht ein gewisser **Zusammenhang** zwischen Niederschlag und Ertrag. Je höher der Niederschlag, desto höher der Ertrag. Man spricht auch von einer positiven **Korrelation**. Die Korrelation ist allerdings nicht perfekt, da der Ertrag nicht ausschließlich durch den Niederschlag bestimmt wird. Ziel der weiteren statistischen Auswertung kann es nun sein, den Zusammenhang zwischen den beiden Variablen zu quantifizieren. Hierzu können die Verfahren der Korrelation und Regression verwendet werden, die in diesem Kapitel besprochen werden.

6.1 Die Pearsonsche Produkt-Moment Korrelation

Die Korrelation ist wie folgt definiert:

$$\rho = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

wobei σ_{xy} die sog. **Kovarianz** von x_i und y_i ist, während σ_x^2 und σ_y^2 die Varianzen von x_i und y_i sind. Gibt es keinen Zusammenhang zwischen x_i und y_i , ist die Kovarianz gleich Null. Gibt es dagegen einen Zusammenhang, weicht die Kovarianz um so stärker von Null ab, je stärker der Zusammenhang ist. Die Division durch die Wurzel aus dem Produkt der Varianzen ist eine **Standardisierung**, die bewirkt, dass die **Korrelation** eine dimensionslose Maßzahl zwischen -1 (negativer Zusammenhang) und 1 (positiver Zusammenhang) ist. Die Korrelation kann auch als eine **standardisierte Kovarianz** bezeichnet werden. Wir kommen auf den beschränkten Wertebereich der Korrelation zurück, wenn wir deren Schätzung besprochen haben.

Nun zur Schätzung der Korrelation. Die Varianz wird bekanntermaßen wie folgt geschätzt:

$$s_x^2 = \frac{SQ_x}{n-1} \quad \text{und} \quad s_y^2 = \frac{SQ_y}{n-1}$$

mit

$$SQ_x = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{und} \quad SQ_y = \sum_{i=1}^n (y_i - \bar{y})^2$$

wobei (x_i, y_i) das i -te Messwertpaar ist. Die Kovarianz wird geschätzt durch

$$s_{xy} = \frac{SP_{xy}}{n-1}$$

$$SP_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Die Korrelation wird geschätzt durch

$$r = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}$$

SP_{xy} ist die Summe der Kreuzprodukte. Man entnimmt der Formel für SP_{xy} leicht folgendes: Wenn sowohl x_i als auch y_i größer ist als der Mittelwert, ist das Kreuzprodukt $(x_i - \bar{x})(y_i - \bar{y})$ größer als Null. Ebenso ist das Kreuzprodukt größer als Null, wenn beide Abweichungen negativ sind. Hieraus folgt, dass die Summe der Kreuzprodukte und somit auch die Korrelation im Falle eines positiven Zusammenhanges positiv werden muss. Haben die Abweichungen dagegen umgekehrte Vorzeichen, so wird der Beitrag zur Summe der Kreuzprodukte negativ. Ist schließlich das Vorzeichen der Abweichung für x_i unabhängig vom Vorzeichen der Abweichung für y_i , dann wird die Summe der Kreuzprodukte mehr oder weniger nahe bei Null liegen.

Nun zur Erklärung, warum die Korrelation zwischen -1 und 1 liegen muss. Man überzeugt sich leicht, dass die Korrelation auch ausgedrückt werden kann als Kovarianz der z -transformierten Variablen x und y :

$$r = \frac{\sum_{i=1}^n z_{xi} z_{yi}}{(n-1)}$$

wobei

$$z_{xi} = \frac{x_i - \bar{x}}{s_x} \quad \text{und} \quad z_{yi} = \frac{y_i - \bar{y}}{s_y}$$

Es ist klar, dass

$$\sum_{i=1}^n (z_{xi} - z_{yi})^2 \geq 0$$

ist, da hier quadrierte Differenzen summiert werden, und das Quadrat einer beliebigen Zahl größer oder gleich Null sein muss. Nun ist aber

$$\sum_{i=1}^n (z_{xi} - z_{yi})^2 = \sum_{i=1}^n z_{xi}^2 - 2 \sum_{i=1}^n z_{xi} z_{yi} + \sum_{i=1}^n z_{yi}^2 = (n-1) - 2(n-1)r + (n-1) \geq 0$$

woraus durch Umformung folgt, dass $r \leq 1$ ist (Achtung: Bei Division beider Seiten einer Ungleichung durch eine negative Zahl muss "größer gleich" in "kleiner gleich" umgewandelt werden und umgekehrt). Analog folgt aus

$$\sum_{i=1}^n (z_{xi} + z_{yi})^2 = \sum_{i=1}^n z_{xi}^2 + 2 \sum_{i=1}^n z_{xi} z_{yi} + \sum_{i=1}^n z_{yi}^2 = (n-1) + 2(n-1)r + (n-1) \geq 0,$$

dass $r \geq -1$ sein muss. Somit haben wir gezeigt, dass

$$-1 \leq r \leq 1$$

Man kann auch folgende einfache Betrachtung anstellen: Bei perfektem linearen Zusammenhang ist entweder $z_{xi} = z_{yi}$, so dass $r = s_z^2 = 1$ ist, oder es ist $z_{xi} = -z_{yi}$, so dass $r = -s_z^2 = -1$ ist. Somit muss r zwischen -1 und 1 liegen. In Abb. 6.2 werden einige Beispiele für Werte der Korrelation bei verschiedenen Ausprägungen und Vorzeichen des Zusammenhanges zwischen X und Y gegeben.

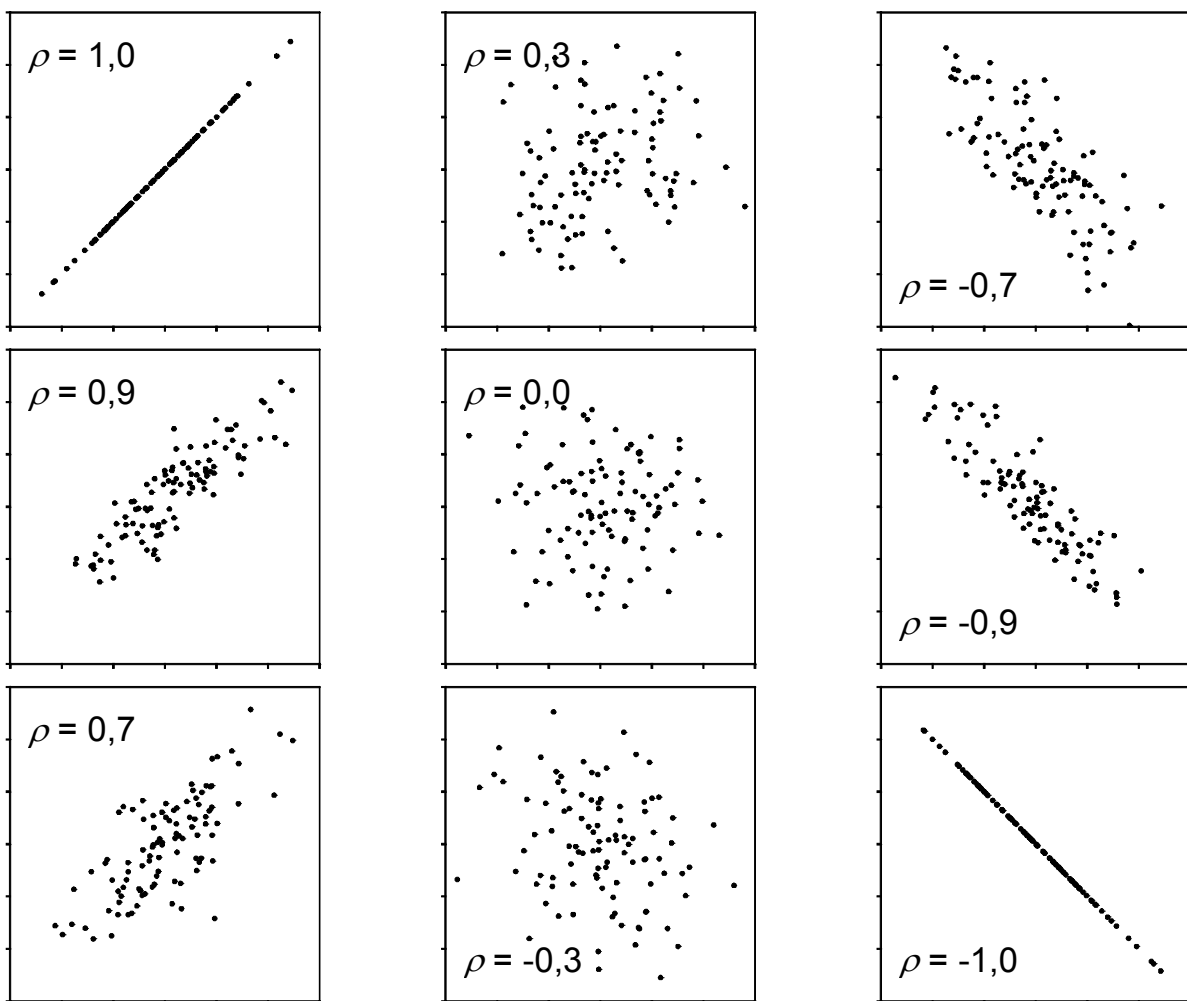


Abb. 6.2: Einige Beispiele (simulierte Daten) für verschiedene Werte der Korrelation.

Es ist wichtig, zu beachten, dass die Korrelation ein Maß für den linearen Zusammenhang zweier Variablen ist. Bei nichtlinearen Zusammenhängen sollte ein anderes Korrelationsmaß verwendet werden, z.B. die Rangkorrelation von Spearman.

Für die praktische Anwendung ist folgende Rechenformel für die Korrelation hilfreich:

Schätzen des Korrelationskoeffizienten

$$r = \frac{SP_{xy}}{\sqrt{SQ_x SQ_y}}$$

$$SP_{xy} = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right) \left(\sum_{i=1}^n y_i\right)}{n} \quad (\text{Summe der Kreuzprodukte})$$

$$SQ_x = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} \quad (\text{Summe der Quadrate für X})$$

$$SQ_y = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} \quad (\text{Summe der Quadrate für Y})$$

$(x_i, y_i) = i\text{-tes Messwertpaar.}$

Beispiel: Für die Regendaten (y = Erträge und x = Regenmenge zwischen April und Juni) finden wir

$$SP_{xy} = 99737,8 - \frac{3623 \cdot 687,6}{26} = 3923,38$$

$$SQ_x = 553187 - \frac{3623^2}{26} = 48335,88$$

$$SQ_y = 18988,96 - \frac{687,6^2}{26} = 804,58$$

$$r = \frac{3923,38}{\sqrt{48335,88 \cdot 804,88}} = 0,63$$

Die Korrelation ist positiv, da der Zusammenhang zwischen Ertrag und Niederschlag positiv ist. Da $r < 1$ ist, ist der Zusammenhang aber nicht perfekt. Im weiteren ist die Frage zu klären, ob der Stichprobenkorrelationskoeffizient r signifikant vom Wert $\rho = 0$ abweicht, ob es also überhaupt einen signifikanten Zusammenhang gibt.

Test der Korrelation ρ

Frage: Gibt es einen echten Zusammenhang zwischen zwei Variablen X und Y?

Voraussetzung: Die Daten sind bivariat normalverteilt

$H_0: \rho = 0$

$H_A: \rho \neq 0$

Rechenweg:

(1) Berechne $t_{Vers} = \frac{|r|}{\sqrt{1-r^2}} \sqrt{n-2}$

(2) Lies in der t -Tabelle $t_{Tab}(FG; \alpha)$ ab (Tab. II, zweiseitig) wobei α das Signifikanzniveau und $FG = n - 2$ die Freiheitsgrade sind.

(3) Vergleiche t_{Vers} mit t_{Tab} :

Falls $t_{Vers} \leq t_{Tab} \Rightarrow H_0 (\rho = 0)$ (Kein Zusammenhang)

Falls $t_{Vers} > t_{Tab} \Rightarrow H_1 (\rho \neq 0)$ (Signifikanter Zusammenhang)

Man beachte, dass der Test die Annahme macht, dass die Daten einer bivariaten Normalverteilung folgen. Außerdem erfasst die Korrelation nur lineare Zusammenhänge, und es muss angenommen werden, dass kein nichtlinearer Zusammenhang besteht. Die (geschätzte) bivariate Normalverteilung für die Regendaten ist in der folgenden Graphik wiedergegeben.

Wahrscheinlichkeitsdichte

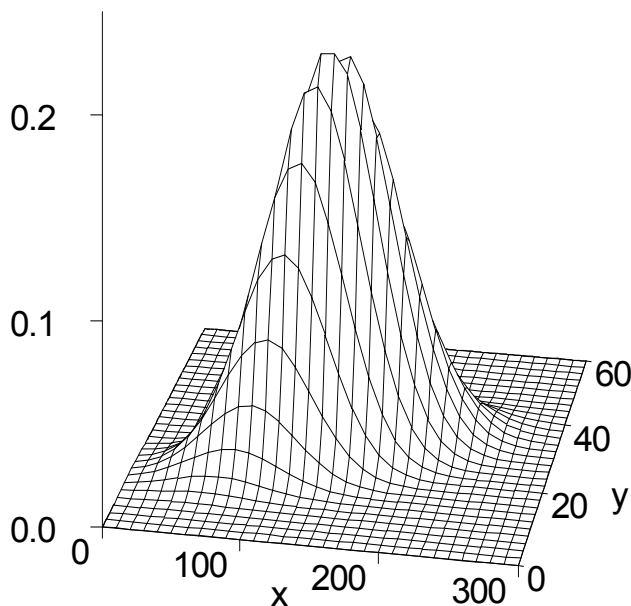


Abb.6.2(b): Geschätzte bivariate Normalverteilung für die Regendaten (F = Wahrscheinlichkeitsdichte).

Beispiel: Wir wollen die Korrelation zwischen Regenmenge und Ertrag zum 5%-Niveau testen.

$r = 0,63, n = 26, \alpha = 5\%$

$$t_{Vers} = \frac{|0,63|}{\sqrt{1-0,63^2}} \sqrt{26-2} = 3,97$$

$t_{Tab} = 2,064 < t_{Vers} \Rightarrow$ Die Korrelation ist signifikant

Eine signifikante Korrelation sagt zunächst nichts über die **Kausalität** des Zusammenhangs aus. Sie besagt lediglich, dass ein Zusammenhang besteht, ohne dass damit geklärt wäre, wie es zu diesem Zusammenhang kommt. Insbesondere ist damit nichts gesagt über eine **Ursache-Wirkungs-Beziehung** oder **Kausalbeziehung**.

Beispiel: In Industrieländern ist oft eine signifikante Korrelation zwischen der Zahl der Geburten und der Zahl der Störche pro Flächeneinheit festgestellt worden. Eine Ursache-Wirkungs-Beziehung besteht hier nicht. Vielmehr ist zu vermuten, dass sowohl Geburten- als auch Storchzahlen beispielsweise vom Grad der Industrialisierung beeinflusst werden. Andere Erklärungen sind möglich.

Beispiel: Es kann oft eine hohe Korrelation zwischen der Größe des Schadens eines Feuers und der Zahl der an der Brandlöschung beteiligten Feuerwehrmänner festgestellt werden.

Man spricht in Fällen wie den vorangegangenen auch von einer **Scheinkorrelation**, obwohl diese Bezeichnung den Punkt nicht ganz trifft, denn die Korrelation ist ja real. Problematisch ist lediglich die fälschliche Interpretation im Sinne einer Ursache-Wirkungs-Beziehung. In beiden Fällen liegt eine Abhängigkeit von einer dritten Größe Z vor, welche zu der Scheinkorrelation von X und Y führt.

Beispiel: Bei den Regendaten liegt von der Sachlage her eine Ursache-Wirkungs-Beziehung auf der Hand: Die Höhe des Niederschlages bedingt den Ertrag, und nicht umgekehrt. Es liegt hier eine **einseitige Abhängigkeit** vor. Allerdings liefern die Regendaten, die durch eine Erhebung ermittelt wurden, keinen strengen Beweis, dass die Regenmenge wirklich ursächlich für den Ertrag verantwortlich ist, weil in dieser Erhebung viele Umweltfaktoren zufällig variieren, die den Ertrag beeinflussen können (Temperatur, Aussaattermin, Wärmesumme, Bodenzustand zum Zeitpunkt der Aussaat, Keimfähigkeit des Saatgutes, etc.). Um den Einfluss der Regenmenge zweifelsfrei nachweisen zu können, ist ein Experiment erforderlich. Es ist generell festzuhalten, dass **Erhebungen** im allgemeinen nicht für den Nachweis von Kausalbeziehungen geeignet sind, auch wenn eine solche noch so plausibel erscheint. Erhebungen können lediglich Hinweise für die Bildung von Hypothesen über Kausalbeziehungen geben. Diese Hypothesen müssen dann in einem **Experiment** überprüft werden, in welchem der ursächliche Faktor gezielt variiert wird, während alles andere möglichst konstant gehalten wird (**ceteris paribus Klausel**). Erst wenn sich unter diesen Bedingungen der Zusammenhang reproduzieren lässt, ist die Kausalbeziehung nachgewiesen. Das für den Nachweis notwendige Konstanthalten aller anderen Einflussfaktoren ist in einer Erhebung generell nicht oder nur sehr eingeschränkt möglich. Im Fall der Regendaten ist es beispielsweise unmöglich, die Temperaturbedingungen konstant zu halten. Wir werden auf diesen Punkt in Abschnitt 6.2 zurückkommen.

Beispiel: Es kann oft eine Korrelation zwischen Pflanzenlänge und Blattflächenindex (BFI) (Blattfläche pro Standfläche) bei Raps festgestellt werden. Diese beiden Merkmale bedingen sich gegenseitig, es besteht also eine **zweiseitige Abhängigkeit**. Je länger die Pflanze, desto mehr Verzweigungen bildet die Pflanze und umso mehr Blätter werden gebildet. Umgekehrt ermöglicht ein höherer Blattflächenindex eine erhöhte Assimilation und somit eine bessere Bedingung für das Längenwachstum der Pflanze.

Neben einem Test der Korrelation können wir auch ein Vertrauensintervall berechnen. R. A. Fisher hat durch eine komplizierte Approximation (A. Stuart & K. Ord: Kendall's advanced theory of statistics. 6th edition. Volume 1, § 16.33) gezeigt, dass

$$q = \frac{1}{2} \ln \left[\frac{1+r}{1-r} \right]$$

für nicht zu kleines n näherungsweise einer Normalverteilung mit Mittelwert

$$\theta = \frac{1}{2} \ln \left[\frac{1+\rho}{1-\rho} \right]$$

und Varianz

$$\sigma_q^2 = \frac{1}{n-3}$$

Folgt (der Term $n-3$ im Zähler der Varianz hat nichts mit Freiheitsgraden zu tun). Daher ist

$$z = \frac{q-\theta}{\sigma_q}$$

näherungsweise standardnormalverteilt. Somit sind

$$\theta_u = q - z_{1-\alpha/2} \sigma_q \text{ und } \theta_o = q + z_{1-\alpha/2} \sigma_q$$

die $(1-\alpha)100\%$ -Vertrauensgrenzen für θ , wobei $z_{1-\alpha/2}$ das $(1-\alpha/2)100\%$ -Quantil der Standardnormalverteilung ist. Die entsprechenden Grenzen für ρ erhält man durch Rücktransformation der Grenzen θ_u und θ_o . Die praktisch durchzuführenden Berechnungen enthält der folgende Kasten.

Vertrauensintervall für die Korrelation ρ

Berechne:

$$q = \frac{1}{2} \ln \left[\frac{1+r}{1-r} \right]$$

wobei $\ln()$ der natürliche Logarithmus ist (Basis e)

$$\theta_u = q - \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \text{ und } \theta_o = q + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}$$

wobei

$z_{1-\alpha/2} = (1-\alpha/2)100\%$ -Quantil der Standardnormalverteilung (Tab. III)

Die $(1-\alpha)100\%$ -Vertrauensgrenzen sind gegeben durch

$$\rho_u = \frac{e^{2\theta_u} - 1}{e^{2\theta_u} + 1} \quad \text{und} \quad \rho_o = \frac{e^{2\theta_o} - 1}{e^{2\theta_o} + 1}$$

Voraussetzung: Die Daten sind bivariat normalverteilt.

Beispiel: $r = 0,63$, $n = 26$, $\alpha = 5\%$

$$q = \frac{1}{2} \ln \left[\frac{1+0,63}{1-0,63} \right] = 0,741$$

$z_{1-\alpha/2} = 1,96$ (siehe Tab. III: $\alpha = 5\% = 0,05 \Rightarrow \gamma = 1-\alpha/2 = 1-0,05/2 = 0,975$
 $\Rightarrow z_{1-\alpha/2} = z_\gamma = z_{0,975} = 1,95996 \approx 1,96$)

$$\theta_u = 0,741 - \frac{1,96}{\sqrt{23}} = 0,741 - 0,409 = 0,332 \quad \text{und} \quad \theta_o = 0,741 + 0,409 = 1,150$$

$$\rho_u = \frac{e^{2 \cdot 0,332} - 1}{e^{2 \cdot 0,332} + 1} = \frac{1,943 - 1}{1,943 + 1} = 0,32 \quad \text{und} \quad \rho_o = \frac{e^{2 \cdot 1,150} - 1}{e^{2 \cdot 1,150} + 1} = \frac{9,974 - 1}{9,974 + 1} = 0,82$$

Mit 95%-iger Wahrscheinlichkeit überdeckt das Intervall von 0,32 bis 0,82 die Korrelation ρ . Wir sehen, dass die Schätzung der Korrelation relativ ungenau ist, u.a. aufgrund des geringen Stichprobenumfanges.

6.2 Regression

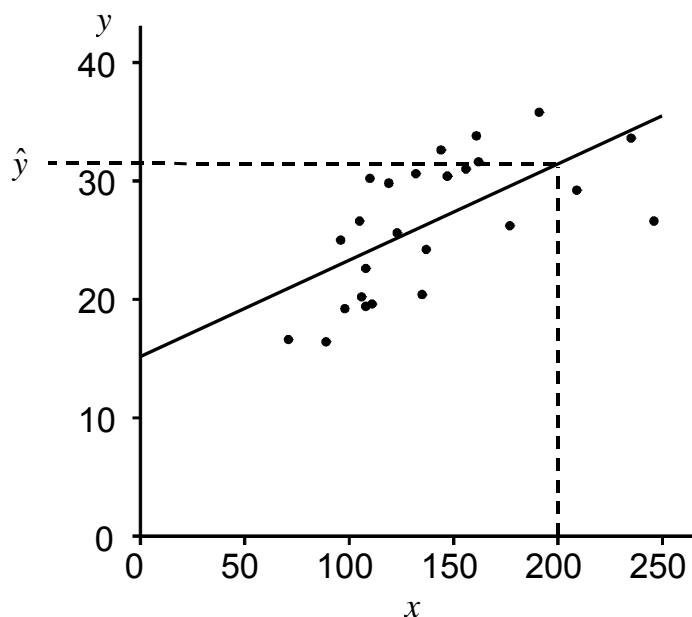


Abb. 6.3: Regressionsgerade für die Regendaten, mit Prognose des Ertrages (\hat{y}) bei Niederschlag $x = 200$ mm.

Zur besseren Beschreibung eines linearen Zusammenhanges ist es oft hilfreich, eine Gerade durch die Punktwolke zu legen. Man spricht in diesem Zusammenhang von einer **Regressionsgeraden**. Neben einer Beschreibung des Zusammenhanges kann eine Regression aber auch zur Vorhersage (Prognose) dienen.

Beispiel: Für die Regendaten ist eine Regressionsgerade in der Abb. 6.3 dargestellt.

Die Regressionsgerade hilft zum einen bei der Interpretation der Punktwolke, indem sie den linearen Trend veranschaulicht. Sie kann hier außerdem für eine Ertragsprognose in einem neuen Jahr genutzt werden, vorausgesetzt, wir haben für dieses Jahr die Niederschlagsmenge zwischen April und Juni (x) ermittelt. Bei einem Niederschlag vom $x = 200$ mm erwarten wir einen Ertrag von etwa $\hat{y} = 32$ dt/ha, wie ein Blick auf Abb. 6.3 zeigt. Die Vorhersage wird mit einem "Dach" auf dem y kenntlich gemacht. Eine Regressionsgerade ist also in verschiedener Hinsicht sehr hilfreich.

Die nun zu klärende Frage ist, wie man am besten eine Gerade durch die Punktwolke legt. Dazu betrachten wir zunächst die Geradengleichung. Sie ist gegeben durch

$$\eta = E(y) = \alpha + \beta x$$

wobei

$E(y)$ = Erwartungswert von y

α = Achsenabschnitt

β = Steigung

Die wesentliche Interpretation der Steigung β ist folgende: Steigt x um eine Einheit, so steigt der erwartete Wert für y um β Einheiten. Dies veranschaulicht das

Steigungsdreieck in Abb. 6.4(a). Der Achsenabschnitt α ist der erwartete Wert für y wenn $x = 0$ ist.

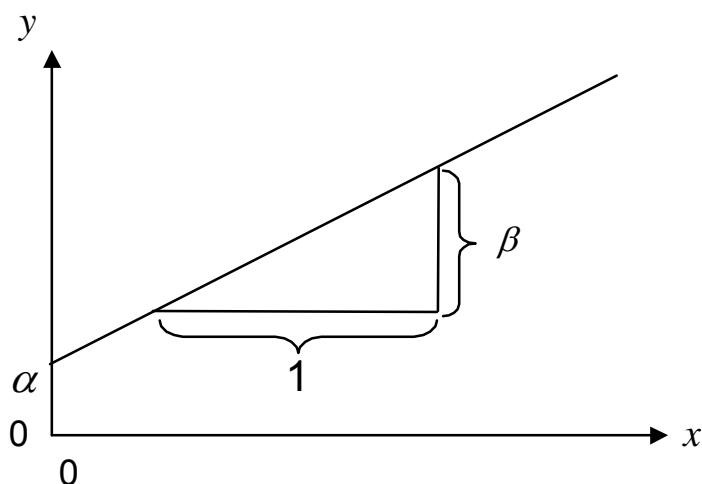


Abb. 6.4(a): Veranschaulichung der Interpretation der Regressionsgleichung $\eta = \alpha + \beta x$ mit Steigungsdreieck.

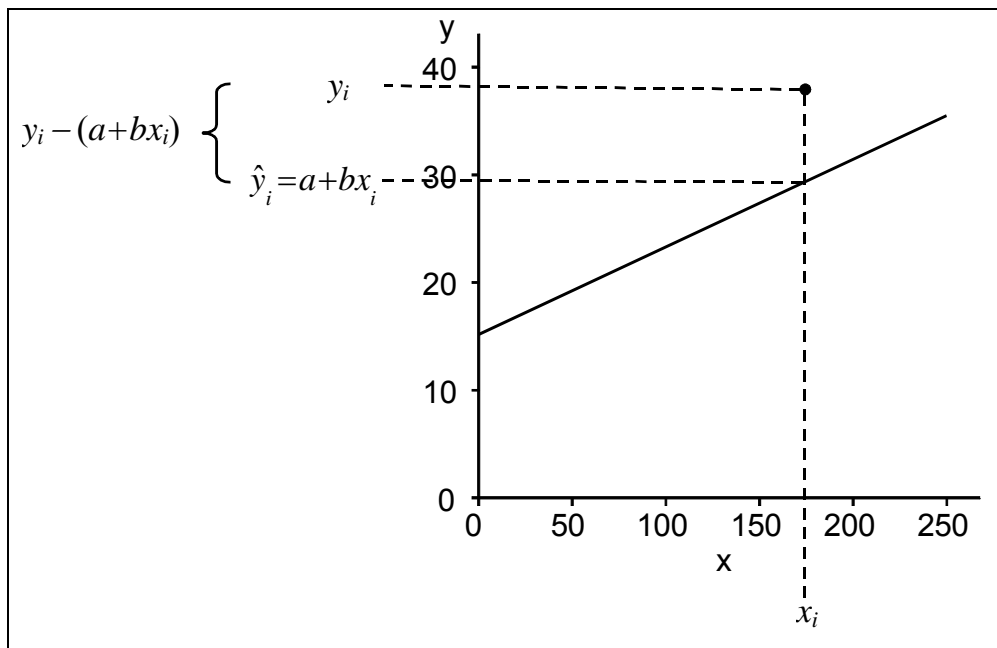


Abb. 6.4(b): Veranschaulichung der Abweichung eines Datenpunktes von der geschätzten der Regressionsgerade.

Eine Gerade legen wir nun so durch die Punktwolke, dass die vertikale Abweichung zwischen Gerade und den einzelnen Punkten möglichst klein wird. Die vertikale Abweichung des Punktes für die i -te Beobachtung von der zu schätzenden Regressionsgeraden ist gegeben durch

$$y_i - (a + bx_i) \quad ,$$

wobei a und b Schätzwerte für die Parameter α und β sind [siehe Abb. 6.4(b)]. Diese Abweichungen sollen minimiert werden. Bei der Regression wird die vertikale Abweichung betrachtet, weil es bei der Regression um die möglichst genaue Schätzung von y_i mittels der angepassten Regressionsgerade geht.

Um die Minimierung der vertikalen Abweichungen praktikabel zu gestalten, suchen wir nach einem Kriterium für die Güte der Anpassung der Geraden an die Punkte, welches dann optimiert wird. Zu denken wäre zunächst an die Summe der Abweichungen $y_i - (a + bx_i)$. Da die Abweichungen aber für gut passende Geraden mal positiv und mal negativ werden können, ist dies kein geeignetes Maß. Um das Problem der wechselnden Vorzeichen auszuschließen, können stattdessen die Abweichungen quadriert und dann summiert werden. Dies führt zur bereits von der Varianz her bekannten **Summe der Abweichungsquadrate** oder auch **Summe der Fehlerquadrate**:

$$SQ_{Fehler} = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

Dieses ist nun das Zielkriterium, welches minimiert werden soll. Wir bestimmen die Werte für a und b so, dass SQ_{Fehler} minimal wird. Hierzu können wir SQ_{Fehler} als Funktion der zwei Variablen a und b betrachten. Das Optimum einer Funktion

mehrerer Variablen bestimmt man, indem die partiellen Ableitungen der Funktion nach den Variablen gleich Null gesetzt werden (bei der Berechnung der partiellen Ableitung nach einer Variablen werden alle anderen Variablen als Konstanten betrachtet und dann die gewöhnliche Ableitung berechnet). In unserem Fall finden wir:

$$\frac{\partial S Q_{Fehler}}{\partial b} = 2 \sum_{i=1}^n [y_i - (a + bx_i)](-1)x_i = 0$$

$$\frac{\partial S Q_{Fehler}}{\partial a} = 2 \sum_{i=1}^n [y_i - (a + bx_i)](-1) = 0$$

Diese Gleichungen sind die sog. **Normalengleichungen**. Auflösen der Normalengleichungen nach den Unbekannten a und b liefert die Kleinstquadratschätzungen für die Parameter α und β . Lösen von $\partial S Q_{Fehler} / \partial a = 0$ nach a liefert:

$$\sum_{i=1}^n [y_i - (a + bx_i)] = n\bar{y} - n(a + b\bar{x}) = 0 \Leftrightarrow a = \bar{y} - b\bar{x}$$

Einsetzen dieser Gleichung in $\partial S Q_{Fehler} / \partial b = 0$ liefert:

$$\begin{aligned} \frac{\partial S Q_{Fehler}}{\partial b} &= \sum_{i=1}^n [y_i - (\bar{y} - b\bar{x} + bx_i)]x_i \\ &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - b \left[\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right] \\ &= \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} - b \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right] \\ &= SP_{xy} - b SQ_x = 0 \\ \Leftrightarrow b &= \frac{SP_{xy}}{SQ_x} \end{aligned}$$

Die hier beschriebene Methode heißt **Methode der kleinsten Quadrate**. Die Kleinstquadratschätzungen nach Auflösen der Normalengleichungen sind noch einmal in dem folgenden Kasten zusammengefaßt:

Schätzung der Regressionsgerade (Methode der kleinsten Quadrate):

$$\begin{aligned} b &= \frac{SP_{xy}}{SQ_x} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

wobei

$$SP_{xy} = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}$$

$$SQ_x = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$$

Geschätzte Regressionsgerade:

$$\hat{y} = a + bx$$

Man beachte, dass SQ_x und SP_{xy} dieselben Größen sind, die auch zur Berechnung der Korrelation benötigt wurden.

Beispiel: Für die Regendaten finden wir:

$$SP_{xy} = 3923,38$$

$$SQ_x = 48335,88$$

$$\bar{x} = 139,35$$

$$\bar{y} = 26,45$$

$$b = \frac{3923,38}{48335,88} = 0,0812$$

$$a = 26,45 - 0,0812 \cdot 139,35 = 15,14$$

$$\hat{y} = 15,14 + 0,0812x$$

Dies ist die Gleichung, welche für das Zeichnen der Gerade in Abb. 6.3 verwendet wurde. Die Steigung kann wie folgt interpretiert werden: steigt die Regenmenge um einen mm, so steigt der Ertrag um 0,0812 dt/ha = 8,12 kg/ha. Außerdem können wir die Regressionsgleichung zur Prognose nutzen. Wenn in einem neuen Jahr von April bis Juni 200 mm Regen gefallen sind, so schätzen wir den erwarteten Ertrag nach

$$\hat{y} = 15,14 + 0,0812 \cdot 200 = 31,37$$

Wir erwarten also einen Ertrag von 31,37 dt/ha, was nicht heißt, dass dies der tatsächlich realisierte Ertrag sein wird. Die Schätzung besagt lediglich, dass wir bei einem Niederschlag von 200 mm im Mittel 31,37 dt/ha erwarten. Da der Ertrag von vielen Faktoren beeinflusst wird, kann in verschiedenen Jahren mit derselben Niederschlagsmenge von 200 mm mal ein niedrigerer, mal ein höherer Ertrag als 31,37 dt/ha realisiert werden. 31,37 dt/ha ist einfach die beste Schätzung, die wir im Voraus aufgrund der historischen Daten machen können (Die Daten sind hier sehr alt und müssten durch neuere Daten aktualisiert werden, weil das Ertragsniveau im Laufe der Jahre aufgrund von Fortschritten in der Züchtung und im Pflanzenbau gestiegen ist).

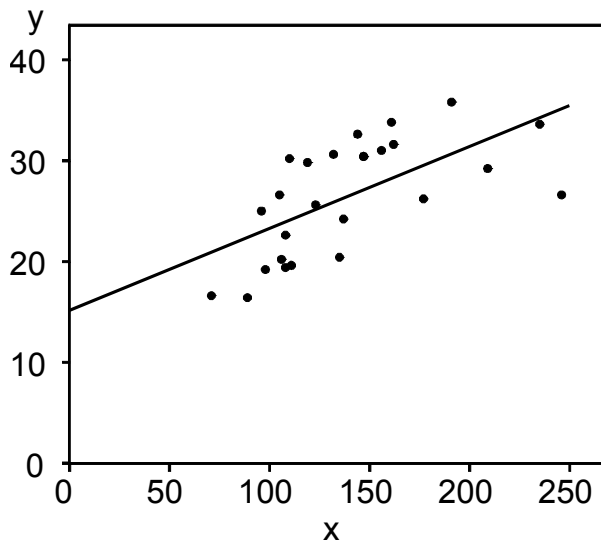


Abb. 6.4(c): Lineare Regression für Regendaten mit Streudiagramm [x = Regenmenge zwischen April und Juni in (mm); y = Ertrag (dt/ha)].

Im Zusammenhang mit der Prognose ist folgende Bezeichnungsweise üblich:

x = Prädiktorvariable, Einflussvariable (unabhängige Variable)

y = Zielvariable (abhängige Variable)

Das Wortpaar "abhängig/unabhängig" impliziert eine Ursache-Wirkungs-Beziehung der Form

$$x \rightarrow y$$

Eine solche muss aber nicht notwendigerweise gegeben sein, wie bereits im Zusammenhang mit der Korrelation in Abschnitt 6.1 diskutiert wurde. Ich bevorzuge daher im allgemeinen das Begriffspaar "Prädiktorvariable/Zielvariable" bzw. Einflussvariable/Zielvariable). Um eine Ursache-Wirkungs-Beziehung eindeutig nachzuweisen, muss die Prädiktorvariable in einem Versuch gezielt variiert werden, während alle anderen Umweltfaktoren konstant gehalten oder durch Randomisation ausgeschaltet werden. Bei der Erhebung zum Zusammenhang zwischen Ertrag und Regenmenge ist dies beispielsweise nicht gegeben. Es ist daher nicht auszuschließen, dass nicht die variierende Regenmenge, sondern ausschließlich andere Umweltfaktoren, die auch von Jahr zu Jahr schwankten, für die Ertragsbildung verantwortlich sind. Dies mag in diesem Beispiel als spitzfindig erscheinen, weil unsere praktische Erfahrung eindeutig zu beweisen scheint, dass der Regen ursächlich mit dem Ertrag zusammenhängt. Und trotzdem liefern die Erhebungsdaten nur einen Hinweis, aber keinen Beweis, dass diese Annahme zutrifft. Um den Einfluss der Wasserzufuhr experimentell einwandfrei nachzuweisen, kann beispielsweise ein Experiment durchgeführt werden, bei dem die Versuchsfläche überdacht wird und die Wasserzufuhr künstlich in mehreren Abstufungen variiert wird. Wir wiederholen hiermit die wichtige Feststellung, dass der strenge Nachweis von Kausalbeziehungen nur in einem Experiment möglich ist, nicht aber in einer Erhebung. Allerdings können in einer Erhebung weitere theoretische Überlegungen neben der Signifikanz eines Zusammenhanges sehr wohl weitere Indizien für eine Kausalbeziehung liefern.

Bei der linearen Regression ist weiterhin zu beachten, dass die Zuordnung der Variablen als x - oder y -Variable wichtig ist. Die Lage der angepassten Geraden ändert sich, wenn man x und y vertauscht. Die vorherzusagende Variable ist immer als y -Variable zu wählen. Außerdem ist es in bei geplanten Versuchen oft so, dass die beobachteten Stufen von x gezielt ausgewählt werden können, z.B. die Aufwandmengen für einen Dünger. Die x -Variable ist in solchen Fällen keine Zufallsvariable, wohl aber der Ertrag als "abhängige" Variable. Die Düngermenge als fixe Größe muss daher als x -Variable verrechnet werden. Als y -Variable ist in solchen Fällen immer diejenige Variable zu verwenden, die eine Zufallsgröße darstellt, hier der Ertrag.

6.2.1 Streuungszerlegung

Wenden wir uns nun der **Streuungszerlegung** für die lineare Regression zu. Die Streuungszerlegung ist die Grundlage der Varianzanalyse für die lineare Regression, die im folgenden hergeleitet wird.

Wir betrachten zunächst die Abweichung einer Beobachtung vom Gesamtmittel. Diese kann wie folgt aufgespalten werden:

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

wobei $\hat{y}_i = a + bx_i$ der durch die Regression an der Stelle x_i vorhergesagte Erwartungswert ist und $\bar{y} = \sum_i y_i / n$. Die Gesamtabweichung lässt sich also aufspalten in eine Abweichung der Beobachtung von der Regression $y_i - \hat{y}_i$ (Fehler) und eine Abweichung des durch die Regression vorhergesagten Wertes vom Gesamtmittel $\hat{y}_i - \bar{y}$ (durch Regression erklärter Teil). Die Aufspaltung der Gesamtstreuung in zwei Komponenten kann auch der Abb. 6.5 entnommen werden.

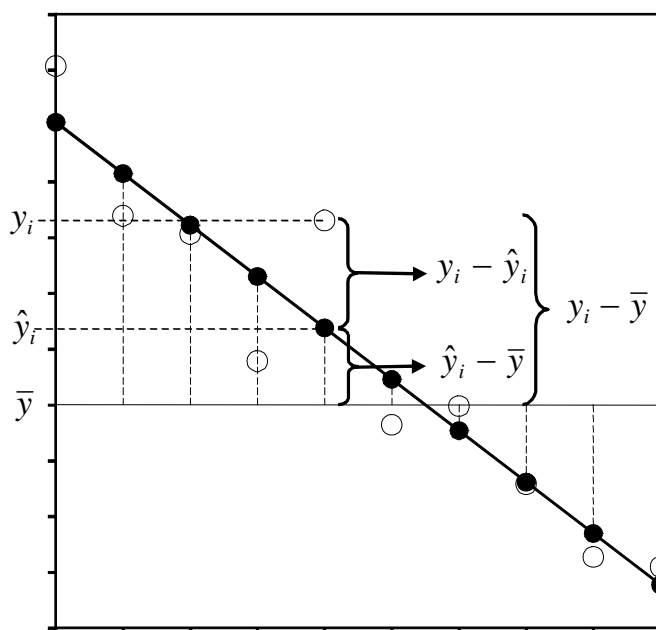


Abb. 6.5: Zerlegung der Gesamtabweichung $y_i - \bar{y}$ in einen durch die Regression erklärten Teil $\hat{y}_i - \bar{y}$ und eine Abweichung von der Regression oder Fehler $y_i - \hat{y}_i$.

Analog zu der einzelnen Abweichung vom Gesamtmittel lässt sich SQ_y aufspalten (**Quadratsummenzerlegung**). Es gilt

$$\begin{aligned} SQ_y &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \end{aligned} \quad (6.2.1)$$

wobei

$$\hat{y}_i = a + bx_i = \bar{y} - b\bar{x} + bx_i = \bar{y} + b(x_i - \bar{x})$$

Die Summe der Kreuzprodukte in (6.2.1) ist gleich Null, wie wir im folgenden sehen. Es ist

$$y_i - \hat{y}_i = y_i - \bar{y} - b(x_i - \bar{x}) \quad \text{und} \quad (6.2.2)$$

$$\hat{y}_i - \bar{y} = b(x_i - \bar{x}) \quad (6.2.3)$$

so dass

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^n [y_i - \bar{y} - b(x_i - \bar{x})]b(x_i - \bar{x}) \\ &= b \sum_{i=1}^n [(y_i - \bar{y})(x_i - \bar{x}) - b(x_i - \bar{x})^2] = b SP_{xy} - b^2 SQ_x \end{aligned}$$

Da $b SP_{xy} = (SP_{xy})^2 / SQ_x = (SP_{xy} / SQ_x)^2 SQ_x = b^2 SQ_x$ ist, folgt, dass $b SP_{xy} = b^2 SQ_x$, so dass die Summe der Kreuzprodukte in (6.2.1) gleich Null ist. Somit ist

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ SQ_y &= SQ_{Fehler} + SQ_{Regression} \end{aligned}$$

Die Gesamtstreuung zerfällt also in einen durch die Regression erklärten Teil ($SQ_{Regression}$) und die Streuung um die Regression, also den unerklärten Teil (SQ_{Fehler}). Zur einfachen Berechnung nutzen wir, dass aus (6.2.3) folgt:

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SQ_{Regression} = b^2 SQ_x$$

Hiermit finden wir SQ_{Fehler} einfach durch Differenzbildung:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = SQ_{Fehler} = SQ_y - b^2 SQ_x$$

$SQ_{Regression}$ ist der durch die Regression erklärte Anteil von SQ_y . Diesen Anteil können wir auch relativ ausdrücken und als **Bestimmtheitsmaß** bezeichnen. Es kann gezeigt werden, dass

$$B = \frac{SQ_{Regression}}{SQ_y} = r^2$$

Das **Bestimmtheitsmaß** für die einfache lineare Regression ist definiert durch

$$B = \frac{SQ_{Regression}}{SQ_y} = r^2$$

wobei

$$SQ_y = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{und} \quad SQ_{Regression} = b^2 SQ_x$$

Es entspricht dem quadrierten Korrelationskoeffizienten (r^2) und gibt den Anteil der gesamten Streuung der Zielvariablen (y) an, der durch die Prädiktorvariable (x) erklärt werden kann.

Die Streuungszerlegung kann in einer Varianzanalyse-Tabelle zusammengefaßt werden:

Varianzanalyse-Tabelle:

Ursache	FG	SQ	MQ	E(MQ)	F
Regression	1	$SQ_{Regression} = b^2 SQ_x$	$SQ_{Regression}$	$\sigma^2 + \beta^2 SQ_x$	$F_{Vers} = \frac{b^2 SQ_x}{s^2}$
Fehler	$n - 2$	$SQ_{Fehler} = SQ_y - b^2 SQ_x$	s^2	σ^2	

wobei

$$s^2 = MQ_{Fehler} = SQ_{Fehler} / (n - 2)$$

Wenn die Steigung $\beta = 0$ ist, so folgt F_{Vers} einer F -Verteilung mit 1 und $(n - 2)$ Freiheitsgraden. Dies führt zu folgender Entscheidungsregel:

Fragestellung: Ist die Steigung der Regression signifikant von Null verschieden?

$$H_0: \beta = 0$$

$$H_A: \beta \neq 0$$

Berechne:

$$F_{Vers} = \frac{b^2 SQ_x}{s^2}$$

mit

$$s^2 = MQ_{Fehler} = SQ_{Fehler} / (n - 2)$$

$$SQ_{Fehler} = SQ_y - b^2 SQ_x$$

Bestimme:

α = Signifikanzniveau

$$F_{Tab} = F_{(1 - \alpha; 1; n - 2)} \quad (\text{Tab. VI})$$

$$F_{Vers} > F_{Tab} \Rightarrow \text{verwerfe } H_0$$

$$F_{Vers} \leq F_{Tab} \Rightarrow \text{behalte } H_0 \text{ bei}$$

Voraussetzung: Abweichungen von der Geraden sind normalverteilt.

Man beachte, dass die zu testende Nullhypothese bezüglich β eine zweiseitige Alternative hat (Abweichung von H_0 , wenn $\beta < 0$ oder $\beta > 0$). Allerdings ist eine einseitige F-Verteilung für den Test zu wählen: Grund: F_{Vers} wird groß wenn β^2 groß wird [siehe $E(MQ_{Regression})$], und dies ist immer dann der Fall, wenn β stark von Null abweicht, egal ob diese Abweichung positiv oder negativ ist. Daher ist der Ablehnungsbereich für F_{Vers} einseitig.

Beispiel: Für die Regendaten finden wir

$$SP_{xy} = 3923,38$$

$$SQ_x = 48335,88$$

$$SQ_y = 804,58$$

$$b = 0,0812$$

$$SQ_{Regression} = 0,0812^2 \cdot 48335,88 = 318,70$$

$$SQ_{Fehler} = 804,58 - 318,70 = 485,88$$

Varianzanalyse-Tabelle:

Ursache	FG	SQ	MQ	F_{Vers}
Regression	1	318,70	318,70	15,74
Fehler	24	485,88	20,25	

$$F_{Tab} = F_{(\alpha = 5\%; 1; 24)} = 4,26 < F_{Vers} = 15,74$$

\Rightarrow Die Steigung ist signifikant von Null verschieden

Das Bestimmtheitsmaß beträgt $B = 318,70/804,58 = 0,396 = 39,6\%$. Alternativ berechnen wir dies auch mittels des in Abschnitt 6.1 ermittelten Korrelationskoeffizienten: $r = 0,629 \Rightarrow B = r^2 = 0,629^2 = 0,396$. Somit kann rund 40% der Ertragschwankungen durch die Regenmenge zwischen April und Juni erklärt werden.

Dem F-Test der obigen Varianzanalyse liegt das folgende **lineare Modell** zugrunde:

$$y_i = \alpha + \beta x_i + e_i$$

wobei

y_i = i -ter Meßwert der Zielvariable

x_i = i -ter Meßwert der Einflussvariable

e_i = Abweichung des i -ten Meßwertes der Zielvariable von der Regressionsgeraden; normalverteilt mit Mittelwert 0 und Varianz σ^2

Die Geradengleichung besagt folgendes: Wenn die Einflussvariable den Wert x_i hat, erwarten wir für die Zielvariable im Mittel den Wert $\alpha + \beta x_i$. Diesen Erwartungswert kann man direkt anhand der graphischen Darstellung der Regressionsgeraden ablesen, wie in Abb. 6.6 gezeigt. Der im Einzelfall beobachtete Wert y_i entspricht aber nicht exakt dem Erwartungswert, sondern er weicht mehr oder weniger stark davon ab. Die Abweichung e_i folgt einer Normalverteilung mit Mittelwert 0 und Varianz σ^2 .

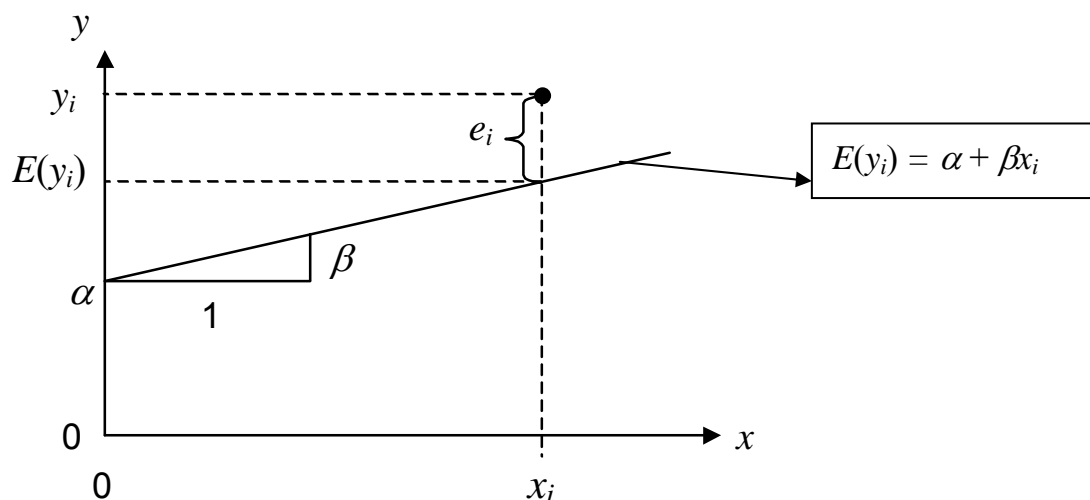


Abb. 6.6: Schematische Darstellung des Modells der linearen Regression.

6.2.2 t-Tests und Vertrauensintervalle

Anstelle des F-Tests der Varianzanalyse kann der Regressionskoeffizient auch mit einem t-Test geprüft werden. Dies liegt daran, dass

$$t = \sqrt{F}$$

einer t-Verteilung folgt, falls die F-Statistik nur einen Freiheitsgrad im Zähler hat, wie in unserem Fall. Der t-Test kann wie folgt durchgeführt werden:

Fragestellung: Ist die Steigung der Regression signifikant von Null verschieden?

$$H_0: \beta = 0$$

$$H_A: \beta \neq 0$$

Berechne:

$$t_{Vers} = \frac{|b| \sqrt{SQ_x}}{s}$$

mit

$$s^2 = SQ_{Fehler} / (n - 2)$$

$$SQ_{Fehler} = SQ_y - b^2 SQ_x$$

Bestimme:

$$t_{Tab}(FG = n - 2; \alpha) \quad (\text{Tab. II, zweiseitig})$$

α = Signifikanzniveau (nicht zu verwechseln mit dem Achsenabschnitt)

$$t_{Vers} > t_{Tab} \Rightarrow \text{verwerfe } H_0$$

$$t_{Vers} \leq t_{Tab} \Rightarrow \text{behalte } H_0 \text{ bei}$$

Voraussetzung: y ist normalverteilt.

Beispiel: Für die Regendaten finden wir

$$SQ_x = 48335,88$$

$$SQ_y = 804,58$$

$$b = 0,0812$$

$$SQ_{Fehler} = 804,58 - 0,0812^2 \cdot 48335,88 = 485,88$$

$$s^2 = 485,88 / 24 = 20,25$$

$$n = 26$$

$$t_{Vers} = \frac{0,0812 \sqrt{48335,88}}{\sqrt{20,25}} = 3,97$$

$$t_{Tab}(FG = 24; \alpha = 5\%) = 2,064 < t_{Vers} \Rightarrow \text{Steigung ist signifikant von Null verschieden}$$

Für dieselben Daten hatten wir in Abschnitt 6.1 bereits $H_0: \rho = 0$ geprüft und denselben t-Wert erhalten ($t_{Vers} = 3,97$)! Dies ist kein Zufall. Wenn die Korrelation Null ist, muss dies auch für den Regressionskoeffizienten gelten. Außerdem kann gezeigt werden, dass

$$t_{Vers} = \frac{|b|\sqrt{SQ_x}}{s} = \frac{|SP_{xy}|\sqrt{SQ_x}}{SQ_x\sqrt{SQ_y - SP_{xy}^2 / SQ_x}}\sqrt{n-2} = \frac{|SP_{xy}|}{\sqrt{SQ_x SQ_y} \sqrt{1 - SP_{xy}^2 / (SQ_x SQ_y)}}\sqrt{n-2}$$

$$= \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2}$$

ist, was der t-Statistik für den Test des Korrelationskoeffizienten entspricht. Beide Tests liefern also immer exakt dasselbe Ergebnis.

Neben einem Test können für verschiedene Größen Vertrauensintervalle berechnet werden, wie der folgende Kasten zeigt:

Vertrauensintervalle

Steigung: $b \pm t_{Tab} \frac{s}{\sqrt{SQ_x}}$

Regressionslinie an Stelle x_0 : $a + bx_0 \pm t_{Tab} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}}$

Vorhersage eines neuen Einzelwertes an Stelle x_0 (Prognoseintervall):

$$a + bx_0 \pm t_{Tab} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}}$$

$t_{Tab}(FG = n - 2; \alpha)$ (Tab. II, zweiseitig)

α = Signifikanzniveau (nicht zu verwechseln mit dem Achsenabschnitt)

Beispiel: Für die Regendaten berechnen wir alle oben genannten Vertrauensintervalle ($\alpha = 5\%$).

Regressionskoeffizient:

$$0,0812 \pm 2,064 \frac{\sqrt{20,25}}{\sqrt{48335,88}} = 0,0812 \pm 0,0422 = (0,0390; 0,1234)$$

Die 95% Vertrauensgrenzen für den wahren Regressionskoeffizienten β sind 0,0390 und 0,1234.

Für die Regressionsgerade sowie für die Vorhersage von y können wir ein Vertrauensband berechnen, indem für verschiedene Werte x_0 die Vertrauensgrenzen berechnet und dann zu Kurven verbunden werden. Dies ist von Hand sehr mühsam, so dass wir hier nur einen Punkt $x_0 = 200$ betrachten.

Regressionsgerade:

$$SQ_x = 48335,88$$

$$b = 0,0812$$

$$a = 15,14$$

$$a + bx_0 = 15,14 + 0,0812 \cdot 200 = 31,37$$

$$t_{Tab} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}} = 2,064 \cdot \sqrt{20,25} \sqrt{\frac{1}{26} + \frac{(200 - 139,35)^2}{48335,88}} = 9,288 \sqrt{0,1178} = 3,15$$

Grenzen: $31,37 - 3,15 = 28,20$ und $31,37 + 3,15 = 34,50$.

Vorhersage eines neuen Wertes:

$$t_{Tab} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}} = 2,064 \cdot \sqrt{20,25} \sqrt{1 + \frac{1}{26} + \frac{(200 - 139,35)^2}{48335,88}} = 9,288 \sqrt{1,1178} = 9,81$$

Grenzen: $31,37 - 9,81 = 21,54$ und $31,37 + 9,81 = 41,16$.

Die Grenzen für die Vorhersage sind viel weiter als die Grenzen für die Regression. Der Grund ist, dass die Regressionslinie nur den Erwartungswert an einer Stelle x_0 wiedergibt, während bei der Vorhersage eines neuen Wertes noch die Streuung der Einzelwerte um diesen Erwartungswert in Rechnung zu stellen ist. In den Abbildungen 6.7 und 6.8 sind für beide Fälle Vertrauensbänder für beliebige Werte x_0 gezeigt.

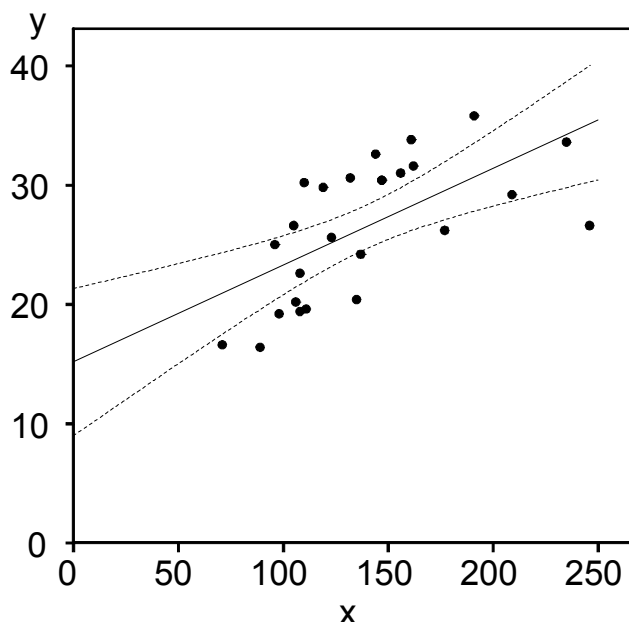


Abb. 6.7: Regressionsgerade für Regendaten mit 95%-Vertrauensintervall für den Verlauf der Regressionsgerade.

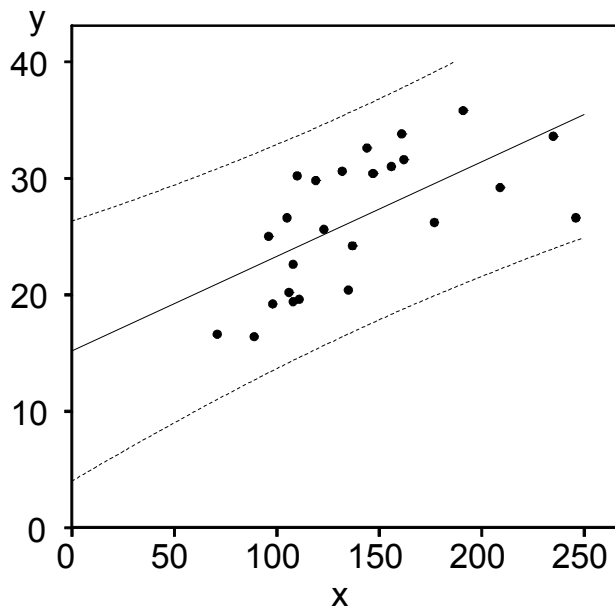


Abb. 6.8: Regressionsgerade für Regendaten mit 95%-Vertrauensintervall für Vorhersage eines neuen y -Wertes bei gegebenem x -Wert.

Wichtig ist hier der Hinweis, dass das Vertrauensband (Konfidenzband) in den Abbildungen 6.7 und 6.8 kein simultanes Konfidenzband ist, welches simultan, also versuchsbezogen, an allen x -Stellen den Fehler 1. Art einhalten wurde. Hierzu wären ähnliche Modifikationen notwendig, wie die des Tukey-Test gegenüber des LSD-Tests (siehe Kap. 4.5).

Extrapolation: Im obigen Beispiel haben wir eine Vorhersage des Ertrages mit der Regenmenge betrachtet. Eine Vorhersage ist nur für solche x -Werte sinnvoll und zulässig, die sich im Bereich der beobachteten x -Werte befinden. Man spricht hierbei von **Interpolation**. Es ist dagegen nicht zulässig, eine Vorhersage für x -Werte außerhalb des beobachteten Bereichs zu machen (**Extrapolation**). Es ist beispielsweise nicht zulässig, eine Vorhersage für $x = 0$ mm Regen zu machen, weil dieser x -Wert deutlich unterhalb der beobachteten x -Werte liegt. Ob der Ertragsverlauf bei niedrigen Niederschlägen tatsächlich linear verläuft, was für diese Voraussage (Extrapolation) angenommen werden müßte, können wir nicht nachweisen, da in diesem Bereich keine x -Werte beobachtet wurden (Dies gilt ungeachtet der Tatsache, dass wir hier aus biologischen Gründen natürlich $y = 0$ bei $x = 0$ erwarten würden).

Regression zum Mittel: Man beachte, dass eine angepasste Regressionsgerade immer durch den Punkt (\bar{x}, \bar{y}) gehen muss. Dies folgt aus der Tatsache, dass der für die Stelle \bar{x} vorhergesagte Wert gleich \bar{y} ist, denn es gilt:

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

Schauen wir uns nun den Mittelwert und die Varianz der für die beobachteten Datenpunkte vorhergesagten Werte

$$\hat{y}_i = a + bx_i$$

an. Wir finden

$$\bar{\hat{y}} = \frac{\sum_{i=1}^n \hat{y}_i}{n} = \frac{\sum_{i=1}^n (a + bx_i)}{n} = a + b\bar{x} = \bar{y} - b\bar{x} + b\bar{x} = \bar{y}$$

Die vorhergesagten y -Werte haben also denselben Mittelwert wie die beobachteten y -Werte! Außerdem finden wir

$$s_{\hat{y}}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{n-1} = \frac{\sum_{i=1}^n (a + bx_i - a - b\bar{x})^2}{n-1} = \frac{b^2 SQ_x}{n-1} = \frac{(SP_{xy})^2}{(n-1)SQ_x} = \frac{(SP_{xy})^2}{(n-1)SQ_x SQ_y} SQ_y = r^2 s_y^2$$

Da $r^2 \leq 1$ ist, gilt $s_{\hat{y}}^2 \leq s_y^2$. Die vorhergesagten y -Werte streuen also weniger als die beobachteten y -Werte! Da beide denselben Mittelwert \bar{y} haben, müssen die vorhergesagten Werte tendenziell näher beim Mittel liegen als die Originalwerte. Man spricht in diesem Zusammenhang von der **"Regression zum Mittel"** (Rückfall oder Rückschritt zum Mittel).

Beispiel: Die "Regression zum Mittel" wurde erstmals von Francis Galton in der Genetik beobachtet. Er untersuchte den Zusammenhang zwischen der Körpergröße von Eltern und ihren Kindern (im Erwachsenenalter). Er fand einen linearen Zusammenhang. Die Steigung der Regression war jedoch nicht genau Eins, wie man hätte erwarten können, sondern kleiner als Eins. Die Kinder wichen im Schnitt weniger stark vom Mittel ab als ihre Eltern. Es hatte also offenbar eine "Regression" zurück zum Mittel stattgefunden (Lynch, M. und Walsh, B. 1998, Genetics and the analysis of quantitative traits, Sinauer, Sunderland).

Die oben abgeleitete Beziehung $s_{\hat{y}}^2 = r^2 s_y^2$ hat noch eine weitere Interpretation: Von der Gesamtstreuung der beobachteten Werte (s_y^2) wird immer nur ein Teil durch die Regression erklärt. Dieser Anteil ist gegeben durch die Varianz der vorhergesagten Werte ($s_{\hat{y}}^2$). Der relative Anteil der erklärten Streuung ist $B = r^2$ (Bestimmtheitsmaß), wie wir bereits weiter oben festgestellt hatten.

6.2.3 Inverse Regression

Bisher hatten wir die Regression u.a. benutzt, um y für einen gegebenen Wert von x vorherzusagen. Dabei ist es in vielen Fällen so, dass die Stufen von x fest vorgegeben sind, während y eine Zufallsvariable ist. Es gibt aber Fälle, in denen die umgekehrte Zielstellung vorliegt, man also für gegebenes y den zugehörigen x -Wert schätzen möchte. Man spricht hier auch von Kalibration. Man kann in solchen Fällen aber die Rollen von y und x nicht einfach vertauschen, wenn es weiterhin so ist, dass die Stufen von x fest vorgegeben werden und y die Zufallsvariable ist. Denn durch vertauschen hätte man dann als Zielvariable plötzlich eine feste Größe und damit keine Zufallsvariable mehr, während die Einflussvariablen dann Zufallsvariablen

wären. Man wird dann daher weiterhin eine Regression von y auf x durchführen, muss aber bei der Vorhersage von x bei gegebenem y ein anderes Verfahren anwenden als im umgekehrten Fall.

Beispiel: In einer Laboruntersuchung soll die Konzentration von Leucin colorimetrisch bestimmt werden. Hierzu wird eine Verdünnungsreihe mit jeweils bekannter Leucin-Konzentration (x) erzeugt. Für jede Konzentration wird dann die Extinktion (y) mittels eines Colorimeters bestimmt. Es wird dann mit Hilfe der linearen Regression eine Eichkurve bestimmt. Diese dient dazu, an neuen Proben anhand der gemessenen Extinktion (y) die Konzentration von Leucin (x) zu schätzen (Linder, Elementare statistische Methoden, S. 176).

Leucin-Konzentration x (mmol)	Extinktion y
0,02	0,08
0,05	0,15
0,10	0,29
0,30	0,88
0,40	1,13
0,50	1,42
0,60	1,69

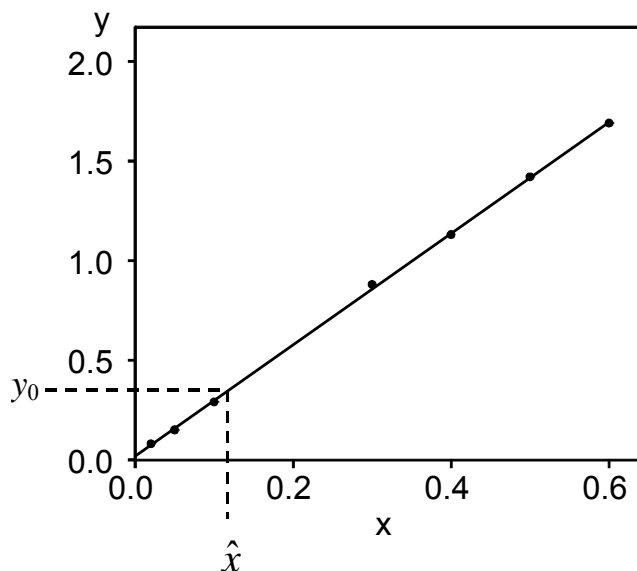


Abb. 6.9: Leucindaten mit Veranschaulichung des inversen Regressionsproblems.

Während wir bisher die Prognose für y bei gegebenem x -Wert x_0 nach

$$\hat{y} = a + bx_0$$

durchgeführt haben, ist das Problem nun umgekehrt. Zur Schätzung von x lösen wir für gegebenen Wert y_0 die Gleichung

$$y_0 = a + b\hat{x}$$

nach x auf und finden

$$\hat{x} = \frac{y_0 - a}{b}$$

Für einen gegebenen y -Wert y_0 schätzt man den x -Wert nach

$$\hat{x} = \frac{y_0 - a}{b}$$

Beispiel: Für die Leucindaten finden wir

$$a = 0,018525 \text{ und } b = 2,797120$$

Wird nun in einer neuen Probe eine Extinktion von 0,35 gemessen, so schätzen wir die Leucinkonzentration nach

$$\hat{x} = \frac{0,35 - 0,018525}{2,797120} = 0,119$$

Die geschätzte Leucinkonzentration beträgt also 0,119 mmol.

Für die Berechnung eines Vertrauensintervalls für die Prognose von x wurde von Fieller (sprich: "Feiler") eine Methode vorgeschlagen. Diese nutzt aus, dass für ein gegebenes x_0 der vorhergesagte Wert

$$a + bx_0$$

einer Normalverteilung folgt mit Erwartungswert

$$\alpha + \beta x_0$$

und Varianz

$$\sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x} \right)$$

Eine neue Beobachtung y_0 an der Stelle x_0 hat denselben Erwartungswert

$$\alpha + \beta x_0$$

und die Varianz σ^2 .

Deswegen hat die Differenz

$$y_0 - (a + bx_0)$$

den Erwartungswert Null und die Varianz

$$\sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x} \right)$$

Die Statistik

$$t = \frac{y_0 - (a + bx_0)}{s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}}}$$

hat daher eine t -Verteilung mit $n-2$ Freiheitsgraden. Ein $(1-\alpha)100\%$ -Vertrauensintervall für x_0 ist nun gegeben durch alle Werte x_0 , welche die Ungleichung

$$|t| = \frac{|y_0 - (a + bx_0)|}{s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}}} = \frac{|y_0 - \bar{y} - b(x_0 - \bar{x})|}{s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}}} \leq t_{Tab}$$

erfüllen, wobei t_{Tab} der kritische Wert der t -Verteilung mit $n-2$ Freiheitsgraden zum Signifikanzniveau α ist. Die resultierenden Vertrauensgrenzen sind gegeben durch die Lösung der Gleichung

$$\frac{|y_0 - \bar{y} - b(x_0 - \bar{x})|}{s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}}} = t_{Tab}$$

nach x_0 . Einfacher ist es, die Gleichung zunächst nach

$$d = x_0 - \bar{x}$$

zu lösen und hieraus die Grenzen für x_0 zu berechnen. Wichtig ist, dass eine Lösung nur existiert, wenn $H_0: \beta = 0$ zum Niveau α verworfen wurde. Die Lösungen d_1 und d_2 sind dann Lösungen der folgenden quadratischen Gleichung:

$$d^2 \left(b^2 - \frac{(t_{Tab})^2 s^2}{SQ_x} \right) - 2db(y_0 - \bar{y}) + \left\{ (y_0 - \bar{y})^2 - (t_{Tab})^2 s^2 \left(1 + \frac{1}{n} \right) \right\} = 0$$

Die Vertrauensgrenzen sind dann

$$[\bar{x} + d_1, \bar{x} + d_2]$$

Wir fassen das Ganze in einem Formelkasten zusammen.

Fragestellung: Für einen gegebenen y -Wert y_0 wurde der zugehörige x_0 Wert geschätzt. Für die Schätzung soll ein Vertrauensintervall berechnet werden.

Rechenweg:

Löse die quadratische Gleichung

$$d^2 \left(b^2 - \frac{(t_{Tab})^2 s^2}{SQ_x} \right) - 2db(y_0 - \bar{y}) + \left\{ (y_0 - \bar{y})^2 - (t_{Tab})^2 s^2 \left(1 + \frac{1}{n} \right) \right\} = 0$$

nach d , wobei $t_{Tab}(FG = n - 2; \alpha)$ der kritische t-Wert, α das Signifikanzniveau, und $s^2 = MQ_{Fehler}$ der übliche Varianzschätzer aus der Varianzanalyse ist (Tab. II, zweiseitig). Bezeichne die Lösungen mit d_1 und d_2 .

Die $100(1-\alpha)$ -Vertrauensgrenzen sind gegeben durch

$$[\bar{x} + d_1, \bar{x} + d_2]$$

Voraussetzung: $H_0: \beta = 0$ wurde zum Niveau α verworfen.

Beispiel: Für die Leucindaten finden wir für $y_0 = 0,35$

$$b = 2,79712$$

$$n = 7$$

$$SQ_x = 0,31849$$

$$s^2 = 0,00015535$$

$$\bar{x} = 0,28143$$

$$\bar{y} = 0,80571$$

$$t_{Tab}(\alpha=5\%; FG = 5) = 2,571$$

$$d^2 \left(2,79712^2 - \frac{2,571^2 0,00015535}{0,31849} \right) - 2 \cdot d \cdot 2,79712 \cdot (0,35 - 0,80571) + \left\{ (0,35 - 0,80571)^2 - 2,571^2 0,00015535 \left(1 + \frac{1}{7} \right) \right\} = 0$$

$$7,82066d^2 + 2,54935d + 0,20650 = 0$$

Eine quadratische Gleichung der Form

$Ad^2 + Bd + C = 0$ hat die Lösungen

$$d_{1,2} = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$

Also finden wir

$$d_{1,2} = \frac{-2,54935 \pm \sqrt{2,54935^2 - 4 \cdot 7,82066 \cdot 0,20650}}{2 \cdot 7,82066}$$

$$d_1 = -0,17568$$

$$d_2 = -0,15030$$

Die Vertrauensgrenzen für x sind

$$0,28143 - 0,17568 = 0,10575 \text{ und}$$

$$0,28143 - 0,15030 = 0,13113$$

Mit 95%iger Wahrscheinlichkeit liegt bei einem Extinktionswert von $y_0 = 0,35$ die Leucinkonzentration zwischen 0,106 und 0,131 mmol.

6.3 Vergleich von Korrelation und Regression

Im folgenden werden zusammenfassend einige Eigenschaften von Regression und Korrelation gegenübergestellt.

(1) Sowohl Korrelation als auch Regression eignen sich nur zur Erfassung **linearer Zusammenhänge**. Sie sind nicht geeignet, nichtlineare Zusammenhänge aufzudecken. Beide machen die Voraussetzung der Normalverteilung zumindest einer der beiden Variablen.

(2) Die Korrelation beschreibt den Zusammenhang zwischen zwei Variablen mit einer **dimensionslosen Maßzahl**. Die Daten müssen einer bivariaten Normalverteilung folgen. Sowohl x als auch y müssen als zufallsverteilt betrachtet werden können.

(3) Die Regression beschreibt, wie sich y ändert, falls x um eine Einheit erhöht wird. Die Zielvariable y muss hier einer Normalverteilung folgen und als zufallsverteilt betrachtet werden können. Die Einflussvariable x kann dagegen (muss aber nicht; siehe Regendaten) sogar in einem Experiment systematisch variiert werden. In diesem Fall ist nur die Regression, nicht aber die Korrelation ein geeignetes Auswertungsverfahren.

Beispiel: In einem Feldversuch wird die Wirkung einer Mistdüngung (x) auf den Ertrag von Weizen (y) geprüft. Hierzu werden systematisch die Gaben 0, 2, 4, 6, 8 und 10 kg pro Parzelle geprüft. Dieser Versuch kann mittels Regression, nicht aber mittels Korrelation ausgewertet werden.

(4) Die Güte der Anpassung einer Regressionsgeraden kann durch das Bestimmtheitsmaß $B = r^2$ beschrieben werden. Das Bestimmtheitsmaß gibt an, welcher Teil der Gesamtstreuung der y -Variable durch die Regression auf x erklärt werden kann. Die Korrelation r sollte dagegen nicht zur Beschreibung der Güte der Anpassung einer Regression verwendet werden, insbesondere dann nicht, wenn die Einflussvariable x systematisch variiert wurde.

(5) Die Regression kann für Prognosen herangezogen werden, die Korrelation nicht.

(6) Der Regressionskoeffizient b sagt nichts über die Stärke des Zusammenhanges aus, wohl aber das Bestimmtheitsmaß.

(7) In vielen Lehrbüchern findet man die Aussage, eine Regression solle nur durchgeführt werden, wenn eine einseitige Ursache-Wirkungs-Beziehung der Form $x \rightarrow y$ bestehe. Diese Aussage ist m.E. nicht zutreffend. Wichtig ist aber, dass die vorherzusagende Variable als y gewählt wird.

Beispiel: Bei Futter- und Nahrungsgetreide wird ein möglichst geringer Besatz mit Pilztoxinen gefordert. Manche Toxine, wie das Ergosterol, sind sehr leicht zu messen, während andere, wie das Nivalenol, nur sehr schwer messbar sind. Es bietet sich daher an, für eine Zahl von Proben beide Toxine zu bestimmen und dann eine Regression des Ergosterolgehaltes (x) auf den Nivalenolgehalt (y) durchzuführen. Die Regressionsgerade kann dann genutzt werden, um in Routineuntersuchungen, bei denen aus Kostengründen nur das Ergosterol (x) gemessen wird, den Gehalt an Nivalenol (y) vorherzusagen. Obwohl hier die Regression sehr sinnvoll eingesetzt werden kann, besteht keine eindeutige Kausalbeziehung zwischen den beiden Variablen. Beide sind vielmehr kausal beeinflusst von der Stärke des Pilzbefalls.

Beispiel: Bei Bäumen findet man oft eine sehr enge Beziehung zwischen Brusthöhenumfang des Stammes und seines Volumens. Das Stammvolumen ist von unmittelbarem ökonomischen Interesse, lässt sich aber nur mit großem Aufwand vor dem Fällen des Baumes bestimmen. Daher wird die enge Beziehung zum Brusthöhenumfang zur indirekten Messung des Volumens herangezogen. Hierbei wird eine Regression des Volumens (y) auf den Brusthöhenumfang (x) durchgeführt und die Regressionsgleichung anschließend für eine Volumenprognose herangezogen. Hiermit ist in keiner Weise eine Ursache-Wirkungs-Beziehung zwischen Brusthöhenumfang und Volumen impliziert, und eine solche Beziehung gibt es auch nicht.

6.4 Nichtlineare Regression durch Transformation der Variablen

In Abschnitt 6.2 wurde die lineare Regression behandelt. Hierbei wurde ein zumindest annähernd linearer Zusammenhang angenommen. Es gibt viele Zusammenhänge, die sich eher durch eine nichtlineare Funktion beschreiben lassen. Mit Hilfe der Methode der Kleinsten Quadrate können prinzipiell beliebige, nichtlineare Funktionen angepasst werden. Es gibt hierzu verschiedene Methoden, die an dieser Stelle nicht umfassend dargestellt werden können. Am einfachsten liegt der Fall, wenn sich das Modell, welches den nichtlinearen Zusammenhang beschreibt, durch einfache Transformationen linearisieren lässt, d.h., in ein Modell überführen lässt, welches linear in x ist. Dies hat den Vorteil, dass die in Abschnitt 6.2 beschriebenen Methoden verwendet werden können. Dieses einfache Verfahren soll hier besprochen werden, während aufwendigere Verfahren (Polynomregression, eigentliche nichtlineare Regression) später behandelt werden.

Beispiel: An Pflanzen der Reissorte IR8 wurden die Licht-Transmissions-Rate (y) sowie der Blattflächenindex (x) gemessen (Gomez und Gomez, 1984, S. 390).

y	x	$\log(y)$
75,0	0,50	4,31749
72,0	0,60	4,27667
42,0	1,80	3,73767
29,0	2,50	3,36730
27,0	2,80	3,29584
10,0	5,45	2,30259
9,0	5,60	2,19722
5,0	7,20	1,60944
2,0	8,75	0,69315
2,0	9,60	0,69315
1,0	10,40	0,00000
0,9	12,00	-0,10536

Die Daten sind in Abb. 6.10(a) graphisch dargestellt. Der Zusammenhang ist eindeutig nichtlinear. Das Streudiagramm legt nahe, dass der systematische Anteil des Zusammenhangs durch eine Exponentialfunktion der Form

$$\eta = \alpha \exp(\beta x)$$

zu beschreiben ist, wobei "exp" die Exponentialfunktion zur Basis e (Euler'sche Konstante = 2,7182...) ist. Dieses Modell lässt sich durch Logarithmieren linearisieren:

$$\log(\eta) = \log(\alpha) + \beta x = \alpha' + \beta x$$

mit $\alpha' = \log(\alpha)$. Letztere Bezeichnung wird eingeführt, damit man besser erkennt, dass es sich um ein lineares Modell handelt. In diesem Skript ist mit $\log(\cdot)$ immer der Logarithmus zur Basis e (Eulersche Konstante = 2,71828...) gemeint. Vor dem Gleichheitszeichen steht hier "eta" (η) für den systematischen Teil des Modells, und nicht y , weil das Modell noch keinen Fehlerterm hat.

Die Tatsache, dass das exponentielle Modell sich durch Logarithmieren linearisieren lässt, lässt erwarten, dass wir auch für die logarithmierten Daten (y) ein lineares Modell annehmen können. Wir logarithmieren daher die Daten:

$$y' = \log(y)$$

und passen das lineare Modell

$$y' = \log(\eta) + e = \alpha' + \beta x + e$$

mit der Methode der Kleinsten Quadrate aus Abschnitt 6.2 an [Abb. 6.10(b)]. Für die Durchführung dieser Regression benötigen wir also lediglich $\log(y)$ und x , nicht aber die ursprünglichen y -Werte. Man beachte, dass jetzt ein Fehlerterm hinzugefügt wurde, weil wir jetzt y statt η betrachten. Die Schätzwerte für α' und β sind:

$$a' = 4,458 \text{ und}$$

$$b = -0,4034$$

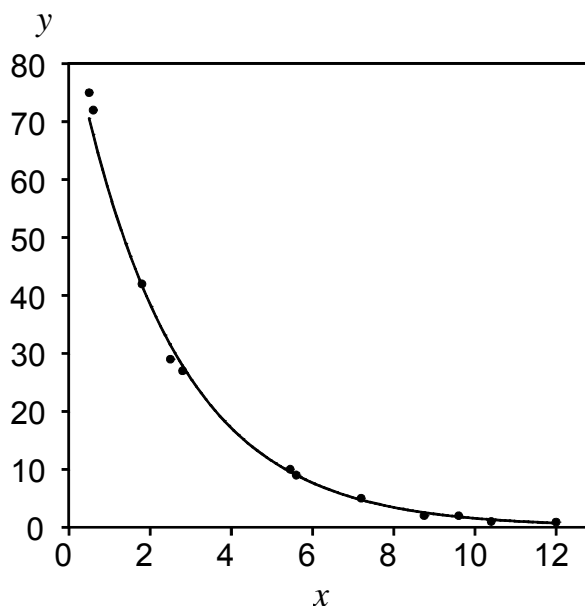


Abb. 6.10(a): Plot der Licht-Transmissions-Rate (y) gegen den Blattflächenindex (x).

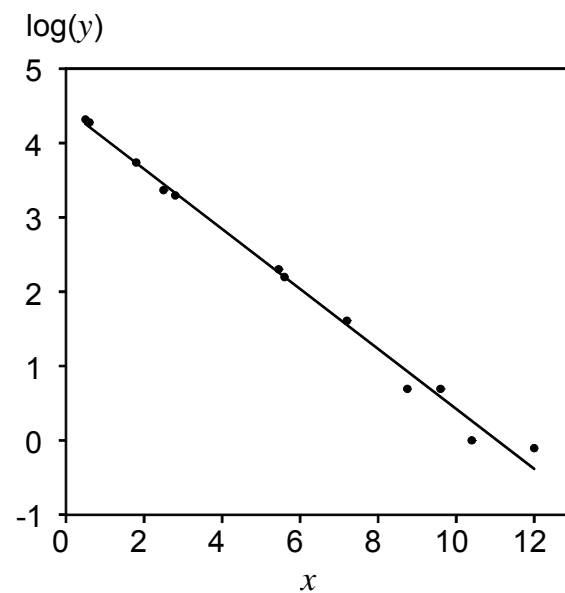


Abb. 6.10(b): Plot der logarithmierten Licht-Transmissions-Rate [$\log(y)$] gegen den Blattflächenindex (x).

Um den Schätzwert für α zu finden, muss zurücktransformiert werden:

$$a = \exp(a') = \exp(4,458) = 86,31$$

Das hierdurch angepasste exponentielle Modell lautet:

$$\hat{y} = 86,31 \exp(-0,4034x)$$

Mit dieser geschätzten Funktion wurde die Kurve in Abb. 6.10(a) gezeichnet.

Rechnen mit Logarithmen - kurze Rekapitulation

Die Funktion $\log(\cdot)$ bezeichnet den natürlichen Logarithmus zur Basis e (Eulersche Konstante = 2,7182...)

$$c = a \cdot b \quad \Leftrightarrow \log(c) = \log(a) + \log(b)$$

$$c = a^b \quad \Leftrightarrow \log(c) = b \cdot \log(a)$$

$$b = \exp(a) = e^a \quad \Leftrightarrow \log(b) = a \cdot \log(e) = a$$

Beispiel: Ratkowski (1983, S. 52) beschreibt ein mehrortiges Experiment mit Zwiebeln zur Ermittlung der Abhängigkeit von Einzelpflanzenenertrag (y ; g pro Pflanze) von der Pflanzendichte (x ; Pflanzen pro m^2). Die Daten für den Ort Virginia sind wie folgt:

x	y	x	y	x	y	x	y
18,78	272,15	38,55	154,10	61,78	96,52	92,91	70,93
21,25	235,23	39,54	124,17	61,78	94,71	101,81	60,99
23,23	180,47	39,54	146,28	63,75	99,86	103,78	74,09
27,18	177,31	41,02	105,47	67,71	93,37	115,15	49,45
30,15	141,28	42,50	139,24	71,66	89,78	123,06	56,65
31,67	169,39	43,98	148,31	77,59	69,34	144,31	47,84
32,12	138,17	45,47	110,44	80,56	73,74	155,68	40,03
32,62	171,81	49,92	90,72	86,49	75,17	158,15	38,70
32,62	112,02	50,90	102,62	88,46	72,98		
33,61	156,09	53,87	107,36	89,45	79,94		
37,07	137,29	57,82	92,66	90,93	79,13		

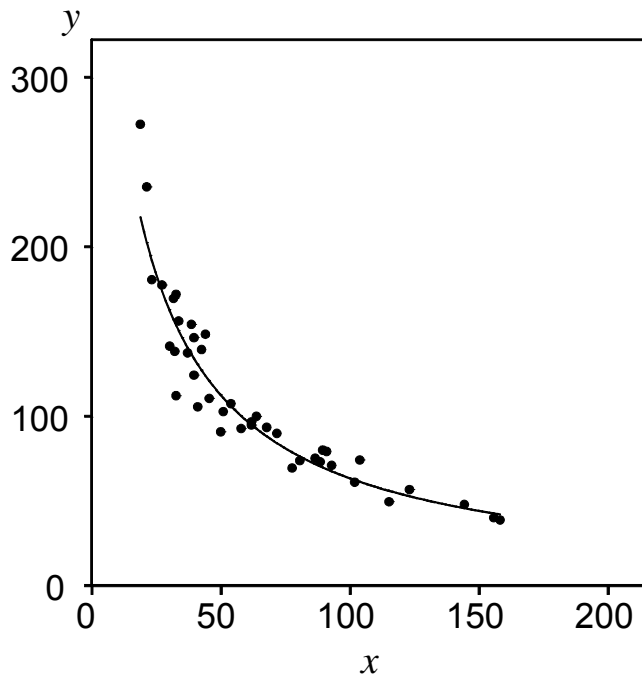


Abb. 6.11(a): Plot des Einzelpflanzenenertrages (y) gegen die Pflanzendichte (x).

Für diese Daten kann das folgende Modell angepasst werden:

$$\eta = \frac{1}{\alpha + \beta x}$$

Der Parameter α hat folgende biologische Interpretation: $1/\alpha$ ist der Ertrag der Einzelpflanze, wenn der Standraum unendlich groß wird, wenn also $x = 0$ wird. Es ist somit als das Ertragspotential der Pflanze bei Abwesenheit von Konkurrenz zu interpretieren. Die folgende inverse Transformation linearisiert das Modell:

$$\frac{1}{\eta} = \alpha + \beta x$$

Daher führen wir eine lineare Regression von $1/y$ gegen x durch und passen folgendes Modell an:

$$y' = \frac{1}{y} = \alpha + \beta x + e$$

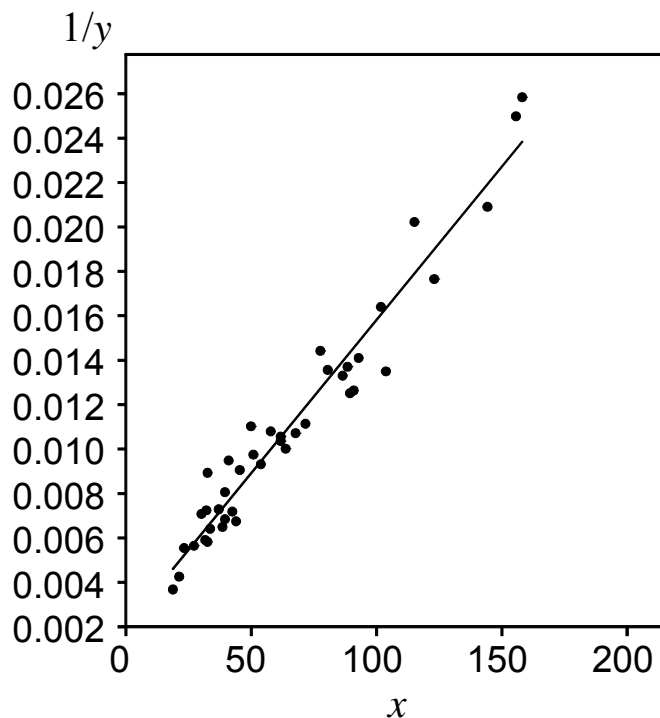


Abb. 6.11(b): Plot von $1/y$ gegen die Pflanzendichte (x).

Wir finden:

$$a = 0,002009$$

$$b = 0,000138$$

Damit ist der maximale Einzelpflanzenenertrag gleich $1/a = 1/0,002009 = 497,76$.

Wir können ein Vertrauensintervall für α berechnen, woraus durch Transformation ein Intervall für $1/\alpha$ erhalten wird. Man beachte, dass α der Achsenabschnitt ist. Dies ist der Wert der Regression bei $x_0 = 0$. Aus Abschnitt 6.2 wissen wir, dass ein Vertrauensintervall für die Regressionsgerade an Stelle x_0 gegeben ist durch:

$$a + bx_0 \pm t_{Tab} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}}$$

Für den Spezialfall $x_0 = 0$ (Regressionsgerade schneidet die y -Achse = Achsenabschnitt) finden wir

$$a \pm t_{Tab} s \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{SQ_x}}$$

In unserem Fall haben wir

$$a = 0,002009$$

$$n = 41$$

$$s = 0,00122$$

$$\bar{x} = 64,83951$$

$$SQ_x = 54466,11$$

$$t_{Tab} = 2,023$$

Das 95%-Vertrauensintervall ist

$$0,002009 \pm 2,023 * 0,00122 \sqrt{\frac{1}{41} + \frac{64,83951^2}{54466,11}} = 0,002009 \pm 0,000787 = (0,001222; 0,002796)$$

Die Grenzen für $1/\alpha$ sind somit

$$1/0,001222 = 818 \text{ und}$$

$$1/0,002796 = 358$$

Der wahre Maximalertrag je Einzelpflanze liegt mit 95% Wahrscheinlichkeit zwischen 358 g und 818 g.

Beispiel: Die Reaktionseigenschaften von Enzymen werden untersucht, indem das Enzym mit verschiedenen Konzentrationen des Substrats versetzt wird und die Reaktionsgeschwindigkeit v gemessen wird. Die untenstehende Tabelle zeigt den Fall eines Enzyms, welches Zucker in eine andere Verbindung umwandeln kann (Christensen HN, Palmer GA, 1974, Lehrprogramm Enzymkinetik, Verlag Chemie, Weinheim, S.37). Die Reaktionsgeschwindigkeit des Enzyms wurde für zwei Zucker untersucht: Glucose und Mannose.

Zucker-Konzentration $10^{-4} \text{ mol l}^{-1}$	v_{Glucose} 10^3 min^{-1}	v_{Mannose} 10^3 min^{-1}
x	y	y
0,1	0,150	0,082
0,2	0,256	0,150
1,0	0,600	0,450
3,0	0,770	0,670
5,0	0,818	0,750

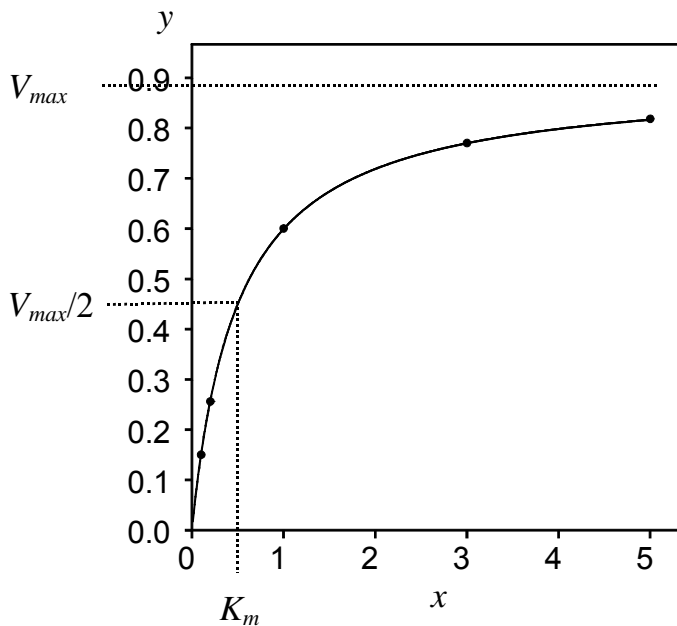


Abb. 6.12(a): Plot der Reaktionsgeschwindigkeit (y) gegen die Substratkonzentration (x) für Glucose.

Die Daten für Glucose sind in Abb. 6.12(a) wiedergegeben. Bei sog. Reaktionen erster Ordnung kann die Abhängigkeit der Reaktionsgeschwindigkeit y von der Konzentration x des Substrates durch die sog. Michaelis-Menten-Gleichung beschrieben werden:

$$\eta = \frac{x \cdot V_{\max}}{K_m + x}$$

wobei K_m die Michaelis-Menten-Konstante ist und V_{\max} die maximale Reaktionsgeschwindigkeit. Die Konstante K_m gibt an, bei welcher Konzentration die halbe Reaktionsgeschwindigkeit erreicht ist [siehe Abb. 6.12(b)]:

$$\frac{V_{\max}}{2} = \frac{x \cdot V_{\max}}{K_m + x} \Leftrightarrow x = K_m$$

Sie ist ein Maß für die Affinität zwischen Substrat und Enzym. Die beiden Konstanten K_m und V_{\max} werden zur Charakterisierung von Enzym-Substrat-Relationen herangezogen.

Die Michaelis-Menten-Gleichung kann durch folgende Umformung in eine lineare Gleichung umgewandelt werden:

$$\eta = \frac{x \cdot V_{\max}}{K_m + x} \Leftrightarrow \frac{1}{\eta} = \frac{1}{V_{\max}} + \frac{K_m}{x \cdot V_{\max}} = \alpha + \beta \frac{1}{x}$$

mit

$$\alpha = \frac{1}{V_{\max}} \quad \text{und} \quad \beta = \frac{K_m}{V_{\max}}$$

Eine Plot von $1/y$ gegen $1/x$ sollte daher annähernd linear sein, wenn dieses Modell passt. Dies trifft im vorliegenden Beispiel zu, wie Abb. 6.12(b) zeigt.

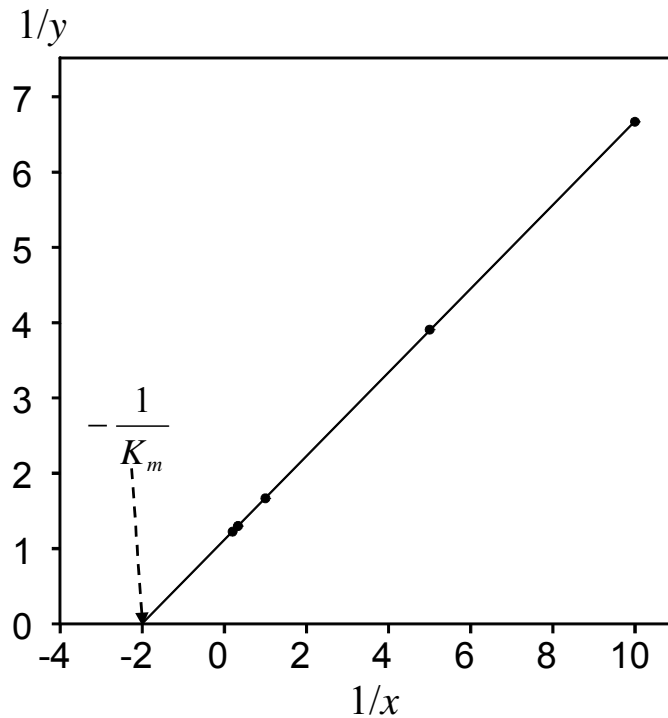


Abb. 6.12(b): $1/y$ gegen $1/x$.

Eine lineare Regression von $1/y$ gegen $1/x$ liefert folgende Schätzwerte:

$$a = 1,114129$$

$$b = 0,555855$$

Hiermit finden wir folgende Schätzungen:

$$V_{\max} = 1/a = 1/1,114129 = 0,89756$$

$$K_m = b \cdot V_{\max} = 0,555855 \cdot 0,89756 = 0,4989$$

Mit diesen Parameterwerten wurde die Kurve in Abb. 6.12(a) gezeichnet.

Zum Vergleich haben wir die Parameter für Mannose geschätzt:

$$a = 1,117814$$

$$b = 1,108114$$

Hiermit finden wir folgende Schätzungen für V_{\max} und K_m :

$$V_{\max} = 1/a = 1/1,117814 = 0,89603$$

$$K_m = b \cdot V_{\max} = 1,108114 \cdot 0,89603 = 0,9913$$

Offenbar ist die Affinität des Enzyms zu Mannose ($K_m = 0,9913$) geringer als zu Glucose ($K_m = 0,4989$).

Ein Vertrauensintervall für K_m kann mit den Methoden in Abschnitt 6.2.3 berechnet werden. Hierzu beachte man, dass der Schnittpunkt der Geraden in Abb. 6.12(b) mit der Abszisse ($1/x$) gleich

$$-\frac{1}{K_m}$$

ist. Dies sieht man, indem

$$\frac{1}{\eta} = \frac{1}{V_{\max}} + \frac{K_m}{x \cdot V_{\max}} = 0$$

gesetzt wird, was nach Auflösen folgendes ergibt:

$$\frac{1}{x} = -\frac{1}{K_m}$$

Daher können wir die Methoden in Abschnitt 6.2.3 verwenden, um zunächst ein Vertrauensintervall für $-1/K_m$ zu berechnen, welches dann transformiert wird, um ein Intervall für K_m zu erhalten. Die Berechnung wird hier aus Platzgründen nicht aufgeführt.

Beispiel: Mead R, Curnow RN und Hasted AM (1993, Statistical methods in agriculture and experimental biology, S. 336) beschreiben ein Laborexperiment zur Bestimmung der sog. LD_{50} eines Insektizides. Die LD_{50} ist diejenige Dosis, bei welcher 50% der Insekten sterben. Somit ist die LD_{50} ein Maß für die Wirksamkeit des Insektizides. Das Insektizid wird im Labor in verschiedenen Dosen gegen Insekten eingesetzt (Larven in diesem Fall). Die Dosen werden in gleichem Abstand auf einer logarithmischen Skala gewählt, weil sich herausgestellt hat, dass auf dieser Skala die Modellierung oft einfach ist. Bei jeder Dosis wird das Insektizid auf $m = 20$ Larven angewendet und die Zahl der getöteten Larven ausgezählt.

Dosis (%) (x)	Beobachtete Zahl toter Larven (z)	Beobachtete Mortalität (y = z/m)	Log-Dosis [x' = log(x)]	Empirische Logits (y' = elogit)
0,375	0	0,00	-0,98083	-3,71357
0,75	1	0,05	-0,28768	-2,56495
1,5	8	0,40	0,40547	-0,38566
3	11	0,55	1,09861	0,19106
6	16	0,80	1,79176	1,29928
12	18	0,90	2,48491	2,00148
24	20	1,00	3,17805	3,71357

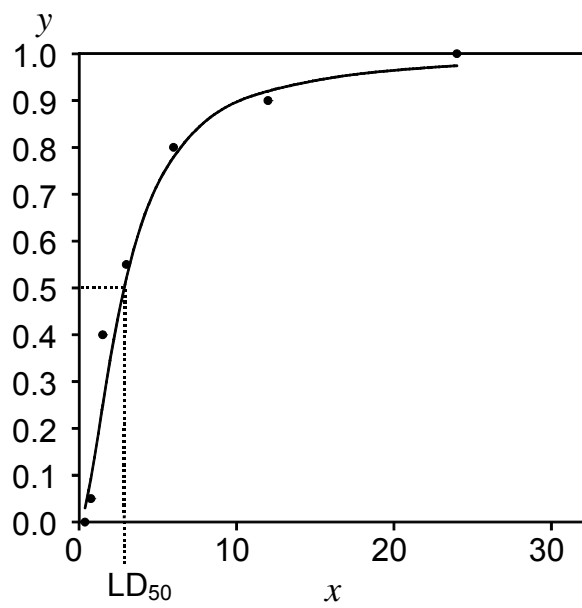


Abb. 6.13(a): Plot des Anteils gestorbener Insektenlarven (y) gegen die Dosis (x).

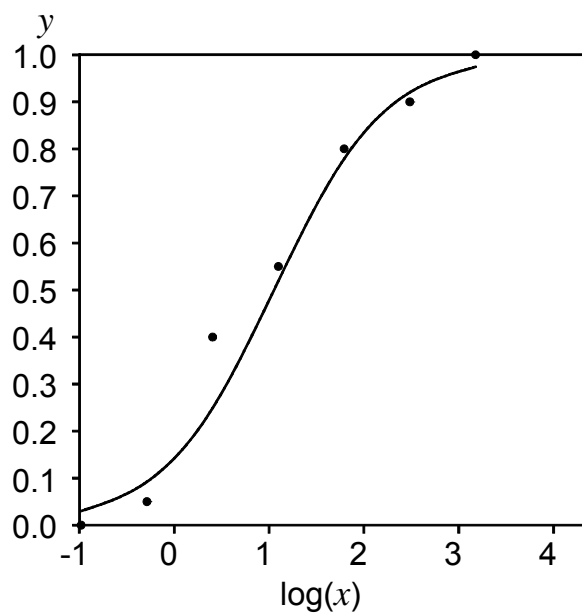


Abb. 6.13(b): Plot des Anteils gestorbener Insektenlarven (y) gegen die logarithmierte Dosis (x).

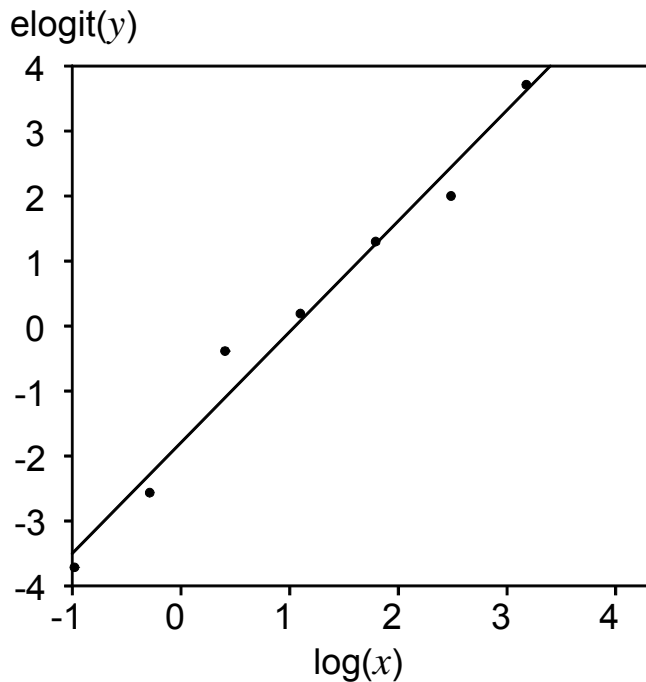


Abb. 6.13(c): Plot der empirischen Logits des Anteils gestorbener Insektenlarven (y) gegen die logarithmierte Dosis (x).

Der Plot des Anteils getöteter Larven (y) gegen die Dosis des Insektizids (x) in Abb. 6.13(a) zeigt eine deutlich nichtlineare Beziehung. Dies wird auch durch die angepasste Kurve (Erklärung unten) untermauert. Da der Anteil der getöteten Larven zwischen 0 und 1 liegen muss, nähert sich der Anteil getöteter Larven (y) mit steigender Dosis asymptotisch der 1.

Ziel der Auswertung ist es hier, eine möglichst einfache Auswertung zur Bestimmung der LD_{50} zu präsentieren, die mit Hilfe der einfachen linearen Regression zu bewerkstelligen ist. Dazu ist es notwendig, die Beziehung zwischen y und x durch eine geeignete Transformation zu linearisieren. Eine Logarithmierung der Dosis führt zu einer S-förmigen Kurve [Abb 6.13(b)]. Eine solche Kurve kann gut durch eine logistische Funktion modelliert werden:

$$\eta_i = \frac{\exp(\alpha + \beta x'_i)}{1 + \exp(\alpha + \beta x'_i)}$$

wobei

$\exp(.)$ = e-Funktion = Exponentialfunktion zur Basis e

η_i = Erwarteter Anteil getöteter Insekten

x'_i = $\log(x_i)$ (Logarithmus zur Basis e)

x_i = i -te Dosisstufe

Dieses Modell kann durch Transformation in eine lineare Beziehung überführt werden:

$$\frac{\eta_i}{1-\eta_i} = \frac{\frac{\exp(\alpha + \beta x'_i)}{1 + \exp(\alpha + \beta x'_i)}}{\frac{1}{1 + \exp(\alpha + \beta x'_i)}} = \frac{\exp(\alpha + \beta x'_i)}{1 + \exp(\alpha + \beta x'_i)} \times \frac{1 + \exp(\alpha + \beta x'_i)}{1} = \exp(\alpha + \beta x'_i)$$

Logarithmieren ergibt

$$\log\left(\frac{\eta_i}{1-\eta_i}\right) = \text{logit}(\eta_i) = \alpha + \beta x'_i$$

Die Funktion $\log[\eta_i/(1-\eta_i)]$ wird auch als Logit bezeichnet.

Zur Auswertung ersetzen wir den erwarteten Anteil getöteter Larven η_i durch den beobachteten Anteil y_i . Um diese immer berechnen zu können, darf der Fall $y_i = 0$ oder $y_i = 1$ nicht eintreten, denn der Logarithmus von Null ist nicht definiert. Daher berechnet man besser die leicht korrigierten, sog. empirischen Logits (McCullagh und Nelder, 1989, S. 106):

$$y'_i = \text{elogit}(y_i) = \log\left(\frac{y_i + \frac{1}{2m}}{1 - y_i + \frac{1}{2m}}\right)$$

wobei m die Zahl der je Variante behandelten Larven ist (hier: $m = 20$). Die empirischen Logits sind in Abb. 6.13(c) gegen $\log(x)$ abgetragen. Es ergibt sich erwartungsgemäß ein annähernd linearer Zusammenhang. Daher können wir die Parameter α und β des logistischen Modells durch lineare Regression nach dem transformierten, einfachen linearen Regressionsmodell

$$y'_i = \text{elogit}(y_i) = \alpha + \beta x'_i + e_i$$

schätzen. Wir finden folgende geschätzte Regressionsgerade:

$$\hat{y}' = a + bx'$$

mit

$$a = -1,796$$

$$b = 1,705$$

Mit diesen Schätzwerten haben wir die Kurven in 6.13(a) und 6.13(b) sowie die Gerade in Abb. 6.13(c) gezeichnet.

Die LD_{50} ist nun diejenige Konzentration x , bei welcher der Erwartungswert für y gleich 0,5 wird. Wir haben es hier also mit einem inversen Regressionsproblem zu tun. Einsetzen in die Formel für die empirischen Logits liefert:

$$y'_0 = \text{elogit}(0,5) = \log \left(\frac{0,5 + \frac{1}{2m}}{1 - 0,5 + \frac{1}{2m}} \right) = \log(1) = 0$$

Um die LD_{50} zu ermitteln, müssen wir für die lineare Regression in Abb. 6.13(c) den x' -Wert zu $y' = 0$ finden. Dieses inverse Regressionsproblem hatten wir schon in 6.2.3 besprochen. Wir finden

$$\hat{x}' = \frac{y'_0 - a}{b} = \frac{1.796}{1.705} = 1,053$$

Nun müssen wir noch die logarithmische Transformation rückgängig machen:

$$\hat{x} = \exp(\hat{x}') = \exp(1,053) = 2,866$$

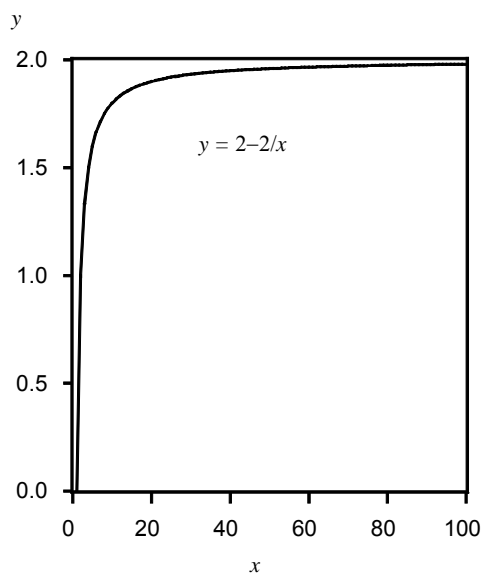
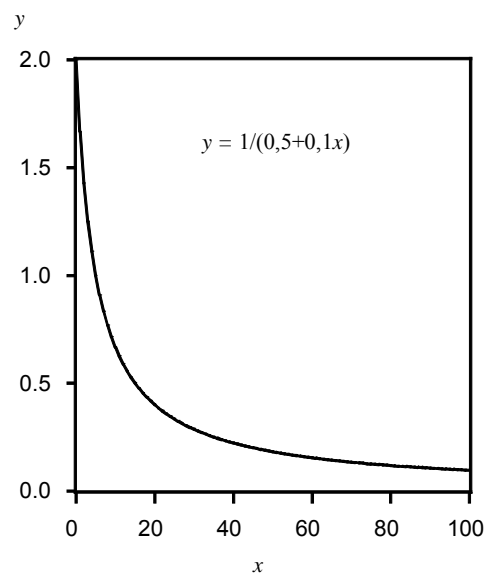
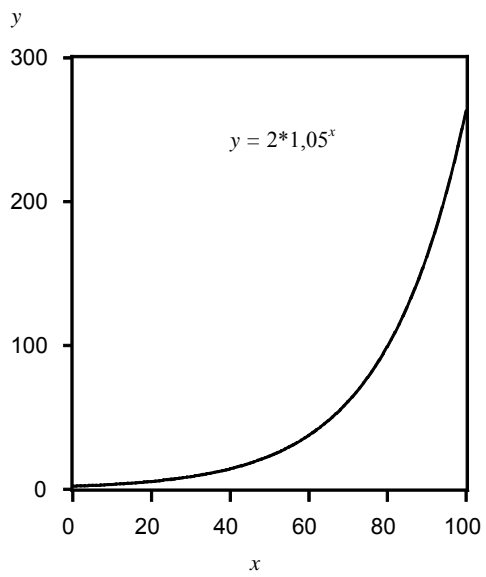
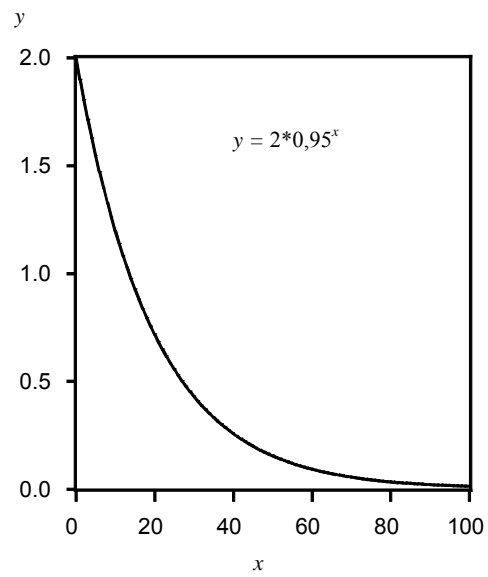
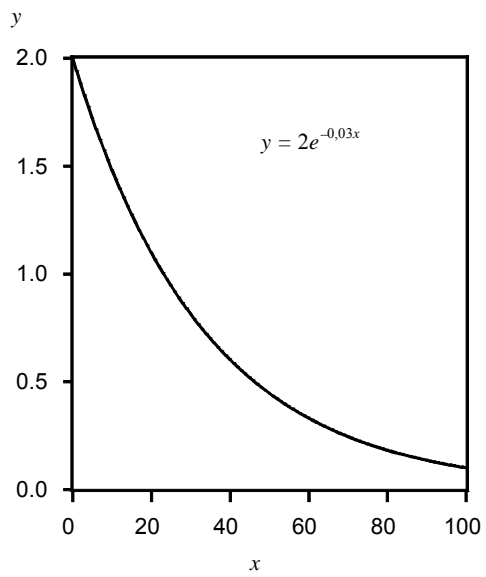
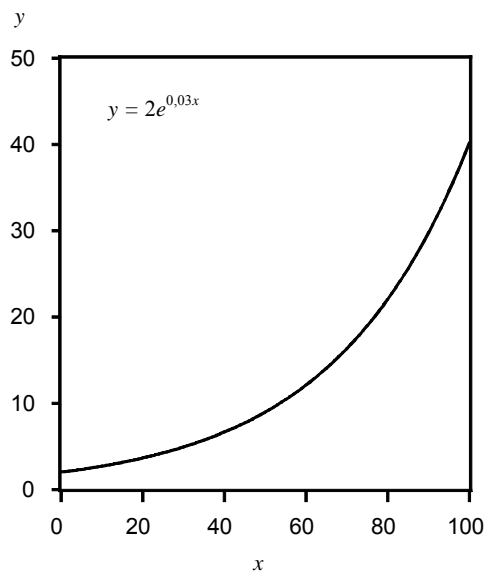
Bei einer Konzentration von 2,866 erwarten wir also 50% abgetötete Larven ($LD_{50} = 2,866$). Das nach der Methode in Abschnitt 6.2.3 berechnete 95%Vertrauensintervall für die logarithmierte Dosis \hat{x}' hat die Grenzen 0,278 und 1,823. Hieraus ergeben sich für die LD_{50} die Grenzen $1,32 = \exp(0,278)$ und $6,19 = \exp(1,823)$.

Abschließende Bemerkung: Wir haben hier exemplarisch vier Funktionen kennengelernt, die sich durch Transformation linearisieren ließen. Eine wichtige Voraussetzung ist dabei, dass die Transformationen **monoton** sind (die Daten müssen nach der Transformation dieselbe Rangfolge haben wie vorher). Es gibt eine Reihe anderer nichtlinearer Funktionen mit dieser Eigenschaft. In Tab 6.4.1 ist eine kleine Übersicht gegeben. Falls eine geeignete Funktion sich nicht linearisieren lässt, muss das Verfahren der eigentlichen nichtlinearen Regression verwendet werden (Abschnitt 6.12). Ein Beispiel für eine nicht linearisierbare Funktion ist die logistische Wachstumsfunktion

$$\eta = \alpha + \beta \exp(\gamma x)$$

Tab. 6.4.1: Einige Funktionen, die sich durch Transformation linearisieren lassen.

Funktion	Datentrans- formation	Modell für trans- formierte Daten	Transformation der Parameter
$y = \alpha e^{\beta x}$	$y' = \log(y)$	$y' = \alpha' + \beta x$	$\alpha' = \log(\alpha)$
$y = \alpha \beta^x$	$y' = \log(y)$	$y' = \alpha' + \beta' x$	$\alpha' = \log(\alpha)$ $\beta' = \log(\beta)$
$y = 1/(\alpha + \beta x)$	$y' = 1/y$	$y' = \alpha + \beta x$	-
$y = \alpha + \beta/x$	$x' = 1/x$	$y = \alpha + \beta x'$	-
$y = (\alpha + \beta/x)^{-1}$	$y' = 1/y$ $x' = 1/x$	$y' = \alpha + \beta x'$	-



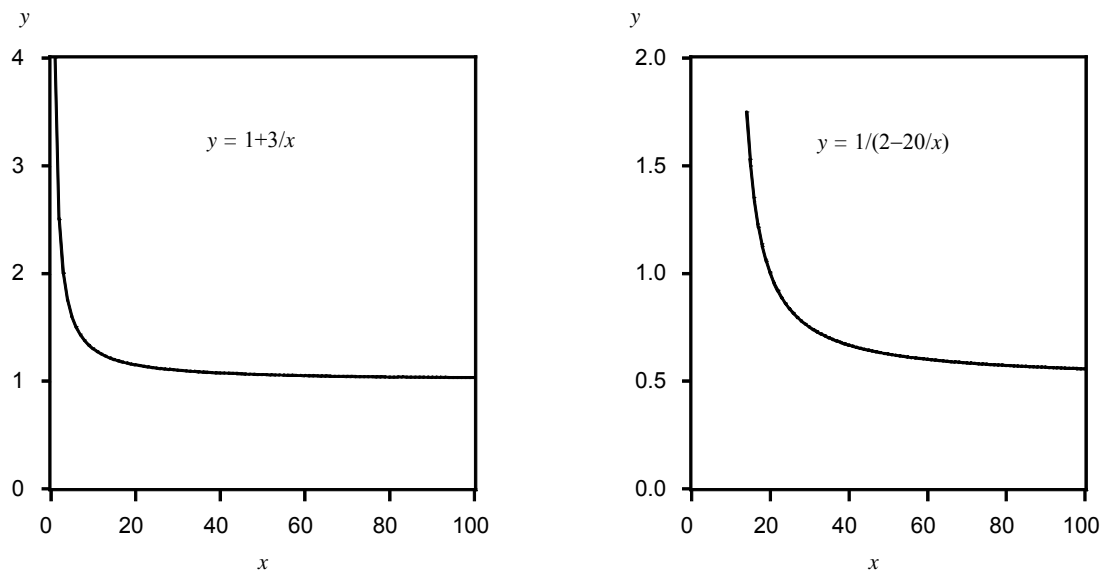


Abb. 6.4.1: Einige Bilder der Funktionen in Tab. 6.4.1.

6.5 Korrelation bei nichtlinearen Zusammenhängen

Beispiel: In einer Erhebung in Australien wurden das Alter (x) und das Gewicht der Augenlinse (y) bei 71 Hasen ermittelt. Ziel der Untersuchung war es, herauszufinden, ob ein Zusammenhang zwischen beiden Variablen besteht (Ratkowski, 1983, S. 108). Die Daten sind in Abb. 6.14 dargestellt. Der Zusammenhang ist eindeutig nicht-linear, so dass die Pearsonsche Produkt-Moment-Korrelation kein geeignetes Maß ist, um den Zusammenhang zu beschreiben.

Anstelle der Pearsonschen Produkt-Moment-Korrelation kann die sog. Spearmansche Rangkorrelation verwendet werden. Diese ist anwendbar nicht nur für lineare Zusammenhänge sondern für jede Art von monotonem, nichtlinearem Zusammenhang, wie im Hasenbeispiel. Hinzu kommt, dass für den Test der Rangkorrelation (siehe unten) keine bivariate Normalverteilung angenommen werden muss wie bei der Produkt-Moment-Korrelation. Man bezeichnet den Test der Rangkorrelation daher auch als **verteilungsfrei** oder **nichtparametrisch** (siehe auch Kap. 11). Die Rangkorrelation ist nicht nur auf metrische, sondern auch auf ordinale Daten anwendbar, wie im zweiten Beispiel erläutert wird.

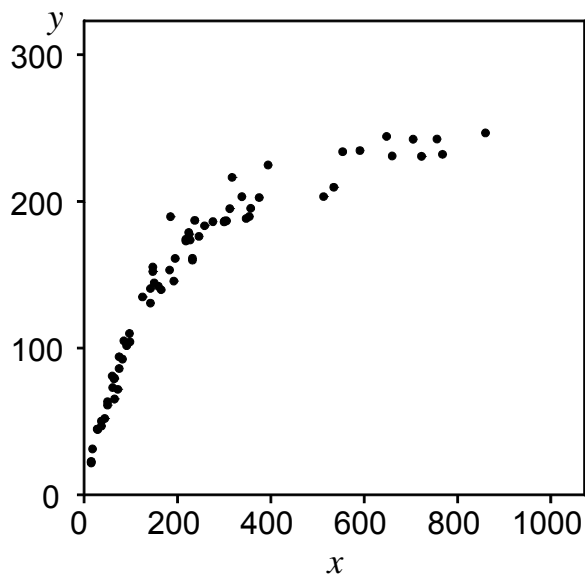


Abb. 6.14: Plot des Gewichts der Augenlinse (y ; mg) gegen das Alter von Hasen (x ; Tage).

Tab. 6.5.1: Hasendaten mit Rängen

X	Y	Ränge		X	Y	Ränge	
		R(X)	R(Y)			R(X)	R(Y)
15	21,66	2	1	195	161,10	37	37
15	22,75	2	3	218	174,18	38,5	42
15	22,30	2	2	218	173,03	38,5	39
18	31,25	4	4	219	173,54	40	40
28	44,79	5	6	224	178,86	41	45
29	44,55	6	5	225	177,68	42	44
37	50,25	7,5	8	227	173,73	43	41
37	46,88	7,5	7	232	159,98	44,5	36
44	52,03	9	9	232	161,29	44,5	38
50	63,47	10,5	11	237	187,07	46	51
50	61,13	10,5	10	246	176,13	47	43
60	81,00	12	17	258	183,40	48	46
61	73,09	13	14	276	186,26	49	48
64	79,09	14	15	300	186,09	50	47
65	79,51	15,5	16	301	186,70	51	49
65	65,31	15,5	12	305	186,80	52	50
72	71,90	17	13	312	195,10	53	55
75	86,10	18,5	18	317	216,41	54	61
75	94,10	18,5	20	338	203,23	55	58
82	92,50	20	19	347	188,38	56	52
85	105,00	21	24	354	189,70	57	54
91	101,70	22,5	21	357	195,31	58	56
91	102,90	22,5	22	375	202,63	59	57
97	110,00	24	25	394	224,82	60	62
98	104,30	25	23	513	203,30	61	59
125	134,90	26	27	535	209,70	62	60
142	130,68	27,5	26	554	233,90	63	66
142	140,58	27,5	29	591	234,70	64	67
147	155,30	29,5	35	648	244,30	65	70
147	152,20	29,5	33	660	231,00	66	64
150	144,50	31	31	705	242,40	67	68
159	142,15	32	30	723	230,77	68	63
165	139,81	33	28	756	242,57	69	69
183	153,22	34	34	768	232,12	70	65
185	189,66	35	53	860	246,70	71	71
192	145,72	36	32				

Berechnung und Test der **Spearman'schen Rangkorrelation**:

H_0 : Es besteht kein monotoner Zusammenhang zwischen x und y

H_1 : Es besteht ein monotoner Zusammenhang zwischen x und y

(1) Berechne die Ränge der beiden Variablen

$R(x_i)$ = Rang von x_i

$R(y_i)$ = Rang von y_i

Bei Rangbindungen ordne den Mittel-Rang zu.

(2) Berechne die Pearson'sche Produkt-Moment-Korrelation der Ränge nach der Formel in Abschnitt 6.1. Bezeichne diesen mit r_s .

Wenn es für keine der beiden Variablen Rangbindungen gibt, kann die folgende einfache Rechenformel benutzt werden:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

wobei $d_i = R(x_i) - R(y_i)$

(3) Teste H_0 : Kein monotoner Zusammenhang.

(i) $n > 30$: Test wie bei normalverteilten Daten: Verwerfe H_0 , falls

$$t_{Vers} = \frac{|r_s|}{\sqrt{1 - r_s^2}} \sqrt{n - 2} > t_{Tab},$$

wobei t_{Tab} der kritische t-Wert bei $FG = n - 2$ Freiheitsgraden ist. Dies Verfahren ist anwendbar unabhängig davon ob Rangbindungen vorliegen.

(ii) $n \leq 30$ und keine Rangbindungen: Führe einen exakten Test durch:

Verwerfe H_0 falls $|r_s| > r_{Tab}$

wobei r_{Tab} aus Tab. IX (Bortz et al. 1990, S.749) abzulesen ist.

Tab. IX: Kritische Werte für die Spearman'sche Rangkorrelation bei $\alpha = 0,05$.

n	r_{Tab}	n	r_{Tab}	n	r_{Tab}	n	r_{Tab}	n	r_{Tab}
6	0,886	11	0,623	16	0,507	21	0,438	26	0,392
7	0,786	12	0,591	17	0,490	22	0,428	27	0,385
8	0,738	13	0,566	18	0,476	23	0,418	28	0,377
9	0,683	14	0,545	19	0,462	24	0,409	29	0,370
10	0,648	15	0,525	20	0,450	25	0,400	30	0,364

(iii) $n \leq 30$ und Rangbindungen: Hier wäre ein exakter Test zu bevorzugen, aber die exakten Werte in Tab. IX sind hier nicht anwendbar. Je nach Bindungen müsste neu die exakte Verteilung ausgerechnet werden. Hilfsweise kann man vereinfachend wie im Fall $n > 30$ vorgehen.

Beispiel: Für die Hasendaten berechnen wir zunächst die Ränge. Exemplarisch betrachten wir die Variable x , die in der obenstehenden Tabelle bereits der Größe nach geordnet ist. Der kleinste Wert ist 15. Dieser kommt dreimal vor. Daher müssen diese drei Werte die Ränge 1, 2 und 3 zugeordnet bekommen. Da jeweils derselbe Wert vorliegt, vergeben wir den Mittelrang 2 (Rangbindung). Der zweitgrößte Wert ist die 18. Dieser erhält den Rang 4. Der drittgrößte Wert ist die 26, dieser bekommt den Rang 5, usw. Mit den Rängen berechnen wir folgende Produkt-Moment-Korrelation (siehe Abschnitt 6.1), die der Rangkorrelation entspricht:

$$r_s = 0,98463$$

Hierfür finden wir

$$t_{Vers} = \frac{|r_s|}{\sqrt{1-r_s^2}} \sqrt{n-2} = \frac{0,98463}{\sqrt{1-0,98463^2}} \sqrt{69} = 46,83 > t_{Tab} = 1,995$$

Es besteht ein hoch signifikanter Zusammenhang. Zum Vergleich berechnen wir die Produkt-Moment-Korrelation der Originaldaten. Wir finden

$$r = 0,86763$$

Dieser Wert ist deutlich kleiner als die Rangkorrelation r_s , was zeigt, dass die Produkt-Moment-Korrelation der Originaldaten nicht geeignet ist, den sehr engen, aber nicht-linearen Zusammenhang adäquat zu beschreiben.

Beispiel: Der Gesundheitszustand von 7 Laborratten nach einer Mangelernährung wird von 2 Personen beurteilt, indem die Ratten in eine Rangfolge gebracht werden (Snedecor und Cochran, 1967). Es soll die Korrelation der beiden Beurteilungen geprüft werden.

Ratte Nr.	Person 1 $R(x_i) = x_i$	Person 2 $R(y_i) = y_i$	Rang- differenz d_i	d_i^2
1	4	4	0	0
2	1	2	-1	1
3	6	5	1	1
4	5	6	-1	1
5	3	1	2	4
6	2	3	-1	1
7	7	7	0	0

$$\sum_{i=1}^n d_i^2 = 8$$

Da hier direkt die Ränge beobachtet werden, ist $R(x_i) = x_i$ und $R(y_i) = y_i$. Außerdem liegen keine Rangbindungen vor, so dass wir die vereinfachte Rechenformel benutzen können:

$$r_s = 1 - \frac{6 \cdot 8}{7(7^2 - 1)} = 0,857$$

Wegen des kleinen Stichprobenumfanges muss exakt getestet werden. Für $n = 7$ finden wir $r_{Tab} = 0,786 < r_s$. Die Rangkorrelation ist signifikant.

6.6 Test auf Linearität

Die lineare Regression ist sehr verbreitet, vor allem, weil sie einfach durchzuführen ist. Daher ist bei einer Regression sinnvoll zu prüfen, ob eine lineare Regression zur Beschreibung des Zusammenhangs ausreicht. Die Linearität kann mittels eines Tests formal geprüft werden, falls für mindestens einen x -Wert mehrere y -Werte vorliegen, weil dann eine von der Regression unabhängige Schätzung der Varianz möglich ist. Man kann das Verfahren auch auf nichtlineare Modelle erweitern, so z.B. in der Polynomregression (siehe Abschnitt 8.9).

Beispiel: Es wurde ein Versuch durchgeführt, um den Einfluss verschiedener Stickstoff (N) Mengen auf den Wurzelmasseertrag von Zuckerrüben zu erfassen. Der Versuch wurde in einer Blockanlage durchgeführt, aber wir wollen hier zur Illustration annehmen, dass der Versuch vollständig randomisiert wurde (Petersen, 1994).

Tab. 6.6.1: Wurzelmasseertrag (t/ha) bei Zuckerrüben in Abhängigkeit von der gedüngten Stickstoffmenge.

Wiederholung	Düngerstufe (kg N /ha)				
	0	35	70	105	140
1	9,9	20,3	27,5	31,4	28,1
2	7,8	22,6	30,3	27,2	25,7
3	10,7	23,9	29,2	33,4	31,9

Die Rohdaten sind in Abb. 6.15 graphisch dargestellt.

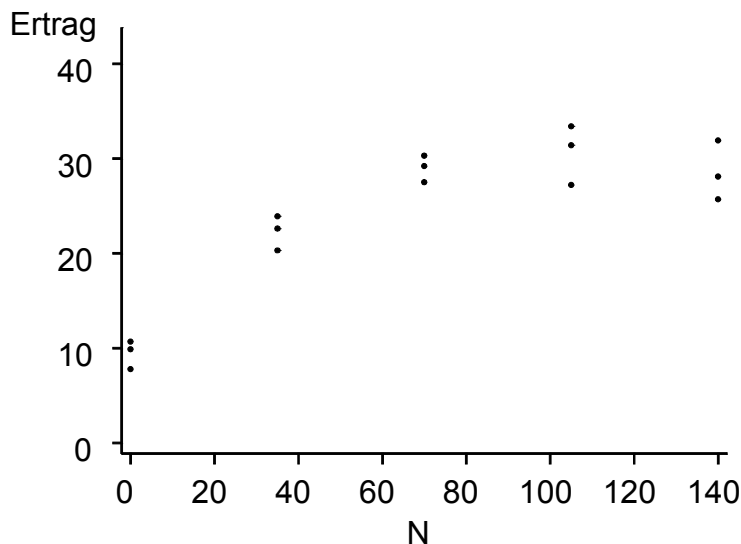


Abb. 6.15: Plot der Rübenenerträge gegen die N-Menge.

Falls je x -Stufe mehrere y -Werte vorliegen, kann man die Fehlervarianz auf zwei Wegen schätzen. Zum einen kann die Streuung der Einzelwerte um die Stufenmittelwerte berechnet werden. Diese ist von der geschätzten Regressionsgerade unabhängig. Zum anderen kann die Streuung der Stufenmittelwerte um die Regression betrachtet werden. Diese beiden Komponenten ergeben sich durch eine Zerlegung der "Streuung um die Regression":

$$\sum_{i=1}^t \sum_{j=1}^r (y_{ij} - a - bx_i)^2 = \sum_{i=1}^t \sum_{j=1}^r [(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - a - bx_i)]^2 = \sum_{i=1}^t \sum_{j=1}^r (y_{ij} - \bar{y}_{i.})^2 + r \sum_{i=1}^t (\bar{y}_{i.} - a - bx_i)^2$$

Einzelwerte um
Mittelwerte

Mittelwerte um die
Regression

Hierbei ist:

y_{ij} = j -ter Messwert bei der Stufe x_i . ($i = 1, \dots, t$; $j = 1, \dots, r$)

t = Zahl der Stufen, r = Zahl der Wiederholungen je Stufe

Die Streuung der Mittelwerte um die Regression hängt ausschließlich vom Versuchsfehler ab, sofern die Annahme einer linearen Regression zutrifft. Ist der Zusammenhang zwischen x und y dagegen nicht-linear, so wird die Streuung der Mittelwerte um die Regression größer sein, als allein aufgrund des Versuchsfehlers zu erwarten wäre. Im Gegensatz hierzu ist die Streuung der Einzelwerte um die Stufenmittelwerte unabhängig davon, ob das lineare Modell zutrifft oder nicht. Die gesamte Streuungszerlegung ist für die Rübensdaten in Abb. 6.16 veranschaulicht. Die Varianzanalyse-Tabelle ist wie folgt:

Ursache	Freiheitsgrade	SQ
Auf der Regression	1	$SQ_{Auf} = r \sum_{i=1}^t (a + bx_i - \bar{y}_{..})^2 = b^2 SQ_x$
Um die Regression (Nichtlinearität)	$t - 2$	$SQ_{Um} = r \sum_{i=1}^t (\bar{y}_{i.} - a - bx_i)^2 = \sum_{i=1}^t y_{i.}^2 / r - y_{..}^2 / (rt) - SQ_{Auf}$
Fehler	$t(r - 1)$	$SQ_{Fehler} = \sum_{i=1}^t \sum_{j=1}^r (y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^t \sum_{j=1}^r y_{ij}^2 - \sum_{i=1}^t y_{i.}^2 / r$

t = Zahl der Stufen von x ; r = Zahl der Wiederholungen je Stufe von x .

$$b = \frac{SP_{xy}}{SQ_x}$$

$$SP_{xy} = \sum_{i=1}^t \sum_{j=1}^r (x_i - \bar{x}_{.}) (y_{ij} - \bar{y}_{i.}) = \sum_{i=1}^t \sum_{j=1}^r x_i y_{ij} - \frac{\left(\sum_{i=1}^t \sum_{j=1}^r x_i \right) \left(\sum_{i=1}^t \sum_{j=1}^r y_{ij} \right)}{rt}$$

$$SQ_x = \sum_{i=1}^t \sum_{j=1}^r (x_i - \bar{x}_{.})^2 = \left[\sum_{i=1}^t \sum_{j=1}^r x_i^2 - \frac{\left(\sum_{i=1}^t \sum_{j=1}^r x_i \right)^2}{rt} \right]$$

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

H_0 : Regression ist linear

Berechne: $F_{Vers} = \frac{SQ_{Um} / (t-2)}{SQ_{Fehler} / [t(r-1)]}$

Bestimme: $F_{Tab} = F_{(1-\alpha; t-2, t(r-1))}$ (Tab. VI)

$F_{Vers} > F_{Tab} \Rightarrow$ Kurvenverlauf nichtlinear

$F_{Vers} \leq F_{Tab} \Rightarrow$ Keine signifikante Abweichung von der Linearität

Eine Bemerkung zu den Freiheitsgraden (FG). Der Fehler hat dieselben FG wie bei einer einfachen Varianzanalyse, und die Berechnung des SQ_{Fehler} ist identisch. Die $(t-1)$ Behandlungs-Freiheitsgrade der einfachen Varianzanalyse werden hier in zwei Komponenten zerlegt, genauso wie das Behandlungs-SQ: ein FG wird für den Regressionskoeffizienten verwendet, die verbleibenden $(t-2)$ FG bleiben für die Abweichung von der Linearität (Lack-of-fit).

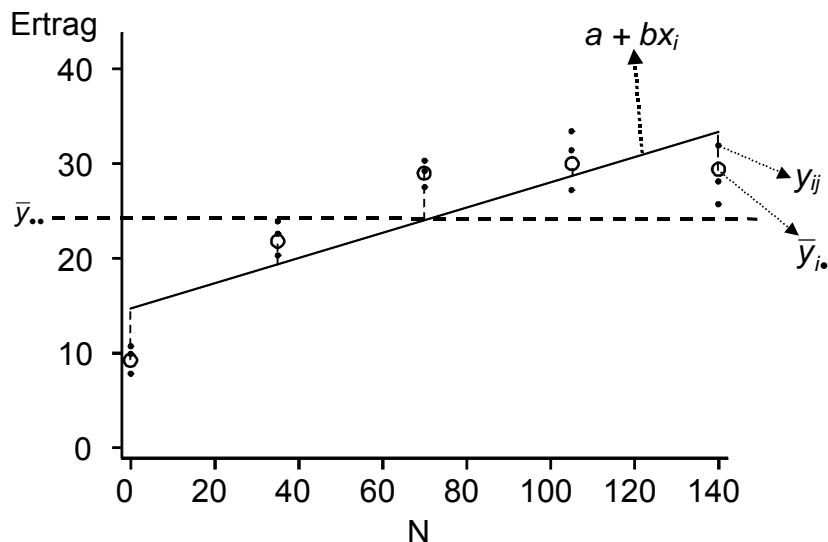


Abb. 6.16: Veranschaulichung der Streuungszerlegung für den Test auf Linearität:

- "Auf der Regression": Abweichung Regressionsgerade von Horizontale (gestrichelt); $(a + bx_i) - \bar{y}_{..}$.
- "Um die Regression": Mittelwerte (Kreise) um die Regression; $\bar{y}_{i.} - (a + bx_i)$
- "Fehler": Einzelwerte (schwarze Punkte) um Mittelwerte (Kreise); $y_{ij} - \bar{y}_{i.}$

Der Varianzanalyse liegt folgendes Modell zugrunde:

$$y_{ij} = \alpha + \beta x_i + \delta_i + e_{ij}$$

wobei

α = Achsenabschnitt

β = Steigung

x_i = Wert der i -ten Stufe der Einflußvariable (hier: N-Menge)

δ_i = Systematische Abweichung (der Mittelwerte) von der Regression für i -te Stufe der Einflußvariable

e_{ij} = Fehler von y_{ij}

Die Nullhypothese H_0 : "linearer Kurvenverlauf" entspricht

$H_0: \delta_i = 0$ für alle i

Beispiel:

x_i	y_{ij}	x_i^2	y_{ij}^2	$x_i y_{ij}$
0	9,9	0	98,01	0,0
0	7,8	0	60,84	0,0
0	10,7	0	114,49	0,0
35	20,3	1225	412,09	710,5
35	22,6	1225	510,76	791,0
35	23,9	1225	571,21	836,5

70	27,5	4900	756,25	1925,0
70	30,3	4900	918,09	2121,0
70	29,2	4900	852,64	2044,0
105	31,4	11025	985,96	3297,0
105	27,2	11025	739,84	2856,0
105	33,4	11025	1115,56	3507,0
140	28,1	19600	789,61	3934,0
140	25,7	19600	660,49	3598,0
140	31,9	19600	1017,61	4466,0
1050	359,9	110250	9603,45	30086,0

$$r = 3; t = 5$$

$$SP_{xy} = 30086,0 - 1050 \cdot 359,9 / 15 = 4893$$

$$SQ_x = 110250 - 1050^2 / 15 = 36750$$

$$b = 4893 / 36750 = 0,133143$$

$$\bar{x}_. = 1050 / 15 = 70$$

$$\bar{y}_{..} = 359,9 / 15 = 23,99$$

$$a = 23,99 - 0,133143 \cdot 70 = 14,673$$

$$SQ_{Auf} = b^2 SQ_x = 0,133143^2 \cdot 36750 = 651,468$$

Wiederholung	Düngerstufe (kg N /ha)					
	0	35	70	105	140	
1	9,9	20,3	27,5	31,4	28,1	
2	7,8	22,6	30,3	27,2	25,7	
3	10,7	23,9	29,2	33,4	31,9	
$y_{i.}$	28,4	66,8	87,0	92,0	85,7	$y_{..} = 359,9$

$$SQ_{Um} = \sum_{i=1}^t y_{i.}^2 / r - y_{..}^2 / (rt) - SQ_{Auf}$$

$$= (28,4^2 + 66,8^2 + 87,0^2 + 92,0^2 + 85,7^2) / 3 - 359,9^2 / 15 - 651,468 = 262,095$$

$$SQ_{Fehler} = \sum_{i=1}^t \sum_{j=1}^r y_{ij}^2 - \sum_{i=1}^t y_{i.}^2 / r = 9603,45 - (28,4^2 + 66,8^2 + 87,0^2 + 92,0^2 + 85,7^2) / 3 = 54,687$$

Varianzanalyse:

Ursache	FG	SQ	MQ	F_{Vers}
Auf der Regression	1	651,468	651,468	$F_{Vers} = 15,98$
Um die Regression (Nichtlinearität)	3	262,095	87,364	
Fehler	10	54,687	5,469	

$$\alpha = 5\%, 1-\alpha = 0,95; F_{Tab} = F(0,95; 3; 10;) = 3,71 < F_{Vers}$$

Es besteht eine signifikante Abweichung von der Linearität. Daher sollte für diese Daten eine nichtlineare Regression in Betracht gezogen werden (Abschnitte 6.4, 6.11 und 6.12).

6.7 Residuen, Modellvoraussetzungen und Ausreisser

6.7.1 Was sind Residuen?

Bei der Regression ebenso wie bei der Varianzanalyse wird ein lineares Modell angenommen, wobei der Fehlerterm normalverteilt sein soll. Für die lineare Regression ist das Modell

$$y_i = \alpha + \beta x_i + e_i$$

wobei e_i einer Normalverteilung mit Mittelwert 0 und Varianz σ^2 folgt. Um die Modellannahmen prüfen zu können, ist es notwendig, die Fehler e_i zu schätzen. Wären die Parameter α und β bekannt, könnten die Fehler direkt durch Differenzbildung berechnet werden:

$$e_i = y_i - (\alpha + \beta x_i)$$

Da in der Praxis die Parameter geschätzt werden müssen, setzen wir anstelle der Parameter die Kleinstquadratschätzwerte a und b ein. Die sich daraus ergebenden **geschätzten Fehler** werden als **Residuen** (auch **Roh-Residuen**) bezeichnet:

$$r_i = \hat{e}_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

Man beachte, dass das SQ_{Fehler} der Regression (Abschnitt 6.2) sich berechnen lässt als

$$SQ_{Fehler} = \sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Wenn die Beziehung von y_i und x_i tatsächlich linear ist, so weichen die Residuen nur zufällig von Null ab. Es gilt also

$$E(r_i) = 0$$

Falls die Fehler e_i mit konstanter (homogener) Varianz verteilt sind, so weisen die Residuen, also die Schätzungen von e_i , keine homogene Varianz auf. Es gilt:

$$\text{var}(r_i) = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{SQ_x} \right]$$

Aus diesem Grunde ist es sinnvoll, die Residuen zu standardisieren ("**studentisieren**"). Die sog. studentisierten Residuen sind gegeben durch

$$c_i = \frac{r_i}{\sqrt{\sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{SQ_x} \right]}}$$

In praktischen Anwendungen wird die unbekannte Varianz durch ihren Schätzwert aus der Varianzanalyse (s^2) ersetzt. Die meisten Statistikpakete verfügen über Optionen zur Berechnung und Ausgabe von Residuen. Residuen eignen sich als diagnostisches Hilfsmittel zum Aufspüren von Abweichungen von den Modellvoraussetzungen, wie im folgenden exemplarisch gezeigt wird. Eine Residuenanalyse ist ein Standardwerkzeug bei der Analyse linearer Modelle. Hier wird jeweils anhand von Residuenplots eine Diagnose gegeben, während das sich daraus ergebende Vorgehen nur angedeutet wird. Eine wichtige Methode bei Verletzung von Voraussetzungen ist eine Transformation der Daten. Dies wird in Kap. 7 näher besprochen.

6.7.2 Residuen-Plots

Plot gegen den geschätzten Wert: Trägt man die studentisierten Residuen gegen die geschätzten Werte $a + bx_i$ ab, so erhält man Aufschluss über mögliche Abweichungen von den Modellannahmen wie Varianzhomogenität und Linearität.

Q-Q-Plots: Außerdem lässt sich die Normalverteilung prüfen, indem die n studentisierten Residuen zunächst der Größe nach geordnet werden. Die der Größe nach geordneten Residuen bezeichnen wir mit

$$c_{(1)}, c_{(2)}, \dots, c_{(i)}, \dots, c_{(n)}$$

Die geordneten Residuen werden gegen den Erwartungswert des kleinsten, zweitkleinsten, ... größten Wertes einer Stichprobe vom Umfang n aus einer Standardnormalverteilung abgetragen. Diese Erwartungswerte (u_i) werden auch als "Normal Scores" bezeichnet. Falls die Annahme der Normalverteilung zutrifft, sollten die Punkte etwa auf der Diagonalen $c_{(i)} = u_i$ liegen. Die "Normal Scores" berechnen sich, indem für jedes i der Anteil

$$p_i = (i - 1/2)/n$$

und für das resultierende p_i die Ordinate u_i nach

$$p_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u_i} \exp(-x^2/2) dx$$

berechnet wird. Hierfür wird man in der Regel ein Computerprogramm verwenden. Der Plot von $c_{(i)}$ gegen u_i wird als Q-Q-Plot bezeichnet (Quantil-Quantil-Plot), weil hier die empirischen Quantile der Residuen gegen die theoretischen Quantile der Normalverteilung abgetragen werden.

Weitere Möglichkeiten zur Verwendung von Residuen zu diagnostischen Zwecken können z.B. bei Atkinson (1985: Plots, transformations and regression. Clarendon Press, London) nachgeschlagen werden. Hier sollen die beiden erwähnten Residuenplots anhand von Beispielen erläutert werden.

Beispiel (keine auffälligen Abweichungen): Regendaten vom Anfang des Kapitel 6. Die Rohdaten finden sich in Abb. 6.17(a). Der Plot der Residuen gegen den vorhergesagten Wert in Abb. 6.17(b) zeigt keine Verletzungen der Voraussetzungen an, da die Punkte etwa in einer Ellipse mit horizontaler Symmetrieachse liegen. Der Q-Q-Plot in Abb. 6.17(c) zeigt die Punkte etwa auf einer Geraden, so dass kein Hinweis auf Abweichung von der Normalverteilung vorliegt. Allerdings fällt ein etwas schlangenförmiger Verlauf auf, der nicht näher erklärt werden kann.

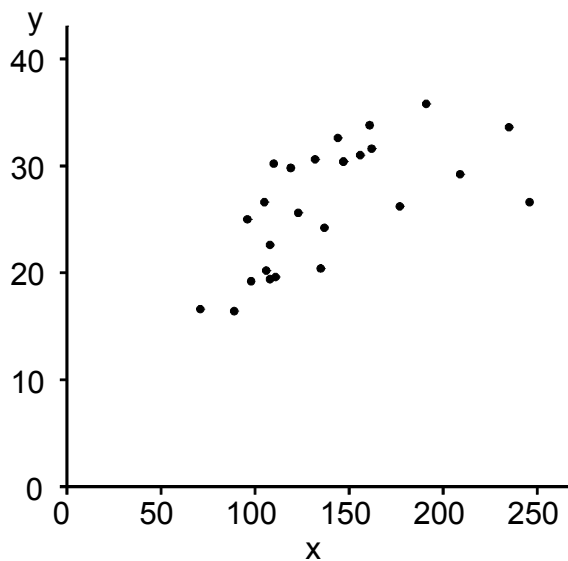


Abb. 6.17(a): Plot der Erträge (y ; dt/ha) gegen die in den Monaten April bis Juni gefallene Regenmenge (x ; mm).

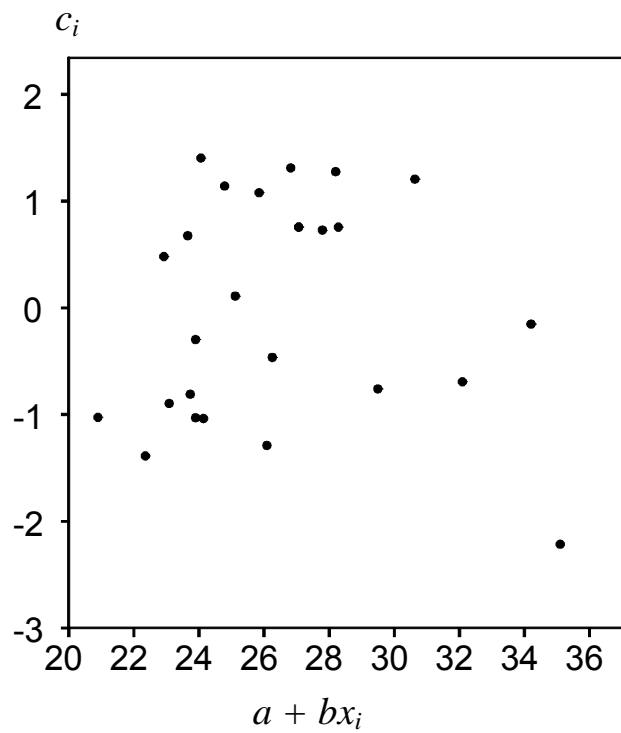


Abb. 6.17(b): Plot der Residuen c_i gegen den geschätzten Wert $a + bx_i$ für Regendaten.

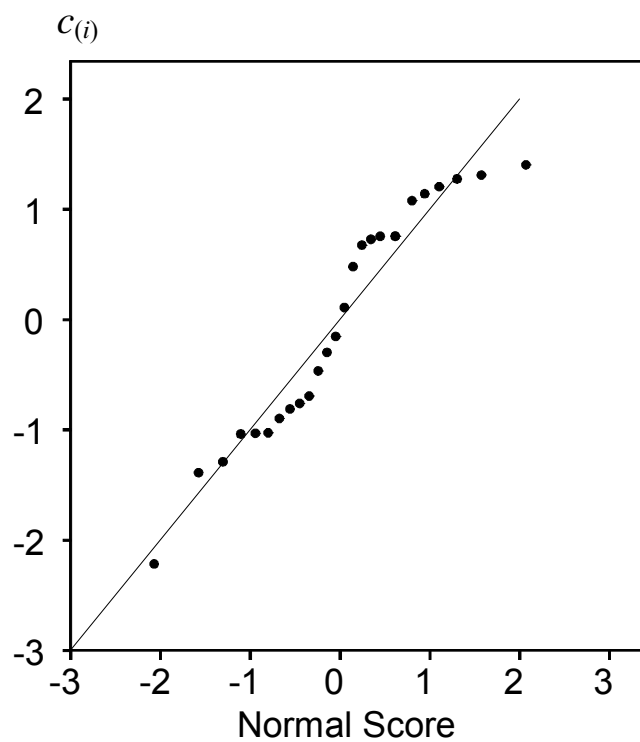


Abb. 6.17(c): Plot der geordneten Residuen $c_{(i)}$ gegen den Normal Score (u_i) für Regendaten (Q-Q-Plot).

Tab. 6.7.1: Berechnung der studentisierten Residuen (c_i) und Quantile der Normalverteilung (u_i).

x_i	y_i	c_i	(i)	p_i	u_i
177	26,2	-0,76000	9,0	0,32692	-0,44843
96	25,0	0,47934	15,0	0,55769	0,14512
144	32,6	1,30913	25,0	0,94231	1,57444
105	26,6	0,67519	16,0	0,59615	0,24340
111	19,6	-1,03896	4,0	0,13462	-1,10484
135	20,4	-1,29035	3,0	0,09615	-1,30378
209	29,2	-0,69434	10,0	0,36538	-0,34410
161	33,8	1,27451	24,0	0,90385	1,30378
246	26,6	-2,21708	1,0	0,01923	-2,06990
108	22,6	-0,29815	12,0	0,44231	-0,14512
137	24,2	-0,46584	11,0	0,40385	-0,24340
71	16,6	-1,02700	6,0	0,21154	-0,80109
119	29,8	1,13926	22,0	0,82692	0,94208
108	19,4	-1,03104	5,0	0,17308	-0,94208
132	30,6	1,07697	21,0	0,78846	0,80109
89	16,4	-1,38881	2,0	0,05769	-1,57444
147	30,4	0,75562	19,5	0,73077	0,61514
98	19,2	-0,89815	7,0	0,25000	-0,67449
106	20,2	-0,81179	8,0	0,28846	-0,55788
123	25,6	0,10923	14,0	0,51923	0,04822
156	31,0	0,72774	17,0	0,63462	0,34410
191	35,8	1,20457	23,0	0,86538	1,10484
162	31,6	0,75535	18,0	0,67308	0,44843
235	33,6	-0,15431	13,0	0,48077	-0,04822
147	30,4	0,75562	19,5	0,73077	0,61514
110	30,2	1,40341	26,0	0,98077	2,06990

Beispiel (Nichtlinearität): An Pflanzen der Reissorte IR8 wurden die Licht-Transmissions-Rate (y) sowie der Blattflächenindex (x) gemessen (Gomez und Gomez, 1984, S. 390). Dieses Beispiel wurde auch im Abschnitt 6.4 betrachtet, wo eine nichtlineare Regression durchgeführt wurde. Sowohl der Plot gegen den vorhergesagten Wert [Abb. 6.18(b)] als auch der Q-Q-Plot [Abb. 6.18(c)] zeigen eine Abweichung von den Modellvoraussetzungen, die hier durch die Nichtlinearität bedingt ist. **Weiterer Analyseschritt:** Nichtlineare Regression.

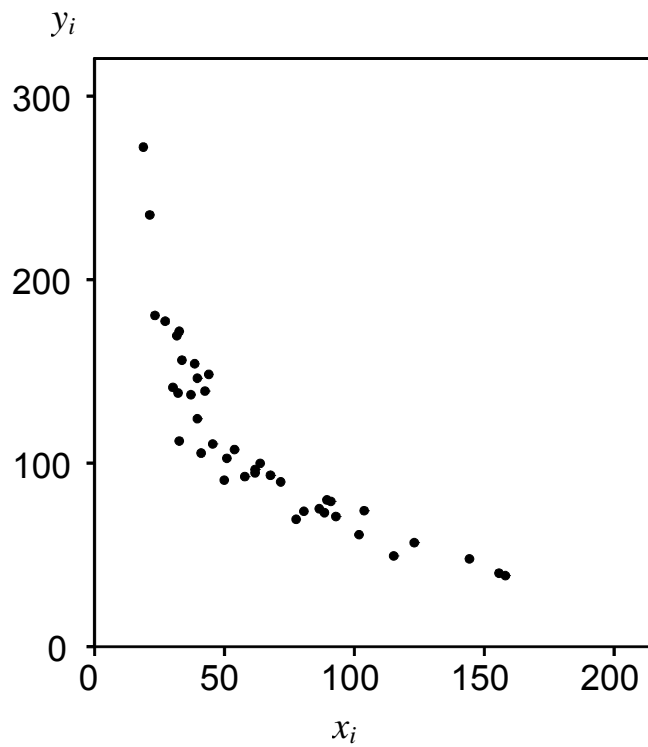


Abb. 6.18(a): Plot der Licht-Transmissions-Rate (y) gegen den Blattflächenindex (x) für Reisdaten.

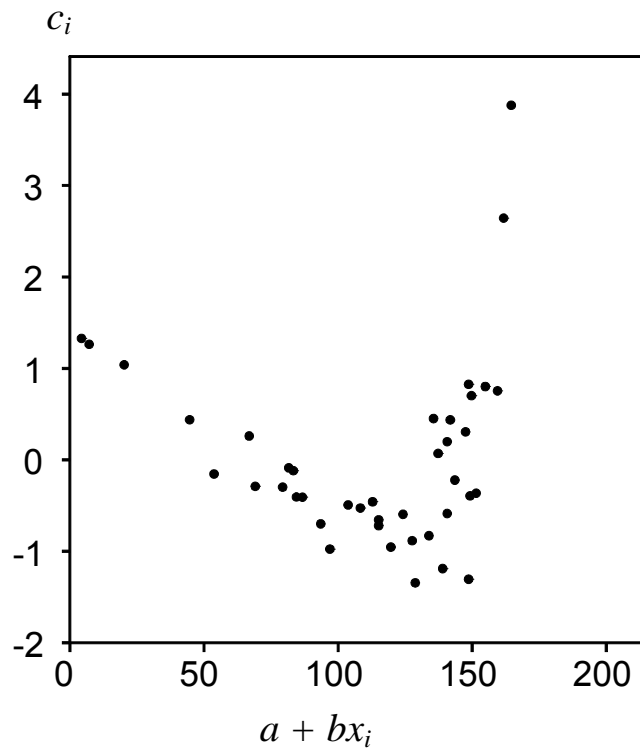


Abb. 6.18(b): Plot der Residuen c_i gegen den geschätzten Wert $a + bx_i$ für Reisdaten.

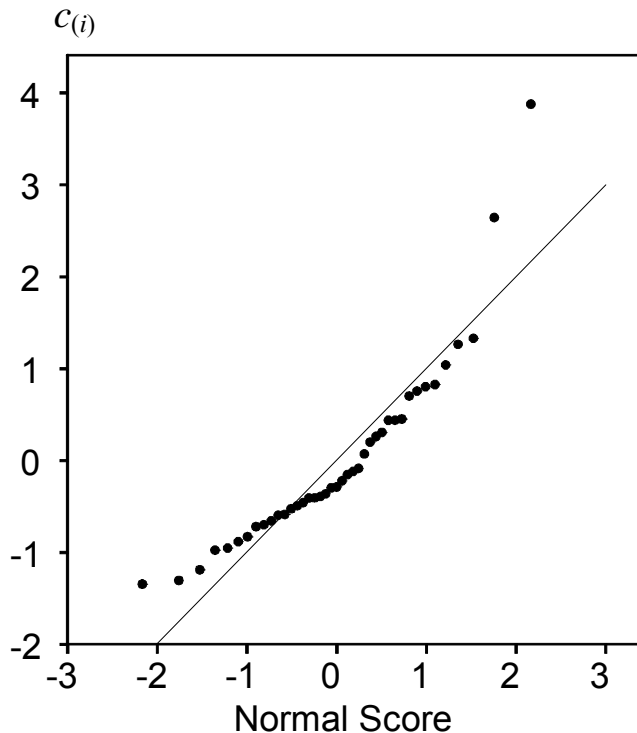


Abb. 6.18(c): Plot der geordneten Residuen $c_{(i)}$ gegen den Normal Score (u_i) für Reisdaten (Q-Q-Plot).

Beispiel (Ausreißer): In einer Erhebung wurden auf 91 Feldern der Ertrag (x) und der Proteingehalt (y) von Weizen ermittelt (Snedecor GW, Cochran WG 1967 Statistical methods. Sixth edition. Iowa State University Press, Ames, S. 454). Die Proteingehalte (y) sind in Abb. 6.19(a) gegen den Ertrag (x) geplottet, zusammen mit einem Polynom 2. Grades ($\eta_i = \alpha + \beta_1 x_i + \beta_2 x_i^2$), welches an die Daten angepasst wurde, um den Einfluß des Ertrages auf den Proteingehalt zu beschreiben. Der Q-Q-Plot weist einen Ausreißer aus. **Weiterer Analyseschritt:** Wird diese Beobachtung gelöscht, zeigt der Q-Q-Plot keine Abweichung mehr an. Der Ausreißer hat in diesem Fall eine nicht zu vernachlässigende Auswirkung auf die Varianzanalyse für die Regression, weil er den Versuchsfehler aufbläht. Die Residuenanalyse erhöht hier die Aussagekraft der anschließenden Varianzanalyse.

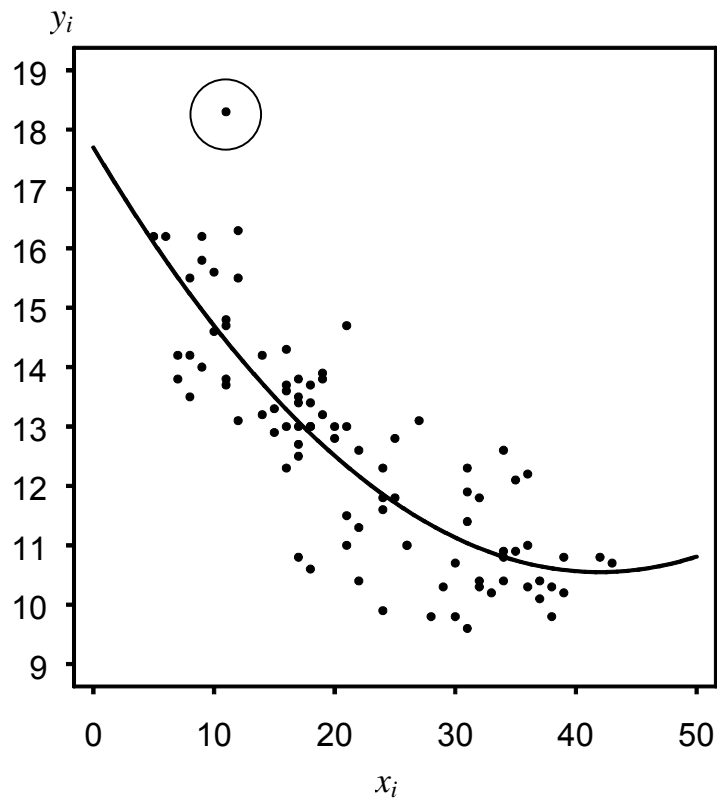


Abb. 6.19(a): Plot des Proteingehaltes (y) gegen den Ertrag (x). Daten mit Ausreißer (umkreist) und angepasstem Polynom 2. Grades für Weizendaten.

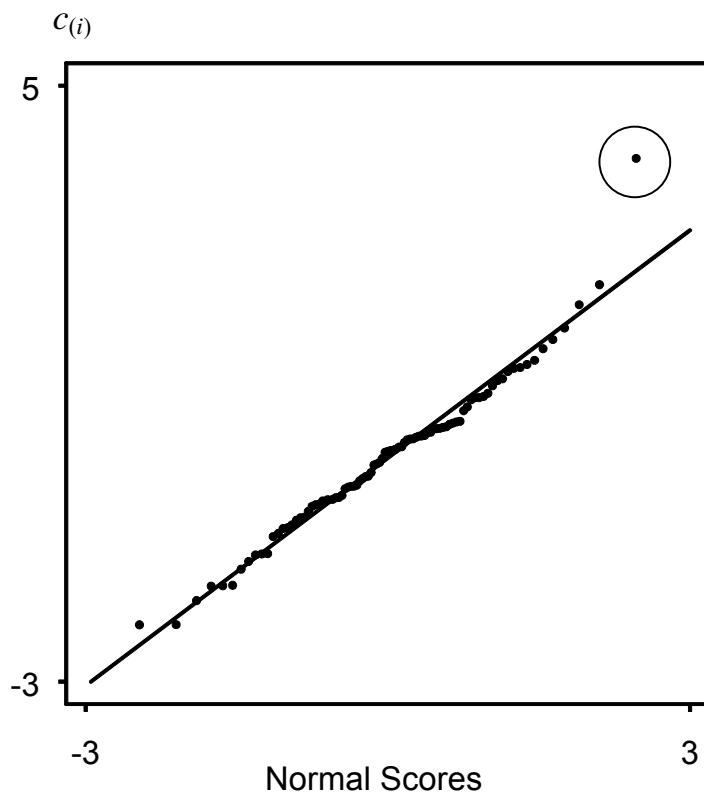


Abb. 6.19(b): Plot der geordneten Residuen $c(i)$ gegen den Normal Score (u_i) für Regendaten (Q-Q-Plot) für Weizendaten (Ausreißer eingekreist).

Beispiel (Varianzheterogenität): Um zu zeigen, wie sich Varianzheterogenität auf die Residuenplots auswirkt, wurden Daten simuliert. Hierzu wurde folgendes Modell verwendet:

$$y_i = 10 + 3x_i + e_i$$

$$x_i = i; i = 1, \dots, 100$$

e_i normalverteilt mit Mittelwert Null und Varianz $\text{var}(e_i) = x_i^2$.

Wir haben hier also eine lineare Regression mit $\alpha = 10$ und $\beta = 3$ vorliegen. Die Varianzen sind nicht homogen, sondern steigen mit dem Quadrat von x_i an. Diese zunehmende Varianz mit der Höhe von x_i ist auch den simulierten Daten anzusehen, die in Abb. 6.20(a) wiedergegeben sind. Der Plot der Residuen gegen den vorhergesagten Wert [Abb. 6.20(b)] hat die Form eines Megaphons und zeigt so die zunehmende Varianz deutlich an. **Weiterer Analyseschritt:** Varianzstabilisierende Transformation suchen. Problem: Nach Transformation der Daten wird das Modell nichtlinear. Daher z.B. besser eine gewichtete lineare Regression durchführen, bei der Beobachtungen mit kleiner Varianz ein relativ höheres Gewicht erhalten (Atkinson 1985: Plots, transformations and regression. Clarendon Press, London).

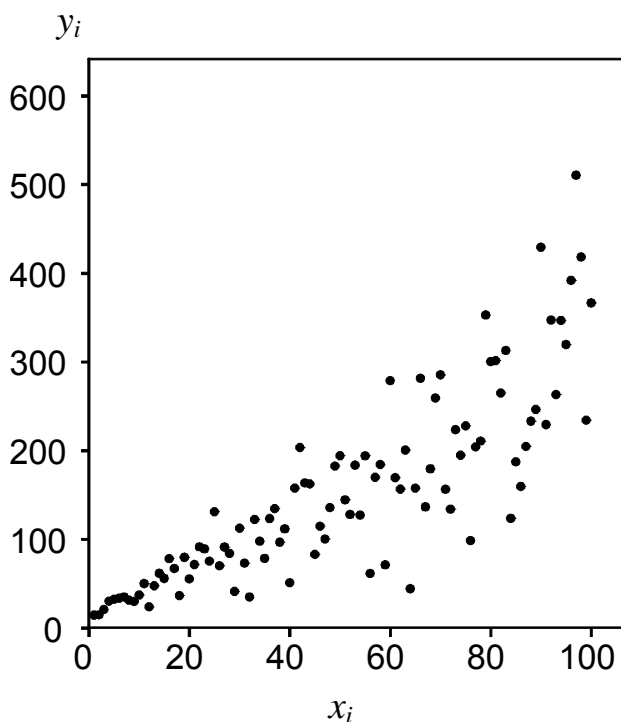


Abb. 6.20(a): Plot von y_i gegen x_i . Daten simuliert nach $y_i = 10 + 3x_i + e_i$, wobei $\text{var}(e_i) = x_i^2$.

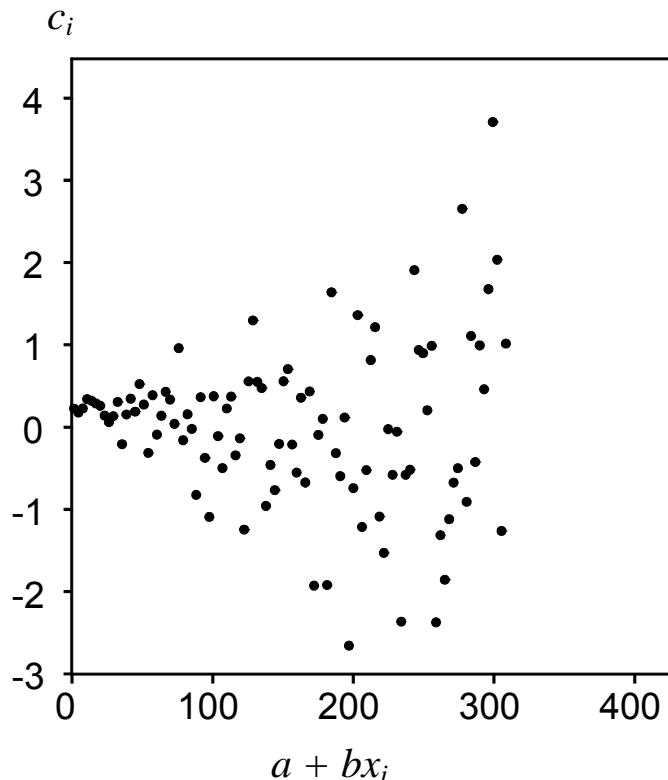


Abb. 6.20(b): Plot der Residuen c_i gegen geschätzten Wert $a + bx_i$. Daten simuliert nach $y_i = 10 + 3x_i + e_i$, wobei $\text{var}(e_i) = x_i^2$.

Beispiel (Abweichung von der Normalverteilung): Um zu zeigen, wie sich eine Abweichung von der Normalverteilung auf die Residuenplots auswirkt, wurden Daten simuliert. Hierzu wurde folgendes Modell verwendet:

$$y_i = x_i + e_i$$

$$x_i = 1; 1,1; 1,2; \dots, 10$$

$$e_i = \exp(f_i)$$

$$f_i \sim N(0, 1)$$

Die Fehler folgen hier einer log-Normalverteilung, da f_i standardnormalverteilt ist (Mittelwert Null und Varianz Eins). Der Erwartungswert folgt einer linearen Regression mit $\alpha = 0$ und $\beta = 1$.

Sowohl der Plot der Rohdaten [Abb. 6.21(a)] als auch der Plot der Residuen gegen den vorhergesagten Wert [Abb. 6.21(b)] deuten an, dass die Fehler eine sehr schiefe Verteilung aufweisen, so dass sehr große positive Abweichungen von der Regression auftreten, während sehr viele Abweichungen negativ, aber klein sind. Diese Tatsache macht sich sehr deutlich im Q-Q-Plot bemerkbar [Abb. 6.21(c)].

Weiterer Analyseschritt: Logarithmische Transformation. Problem: Nach Transformation der Daten wird Modell nichtlinear. Daher ist es besser, auch das Regressionsmodell transformieren ("transform-both-sides"; Carroll RJ, Ruppert D 1988: Transformation and weighting in regression. Chapman and Hall, London).

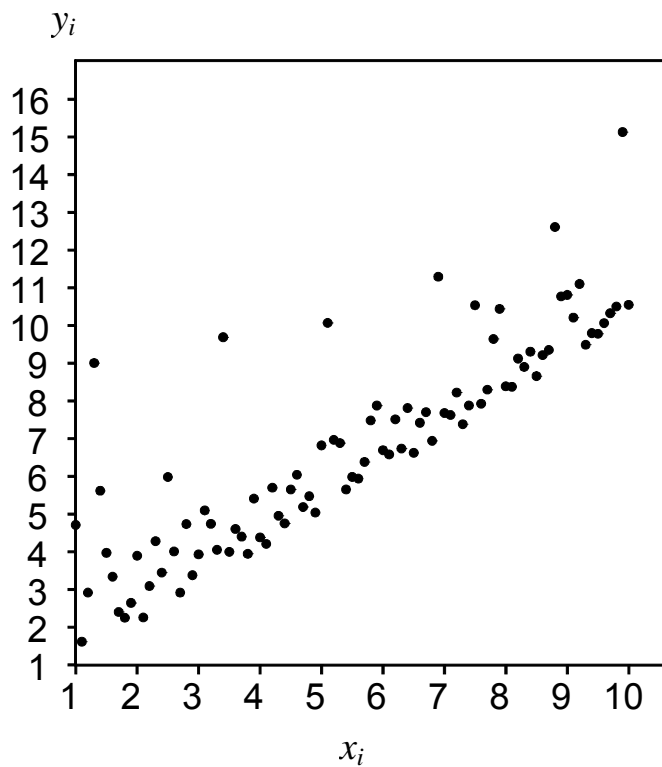


Abb. 6.21(a): Plot von y_i gegen x_i . Daten simuliert nach $y_i = x_i + e_i$, wobei $e_i = \exp(f_i)$ und $f_i \sim N(0, 1)$ (f_i ist standardnormalverteilt).

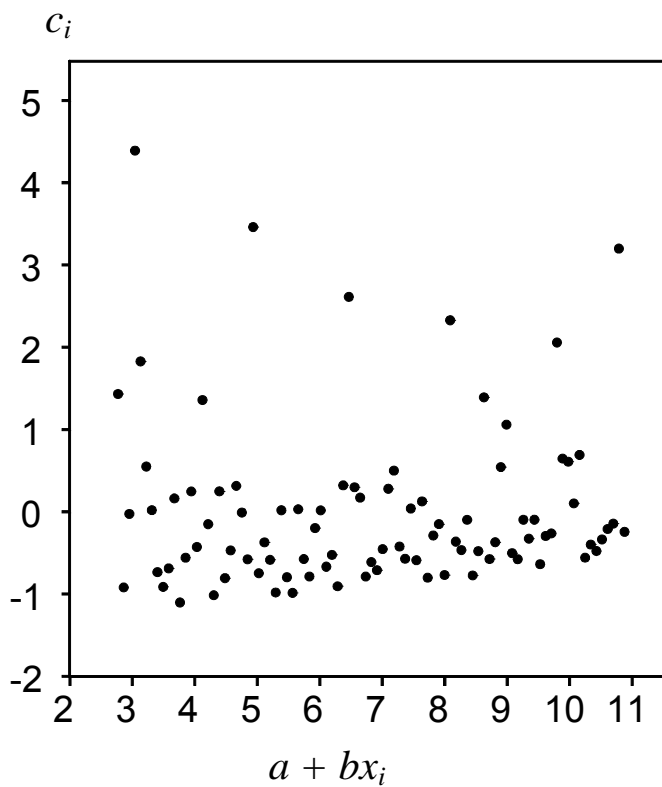


Abb. 6.21(b): Plot der Residuen c_i gegen geschätzten Wert $a + bx_i$. Daten simuliert nach $y_i = x_i + e_i$, wobei $e_i = \exp(f_i)$ und $f_i \sim N(0, 1)$ (f_i ist standardnormalverteilt).

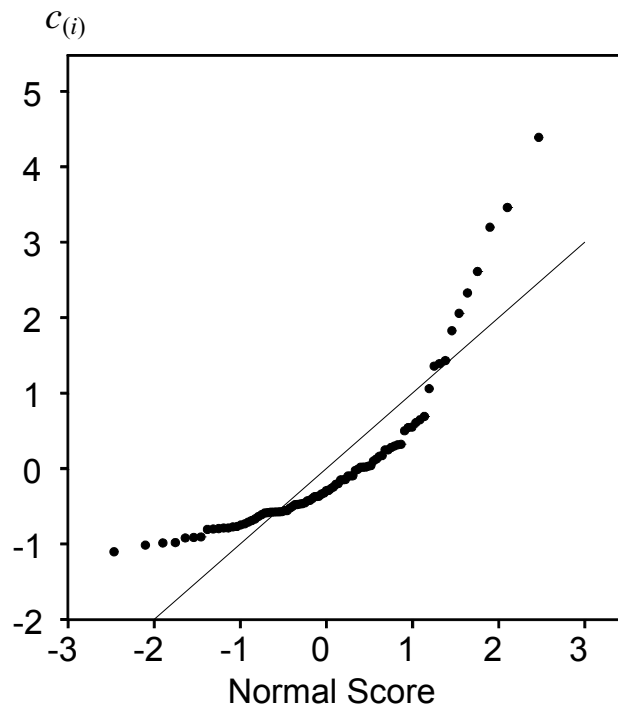


Abb. 6.21(c): Plot der geordneten Residuen $c_{(i)}$ gegen den Normal Score (u_i) für Regendaten (Q-Q-Plot). Daten simuliert nach $y_i = x_i + e_i$, wobei $e_i = \exp(f_i)$ und $f_i \sim N(0, 1)$ (f_i ist standardnormalverteilt).

6.8 Lineare Modelle in Matrizenschreibweise

Wir haben bisher verschiedene lineare Modelle kennen gelernt, und zwar im Zusammenhang mit der linearen Regression und der Varianzanalyse (Kap. 4). Diese Modelle können kompakt in Matrizen-Schreibweise ausgedrückt werden. Die Matrizenschreibweise wird hier eingeführt, weil sie die Behandlung weiterführender Modelle entscheidend vereinfacht und vereinheitlicht, wie am Beispiel der multiplen Regression (Abschnitt 6.10) und der Polynomregression (Abschnitt 6.11) deutlich wird. Es soll gezeigt werden, dass die Vielzahl der in der Praxis verwendeten linearen Modelle unter ein gemeinsames Dach gefasst werden kann. Die hier und im folgenden Abschnitt 6.9 präsentierten allgemeinen Resultate sind auch deshalb für die Anwendung relevant, weil diese in allen gängigen Statistik-Paketen für die Analyse linearer Modelle genutzt werden. Wichtige Grundregeln der Matrizenrechnung sind im Anhang dieses Kapitels beschrieben.

Tab. 6.8.1: Regression von Ertrag (y) auf Regenmenge zwischen April und Juni (x).

Jahr	i	x	y	Modell (symbolisch)	Modell (mit Daten)
1900	1	177	26,2	$y_1 = \alpha + \beta x_1 + e_1$	$26,2 = \alpha + 177\beta + e_1$
1901	2	96	25,0	$y_2 = \alpha + \beta x_2 + e_2$	$25,0 = \alpha + 96\beta + e_2$
1902	3	144	32,6	$y_3 = \alpha + \beta x_3 + e_3$	$32,6 = \alpha + 144\beta + e_3$
1903	4	105	26,6	$y_4 = \alpha + \beta x_4 + e_4$	$26,6 = \alpha + 105\beta + e_4$
1904	5	111	19,6	$y_5 = \alpha + \beta x_5 + e_5$	$19,6 = \alpha + 111\beta + e_5$
1905	6	135	20,4	$y_6 = \alpha + \beta x_6 + e_6$	$20,4 = \alpha + 135\beta + e_6$
1906	7	209	29,2	$y_7 = \alpha + \beta x_7 + e_7$	$29,2 = \alpha + 209\beta + e_7$
1907	8	161	33,8	$y_8 = \alpha + \beta x_8 + e_8$	$33,8 = \alpha + 161\beta + e_8$
1908	9	246	26,6	$y_9 = \alpha + \beta x_9 + e_9$	$26,6 = \alpha + 246\beta + e_9$
1909	10	108	22,6	$y_{10} = \alpha + \beta x_{10} + e_{10}$	$22,6 = \alpha + 108\beta + e_{10}$
1910	11	137	24,2	$y_{11} = \alpha + \beta x_{11} + e_{11}$	$24,2 = \alpha + 137\beta + e_{11}$
1911	12	71	16,6	$y_{12} = \alpha + \beta x_{12} + e_{12}$	$16,6 = \alpha + 71\beta + e_{12}$
1912	13	119	29,8	$y_{13} = \alpha + \beta x_{13} + e_{13}$	$29,8 = \alpha + 119\beta + e_{13}$
1913	14	108	19,4	$y_{14} = \alpha + \beta x_{14} + e_{14}$	$19,4 = \alpha + 108\beta + e_{14}$
1914	15	132	30,6	$y_{15} = \alpha + \beta x_{15} + e_{15}$	$30,6 = \alpha + 132\beta + e_{15}$
1915	16	89	16,4	$y_{16} = \alpha + \beta x_{16} + e_{16}$	$16,4 = \alpha + 89\beta + e_{16}$
1916	17	147	30,4	$y_{17} = \alpha + \beta x_{17} + e_{17}$	$30,4 = \alpha + 147\beta + e_{17}$
1917	18	98	19,2	$y_{18} = \alpha + \beta x_{18} + e_{18}$	$19,2 = \alpha + 98\beta + e_{18}$
1918	19	106	20,2	$y_{19} = \alpha + \beta x_{19} + e_{19}$	$20,2 = \alpha + 106\beta + e_{19}$
1919	20	123	25,6	$y_{20} = \alpha + \beta x_{20} + e_{20}$	$25,6 = \alpha + 123\beta + e_{20}$
1920	21	156	31,0	$y_{21} = \alpha + \beta x_{21} + e_{21}$	$31,0 = \alpha + 156\beta + e_{21}$
1921	22	191	35,8	$y_{22} = \alpha + \beta x_{22} + e_{22}$	$35,8 = \alpha + 191\beta + e_{22}$
1922	23	162	31,6	$y_{23} = \alpha + \beta x_{23} + e_{23}$	$31,6 = \alpha + 162\beta + e_{23}$
1923	24	235	33,6	$y_{24} = \alpha + \beta x_{24} + e_{24}$	$33,6 = \alpha + 235\beta + e_{24}$
1924	25	147	30,4	$y_{25} = \alpha + \beta x_{25} + e_{25}$	$30,4 = \alpha + 147\beta + e_{25}$
1925	26	110	30,2	$y_{26} = \alpha + \beta x_{26} + e_{26}$	$30,2 = \alpha + 110\beta + e_{26}$

Beispiel: Im Zusammenhang mit der Regression lautete ein einfaches lineares Modell:

$$y_i = \alpha + \beta x_i + e_i$$

wobei i ein Index für die Beobachtungen ist. Für das Beispiel der Regendaten (Abb. 6.1; Anfang von Kap. 6) wurde eine Regression des Ertrages (y_i) auf die Regenmenge durchgeführt. In der Tabelle 6.8.1 ist das Modell für jede einzelne Beobachtung explizit aufgeschlüsselt. Man kann das Modell für alle Beobachtungen auch kompakt in Matrizen-Schreibweise ausdrücken:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \\ y_{17} \\ y_{18} \\ y_{19} \\ y_{20} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \\ y_{26} \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \\ 1 & x_7 \\ 1 & x_8 \\ 1 & x_9 \\ 1 & x_{10} \\ 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{13} \\ 1 & x_{14} \\ 1 & x_{15} \\ 1 & x_{16} \\ 1 & x_{17} \\ 1 & x_{18} \\ 1 & x_{19} \\ 1 & x_{20} \\ 1 & x_{21} \\ 1 & x_{22} \\ 1 & x_{23} \\ 1 & x_{24} \\ 1 & x_{25} \\ 1 & x_{26} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \\ e_{17} \\ e_{18} \\ e_{19} \\ e_{20} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{25} \\ e_{26} \end{pmatrix} \quad \text{bzw.} \quad \begin{pmatrix} 26,2 \\ 25,0 \\ 32,6 \\ 26,6 \\ 19,6 \\ 20,4 \\ 29,2 \\ 33,8 \\ 26,6 \\ 22,6 \\ 24,2 \\ 16,6 \\ 29,8 \\ 19,4 \\ 30,6 \\ 16,4 \\ 30,4 \\ 19,2 \\ 20,2 \\ 25,6 \\ 31,0 \\ 35,8 \\ 31,6 \\ 33,6 \\ 30,4 \\ 30,2 \end{pmatrix} = \begin{pmatrix} 1 & 177 \\ 1 & 96 \\ 1 & 144 \\ 1 & 105 \\ 1 & 111 \\ 1 & 135 \\ 1 & 209 \\ 1 & 161 \\ 1 & 246 \\ 1 & 108 \\ 1 & 137 \\ 1 & 71 \\ 1 & 119 \\ 1 & 108 \\ 1 & 132 \\ 1 & 89 \\ 1 & 147 \\ 1 & 98 \\ 1 & 106 \\ 1 & 123 \\ 1 & 156 \\ 1 & 191 \\ 1 & 162 \\ 1 & 235 \\ 1 & 147 \\ 1 & 110 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \\ e_{17} \\ e_{18} \\ e_{19} \\ e_{20} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{25} \\ e_{26} \end{pmatrix}$$

Mit

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \\ y_{17} \\ y_{18} \\ y_{19} \\ y_{20} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \\ y_{26} \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \\ 1 & x_7 \\ 1 & x_8 \\ 1 & x_9 \\ 1 & x_{10} \\ 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{13} \\ 1 & x_{14} \\ 1 & x_{15} \\ 1 & x_{16} \\ 1 & x_{17} \\ 1 & x_{18} \\ 1 & x_{19} \\ 1 & x_{20} \\ 1 & x_{21} \\ 1 & x_{22} \\ 1 & x_{23} \\ 1 & x_{24} \\ 1 & x_{25} \\ 1 & x_{26} \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \\ e_{17} \\ e_{18} \\ e_{19} \\ e_{20} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{25} \\ e_{26} \end{pmatrix}$$

und

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

kann dies kompakt geschrieben werden als

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad .$$

Die Matrix \mathbf{X} wird auch als **Design-Matrix** bezeichnet.

Beispiel: In Kap. 4 hatten wir in Zusammenhang mit der einfachen Varianzanalyse einen Versuch mit 5 Sorten in 4 Wiederholungen betrachtet. Das Modell lautete:

$$y_{ij} = \mu + \tau_i + e_{ij}$$

wobei

y_{ij} = j -te Wiederholung der i -ten Sorte

μ = Gesamteffekt

τ_i = Effekt der i -ten Sorte

e_{ij} = Fehler von y_{ij}

Die Daten waren wie folgt:

Messwerte	Sorte					
	A	B	C	D	E	
1	$y_{11} = 31$	$y_{21} = 21$	$y_{31} = 27$	$y_{41} = 34$	$y_{51} = 24$	
2	$y_{12} = 32$	$y_{22} = 23$	$y_{32} = 29$	$y_{42} = 32$	$y_{52} = 23$	
3	$y_{13} = 37$	$y_{23} = 25$	$y_{33} = 34$	$y_{43} = 31$	$y_{53} = 27$	
4	$y_{14} = 32$	$y_{24} = 19$	$y_{34} = 34$	$y_{44} = 27$	$y_{54} = 26$	
Summe	$y_{1\bullet} = 132$	$y_{2\bullet} = 88$	$y_{3\bullet} = 124$	$y_{4\bullet} = 124$	$y_{5\bullet} = 100$	$y_{\bullet\bullet} = 568$
Mittelwert	$\bar{y}_{1\bullet} = 33$	$\bar{y}_{2\bullet} = 22$	$\bar{y}_{3\bullet} = 31$	$\bar{y}_{4\bullet} = 31$	$\bar{y}_{5\bullet} = 25$	$\bar{y}_{\bullet\bullet} = 28,4$

In Matrizenschreibweise lautet das Modell

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \\ y_{41} \\ y_{42} \\ y_{43} \\ y_{44} \\ y_{51} \\ y_{52} \\ y_{53} \\ y_{54} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \\ e_{41} \\ e_{42} \\ e_{43} \\ e_{44} \\ e_{51} \\ e_{52} \\ e_{53} \\ e_{54} \end{pmatrix}$$

bzw. $y = X\beta + e$ mit

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \\ y_{41} \\ y_{42} \\ y_{43} \\ y_{44} \\ y_{51} \\ y_{52} \\ y_{53} \\ y_{54} \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \\ e_{41} \\ e_{42} \\ e_{43} \\ e_{44} \\ e_{51} \\ e_{52} \\ e_{53} \\ e_{54} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \end{pmatrix}$$

6.8.1 Die Normalengleichungen und ihre Lösung

Beispiel: Kommen wir zurück zur linearen Regression. Die Parameter werden mit der Methode der Kleinsten Quadrate geschätzt. Die Summe der Fehlerquadrate ist gleich

$$SQ_{\text{Fehler}} = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

wobei a und b Schätzungen für α und β sind, die so gewählt werden, dass das SQ_{Fehler} minimal wird. Zur Lösung dieser Optimierungsaufgabe berechnet man die

Ableitung von SQ_{Fehler} nach a und b und setzt diese gleich Null, was zu den Normalengleichungen führt:

$$\frac{\partial SQ_{Fehler}}{\partial b} = 2 \sum_{i=1}^n [y_i - (a + bx_i)](-1)x_i = 0$$

$$\frac{\partial SQ_{Fehler}}{\partial a} = 2 \sum_{i=1}^n [y_i - (a + bx_i)](-1) = 0$$

Wir können das SQ_{Fehler} in Matrizenschreibweise ausdrücken. Hierzu nutzen wir, dass der Vektor

$$\mathbf{y} - \mathbf{Xb} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 - (a + bx_1) \\ y_2 - (a + bx_2) \\ \vdots \\ y_n - (a + bx_n) \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix} = \mathbf{r}$$

die Residuen $r_i = y_i - (a + bx_i)$ enthält. Die Summe der Fehlerquadrate entspricht der Summe der quadrierten Residuen, also

$$SQ_{Fehler} = \sum_{i=1}^n [y_i - (a + bx_i)]^2 = \sum_{i=1}^n r_i^2 = \mathbf{r}'\mathbf{r} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$$

Diese Form der SQ_{Fehler} gilt ganz allgemein für jedes lineare Modell.

Die Summe der Fehlerquadrate im linearen Modell hat die Form

$$SQ_{Fehler} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$$

Zur Berechnung der Ableitung der SQ_{Fehler} nach den Parametern ist es zunächst hilfreich, die Matrix \mathbf{X} zu partitionieren als

$$\mathbf{X} = (\mathbf{1} \quad \mathbf{x})$$

wobei

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} \quad \text{und} \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} \quad \text{sind.}$$

Für die Ableitungen findet man

$$\frac{\partial(\mathbf{X}\mathbf{b})}{\partial a} = \begin{pmatrix} \mathbf{1} & \mathbf{x} \end{pmatrix} \begin{pmatrix} \partial a / \partial a \\ \partial b / \partial a \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{x} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbf{1}$$

$$\frac{\partial(\mathbf{X}\mathbf{b})}{\partial b} = \begin{pmatrix} \mathbf{1} & \mathbf{x} \end{pmatrix} \begin{pmatrix} \partial a / \partial b \\ \partial b / \partial b \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{x} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \mathbf{x}$$

Somit sind die Normalengleichungen (Produktregel!)

$$\frac{\partial S_{\text{Fehler}}}{\partial a} = - \left(\frac{\partial(\mathbf{X}\mathbf{b})}{\partial a} \right)' (\mathbf{y} - \mathbf{X}\mathbf{b}) - (\mathbf{y} - \mathbf{X}\mathbf{b})' \left(\frac{\partial(\mathbf{X}\mathbf{b})}{\partial a} \right) = -2\mathbf{1}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = 0$$

$$\frac{\partial S_{\text{Fehler}}}{\partial b} = - \left(\frac{\partial(\mathbf{X}\mathbf{b})}{\partial b} \right)' (\mathbf{y} - \mathbf{X}\mathbf{b}) - (\mathbf{y} - \mathbf{X}\mathbf{b})' \left(\frac{\partial(\mathbf{X}\mathbf{b})}{\partial b} \right) = -2\mathbf{x}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = 0$$

Kombination der beiden Gleichungen ergibt nach Division durch -2

$$\begin{pmatrix} \mathbf{1} & \mathbf{x} \end{pmatrix}' (\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$$

und somit

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$$

Dies ist die Matrix-Form der Normalengleichungen. Lösen dieser Gleichung nach \mathbf{b} liefert die Kleinstquadratschätzung der Regressionsparameter. Falls die Matrix $\mathbf{X}'\mathbf{X}$ invertierbar ist, was bei der linearen Regression der Fall ist, finden wir:

$$(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{b} \Leftrightarrow \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}, \text{ da}$$

$$(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X} = \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Generell gilt folgendes für die Kleinst-Quadrat-Schätzung der Parameter eines linearen Modells:

Die Kleinst-Quadrat-Lösung des Parametervektors $\boldsymbol{\beta}$ im linearen Modell

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

ist (sofern \mathbf{X} von vollem Rang ist) gegeben durch

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

wobei $(\mathbf{X}'\mathbf{X})^{-1}$ die Inverse von $\mathbf{X}'\mathbf{X}$ ist.

Diese Gleichung für die Kleinstquadratschätzung der Parameter ist generell gültig, sofern die Matrix $X'X$ invertierbar ist. Dies ist immer dann der Fall, wenn die Matrix X von vollem Rang ist, es also keine linearen Abhängigkeiten zwischen den Spalten von X gibt. Man beachte, dass im Beispiel des varianzanalytischen Modells eine solche Abhängigkeit besteht, da die erste Spalte gleich der Summe der übrigen Spalten ist. Auf dieses Problem gehen wir gesondert ein.

Im folgenden soll nun gezeigt werden, dass das obige allgemeine Resultat im Fall der linearen Regression zu bekannten Schätzformeln führt. Für die Berechnung der Kleinst-Quadrat-Lösung der linearen Regression sind folgende Ausdrücke hilfreich:

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \quad \text{und} \quad X'y = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

Die Inverse einer symmetrischen Matrix der Form $\begin{pmatrix} c & e \\ e & d \end{pmatrix}$ ist (siehe auch Anhang zu diesem Kapitel), wie man leicht nachprüft,

$$\begin{pmatrix} c & e \\ e & d \end{pmatrix}^{-1} = \frac{1}{cd - e^2} \begin{pmatrix} d & -e \\ -e & c \end{pmatrix}$$

so dass

$$(X'X)^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} = \frac{1}{nSQ_x} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

Hiermit finden wir für die KleinstquadratLösung:

$$\begin{aligned} b &= (X'X)^{-1} X'y = \frac{1}{nSQ_x} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\ &= \frac{1}{nSQ_x} \begin{pmatrix} \left(\sum_{i=1}^n x_i^2 \right) \left(\sum_{i=1}^n y_i \right) - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n x_i y_i \right) \\ - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) + n \sum_{i=1}^n x_i y_i \end{pmatrix} \end{aligned}$$

Weitere Umformung ergibt:

$$- \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) + n \sum_{i=1}^n x_i y_i = nSP_{xy}$$

und

$$\begin{aligned} & \left(\sum_{i=1}^n x_i^2 \right) \left(\sum_{i=1}^n y_i \right) - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n x_i y_i \right) = \\ & \underbrace{\left(\sum_{i=1}^n x_i^2 \right) \left(\sum_{i=1}^n y_i \right) - n^{-1} \left(\sum_{i=1}^n x_i \right)^2 \left(\sum_{i=1}^n y_i \right)}_{nSQ_x \bar{y}.} + \underbrace{n^{-1} \left(\sum_{i=1}^n x_i \right)^2 \left(\sum_{i=1}^n y_i \right) - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n x_i y_i \right)}_{nSP_{xy} \bar{x}.} \\ & = \quad \quad \quad nSQ_x \bar{y}. \quad \quad \quad - \quad \quad \quad nSP_{xy} \bar{x}. \end{aligned}$$

so dass

$$\mathbf{b} = \frac{1}{nSQ_x} \begin{pmatrix} nSQ_x \bar{y}. - nSP_{xy} \bar{x}. \\ nSP_{xy} \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

mit

$$b = \frac{SP_{xy}}{SQ_x} \quad \text{und}$$

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

Dies sind genau die Rechenformeln für die lineare Regression, wie sie in Abschnitt 6.2 verwendet werden!

Beispiel: Anstatt mit den Rechenformeln aus Abschnitt 6.2 zu rechnen, können wir auch die allgemeinen Matrix-Resultate aus dem aktuellen Abschnitt verwenden (Ein PC Statistik-Programm für lineare Modelle tut genau das!). Für die Regendaten finden wir:

$$\mathbf{X}\mathbf{X} = \begin{pmatrix} 26 & 3623 \\ 3623 & 553187 \end{pmatrix} \quad \text{und} \quad \mathbf{X}\mathbf{y} = \begin{pmatrix} 687,6 \\ 99737,8 \end{pmatrix}$$

Mit

$$\begin{pmatrix} c & e \\ e & d \end{pmatrix}^{-1} = \frac{1}{cd - e^2} \begin{pmatrix} d & -e \\ -e & c \end{pmatrix}$$

ist

$$(\mathbf{X}\mathbf{X})^{-1} = \frac{1}{26 \cdot 553187 - 3623^2} \begin{pmatrix} 553187 & -3623 \\ -3623 & 26 \end{pmatrix} = \begin{pmatrix} 0,440178 & -0,002883 \\ -0,002883 & 0,0000207 \end{pmatrix}$$

und

$$\mathbf{b} = (\mathbf{X}\mathbf{X})^{-1} \mathbf{X}\mathbf{y} = \begin{pmatrix} 0,440178 & -0,002883 \\ -0,002883 & 0,0000207 \end{pmatrix} \begin{pmatrix} 687,6 \\ 99737,8 \end{pmatrix} = \begin{pmatrix} 15,1355 \\ 0,08117 \end{pmatrix}$$

Die Kleinstquadratschätzung der Geradengleichung lautet also:

$$\hat{y} = 15,14 + 0,0812x$$

Dies hatten wir bereits in Abschnitt 6.2 mit skalaren Methoden gefunden.

Beispiel: Bei der Varianzanalyse des Sortenversuches mit $t = 5$ Sorten und $r = 4$ Wiederholungen je Sorte (vollständig randomisiert) ist die Designmatrix \mathbf{X} nicht von vollem Rang, weil lineare Abhängigkeiten zwischen den Spalten bestehen. Daher ist die Matrix $\mathbf{X}'\mathbf{X}$ nicht invertierbar, und es gibt keine eindeutigen Kleinst-Quadrat-Lösungen. Eine Lösung dieses Problems besteht in der Verwendung einer sog. generalisierten Inversen von $\mathbf{X}'\mathbf{X}$ zur Lösung der Normalengleichungen. Hier soll der Einfachheit halber eine alternative Lösung dargestellt werden. Sie besteht darin, eine Restriktion für die Parameter einzuführen. Beispielsweise können wir ohne Verlust der Allgemeinheit

$$\tau_5 = 0$$

setzen und diesen Effekt aus dem Parametervektor nehmen. Die Begründung für dieses Vorgehen soll an einem einfachen Beispiel geliefert werden. Hierzu betrachten wir nur die ersten beiden Sorten. Deren Erwartungswerte sind nach dem varianzanalytischen Modell gegeben durch

$$\mu_1 = \mu + \tau_1$$

$$\mu_2 = \mu + \tau_2$$

Angenommen, die Erwartungswerte sind

$$\mu_1 = \mu + \tau_1 = 40$$

$$\mu_2 = \mu + \tau_2 = 50$$

Es gibt nun unendlich viele Möglichkeiten, den drei Parametern μ , τ_1 und τ_2 Werte zuzuweisen, so dass sich die genannten Erwartungswerte ergeben. Hierzu drei Beispiele:

$$\mu = 100$$

$$\tau_1 = -60$$

$$\tau_2 = -50$$

$$\mu = 45$$

$$\tau_1 = -5$$

$$\tau_2 = 5$$

$$\mu = 50$$

$$\tau_1 = -10$$

$$\tau_2 = 0$$

In allen drei Fällen finden wir

$$\mu_1 = \mu + \tau_1 = 40$$

$$\mu_2 = \mu + \tau_2 = 50$$

$$\mu_1 - \mu_2 = -10$$

$$\tau_1 - \tau_2 = -10$$

Während also die Werte für die Effekte nicht eindeutig zu bestimmen sind, ergibt sich immer derselbe Erwartungswert. Außerdem ist die Differenz der Behandlungseffekte τ_1 und τ_2 immer gleich der Differenz der Erwartungswerte.

Zur Schätzung der Parameter kann man nun ohne Verlust der Allgemeinheit einen der drei Effekte auf einen beliebigen Wert festlegen.

Beim Sortenversuch (5 Sorten) verwenden wir nun diesen Überlegungen entsprechend

$$\tau_5 = 0$$

Wir haben dann folgendes Modell:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \\ y_{41} \\ y_{42} \\ y_{43} \\ y_{44} \\ y_{51} \\ y_{52} \\ y_{53} \\ y_{54} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \\ e_{41} \\ e_{42} \\ e_{43} \\ e_{44} \\ e_{51} \\ e_{52} \\ e_{53} \\ e_{54} \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{e}$$

Hier ist nun \mathbf{X} von vollem Rang, da keine linearen Abhängigkeiten zwischen den Spalten von \mathbf{X} mehr bestehen. Die Normalgleichungen lauten

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$$

wobei

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} rt & r & r & r & r \\ r & r & 0 & 0 & 0 \\ r & 0 & r & 0 & 0 \\ r & 0 & 0 & r & 0 \\ r & 0 & 0 & 0 & r \end{pmatrix} = r \begin{pmatrix} t & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \\ \hat{\tau}_4 \end{pmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{3.} \\ y_{4.} \end{pmatrix}$$

t = Zahl der Behandlungen

r = Zahl der Wiederholungen je Behandlung

Die Hut-Notation in \mathbf{b} symbolisiert die Kleinst-Quadrat-Lösung des betreffenden Parameters. Mit

$$(\mathbf{X}'\mathbf{X})^{-1} = r^{-1} \begin{pmatrix} 1 & -1 & -1 & -1 & -1 \\ -1 & 2 & 1 & 1 & 1 \\ -1 & 1 & 2 & 1 & 1 \\ -1 & 1 & 1 & 2 & 1 \\ -1 & 1 & 1 & 1 & 2 \end{pmatrix}$$

(diese Form der Inversen kann man mit dem Computer errechnen) finden wir

$$\begin{aligned} \mathbf{b} = \begin{pmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \\ \hat{\tau}_4 \end{pmatrix} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = r^{-1} \begin{pmatrix} 1 & -1 & -1 & -1 & -1 \\ -1 & 2 & 1 & 1 & 1 \\ -1 & 1 & 2 & 1 & 1 \\ -1 & 1 & 1 & 2 & 1 \\ -1 & 1 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{3.} \\ y_{4.} \end{pmatrix} \\ &= r^{-1} \begin{pmatrix} y_{5.} \\ y_{1.} - y_{5.} \\ y_{2.} - y_{5.} \\ y_{3.} - y_{5.} \\ y_{4.} - y_{5.} \end{pmatrix} = \begin{pmatrix} \bar{y}_{5.} \\ \bar{y}_{1.} - \bar{y}_{5.} \\ \bar{y}_{2.} - \bar{y}_{5.} \\ \bar{y}_{3.} - \bar{y}_{5.} \\ \bar{y}_{4.} - \bar{y}_{5.} \end{pmatrix} = \begin{pmatrix} 25 \\ 8 \\ -3 \\ 6 \\ 6 \end{pmatrix} \end{aligned}$$

Hieraus ermitteln wir die Kleinst-Quadrat-Schätzungen der Erwartungswerte:

$$\hat{\mu}_1 = \hat{\mu} + \hat{\tau}_1 = \bar{y}_{5.} + \bar{y}_{1.} - \bar{y}_{5.} = \bar{y}_{1.} = 33$$

$$\hat{\mu}_2 = \hat{\mu} + \hat{\tau}_2 = \bar{y}_{5.} + \bar{y}_{2.} - \bar{y}_{5.} = \bar{y}_{2.} = 22$$

$$\hat{\mu}_3 = \hat{\mu} + \hat{\tau}_3 = \bar{y}_{5.} + \bar{y}_{3.} - \bar{y}_{5.} = \bar{y}_{3.} = 31$$

$$\hat{\mu}_4 = \hat{\mu} + \hat{\tau}_4 = \bar{y}_{5.} + \bar{y}_{4.} - \bar{y}_{5.} = \bar{y}_{4.} = 31$$

$$\hat{\mu}_5 = \hat{\mu} + \hat{\tau}_5 = \bar{y}_{5.} + 0 = \bar{y}_{5.} = 25 \quad (\text{Restriktion } \tau_5 = 0!)$$

Die Kleinst-Quadrat-Schätzungen der Erwartungswerte der Behandlungen entsprechen also den Stichprobenmittelwerten der Behandlungen.

Übrigens hätten wir dasselbe Ergebnis mit jeder beliebigen anderen Restriktion erhalten, z.B. $\tau_1 = 0$, $\tau_3 = 1000$ oder $\sum_i \tau_i = 0$. Schließlich könnten wir auch einfach den Gesamteffekt μ aus dem Modell nehmen, so dass $\tau_i = \mu_i$ ist.

6.8.2 Vertrauensintervalle und Tests für Linearkombinationen der Parameter

Im linearen Modell sind wir oft interessiert an Vertrauensintervallen oder Tests für Linearkombinationen der Parameter, die geschrieben werden können als

$$\lambda = \mathbf{k}'\boldsymbol{\beta}$$

Hierbei ist $\boldsymbol{\beta}$ der Parametervektor und \mathbf{k} ein Vektor mit Konstanten. Die meisten in einer statistischen Analyse interessierenden Größen haben diese Form, z.B. Mittelwertdifferenzen oder Schätzungen durch eine Regressionsgerade. Es ist von Vorteil, diese Größen in der allgemeinen Form $\lambda = \mathbf{k}'\boldsymbol{\beta}$ zu schreiben, weil dann wieder allgemeine Resultate für lineare Modelle angegeben werden können, mit denen sich jeder Spezialfall behandeln lässt. Die Wahl des Vektors \mathbf{k} hängt von der jeweiligen Fragestellung ab und muss vom Benutzer erfolgen. Im einfachsten Fall will man die einzelnen Parameterschätzungen beurteilen, so dass \mathbf{k} ein Vektor mit lauter Nullen und einer Eins an der Stelle des in Frage stehenden Parameters ist. Einige weitere Beispiele für die Wahl von \mathbf{k} werden im folgenden gegeben. In Statistik-Paketen müssen die Koeffizienten für das jeweilige \mathbf{k} separat angegeben werden, in SAS z.B. über die ESTIMATE Anweisung. Um Tests und Vertrauensintervalle zu berechnen, werden für die Parameter die Kleinst-Quadrat-Schätzungen eingesetzt:

Die Kleinst-Quadrat-Schätzung der Linearkombination $\lambda = \mathbf{k}'\boldsymbol{\beta}$, ist gegeben durch

$$\hat{\lambda} = \mathbf{k}'\mathbf{b}$$

wobei \mathbf{k} ein vom Benutzer zu bestimmender Koeffizientenvektor ist und

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad .$$

Beispiel: In der Varianzanalyse ist der Vergleich von Behandlungsmittelwerten und damit von Behandlungseffekten von Interesse, z.B.

$$\lambda = \mu_1 - \mu_2 = \tau_1 - \tau_2 = \mathbf{k}'\boldsymbol{\beta} = (0 \quad 1 \quad -1 \quad 0 \quad 0) \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix}$$

wobei $\mathbf{k}' = (0 \quad 1 \quad -1 \quad 0 \quad 0)$ ist. Mit

$$\mathbf{b} = \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \\ \hat{\tau}_4 \end{pmatrix} = \begin{pmatrix} 25 \\ 8 \\ -3 \\ 6 \\ 6 \end{pmatrix}$$

finden wir

$$\hat{\lambda} = \hat{\tau}_1 - \hat{\tau}_2 = \mathbf{k}'\mathbf{b} = (0 \quad 1 \quad -1 \quad 0 \quad 0) \begin{pmatrix} 25 \\ 8 \\ -3 \\ 6 \\ 6 \end{pmatrix} = 8 + 3 = 11$$

Sorte 1 ist also um 11 dt/ha besser als Sorte 2. Für diesen Vergleich kann ein Vertrauensintervall berechnet werden. Außerdem kann ein Test der Nullhypothese

$$H_0: \tau_1 - \tau_2 = 0$$

durchgeführt werden. Test und Vertrauensintervall sind hierbei immer äquivalent: Sofern das Intervall für $\mathbf{k}'\boldsymbol{\beta}$ die Null enthält, ist der entsprechende Test nicht signifikant.

Beispiel: In der linearen Regression ist ein Vertrauensintervall bzw. Test des Regressionskoeffizienten von Interesse. Dieser kann in der allgemeinen Form erhalten werden als

$$\lambda = \beta = \mathbf{k}'\boldsymbol{\beta} = (0 \quad 1) \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

wobei $\mathbf{k}' = (0 \quad 1)$. Mit

$$\mathbf{b} = \hat{\boldsymbol{\beta}} = \begin{pmatrix} 15,1355 \\ 0,08117 \end{pmatrix}$$

finden wir

$$\hat{\lambda} = \hat{\beta} = \mathbf{k}'\mathbf{b} = (0 \quad 1) \begin{pmatrix} 15,1355 \\ 0,08117 \end{pmatrix} = 0,08117$$

Desweiteren ist der Erwartungswert an der Stelle x_0 von Interesse. Wollen wir im Beispiel der Regendaten den erwarteten Ertrag bei einer Regenmenge von $x_0 = 200$ mm (April bis Juni) vorhersagen (siehe Abschnitt 6.2), so entspricht dies

$$\lambda = \alpha + \beta x_0 = \mathbf{k}'\boldsymbol{\beta} = \begin{pmatrix} 1 & x_0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 & 200 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

wobei $\mathbf{k}' = (1 \quad x_0) = (1 \quad 200)$ ist. Diese Funktion schätzen wir mit

$$\hat{\lambda} = \hat{\alpha} + \hat{\beta}x_0 = \mathbf{k}'\mathbf{b} = \begin{pmatrix} 1 & 200 \end{pmatrix} \begin{pmatrix} 15,1355 \\ 0,08117 \end{pmatrix} = 1 \cdot 15,1355 + 0,08117 \cdot 200 = 31,37$$

Abschließend sei hier noch einmal betont, dass der Koeffizientenvektor k jeweils vom Benutzer in Abhängigkeit von der Fragestellung zu bestimmen ist.

Um für die obigen Kleinstquadratschätzungen von Linearkombinationen der Parameter Vertrauensintervalle und Tests zu berechnen, benötigen wir zunächst die Varianz der Schätzung. Hierzu müssen wir von der Fehlervarianz der Beobachtungen y ausgehen. Im linearen Modell gilt für die Varianz der Beobachtungen:

$$\text{var}(\mathbf{y}) = \text{var} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \text{var}(\mathbf{e}) = \mathbf{I}\sigma^2 = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & \sigma^2 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & \sigma^2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \sigma^2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \sigma^2 \end{pmatrix}$$

Auf der Diagonalen stehen die Varianzen der Beobachtungen (alle gleich σ^2), jenseits der Diagonale die Kovarianzen (alle gleich Null). In der ersten Zeile und Spalte steht beispielsweise die Varianz der ersten Beobachtung, während in der ersten Zeile und zweiten Spalte die Kovarianz der ersten und der zweiten Beobachtung steht. Dies bedeutet, dass alle Beobachtungen dieselbe Varianz σ^2 haben und dass alle Beobachtungen unkorreliert sind (unabhängig bei Normalverteilung).

Unter diesen Verteilungsannahmen gilt für die Varianz der Kleinst-Quadratschätzung (falls \mathbf{X} von vollem Rang ist, also die Spalten von \mathbf{X} linear unabhängig sind):

$$\text{var}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2$$

Die Varianz σ^2 schätzen wir mittels der Summe der Fehlerquadrate:

$$s^2 = \hat{\sigma}^2 = \frac{SQ_{\text{Fehler}}}{FG_{\text{Fehler}}}$$

wobei

$$SQ_{Fehler} = (y - Xb)'(y - Xb)$$

$$FG_{Fehler} = n - Rang(X)$$

n = Zahl der Beobachtungen

Falls alle Spalten einer Matrix linear unabhängig sind, und auf diesen einfachen Fall wollen wir uns hier beschränken, so ist der Rang der Matrix gleich der Zahl der Spalten. Es gilt für die Kleinst-Quadratschätzung der Linearkombination $k'\beta$:

$$\text{var}(k'\hat{\beta}) = \text{var}(k'b) = k'(X'X)^{-1} k \sigma^2$$

Diese Varianz wird geschätzt durch

$$\hat{\text{var}}(k'\hat{\beta}) = k'(X'X)^{-1} k s^2$$

Die Quadratwurzel aus der Varianz einer Parameterschätzung für $\lambda = k'\beta$ wird als **Standardfehler** bezeichnet:

$$\text{Standardfehler}(k'\hat{\beta}) = \sqrt{k'(X'X)^{-1} k s^2}$$

Nähere Erläuterung: Das obige Resultat über den Standardfehler basiert auf einem allgemeinen Resultat über die Varianz-Kovarianz-Matrix einer Linearkombinationen von Zufallsvariablen (siehe Anhang F). Dieses besagt, dass für

$$z = L'y,$$

wobei y ein Vektor mit Zufallsvariablen und L eine Matrix mit Koeffizienten ist, gilt:

$$\text{var}(z) = L' \text{var}(y) L.$$

Exkurs: Wenn zum Beispiel

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \text{ und } \text{var}(y) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}, \text{ wobei } \sigma_1^2 \text{ und } \sigma_2^2 \text{ die Varianzen von } y_1 \text{ und } y_2$$

sind und σ_{12} deren Kovarianz, so gilt für die Differenz $z = y_1 - y_2 = L'y$ mit $L' = (1 \quad -1)$:

$$\text{var}(z) = (1 \quad -1) \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}.$$

Die Varianz einer Differenz ist also umso kleiner, je größer die Kovarianz der Beobachtungen y_1 und y_2 ist. Diese Tatsache wird bei der Planung eines Versuchs nach einer verbundenen Stichprobe (Statistik-Skript, Abschnitt 3.12) ausgenutzt: Immer wenn es möglich ist, zwei verschiedene Behandlungen bei derselben

Beobachtungseinheit zu messen, führt dies zu einem Gewinn an Genauigkeit im Vergleich zu einem unverbundenen Design, wenn Beobachtungen von derselben Einheit eine positive Kovarianz aufweisen, was normalerweise der Fall ist.

Zurück zum linearen Modell: Im vorliegenden Fall der Schätzung einer Linearkombination der Parameter des linearen Modells wird das allgemeine Resultat wie folgt verwendet. Es gilt:

$$\mathbf{k}'\hat{\boldsymbol{\beta}} = \mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \mathbf{L}'\mathbf{y} \quad \text{mit} \quad \mathbf{L}' = \mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \quad \text{und} \quad \text{var}(\mathbf{y}) = \mathbf{I}\sigma^2 \quad \text{und daher}$$

$$\begin{aligned} \text{var}(\mathbf{k}'\hat{\boldsymbol{\beta}}) &= \text{var}(\mathbf{L}'\mathbf{y}) = \mathbf{L}' \text{var}(\mathbf{y}) \mathbf{L} = \mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \sigma^2 \mathbf{I} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{k} = \\ &= \sigma^2 \mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{k} = \sigma^2 \mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{k} \end{aligned}$$

Beispiel: lineare Regression

Benutze:

$$\begin{pmatrix} c & e \\ e & d \end{pmatrix}^{-1} = \frac{1}{cd - e^2} \begin{pmatrix} d & -e \\ -e & c \end{pmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} = \frac{1}{nSQ_x} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

Regressionskoeffizient:

$$\mathbf{k}' = (0 \quad 1)$$

$$\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{k} = \frac{1}{nSQ_x} (0 \quad 1) \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{nSQ_x} (0 \quad 1) \begin{pmatrix} -\sum_{i=1}^n x_i \\ n \end{pmatrix} = \frac{1}{SQ_x}$$

$$\text{Standardfehler}(\mathbf{k}'\hat{\boldsymbol{\beta}} = \hat{\beta}) = \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{k} s^2} = \frac{s}{\sqrt{SQ_x}}$$

(Genau diese Formel wird bei der Berechnung eines Vertrauensintervalls für den Regressionskoeffizienten in Abschnitt 6.2.2 verwendet!)

Vorhersage bei x_0 :

$$\mathbf{k}' = (1 \ x_0)$$

$$\mathbf{k}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{k} = \frac{1}{nSQ_x} (1 \ x_0) \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} 1 \\ x_0 \end{pmatrix}$$

$$= \frac{1}{nSQ_x} (1 \ x_0) \begin{pmatrix} \sum_{i=1}^n x_i^2 - x_0 \sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i + nx_0 \end{pmatrix}$$

$$= \frac{1}{nSQ_x} \left(\sum_{i=1}^n x_i^2 - x_0 \sum_{i=1}^n x_i - x_0 \sum_{i=1}^n x_i + nx_0^2 \right)$$

$$= \frac{1}{nSQ_x} \left(\sum_{i=1}^n x_i^2 - 2x_0 \sum_{i=1}^n x_i + nx_0^2 \right)$$

$$= \frac{1}{nSQ_x} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 + n\bar{x}^2 + 2nx_0\bar{x} + nx_0^2 \right)$$

$$= \frac{1}{nSQ_x} (SQ_x + n(x_0 - \bar{x})^2)$$

$$= \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}$$

$$\text{Standardfehler}(\mathbf{k}'\hat{\boldsymbol{\beta}} = \hat{\alpha} + \hat{\beta}x_0) = \sqrt{\mathbf{k}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{k}s^2} = s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQ_x}}$$

(Genau diese Formel wird bei der Berechnung eines Vertrauensintervalls für den Erwartungswert an der Stelle x_0 in Abschnitt 6.2.2 verwendet!)

Mit der allgemeinen Formel für den Standardfehler sind wir nun in der Lage, ein Vertrauensintervall und einen Test für $\lambda = \mathbf{k}'\boldsymbol{\beta}$ zu berechnen:

Ein $100(1-\alpha)\%$ Vertrauensintervall für $\lambda = \mathbf{k}'\boldsymbol{\beta}$ ist gegeben durch

$$\lambda_u = \hat{\lambda} - t_{Tab} \sqrt{\mathbf{k}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{k}s^2}$$

$$\lambda_o = \hat{\lambda} + t_{Tab} \sqrt{\mathbf{k}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{k}s^2}$$

wobei t_{Tab} der kritische Wert der t -Verteilung mit $FG_{Fehler} = n - \text{Rang}(\mathbf{X})$ Freiheitsgraden und Irrtumswahrscheinlichkeit α ist [Tab. II],

$$s^2 = SQ_{Fehler} / FG_{Fehler}$$

$$SQ_{Fehler} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

$$FG_{Fehler} = n - \text{Rang}(\mathbf{X})$$

n = Zahl der Beobachtungen in \mathbf{y}

Test der Nullhypothese

$$H_0: \lambda = \mathbf{k}'\boldsymbol{\beta} = 0$$

(1) Berechne

$$t_{Vers} = \frac{|\hat{\lambda}|}{\sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2}}$$

(2) Bestimme t_{Tab} wie bei Vertrauensintervall für vorgegebenes α

(3) Falls

$$t_{Vers} > t_{Tab} \Rightarrow \text{verwerfe } H_0$$

$$t_{Vers} \leq t_{Tab} \Rightarrow \text{behalte } H_0 \text{ bei}$$

Äquivalenz von Vertrauensintervall und t-Test: Anhand dieser beiden allgemeinen Verfahren zur Berechnung eines Vertrauensintervalls sowie eines Tests für die Linearkombination $\lambda = \mathbf{k}'\boldsymbol{\beta}$ soll hier kurz die Äquivalenzbeziehung der beiden Verfahren erläutert werden. Es gilt:

- Enthält das Vertrauensintervall für $\lambda = \mathbf{k}'\boldsymbol{\beta}$ den Wert Null, so ist die Nullhypothese $H_0: \lambda = \mathbf{k}'\boldsymbol{\beta}$ nicht zu verwerfen.
- Enthält das Intervall dagegen nicht die Null, ist H_0 zu verwerfen.

Beweis: Die Nullhypothese wird dann beibehalten, wenn

$$t_{Vers} = \frac{|\hat{\lambda}|}{\sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2}} \leq t_{tab}.$$

Äquivalenzumformung:

$$\begin{aligned} \frac{|\hat{\lambda}|}{\sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2}} \leq t_{tab} &\Leftrightarrow |\hat{\lambda}| \leq t_{tab} \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} \\ &\Leftrightarrow \left(\hat{\lambda} \leq t_{tab} \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} \quad \text{und} \quad \hat{\lambda} \geq -t_{tab} \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} \right) \\ &\Leftrightarrow \left(\hat{\lambda} - t_{tab} \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} \leq 0 \quad \text{und} \quad \hat{\lambda} + t_{tab} \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} \geq 0 \right) \end{aligned}$$

Die beiden resultierenden Ungleichungen sind Aussagen über die Vertrauensgrenzen! Die erste Ungleichung sagt, dass die untere Grenze des Vertrauensintervalls kleiner gleich Null ist, die zweite sagt, dass die obere Grenze größer gleich Null ist. Es folgt, dass das Intervall die Null enthält.

Beispiel: Regendaten, lineare Regression, Prognose bei $x_0 = 200$:

Vertrauensintervall:

$$\hat{\lambda} = \hat{\alpha} + \hat{\beta}x_0 = \mathbf{k}'\mathbf{b} = \begin{pmatrix} 1 & 200 \end{pmatrix} \begin{pmatrix} 15,1355 \\ 0,08117 \end{pmatrix} = 1 \cdot 15,1355 + 0,08117 \cdot 200 = 31,37$$

$$SQ_{Fehler} = 486,12666$$

$$\text{Rang}(\mathbf{X}) = 2 \text{ (zwei Spalten in } \mathbf{X} \text{)}$$

$$n = 26$$

$$FG_{Fehler} = 24$$

$$s^2 = SQ_{Fehler} / FG_{Fehler} = 486,13 / 24 = 20,25$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0,440178 & -0,002883 \\ -0,002883 & 0,0000207 \end{pmatrix}$$

$$\mathbf{k}' = \begin{pmatrix} 1 & 200 \end{pmatrix}$$

$$\sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} = \sqrt{\begin{pmatrix} 1 & 200 \end{pmatrix} \begin{pmatrix} 0,440178 & -0,002883 \\ -0,002883 & 0,0000207 \end{pmatrix} \begin{pmatrix} 1 \\ 200 \end{pmatrix} 20,25} = 1,523$$

$$t_{Tab} = 2,064$$

$$\lambda_u = \hat{\lambda} - t_{Tab} \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} = 31,37 - 2,064 \cdot 1,523 = 28,23$$

$$\lambda_o = \hat{\lambda} + t_{Tab} \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} = 31,37 + 2,064 \cdot 1,523 = 34,51$$

Bis auf Rundungsfehler sind diese Werte identisch mit den in Abschnitt 6.2 berechneten Vertrauensgrenzen (siehe Anhang Kap. 6!).

Beispiel: Lineare Regression, Regressionskoeffizient.

Vertrauensintervall:

$$\hat{\lambda} = 0,0812$$

$$SQ_{Fehler} = 486,12666$$

$$\text{Rang}(\mathbf{X}) = 2 \text{ (zwei Spalten in } \mathbf{X} \text{)}$$

$$n = 26$$

$$FG_{Fehler} = 24$$

$$s^2 = SQ_{Fehler} / FG_{Fehler} = 486,13 / 24 = 20,25$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0,440178 & -0,002883 \\ -0,002883 & 0,0000207 \end{pmatrix}$$

$$\mathbf{k}' = \begin{pmatrix} 0 & 1 \end{pmatrix}$$

$$\sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} = \sqrt{\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 0,440178 & -0,002883 \\ -0,002883 & 0,0000207 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} 20,25} = 0,02047$$

$$t_{Tab} = 2,064$$

$$\lambda_u = \hat{\lambda} - t_{Tab} \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} = 0,0812 - 2,064 \cdot 0,02047 = 0,0389$$

$$\lambda_o = \hat{\lambda} + t_{Tab} \sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}s^2} = 0,0812 + 2,064 \cdot 0,02047 = 0,1234$$

Bis auf Rundungsfehler sind diese Werte identisch mit den in Abschnitt 6.2 berechneten Vertrauensgrenzen (siehe Anhang Kap. 6!).

Test:

$$t_{Vers} = \frac{|\hat{\lambda}|}{\sqrt{\mathbf{k}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}_S^2}} = \frac{0,0812}{0,02047} = 3,97 > t_{Tab} = 2,064$$

Der Regressionskoeffizient ist signifikant von Null verschieden ($H_0: \beta = 0$ wird verworfen). Dies Ergebnis deckt sich mit dem Ergebnis in Abschnitt 6.2 (t-Test für den Regressionskoeffizienten). Und das Ergebnis des Tests deckt sich damit, dass das Vertrauensintervall für λ nicht die Null enthält.

Exkurs: Verwendung eines Statistik-Paketes. Um die Regression mittels eines Statistik-Paketes durchzuführen, muss einer Routine für lineare Modelle vor allem das lineare Modell mitgeteilt werden. Hier sind exemplarisch die Anweisungen für die REG Prozedur des Statistik-Paketes SAS (Statistical Analysis System) wiedergegeben:

```
data t;
input jahr x y;
datalines;
1911      71      16.6
1915      89      16.4
1901      96      25.0
1917      98      19.2
1903     105      26.6
1918     106      20.2
1913     108      19.4
1909     108      22.6
1925     110      30.2
1906     111      19.6
1912     119      29.8
1919     123      25.6
1914     132      30.6
1905     135      20.4
1910     137      24.2
1902     144      32.6
1916     147      30.4
1924     147      30.4
1920     156      31.0
1907     161      33.8
1922     162      31.6
1900     177      26.2
1921     191      35.8
1906     209      29.2
1923     235      33.6
1908     246      26.6
;
proc reg;
model y=x;
run;
```


Output:

The REG Procedure
Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	318.45795	318.45795	15.72	0.0006
Error	24	486.12666	20.25528		
Corrected Total	25	804.58462			

Root MSE	4.50059	R-Square	0.3958
Dependent Mean	26.44615	Adj R-Sq	0.3706
Coeff Var	17.01792		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	1	15.13554	2.98596	5.07	<.0001	8.97283	21.29825
x	1	0.08117	0.02047	3.97	0.0006	0.03892	0.12342

Die Kleinstquadratschätzung für β findet sich unter "x" in der letzten Zeile des Output. Hier wird auch der t_{Vers} -Wert von 3,97 ausgegeben, versehen mit einem p-Wert von 0,0006. Letzterer zeigt Signifikanz an, da er kleiner als $\alpha = 5\%$ ist (siehe Abschnitt 3.15 sowie Anhang D). Das Vertrauensintervall wird in derselben Zeile angegeben. Im oberen Teil des Output findet sich die Varianzanalyse-Tabelle.

Um ein Vertrauensintervall für die Vorhersage an der Stelle $x_0 = 200$ zu erhalten, verwenden wir die MIXED Prozedur wie folgt:

```
proc mixed;
model y=x;
estimate 'Ertrag bei 200 mm NS' intercept 1 x 200/cl;
run;
```

Output:

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
Ertrag bei 200 mm NS	31.3694	1.5234	24	20.59	<.0001	0.05	28.2253	34.5135

In der ESTIMATE Anweisung müssen die Koeffizienten der Vektors $k' = (1 \ 200)$ angegeben werden. "Intercept" steht für den Achsenabschnitt (α), die Steigung (β) wird mit der Bezeichnung der Einflussvariable (X) angesprochen. Ohne weiter auf Details der SAS Anweisungen einzugehen, wird an diesem Beispiel deutlich, dass man sich in der Matrix-Notation für das allgemeine lineare Modell auskennen muss, um die Prozeduren für lineare Modelle bei etwas weitergehenden Auswertungen (hier: Vertrauensintervall für eine Vorhersage) adäquat bedienen zu können. Die

Prozeduren verwenden allgemeine Resultate für das lineare Modell, wie sie in diesem und dem nächsten Abschnitt exemplarisch vorgestellt werden.

Beispiel: Varianzanalyse, Vergleich der Sorten 1 und 2:

$$\hat{\lambda} = 11$$

$$SQ_{Fehler} = 116,0$$

$$\text{Rang}(X) = 5 \text{ (fünf Spalten in } X)$$

$$n = t * r = 20 = \text{Gesamtzahl der Beobachtungen}$$

$$FG_{Fehler} = n - \text{Rang}(X) = 15$$

$$s^2 = SQ_{Fehler} / FG_{Fehler} = 116,0 / 15 = 7,733$$

$$(X'X)^{-1} = 4^{-1} \begin{pmatrix} 1 & -1 & -1 & -1 & -1 \\ -1 & 2 & 1 & 1 & 1 \\ -1 & 1 & 2 & 1 & 1 \\ -1 & 1 & 1 & 2 & 1 \\ -1 & 1 & 1 & 1 & 2 \end{pmatrix} \quad (\text{vgl. S. 79})$$

$$k' = (0 \ 1 \ -1 \ 0 \ 0)$$

$$\sqrt{k'(X'X)^{-1}ks^2} = \sqrt{(0 \ 1 \ -1 \ 0 \ 0)4^{-1} \begin{pmatrix} 1 & -1 & -1 & -1 & -1 \\ -1 & 2 & 1 & 1 & 1 \\ -1 & 1 & 2 & 1 & 1 \\ -1 & 1 & 1 & 2 & 1 \\ -1 & 1 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}} 7,733 = 1,966$$

$$t_{Tab} = 2,131$$

$$\lambda_u = \hat{\lambda} - t_{Tab} \sqrt{k'(X'X)^{-1}ks^2} = 11 - 2,131 * 1,966 = 6,81$$

$$\lambda_o = \hat{\lambda} + t_{Tab} \sqrt{k'(X'X)^{-1}ks^2} = 11 + 2,131 * 1,966 = 15,19$$

Test:

$$t_{Vers} = \frac{|\hat{\lambda}|}{\sqrt{k'(X'X)^{-1}ks^2}} = \frac{11}{1,966} = 5,59 > t_{Tab} = 2,131$$

Die Sorten 1 und 2 sind signifikant verschieden.

SAS Anweisungen

```
data;
input trt$ y;
cards;
a 31
a 32
a 37
a 32
b 21
b 23
b 25
b 19
c 27
```

```

c 29
c 34
c 34
d 34
d 32
d 31
d 27
e 24
e 23
e 27
e 26
;
proc mixed;
class trt;
model y=trt;
lsmeans trt/cl pdiff;
run;

```

Output:

Differences of Least Squares Means

Effect	trt	_trt	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
trt	a	b	11.0000	1.9664	15	5.59	<.0001	0.05	6.8088	15.1912
trt	a	c	2.0000	1.9664	15	1.02	0.3252	0.05	-2.1912	6.1912
trt	a	d	2.0000	1.9664	15	1.02	0.3252	0.05	-2.1912	6.1912
trt	a	e	8.0000	1.9664	15	4.07	0.0010	0.05	3.8088	12.1912
trt	b	c	-9.0000	1.9664	15	-4.58	0.0004	0.05	-13.1912	-4.8088
trt	b	d	-9.0000	1.9664	15	-4.58	0.0004	0.05	-13.1912	-4.8088
trt	b	e	-3.0000	1.9664	15	-1.53	0.1479	0.05	-7.1912	1.1912
trt	c	d	0	1.9664	15	0.00	1.0000	0.05	-4.1912	4.1912
trt	c	e	6.0000	1.9664	15	3.05	0.0081	0.05	1.8088	10.1912
trt	d	e	6.0000	1.9664	15	3.05	0.0081	0.05	1.8088	10.1912

6.9 Vergleich von geschachtelten Modellen mittels F-Test

Beispiel: In der Varianzanalyse verwenden wir das Modell

$$y_{ij} = \mu + \tau_i + e_{ij}$$

Unter der Nullhypothese

$$H_0: \tau_1 = \tau_2 = \tau_3 = \dots \quad (\text{keine Behandlungsunterschiede})$$

hat jede Sorte denselben Erwartungswert, so dass das **reduzierte Modell**

$$y_{ij} = \mu + e_{ij}$$

gilt. Demgegenüber wird das Modell unter der Alternativhypothese als **volles Modell** bezeichnet. Das reduzierte Modell ist ein Spezialfall des vollen Modells. Das reduzierte Modell ist innerhalb des vollen Modells "geschachtelt". Der F-Test der Varianzanalyse prüft, ob das reduzierte Modell zur Beschreibung der Daten ausreicht, oder ob das volle Modell vorzuziehen ist.

Beispiel: Bei der linearen Regression verwenden wir das Modell

$$y_i = \alpha + \beta x_i + e_i$$

Unter der Nullhypothese

$$H_0: \beta = 0 \quad (\text{kein Zusammenhang zwischen } x \text{ und } y)$$

gilt das reduzierte Modell

$$y_i = \alpha + e_i$$

Der F-Test der Regressionsanalyse prüft, ob das reduzierte Modell zur Beschreibung der Daten ausreicht, oder ob das volle Modell vorzuziehen ist.

Generell kann ein Vergleich eines reduzierten mit einem vollen Modell mittels eines F-Tests durchgeführt werden. Für den Vergleich werden jeweils das volle und das reduzierte Modell mit der Methode der kleinsten Quadrate angepasst und das SQ_{Fehler} bestimmt. Generell ist das SQ_{Fehler} für das volle Modell kleiner als (oder gleich dem) für das reduzierte. Ob die Reduktion des SQ_{Fehler} gegenüber dem reduzierten Modell signifikant ist, kann mittels F-Test wie folgt geprüft werden:

F-Test zum Vergleich eines reduzierten mit einem vollen linearen Modell

H_0 : reduziertes Modell gilt

H_A : volles Modell gilt, aber reduziertes nicht

(1) Berechne SQ_{Fehler} für das volle und das reduzierte Modell

(2) Bestimme Fehlerfreiheitsgrade (FG_{Fehler}) für das volle und das reduzierte Modell

(3) Berechne

$$F_{Vers} = \frac{(SQ_{Fehler}^{red} - SQ_{Fehler}^{voll}) / (FG_{Fehler}^{red} - FG_{Fehler}^{voll})}{SQ_{Fehler}^{voll} / FG_{Fehler}^{voll}}$$

(4) Bestimme $F_{Tab} = F(1-\alpha, FG_1, FG_2)$ (Tab. VI)

$$FG_1 = FG_{Fehler}^{red} - FG_{Fehler}^{voll}; FG_2 = FG_{Fehler}^{voll}$$

(5) Falls $F_{Vers} > F_{Tab}$, verwirfe H_0 ,
falls $F_{Vers} \leq F_{Tab}$, behalte H_0 bei

Beispiel: Für die Regendaten passen wir reduziertes und volles Modell an und finden folgende SQ_{Fehler} :

Modell	Bezeichnung	FG_{Fehler}	SQ_{Fehler}
(1) $y_i = \alpha + e_i$	reduziert	$25 = (n - 1)$	804,58462
(2) $y_i = \alpha + \beta x_i + e_i$	voll	$24 = (n - 2)$	486,12666

Die Fehler-SQ und -FG des vollen Modells hatten wir bereits in Zusammenhang mit der linearen Regression in Abschnitt 6.2 berechnet. Die Fehler-FG entsprechen der Zahl der Beobachtungen (n), abzüglich der Zahl der Parameter (Spalten von X !). Das reduzierte Modell impliziert, dass die Daten einer einfachen Stichprobe vom Umfang n aus einer Normalverteilung mit Mittelwert α und Varianz $\text{var}(e_i) = \sigma^2$ entstammen. Das SQ_{Fehler} ist daher einfach zu berechnen als

$$SQ_{Fehler}^{red} = \sum_{i=1}^n (y_i - \bar{y}.)^2$$

Man beachte, dass

$$\alpha = \bar{y}.$$

die Kleinstquadratschätzung von α unter dem reduzierten Modell ist.

Die zugehörigen Freiheitsgrade sind bekanntermaßen bei einer einfachen Stichprobe ($n - 1$), was ebenfalls der Zahl der Beobachtungen, abzüglich der Zahl der Parameter entspricht. Dieser Wert liegt um Eins höher als die Fehlerfreiheitsgrade des vollen Modells ($n - 2$). Der zusätzliche Parameter im vollen Modell (β) hat einen Freiheitsgrad verbraucht.

$$F_{Vers} = \frac{(SQ_{Fehler}^{red} - SQ_{Fehler}^{voll}) / (FG_{Fehler}^{red} - FG_{Fehler}^{voll})}{SQ_{Fehler}^{voll} / FG_{Fehler}^{voll}} = \frac{(804,58462 - 486,12666) / (25 - 24)}{486,12666 / 24} = 15,72$$

$$FG_1 = FG_{Fehler}^{red} - FG_{Fehler}^{voll} = 1; FG_2 = FG_{Fehler}^{voll} = 24;$$

$$F_{Tab} = F(1-\alpha, FG_1, FG_2) = F(0,95, 1, 24) = 4,26 < F_{Vers}$$

\Rightarrow Die Steigung β ist signifikant von Null verschieden.

Man beachte, dass der F-Test identisch ist mit dem in Abschnitt 6.2.1 vorgestellten und angewendeten F-Test für $H_0: \beta = 0$. Die Quadratsummen sind leicht verschieden aufgrund von Rundungsfehlern bei der Berechnung der SQ in 6.2.1. Hier wurde dagegen mit dem PC gerechnet, was die leichten numerischen Unterschiede erklärt.

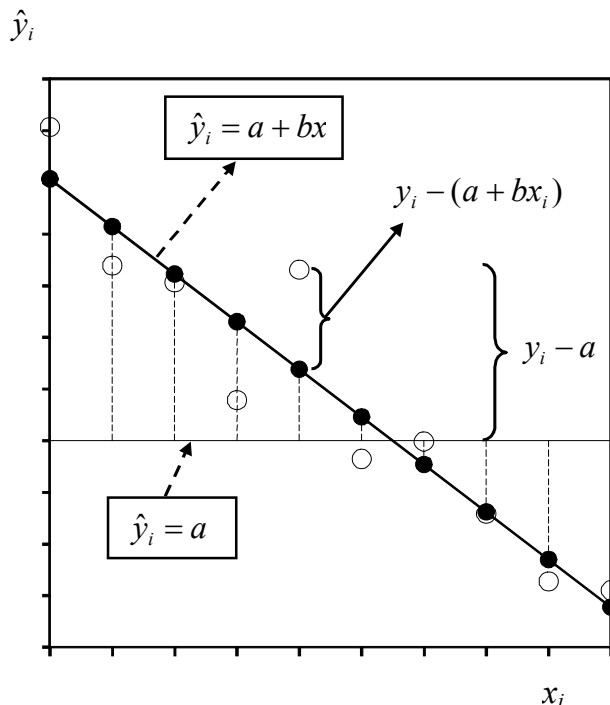


Abb. 6.9.1: Vergleich von Kleinstquadratschätzung des vollen ($a + bx_i$) und reduzierten (a) Modells bei der linearen Regression (Kreise = Daten).

Am Beispiel der linearen Regression soll auch verdeutlicht werden, warum das SQ_{Fehler} für das volle Modell kleiner sein muss als für das reduzierte. Abb. 6.9.1 zeigt exemplarisch, dass das volle Modell ($a + bx_i$) wegen seiner größeren Flexibilität "näher" an die Daten heranreicht als das reduzierte (Horizontale durch $a = \bar{y}$). Beim vollen Modell ist die Gerade geneigt, wobei die Neigung so ist, dass das SQ_{Fehler} minimiert wird. Beim reduzierten Modell besteht dagegen die Restriktion, dass die Gerade horizontal verlaufen muss, weshalb sie nicht so "nah" an die Daten heranreicht bzw. nicht so gut an die Daten passt. Eine bessere Anpassung des vollen Modells ist übrigens auch dann gegeben, wenn in Wirklichkeit kein Zusammenhang zwischen x und y besteht. In diesem Fall spiegelt die von Null abweichende geschätzte Steigung b nur Zufallsschwankungen wieder. Der F-Test prüft, ob die immer zu beobachtende Reduktion von SQ_{Fehler} im Rahmen der Zufallsschwankung der Erhebung/des Versuchs liegt, oder ob die Reduktion so substantiell ist, dass von der Ungültigkeit des reduzierten Modells ausgegangen werden muss.

Beispiel: Beim vollständig randomisierten Sortenversuch mit $t = 5$ Sorten und $r = 4$ Wiederholungen je Sorte finden wir:

Modell	Bezeichnung	FG_{Fehler}	SQ_{Fehler}
(1) $y_{ij} = \mu + e_{ij}$	reduziert	$19 = (rt - 1)$	464,8
(2) $y_{ij} = \mu + \tau_i + e_{ij}$	voll	$15 = t(r - 1)$	116,0

Man beachte, dass auch hier (wie bei der Regression) das reduzierte Modell einer Normalverteilung mit konstanten Mittelwert (μ) und Varianz σ^2 entspricht. Demzufolge ist

$$SQ_{Fehler}^{red} = \sum_{i=1}^t \sum_{j=1}^r (y_{ij} - \bar{y}_{..})^2$$

Die FG des reduzierten Modells entsprechen der Zahl der Beobachtungen ($n = rt$) minus Eins (für den Parameter μ). Die FG für das volle Modell können auf zwei Wegen erklärt werden:

1) Wir können für jede Behandlung ein Fehler-SQ berechnen. Dieses hat $(r-1)$ FG. Führen wir die Fehler-SQ der Behandlungen zusammen ("poolen"), so haben wir t mal $(r-1)$, also $t(r-1)$ Fehler-FG.

2) Das volle Modell hat $t+1$ Parameter. Allerdings muss wegen der Überparametrisierung eine Restriktion eingeführt werden, so dass wir nur t freie Parameter haben. Anders betrachtet: *De facto* liegt alle Information in den Mittelwerten, und davon gibt es t Stück. Die Zahl der Beobachtungen (rt) minus Zahl der freien Parameter (t) ergibt $rt-t = t(r-1)$.

Wir finden

$$F_{Vers} = \frac{(SQ_{Fehler}^{red} - SQ_{Fehler}^{voll}) / (FG_{Fehler}^{red} - FG_{Fehler}^{voll})}{SQ_{Fehler}^{voll} / FG_{Fehler}^{voll}} = \frac{(464,8 - 116,0) / (19 - 15)}{116,0 / 15} = 11,28$$

$$FG_1 = FG_{Fehler}^{red} - FG_{Fehler}^{voll} = t - 1 = 4; \quad FG_2 = FG_{Fehler}^{voll} = t(r - 1) = 15;$$

$$F_{Tab} = F(1 - \alpha, FG_1, FG_2) = F(0,95, 4, 15) = 3,06 < F_{Vers}$$

Dieser Test ist identisch mit dem in Abschnitt 4 besprochenen F-Test der einfachen Varianzanalyse.

Die in den Abschnitten 6.8 und 6.9 angesprochen Techniken können direkt in den nächsten beiden Abschnitten verwendet werden.

6.10 Multiple lineare Regression

Beispiel: In einem Versuch mit Ratten wurde das Endgewicht (y_i ; in g) von 35 Tieren in Abhängigkeit vom Anfangsgewicht (x_{1i} ; in g) und vom Futterverzehr (x_{2i} ; in g) untersucht. Die Daten waren wie folgt (Linder und Berchtold II, S. 125):

x_1 (g)	x_2 (g)	y (g)
55,8	289	114,8
45,8	316	109,7
48,1	304	111,3
43,3	299	126,0
50,1	353	144,7
40,1	298	121,0
47,1	303	125,2
51,0	312	113,7
53,7	333	138,5
41,2	280	105,8
40,2	287	117,7
46,4	338	140,0
45,9	298	117,1
38,0	302	103,0
56,0	355	137,3
32,4	307	109,7
37,5	342	136,3
45,9	310	121,2
40,7	280	104,5
36,4	283	104,0
46,9	305	120,2
42,2	296	120,5
43,4	290	118,9
45,0	224	126,4
43,8	353	125,4
47,8	282	109,4
50,4	288	121,2
37,9	266	109,2
46,0	318	141,1
42,8	335	131,3
50,7	304	128,5
59,6	296	138,8
43,8	292	109,0
65,4	320	113,7
39,3	305	116,2

Um den Einfluss von Anfangsgewicht und Futteraufnahme auf das Endgewicht zu untersuchen, können zunächst x - y -Plots der Daten herangezogen werden (Abb. 6.10.1 und Abb. 6.10.2).

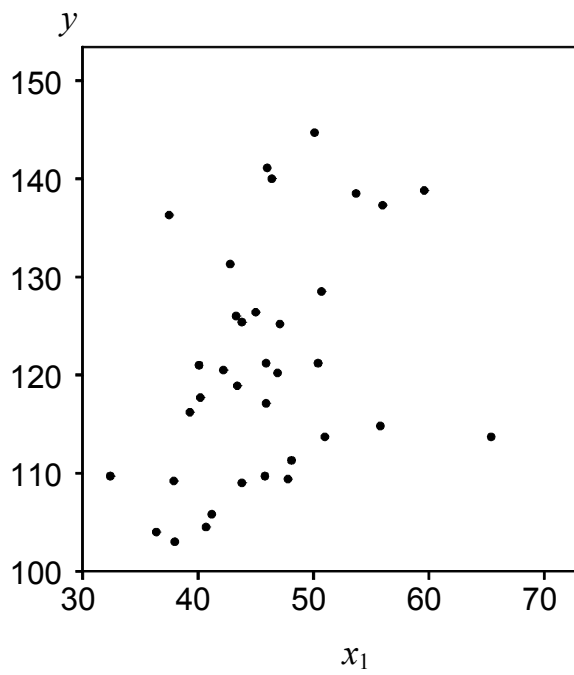


Abb. 6.10.1: Plot des Endgewichtes (y ; in g) gegen das Anfangsgewicht (x_1 ; in g).

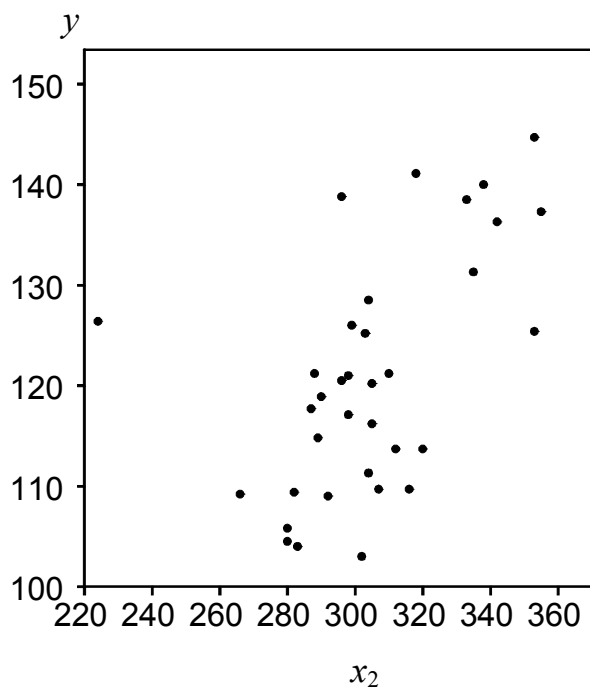


Abb. 6.10.2: Plot des Endgewichtes (y ; in g) gegen die Futteraufnahme (x_2 ; in g).

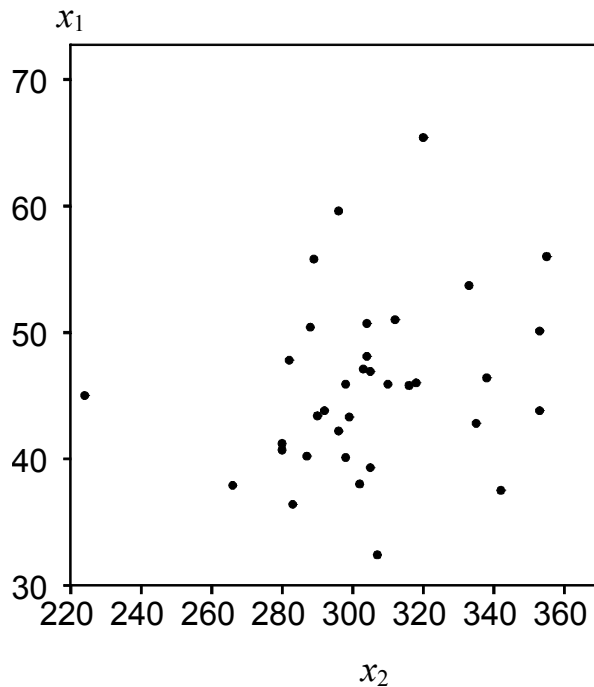


Abb. 6.10.3: Plot des Anfangsgewichtes (x_1 ; in g) gegen Futteraufnahme (x_2 ; in g).

Die Plots von y gegen x_1 und x_2 deuten an, dass jeweils ein gewisser Zusammenhang besteht, der durch eine lineare Regression zumindest näherungsweise modelliert werden kann. Da hier zwei Einflussvariablen zu modellieren sind, kann man ein multiples Regressionsmodell der Form

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

zugrundelegen. In Matrizenform lautet das Modell:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \quad \text{bzw.} \quad \begin{pmatrix} 114,8 \\ 109,7 \\ \vdots \\ 116,2 \end{pmatrix} = \begin{pmatrix} 1 & 55,8 & 289 \\ 1 & 45,8 & 316 \\ \vdots & \vdots & \vdots \\ 1 & 39,3 & 305 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Wir finden:

$$X'X = \begin{pmatrix} 35 & 1600,6 & 10663 \\ 1600,6 & 74797,86 & 489081,2 \\ 10663 & 489081,2 & 3272317 \end{pmatrix}$$

$$X'y = \begin{pmatrix} 4231,3 \\ 194519,87 \\ 1294893,4 \end{pmatrix}$$

$$\mathbf{b} = (\mathbf{X}\mathbf{X})^{-1} \mathbf{X}\mathbf{y} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 34,625696 \\ 0,4387352 \\ 0,2173085 \end{pmatrix}$$

[Inverse $(\mathbf{X}\mathbf{X})^{-1}$ hier mit dem PC berechnet.] Das geschätzte Modell lautet also:

$$\hat{y} = 34,63 + 0,4387x_1 + 0,2173x_2$$

Die Regressionskoeffizienten haben folgende Interpretation:

β_1 : Steigt das Anfangsgewicht um ein Gramm, steigt das Endgewicht um 0,4387 Gramm (bei konstant gehaltener Futteraufnahme).

β_2 : Steigt die Futteraufnahme um ein Gramm, steigt das Endgewicht um 0,2173 Gramm (bei konstant gehaltenem Anfangsgewicht).

Um den Regressionskoeffizienten β_1 auf Signifikanz zu prüfen, passen wir folgende Modellsequenz an:

Modell	Bezeichnung	FG_{Fehler}	SQ_{Fehler}
(1) $y_i = \alpha + \beta_2 x_{2i} + e_i$	reduziert	$33 = n - 2$	3370,98
(2) $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$	voll	$32 = n - 3$	3079,94

Die Zahl der Fehler-FG für ein Modell entspricht auch hier der Zahl der Beobachtungen, abzüglich der Zahl der Parameter. Das volle Modell unterscheidet sich von dem reduzierten Modell durch die Hinzunahme des Regressionsterms für das Anfangsgewicht ($\beta_1 x_{1i}$). Unter der Nullhypothese

$$H_0: \beta_1 = 0$$

kann dieser wegfallen und somit das reduzierte Modell verwendet werden. Wir finden:

$$F_{Vers} = \frac{(SQ_{Fehler}^{red} - SQ_{Fehler}^{voll}) / (FG_{Fehler}^{red} - FG_{Fehler}^{voll})}{SQ_{Fehler}^{voll} / FG_{Fehler}^{voll}} = \frac{(3370,98 - 3079,94) / (33 - 32)}{3079,94 / 32} = 3,02$$

$$FG_1 = FG_{Fehler}^{red} - FG_{Fehler}^{voll} = 1; \quad FG_2 = FG_{Fehler}^{voll} = 32;$$

$$F_{Tab} = F(1 - \alpha, FG_1, FG_2) = F(0,95, 1, 32) = 4,17 > F_{Vers}$$

Der Einfluss des Anfangsgewichts ist nicht signifikant. Um den Regressionskoeffizienten β_2 auf Signifikanz zu prüfen, passen wir folgende Modellsequenz an:

Modell	Bezeichnung	FG_{Fehler}	SQ_{Fehler}
(1) $y_i = \alpha + \beta_1 x_{1i} + e_i$	reduziert	$33 = n - 2$	4140,07
(2) $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$	voll	$32 = n - 3$	3079,94

Das volle unterscheidet sich von dem reduzierten Modell durch die Hinzunahme des Regressionsterms für die Futteraufnahme ($\beta_2 x_{2i}$). Unter der Nullhypothese

$$H_0: \beta_2 = 0$$

kann dieser wegfallen und somit das reduzierte Modell verwendet werden. Wir finden:

$$F_{Vers} = \frac{(SQ_{Fehler}^{red} - SQ_{Fehler}^{voll}) / (FG_{Fehler}^{red} - FG_{Fehler}^{voll})}{SQ_{Fehler}^{voll} / FG_{Fehler}^{voll}} = \frac{(4140,07 - 3079,94) / (33 - 32)}{3079,94 / 32} = 11,01$$

$$FG_1 = FG_{Fehler}^{red} - FG_{Fehler}^{voll} = 1; FG_2 = FG_{Fehler}^{voll} = 32;$$

$$F_{Tab} = F(1 - \alpha, FG_1, FG_2) = F(0,95, 1, 32) = 4,17 < F_{Vers}$$

Die Nullhypothese ($H_0: \beta_2 = 0$) wird verworfen, und wir schließen, dass die Futteraufnahme einen signifikanten Einfluss auf das Endgewicht hat.

Da nur x_2 signifikant war, wählen wir abschließend das Modell

$$y_i = \alpha + \beta_2 x_{2i} + e_i$$

Dieses hat die Kleinst-Quadrat-Schätzung

$$\hat{y} = 46,55 + 0,2440 x_2$$

6.10.1 Sequentieller Modellaufbau - Varianzanalyse

Wenn man ein lineares Modell mit einem Statistik-Paket wie z.B. SAS anpasst, wird immer automatisch eine Varianzanalyse-Tabelle erstellt, in welcher für jeden Term im Modell eine Streuungsursache mit Freiheitsgraden und Quadratsummen ausgewiesen wird. Dies soll anhand der multiplen Regression für das Ratten-Beispiel erläutert werden. Insgesamt können für den Ratten-Datensatz vier verschiedene Modelle betrachtet werden, je nachdem ob keine (Gleichung 1 unten), nur eine (Gl. 2 und 2') oder beide Einflussvariablen (Gl. 3) in das Modell aufgenommen werden:

$$(1) y_i = \alpha + e_i$$

$$(2) y_i = \alpha + \beta_1 x_{1i} + e_i$$

$$(2') y_i = \alpha + \beta_2 x_{2i} + e_i$$

$$(3) y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

Jedes dieser Modelle kann mit der Methode der kleinsten Quadrate angepasst werden, so dass für jedes Modell ein SQ_{Fehler} erhalten wird.

Eine Varianzanalyse-Tabelle kann nun mit einer Sequenz von Modellen aufgebaut werden, in der sukzessive immer ein weiterer Modellterm aufgenommen wird. Für das Ratten-Beispiel gibt es zwei Sequenzen:

1. Sequenz (erst kommt x_1 in das Modell, dann x_2):

Modell	FG_{Fehler}	SQ_{Fehler}
(1) $y_i = \alpha + e_i$	$FG^{(1)} = n - 1$	$SQ_{\text{Fehler}}^{(1)}$
(2) $y_i = \alpha + \beta_1 x_{i1} + e_i$	$FG^{(2)} = n - 2$	$SQ_{\text{Fehler}}^{(2)}$
(3) $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$	$FG^{(3)} = n - 3$	$SQ_{\text{Fehler}}^{(3)}$

2. Sequenz (erst kommt x_2 in das Modell, dann x_1):

Modell	FG_{Fehler}	SQ_{Fehler}
(1) $y_i = \alpha + e_i$	$FG^{(1)} = n - 1$	$SQ_{\text{Fehler}}^{(1)}$
(2') $y_i = \alpha + \beta_2 x_{i2} + e_i$	$FG^{(2')} = n - 2$	$SQ_{\text{Fehler}}^{(2')}$
(3) $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$	$FG^{(3)} = n - 3$	$SQ_{\text{Fehler}}^{(3)}$

Jede Sequenz liefert auch eine Sequenz von geschachtelten Modellen, mit denen man den F-Test aus Abschnitt 6.9 durchführen kann. So liefert die 1. Sequenz z.B. das geschachtelte Modell-Paar (Modelle 1 und 2)

Reduziertes Modell (1): $y_i = \alpha + e_i$
 Volles Modell (2): $y_i = \alpha + \beta_1 x_{i1} + e_i$

Hiermit könnte $H_0: \beta_1 = 0$ geprüft werden (Dieser Test wird aber so nicht von Statistik-Paketen ausgewiesen; siehe unten). Außerdem liefert die 1. Sequenz das geschachtelte Modell-Paar (Modelle 2 und 3)

Reduziertes Modell (2): $y_i = \alpha + \beta_1 x_{i1} + e_i$
 Volles Modell (3): $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$

Hiermit kann $H_0: \beta_2 = 0$ geprüft werden.

Somit liefert die 1. Sequenz für beide Regressionsterme einen Test. Allerdings ignoriert das erste Modell-Paar (Modelle 1 und 2), welches einen Test für x_1 liefert, den Einfluß der zweiten Einflussvariable (x_2). Dies ist dann ein Problem, wenn die beiden Einflussvariablen x_1 und x_2 untereinander korreliert sind, besonders dann, wenn die Korrelation sehr hoch ist (siehe Abschnitt 6.10.4). Denn häufig möchte man wissen, welchen Erklärungswert die Variable x_1 zusätzlich hat, wenn die Variable x_2 bereits im Modell ist. Besser ist es dann, für den Test von x_1 die zweite Modell-Sequenz zu betrachten, die das Modell-Paar (Modelle 2' und 3)

Reduziertes Modell (2'): $y_i = \alpha + \beta_2 x_{i2} + e_i$
 Volles Modell (3): $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$

liefert. Hiermit erhalten wir dann einen Test für x_1 , der um den Einfluß von (x_2) bereinigt ist.

Für beide Sequenzen kann nun jeweils eine Varianzanalyse-Tabelle wie folgt berechnet werden:

1. Sequenz:

Ursache	(Erläuterung)	FG	SQ
β_1	$(x_1, \text{ ignoriere } x_2)$	$FG^{(1)} - FG^{(2)} = 1$	$SQ_{Fehler}^{(1)} - SQ_{Fehler}^{(2)}$
$\beta_2 \beta_1$	$(x_2, \text{ bereinigt um } x_1)$	$FG^{(2)} - FG^{(3)} = 1$	$SQ_{Fehler}^{(2)} - SQ_{Fehler}^{(3)}$
Fehler		$FG^{(3)}$	$SQ_{Fehler}^{(3)}$

Die Ausgabe des Statistik-Paketes SAS für diese Varanzanalyse ist wie folgt:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x1	1	645.710370	645.710370	6.71	0.0143
x2	1	1060.128439	1060.128439	11.01	0.0023
Error	32	3079.940048	96.248127		

(Source = Ursache, Error = Fehler, DF = Freiheitsgrade, Type I SS = SQ, Mean Square = MQ). Diese Sequenz liefert einen F-Test für x_2 , der um den Einfluss von x_1 bereinigt ist. Die F-Statistik ist

$$F_{Vers} = \frac{(SQ_{Fehler}^{(2)} - SQ_{Fehler}^{(3)})/1}{SQ_{Fehler}^{(3)} / FG_{Fehler}^{(3)}}$$

Dieser Test entspricht dem in 6.9 beschriebenen Vorgehen. Darüber hinaus weist ein Statistik-Paket noch diesen Test für x_1 aus, der aber nicht um den Einfluß von x_2 bereinigt ist (und daher nicht zu bevorzugen ist):

$$F_{Vers} = \frac{(SQ_{Fehler}^{(1)} - SQ_{Fehler}^{(2)})/1}{SQ_{Fehler}^{(3)} / FG_{Fehler}^{(3)}}$$

Möchte man jedoch wissen, welchen Erklärungswert x_1 hat, wenn bereits um x_2 bereinigt wurde, ist die 2. Sequenz zu bevorzugen.

2. Sequenz:

Ursache	(Erläuterung)	FG	SQ
β_2	$(x_2, \text{ ignoriere } x_1)$	$FG^{(1)} - FG^{(2)} = 1$	$SQ_{Fehler}^{(1)} - SQ_{Fehler}^{(2)}$
$\beta_1 \beta_2$	$(x_1, \text{ bereinigt um } x_2)$	$FG^{(2)} - FG^{(3)} = 1$	$SQ_{Fehler}^{(2)} - SQ_{Fehler}^{(3)}$
Fehler		$FG^{(3)}$	$SQ_{Fehler}^{(3)}$

SAS-Ausgabe:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x2	1	1414.794674	1414.794674	14.70	0.0006
x1	1	291.044135	291.044135	3.02	0.0917
Error	32	3079.940048	96.248127		

Diese Sequenz liefert einen F-Test für x_1 , der um den Einfluss von x_2 bereinigt ist. Die F-Statistik ist

$$F_{Vers} = \frac{(SQ_{Fehler}^{(2')} - SQ_{Fehler}^{(3)})/1}{SQ_{Fehler}^{(3)} / FG_{Fehler}^{(3)}}$$

Dieser Test entspricht dem in 6.9 beschriebenen Vorgehen. Darüber hinaus weist ein Statistik-Paket noch diesen Test für x_2 aus, der aber nicht um den Einfluß von x_1 bereinigt ist (und daher nicht zu bevorzugen ist):

$$F_{Vers} = \frac{(SQ_{Fehler}^{(1)} - SQ_{Fehler}^{(2')})/1}{SQ_{Fehler}^{(3)} / FG_{Fehler}^{(3)}}$$

Ist jedoch der zusätzliche Erklärungswert von x_2 von Interesse, ist die 1. Sequenz zu bevorzugen, weil dort zu erst um x_1 bereinigt wird.

Bemerkung: In dem besonderen Fall, dass die beiden Einflußvariablen x_1 und x_2 **unkorreliert** sind, liefern beide Sequenzen exakt diesselben F-Tests für x_1 und x_2 . Man sagt dann auch, dass x_1 und x_2 **orthogonal** sind.

6.10.2 Erweiterung auf mehr als zwei Variablen

Die multiple Regression kann auf mehr als zwei Einflussvariablen erweitert werden, wobei dieselben Methoden anwendbar sind. Das Modell lautet im allgemeinen Fall:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + e_i$$

Bei vielen potentiellen Einflussvariablen sind sehr viele F-Tests durchzuführen, um zu entscheiden, welche Variablen in das Modell aufgenommen werden sollen. Das Problem der Variablenselektion wird in diesen Fällen der wichtigste Aspekt der Analyse.

Die Modellselektion wirft eine Reihe von Problemen auf, z.B.

- **Multiples Testen** und damit Schwierigkeit der Einhaltung von Irrtumswahrscheinlichkeiten
- Hohe Korrelation von Einflussvariablen (**Multikollinearität**) und damit Austauschbarkeit von Variablen \Rightarrow es gibt kein eindeutig zu bevorzugendes Modell

- Gefahr der Überanpassung (**Overfitting**) \Rightarrow Das ausgewählte Modell hat mehr Parameter, als mit der vorhandenen Datenbasis sinnvoll geschätzt werden können (Die Zahl der Parameter sollte wesentlich kleiner sein als die Zahl der Beobachtungen)
- Es gibt unzählige **Modellselektionskriterien** und -verfahren, von denen keines als generell überlegen gelten kann. Daher steht der Anwender vor dem Problem, ein geeignetes Verfahren unter der Vielzahl der angebotenen auszuwählen (oft sind die Optionen in Statistik-Paketen auf Knopfdruck leicht verfügbar).

Zu weiteren Details sei auf einschlägige Bücher wie das von N. Draper und H. Smith (1998, Applied regression analysis, 3rd edition, Wiley, New York) hingewiesen. Das Problem der Multikollinearität sowie die Variablen-Selektion werden in den Abschnitten 6.10.4 und 6.10.5 etwas näher besprochen.

6.10.3 Das multiple Bestimmtheitsmaß

Wie bei der einfachen linearen Regression können wir auch für die multiple Regression (sowie für jedes beliebige lineare Modell) ein Maß für den Erklärungsgrad des Modells angeben. Hierbei können wir einen Vergleich zweier geschachtelter Modelle zugrundelegen, wobei das reduzierte Modell das einfachste lineare Modell überhaupt ist:

$$(1) y_i = \alpha + e_i \quad (\text{reduziertes Modell - "Nullmodell"})$$

Nach diesem Modell hat keine der x-Variablen einen Einfluss auf die Zielvariable. Dieses Modell wird im folgenden daher als "Nullmodell" bezeichnet. Das SQ_{Fehler} dieses Modells ist gegeben durch

$$SQ_{\text{Fehler}}^{\text{Null}} = \sum_{i=1}^n (y_i - \bar{y}.)^2$$

Für das gewählte volle Modell mit Einflussvariablen

$$(2) y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + e_i \quad (\text{volles Modell})$$

ist das SQ_{Fehler} gegeben durch

$$SQ_{\text{Fehler}} = \sum_{i=1}^n [y_i - (a + b_1 x_{i1} + b_2 x_{i2} + \dots)]^2$$

Für das volle Modell ist das SQ_{Fehler} reduziert gegenüber dem Nullmodell. Je stärker die Reduktion des SQ_{Fehler} , desto höher der Erklärungsgrad des Modells. Die Reduktion des SQ_{Fehler} entspricht dem Anteil der Gesamtstreuung, der durch die aufgenommenen Einflussvariablen erklärt werden kann. Im Extremfall entspricht die Reduktion dem SQ_{Fehler} des Nullmodells, dann nämlich, wenn das volle Modell keine Reststreuung mehr aufweist.

Das multiple Bestimmtheitsmaß für das Modell

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + e_i$$

ist definiert als

$$R^2 = \frac{SQ_{Fehler}^{Null} - SQ_{Fehler}^{voll}}{SQ_{Fehler}^{Null}}$$

wobei

$$SQ_{Fehler}^{Null} = \sum_{i=1}^n (y_i - \bar{y}.)^2$$

$$SQ_{Fehler} = \sum_{i=1}^n [y_i - (a + b_1 x_{i1} + b_2 x_{i2} + \dots)]^2$$

a, b_1, b_2, \dots sind die Kleinst-Quadrat-Lösungen für $\alpha, \beta_1, \beta_2 \dots$

Beispiel: Ratten-Daten

$$SQ_{Fehler}^{Null} = \sum_{i=1}^n (y_i - \bar{y}.)^2 = 4785,78$$

$$SQ_{Fehler} = \sum_{i=1}^n [y_i - (a + b_1 x_{i1} + b_2 x_{i2})]^2 = 3079,94$$

$$R^2 = \frac{SQ_{Fehler}^{Null} - SQ_{Fehler}}{SQ_{Fehler}^{Null}} = \frac{4785,78 - 3079,94}{4785,78} = 0,3564 = 35,64\%$$

Also können ca. 36% der Gesamtstreuung durch das multiple Regressionsmodell erklärt werden.

6.10.4 Multikollinearität

Beispiel: An 31 Schwarzkirschenbäumen im Allegheny National Forest, Pennsylvania wurden folgende Daten erhoben (modifiziert nach Originaldatensatz verfügbar unter <http://www.statsci.org/data/general/cherry.html>):

VOLUME = Volumen des Baumes (Cubic Feet)

HEIGHT = Höhe des Baumes (Feet)

DIAM = Durchmesser (Inches) in 54 Inches über dem Boden

DIAM2 = Durchmesser (Inches) in 30 Inches über dem Boden

Die Daten wurden erhoben, um das Volumen eines Baumes aus seiner Höhe und seinem Durchmesser abschätzen zu können. Eine solche Abschätzung ist hilfreich, da Durchmesser und Höhe leichter direkt zu bestimmen sind als das Volumen. Wir betrachten hier nur die Regression von VOLUME auf die beiden Durchmesser (DIAM und DIAM2):

$$\text{VOLUME} = \alpha + \beta_1 \times \text{DIAM} + \beta_2 \times \text{DIAM2}$$

HEIGHT könnte ebenfalls in die Gleichung aufgenommen werden, und dies wird oft in der Praxis auch gemacht. Hier lassen wir die Höhe der Einfachheit halber außer acht. Es soll geprüft werden, ob einer der beiden Durchmesser signifikant ist (β_1 and β_2).

Für den Signifikanztest werden zwei Modellsequenzen betrachtet (vergleiche Abschnitt 6.10.1!):

Test für DIAM, bereinigt um DIAM2:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Diam2	1	7532.412772	7532.412772	403.87	<.0001
Diam	1	51.450669	51.450669	2.76	0.1079
Error	28	522.220430	18.650730		

Test für DIAM2, bereinigt um DIAM:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Diam	1	7581.781332	7581.781332	406.51	<.0001
Diam2	1	2.082109	2.082109	0.11	0.7408
Error	28	522.220430	18.650730		

Beide Variablen sind hier nicht signifikant, wenn jeweils um die andere Variable bereinigt wird, d.h. die andere Variable wird in der Modellsequenz zuerst angepasst. Führen wir dagegen die Regression mit jeweils nur einer der beiden Variablen durch, so sind beide hoch signifikant:

Für DIAM alleine:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Diam	1	7581.781332	7581.781332	419.36	<.0001
Error	29	524.302539	18.079398		

Für DIAM2 alleine:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Diam2	1	7532.412772	7532.412772	380.78	<.0001
Error	29	573.671099	19.781762		

Dies Ergebnis ist überraschend, und die verschiedenen Analysen scheinen sich zu widersprechen. Wie soll man nun entscheiden, ob und wenn ja welche Variable in das Vorhersagemodell aufgenommen werden soll?

Die widersprüchlichen Ergebnisse sind hier ein Resultat der hohen Korrelation zwischen DIAM und DIAM2 (siehe auch Abb. 6.10.5): $r_{(\text{DIAM}, \text{DIAM2})} = 0,995$.

Eine hohe Korrelation der Einflussvariablen wird als **Multikollinearität** bezeichnet. Wegen der hohen Korrelation von DIAM und DIAM2 ist der zusätzliche Informationsgewinn durch Hinzunahme von DIAM2, wenn DIAM bereits im Modell ist, äußerst gering. Die Reduktion des SQ_{Fehler} ist demzufolge minimal. Das gleiche gilt für die umgekehrte Betrachtung, also Aufnahme von DIAM nach DIAM2. Aus praktischer Sicht reicht es für die Vorhersage des Volumens völlig, nur eine der beiden Einflussvariablen in das Modell zu nehmen.

Etwas einfach formuliert kann man folgendes sagen: Wenn beide Einflussvariablen im Modell sind, kann die Varianzanalyse nicht entscheiden, welche der beiden besser für die Vorhersage ist, weil beide fast identisch sind und weitgehend dieselbe Information liefern.

Abb. 6.10.4 zeigt einen Plot von VOLUME gegen DIAM und DIAM2. Man sieht deutlich, dass die Punkte etwa in einer gedachten dünnen Zigarre oder einem dünnen Zeppelin liegen. Das multiple lineare Regressionsmodell $VOLUME = \alpha + \beta_1 \times DIAM + \beta_2 \times DIAM2$ beschreibt nun eine Ebene, die durch die Zigarre/den Zeppelin gelegt wird. Da die Punkte dicht gedrängt in der Zigarre liegen, wird die Kleinst-Quadrat-Schätzung durch die Längsachse der Zigarre verlaufen. Allerdings wird sich wegen der Lage der Punkte das SQ_{Fehler} wenig ändern, wenn man die Ebene um diese Achse dreht. Somit liefern die Daten wenig Information darüber, wie die Ebene am besten durch die Achse zu legen ist. Eine Drehung um die Achse hat somit wenig Wirkung auf die Vorhersagegenauigkeit des Modells. Allerdings hat eine Drehung um die Achse eine sehr deutliche Veränderung der Regressionskoeffizienten β_1 und β_2 zur Folge.

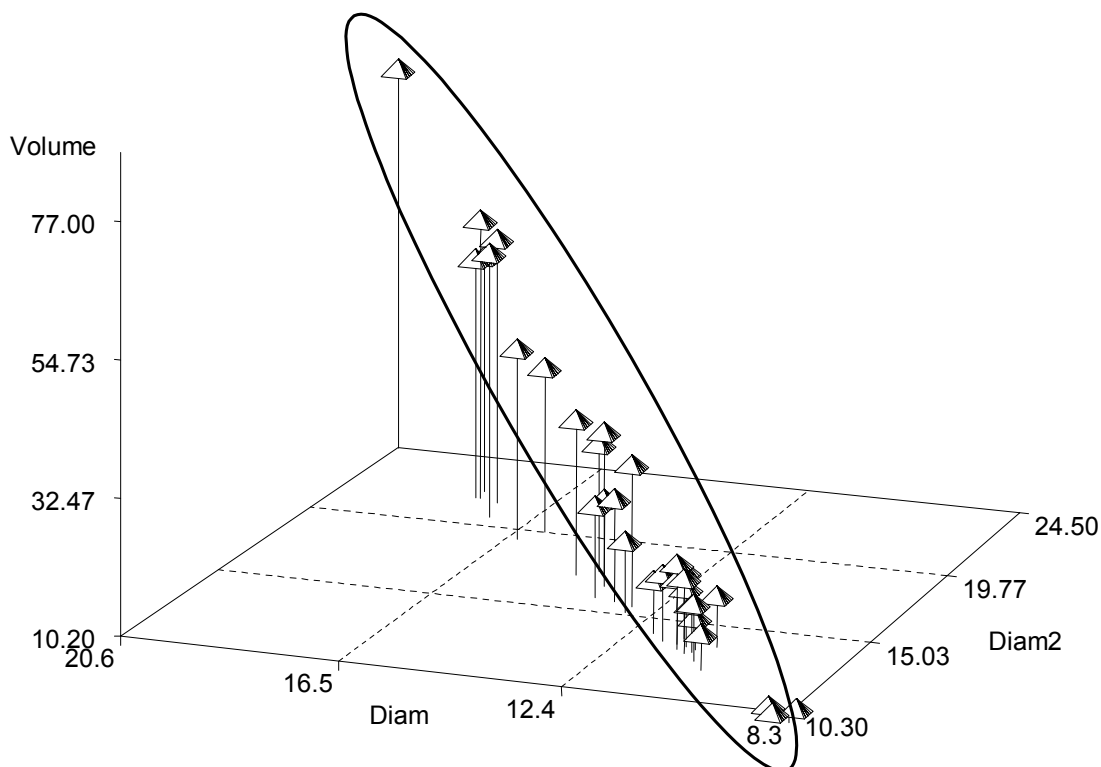


Abb. 6.10.4: 3-D Plot von VOLUME gegen DIAM und DIAM2.

Aus der geringen Information über die genaue Lage der Ebene und der nur kleinen Änderung des SQ_{Fehler} bei Drehung der Ebene um die Längsachse der Zigarre erklären sich die großen Standardfehler der Regressionskoeffizienten und die sog. **Varianzinflation**, die wir im folgenden betrachten.

Mit DIAM und DIAM2 zusammen:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-37.01550833	3.42469843	-10.81	<.0001
Diam	4.22090991	2.54131490	1.66	0.1079
Diam2	0.70699000	2.11596994	0.33	0.7408

$$R^2 = 0,94$$

Mit DIAM alleine:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-36.94345912	3.36514495	-10.98	<.0001
Diam	5.06585642	0.24737695	20.48	<.0001

$$R^2 = 0,94$$

Mit DIAM2 alleine:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-36.82529408	3.52503913	-10.45	<.0001
Diam2	4.20421886	0.21545211	19.51	<.0001

$$R^2 = 0,93$$

Die Standardfehler im vollen Modell (DIAM und DIAM2) sind **aufgebläht (inflated)**, und zwar etwa um den Faktor 10 im Vergleich zu den einfachen Regressionen jeweils nur mit DIAM oder DIAM2. Dieses Phänomen wird als **Varianzinflation** bezeichnet (Varianz = quadrierter Standardfehler). Wegen der Varianzinflation ist keiner der Einflussvariablen bei der gemeinsamen Regression nach dem t-Test signifikant. Für den t-Test werden hier die allgemeinen Methoden aus Abschnitt 6.8 verwendet. Die p-Werte (siehe Anhang D) der t-Tests entsprechen übrigens exakt den p-Werten der "richtigen" F-Tests bei der gemeinsamen Regression auf DIAM und DIAM2.

Die Bestimmtheitsmaße R^2 sind etwa gleich für alle drei Modelle. Wegen der hohen Korrelation zwischen DIAM und DIAM2 reicht einer der beiden für die Vorhersage des Volumens. Das R^2 für DIAM alleine ist etwas größer als für DIAM2 alleine. Daher verwenden wir die Regression auf DIAM:

$$\text{VOLUME} = -36,94 + 5,07 \times \text{DIAM}$$

Bei mehr als zwei Einflussvariablen ist die Identifikation von Multikollinearität etwas schwieriger. Es gibt hier verschiedene Ansätze wie die Berechnung von **Varianzinflationsfaktoren** (VIF) oder die sog. "**condition number**", die aus den Eigenwerten der Matrix $X'X$ berechnet wird. Details finden sich bei Draper & Smith (1998).

6.10.5 Variablenselektion

Bei der multiplen Regression mit einer Zielvariablen y und mehreren Einflussvariablen x_1, x_2, x_3, \dots besteht das Problem der Modellselektion darin, diejenigen Einflussvariablen auszuwählen, die in das Modell aufgenommen werden sollen. Die wichtigste Methode der Modellwahl besteht in der Berücksichtigung von Fachwissen. Mit solcher Information kann in manchen Fällen von vornherein entschieden werden, welche Variablen in das Modell gehören. Oft ist es allerdings gerade ein Ziel der Untersuchung, herauszufinden, welche Einflussvariablen einen Einfluss haben und somit ins Modell gehören. Eine Möglichkeit besteht darin, das volle Modell mit allen Einflussvariablen schätzen und einfach die Signifikanz der Regressionskoeffizienten (t-Tests) zu betrachten, wobei die allgemeine Methode aus Abschnitt 6.8 verwendet wird (Test von $H_0: \lambda = k'\beta = 0$). Für das Beispiel mit den Baumvolumina aus Abschnitt 6.10.4 erhalten wir folgendes Ergebnis:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-37.01550833	3.42469843	-10.81	<.0001
Diam	4.22090991	2.54131490	1.66	0.1079
Diam2	0.70699000	2.11596994	0.33	0.7408

Für beide Einflussvariablen (DIAM und DIAM2) sind die t-Tests nicht signifikant. Aufgrund dieses Ergebnisses würde keiner der beiden Einflussvariablen ins Modell aufgenommen werden. Der Grund ist wie oben besprochen die Multikollinearität und die durch sie bedingte Varianzinflation. Offensichtlich ist die Betrachtung aller t-Tests basierend auf dem vollen Modell keine gute Strategie zur Modellwahl. Dies gilt umso mehr, wenn sehr viele Einflussvariablen vorliegen. Bei Multikollinearität kann es passieren, dass Einflussvariablen signifikant werden, wenn man eine oder mehrere der Einflussvariablen aus dem Modell nimmt, wie das Baum-Beispiel gezeigt hat. Wir werden gleich zu einem weiteren solchen Beispiel kommen.

Es gibt eine ganze Reihe anderer Ansätze zur Modellselektion, welche alternativ verwendet werden können. Eine eindeutige Empfehlung zu Gunsten eines der Verfahren ist schwer zu geben. Hier sollen lediglich einige der Ansätze beschrieben werden, um einen Eindruck der verschiedenen Möglichkeiten zu geben. Den meisten Verfahren sind zwei Bausteine/Komponenten gemeinsam:

1. Ein statistisches Kriterium, mit welchem die Güte eines Modells beurteilt wird
2. Ein Algorithmus, mit welchem die näher zu untersuchenden Modelle ausgewählt werden

Zum 2. Punkt ist zu sagen, dass prinzipiell für alle möglichen Modelle das ausgewählte Gütekriterium berechnet werden kann. Das Problem dabei ist, dass die Zahl der zu betrachtenden Modelle sehr hoch werden kann. Bei p Einflussvariablen gibt es 2^p Modelle. Für $p = 10$ sind das z.B. $2^{10} = 1024$ Modelle. Dies bedeutet, dass eine Betrachtung aller möglichen Modelle sehr rechenintensiv werden kann. Praktikabler ist es daher, nur einen Teil aller möglichen Modelle zu untersuchen.

Statistische Kriterien zur Modellwahl

- Bestimmtheitsmaß (R^2)
- Adjustiertes Bestimmtheitsmaß (adjustiertes R^2) (*größer ist besser*)
- Mittleres Abweichungsquadrat (s^2) (*kleiner ist besser*)
- Akaikes Informationskriterium (AIC) (*kleiner ist besser*)
- Mallows C_p (*kleiner ist besser*)

Das Bestimmtheitsmaß ist bereits in Abschnitt 6.10.3 vorgestellt worden. Grob gesagt gilt: Je höher das Bestimmtheitsmaß, desto „besser“ das Modell. Hierbei ist allerdings zu berücksichtigen, dass das Bestimmtheitsmaß (R^2) mit jeder hinzu kommenden Einflussvariable ansteigen muss, weil gleichzeitig immer das SQ_{Fehler} sinkt. Daher ist das unkorrigierte R^2 nur zum Vergleich von Modellen mit derselben Anzahl von Einflussvariablen geeignet. Zum Vergleich von Modellen mit unterschiedlicher Anzahl von Einflussvariablen ist das **adjustierte Bestimmtheitsmaß (adj. R^2)** zu bevorzugen. Dieses Maß berücksichtigt die Zahl der Parameter im Modell und kann aus dem Bestimmtheitsmaß R^2 berechnet werden:

$$R^2(adj) = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Hierbei ist p die Zahl der Einflussvariablen und n die Zahl der Beobachtungen. Dieses Maß ergibt sich, wenn man die Restvarianz des Nullmodells (ohne x-Variablen), gegeben durch

$$s_{Null}^2 = \frac{SQ_{Fehler}^{Null}}{n-1},$$

mit der Restvarianz $s^2 = \frac{SQ_{Fehler}}{n-p-1}$ des gerade betrachteten Modell mit p x-Variablen vergleicht. Der erklärte Teil der Varianz ist dann

$$R^2(adj) = 1 - \frac{s^2}{s_{Null}^2} = 1 - \frac{n-1}{n-p-1} \times \frac{SQ_{Fehler}}{SQ_{Fehler}^{Null}} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Hierbei haben wir $R^2 = \frac{SQ_{Fehler}^{Null} - SQ_{Fehler}}{SQ_{Fehler}^{Null}} = 1 - \frac{SQ_{Fehler}}{SQ_{Fehler}^{Null}}$ benutzt. Die Reihenfolge der

Werte von $R^2(adj)$ für verschiedene Modelle hängt nur von der Restvarianz s^2 ab. Diese kann daher auch direkt als Modellselektionskriterium verwendet werden.

Die Restvarianz, oder das **mittlere Abweichungsquadrat** ergibt sich aus der Varianzanalyse für das jeweils angepasste Modell. Es berechnet sich als Quotient des SQ_{Fehler} und der Fehler-Freiheitsgrade (FG_{Fehler}):

$$s^2 = \frac{SQ_{Fehler}}{n-p-1} \quad (\text{kleiner ist besser}).$$

Hierbei ist n die Zahl der Beobachtungen und p die Zahl der Einflussvariablen im Modell. Je kleiner s^2 ist, desto besser ist die Anpassung. Wenn ein Modell sukzessive aufgebaut wird, wobei die Einflussvariablen mit dem größten Erklärungswert zuerst aufgenommen werden, so sinkt s^2 kontinuierlich ab, bis Variablen aufgenommen werden, die keinen Einfluss haben: Dann stabilisiert sich der Wert von s^2 , und auch der von $R^2(adj)$. Sowohl s^2 als auch $R^2(adj)$, welches von s^2 abhängt, haben einen **Strafterm**, der mit der Zahl der Parameter p ansteigt. Dies sieht man gut durch Logarithmieren der Restvarianz s^2 :

$$\log(s^2) = \log(SQ_{Fehler}) - \log(n - p - 1).$$

Die Aufnahme einer weiteren Variable reduziert den Term $\log(SQ_{Fehler})$, erhöht aber gleichzeitig den Term $\log(n - p - 1)$. Wenn die Variable keinen großen erklärenden Wert hat, so halten sich beide Terme etwa die Waage. Zwei weitere Selektionskriterien, die einen Strafterm haben, sind wie folgt definiert:

Akaike Information Criterion (AIC) (kleiner ist besser):

$$AIC = n \log(SQ_{Fehler} / n) + 2(p + 1)$$

Mallows C_p für ein Modell mit p Einflussvariablen berechnet sich wie folgt:

$$C_p = \frac{SQ_{Fehler}}{s_{voll}^2} + 2(p + 1) - n$$

Hierbei ist SQ_{Fehler} die Summe der Abweichungsquadrate für das betrachtete Modell, s_{voll}^2 ist das mittlere Abweichungsquadrat für das volle Modell (Modell mit allen Einflussvariablen), p = Zahl der Einflussvariablen im Modell und n = Zahl der Beobachtungen. Je kleiner Mallows C_p , desto besser das Modell.

Für ein "wahres" Modell, also ein Modell, welches die Daten erzeugt haben könnte, gilt $E(C_p) \approx p+1$. Im Mittel erwarten wir also, dass der Wert für C_p nahe $p+1$ liegt, falls das betrachtete Modell zutrifft. Dies gilt auch für Modelle, die zusätzlich überflüssige Variablen im Modell haben. Für das volle Modell gilt sogar exakt $C_p = p+1$. Dagegen kann Mallows C_p noch substantiell sinken, solange wichtige Variablen fehlen.

Mallows C_p eignet sich vor allem zur Auswahl von Modellen, die zur Vorhersage genutzt werden sollen. Vorhersage bedeutet, dass das aus den Daten geschätzte

Modell zur Vorhersage neuer Daten verwendet wird. Bei einer solchen Vorhersage begeht man in der Regel einen Vorhersagefehler. Man könnte nun zunächst annehmen, dass es am besten ist, alle Einflussvariablen in das Modell zu nehmen, die irgendeinen Einfluss haben. Dies ist aber nicht so. Das Problem besteht darin, dass die Genauigkeit der Schätzung eines einzelnen Regressionsparameters umso schlechter ist, je mehr Regressionsparameter insgesamt geschätzt werden müssen. Je mehr Parameter das Modell also hat, desto größer ist die Varianz jeder einzelnen Parameterschätzung. Aus diesem Grund kann es vorteilhaft sein, einige Einflussvariablen aus dem Modell zuzunehmen, selbst wenn sie einen (kleinen) Einfluss haben. Hiermit handelt man sich zwar eine verzerrte Vorhersage (Bias) ein, weil Variablen ignoriert werden, die erklärenden Wert haben, reduziert aber gleichzeitig die Größe der Varianz der einzelnen Parameterschätzung.

Für die Vorhersage ist aber der mittlere quadratische Vorhersagefehler entscheidend (Mean Squared Error of Prediction; *MSEP*). Mallows C_p ist ein Schätzwert für den *MSEP*. Der *MSEP* setzt sich aus Varianz und Verzerrung (Bias) zusammen. Daher müssen bei der Beurteilung eines geschätzten Modells Verzerrung und Varianz gegeneinander abgewogen werden. Dies soll anhand des Bildes einer Zielscheibe verdeutlicht werden (Abb. 6.10.5). Trifft man "im Durchschnitt" das Zentrum der Scheibe, so liegt keine Verzerrung vor. Dabei kann die Streuung (Varianz) groß oder klein sein, wie in der linken und der mittleren Scheibe dargestellt ist. Trifft man dagegen im Mittel daneben (rechte Scheibe), so liegt eine Verzerrung vor.

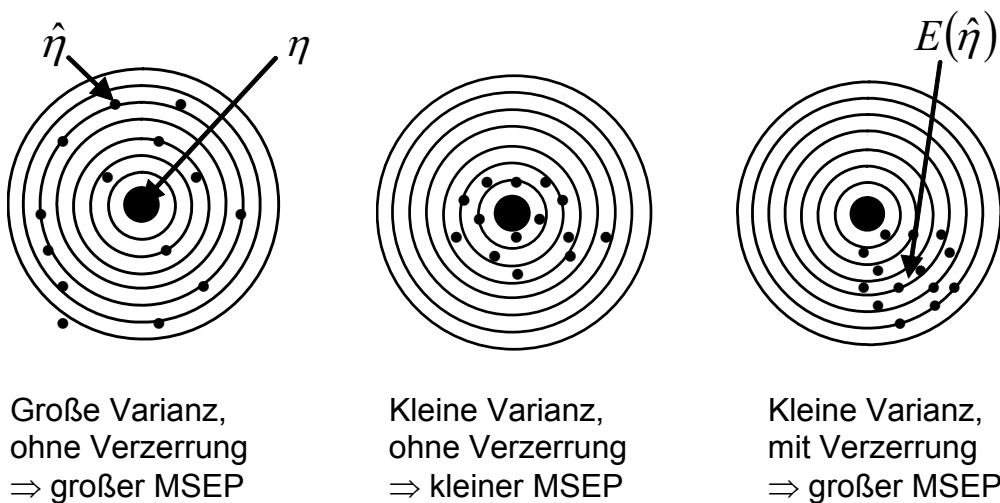


Abb. 6.10.5: Verzerrung (Bias) und Varianz von Modellschätzungen ($\hat{\eta}$) des wahren Modells η illustriert am Bild einer Zielscheibe. Jeder Einschuß entspricht einem Experiment.

Auf das Bild der Zielscheiben kommen wir gleich zurück, nachdem die Begriffe etwas formalisiert worden sind. Im folgenden verwenden wir die nachstehenden Bezeichnungen:

- η = wahres Modell
- $\hat{\eta}$ = Schätzwert des Modells, basierend auf Daten und Modellannahmen, die nicht notwendigerweise das "wahre Modell" treffen müssen

$E(\hat{\eta})$ = Erwartungswert der Modellschätzung

Im Fall eines linearen Modells haben wir

$$\eta = X\beta$$

und

$$\hat{\eta} = X\hat{\beta}.$$

Für die Güte der Modellanpassung ist nun die Differenz des wahren Modells und seiner Schätzung relevant, also die Diskrepanz

$$\hat{\eta} - \eta,$$

wobei es wie gesagt sein kann, dass wir den falschen Modelltyp für $\hat{\eta}$ gewählt haben. Der Erwartungswert für das Quadrat dieser Diskrepanz ist der sog. mittlere quadratische Vorhersagefehler:

$$MSEP(\hat{\eta}) = E\{[\hat{\eta} - \eta]^2\}.$$

Man kann diesen MSEP nun in zwei Komponenten zerlegen, von denen die eine von einer **systematischen Diskrepanz** und die andere von einer **zufälligen Diskrepanz** zwischen dem wahren Modell und seiner Schätzung abhängt. Hierzu betrachten wir zunächst die Identität

$$\hat{\eta} - \eta = \hat{\eta} - E(\hat{\eta}) + E(\hat{\eta}) - \eta$$

Gesamte Diskrepanz	Zufällige Diskrepanz	Systematische Diskrepanz
-----------------------	-------------------------	-----------------------------

(Bias, Verzerrung)

Wenn die Modellannahmen nicht perfekt stimmen, wenn also beispielsweise eine wichtige Einflußvariable im Regressionsmodell fehlt, so gilt für die systematische Diskrepanz, die auch **Verzerrung** oder **Bias** genannt wird:

$$Bias(\hat{\eta}) = E(\hat{\eta}) - \eta \neq 0.$$

In Worten: die Modellschätzung liegt systematisch über oder unter dem wahren Wert. Eine Verzerrung (Bias) ist jedoch nicht notwendigerweise problematisch, sofern die Modellschätzung eine kleine Varianz aufweist (siehe unten). Entscheidend ist die Größe des mittleren quadratischen Vorhersagefehlers (MSEP). Dieses läßt sich mit der obigen Identität wie folgt ausdrücken:

$$E[(\hat{\eta} - \eta)^2] = E\{[\hat{\eta} - E(\hat{\eta})]^2\} + [E(\hat{\eta}) - \eta]^2$$

Gesamte Diskrepanz	Zufällige Diskrepanz	Systematische Diskrepanz
Mittlerer quadratischer Vorhersagefehler	Varianz	Bias ²

Eine Verzerrung (Bias) ist also tolerierbar, sofern die Modellschätzung eine kleine Varianz aufweist, also wenn

$$Varianz(\hat{\eta}) = E\{[\hat{\eta} - E(\hat{\eta})]^2\}$$

deutlich kleiner ist als für andere angenommene Modelle. Denn für die Vorhersagegüte eines Modells ist gewissermaßen der Nettoeffekt von Varianz und Verzerrung ausschlaggebend, der MSEP. Dieser hängt von Varianz und Verzerrung wie oben hergeleitet wie folgt ab:

$$MSEP(\hat{\eta}) = Varianz(\hat{\eta}) + [Bias(\hat{\eta})]^2.$$

Die entscheidende Frage ist nun, ob die Verzerrung (Bias) durch Auslassen einer Variablen größer oder kleiner ist als der Effekt der Reduzierung der Varianz der Parameterschätzungen. Wenn der Effekt der Reduzierung der Varianz größer ist als die Zunahme an Bias, so ist der Nettoeffekt eine Reduzierung des mittleren quadratischen Vorhersagefehlers (MSEP), und es lohnt sich, die Variable auszulassen. Um diese Zusammenhänge zu veranschaulichen, wird wieder das Bild der Zielscheibe verwendet (Abb. 6.10.5). Der Zielpunkt entspricht dabei dem wahren Modell, η , das wir schätzen wollen. Unser Gewehr entspricht nun einer bestimmten Modellannahme, und ein Schuss aus dem Gewehr entspricht der Schätzung $\hat{\eta}$ dieses Modells für einen realen Datensatz, basierend auf einem Experiment. Mehrere Schüsse implizieren daher eine vollständige Wiederholung des Experimentes oder der Erhebung. Bei der ersten Scheibe treffen wir im Mittel das richtige Ziel, aber mit einer großen Varianz; es liegt keine Verzerrung vor. Bei der zweiten Scheibe treffen wir mit größerer Genauigkeit, also mit kleinerer Varianz, wieder ohne Verzerrung. Bei der letzten Scheibe ist nun die Varianz ebenfalls kleiner als bei der ersten, aber wir schießen systematisch daneben. Dies entspricht einer Verzerrung. Ob nun das Einschussbild der ersten oder der letzten Scheibe besser ist, lässt sich nicht ohne weiteres sagen. Beim einen haben wir die größere Varianz, dafür haben wir beim anderen die systematische Verzerrung zu veranschlagen. Zur zusammenfassenden Beurteilung können wir den mittleren quadratischen Abstand vom Zielpunkt berechnen, und dies entspricht dem MSEP. Es kann durchaus sein, dass der Effekt der Verzerrung bei der dritten Zielscheibe geringer ist als die vergrößerte Varianz bei der ersten Zielscheibe. Übertragen auf das Problem der Modellwahl entspricht die dritte Zielscheibe der Anpassung eines einfachen Modells, das möglicherweise deutlich einfacher ist als das "wahre" Modell. Die erste Scheibe entspricht dagegen der Anpassung eines komplexen, möglicherweise des "wahren" Modells, allerdings bei begrenzter Datengrundlage. Hier kann es eben von Vorteil

sein, das einfachere Modell zu verwenden, wenn die Reduktion der Varianz größer ist als die Zunahme des quadrierten Bias.

Die hier besprochenen Selektionskriterien haben als Gemeinsamkeit, dass ein Strafterm für die Zahl der Parameter in das Kriterium eingeht. Damit wirken diese Kriterien der Gefahr einer **Überanpassung (Overfitting)** entgegen. Mit Overfitting ist gemeint, dass das selektierte Modell zu viele Parameter hat und der oben beschriebene Varianz-Bias-Tradeoff negativ wird, wodurch sich das *MSEP* erhöht. Wir wollen das Problem des Overfitting an einem einfachen Beispiel erläutern. Angenommen, es besteht kein Zusammenhang zwischen einer Zielvariable y und einer Einflussvariable x . Das "wahre" Modell lässt sich also schreiben als

$$y_i = \alpha + e_i$$

und somit

$$\eta_i = \alpha.$$

Angenommen, es gilt $\alpha = 2$. Wir beobachten nun zwei Datenpunkte bei $x_1 = 10$ und $x_2 = 20$, die von diesem wahren Modell generiert werden. Wegen des Fehlers e streuen diese Punkte um $\alpha = 2$. Die beobachteten Punkte sind wie folgt:

x_i	y_i
10	1,8
20	2,5

Wir betrachten nun zwei alternative Schätzungen. Zum einen schätzen wir nach dem wahren Modell und verwenden also

$$\hat{\eta}_i = \hat{\alpha} = \bar{y}_\bullet.$$

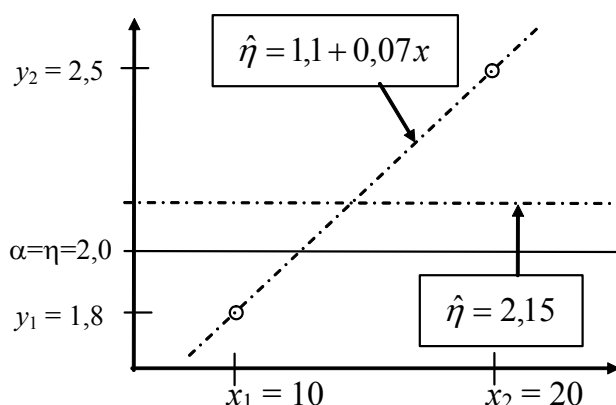


Abb. 6.10.6: Overfitting

Zum anderen passen wir ein lineares Regressionsmodell der Form

$$y_i = \alpha + \beta x_i + e_i$$

an (Abb. 6.10.6). Hier ist also

$$\hat{\eta}_i = \hat{\alpha} + \hat{\beta}x_i.$$

Da nur 2 Datenpunkte vorliegen, geht die angepasste Gerade genau durch beide Punkte, und SQ_{Fehler} wird gleich 0. Das angepasste Modell lautet $\hat{\eta} = 1,1 + 0,07x$. Es sagt an den beiden Stellen $x_1 = 10$ und $x_2 = 20$ jeweils genau die beobachteten Werte ($y_1 = 1,8$ und $y_2 = 2,5$) voraus. Die "wahren" Werte sind aber gemäß des erzeugenden Modells jeweils gleich $\alpha = 2,0$. Somit ergibt sich eine deutliche Diskrepanz zwischen wahrem Modell und angepasstem Modell. Die Summe der beiden quadrierten Abweichungen beträgt

$$\hat{MSEP} = n^{-1} \sum_{i=1}^n (\hat{\eta}_i - \eta_i)^2 = [(1,8 - 2,0)^2 + (2,5 - 2,0)^2] / 2 = 0,145$$

Passen wir dagegen das wahre Modell $y_i = \alpha + e_i$ an, so erhalten wir

$$\hat{\eta} = \hat{\alpha} = \frac{1,8 + 2,5}{2} = 2,15$$

Hierfür ist die Summe der Abweichungsquadrate zwischen wahrem Modell ($\alpha = 2,0$) und vorhergesagtem Wert gleich

$$\hat{MSEP} = n^{-1} \sum_{i=1}^n (\hat{\eta}_i - \eta_i)^2 = [(2,15 - 2,0)^2 + (2,15 - 2,0)^2] / 2 = 0,0225$$

Die Abweichung ist also viel kleiner als für das Regressionsmodell ($0,0225 \ll 0,145$). **Obwohl das Regressionsmodell besser an die Daten passt** ($SQ_{Fehler} = 0$, $R^2 = 1$), **passt es viel schlechter an das wahre (erzeugende) Modell** ($\hat{MSEP} = 0,145$).

In diesem Beispiel verursacht das falsche Modell ($y_i = \alpha + \beta x_i + e_i$) übrigens keine Verzerrung, sondern lediglich eine Erhöhung der Varianz der Schätzung. Es läßt sich zeigen dass gilt:

$$E(\hat{\eta}_i) = E(\hat{\alpha} + \hat{\beta}x) = \eta.$$

Zwar ist dies ein sehr überzogenes Beispiel. In der Praxis werden wir mehr als nur zwei Datenpunkte haben. Das Beispiel zeigt jedoch folgendes: Verlassen wir uns bei der Modellwahl allein auf das R^2 , so kommt es zu einer Überanpassung (Overfitting). Verwendung von Selektionskriterien wie Mallows' C_p , adj. R^2 , AIC oder s^2 wirken dieser Gefahr entgegen.

Strategien zur Modellwahl

Beispiel: In einer Untersuchung wurde der Einfluss verschiedener Mineralien in Steinschlacke auf die Eigenschaften von Zement untersucht. Eine der Eigenschaften war die Erwärmung bei der Herstellung des Zement. Die Daten sind im folgenden wiedergegeben (Draper NR and Smith S 1981 Applied linear regression. 2nd Edition. Wiley, New York)

x_1	x_2	x_3	x_4	y
7	26	6	60	78,5
1	29	15	52	74,3
11	56	8	20	104,3
11	31	8	47	87,6
7	52	6	33	95,9
11	55	9	22	109,2
3	71	17	6	102,7
1	31	22	44	72,5
2	54	18	22	93,1
21	47	4	26	115,9
1	40	23	34	83,8
11	66	9	12	113,3
10	68	8	12	109,4

x_1 = 3 CaO•Al₂O₃ (Tri-Calzium-Aluminat)

x_2 = 3 CaO•SiO₂ (Tri-Calzium-Silikat)

x_3 = 4 CaO•Al₂O₃•Fe₂O₃ (Tetra-Calzium-Ferralit)

x_4 = 2 CaO•SiO₂ (Di-Calzium-Silikat)

y = Wärmeentwicklung (Kalorien) pro Gramm Zement

Die Einflussvariablen x_1 bis x_4 wurden in Prozent des Gewichts der Steinschlacke, aus welcher der Zement gemacht wurde, gemessen. Für die Herstellung von Zement sollte der Einfluss der verschiedenen Mineralbestandteile ermittelt werden. Hierfür kommt u.a. eine multiple lineare Regression in Frage:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + e_i$$

Das Ergebnis der Anpassung dieses vollen Modells ist wie folgt:

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	62.405369	70.07095921	0.891	0.3991
X1	1	1.551103	0.74476987	2.083	0.0708
X2	1	0.510168	0.72378800	0.705	0.5009
X3	1	0.101909	0.75470905	0.135	0.8959
X4	1	-0.144061	0.70905206	-0.203	0.8441

Keiner der Regressionskoeffizienten ist signifikant! Nach diesem Ergebnis könnten wir geneigt sein, alle Variablen aus dem Modell zu nehmen. Hierzu ist zu bemerken, dass wir hier einen offensichtlichen Fall von Multikollinearität haben. Denn es gilt $x_1 +$

$x_2 + x_3 + x_4 \approx 100\%$. Aus diesem Grund besteht eine hohe Abhängigkeit (Multikollinearität) unter den „unabhängigen“ Variablen (Einflussvariablen). Hierdurch ist die Designmatrix X nahezu singulär (nicht von vollem Rang). Dies ist der Grund für die relativ hohen Standardfehler (Varianzinflation) und die Nicht-Signifikanz aller Regressionsparameter. Lassen wir z.B. x_3 weg, so erhalten wir folgendes Ergebnis:

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	71.648307	14.14239348	5.066	0.0007
X1	1	1.451938	0.11699759	12.410	0.0001
X2	1	0.416110	0.18561049	2.242	0.0517
X4	1	-0.236540	0.17328779	-1.365	0.2054

Nun ergeben sich einige Signifikanzen. Das Beispiel zeigt, dass bei Multikollinearität die t-Tests für einzelne Regressionsparameter des vollen Modells irreführend sein können (Man beachte auch wieder die sehr großen Standardfehler für die Regressionsparameter im vollen Modell). Daher sind, wie oben bereits erwähnt, andere Strategien zur Modellselektion zu bevorzugen.

All subsets regression: Die einfachste Strategie besteht darin, alle möglichen Regressionsmodelle durchzurechnen und das Gütekriterium (s^2 , Mallows C_p , adj. R^2) zu berechnen. Man wählt dann dasjenige Modell mit dem günstigsten Wert des Selektionskriteriums. In unserem Beispiel soll Mallows C_p verwendet werden. Das Modell mit den Variablen x_1 und x_2 hat das kleinste C_p und ist daher zu bevorzugen:

C(p)	R-square	Variables in Model	
		In	
2.67824	0.97867837	2	X1 X2
3.01823	0.98233545	3	X1 X2 X4
3.04128	0.98228468	3	X1 X2 X3
3.49682	0.98128109	3	X1 X3 X4
5.00000	0.98237562	4	X1 X2 X3 X4
5.49585	0.97247105	2	X1 X4
7.33747	0.97281996	3	X2 X3 X4
22.37311	0.93528964	2	X3 X4
62.43772	0.84702542	2	X2 X3
138.22592	0.68006041	2	X2 X4
138.73083	0.67454196	1	X4
142.48641	0.66626826	1	X2
198.09465	0.54816675	2	X1 X3
202.54877	0.53394802	1	X1
315.15428	0.28587273	1	X3

Wie bereits bemerkt, müssen bei einer großen Zahl von Einflussvariablen sehr viele Modelle untersucht werden. Daher sind andere Selektionsstrategien vorgeschlagen worden, die den Rechenaufwand erheblich reduzieren.

Best subsets regression: Es gibt verschiedene Algorithmen, welche die K besten Modelle auffinden und nur für diese die vollständige Berechnung durchführen. Der

populärste ist der sog. *leaps-and-bounds* Algorithmus von Furnival und Wilson (siehe Draper und Smith, 1981). Dem Computer-Programm muss lediglich die Zahl K der besten Modelle angegeben werden, die berechnet werden sollen.

Backward elimination: Bei diesem Verfahren wird vom vollen Modell ausgegangen. Es werden die t-Tests (oder äquivalent alle F-Tests) für alle Regressionsparameter betrachtet. Die Einflussvariable mit dem betraglich kleinsten t-Wert wird aus dem Modell genommen. Im Beispiel ist das x_3 . Mit dem reduzierten Modell wird dann eine erneute Varianzanalyse sowie die dazugehörigen t-Tests für die verbleibenden Regressionsparameter berechnet. In unserem Beispiel hat in diesem 2. Schritt x_4 den betraglich kleinsten t-Wert ($|t| = 1,365$). Die Elimination von Einflussvariablen wird solange fortgesetzt, bis alle noch verbleibenden Variablen einen signifikanten t-Test liefern. Im Zement-Beispiel werden auf diese Weise die Variablen x_3 und x_4 aus dem Modell genommen, so dass das Modell

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

ausgewählt wird.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x3	3	0.0000	0.9823	3.0182	0.02	0.8959
2	x4	2	0.0037	0.9787	2.6782	1.86	0.2054

All variables left in the model are significant at the 0.1000 level.

Das Modell, in dem x_3 und x_4 herausgelassen werden, hat unter den untersuchten Modellen den kleinsten Wert für Mallows C_p und ist daher zu bevorzugen. Obwohl der Selektionsalgorithmus auf t-Tests beruht, sollte das Modellwahlkriterium, und nicht die p-Werte, für die Beurteilung der Modelle verwendet werden.

Forward selection: Dies ist die Umkehrung der *backward elimination*. Man untersucht zunächst alle Modelle mit nur einer Einflussvariable und wählt das Modell mit dem höchsten t-Wert für den Regressionskoeffizienten. Ausgehend von diesem Modell wird eine Variable hinzugefügt. Dabei werden alle verbleibenden Einflussvariablen ausprobiert. Diejenige mit dem höchsten t-Wert wird in das Modell aufgenommen. Das Verfahren wird so lange fortgesetzt, bis keine hinzukommende Variable mehr einen signifikanten t-Wert ergibt. Mit diesem Verfahren wird im Zement-Beispiel ein Modell mit den Variablen x_1 , x_2 und x_4 ausgewählt. Dieses Modell hat auch den günstigsten Wert für Mallows C_p .

Man beachte, dass bei der Vorwärts-Selektion das beste Modell (all subsets) nicht betrachtet wird (Modell mit x_1 und x_2). Es ist eine typische Eigenschaft sowohl von Vorwärts-Selektion als auch von Rückwärts-Elimination, dass das nach dem all subsets Verfahren identifizierte beste Modell nicht immer gefunden wird. Dies ist ein wesentlicher Nachteil dieser beiden Verfahren.

No other variable met the 0.50 significance level for entry into the model.

Summary of Forward Selection Procedure for Dependent Variable Y

Step	Variable Entered	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X4	1	0.6745	0.6745	138.7308	22.7985	0.0006
2	X1	2	0.2979	0.9725	5.4959	108.2239	0.0001
3	X2	3	0.0099	0.9823	3.0182	5.0259	0.0517

Bemerkungen: (i) Im Output zur Vorwärts-Selektion und zur Rückwärts-Elimination taucht ein F-test auf. Dieser ist äquivalent zu dem t-Test für den Regressionskoeffizienten. (ii) Bei *backward elimination* und *forward selection* muss ein Signifikanzniveau für die t-Tests (F-Tests) vorgegeben werden. Die Voreinstellung der hier verwendeten SAS Prozedur REG ist $\alpha = 0,10$ für *backward elimination* und $\alpha = 0,50$ bei *forward selection*. Diese Signifikanzniveaus sind viel liberaler (man findet mehr Signifikanzen) als die üblichen 5%. Hinzu kommt, dass wegen des Selektionsalgorithmus die wahren Irrtumswahrscheinlichkeiten nicht mit den nominellen übereinstimmen. Die Tests sollten daher lediglich als Screening-Instrument aufgefasst werden. Eine vorgegebene Irrtumswahrscheinlichkeit kann dabei nicht eingehalten werden. **Statt auf die p-Werte der F-Tests sollte man auf den Wert des Modellwahlkriteriums schauen, hier Mallows C_p .**

Stepwise regression: Dies Verfahren stellt eine Kombination von **forward selection** und **backward elimination** dar. So kann eine Variable, die bereits in das Modell aufgenommen wurde (forward selection), zu einem späteren Zeitpunkt wieder gelöscht werden (backward elimination). Die Prozedur endet, wenn keine der verbleibenden Variablen das Kriterium für die Aufnahme in das Modell erreicht. Üblicherweise wird über Aufnahme oder Elimination durch einen t-Test (bzw. F-Test) entschieden. Zur Illustration hier eine Auswertung, bei der das Signifikanzniveau $\alpha = 10\%$ beträgt.

Number of Observations Read 13
Number of Observations Used 13

Stepwise Selection: Step 1

Variable x4 Entered: R-Square = 0.6745 and C(p) = 138.7308

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1831.89616	1831.89616	22.80	0.0006
Error	11	883.86692	80.35154		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	117.56793	5.26221	40108	499.16	<.0001
x4	-0.73816	0.15460	1831.89616	22.80	0.0006

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable x1 Entered: R-Square = 0.9725 and C(p) = 5.4959

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2641.00096	1320.50048	176.63	<.0001
Error	10	74.76211	7.47621		
Corrected Total	12	2715.76308			

Stepwise Selection: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	103.09738	2.12398	17615	2356.10	<.0001
x1	1.43996	0.13842	809.10480	108.22	<.0001
x4	-0.61395	0.04864	1190.92464	159.30	<.0001

Bounds on condition number: 1.0641, 4.2564

Stepwise Selection: Step 3

Variable x2 Entered: R-Square = 0.9823 and C(p) = 3.0182

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2667.79035	889.26345	166.83	<.0001
Error	9	47.97273	5.33030		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	71.64831	14.14239	136.81003	25.67	0.0007
x1	1.45194	0.11700	820.90740	154.01	<.0001
x2	0.41611	0.18561	26.78938	5.03	0.0517
x4	-0.23654	0.17329	9.93175	1.86	0.2054

Bounds on condition number: 18.94, 116.36

Stepwise Selection: Step 4

Variable x4 Removed: R-Square = 0.9787 and C(p) = 2.6782

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2657.85859	1328.92930	229.50	<.0001
Error	10	57.90448	5.79045		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	52.57735	2.28617	3062.60416	528.91	<.0001
x1	1.46831	0.12130	848.43186	146.52	<.0001
x2	0.66225	0.04585	1207.78227	208.58	<.0001

Bounds on condition number: 1.0551, 4.2205

All variables left in the model are significant at the 0.1000 level.

No other variable met the 0.1000 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		1	0.6745	0.6745	138.731	22.80	0.0006
2	x1		2	0.2979	0.9725	5.4959	108.22	<.0001
3	x2		3	0.0099	0.9823	3.0182	5.03	0.0517
4		x4	2	0.0037	0.9787	2.6782	1.86	0.2054

Es werden zunächst x_4 , x_1 und x_2 aufgenommen. Im Modell, welches diese drei Variablen umfasst, ist jedoch x_4 nicht mehr signifikant bei $\alpha = 10\%$ und wird daher aus dem Modell genommen. Danach ist weder die Aufnahme von x_3 , noch die von x_4 angezeigt, da die entsprechenden F-Tests nicht signifikant sind. Daher werden die Variablen x_1 und x_2 selektiert. Interessant ist, dass x_4 zwar als erste Variable selektiert wird, dann aber am Ende wieder aus dem Modell genommen wird. Grund hierfür ist die Multikollinearität zwischen den Einflussvariablen.

Partielles R^2 : Das partielle Bestimmtheitsmaß (R^2) gibt an, um welchen Betrag sich das Bestimmtheitsmaß durch die Aufnahme oder Elimination einer Variable ändert.

Abschließende Bemerkung: Oft gibt es nicht nur ein bestes Modell sondern viele beste Modelle. Wenn unterschiedliche Modelle mit verschiedener Anzahl von Einflussvariablen etwa gleiche Werte für z.B. Mallows C_p ergeben, so ist jeweils das einfachere Modell zu bevorzugen. Es ist oft vorteilhaft eine Reihe von besten

Modellen (siehe z.B. *best subset selection*) zu betrachten, um zu sehen, welche Einflussvariablen einen Einfluss auf die Zielvariable haben.

Beispiel: (Backhaus et al. 2000 Multivariate Analysemethoden. Springer, Berlin). Eine Marktforscherin eines Margarineherstellers will untersuchen, welche Faktoren den Absatz von Margarine beeinflussen. Sie wählt hierzu zufällig 37 Verkaufsgebiete (geographische Regionen) und erhebt eine Zahl von Variablen, welche den Absatz beeinflussen könnten und auf die der Hersteller einen Einfluss hat:

Absatz	Verkaufte Menge (Kartons pro Gebiet)
Preis	Preis der Margarine (DM/Karton)
Werbung	Ausgaben für Werbung (DM/Gebiet)
Besuche	Zahl der Besuche eines Vertreters

Die Marktforscherin will wissen, wie der Absatz von diesen Variablen beeinflusst wird. Es wird eine Gleichung benötigt, die den Absatz als Funktion dieser Variablen vorhersagt.

Tab. 6.10: Vier Beobachtungen aus den Margarine-Daten.

Absatz	Preis	Werbung	Besuche
2585	12,50	2000	109
1819	10,00	550	107
1647	9,95	1000	99
1496	11,50	800	70

Mit der *all subsets* Methode und dem adjustierten R^2 finden wir folgendes Ergebnis:

The REG Procedure				
Model: MODEL1				
Dependent Variable: absatz				
Adjusted R-Square Selection Method				
Number in Model	Adjusted R-Square	R-Square	Variables in Model	
3	0.8327	0.8466	preis werbung besuche	
2	0.7989	0.8101	werbung besuche	
2	0.6690	0.6874	preis werbung	
1	0.6470	0.6568	werbung	
2	0.2503	0.2920	preis besuche	
1	0.2363	0.2575	besuche	
1	-.0011	0.0268	preis	

Es werden alle drei Einflussvariablen benötigt. Die Kleinst-Quadrat-Schätzung lautet:

$$\text{Absatz} = 763,65 - 45.177 \times \text{Preis} + 0,55111 \times \text{Werbung} + 9,7055 \times \text{Besuche}$$

SAS Anweisungen

```
data;
input
absatz preis werbung besuche;
datalines;
2585 12.5 2000 109
1819 10 550 107
1647 9.95 1000 99
1496 11.5 800 70
921 12 0 81
2278 10 1500 102
1810 8 800 110
1987 9 1200 92
1612 9.5 1100 87
1913 12.5 1300 79
2118 8.5 1550 75
1438 12 550 106
1834 9.5 1980 66
1869 9 1600 80
1574 7 500 90
2597 11 2000 120
2026 10 1680 95
2016 9.5 1700 92
1566 10 1400 65
2169 13 1800 90
1996 11 1600 76
2501 8 2000 89
2604 8.5 1800 108
1277 10 460 78
1789 9 800 88
1824 11 1460 87
1813 12 1300 103
1513 11.5 600 89
1172 13 750 68
1987 9 900 106
2056 10.5 1250 96
1513 9 850 78
1756 12.5 950 86
2007 13 1500 125
2079 11 1850 109
1664 9.9 1200 60
1699 12.5 1600 79
;
proc reg;
model absatz=preis werbung besuche/best=8 selection=adjrsq;
run;
```

6.11 Polynomregression

Beispiel: In einem Feldversuch wurde die Wirkung verschiedener Kalkgaben auf den Ertrag von Weizen untersucht (Linder/Berchtold 1982 Statistische Methoden II. Birkhäuser, Basel). Die Kalkgaben (x) lagen zwischen 0 und 8 t/ha. Im folgenden sind die Erträge in dt/ha (y) wiedergegeben.

Kalkgabe (t/ha) x	Ertrag (dt/ha) y
0	44,4
2	54,6
4	63,8
6	65,7
8	68,9

Die fünf Erträge stellen jeweils Mittelwerte über 5 Wiederholungen dar. Der Versuch wurde als lateinisches Quadrat angelegt. Eine adäquate Auswertung dieser Daten berücksichtigt diese Versuchsanlage. Hier soll jedoch die Versuchsanlage nicht berücksichtigt werden. Diese Vereinfachung hat den Zweck, dass die volle Aufmerksamkeit der Modellierung des Zusammenhanges zwischen x und y gewidmet werden kann. Ein Plot der Daten ergibt folgendes Bild:

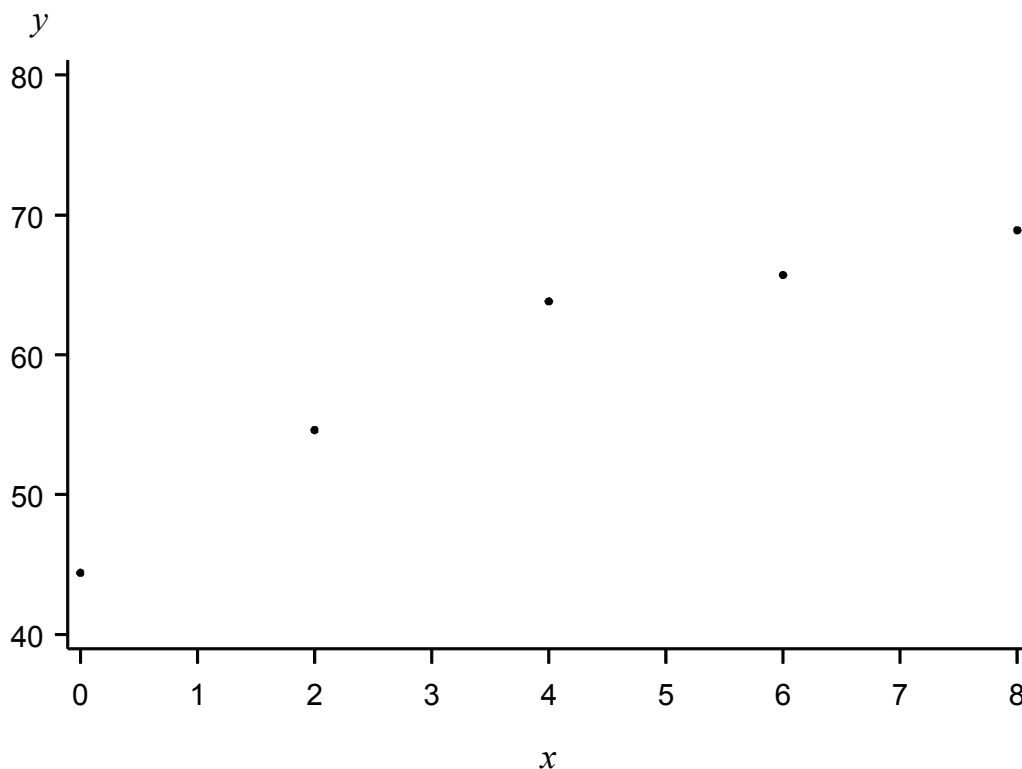


Abb. 6.11.1: Plot der Weizenerträge (y) gegen die Kalkgaben (x).

Es ist offensichtlich, dass der Zusammenhang zwischen Ertrag (y) und Kalkgabe (x) nichtlinear ist. Aus diesem Grunde ist eine nichtlineare Regression angezeigt. Hier soll ein Polynom angepasst werden.

Ein Polynom ist eine Funktion der Form

$$\eta(x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \dots$$

Ein Polynom ist eine Linearkombination von Potenzen der Einflußvariable x , einschließlich $x^0 = 1$ (Dies sieht man, wenn man $\alpha = \beta_0 x^0$ setzt). Polynome sind dazu geeignet, nichtlineare Zusammenhänge zu beschreiben. Der einfachste Fall ist ein quadratisches Polynom

$$\eta(x) = \alpha + \beta_1 x + \beta_2 x^2$$

Im folgenden ist eine quadratische Funktion der Form

$$\eta(x) = 1 + 2x - 0,2x^2$$

wiedergegeben.

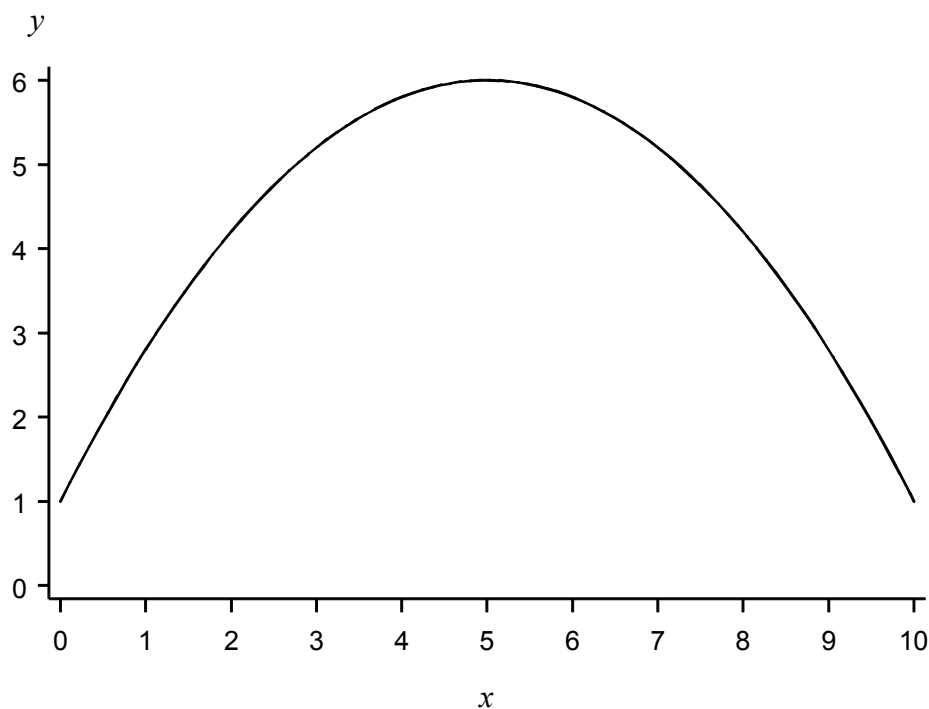


Abb. 6.11.2: Bild der Funktion $\eta(x) = 1 + 2x - 0,2x^2$.

Die Abb. 6.11.2 zeigt, dass ein Polynom zur Beschreibung nichtlinearer Zusammenhänge geeignet ist.

Wie können wir nun an die Kalkdaten eine quadratische Funktion anpassen? Betrachten der Gleichung $\eta(x) = \alpha + \beta_1 x + \beta_2 x^2$ zeigt, dass eine quadratische Funktion als multiple Regression mit zwei Einflussvariablen $x_1 = x$ und $x_2 = x^2$ aufgefaßt werden kann. Und in der Tat kann die Schätzung der Parameter mit Hilfe einer multiplen Regression durchgeführt werden.

Die Design-Matrix und der Daten-Vektor haben die Form

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 8 & 64 \end{pmatrix} \quad y = \begin{pmatrix} 44,4 \\ 54,6 \\ 63,8 \\ 65,7 \\ 68,9 \end{pmatrix}$$

Die Kleinst-Quadrat-Lösung ergibt

$$b = (X^T X)^{-1} X^T y = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 44,42 \\ 6,05 \\ -0,380 \end{pmatrix}$$

Somit ist das geschätzte Modell:

$$\hat{y} = 44,42 + 6,05x - 0,380x^2$$

Diese Funktion zeichnen wir durch die Punktwolke in Abb. 6.11.1 (siehe Abb. 6.11.3).

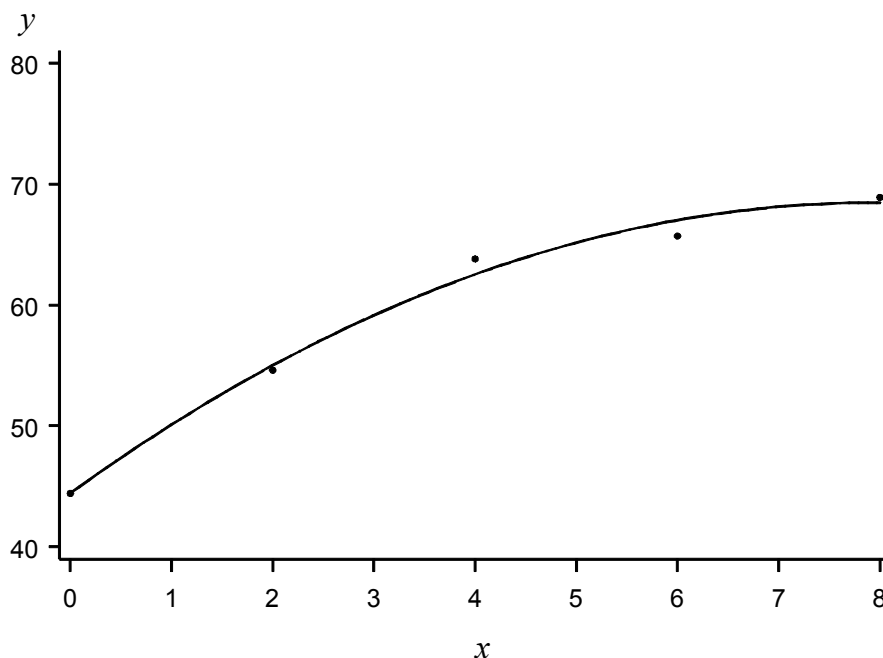


Abb. 6.11.3: Plot der Weizenerträge (y) gegen die Kalkgaben (x) mit angepasster Regressionslinie (quadratisches Polynom $y = 44,42 + 6,05x - 0,38x^2$).

Zum Test des Polynoms führen wir einen sequentiellen Modellaufbau durch, wobei jeder hinzugenommene Term mit dem F-Test aus Abschnitt 6.9 getestet wird. Die Sequenz beginnt mit dem einfachsten Modell, zu dem dann Polynomialterme

wachsenden Grades hinzugefügt werden. Im folgenden ist die Sequenz bis x^2 wiedergegeben.

Modell	FG_{Fehler}	SQ_{Fehler}
(1) $y_i = \alpha + e_i$	$n-1 = 4$	397,31
(2) $y_i = \alpha + \beta_1 x_i + e_i$	$n-2 = 3$	36,11
(3) $y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + e_i$	$n-3 = 2$	3,70

Die Fehler-FG entsprechen der Zahl Beobachtungen (n), abzüglich der Zahl der Spalten der jeweiligen Design-Matrix, also der Zahl der linearen Modellparameter [$n - \text{Rang}(\mathbf{X})$].

Vergleich von Modell (1) und (2) liefert:

$$F_{Vers} = \frac{(SQ_{Fehler}^{red} - SQ_{Fehler}^{voll}) / (FG_{Fehler}^{red} - FG_{Fehler}^{voll})}{SQ_{Fehler}^{voll} / FG_{Fehler}^{voll}} = \frac{(397,31 - 36,11) / (4 - 3)}{36,11 / 3} = 30,01$$

$$F_{Tab} = F(0,95, 1, 3) = 10,13 < F_{Vers}$$

Der lineare Term ist also signifikant.

Der Test des quadratischen Terms erfolgt durch Vergleich der Modelle (2) und (3):

$$F_{Vers} = \frac{(36,11 - 3,70) / (3 - 2)}{3,70 / 2} = 17,51$$

$$F_{Tab} = F(0,95, 1, 2) = 18,51 > F_{Vers}$$

Der quadratische Term ist also gerade nicht signifikant. Insofern würde hier die Anpassung eines linearen Modells gerechtfertigt sein. Eine Auswertung, welche auf den Einzelwerten beruht und die Versuchsanlage (lateinisches Quadrat) berücksichtigt, liefert hier allerdings ein signifikantes Ergebnis (nicht gezeigt). Ein weiterer Test des kubischen Terms x^3 ist nicht signifikant, so dass das quadratische Modell als adäquat gelten kann.

Man beachte, dass wir hier den linearen Term durch Vergleich mit dem reduzierten Modell (1) prüfen. In Analogie zum Test der Terme in der multiplen Regression könnten wir daran denken, stattdessen das volle Modell

$$(3) \quad y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + e_i$$

gegen das reduzierte Modell

$$(2') \quad y_i = \alpha + \beta_2 x_i^2 + e_i$$

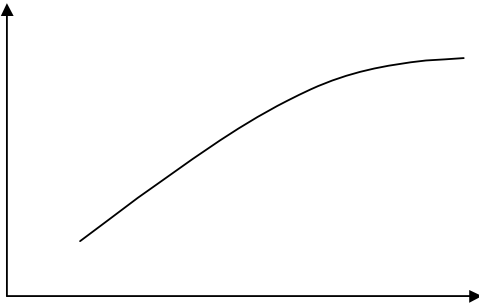
zu prüfen. Dieses reduzierte (quadratische) Modell ist jedoch biologisch nicht sinnvoll, da es bei $x = 0$ ein Minimum/Maximum aufweist, wohingegen wir ein Maximum bei $x \approx 8$ erwarten. Es besteht hier im Gegensatz zur multiplen Regression eine **Hierarchie zwischen den Polynomialtermen**: Ein quadratisches Modell sollte in der Regel auch einen linearen Term haben, so dass das Maximum/Minimum an einem Punkt $x \neq 0$ liegen kann. Daher vergleichen wir zum Test des linearen Terms Modell (2) mit Modell (1), und nicht Modell (3) mit Modell (2'). Ebenso prüft man das kubische Modell

$$(4) y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i$$

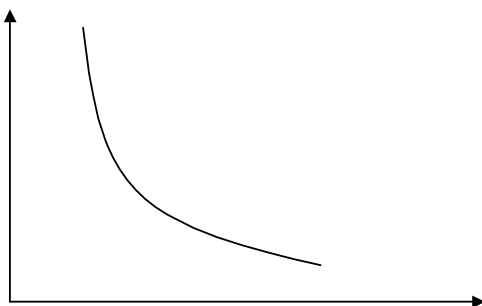
gegen Modell (3). Denn würde man es beispielsweise gegen das Modell

$y_i = \alpha + \beta_1 x_i + \beta_3 x_i^3 + e_i$ prüfen, bei dem der quadratische Term fehlt, so würde dies ein Modell implizieren, bei dem der Wendepunkt bei $x = 0$ liegt (2. Ableitung gleich null!).

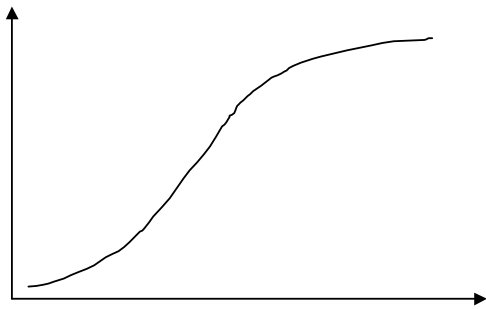
Generell sollte bei der Polynomregression berücksichtigt werden, dass der Grad des Polynoms in der Regel nicht höher als drei sein sollte. Mit Polynomen 2. und 3. Grades lassen sich die häufigsten Kurvenverläufe oft recht gut modellieren, wie die untenstehenden Abbildungen zeigen. Polynome höheren Grades führen in der Regel zu einer Überanpassung an die Daten (Overfitting). Ein Nachteil von Polynomen ist, dass es sich nicht um biologische Modelle handelt, deren Parameter direkt interpretierbar sind. Hinzu kommt das unrealistische Verhalten für x -Werte unterhalb oder oberhalb des beobachteten Wertebereiches: Hier streben die Polynome gegen "plus oder minus Unendlich".



Abnehmender Ertragszuwachs (\Rightarrow Polynom 2. Grades)



Abnehmende Ertragsreduktion (\Rightarrow Polynom 2. Grades)



Sigmoider Verlauf (\Rightarrow Polynom 3. Grades)

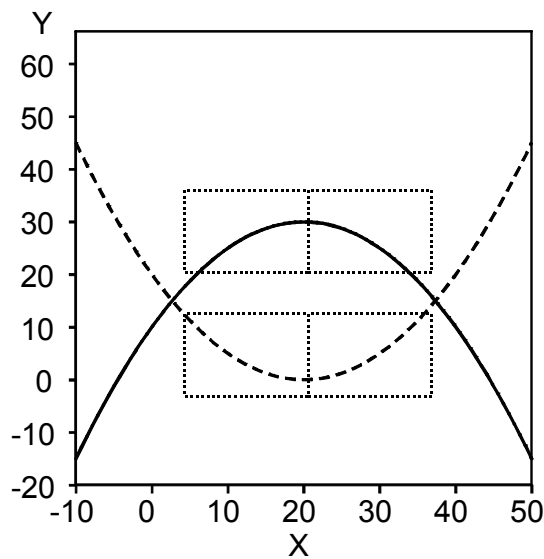


Abb. 6.11.4: Zwei quadratische Polynome. Rechtecke: Geeignete Abschnitte zur Beschreibung nichtlinearer biologischer Zusammenhänge.

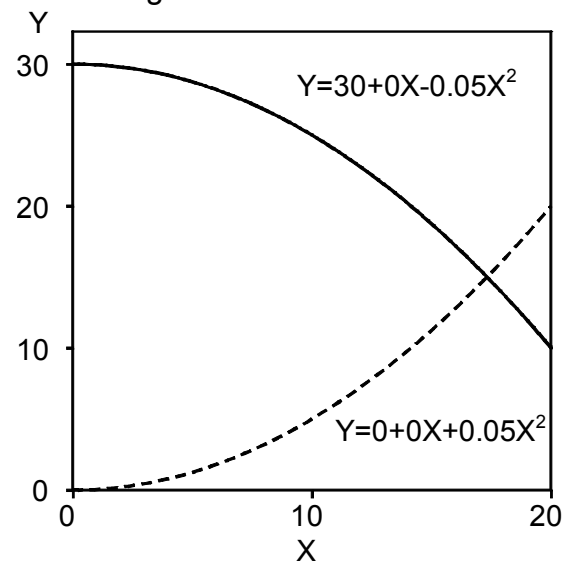
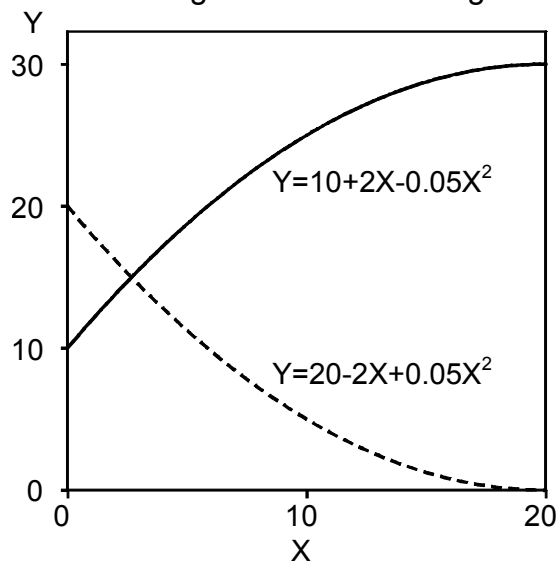


Fig. 6.11.5: Vier quadratische Polynome.

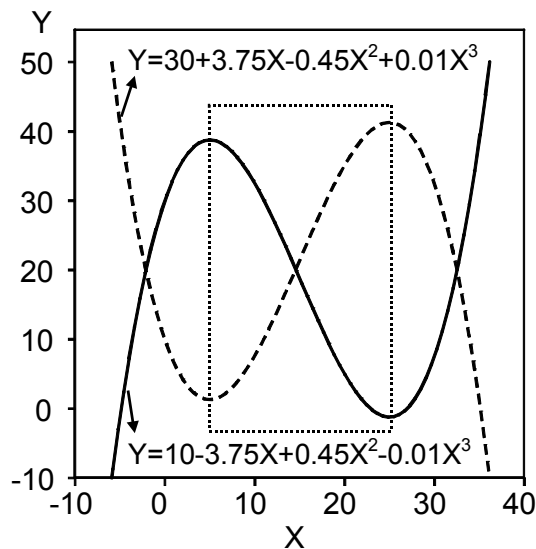


Fig. 6.11.6: Zwei kubische Polynome. Rechteck: Geeignete Abschnitte zur Beschreibung nichtlinearer biologischer Zusammenhänge.

6.12 "Eigentliche" nichtlineare Regression

Bisher haben wir zwei Methoden bzw. Modelle der nichtlinearen Regression kennengelernt:

- Linearisierung durch Transformation der Variablen (Abschnitt 6.4)
- Polynomregression (Abschnitt 6.11)

Beiden Verfahren ist gemeinsam, dass sich die Auswertung auf eine (multiple) lineare Regression zurückführen lässt.

Es gibt Modelle, die weder in die Klasse der Polynome gehören, noch sich durch einfache Transformationen in ein lineares Modell überführen lassen. Bei solchen Modellen ist das Verfahren der **eigentlichen nichtlinearen Regression** anzuwenden.

Beispiel: An die Kalkdaten aus Abschnitt 6.11 lässt sich eine Sättigungskurve der Form

$$f(x) = \eta = \alpha - (\alpha - \beta) \exp(-\gamma x)$$

anpassen, wie in Abb. 6.12.1 gezeigt. Diese Funktion ist auch als **Mitscherlich-Funktion** bekannt. Ein Vergleich mit der Anpassung eines quadratischen Polynoms (Abb. 6.11.3) zeigt, dass beide Anpassungen optisch kaum zu unterscheiden sind. Die Summe der Fehlerquadrate (SQ_{Fehler}) ist mit 3,57 für die Exponentialkurve etwas kleiner als für die quadratische Kurve (3,70).

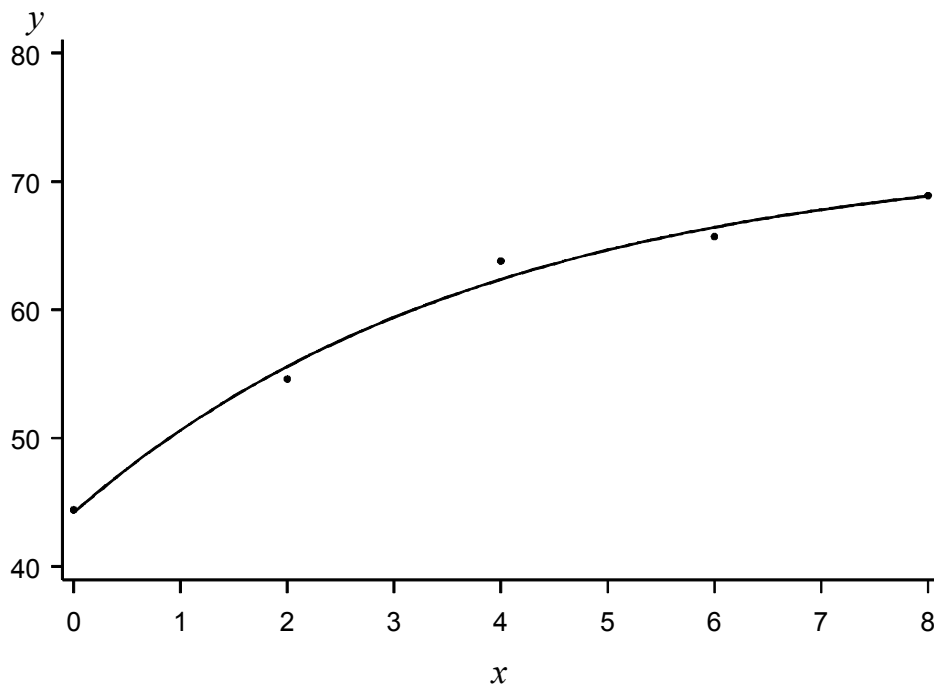


Abb. 6.12.1: Plot der Weizenerträge (y) gegen die Kalkgaben (x) mit angepasster Regressionslinie $\hat{y} = 72,43 - 28,25 \exp(-0,258x)$.

Um das Exponentialmodell anzupassen, ist die Summe der Fehlerquadrate

$$SQ_{Fehler} = \sum_{i=1}^n \{y_i - f(x_i)\}^2 = \sum_{i=1}^n \{y_i - [\alpha + (\alpha - \beta) \exp(-\gamma x_i)]\}^2$$

zu minimieren. Im Gegensatz zur linearen Regression (Polynomregression, multiple Regression) lässt sich dieses Minimierungsproblem nicht explizit lösen, weil das Modell nicht linear in den Parametern, also nicht von der Form $X\beta$ ist. Stattdessen müssen iterative Verfahren verwendet werden, die das SQ_{Fehler} sukzessive in mehreren Iterationsschritten minimieren, wobei die Parameterschätzwerte in jedem Schritt verändert und der Kleinst-Quadrat-Lösung angenähert werden.

6.12.1 Einige "eigentliche" (intrinsisch) nichtlineare Funktionen

Verschiedene nichtlineare Funktionen, die in den Agrarwissenschaften relevant sind, stellen Lösungen einfacher Differentialgleichungen dar, die sich wiederum aus einfachen biologischen, physikalischen oder chemischen Gesetzmäßigkeiten ergeben. Hierfür sollen hier einige Beispiele gegeben werden (siehe auch A. Linder, W. Berchtold 1982 Statistische Methoden II. Varianzanalyse und Regressionsrechnung. Birkhäuser, Basel). Oft steht die Änderung einer Zielvariablen η in Abhängigkeit zum Wert der Zielvariablen :

$$\eta = E(y) = f(x).$$

Die Änderung kann geschrieben werden als die erste Ableitung von η nach x ,

$$\frac{d\eta}{dx}$$

wobei $d\eta$ die Änderung von η und dx die Änderung von x ist. Dies entspricht der Steigung der Funktion $\eta = f(x)$. Wenn die Änderung $d\eta/dx$ proportional zu η ist, wie dies für viele Wachstumsprozesse gilt, so kann man schreiben

$$\frac{d\eta}{dx} = c\eta$$

wobei c ein Proportionalitätsfaktor ist. Die Gleichung ist eine Differentialgleichung, weil in ihr die Funktion $\eta = f(x)$ in Beziehung gesetzt wird zu einer Ableitung derselben Funktion. Man kann nun eine Funktion $\eta = f(x)$ suchen, die diese Differentialgleichung erfüllt. Das Lösen von Differentialgleichungen ist ein eigenes Gebiet der Mathematik, und auf Einzelheiten soll hier nicht eingegangen werden. Eine Lösung der Differentialgleichung ist hier gegeben durch die Exponentialfunktion

$$\eta = f(x) = \alpha \exp(\beta x) \quad (1)$$

Die erste Ableitung dieser Funktion ist

$$\frac{d\eta}{dx} = \alpha \beta \exp(\beta x) = \beta f(x) = \beta \eta$$

was zeigt, dass diese Funktion tatsächlich die Differentialgleichung erfüllt (mit $c = \beta$). In Abb. 6.12.2 ist eine Exponentialfunktion dargestellt.

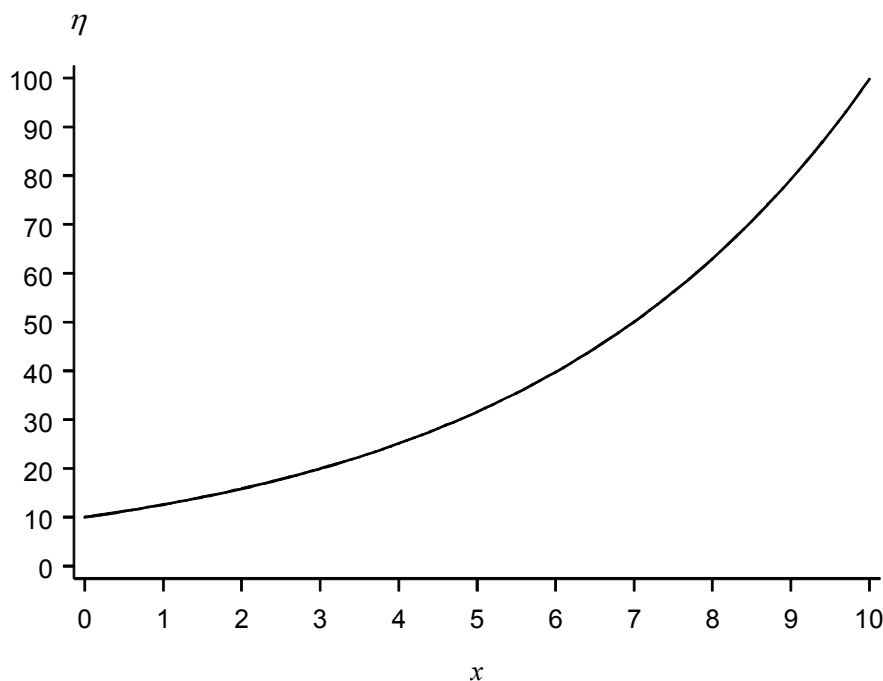


Abb. 6.12.2: Die Funktion $\eta = 10 \exp(0,23x)$.

Viele Wachstumsprozesse können gut durch eine exponentielle Funktion beschrieben werden, wobei x die Zeit ist und y bzw. $\eta = E(y)$ die Biomasse. Exponentielles Wachstum ist meist dann gegeben, wenn ein Sättigungsniveau noch nicht erreicht ist. Dass die Wachstumsrate ohne begrenzende Faktoren proportional zur aktuellen Biomasse η ist, ist biologisch plausibel.

Wenn dagegen ein Sättigungsniveau nahezu erreicht ist, ist die Wachstumsrate eher proportional zum Abstand der Zielvariablen (Biomasse, Länge, etc.) vom Sättigungsniveau. Bezeichnen wir das Sättigungsniveau mit α , so ist die Wachstumsrate dann proportional zu $(\alpha - \eta)$. Dies kann auch durch folgende Differentialgleichung ausgedrückt werden:

$$\frac{d\eta}{dx} = c(\alpha - \eta)$$

wobei c wieder ein Proportionalitätsfaktor ist. Diese Differentialgleichung hat die Lösung

$$\eta = \alpha - (\alpha - \beta) \exp(-\gamma x) \quad (2)$$

Dies ist genau die Funktion, die an die Kalkdaten angepasst wurde (Abb. 6.12.1). Es handelt sich um die Mitscherlich-Funktion.

Oft folgt das Wachstum bei niedriger Biomasse einem exponentiellen Verlauf, während bei hoher Biomasse das Wachstum einer Sättigungskurve folgt. Eine Kombinationen beider Wachstumsverläufe in einer Funktion wird durch folgende Differentialgleichung beschrieben:

$$\frac{d\eta}{dx} = c\eta(\alpha - \eta)$$

Diese Differentialgleichung hat die Lösung

$$\eta = \frac{\alpha}{1 + \beta \exp(-\gamma x)} \quad (3)$$

Die Funktion wird auch als logistische Funktion bezeichnet. Abb. 6.12.3 zeigt einen S-förmigen (sigmoiden) Verlauf dieser Kurve, der typisch ist für viele Wachstumsprozesse.

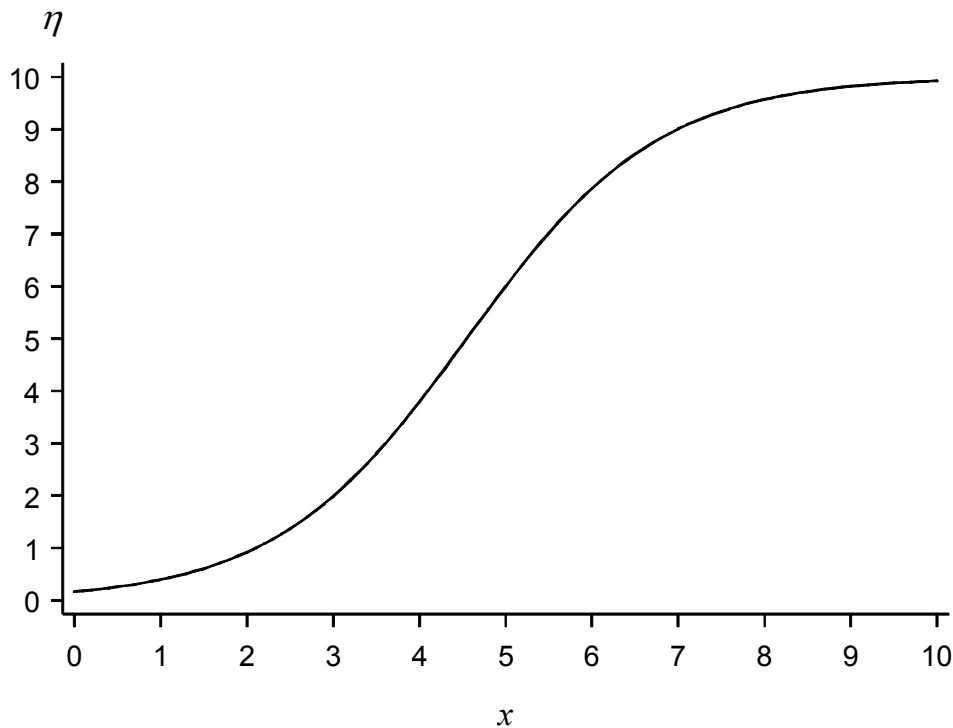


Abb. 6.12.3: Die logistische Funktion $\eta = \frac{10}{1+60\exp(-0,9x)}$.

Bei der logistischen Funktion ist die **relative Wachstumsrate** gegeben durch

$$\frac{d\eta}{dx} \bigg/ \eta = c(\alpha - \eta)$$

Dies ist eine lineare Funktion von η . Bei der Gompertz-Funktion ist dagegen die relative Wachstumsrate proportional zu $\log(\eta)$:

$$\frac{d\eta}{dx} \bigg/ \eta = c[\log(\alpha) - \log(\eta)]$$

Die sich ergebende Differentialgleichung hat die Lösung

$$\eta = \alpha \exp\{-\exp[-\beta(x-\gamma)]\} \quad (4)$$

Weitere sehr gebräuchliche nichtlineare Funktionen mit sigmoidem Verlauf, die sich aus einer Differentialgleichung ergeben, sind die Richards-Funktion und die Gompertz-Funktion (Seber & Wild, 1989). Darüber hinaus gibt es viele andere intrinsisch nichtlineare Modelle, also Modelle, die sich nicht linearisieren lassen. Unter diesen Modellen sind viele, die nicht aus Differentialgleichungen abgeleitet werden können (Seber & Wild, 1989). Auf weitere Modelle soll hier jedoch nicht eingegangen werden.

6.12.2 Schätzen der Parameter

Das Newton-Verfahren (Newton-Raphson-Verfahren)

wie im linearen Fall berechnen wir bei der eigentlichen nichtlinearen Regression zur Bestimmung der Kleinst-Quadrat-Lösung die Normalgleichungen, also die 1. Ableitung des SQ_{Fehler} nach den Parametern, welche gleich Null gesetzt wird. Allerdings hat die Normalgleichung im nichtlinearen Fall keine explizite Lösung. Vielmehr muss die Kleinst-Quadrat-Lösung durch ein iteratives Verfahren bestimmt werden. Insbesondere muss die **Nullstelle** der Normalgleichung bestimmt werden. Ein allgemeines Verfahren zur iterativen Bestimmung der Nullstelle einer Funktion ist das **Newton-Verfahren**. Zur Einführung wenden wir hier das Newton-Verfahren beispielhaft zur Bestimmung der Nullstelle einer einfachen nichtlinearen Funktion an: $f(\theta) = \theta - \exp(-5\theta)$. Wir bestimmen also denjenigen Wert von θ , für den $f(\theta) = 0$ ist. Danach wenden wir das Verfahren zur Lösung der Normalgleichungen an, und damit zur Minimierung der Fehlerquadratsumme SQ_{Fehler} , und zwar an einem Beispiel.

Beispiel: Auffinden der Nullstelle einer Funktion $f(\theta)$

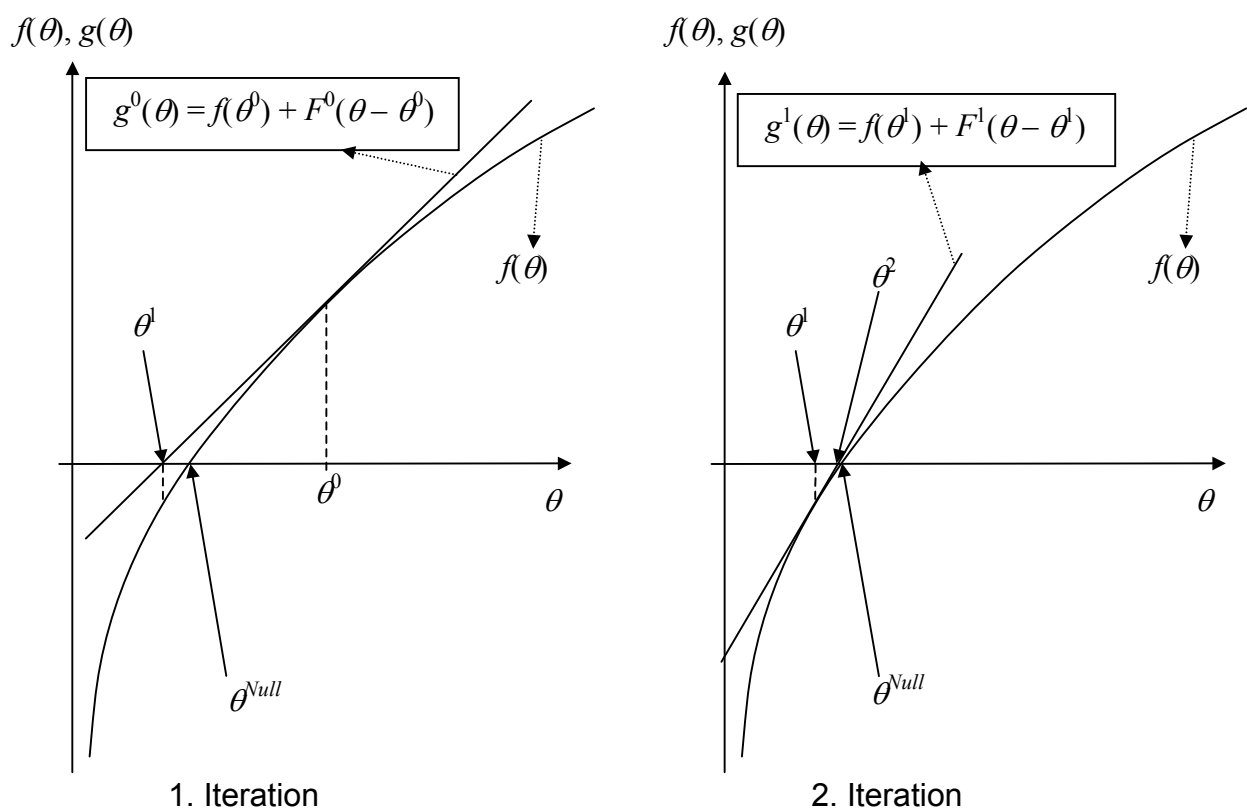


Abb.6.12.4: Schematische Darstellung des Newton-Verfahrens. $F^k = \left. \frac{\partial f(\theta)}{\partial \theta} \right|_{\theta^k}$.

Die Idee des Newton-Verfahrens ist es, die nichtlineare Funktion $f(\theta)$ durch eine lineare Funktion $g(\theta)$ zu approximieren, deren Nullstelle explizit gefunden werden kann. Das Verfahren approximiert die Funktion in der Nähe der Nullstelle durch die Tangente an die Funktion und bestimmt dann die Nullstelle der Tangente als erste Approximation der Nullstelle der nichtlinearen Funktion. An dieser ersten Approximation der Nullstelle wird dann wiederum die Tangente gelegt und deren

Nullstelle bestimmt (2. Approximation). Das Verfahren wird so oft wiederholt, bis sich keine wesentliche Änderung der Approximation der Nullstelle mehr ergibt.

Das Newton-Verfahren ist in Abb. 6.12.4 näher erläutert. Man bestimmt zunächst einen groben ersten Schätzwert θ^0 der Nullstelle θ^{Null} der Funktion $f(\theta)$. Die Funktion $f(\theta)$ wird dann durch die **Tangente** an der Stelle θ^0 approximiert. Die Tangente hat die Funktionsgleichung

$$g(\theta) = f(\theta^0) + F^0(\theta - \theta^0)$$

wobei

$$F^0 = \left. \frac{\partial f(\theta)}{\partial \theta} \right|_{\theta^0} \text{ ist.}$$

F^0 ist die partielle Ableitung von $f(\theta)$ an der Stelle θ^0 , also die Steigung der Tangente an der Stelle θ^0 . Der Achsenabschnitt der Tangente ist $f(\theta^0) - F^0 \theta^0$. Die Tangente an die Funktion $f(\theta)$ approximiert die Funktion in der Nähe von θ^0 . Man spricht in diesem Zusammenhang auch von einer **Taylor-Reihen-Approximation** 1. Ordnung der Funktion $f(\theta)$ an der Stelle $\theta = \theta^0$.

Der Schnittpunkt der Tangente $g(\theta)$ mit der Abszisse (θ -Achse) liefert den neuen Wert θ^1 , der näher an der tatsächlichen Nullstelle liegt als θ^0 . Die neue Approximation der Nullstelle θ^1 lässt sich explizit berechnen durch Lösen von

$$g(\theta^1) = f(\theta^0) + F^0(\theta^1 - \theta^0) = 0$$

\Leftrightarrow

$$\theta^1 = \theta^0 - (F^0)^{-1} f(\theta^0) .$$

Die Berechnung von θ^1 beendet den 1. **Iterationsschritt**. Mit dem neuen Wert θ^1 wird das Verfahren wiederholt, und in der 2. **Iteration** ein neuer Wert θ^2 erhalten, der noch näher an der wahren Nullstelle liegt, usw. Im k -ten Iterationsschritt wird die Parameterschätzung wie folgt adjustiert:

$$\theta^{k+1} = \theta^k - (F^k)^{-1} f(\theta^k) ,$$

wobei $f(\theta^k)$ und F^k der Funktionswert sowie die 1. Ableitung an der Stelle $\theta = \theta^k$ sind. Das Verfahren wird so oft wiederholt, bis sich keine wesentliche Änderung von θ^k mehr ergibt. Das iterative Verfahren kann wie folgt kompakt beschrieben werden:

1. Bestimme eine Grobschätzung für die Nullstelle der Funktion, also einen Wert von θ , der möglichst nahe an der Nullstellen liegt. Setze $k = 0$.
2. Berechne $\theta^{k+1} = \theta^k - (F^k)^{-1} f(\theta^k)$
3. Wenn θ^{k+1} sich im letzten Schritt kaum noch geändert hat, beende das Verfahren. Andernfalls gehe zurück zu 2.

Wenden wir dies nun explizit auf die Funktion $f(\theta) = \theta - \exp(-5\theta)$ an. Zunächst berechnen wir die 1. Ableitung:

$$\frac{\partial f(\theta)}{\partial \theta} = 1 + 5 \exp(-5\theta).$$

Die Iterationen starten wir mit dem Wert $\theta = 0.3$, der relativ nahe an der gesuchten Nullstelle liegt. Diesen Wert bestimmen wir durch Ausprobieren oder durch Zeichnen der Funktion.

k	θ^k	$f(\theta^k)$	F^k	$\theta^{k+1} = \theta^k - (F^k)^{-1}f(\theta^k)$
1	0.30000	0.076870	2.11565	0.26367
2	0.26367	-0.003916	2.33791	0.26534
3	0.26534	-0.000009	2.32675	0.26534
4	0.26534	-0.000000	2.32672	0.26534
5	0.26534	0.000000	2.32672	0.26534

Nach 5 Iterationen ändert sich der Wert von θ^k kaum noch, so dass wir das Verfahren abbrechen und die Lösung $\theta = 0.26534$ verwenden.

Anwendung des Newton-Verfahrens zur Minimierung des SQ_{Fehler}

Wir haben das Newton-Verfahren soeben zur Bestimmung einer Nullstelle betrachtet. Man kann dieses Verfahren ebenso zur Bestimmung des Maximums oder des Minimums einer Funktion benutzen (hier: Minimierung von SQ_{Fehler}). Hierzu kann man die Tatsache nutzen, dass die 1. Ableitung einer Funktion beim Maximum/Minimum gleich Null sein muss. Also kann man das obige Verfahren auf die 1. Ableitung des SQ_{Fehler} anwenden, um dessen Minimum zu finden. Dies soll wiederum an einem einfachen Beispiel erläutert werden.

Beispiel: Angenommen, es liegen folgende Daten vor:

x	y
1	2
2	5
3	10

An diese soll das Modell

$$\eta = \eta(x, \theta) = \exp(\theta x)$$

angepasst werden. Das Modell hat hier den Parameter θ . Das SQ_{Fehler} für das Modell lautet:

$$\begin{aligned} SQ_{Fehler} &= [y_1 - \exp(\theta x_1)]^2 + [y_2 - \exp(\theta x_2)]^2 + [y_3 - \exp(\theta x_3)]^2 \\ &= [2 - \exp(\theta)]^2 + [5 - \exp(2\theta)]^2 + [10 - \exp(3\theta)]^2 \end{aligned}$$

Diese Funktion ist in Abb. 6.12.5 dargestellt. Etwa bei $\theta = 0,8$ liegt das Minimum von SQ_{Fehler} . Der genaue Wert von θ , welcher das SQ_{Fehler} minimiert, ist die gesuchte Kleinst-Quadrat-Lösung, die numerisch zu bestimmen ist. Übrigens wird das SQ_{Fehler} beim Minimum nicht exakt Null, sondern nimmt einen zwar kleinen, aber positiven Wert an.

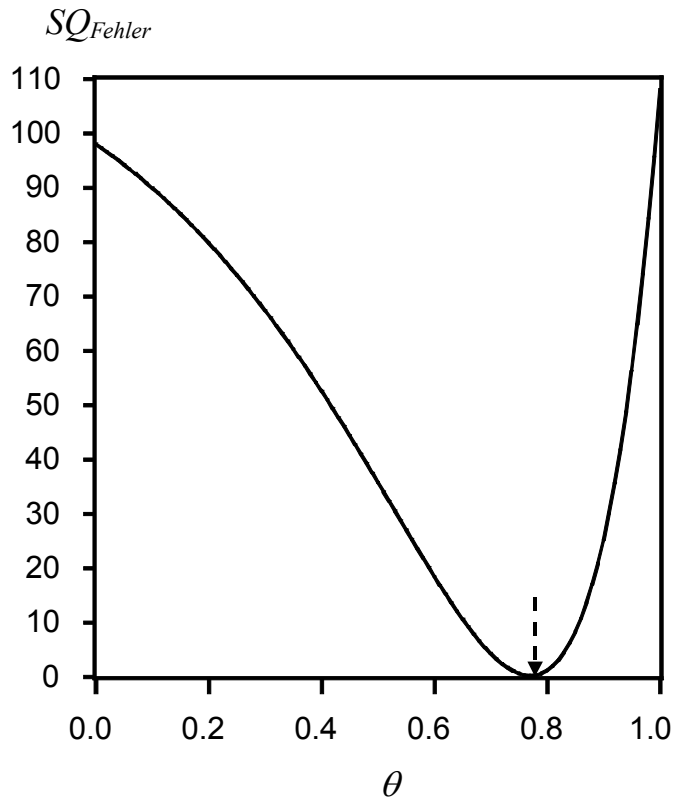


Abb. 6.12.5: Plot von SQ_{Fehler} gegen θ . Minimum bei $\theta = 0,7705$.

Die erste Ableitung nach θ lautet nach etwas Umformung:

$$s = \partial SQ_{Fehler} / \partial \theta = -4 \cdot \exp(\theta) + 2 \cdot \exp(2\theta) \\ - 20 \cdot \exp(2\theta) + 4 \cdot \exp(4\theta) \\ - 60 \cdot \exp(3\theta) + 6 \cdot \exp(6\theta)$$

(Achtung: s steht hier nicht für Standardabweichung). Um SQ_{Fehler} zu minimieren, müssen wir nun die Nullstelle der 1. Ableitung von SQ_{Fehler} , also die Nullstelle von s , finden; denn am Minimum muss gelten:

$$s = \partial SQ_{Fehler} / \partial \theta = 0 \quad .$$

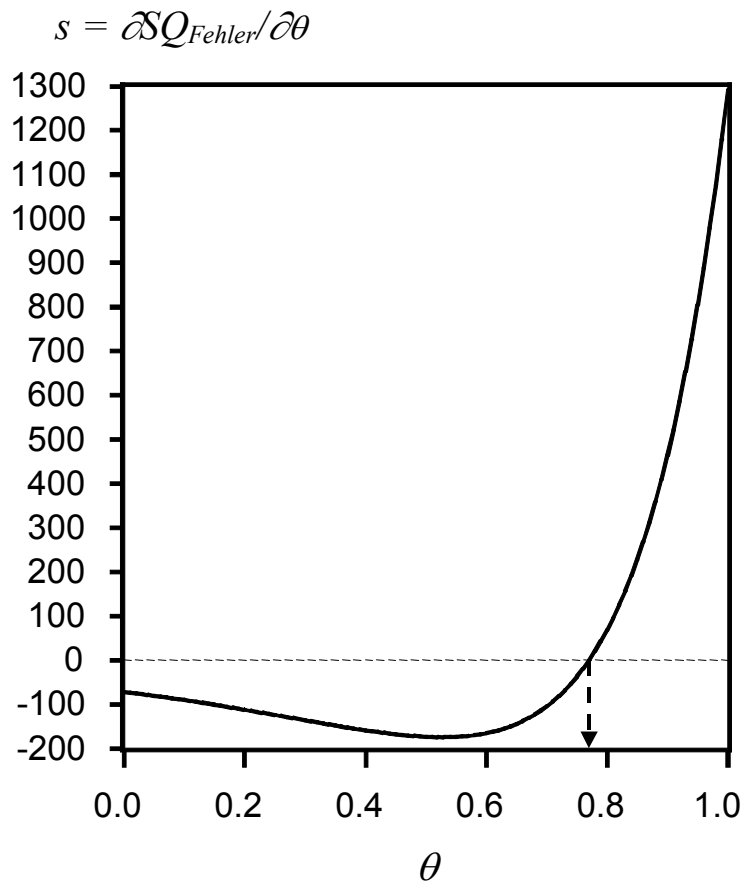


Abb. 6.12.6: Plot von $s = \partial SQ_{Fehler} / \partial \theta$ gegen θ . Nullstelle bei $\theta = 0,7705$.

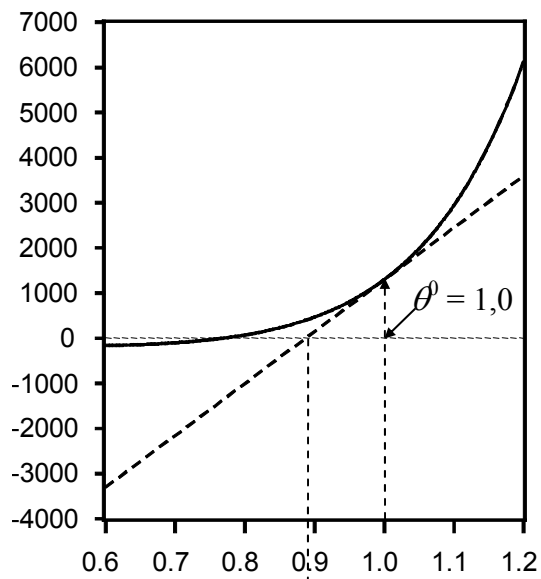
Die 1. Ableitung ist in Abb. 6.12.6 dargestellt. Man sieht, dass hier bei $\theta = 0,7705$, also beim Minimum von SQ_{Fehler} , eine Nullstelle vorliegt. Um nun das Newton-Verfahren zur Bestimmung der Nullstelle von s anzuwenden, benötigen wir die 1. Ableitung von s , was der 2. Ableitung von SQ_{Fehler} entspricht:

$$H = \partial s / \partial \theta = \partial^2 SQ_{Fehler} / \partial \theta^2 = \begin{aligned} & -4 \cdot \exp(\theta) + 4 \cdot \exp(2\theta) \\ & -40 \cdot \exp(2\theta) + 16 \cdot \exp(4\theta) \\ & -180 \cdot \exp(3\theta) + 36 \cdot \exp(6\theta) \end{aligned}$$

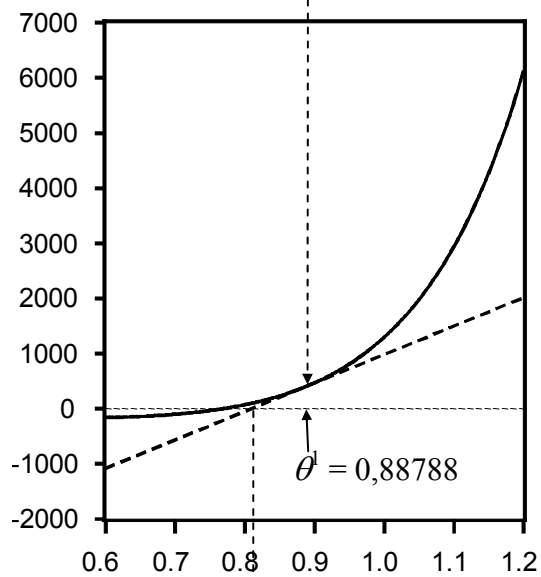
Nach dem Newton-Verfahren berechnen wir den $(k+1)$ -ten Schätzwert aus dem k -ten Schätzwert dann nach

$$\theta^{k+1} = \theta^k - (H^k)^{-1} s(\theta^k)$$

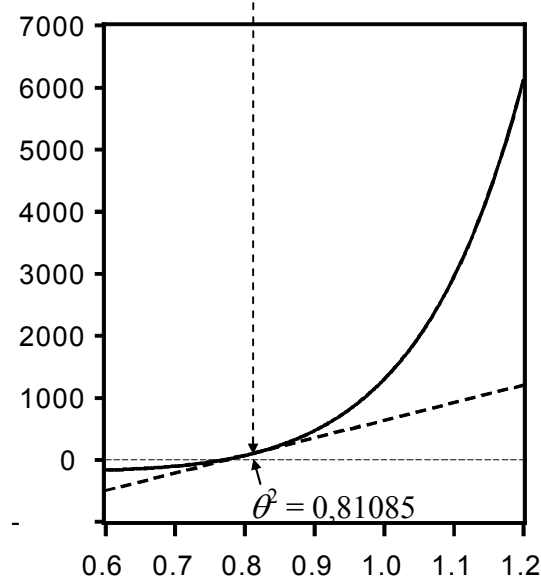
Wir beginnen hier mit dem Startwert $\theta^0 = 1,0$. Die Iterationsschritte sind in Tab. 6.12.1 sowie in Abb. 6.12.7 (Schritte 1 bis 3) dargestellt.



1. Iteration



2. Iteration



3. Iteration

Abb. 6.12.7: Die ersten drei Iterationsschritte.

Tab. 6.12.1: Iterationsschritte zur Schätzung des Parameters θ mit dem Newton-Verfahren.

k	θ^k (alt)	SQ_{Fehler}	$s(\theta^k)$	H^k	$\theta^{k+1} = \theta^k - (H^k)^{-1}s(\theta^k)$
0	1,00000	107,942	1289,96	11504,73	0,88788
1	0,88788	19,911	397,80	5164,28	0,81085
2	0,81085	1,992	97,20	2837,45	0,77659
3	0,77659	0,181	12,58	2130,18	0,77069
4	0,77069	0,143	0,32	2024,02	0,77053
5	0,77053	0,143	0,00	2021,27	0,77053

Wir sehen, dass das SQ_{Fehler} mit jedem Schritt sinkt, anfangs sehr deutlich, und dann in immer kleineren Schritten. Von der 4-ten zur 5-ten Iteration ändert sich der Schätzwert kaum noch; man hat hier Konvergenz erzielt. Die Steigung der Tangente (H^k) wird mit jedem Schritt ebenfalls kleiner.

Im oben beschriebenen Newton-Verfahren wird der Parameterschätzwert im k -ten Schritt nach der Formel

$$\theta^{k+1} = \theta^k - (H^k)^{-1}s(\theta^k)$$

aktualisiert. Dies kann auch wie folgt ausgedrückt werden:

$$\hat{\theta}^{k+1} = \hat{\theta}^k + \delta_{NR}^k$$

wobei

$$\delta_{NR}^k = -(H^k)^{-1}s(\theta^k)$$

die Adjustierung des k -ten Schätzwertes ist. Im obigen Beispiel hat das Modell $\eta = \exp(\theta x)$ nur einen Parameter.

Man kann das Verfahren auch auf Modelle mit mehr als einem Parameter anwenden. Hierzu benötigt man alle 1. Ableitungen des SQ_{Fehler} nach den verschiedenen Parametern, die man in einen Vektor s^k schreibt. Die 2. Ableitungen schreibt man in eine Matrix H^k , die sog. **Hesse-Matrix**. Der k -te Iterationsschritt in Matrizenschreibweise lautet dann

$$\hat{\theta}^{k+1} = \hat{\theta}^k + \delta_{NR}^k,$$

wobei θ der Parametervektor ist, $\delta_{NR}^k = -(H^k)^{-1}s(\theta^k)$ und $(H^k)^{-1}$ die Inverse von H^k .

Das Verfahren endet, wenn δ^k sehr klein wird, es also kaum noch zu einer merklichen Adjustierung kommt. Alternativ sind verschiedene andere Konvergenzkriterien gebräuchlich. Die SAS Prozedur NLIN verwendet folgendes Kriterium:

$$\sqrt{\frac{(n-p)}{p} \mathbf{r}^k{}' \mathbf{F}^k (\mathbf{F}^k{}' \mathbf{F}^k)^{-1} \mathbf{F}^k{}' \mathbf{r}^k}$$

wobei \mathbf{F}_k eine Matrix mit den partiellen Ableitung der angepassten Funktion nach den Parametern ist (siehe unten). Die Iterationen werden beendet, wenn das Kriterium kleiner als 10^{-5} wird. Das Verfahren konvergiert in der Regel zu einem lokalen Minimum von SQ_{Fehler} ; Konvergenz ist allerdings nicht garantiert.

Das hier beschriebene Verfahren wird auch als **Newton-Raphson Verfahren** bezeichnet.

Das Gauß-Newton-Verfahren

Wir kommen nun zu einem zweiten wichtigen Iterations-Verfahren, das ähnlich funktioniert wie das Newton-Raphson-Verfahren, aber eine etwas andere Motivation hat, die nun beschrieben wird. Bei der nichtlinearen wie der linearen Regression ist die zu minimierende Funktion die Summe der Fehler-Quadrate (SQ_{Fehler}). Für das lineare Regressionsmodell

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

bzw.

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i$$

wobei \mathbf{x}_i' die i -te Zeile der Design-Matrix \mathbf{X} ist, minimieren wir

$$SQ_{Fehler} = \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Dieses Minimierungsproblem hat eine explizite Lösung, nämlich $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$, weil das Modell, und somit auch die 1. Ableitung von SQ_{Fehler} nach $\boldsymbol{\beta}$, linear in den Parametern ist.

Das nichtlineare Regressionsmodell kann allgemein geschrieben werden als

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + e_i$$

wobei $\boldsymbol{\theta}$ der Parametervektor ist. In kompakter Form lautet das Modell

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{e}$$

mit

$$f(\mathbf{x}, \boldsymbol{\theta}) = \begin{bmatrix} f(\mathbf{x}_1, \boldsymbol{\theta}) \\ f(\mathbf{x}_2, \boldsymbol{\theta}) \\ \vdots \\ f(\mathbf{x}_n, \boldsymbol{\theta}) \end{bmatrix}$$

Das zu minimierende SQ_{Fehler} ist:

$$SQ_{Fehler} = \sum_{i=1}^n [y_i - f(\mathbf{x}'_i, \boldsymbol{\theta})]^2 = [\mathbf{y} - \mathbf{f}(\mathbf{x}, \boldsymbol{\theta})]' [\mathbf{y} - \mathbf{f}(\mathbf{x}, \boldsymbol{\theta})]$$

Wir suchen also denjenigen Wert des Parametervektors $\boldsymbol{\theta}$, der das SQ_{Fehler} minimiert. Das Problem besteht nun darin, dass $f(\mathbf{x}, \boldsymbol{\theta})$ nicht linear in den Parametern $\boldsymbol{\theta}$ ist wie bei der linearen Regression. Daher hat das Minimierungsproblem keine explizite Lösung. Wenn es nun gelingt, $f(\mathbf{x}, \boldsymbol{\theta})$ in der Nähe der Lösung durch eine Funktion zu approximieren, die linear in $\boldsymbol{\theta}$ ist, dann kann eine approximative Kleinst-Quadrat-Lösung gefunden werden. Eine Möglichkeit der Approximation besteht in der Verwendung einer Taylorreihen-Entwicklung um einen groben Schätzwert $\boldsymbol{\theta}^0$ der Kleinst-Quadrat-Lösung:

$$f(\mathbf{x}, \boldsymbol{\theta}) \approx f(\mathbf{x}, \boldsymbol{\theta}^0) + \mathbf{F}^0 (\boldsymbol{\theta} - \boldsymbol{\theta}^0)$$

wobei

$$\mathbf{F}^0 = \left. \frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}^0} \text{ ist.}$$

Die Matrix \mathbf{F}^0 enthält die partiellen Ableitungen der Funktion $f(\mathbf{x}, \boldsymbol{\theta})$ nach den Parametern an der Stelle $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. Betrachten wir nun die Abweichung $\mathbf{y} - f(\mathbf{x}, \boldsymbol{\theta})$ im Ausdruck für SQ_{Fehler} und setzen die Taylor-Approximation für $f(\mathbf{x}, \boldsymbol{\theta})$ ein, so erhalten wir:

$$\mathbf{y} - f(\mathbf{x}, \boldsymbol{\theta}) \approx \mathbf{y} - [f(\mathbf{x}, \boldsymbol{\theta}^0) + \mathbf{F}^0 (\boldsymbol{\theta} - \boldsymbol{\theta}^0)] = \underbrace{\mathbf{y} - f(\mathbf{x}, \boldsymbol{\theta}^0) + \mathbf{F}^0 \boldsymbol{\theta}^0}_{\mathbf{z}^0} - \mathbf{F}^0 \boldsymbol{\theta}$$

Diese Gleichung zeigt, dass die approximative Kleinst-Quadrat-Lösung berechnet werden kann, indem für **Pseudodaten** \mathbf{z}^0 das lineare Modell $\mathbf{F}^0 \boldsymbol{\theta}$ angepasst wird ($\mathbf{F}^0 \boldsymbol{\theta}$ ist linear in $\boldsymbol{\theta}$). Nach einem **Iterations-Schritt** erhalten wir

$$\hat{\boldsymbol{\theta}}^1 = \left(\mathbf{F}^{0'} \mathbf{F}^0 \right)^{-1} \mathbf{F}^{0'} \mathbf{z}^0$$

Nach $k+1$ Iterationen erhalten wir

$$\hat{\theta}^{k+1} = \left(F^k{}' F^k \right)^{-1} F^k{}' z^k$$

wobei

$$z^k = y - f(x, \theta^k) + F^k \theta^k$$

Dies kann auch wie folgt umgeformt werden:

$$\hat{\theta}^{k+1} = \hat{\theta}^k + \left(F^k{}' F^k \right)^{-1} F^k{}' r^k$$

wobei

$$r^k = y - f(x, \theta^k)$$

Im $(k+1)$ -ten Iterations-Schritt wird also der aktuelle Parameterwert um die Differenz

$$\delta_{GN}^k = \left(F^k{}' F^k \right)^{-1} F^k{}' r^k$$

adjustiert. Wir können den $(k+1)$ -ten Iterationsschritt auch schreiben als:

$$\hat{\theta}^{k+1} = \hat{\theta}^k + \delta_{GN}^k$$

Das Verfahren endet, wenn δ^k sehr klein wird, es also kaum noch zu einer merklichen Adjustierung kommt. Das hier beschriebene Verfahren heißt **Gauß-Newton**-Verfahren, und es ist z.B. die voreingestellte Methode der Prozedur NLIN.

Weder das Gauß-Newton- noch das Newton-Raphson-Verfahren konvergieren mit Sicherheit zur Kleinst-Quadrat-Lösung. Dies gilt vor allem für das Newton-Raphson-Verfahren und für den Fall, dass F^k schlecht konditioniert ist, das heißt, wenn es in den Spalten von F^k eine nahezu lineare Abhängigkeiten gibt. Das Problem ist analog dem der Multikollinearität in der multiplen linearen Regression, denn F^k hat bei der nichtlinearen Regression eine ähnliche Funktion wie die Designmatrix X in der linearen Regression. Die Konvergenz hängt entscheidend von der Wahl der Startwerte θ^0 ab. Außerdem kann eine Reparametrisierung des Modells die Konditionierung der Matrix F^k oft so verbessern, dass eine Konvergenz leichter zu erzielen ist (Ratkowski, 1983; Seber und Wild 1989, § 3.4; Schabenberger und Pierce, 2000).

6.12.3 Startwerte

Die iterative Berechnung der Kleinst-Quadrat-Lösung mittels Gauß-Newton- bzw. Newton-Raphson-Verfahren wird man in der Praxis einem Statistik-Paket überlassen. Wir verwenden hier die NLIN Prozedur von SAS. Man braucht in NLIN ab der SAS

Version 8 die ersten Ableitungen der Funktion nicht mehr anzugeben. Die Prozedur berechnet diese Ableitungen numerisch. Die für den Anwender einzig relevante zusätzliche Arbeit gegenüber der linearen Regression besteht in der Bestimmung geeigneter Startwerte für die Parameter. Hierfür gibt es keine allgemeingültigen Regeln, sondern es ist etwas Intuition gefordert, wie die folgenden beiden Beispiele zeigen.

Beispiel: An die Kalkdaten

x	y
0	44,4
2	54,6
4	63,8
6	65,7
8	68,9

soll die Mitscherlich-Funktion

$$f(x) = \alpha - (\alpha - \beta) \exp(-\gamma x) \quad (\gamma > 0)$$

angepasst werden. Die Funktion hat eine Asymptote bei

$$f(\infty) = \alpha \quad (\infty = \text{"Unendlich"})$$

Eine graphische Betrachtung der Daten deutet auf eine Asymptote bei etwa 70 dt/ha hin. Also wählen wir den Startwert $\alpha = 70$. Der Ertrag bei $x = 0$ ist

$$f(0) = \beta$$

Der beobachtete Ertrag bei $x = 0$ ist 44,4 dt/ha, also wählen wir $\beta = 44,4$ als Startwert. Hiermit ist das vorläufige Modell

$$f(x) = 70 - 25,6 \exp(-\gamma x)$$

Einen Schätzwert für γ können wir durch einen dritten x -Wert erhalten, bei dem ein Ertrag gemessen wurde, z.B. $y = 63,8$ bei $x = 4$. Wir finden

$$63,8 = 70 - 25,6 \exp(-4\gamma)$$

\Leftrightarrow

$$6,2/25,6 = \exp(-4\gamma) \Leftrightarrow \gamma = -\log(6,2/25,6)/4 = 0,3545$$

Die Startwerte sind also:

$$\alpha = 70$$

$$\beta = 44,4$$

$$\gamma = 0,3545$$

Zur Verrechnung mit SAS verwenden wir die folgenden Anweisungen:

```
data;
input x y;
datalines;
0 44.4
2 54.6
4 63.8
6 65.7
8 68.9
;
proc nlin;
model y=alpha-(alpha-beta)*exp(-gamma*x);
parms alpha=70 beta=44.4 gamma=0.3545;
run;
```

Ergebnis:

Iter	alpha	beta	gamma	Sum of Squares
0	70.0000	44.4000	0.3545	9.5678
1	71.7890	44.2346	0.2442	6.8972
2	72.3656	44.1757	0.2593	3.5697
3	72.4350	44.1812	0.2580	3.5688
4	72.4325	44.1807	0.2581	3.5688
5	72.4326	44.1808	0.2581	3.5688

NOTE: Convergence criterion met.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	3	18083.1	6027.7	110.33	0.0090
Residual	2	3.5688	1.7844		
Uncorrected Total	5	18086.7			
Corrected Total	4	397.3			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
alpha	72.4326	3.1864	58.7225	86.1428
beta	44.1808	1.3134	38.5296	49.8319
gamma	0.2581	0.0679	-0.0339	0.5501

Hier wird als Voreinstellung das Gauß-Newton-Verfahren verwendet. Die „Iterations-Geschichte“ findet sich am Beginn des Output. Für die Startwerte $\alpha^0 = 70$, $\beta^0 = 44,4$

und $\gamma^0 = 0,3545$ beträgt die Summe der Abweichungsquadrate $SQ_{Fehler} = 9,5678$. Nach einem Iterationsschritt haben die Parameter die Werte $\alpha^1 = 71,7890$, $\beta^1 = 44,2346$ und $\gamma^1 = 0,2442$. Das SQ_{Fehler} sinkt auf einen Wert von 6,8972. Nach 5 Iterationen konvergiert der Gauß-Newton-Algorithmus und man erhält die Lösungen $\alpha^5 = 72,4326$, $\beta^5 = 44,1808$ und $\gamma^5 = 0,2581$. Das SQ_{Fehler} sinkt in jedem Schritt. Dass das Programm tatsächlich konvergiert hat, sehen wir zum einen daran, dass sich die Parameterwerte im letzten Schritt numerisch kaum mehr verändert haben. Zum anderen bekommen wir die Meldung **NOTE: Convergence criterion met**. Die geschätzte Kurve hat folgende Form:

$$\eta = 72,43 - 28,25 \exp(-0,258x)$$

Die Kurve ist in Abb. 6.12.1 wiedergegeben. Ein Vergleich mit der Anpassung einer quadratischen Gleichung (Abb. 6.11.3) zeigt, dass beide Anpassungen optisch kaum zu unterscheiden sind. Die Summe der Abweichungsquadrate (SQ_{Fehler}) ist mit 3,57 für die Exponentialkurve etwas kleiner als für die quadratische Kurve ($SQ_{Fehler} = 3,70$).

Für die Schätzwerte der Kurve werden auch asymptotische Standardfehler und 95%-Vertrauensintervalle angegeben (siehe 6.12.5). So können wir beispielsweise sagen, dass der wahre Wert für α , der durch $\hat{\alpha} = 72,43$ geschätzt wird, mit 95%iger Wahrscheinlichkeit vom den Grenzen 58,7 und 86,1 eingeschlossen wird. Die Lösung $\hat{\alpha} = 72,43$ deutet an, dass wir durch eine weitere Steigerung der Kalkgabe den Ertrag wahrscheinlich nicht über 72 dt/ha steigern können. Voraussetzung für diese Aussage ist allerdings, dass unsere Funktion den wahren Ertragsverlauf auch über den Bereich der betrachteten Kalkgaben hinaus realistisch beschreibt. Die getroffene Aussage bedeutet eine Extrapolation, und Extrapolationen sind immer mit Vorsicht zu genießen.

Beispiel: Eine Untersuchung zum Nachwuchs von Grünland nach einem Schnitt (y) in Abhängigkeit von der Zeit (x) lieferte folgendes Ergebnis (Ratkowsky, 1983).

x	y
9	8,93
14	10,80
21	18,59
28	22,33
42	39,35
57	56,11
63	61,73
70	64,62
79	67,08

Ein Plot der Daten (Abb. 6.12.8) legt einen sigmoiden (S-förmigen) Verlauf nahe. Daher können wir versuchen, eine logistische Funktion anzupassen:

$$\eta = \frac{\alpha}{1 + \beta \exp(-\gamma x)}$$

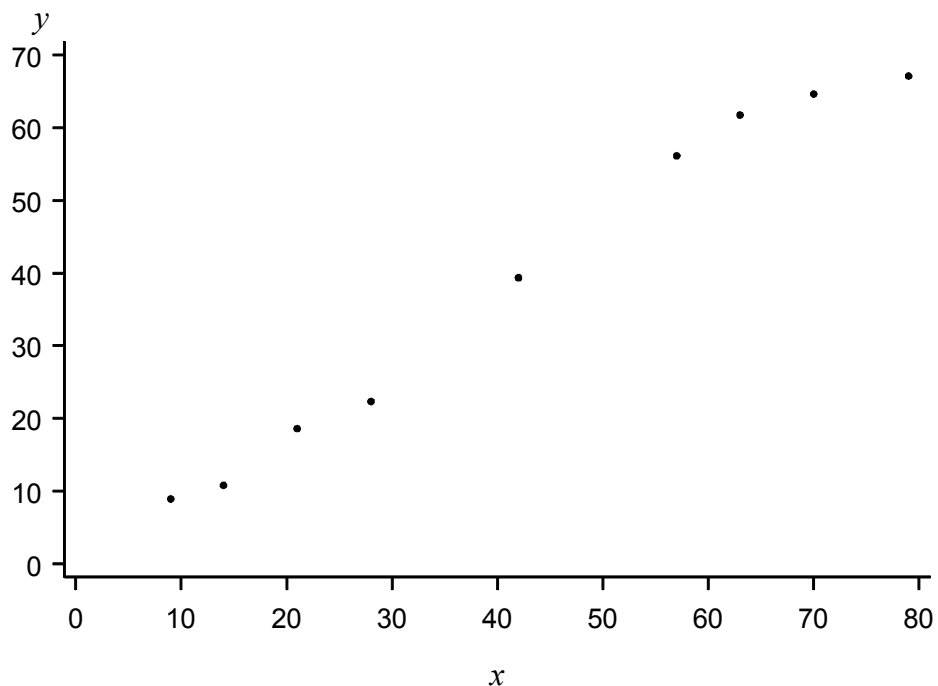


Abb. 6.12.8: Plot der Grünlanddaten. x = Zeit; y = Ertrag (Nachwuchs).

Bei der logistischen Funktion stellt α den maximal erreichbaren Ertrag dar. Dieser Wert scheint im vorliegenden Fall etwa bei $\alpha = 70$ zu liegen, so dass

$$\eta \approx \frac{70}{1 + \beta \exp(-\gamma x)}$$

Dies kann umgeformt werden zu

$$\eta' = \log\left(\frac{70}{\eta} - 1\right) \approx \log(\beta) - \gamma x$$

Dies ist ein lineares Regressionsmodell mit Achsenabschnitt $\log(\beta)$ und Steigung $-\gamma$. Dies legt eine Regression von

$$y' = \log\left(\frac{70}{y} - 1\right)$$

auf x nahe, um Startwerte für β und γ zu finden. Hierzu müssen wir y' für alle Beobachtungen berechnen und dann die Regression durchführen. Ergebnis der Berechnung von y' mit SAS:

x	y	y_strich
9	8.93	1.92260
14	10.80	1.70138
21	18.59	1.01721

28	22.33	0.75837
42	39.35	-0.24986
57	56.11	-1.39614
63	61.73	-2.01014
70	64.62	-2.48584
79	67.08	-3.13430

SAS Anweisungen zur linearen Regression von y' auf x :

```
proc reg data=a;
model y_strich=x;
run;
```

Ausgabe:

```

The REG Procedure
Model: MODEL1
Dependent Variable: y_strich

Parameter Estimates
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.68238	0.07309	36.70	<.0001
x	1	-0.07315	0.00149	-49.06	<.0001

Regression von y' auf x liefert einen Achsenabschnitt von 2,682382 und eine Steigung von -0,073154. Also schätzen wir

$$\log(\hat{\beta}) = 2,682382 \Rightarrow \hat{\beta} = 14,585$$

$$\text{und } \hat{\gamma} = 0,073154$$

Mit diesen Startwerten kann eine nichtlineare Regression durchgeführt werden.

```
proc nlin;
model y=alpha/( 1 + beta*exp(-gamma*x) );
parms alpha=70 beta=14.58 gamma=0.073154;
run;
```

Ergebnis:

```

The NLIN Procedure
```

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
alpha	72.4622	1.7340	68.2192	76.7053
beta	13.7093	1.2105	10.7474	16.6712
gamma	0.0674	0.00345	0.0589	0.0758

Wir sehen, dass die Startwerte schon nah an der Lösung lagen. Die geschätzte Funktion lautet

$$y = \frac{72,46}{1 + 13,71 \exp(-0,0674x)}$$

Diese Funktion ist zusammen mit den Datenpunkten in Abb. 6.12.9 wiedergegeben.

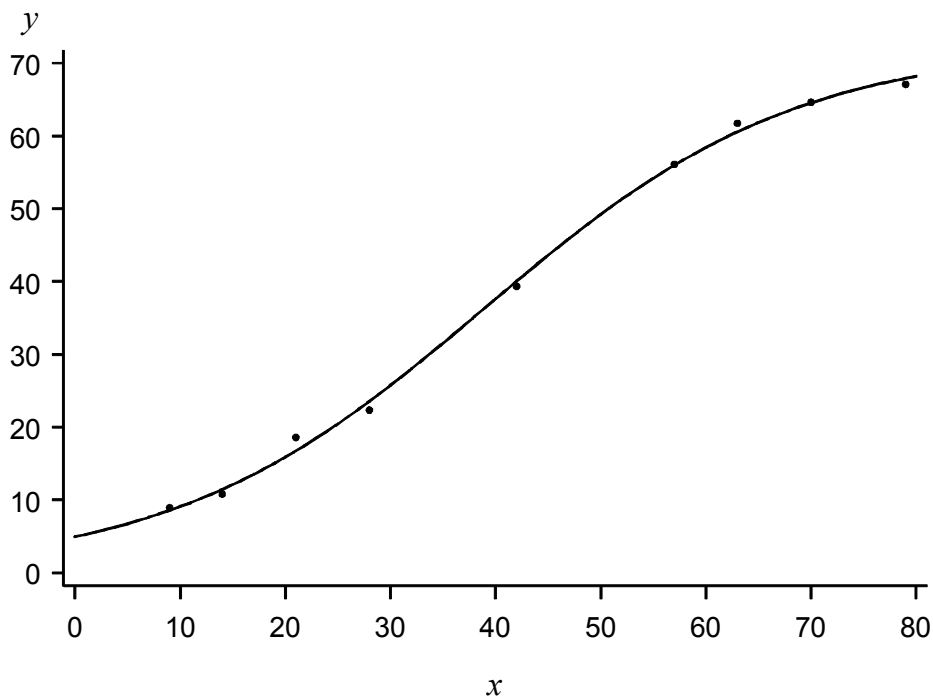


Abb. 6.12.9: Plot der Grünlanddaten. x = Zeit; y = Ertrag (Nachwuchs). Angepasste Funktion: $y = 72,46/[1 + 13,71 \exp(-0,0674x)]$.

6.12.4 Schließende Statistik

Die Verfahren zum Test von Hypothesen und zur Berechnung von Vertrauensintervallen sind weitgehend analog zu denen für das lineare Modell (Abschnitt 6.8 und 6.9), mit der wesentlichen Einschränkung, dass die Verfahren nur asymptotisch (für großes n) gültig sind.

Die Nullhypothese

$$H_0: k' \theta = 0$$

testen wir mit

$$t_{Vers} = \frac{|k' \hat{\theta}|}{ase(k' \hat{\theta})}$$

wobei

$$ase(k'\hat{\theta}) = \sqrt{k'(F'F)^{-1}ks^2}$$

und

$$s^2 = \frac{SQ_{Fehler}}{n-p}$$

mit

p = Zahl der Parameter

n = Zahl der Beobachtungen

ase = asymptotischer Standardfehler der Kleinst-Quadrat-Schätzung von $k'\hat{\theta}$

F = Vektor der ersten Ableitungen von $f(x; \theta)$ an der Stelle $\theta = \hat{\theta}$

"Asymptotisch" heißt, dass der Stichprobenumfang n groß sein muss, damit der ase näherungsweise richtig ist.

H_0 wird verworfen, wenn

$$t_{Vers} > t_{Tab}(\alpha = 5\%, FG = n - p)$$

Ein asymptotisches 95% Vertrauensintervall für $k'\theta$ ist gegeben durch

$$k'\hat{\theta} \pm t_{Tab}ase(k'\hat{\theta})$$

Die Vertrauensintervalle für die Parameter α , β und γ im Kalkbeispiel sind mit dieser allgemeinen Methode berechnet. Man beachte, dass die Matrix F im nichtlinearen Modell eine analoge Rolle spielt wie die Designmatrix X im linearen Modell.

Ebenso können hierarchisch geschachtelte Modelle mittels eines F-Test basierend auf der Reduktion des SQ_{Fehler} verglichen werden wie in Abschnitt 6.9, wobei auch dieser Test nur asymptotisch gültig ist.

Anstelle des **Bestimmtheitsmaßes** (R^2) ist es besser, die **Restvarianz** (s^2) als Maß für die Güte der Anpassung eines Modells sowie zum Vergleich von Modellen zu verwenden, weil bei nichtlinearen Modellen eine Zerlegung der Streuung in zwei Komponenten für „Auf der Regression“ und „Um die Regression“ nicht mehr möglich ist (siehe Schabenberger & Pierce, 2000, p.211-213).

Beispiel: Wir wollen die Nullhypothese testen, dass Kalk keinen Einfluss auf den Ertrag hat. Hierzu betrachten wir folgende Modellsequenz.

Modell	SQ_{Fehler}	FG_{Fehler} ($n - p$)
(0) $f(x) = \beta$	397,3	4
(1) $f(x) = \alpha - (\alpha - \beta) \exp(-\gamma x)$	3,5688	2

$$F_{Vers} = \frac{(397,3 - 3,5688)/2}{3,5688/2} = 110,29 > F_{Tab}(FG_1 = 2, FG_2 = 2, \alpha = 5\%) = 19,00$$

Es besteht ein signifikanter Einfluss der Kalkmenge auf den Ertrag.

6.12.5 Numerische Probleme - „ill-conditioning“

Als Beispiel wird hier ein simulierter Datensatz von Seber und Wild (1989) verwendet, bei dem folgendes Modell zugrundegelegt wurde:

$$\eta = \alpha[1 - \exp(-\gamma x)] \quad (\gamma > 0)$$

Diese Funktion entspricht der Mitscherlich-Funktion mit $\beta = 0$. Die simulierten Daten sind untenstehend aufgeführt.

x	y
1	6,38
2	18,70
3	24,95
4	32,07
5	33,56
6	50,40
7	50,30
8	64,21
9	64,12
10	56,78

Die Daten sind auch in Abb. 6.12.10 dargestellt.

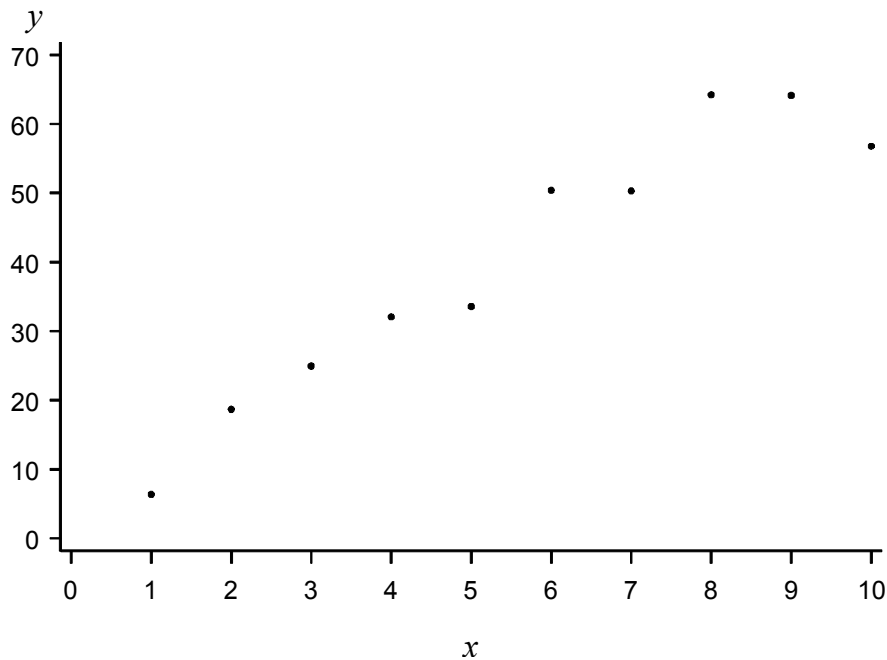


Abb. 6.12.10: Plot der simulierten Daten von Seber und Wild (1989).

An diese Daten soll nun das Modell angepasst werden, mit dem die Daten auch erzeugt wurden. Es ist hier schwierig, gute Anfangswerte zu finden, und dies deutet schon ein Problem an, auf das wir gleich zu sprechen kommen. Der Parameter α stellt das Sättigungsniveau der Kurve dar. Bei den vorliegenden Daten ist dieser Wert schwer abzuschätzen. Eine grobe Schätzung ist etwa 80. Somit haben wir

$$y \approx 80[1 - \exp(-\gamma x)] \Leftrightarrow y' = \log(1 - y/80) = -\gamma x$$

Eine Regression von y' auf x liefert den Startwert $\gamma = 0,169$. Mit diesen Startwerten führen wir eine nichtlineare Regression durch.

```
data;
input x y;
cards;
1 6.38
2 18.7
3 24.95
4 32.07
5 33.56
6 50.4
7 50.3
8 64.21
9 64.12
10 56.78
;
proc nlin;
model y=alpha*(1-exp(-gamma*x));
parms alpha=80 gamma=0.169;
run;
```

Ergebnis:

The NLIN Procedure

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
alpha	108.5	36.0063	25.4240	191.5
gamma	0.0910	0.0429	-0.00789	0.1900

Approximate Correlation Matrix

	alpha	gamma
alpha	1.0000000	-0.9937330
gamma	-0.9937330	1.0000000

Die Lösung ist

$$\eta = 108,42[1 - \exp(-0,091x)].$$

Allerdings fällt zunächst auf, dass die Standardfehler der Parameterschätzwerte relativ groß und die Vertrauensintervalle breit sind. So hat das Intervall für α die Grenzen 25 und 192. Da α das Sättigungsniveau ist, können wir somit über das tatsächliche Sättigungsniveau wenig sagen. Aufschlussreich ist in dieser Hinsicht auch eine Abbildung der geschätzten Kurve mit den Daten in Abb. 6.12.11.

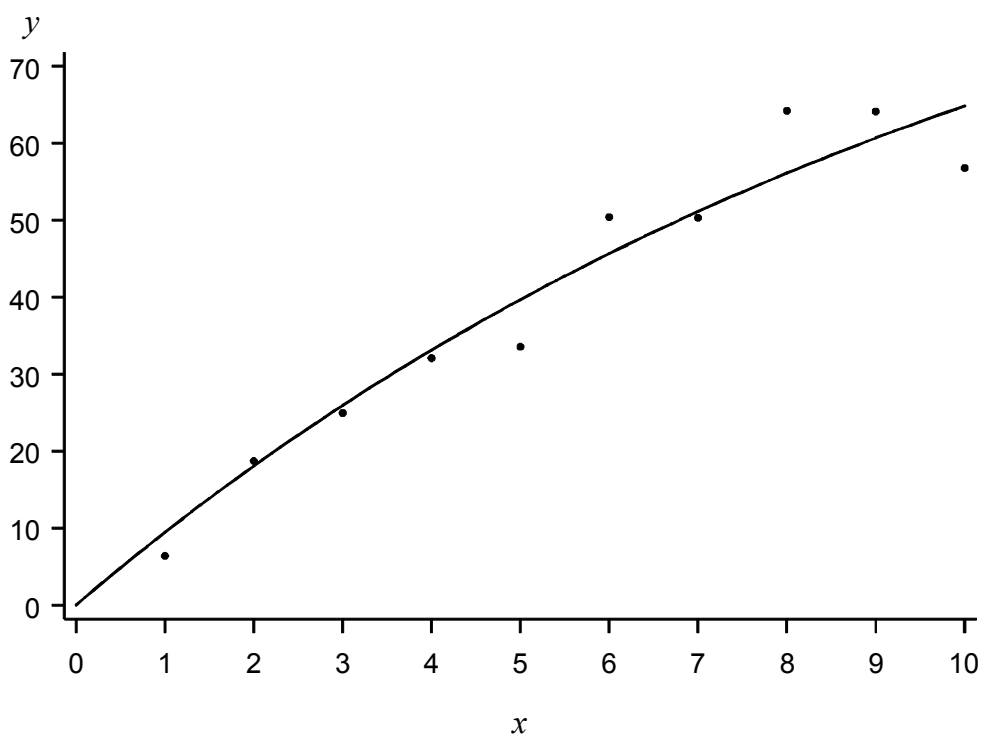


Abb. 6.12.11: Plot der simulierten Daten von Seber und Wild (1989). Angepasste Funktion: $\eta = 108,42[1 - \exp(-0,091x)]$.

Man sieht, dass nach dem Kurvenverlauf zu urteilen der Sättigungsbereich weit jenseits des untersuchten x -Bereiches liegen muss. Dies bedeutet, dass man über die genaue Lage der Asymptote $y = \alpha$ keine oder so gut wie keine Information hat. Egal, wie stark die Kurve jenseits des beobachteten x -Bereiches gekrümmt ist, die Anpassung im beobachteten Bereich wird sich dadurch wenig ändern. Die Anpassung wäre auch nicht sehr viel schlechter, wenn man anstelle der Funktion $\eta = \alpha[1 - \exp(-\gamma x)]$ eine Gerade durch den Ursprung angepasst hätte. Eine solche Funktion hätte gar keine Asymptote mehr. Da die Daten keine Information über die genaue Lage der Asymptote liefern, können wir auch nicht erwarten, dass der Parameter α der Funktion $\eta = \alpha[1 - \exp(-\gamma x)]$ mit großer Genauigkeit gemessen werden kann. Hiervon ist außerdem auch die Genauigkeit der Schätzung des zweiten Parameters γ betroffen. Um dies nachzuvollziehen, ist folgende analytische Betrachtung hilfreich. Die erste Ableitung der Funktion $\eta = \alpha[1 - \exp(-\gamma x)]$ nach x ist

$$\frac{d\eta}{dx} = \alpha\gamma \exp(-\gamma x)$$

Die erste Ableitung entspricht der Steigung der Kurve an der Stelle x . Wenn nun viele der beobachteten Werte von $-\gamma x$ nahe Null sind, wie im vorliegenden Fall, dann ist

$$\frac{d\eta}{dx} \approx \alpha\gamma$$

so dass

$$\eta \approx \alpha\gamma x$$

Daher ist die Steigung der Funktion in diesem Bereich fast unabhängig vom x -Wert, und der Kurvenverlauf entspricht fast einer Gerade. Dabei ist die Steigung dieser Gerade $\alpha\gamma$. Unsere Daten liefern somit im wesentlichen Information über das Produkt der Parameter α und γ , nicht aber über jeden einzelnen der beiden Parameter. Das ist auch der Grund, warum die Schätzwerte von α und γ eine so hohe Korrelation aufweisen ($r = 0,99$) und mit einem so hohen Standardfehler behaftet sind.

Eine andere Sicht auf das Problem bietet die Matrix der ersten Ableitungen der Funktion nach den Parametern, F , die für die Berechnung der Kleinst-Quadrat-Schätzungen benötigt wird (beachte: F hat dieselbe Funktion wie die Design-Matrix X im linearen Modell. Die ersten Ableitungen sind:

$$\begin{aligned}\frac{d\eta}{d\alpha} &= 1 - \exp(-\gamma x) \\ \frac{d\eta}{d\gamma} &= \alpha x \exp(-\gamma x)\end{aligned}$$

Hiermit finden wir folgende F -Matrix:

$$F = \begin{pmatrix} 0,087 & 98,9 \\ 0,166 & 108,6 \\ 0,239 & 247,6 \\ 0,305 & 301,4 \\ 0,366 & 343,9 \\ 0,421 & 376,8 \\ 0,471 & 401,4 \\ 0,517 & 418,8 \\ 0,559 & 430,2 \\ 0,597 & 436,4 \end{pmatrix}$$

Die beiden Spalten dieser Matrix sind hoch korreliert ($r = 0,98$). Somit liegt ein hohes Maß an Multikollinearität vor. Diese führt zu einer sehr großen Ungenauigkeit der Schätzungen der Parameter (Varianzinflation), ähnlich wie bei der multiplen linearen Regression.

Wenn wir die Freiheit haben, den Wertebereich von x auszudehnen, so dass die Asymptote erreicht wird (Versuchsplanung), dann liegt mehr Information über die Asymptote vor, und das Problem reduziert sich. Generell gilt: Wenn ein Modell eine Asymptote hat, dann muss der Wertebereich der Daten den asymptotischen Bereich gut abdecken, um eine verlässliche Parameterschätzung zu erhalten. Wird der Sättigungsbereich nicht erreicht und der Zusammenhang im untersuchten Wertebereich annähernd linear, ist es oft besser, eine lineare Regression durchzuführen.

Abschließend sei noch einmal darauf hingewiesen, dass das beschriebene Problem auftritt, obwohl die Daten sogar nach der Funktion simuliert wurden, die auch zur Auswertung verwendet wird.

***6.12.6 Ein weiteres Beispiel**

Beispiel (Dr. S. Graeff, Institut 340, Uni Hohenheim, im Februar 2003): Mittels Reflexionsmessungen wird versucht, den N-Ernährungsstatus von Pflanzen zu charakterisieren.

Hypothese: Die Reflexion einer Pflanze ändert sich ab einer bestimmten N-Konzentration in der TM.

Beim vorliegenden Beispiel handelt es sich um Maispflanzen, die ab einer Konzentration von $N < 3 \%$ in der TM im Mangel sind (Optimum- und Mangelwerte sind definiert in der Literatur).

Es wurde ein Feldversuch angelegt, mit 6 unterschiedlichen N-Düngungsstufen (0, 20, 40, 80, 120, 160 kg N ha⁻¹). Die Reflexion der Maispflanzen wird zu unterschiedlichen Zeitpunkten in der Vegetationsperiode gemessen. Von der

gemessenen Pflanze wird anschließend chemisch die tatsächlich vorliegende N-Konzentration ermittelt.

Im Beispiel ändert sich der Reflexionsparameter ΔEb deutlich, wenn die N-Konzentration in der TM 3 % unterschreitet. Über eine mathematische Funktion wird versucht, die Änderung des Reflexionsparameters unter verschiedenen N-Konzentrationen in der TM zu beschreiben. Über die angepasste Funktion soll ein Rückschluss gezogen werden, ob bei der Maispflanze ein Stickstoffmangel vorliegt und eine N-Düngung erfolgen muss, oder nicht.

Die Daten sind in Abb. 6.12.12 geplottet.

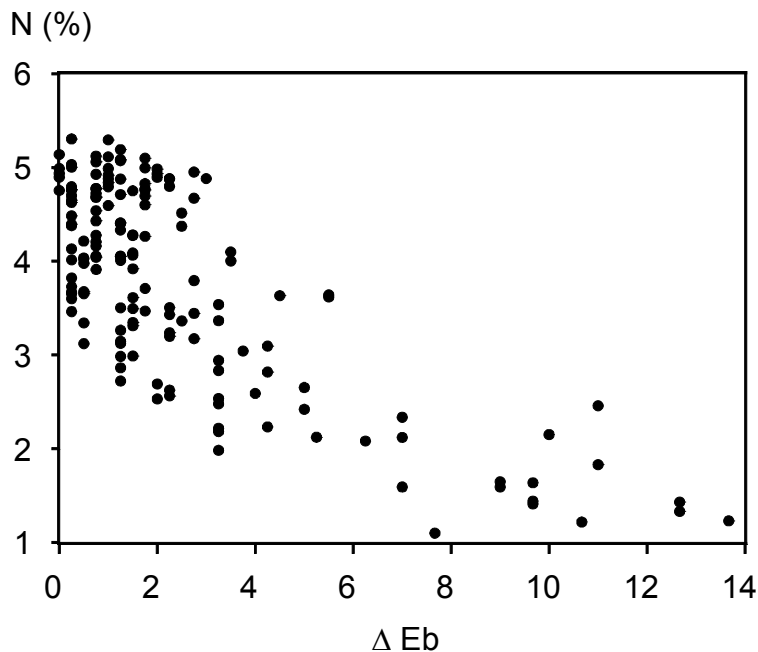


Abb. 6.12.12: Plot der Stickstoffgehalte [N (%)] gegen die Reflexionsmessung [ΔEb]

Eine mögliche Auswertungsstrategie

Das Ziel der Auswertung ist eine Vorhersage des N-Gehaltes aus der Reflexionsmessung (ΔEb). Daher ist eine Regression mit N (%) als Zielvariable und ΔEb als Prädiktorvariable sinnvoll. Sodann ist es von Interesse, denjenigen Wert der Reflexion zu bestimmen, für den der erwartete N Gehalt gleich 3% ist (Schwelle für N Mangel aus Literatur). Dieser Reflexionswert ist dann der Schwellenwert für die Reflexion, dessen Überschreitung einen N-Mangel anzeigt.

Das gestellte Problem ist das einer inversen Regression (siehe Abschnitt 6.2.3). Wir werden hier zeigen, wie das Problem mit Hilfe eines geeignet parametrisierten nichtlinearen Modells zu lösen ist.

Aus Abb. 6.12.5 ergibt sich der Eindruck einer nichtlinearen Beziehung. Diesen Eindruck prüfen wir zunächst mittels eines Polynoms 2. Grades (y = N-Gehalt in %, x = ΔEb).

Dependent Variable: Y

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X	1	107.6569677	107.6569677	221.06	<.0001
X*X	1	3.3737463	3.3737463	6.93	0.0093

Der quadratische Term ist signifikant, was auf eine Abweichung von der Linearität hinweist. Die Daten in Abb. 6.12.9 könnten durch eine Exponentialfunktion der Form

$$\eta = \alpha \exp(\beta x)$$

zu beschreiben sein. In diesem Fall lässt sich das Modell durch eine logarithmische Transformation linearisieren:

$$\eta' = \log(\eta) = \log(\alpha) + \beta x \quad (1)$$

wobei $\log()$ der Logarithmus zur Basis e ist. Wir plotten $y' = \log(y)$ gegen x . Es ergibt sich eine nahezu lineare Beziehung (Abb. 6.12.13).

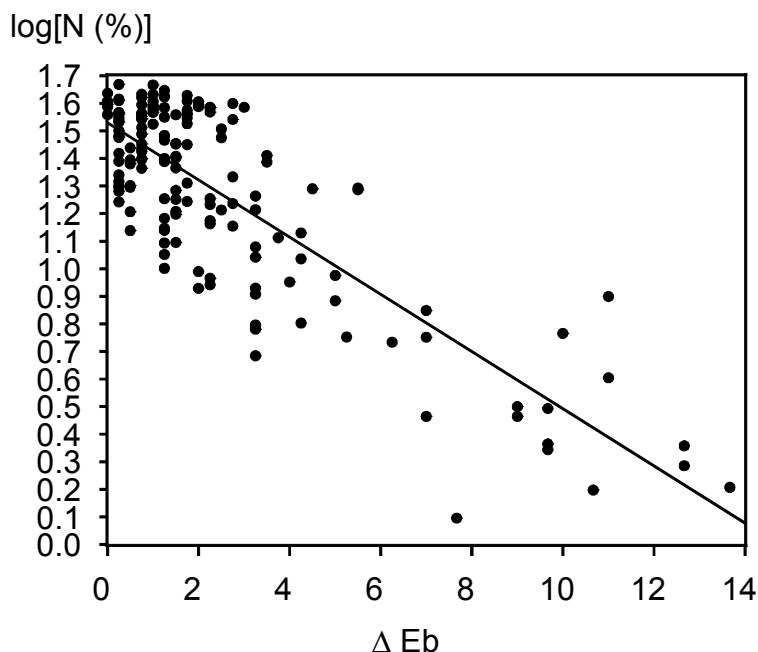


Abb. 6.12.13: Plot der logarithmierten Stickstoffgehalte [N (%)] gegen die Reflexionsmessung [ΔE_b] mit angepasster Regressionsgerade.

Anpassung eines Polynoms zweiten Grades für die Regression von y' auf x ergibt, dass nur noch der lineare Term signifikant ist, was sich mit dem Eindruck aus Abb. 6.12.10 deckt.

Dependent Variable: log_Y

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X	1	13.75853306	13.75853306	338.46	<.0001
X*X	1	0.06165170	0.06165170	1.52	0.2200

Dies Ergebnis führt uns dazu, fortan das exponentielle Modell $\eta = \alpha \exp(\beta x)$ anzunehmen, dessen Analyse auf der logarithmierten Skala durch eine einfache lineare Regression zu erhalten ist. Die Parameterschätzungen auf der logarithmischen Skala für y sind:

Dependent Variable: log_Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13.75853306	13.75853306	337.34	<.0001
Error	156	6.36244737	0.04078492		

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	1.528864716	0.02129310	71.80	<.0001
X	-0.103686881	0.00564531	-18.37	<.0001

Somit lautet das geschätzte Modell:

$$\log(\hat{y}) = 1,52886 - 0,10369x$$

Wollen wir nun denjenigen Wert der Reflexion (x) ermitteln, für den der erwartete N-Gehalt gleich 3% ist, so ist folgende Gleichung zu lösen:

$$\log(3) = 1,52886 - 0,10369x$$

Auflösen nach x ergibt:

$$x = \frac{1,52886 - \log(3)}{0.10369} = 4,14937$$

Die geschätzte Reflexion, bei welcher der erwartete N-Gehalt gleich 3% ist, beträgt somit

$$\Delta E_b = 4,14937$$

Nun ist dies nur eine Schätzung, und es ist daher geboten, ein Vertrauensintervall zu berechnen. Hierzu kann das in Abschnitt 6.2.3 beschriebene Verfahren verwendet werden, was allerdings etwas aufwendig ist. Einfacher ist es, einen alternativen Weg zu verfolgen, der jetzt beschrieben wird.

Die Idee besteht darin, das lineare Modell so zu reparametrisieren, dass es den interessierenden Schwellenwert für die Reflexion als Parameter enthält. Hierzu setzen wir für η den Schwellenwert $\eta_s = 3\%$ ein und finden

$$\eta'_s = \log(\eta_s) = \alpha' + \beta x_s$$

wobei $\alpha' = \log(\alpha)$ ist und x_s den Schwellenwert der Reflexion bezeichnet. Dies lösen wir nach β auf:

$$\beta = \frac{\log(\eta_s) - \alpha'}{x_s}$$

Setzen wir dies in das lineare Modell (1) ein, so finden wir

$$\log(\eta) = \eta' = \alpha' + \frac{\log(\eta_s) - \alpha'}{x_s} x$$

Dieses reparametrisierte Modell hat nun den interessierenden Schwellenwert x_s als Parameter (neben dem Parameter α'). Wir können dieses Modell nun nutzen, um den gesuchten Schwellenwert direkt zu schätzen. Hierbei ist zu beachten, dass das reparametrisierte Modell (2) nichtlinear in den Parametern ist und sich nicht durch eine einfache Transformation linearisieren lässt. Also müssen wir ein Verfahren zur eigentlichen nichtlinearen Regression verwenden.

Wir verwenden die NLIN Prozedur von SAS. Da wir die Parameter bereits für das Modell (1) geschätzt haben, sind wir hier in der komfortablen Situation, perfekte Startwerte für die Parameter zu haben. Am geschätzten Modell ändert sich durch die Reparametrisierung nichts, aber wir bekommen zusätzlich ein 95% Vertrauensintervall für alle Parameter angegeben, so auch für den interessierenden Schwellenwert x_s .

```
proc nlin data=daten;
parms a_prime=1.53 x_s=4.15;
model log_y=a_prime + x*(log(3)-a_prime)/x_s;
run;
```

Ergebnis:

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	2	269.5	134.7	337.34	<.0001
Residual	156	6.3624	0.0408		
Uncorrected Total	158	275.8			
Corrected Total	157	20.1210			
Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits		
a_prime	1.5289	0.0213	1.4868	1.5709	
x_s	4.1495	0.1798	3.7944	4.5047	

Approximate Correlation Matrix		
	a_prime	x_s
a_prime	1.0000000	0.3175929
x_s	0.3175929	1.0000000

Man beachte, dass das MQ_{Fehler} dasselbe ist wie für die lineare Regression nach (1). Dies ist zu erwarten, da die Modelle (1) und (2) äquivalent sind.

Die Schätzung für den Schwellenwert der Reflexion ist $x_s = 4,1495$, was bis auf Rundungsfehler identisch ist mit der linearen Regression.

Das 95%-Vertrauensintervall hat die Grenzen 3,79 und 4,51.

6.13 Lineare Kontraste

Im Zusammenhang mit der Varianzanalyse für einen qualitativen Behandlungsfaktor (z.B. Sorte) haben wir im Anschluss an einen F-Test paarweise multiple Mittelwertvergleiche durchgeführt (Abschnitt 4.5). Je nach Fragestellung kann es nun sein, dass etwas komplexere Mittelwertvergleiche von Interesse sind.

Beispiel: Sokal & Rohlf beschreiben das Ergebnis eines Gewebekulturexperimentes. Der Zweck des Experimentes war zu ermitteln, welche Effekte die Zugabe verschiedener Zuckerarten und -mengen auf die Länge (in okularen Einheiten $\times 0,114 = \text{mm}$) von Erbsengewebeabschnitten auf Nährmedium mit Auxin hat. Die Behandlungen wurden zufällig auf die Versuchseinheiten (Petrischalen) verteilt. Die Ergebnisse waren wie folgt:

	Behandlung				
	1	2	3	4	5
Wiederholung (Petrischale)	Kontrolle	2% Glucose	2% Fructose	1% Glucose + 1% Fructose (Mischung)	2% Saccharose
1	75	57	58	58	62
2	67	58	61	59	66
3	70	60	56	58	65
4	75	59	58	61	63
5	65	62	57	57	64
6	71	60	56	56	62
7	67	60	61	58	65
8	67	57	60	57	65
9	76	59	57	57	62
10	68	61	58	59	67

Der Versuchsansteller ist an folgenden Vergleichen interessiert:

- Kontrolle vs. Durchschnitt aller Zuckerbehandlungen
- Mischung vs. reine Zucker (Fructose und Glucose)

Diese Arten von Vergleichen werden als **lineare Kontraste** bezeichnet.

Zur Auswertung wird man das lineare Modell

$$y_{ij} = \mu + \tau_i + e_{ij}$$

zugrundelegen. Der Erwartungswert einer Behandlung ist gegeben durch

$$E(y_{ij}) = \mu_i = \mu + \tau_i$$

Erwartungswerte können im Fall einer vollständig randomisierten Anlage durch den Stichprobenmittelwert $\bar{y}_{i\cdot}$ geschätzt werden. Die beiden Kontraste können wie folgt geschätzt werden:

Beschreibung des Kontrasts	Kontrast (Schätzung), L
Kontrolle vs. Durchschnitt aller Zuckerbehandlungen	$\bar{y}_{1\cdot} - \frac{\bar{y}_{2\cdot} + \bar{y}_{3\cdot} + \bar{y}_{4\cdot} + \bar{y}_{5\cdot}}{4}$
Reine Zucker (Fructose und Glucose) vs. Mischung	$\bar{y}_{4\cdot} - \frac{\bar{y}_{2\cdot} + \bar{y}_{3\cdot}}{2}$

Die Behandlungsmittelwerte sind:

Behandlung	1	2	3	4	5
Mittel ($\bar{y}_{i\cdot}$)	70,1	59,3	58,2	58,0	64,1

Die Schätzungen des Kontrasts "Kontrolle vs. Durchschnitt aller Zuckerbehandlungen" ist:

$$L_1 = 70,1 - \frac{59,3 + 58,2 + 58,0 + 64,1}{4} = 10,20$$

Im Mittel unterdrücken die Zuckerarten das Wachstum der Erbsen. Der Kontrast "Reine Zucker vs. Mischung" ergibt

$$L_2 = 58,0 - \frac{59,3 + 58,2}{2} = -0,75$$

Der Kontrast ist sehr klein und weist auf einen leichten Vorteil von reinen Zuckern hin.

Die beiden Beispiele können als Spezialfälle eines linearen Kontrasts der Form

$$L = \sum_{i=1}^t c_i \bar{y}_i.$$

betrachtet werden, wobei c_i die Kontrastkoeffizienten sind. Diese erfüllen die Bedingung

$$\sum_{i=1}^t c_i = 0$$

Für die beiden Kontraste sind die Koeffizienten wie folgt:

Beschreibung des Kontrasts	Koeffizienten				
	c_1	c_2	c_3	c_4	c_5
Kontrolle vs. Durchschnitt aller Zuckerbeh.	1	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
Mischung vs. reine Zucker	0	$-\frac{1}{2}$	$-\frac{1}{2}$	1	0

Ein Kontrast kann mittels t-Test geprüft werden, wobei die t -Statistik die Form

$$t_{\text{vers}} = \frac{|L|}{s.e.(L)}$$

hat, wobei $s.e.(L)$ der geschätzte Standardfehler des Kontrastes L ist. Für balancierte Daten gilt:

$$s.e.(L) = \sqrt{\frac{s^2}{r} \sum_{i=1}^t c_i^2}$$

wobei $s^2 = MQ_{\text{Fehler}}$ der einfaktoriellen Varianzanalyse ist (Abschnitt 4.4).

Bemerkung: Ein paarweiser t-Test ist der Spezialfall eines Kontrasts mit Koeffizienten 1 und -1 für die beiden zu vergleichenden Behandlungen und 0 für die übrigen Behandlungen.

Die Varianzanalyse für die Zuckerdaten liefert $s^2 = 5,46$. Hiermit berechnen sich die Standardfehler zu ($r = 10$ Wiederholungen)

$$s.e.(L_1) = \sqrt{\frac{5,46}{10} \left[1^2 + \left(-\frac{1}{4}\right)^2 + \left(-\frac{1}{4}\right)^2 + \left(-\frac{1}{4}\right)^2 + \left(-\frac{1}{4}\right)^2 \right]} = 0,826$$

$$s.e.(L_2) = \sqrt{\frac{5,46}{10} \left[0^2 + \left(-\frac{1}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 + (1)^2 + 0^2 \right]} = 0,905$$

Die t-Werte sind wie folgt:

Beschreibung des Kontrasts	L	s.e.(L)	t_{vers}
Kontrolle vs. Durchschnitt aller Zuckerbeh.	10,20	0,826	12,35
Mischung vs. reine Zucker	-0,75	0,905	0,83

Das Experiment hat $t(r-1) = 45$ Fehler-Freiheitsgrade (t = Zahl der Behandlungen, r = Zahl der Wiederholungen; siehe Abschnitt 4.4, Statistik-Skript), so dass $t_{tab} = 2,014$ ($\alpha = 5\%$). Also ist der erste Kontrast signifikant und der zweite nicht. Also unterdrücken Zucker das Wachstum und es gibt keinen signifikanten Mischungseffekt zwischen Fructose und Glucose.

Ein Kontrast L ist eine Schätzung des entsprechenden Kontrastes der Erwartungswerte $\mu_i = \mu + \tau_i$:

$$\lambda = \sum_{i=1}^t c_i \mu_i$$

Wegen der Bedingung $\sum_{i=1}^t c_i = 0$ gilt

$$\lambda = \sum_{i=1}^t c_i \tau_i$$

Dies ist wiederum ein Spezialfall einer linearen Funktion von Parametern

$$\lambda = \mathbf{k}'\boldsymbol{\beta}$$

im allgemeinen linearen Modell

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Die Berechnung von t-Tests für λ im allgemeinen Fall wurde in 6.8 besprochen. Die hier angegebenen Formeln für lineare Kontraste ergeben sich durch Anwendung dieser allgemeinen Resultate. Im vorliegenden Fall haben wir

$$\boldsymbol{\beta}' = (\mu \quad \tau_1 \quad \tau_2 \quad \tau_3 \quad \tau_4 \quad \tau_5)$$

Für einen linearen Kontrast gilt

$$k' = (0 \quad c_1 \quad c_2 \quad c_3 \quad c_4 \quad c_5)$$

Wir haben hier zunächst den Fall linearer Kontraste bei balancierten Daten in einer vollständig randomisierten Anlage betrachtet. Kontraste sind ganz allgemein in verschiedenen Situationen relevant. Die Beziehung zum allgemeinen linearen Modell weist den Weg, wie Kontraste in solchen Fällen zu testen sind. Für die Umsetzung in Statistik-Paketen ist die Spezifizierung der Kontrastkoeffizienten c_i der entscheidende Schritt.

SAS Anweisungen

```
data;  
input trt length;  
cards;  
1 75  
1 67  
<mehr Daten>  
5 62  
5 67  
;  
proc glm;  
class trt;  
model length=trt;  
means trt;  
estimate 'control vs. all sugars' trt 4 -1 -1 -1 -1/divisor=4;  
estimate 'mixture vs. pure'      trt 0 -1 -1 2 0/divisor=2;  
run;
```

Label für Kontrast

Behandlungsvariable $\Rightarrow \tau_i$

Gemeinsamer Nenner der Kontrastkoeffizienten

Zähler der Kontrastkoeffizienten

In der ESTIMATE Anweisung werden am besten nur ganze Zahlen verwendet. Bei gebrochenen Zahlen kann der gemeinsame Nenner mit der DIVISOR= Option angegeben werden, während die Zähler hinter der Behandlungsvariable aufgelistet werden.

Anhang zu Kap. 6: Einige Grundlagen der Matrizenrechnung

Die folgende Darstellung ist RGD Steel und JH Torrie (1980, Principles and procedures of statistics, McGraw-Hill, New York) entlehnt.

A.1 Matrizen

Eine Matrize oder Matrix ist ein rechteckiges Feld von Zahlen (Skalaren). Sie ist in Zeilen und Spalten strukturiert, wobei in jeder Zelle (Zeilen-Spalten-Kombination) eine Zahl steht. Den Eintrag einer Zelle bezeichnet man auch als **Element**.

Beispiele:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 5 & 1 & 3 \\ 2 & 10 & 7 \end{pmatrix}, C = \begin{pmatrix} 20 \\ 30 \end{pmatrix}, D = (5 \quad 20), E = \begin{pmatrix} 20 & 12 \\ 12 & 24 \end{pmatrix}$$

Matrizen werden symbolisch durch fettgeschriebene Buchstaben repräsentiert. Hat eine Matrize nur eine Spalte, so spricht man von einem Spaltenvektor. Die Matrizen **A** und **E** haben zwei Zeilen und zwei Spalten, die Matrix **B** hat drei Spalten und zwei Zeilen. **C** ist ein Spaltenvektor, **D** ist ein Zeilenvektor. Allgemein kann eine Matrix **A** wie folgt ausgedrückt werden:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1c} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2c} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{ic} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{r1} & a_{r2} & \dots & a_{rj} & \dots & a_{rc} \end{pmatrix}$$

a_{ij} = Element in Zeile i und Spalte j .

Die **Transponierte** **A'** einer Matrix **A** ergibt sich durch Vertauschen der Zeilen und Spalten: Die erste Spalte der Transponierten entspricht beispielsweise der ersten Zeile der ursprünglichen Matrix:

$$A' = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1c} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2c} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{ic} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{r1} & a_{r2} & \dots & a_{rj} & \dots & a_{rc} \end{pmatrix}$$

Beispiel:

$$B = \begin{pmatrix} 5 & 1 & 3 \\ 2 & 10 & 7 \end{pmatrix}, B' = \begin{pmatrix} 5 & 2 \\ 1 & 10 \\ 3 & 7 \end{pmatrix}$$

Eine **quadratische Matrix** ist eine Matrix mit derselben Zahl von Zeilen und Spalten:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2r} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{ir} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{r1} & a_{r2} & \dots & a_{rj} & \dots & a_{rr} \end{pmatrix}$$

Eine **symmetrische Matrix** ist eine quadratische Matrix, für die $a_{ij} = a_{ji}$ ist.

Die **Einheitsmatrix** ist eine symmetrische Matrix mit Einsen auf der Diagonale und Nullen jenseits der Diagonale:

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

A.2 Elementare Matrixoperationen

Die **Addition** von Matrizen erfolgt durch elementweise Addition für alle Elemente:

$$A+B = \begin{pmatrix} a_{11}+b_{11} & \dots & a_{1c}+b_{1c} \\ \dots & \dots & \dots \\ \dots & a_{ij}+b_{ij} & \dots \\ \dots & \dots & \dots \\ a_{r1}+b_{r1} & \dots & a_{rc}+b_{rc} \end{pmatrix}$$

Beispiel:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, E = \begin{pmatrix} 20 & 12 \\ 12 & 24 \end{pmatrix}, A+E = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 20 & 12 \\ 12 & 24 \end{pmatrix} = \begin{pmatrix} 1+20 & 2+12 \\ 3+12 & 4+24 \end{pmatrix} = \begin{pmatrix} 21 & 14 \\ 15 & 28 \end{pmatrix}$$

Wiederholte Anwendung der Regel zur Addition führt zu folgender Regel für die **Multiplikation einer Matrix mit einem Skalar**:

$$\underbrace{A+A+\dots+A}_{k \text{ mal}} = kA = \begin{pmatrix} ka_{11} & \dots & ka_{1c} \\ \dots & \dots & \dots \\ ka_{r1} & \dots & ka_{rc} \end{pmatrix}$$

Diese Definition gilt auch, wenn k keine natürliche Zahl ist.

Die **Subtraktion** von Matrizen erfolgt analog der Addition.

Die **Multiplikation** von Matrizen ist auf eine Art definiert, wie man sie von der Definition der Addition und Subtraktion von Matrizen her nicht erwarten würde. Am besten wird dies zunächst an einem Beispiel gezeigt:

$$AE = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 20 & 12 \\ 12 & 24 \end{pmatrix} = \begin{pmatrix} 1*20+2*12 & 1*12+2*24 \\ 3*20+4*12 & 3*12+4*24 \end{pmatrix} = \begin{pmatrix} 44 & 60 \\ 108 & 132 \end{pmatrix}$$

Allgemein gilt:

$$AB = \begin{pmatrix} \sum_h a_{1h} b_{h1} & \sum_h a_{1h} b_{h2} & \dots & \sum_h a_{1h} b_{hc} \\ \dots & \dots & \dots & \dots \\ \sum_h a_{rh} b_{h1} & \sum_h a_{rh} b_{h2} & \dots & \sum_h a_{rh} b_{hc} \end{pmatrix} = C_{r \times c}$$

wobei A eine $r \times s$ Matrix ist (r Zeilen und s Spalten), während B eine $s \times c$ Matrix ist (s Zeilen und c Spalten). Man beachte, dass für das Beispiel

$$EA = \begin{pmatrix} 20 & 12 \\ 12 & 24 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 20*1+12*3 & 20*2+12*4 \\ 12*1+24*3 & 12*2+24*4 \end{pmatrix} = \begin{pmatrix} 56 & 88 \\ 84 & 120 \end{pmatrix} \neq AE$$

Es gilt also nicht das Kommutativgesetz der Multiplikation bei Skalaren. Die Einheitsmatrix ist das **neutrale Element der Multiplikation** von Matrizen. Es gilt: $AI = A$.

Als Eselsbrücke dafür, wie Matrizen multipliziert werden, hilft vielleicht folgender Spruch: "Zuerst die Zeilen, später die Spalten." Dieser Spruch trifft übrigens auch auf die Reihenfolge der Indizierung der Elemente einer Matrix (a_{ij}) zu.

A.3 Inverse einer Matrix, lineare Abhängigkeiten und Rang einer Matrix

Multipliziert man ein Skalar ($a \neq 0$) mit seinem Kehrwert, so erhält man das Ergebnis 1, das neutrale Element der Multiplikation:

$$a \times \frac{1}{a} = 1, \text{ z.B. } 7 \times \frac{1}{7} = 1$$

Der Kehrwert ist nicht definiert, falls $a = 0$, da die Division durch Null nicht erlaubt ist. In der Matrizenrechnung gibt es eine analoge Operation, die Multiplikation einer Matrix A mit ihrer **Inversen** A^{-1} . Das Ergebnis ist das neutrale Element der Matrizenmultiplikation, die Einheitsmatrix:

$$AA^{-1} = A^{-1}A = I$$

Die Berechnung einer Inversen ist bei größeren Matrizen am besten mit numerischen Methoden zu bewerkstelligen, die hier nicht näher besprochen werden können. Für kleine Matrizen lässt sich die Inverse leicht finden.

Beispiel:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix}; A^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix};$$

$$AA^{-1} = \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Die explizite Multiplikation von A mit der Inversen A^{-1} liefert vier Gleichungen, die leicht paarweise nach den Unbekannten a, b, c und d aufgelöst werden:

$$\begin{array}{ll} a + 2c = 1 & b + 2d = 0 \\ 3a - c = 0 & 3b - d = 1 \end{array}$$

$$a = 1/7, c = 3/7 \quad b = 2/7, d = -1/7$$

Nicht für jede Matrix ist eine Inverse definiert, z.B. für

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 8 \end{pmatrix}$$

Hier finden wir mit

$$AA^{-1} = \begin{pmatrix} 1 & 2 \\ 4 & 8 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

die Gleichungen

$$\begin{array}{ll} a + 2c = 1 & \text{und} \quad b + 2d = 1 \\ 4a + 8c = 0 & 4b + 8d = 0 \end{array}$$

Die erste Gleichung ergibt

$$a = 1 - 2c$$

Einsetzen in die zweite Gleichung führt zu

$$4 - 8c + 8c = 4 \neq 0$$

Die beiden Gleichungen haben keine Lösung, sie sind inkonsistent. Dasselbe Problem trifft für die dritte und vierte Gleichung in b und d zu. Daher hat die zweite A Matrix keine eindeutige Inverse.

Die Matrix hat zwei Spalten, von denen die zweite das zweifache der ersten ist. Daher sind die beiden Spalten **linear abhängig**.

Der **Rang** einer Matrix ist die maximale Zahl unabhängiger Spalten (Zeilen) einer Matrix. Die obige A Matrix hat z.B. den Rang 1. Der Rang einer Matrix wird in der Regel mit numerischen Methoden bestimmt, die hier nicht näher behandelt werden sollen. Eine Matrix, deren Rang kleiner als die Zahl der Spalten (Zeilen) ist, ist nicht von vollem Spalten- (Zeilen-) Rang. Bei einer quadratischen Matrix gilt, dass der Zeilenrang gleich dem Spaltenrang ist. Ist der Rang kleiner als die Dimension der Matrix, so sagt man kurz, dass diese nicht von vollem Rang ist. **Eine quadratische Matrix, die nicht von vollem Rang ist, ist nicht eindeutig invertierbar.**

Ein einfacher Spezialfall ist der einer symmetrischen 2×2 Matrix der Form

$M = \begin{pmatrix} c & e \\ e & d \end{pmatrix}$. Diese hat die Inverse

$$M^{-1} = \begin{pmatrix} c & e \\ e & d \end{pmatrix}^{-1} = \frac{1}{cd - e^2} \begin{pmatrix} d & -e \\ -e & c \end{pmatrix}$$

Man beachte, dass eine Inverse nur dann existiert, wenn der Nenner des Bruchs von M^{-1} , also $cd - e^2$, ungleich Null ist. Der Nenner wird als **Determinante** von M bezeichnet. Die Berechnung der Determinante größerer quadratischer Matrizen ist komplizierter und soll hier nicht besprochen werden (siehe Vorlesung zur Mathematik). Die Bedingung für die Existenz einer eindeutigen Inversen einer symmetrischen Matrix M lautet $\det(M) \neq 0$. Die folgenden Aussagen sind äquivalent:

- (1) $\det(M) = 0$.
- (2) Die Zeilen (Spalten) von M sind linear abhängig.
- (3) M ist nicht von vollem Rang.

Außerdem gilt: Der Rang der Matrix M entspricht der Dimension (Zahl der Spalten/ Zeilen) der größten Sub-Matrix von M , welche eine Determinante ungleich Null hat.

7. Transformationen zur Erzielung der Voraussetzungen

7.1 Beispiel einer einfachen Varianzanalyse

Tab. 7.1: Anzahl Unkräuter je Parzelle in einem vollständig randomisierten Versuch zum Vergleich von drei Herbiziden (A, B, C) und einer Kontrolle (D).

	A	B	C	D
	4	8	25	33
	5	11	28	21
	2	9	20	48
	5	12	15	18
	4	7	14	53
	1	7	30	31
\bar{y}_i	3,5	9	22	34
s_i^2	2,7	4,4	45,2	198,4

In der Tab. 7.1 sind Daten eines vollständig randomisierten Versuches zum Vergleich von 4 Herbizidbehandlungen wiedergegeben (Clewer und Scarisbrick, 2001, Practical statistics and experimental design for plant and crop science, Wiley, New York, S.215). Der Versuch hatte 6 Wiederholungen je Behandlung. Die Struktur und Art der Daten legt nahe, diesen Versuch mittels einer einfachen Varianzanalyse und anschließenden Mittelwertvergleichen auszuwerten. Hierbei würde das folgende lineare Modell zugrundegelegt:

$$y_{ij} = \mu + \tau_i + e_{ij}$$

wobei

y_{ij} = j -te Wiederholung des i -ten Herbizides

μ = Gesamteffekt

τ_i = Effekt der i -tes Herbizides

e_{ij} = Fehler von y_{ij}

Diese Auswertung macht verschiedene wichtige Annahmen bezüglich der Fehler e_{ij} :

- Normalverteilung
- Varianzhomogenität
- Statistische Unabhängigkeit

Die Gültigkeit der letzten Annahme wird durch eine adäquate Randomisation (zufällige Verteilung der 4 Behandlungen auf die 24 Versuchseinheiten = Parzellen) gewährleistet. Ein Blick auf die Daten zeigt allerdings, dass die Varianz der Beobachtungen sich zwischen den Behandlungen relativ deutlich unterscheidet, so dass die zweite Annahme verletzt ist. Näheren Aufschluss, auch hinsichtlich der ersten Annahme, bietet eine **Residuenanalyse** (vergl. Abschnitt 6.7).

Wie am Beispiel der linearen Regression können wir auch für die Varianzanalyse eine Residuenanalyse durchführen, um Abweichungen von den Voraussetzungen auszuspüren. Dieselben Methoden sind deswegen anwendbar, da der Varianzanalyse wie der linearen Regression einfache lineare Modelle zugrundeliegen. Im linearen Modell (Matrizenschreibweise)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

haben die Residuen die allgemeine Form

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

wobei

\mathbf{y} = Beobachtungsvektor

\mathbf{X} = Designmatrix

\mathbf{b} = Kleinstquadratschätzung des Parametervektors $\boldsymbol{\beta}$

Die Residuen sind Schätzwerte des Fehlervektors \mathbf{e} im linearen Modell. Im Fall des linearen Modells der Varianzanalyse haben die Residuen die Form

$$r_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

Wir müssen also im Fall der Varianzanalyse zur Berechnung der Residuen einfach von jeder Beobachtung den jeweiligen Behandlungsmittelwert abziehen. Außerdem können wir die Residuen auch "studentisieren" (siehe Abschnitt 7.2), was aber bei der einfachen Varianzanalyse nicht notwendig ist, da die "Roh-Residuen" r_{ij} bei Gültigkeit der Modellvoraussetzungen varianzhomogen sind (im Unterschied zur Regressionsanalyse).

In den Abbildungen 7.1 und 7.2 sind die Q-Q-Plots sowie Plots gegen den vorhergesagten Wert (hier: Mittelwert $\bar{y}_{i\cdot}$) angegeben. Beide Plots weisen auf eine Verletzung der Modellvoraussetzungen hin. Insbesondere der Plot gegen den vorhergesagten Wert (= Mittelwert) zeigt die Varianzheterogenität an (Varianz steigt mit dem Mittelwert). Infolge der Modellverletzung zeigt auch der Q-Q-Plot eine gewisse Anomalität, die eher auf die Varianzheterogenität als auf eine Abweichung von der Normalverteilung zurückzuführen ist.

Die Verletzung der Varianzhomogenitätsannahme hat zur Folge, dass die Varianzanalyse und nachfolgende Mittelwertvergleiche, z.B. mit dem LSD Test, nicht anwendbar sind. Die Verwendung einer gemeinsamen Grenzdifferenz für alle paarweisen Mittelwertvergleiche setzt nämlich voraus, dass alle Mittelwerte mit derselben Genauigkeit geschätzt werden, und dies ist nur bei homogenen Fehlervarianzen der Fall. Nur dann dürfen wir die einzelnen Stichprobenvarianzen der verschiedenen Behandlungen zu einem Mittelquadrat für den Fehler "poolen", wie wir es in der Varianzanalyse tun. Denn dort ist (Kap. 4)

$$MQ_{Fehler} = \frac{\sum_{i=1}^t s_i^2}{t}$$

wobei

s_i^2 = Stichprobenvarianz der i -ten Behandlung

t = Zahl der Behandlungen

Das MQ_{Fehler} ist ein Schätzwert für die gemeinsame Fehlervarianz σ^2 der Fehler e_{ij} im linearen Modell der Varianzanalyse. Unterscheiden sich nun die Varianzen der Behandlungen, so sind für jeden Vergleich von Behandlungen andere Fehlervarianzen und somit auch eine andere Grenzdifferenz (LSD) anzusetzen. Das Standardverfahren auf Basis einer gemeinsamen LSD wird damit ungültig.

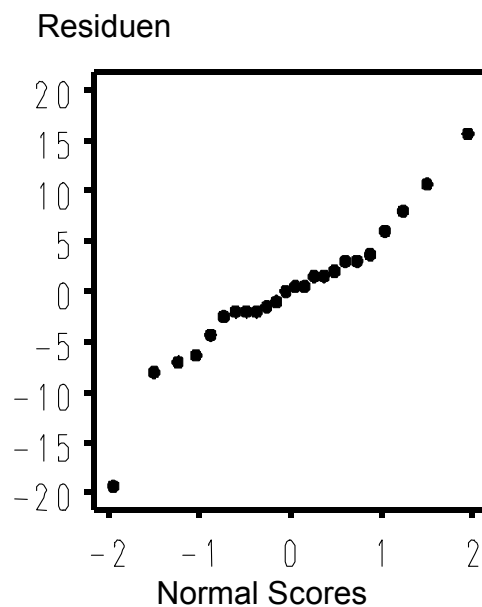


Abb. 7.1: Q-Q-Plot für Residuen der Varianzanalyse für die Herbizid-Daten. Untransformiert.

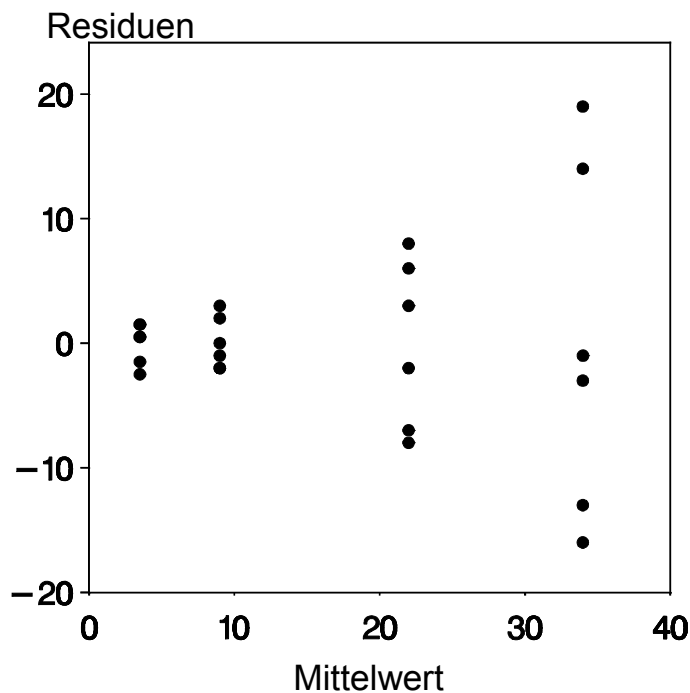


Abb. 7.2: Plot der Residuen gegen den vorhergesagten Wert (Mittelwert).
Untransformierte Daten.

Zur besseren Erfüllung der Voraussetzungen können die Daten transformiert werden. Für Zählwerte ohne feste Beschränkung nach oben, bei denen als Verteilungsmodell die Poisson-Verteilung anzusetzen ist (Abschnitt 5.3), kommen typischerweise die logarithmische und die Wurzel-Transformation in Frage. Die Plots in Abb. 7.3 bis 7.6 zeigen, dass beide Transformationen eine gewisse Varianzstabilisierung erreichen, wobei die log-Transformation etwas besser abschneidet. Allerdings ist der Q-Q-Plot der log-transformierten Daten etwas auffälliger als der für die wurzel-transformierten. Es zeigt sich hier ein generelles Problem bei der Verwendung von Daten-Transformationen: Es sollen mehrere Voraussetzungen gleichzeitig erfüllt werden, und das erwartet man von einer einzigen Transformation! Allerdings kann es sein, dass die eine Transformation besser für das Erreichen der Normalverteilung, eine andere aber besser für das Erreichen von Varianzhomogenität ist. In der Varianzanalyse ist Varianzhomogenität die wichtigere Annahme, so dass hier die log-Transformation vorzuziehen ist.

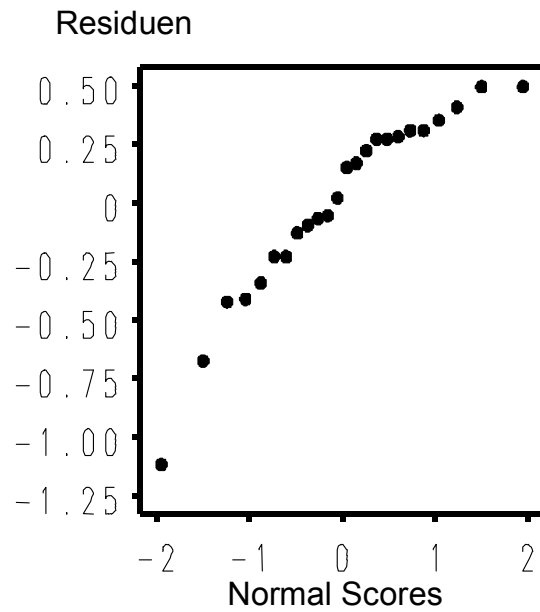


Abb. 7.3: Q-Q-Plot für Residuen der Varianzanalyse für die Herbizid-Daten. Log-transformiert.

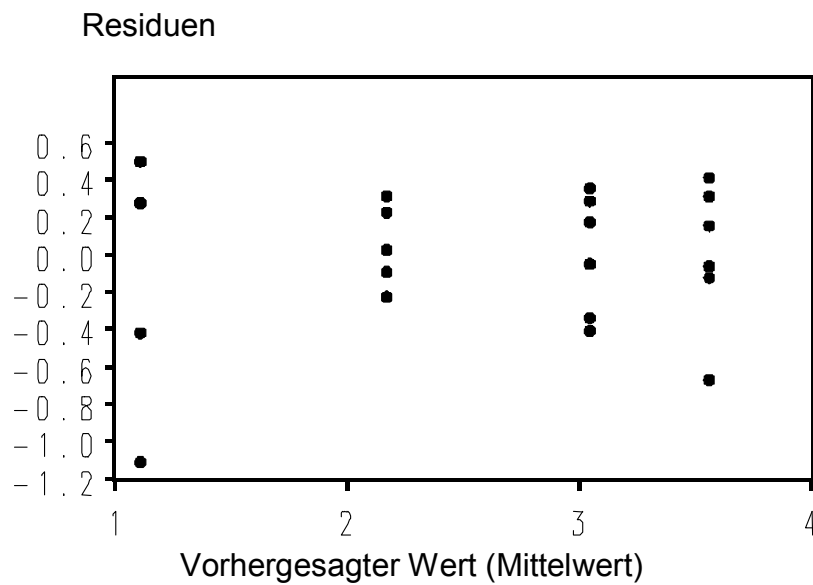


Abb. 7.4: Plot der Residuen gegen den vorhergesagten Wert (Mittelwert). Log-transformierte Daten.

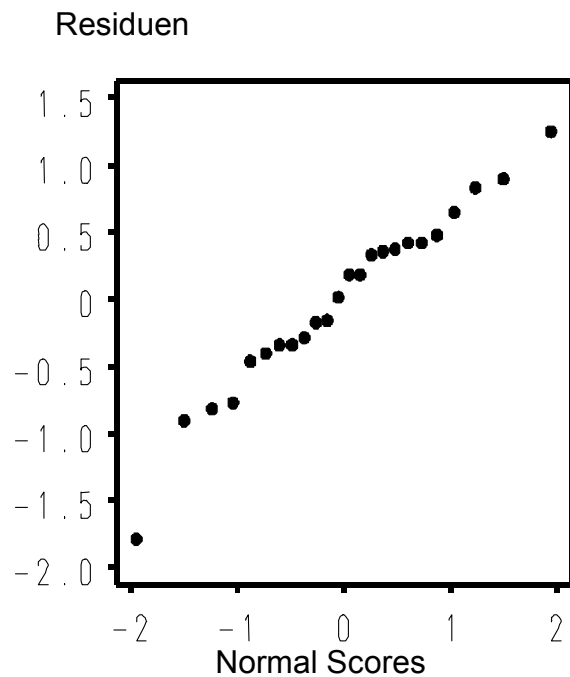


Abb. 7.5: Q-Q-Plot für Residuen der Varianzanalyse für die Herbizid-Daten. Wurzeltransformiert.

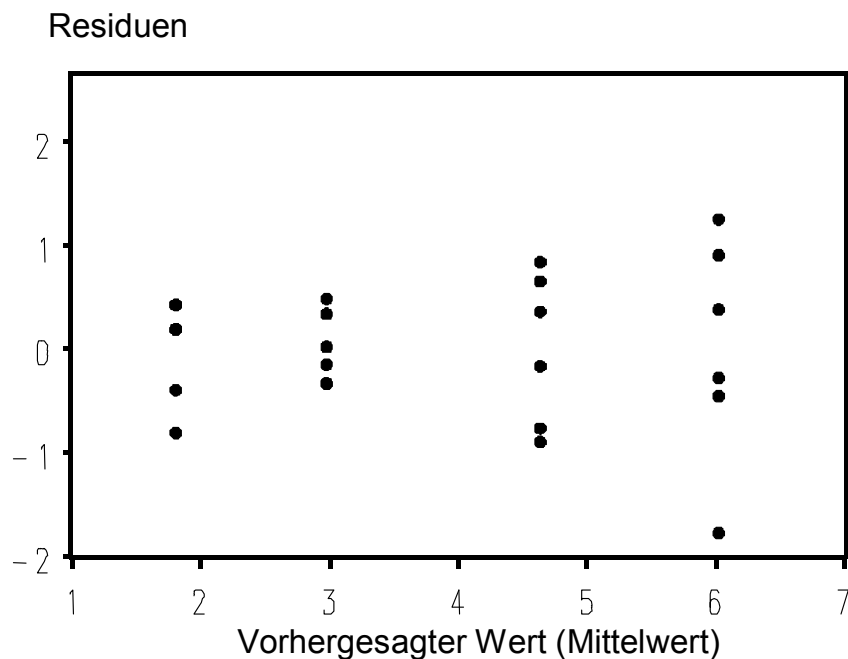


Abb. 7.6: Plot der Residuen gegen den vorhergesagten Wert (Mittelwert). Wurzeltransformierte Daten.

Im vorliegenden Fall können wir die Varianzheterogenität auch näher untersuchen, indem wir für jede Behandlung Mittelwert und Varianz berechnen und die funktionale Beziehung zwischen Varianz (bzw. Standardabweichung) und Mittelwert untersuchen (Tab. 7.2). Es zeigt sich, dass die Varianz mit dem Mittelwert steigt. Dabei ist die Standardabweichung etwa proportional zum Mittelwert, während die Varianz überproportional mit dem Mittelwert steigt. Dies zeigt an, dass eine log-Transformation die Varianz stabilisiert, denn es gilt:

Varianz stabilisiert durch:	
Varianz proportional zu Mittelwert	⇒ Wurzel-Transformation
Standardabweichung proportional zu Mittelwert	⇒ Log-Transformation

Tab. 7.2: Beziehung zwischen Anzahl Unkräuter je Parzelle in einem vollständig randomisierten Versuch zum Vergleich von drei Herbiziden (A, B, C) und einer Kontrolle (D).

	A	B	C	D
\bar{y}_i	3,5	9	22	34
s_i^2	2,7	4,4	45,2	198,4
s_i	1,643	2,098	6,723	14,09
s_i^2 / \bar{y}_i	0,77	0,49	2,05	5,84
s_i / \bar{y}_i	0,47	0,23	0,31	0,41

Für theoretisch Interessierte: Dieser Leitsatz kann mit Hilfe der sog. Delta-Methode abgeleitet werden, welche auf einer Taylor-Reihen-Entwicklung einer Zufallsvariable um ihren Erwartungswert beruht (siehe Anhang F). Anwendung dieser Methode führt zu der Feststellung, dass die Varianz einer mit der Funktion $f(y)$ transformierten Zufallsvariable mit der Varianz der untransformierten Zufallsvariable y näherungsweise wie folgt zusammenhängt (Der Einfachheit halber wird in der Notation nicht zwischen einer Zufallsvariable - i.d.R. Grossbuchstaben - und ihrer Realisation - i.d.R. Kleinbuchstaben - unterschieden):

$$\text{var}[f(y)] \approx \left(\frac{df}{dy} \right)_{y=E(y)}^2 \text{var}(y)$$

wobei $E(y)$ der Erwartungswert (Mittelwert) von y ist.

Beispiel: Logarithmische Transformation

$$f(y) = \log(y)$$

$$\left(\frac{df}{dy} \right)_{y=E(y)} = \frac{1}{E(y)}$$

$$\text{var}[f(y)] \approx \frac{\text{var}(y)}{(E(y))^2}$$

Hieraus folgt, dass bei Proportionalität von $\text{var}(y)$ und $[E(y)]^2$, also von Standardabweichung (Wurzel aus Varianz) und Mittelwert $[E(y)]$ die log-Transformation eine näherungsweise Varianzstabilisierung erzielt.

Beispiel: Wurzel-Transformation

$$f(y) = \sqrt{y}$$

$$\left(\frac{df}{dy}\right)_{y=E(y)} = \frac{1}{2\sqrt{E(y)}}$$

$$\text{var}[f(y)] \approx \frac{\text{var}(y)}{4E(y)}$$

Hieraus folgt, dass bei Proportionalität von Varianz $\text{var}(y)$ und Mittelwert $E(y)$ die Wurzel-Transformation eine näherungsweise Varianzstabilisierung erzielt.

Da bei den Herbizid-Daten näherungsweise Proportionalität zwischen Standardabweichung und Mittelwert besteht, spricht auch dieser Befund für die log-Transformation.

Tab. 7.3: Logarithmisch transformierte Anzahl Unkräuter je Parzelle in einem vollständig randomisierten Versuch zum Vergleich von drei Herbiziden (A, B, C) und einer Kontrolle (D).

	A	B	C	D
	1,39	2,08	3,22	3,50
	1,61	2,40	3,33	3,04
	0,69	2,20	3,00	3,87
	1,61	2,48	2,71	2,89
	1,39	1,95	2,64	3,97
	0	1,95	3,40	3,43
\bar{y}_i	1,11	2,18	3,05	3,45
s_i^2	0,41	0,05	0,10	0,19

Wir führen nun die Varianzanalyse mit den logarithmisch transformierten Werten (Tab. 7.5) durch. Die Stichprobenvarianzen weisen noch eine gewisse Heterogenität auf, was aber wahrscheinlich v.a. am kleinen Stichprobenumfang liegt. Zum Vergleich die Ergebnisse der Varianzanalyse für untransformierte Daten. Der F_{Vers} -Wert bei log-transformierten Daten ist etwas größer geworden. Bei den transformierten Daten finden wir signifikante Mittelwertunterschiede zwischen allen Behandlungen, außer bei dem Paar C, D (Tab. 7.6). Die Herbizide A und B führen also zu einer signifikanten Reduktion des Unkrautbesatzes, das Herbizid C dagegen nicht.

Führen wir fälschlicherweise die Mittelwertvergleiche mit den untransformierten Daten durch (Tab. 7.7), so finden wir im Gegensatz zur Auswertung der transformierten Daten signifikante Unterschiede zwischen C und D, nicht aber zwischen A und B. Nach dieser Auswertung wäre Herbizid C besser als die Kontrolle. Und die Wirkung von A und B wäre nicht zu unterscheiden. Damit weicht das Ergebnis für die untransformierten Daten deutlich von dem bei Auswertung der transformierten Daten ab. Das Beispiel zeigt, dass eine Verletzung der Voraussetzungen zu statistischen Fehlschlüssen führen kann.

Zur Interpretation ist es oft hilfreich, neben den Mittelwerten für die transformierten Daten auch die Mittelwerte der untransformierten Daten darzustellen. Die statistische Auswertung, insbesondere Varianzanalyse und Mittelwertvergleiche erfolgen dann mit den transformierten Daten, während die Interpretation mit Hilfe der Mittelwerte der Originaldaten erfolgen kann (Tab. 7.6). Für die Berechnung der Mittelwerte auf der untransformierten Skala gibt es zwei Optionen:

- (1) Direkte Mittelwertbildung der untransformierten Daten (nur bei einfachen und balancierten Versuchsanlagen). Dies liefert einen Schätzwert des Erwartungswertes.
- (2) Rücktransformation der "transformierten" Mittelwerte (auch bei komplexeren und unbalancierten Datenstrukturen). Dies liefert einen Schätzwert des Medians auf der untransformierten Skala, aber nicht des Mittelwertes (Erwartungswertes) auf der untransformierten Skala!

Man sieht in Tab. 7.6, dass die Medianschätzungen unterhalb der Mittelwert-schätzungen liegen. Dies ist zu erwarten bei rechtsschiefen Verteilungen. Wenn die log-Transformation Normalverteilung erzielt, so folgen die Ausgangsdaten einer log-Normalverteilung, und diese ist rechtsschief.

Tab. 7.4: Varianzanalyse für untransformierte Daten.

Ursache	FG	SQ	MQ	F_{Vers}	$^{\S}p\text{-Wert}$
Behandlungen	3	3361	1120,38	17,88	<0,0001
Fehler	20	1254	62,68		

§ siehe Anhang D

Tab. 7.5: Varianzanalyse für log-transformierte Daten.

Ursache	FG	SQ	MQ	F_{Vers}	$^{\S}p\text{-Wert}$
Behandlungen	3	19,33	6,44	34,22	<0,0001
Fehler	20	3,77	0,19		

§ siehe Anhang D

Tab. 7.6: Mittelwertvergleiche für log-transformierte Daten. Mittelwerte, die mit einem gemeinsamen Buchstaben versehen sind, sind nicht signifikant verschieden.

Behandlung	Mittelwert (log-Skala)	Mittelwert (untransformiert)	Median [§] (untransformiert)
A	1,11 ^c	3,5	3,03
B	2,18 ^b	9,0	8,85
C	3,05 ^a	22,0	21,12
D	3,45 ^a	34,0	31,50
LSD(5%)	0,52		

§ Berechnet durch Rücktransformation der Mittelwerte auf der log-Skala.

Tab. 7.7: Mittelwertvergleiche für untransformierte Daten bei (falscher!!) Annahme der Varianzhomogenität. Mittelwerte, die mit demselben Buchstaben versehen sind, sind nicht signifikant verschieden. **Diese Mittelwertvergleiche beruhen auf falschen Voraussetzungen.**

Behandlung	Mittelwert
A	3,5 ^c
B	9,0 ^c
C	22,0 ^b
D	34,0 ^a
LSD(5%)	9,53

Das anhand des Beispiels erläuterte Vorgehen bei varianzanalytischen Auswertungen kann wie folgt zusammengefasst werden:

- Suche eine geeignete Transformation zur Erfüllung der Voraussetzungen (Residuenanalyse)
- Statistische Auswertung (Varianzanalyse, Mittelwertvergleiche) für transformierte Daten
- Zur besseren Interpretation auch Mittelwerte für untransformierte Daten berechnen und/oder Median durch Rücktransformation der Mittelwerte auf transformierter Skala

7.2 Studentisierte Residuen im allgemeinen linearen Modell

Eine Überprüfung der Modellvoraussetzungen im linearen Modell kann generell mit Hilfe der Residuen erfolgen. In der Regel ist es sinnvoll, die Residuen zu "studentisieren", wie bei der linearen Regression (Abschnitt 6.7). Dies ist notwendig, weil die Roh-Residuen (Annahme: X von vollen Rang)

$$r = y - Xb = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y$$

selbst bei homogener Varianz der Beobachtungen eine Varianzheterogenität aufweisen können. Es gilt:

$$\text{var}(r) = P\sigma^2$$

wobei

$$P = I - X(X'X)^{-1}X'$$

Die Diagonalelemente der Matrix P sind also proportional zur Varianz der Residuen. Bezeichne das i -te Diagonalelement von P mit p_{ii} und das i -te Element des Residuenvektors mit r_i . Das i -te studentisierte Residuum berechnet sich durch

Division des Roh-Residuums durch seine Standardabweichung (Wurzel aus seiner Varianz). Die Formel lautet

$$c_i = \frac{r_i}{s\sqrt{p_{ii}}}$$

(Für σ wurde hier der Schätzer s eingesetzt). Computerprogramme für lineare Modelle berechnen Residuen nach dieser allgemeinen Formel. Hiermit kann man für jedes beliebige lineare Modell leicht die studentisierten Residuen berechnen und somit eine graphische Überprüfung der Modellvoraussetzungen vornehmen.

Beispiel: Für die einfache lineare Regression findet man (vergleiche 6.7 zu Residuen und Resultate in 6.8.2) nach einiger algebraischer Umformung

$$p_{ii} = 1 - \left[\frac{1}{n} + \frac{(x_i - \bar{x}_\cdot)^2}{SQ_x} \right]$$

so dass

$$c_i = \frac{r_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x}_\cdot)^2}{SQ_x}}}$$

Beispiel: In Abschnitt 6.10 hatten wir eine multiple lineare Regression für einen Versuch mit Ratten durchgeführt, bei dem das Endgewicht (y_i ; in g) von 35 Tieren in Abhängigkeit vom Anfangsgewicht (x_{1i} ; in g) und vom Futterverzehr (x_{2i} ; in g) untersucht wurde. Die Daten waren wie folgt (Linder und Berchtold II, S. 125):

x_1	x_2	y
(g)	(g)	(g)
55,8	289	114,8
45,8	316	109,7
48,1	304	111,3
43,3	299	126,0
50,1	353	144,7
40,1	298	121,0
47,1	303	125,2
51,0	312	113,7
53,7	333	138,5
41,2	280	105,8
40,2	287	117,7
46,4	338	140,0
45,9	298	117,1
38,0	302	103,0
56,0	355	137,3
32,4	307	109,7

37,5	342	136,3
45,9	310	121,2
40,7	280	104,5
36,4	283	104,0
46,9	305	120,2
42,2	296	120,5
43,4	290	118,9
45,0	224	126,4
43,8	353	125,4
47,8	282	109,4
50,4	288	121,2
37,9	266	109,2
46,0	318	141,1
42,8	335	131,3
50,7	304	128,5
59,6	296	138,8
43,8	292	109,0
65,4	320	113,7
39,3	305	116,2

Wir hatten nach dem Modell

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

ausgewertet. Mit einem PC Programm für lineare Modelle (hier: SAS Prozedur GLM; siehe Anhang) berechnen wir nun die studentisierten Residuen c_i sowie den vorhergesagten Wert

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i}$$

wobei a , b_1 und b_2 die Kleinst-Quadrat-Schätzungen der Parameter α , β_1 und β_2 sind.

Tab. 7.8: Studentisierte Residuen und vorhergesagter Wert für Ratten-Daten. Berechnet mit SAS Prozedur GLM.

y_i	x_{1i}	x_{2i}	c_i	$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i}$
114,8	55,8	289	-0,77215	121,909
109,7	45,8	316	-1,41988	123,389
111,3	48,1	304	-1,08709	121,791
126,0	43,3	299	0,76716	118,598
144,7	50,1	353	1,24280	133,316
121,0	40,1	298	0,42039	116,977
125,2	47,1	303	0,42077	121,135
113,7	51,0	312	-1,15867	124,801
138,5	53,7	333	0,84867	130,550
105,8	41,2	280	-0,81467	113,548
117,7	40,2	287	0,32183	114,630
140,0	46,4	338	1,22692	128,433
117,1	45,9	298	-0,25071	119,522

103,0	38,0	302	-1,46926	116,925
137,3	56,0	355	0,10745	136,339
109,7	32,4	307	-0,64683	115,554
136,3	37,5	342	1,21268	125,398
121,2	45,9	310	-0,09617	122,129
104,5	40,7	280	-0,92937	113,329
104,0	36,4	283	-0,86523	112,094
120,2	46,9	305	-0,13259	121,481
120,5	42,2	296	0,31550	117,464
118,9	43,4	290	0,23018	116,686
126,4	45,0	224	2,87404	103,046
125,4	43,8	353	-0,56718	130,552
109,4	47,8	282	-0,78540	116,878
121,2	50,4	288	0,19752	119,323
109,2	37,9	266	0,01537	109,058
141,1	46,0	318	1,78467	123,912
131,3	42,8	335	0,54249	126,202
128,5	50,7	304	0,58089	122,931
138,8	59,6	296	1,53195	125,098
109,0	43,8	292	-0,86139	117,296
113,7	65,4	320	-2,28647	132,858
116,2	39,3	305	-0,20428	118,147

Der Q-Q-Plot in Abb. 7.7 sowie der Plot der Residuen gegen den vorhergesagten Wert (Abb. 7.8) weisen keine Auffälligkeiten auf (vergleiche Abschnitt 6.7), abgesehen von einem etwas abgelegenen Wert in letzterem Plot. Diese Beobachtung könnte man etwas genauer untersuchen und z.B. prüfen, ob sich das angepasste Modell wesentlich ändert, wenn man ihn weglässt. Dies ist allerdings nicht der Fall (Ergebnisse nicht gezeigt). Die Residuenanalyse weist somit nicht auf gravierende Abweichungen von den Modellvoraussetzungen hin. Somit ist die statistische Auswertung in Abschnitt 6.10 valide; eine Daten-Transformation ist nicht angezeigt.

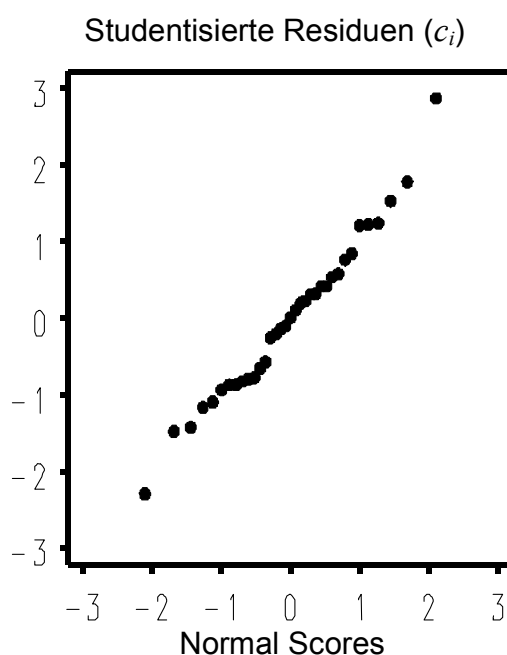


Abb. 7.7: Q-Q-Plot für Versuch mit Ratten für Zielvariable Gewicht.

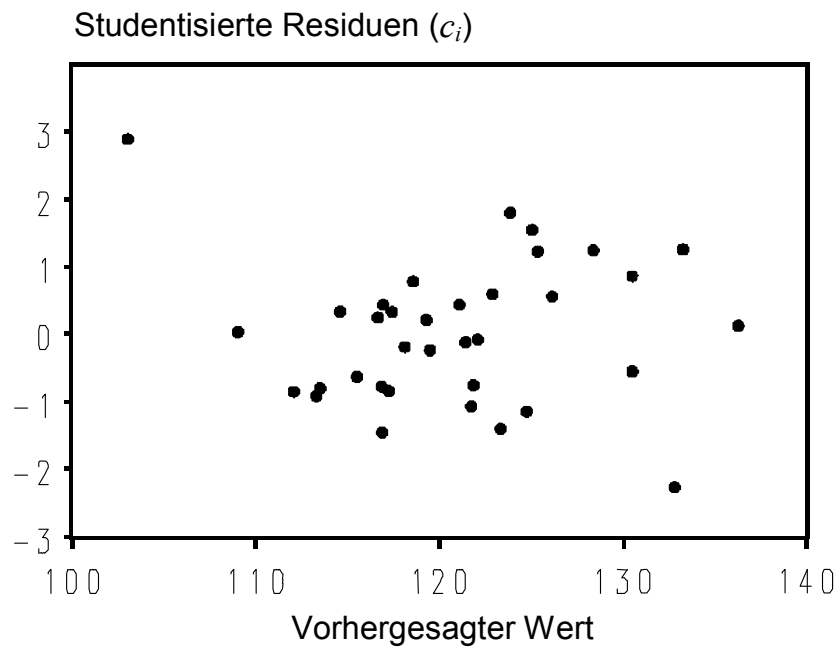


Abb. 7.8: Plot gegen den vorhergesagten Wert [$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i}$] für Versuch mit Ratten.

SAS-Anweisungen

```
data d;
input x1 x2 y;
datalines;
55.8 289 114.8
45.8 316 109.7
48.1 304 111.3
43.3 299 126.0
50.1 353 144.7
40.1 298 121.0
47.1 303 125.2
51.0 312 113.7
53.7 333 138.5
41.2 280 105.8
40.2 287 117.7
46.4 338 140.0
45.9 298 117.1
38.0 302 103.0
56.0 355 137.3
32.4 307 109.7
37.5 342 136.3
45.9 310 121.2
40.7 280 104.5
36.4 283 104.0
46.9 305 120.2
42.2 296 120.5
43.4 290 118.9
```

45.0	224	126.4
43.8	353	125.4
47.8	282	109.4
50.4	288	121.2
37.9	266	109.2
46.0	318	141.1
42.8	335	131.3
50.7	304	128.5
59.6	296	138.8
43.8	292	109.0
65.4	320	113.7
39.3	305	116.2

```

proc glm data=d;
model y=x1 x2;
output out=res student=student p=p;

```

Berechnet studentisierte
Residuen

```

proc print data=res noobs;
var y x1 x2 student p;

```

```

/*Q-Q-Plot*/
proc univariate normal;
var student;
qqplot student/normal;

```

```

/*Plot gegen den Vorhergesagten Wert*/
proc gplot;
plot student*p;
run;

```

7.3 Einige gängige Transformationen

Im Beispiel in Abschnitt 7.1 hatten wir als zu analysierendes Merkmal einen nach oben unbegrenzten Zählwert. Hier ist als Verteilungstyp zunächst an die Poissonverteilung zu denken. Für diesen Verteilungstyp kommt die **Wurzel-Transformation** und die **logarithmische Transformation** in Betracht.

Bei begrenzten Zählwerten bzw. Anteilen (z.B. „ $m = 10$ von $n = 20$ Pflanzen waren befallen mit einem Pilz“) ist die Binomialverteilung das naheliegende Verteilungsmodell. Hier wird häufig die **Arcus-Sinus-Wurzel-Transformation** verwendet. Diese wird auch als **Winkel-Transformation** bezeichnet.

Bei metrischen Merkmalen, welche einen weiten Wertebereich umspannen, ist häufig eine Zunahme der Varianz mit dem Erwartungswert zu beobachten. In solchen Fällen folgen die Daten häufig einer Log-Normalverteilung, so dass eine **logarithmische Transformation** zur Varianzstabilisierung und Erzielung einer annähernden Normalverteilung hilfreich ist.

Tab. 7.9: Einige varianzstabilisierende Transformationen (nach Chatterjee und Price)

Verteilung	Varianz von y als Funktion des Mittelwertes μ	Transformation
Poisson ("nach oben offene Zählwerte")	μ	\sqrt{y} oder $\sqrt{y} + \sqrt{y+1}$ $\log(y)^{\$}$
Binomial ("begrenzte Zählwerte"; $y = m/n$)	$\mu(1-\mu)/n$	$\sin^{-1} \sqrt{y}^*$
Log-normal (metrisch; großer Wertebereich)	$\approx \mu^2$	$\log(y)^{\$}$

* besser: $\sin^{-1} \sqrt{\frac{m+3/8}{n+3/4}}$ -> Normalverteilung und stabile Varianzen. Die Umkehrfunktion kann

entweder im Gradmaß oder im Bogenmaß berechnet werden. Beide Varianzen unterscheiden sich nur durch einen Skalierungsfaktor $360/(2\pi)$; die resultierende Varianzanalyse ist identisch. Wenn die Stichprobenumfänge n für die berechneten Anteile $y = m/n$ nicht konstant sind, ist die Varianz der transformierten Variable nicht konstant, sondern umgekehrt proportional zu n , so dass eine gewichtete Analyse sinnvoll ist (McCullagh & Nelder, 1989; Exercise 4.8).

§: Wenn $y = 0$ vorkommt, eine kleine Zahl zu allen Werten hinzuaddieren, z.B. 1.

Darüber hinaus gibt es eine große Zahl alternativer Transformationen, die fallweise verwendet werden können. In jedem Fall ist eine Residuenanalyse hilfreich bei der Identifikation einer geeigneten Transformation.

Das Vorgehen bei einer varianzanalytischen Auswertung mit transformierten Daten entspricht im wesentlichen immer dem in Abschnitt 7.1 angegebenen Schema.

7.4 Generalisierte lineare Modelle

Mit der Transformation der Daten möchte man die Modellvoraussetzungen möglichst gut erfüllen, u.a.

- Linearität/Additivität des Modells
- Varianzhomogenität
- Normalverteilung

Oft ist es schwierig, eine Transformation zu finden, die alle drei Ziele erreicht. Eine Alternative besteht in der Verwendung sog. generalisierter linearer Modelle (GLM). Diese lassen andere Verteilungen als die Normalverteilung sowie Varianzheterogenität zu. Bei der Suche nach einer geeigneten Transformation kann man sich daher ausschließlich auf die Erzielung der Linearität/Additivität konzentrieren. In die Klasse der GLMs fallen eine ganze Reihe gängiger Modelle, so die Logit- und Probit-Regression für binomialverteilte Daten oder die loglinearen Modelle für poissonverteilte Daten. Näheres findet man bei McCullagh P. und Nelder J. 1989 Generalized linear models. 2nd edition. Chapman and Hall, London.

Hier nur ein kurzer Abriß. Bei Annahme der Normalverteilung lautet das lineare Modell $y = X\beta + e$. Im Falle einer Datentransformation $h(y)$ hoffen wir, dass die transformierten Daten die Voraussetzungen erfüllen, so dass wir das Modell

$$h(y) = \eta = X\beta$$

verwenden können. Bei einem GLM werden dagegen nicht die Daten transformiert, sondern nur der Erwartungswert der Zufallsvariable y . Diese Transformation wird dann dem **linearen Prädiktor**

$$\eta = X\beta$$

gleichgesetzt, so dass

$$g[E(y)] = \eta = X\beta$$

Die Transformation $g()$ des linearen Prädiktors wird als **Link-Funktion** bezeichnet. Durch Anwendung der Umkehrfunktion der Link-Funktion finden wir für den Erwartungswert:

$$E(y) = g^{-1}(\eta) = g^{-1}(X\beta) .$$

Das Modell bildet also den Erwartungswert als eine nicht-lineare Funktion des linearen Prädiktors ab. Die Daten y werden dann so modelliert, dass sie den Erwartungswert $g^{-1}(\eta)$ haben, wobei die Verteilung nun eine andere als die Normalverteilung sein darf, z.B. eine Binomial- oder eine Poisson-Verteilung. Damit hat man eine deutlich größere Flexibilität der Modellierung als beim einfachen linearen Modell, wo eine Datentransformation alle Voraussetzungen gleichzeitig erfüllen muss. Hier können wir uns auf die Annahme der Linearität/Additivität konzentrieren, dabei aber andere Verteilungen zulassen, die dann eine jeweils spezielle Form der Varianzheterogenität zulassen. Bei der Poissonverteilung gilt beispielsweise, dass der Erwartungswert der Varianz entspricht. Wir können die Klasse der GLMs hier nicht vertiefen. Es sollte aber klar geworden sein, dass GLMs bei Verletzungen der üblichen Voraussetzungen der Varianzhomogenität und Normalverteilung eine interessante Alternative zu Datentransformationen sind.

8. Versuchsanlagen

Am Anfang der Versuchsplanung steht die Festlegung der Versuchsfrage: Was will ich mit meinem Experiment herausbekommen? Ausgehend von der Versuchsfrage muss entschieden werden, welche **Prüffaktoren** oder **Behandlungsfaktoren** in einem Versuch untersucht werden sollen. Prüffaktoren sind die Einflussfaktoren, deren Wirkung auf interessierende Zielvariablen untersucht werden soll. Neben dem Prüffaktor selbst muss entschieden werden, welche **Faktorstufen** dieses Faktors untersucht werden sollen. Man spricht auch einfach von **Behandlungen**.

Beispiel: In einem Experiment soll die Wirkung verschiedener Düngervarianten auf das Wachstum von Tomaten untersucht werden. Der Prüffaktor ist hier der Dünger, und die Stufen dieses Faktors sind die im Versuch geprüften verschiedenen Düngervarianten.

Beispiel: In einem Keimversuch werden verschiedene Saatgutbeizungsverfahren bezüglich ihrer Wirkung auf die Keimung von Rapssamen untersucht. Der Prüffaktor ist die Saatgutbeizung und seine Stufen sind die geprüften einzelnen Beizungsverfahren.

Für die Anlage und Durchführung eines Versuchs ist es weiterhin wichtig zu definieren, was die **Randomisationseinheiten** sind, mit denen der Versuch durchgeführt wird. Diese Frage ist zu trennen von der Frage, welche **Behandlungen** in dem Experiment geprüft werden sollen und somit den Versuchseinheiten zugeordnet werden müssen. Für die Planung eines Versuches ist es generell von Vorteil, die beiden Aspekte getrennt zu betrachten, also die Frage, welche Behandlungen geprüft werden sollen und dann die Frage, was die Randomisationseinheiten sind.

Randomisationseinheiten sind denjenigen Einheiten, denen die Stufen der Prüffaktoren gemäß des verwendeten Versuchsplans zufällig zuordnet werden. In der Regel entspricht die Randomisationseinheit außerdem der **Beobachtungseinheit**, also derjenigen Einheit, an der Messungen und Beobachtungen vorgenommen werden.

Beispiel: Es wird ein Fütterungsexperiment mit Schweinen durchgeführt, in dem 5 verschiedene Futtermittelzusätze geprüft werden sollen. Dies sind die 5 Behandlungen. Als Randomisationseinheit kommt zum einen das einzelne Tier in Frage. Es ist allerdings zu berücksichtigen, dass unter Praxisbedingungen die Tiere meist in Gruppen gehalten werden. Damit die Versuchsergebnisse auf die Praxis übertragbar sind, spricht daher vieles dafür, eine Gruppe von Tieren, die gemeinsam in einer Bucht gehalten werden, als Randomisationseinheit zu betrachten. Wird dagegen das Einzeltier als Randomisationseinheit verwendet, so haben die Ergebnisse meist eher Grundlageneigenschaften.

Beispiel: In einem Gewächshausversuch wird die Wirkung dreier verschiedener Düngerbehandlungen auf den Ertrag von Tomaten untersucht. Wenn der Versuch als Gefäßversuch mit einer Tomate pro Gefäß durchgeführt wird, so kann die Einzelpflanze als Versuchseinheit verwendet werden. Werden dagegen mehrere Tomatenpflanzen, zum Beispiel zehn, in größeren Gefäßen (Kisten) gepflanzt, so muss berücksichtigt werden, dass es Nachbarschaftseffekte von Pflanzen in derselben Kiste gibt. Desweiteren wäre es problematisch, bei verschiedenen Pflanzen in derselben Kiste verschiedene Dünger anzuwenden, weil die Dünger sich in der Bodenlösung verteilen können und so Nachbarpflanzen beeinflussen könnten. In dieser Situation muss die ganze Kiste als Randomisationseinheit betrachtet werden.

Dieses Kapitel behandelt drei Prinzipien der Versuchsplanung:

- Wiederholung (muss)
- Randomisation (muss)
- Blockbildung (kann)

Es werden die **randomisierte vollständige Blockanlage** sowie das **Lateinische Quadrat** vorgestellt, die zwei populärsten Versuchsanlagen mit Blockbildung. Hierbei besprechen wir sowohl die Anlage (Randomisation) als auch die Auswertung. Außerdem behandeln wir Anlagen mit unvollständigen oder ungleich großen Blöcken.

8.1 Wiederholung

Ohne Wiederholungen ist es schwierig, verlässliche Aussagen über Behandlungen zu treffen. Stellen wir uns vor, 2 Sorten A und B werden geprüft. Dazu sät Landwirt X die Sorte A auf einem Feld und Landwirt Y sät die Sorte B. Die Erträge seien 75.4 dt/ha für Sorte A und 82.6 für Sorte B. Ob dies jedoch bedeutet, dass die Sorte B die bessere ist, lässt sich nicht sagen, da die Sorten auf verschiedenen Feldern standen und da die Felder von verschiedenen Landwirten bewirtschaftet wurden. Offensichtlich wird der Ertrag nicht nur von der Sorte, also der Behandlung, beeinflusst, sondern wesentlich auch von vielfältigen Umweltfaktoren. Um Umwelt- von Behandlungseffekten trennen zu können, ist es notwendig, dass eine Behandlung auf mehr als einer Randomisationseinheit geprüft wird. Nehmen wir also an, dass Sorte A auf drei Feldern bei drei Landwirten geprüft wurde mit den Erträgen 75.4, 71.2 und 79.5 dt/ha und dass die Sorte B bei drei anderen Landwirten die Erträge 82.6, 89.1 und 93.2 dt/ha erbrachte. Diese Ergebnisse bieten eine etwas bessere Grundlage für die Einschätzung, dass Sorte B die bessere sein könnte, obschon die umweltbedingte Variabilität beträchtlich ist. Genauerer Aufschluß bietet ein Signifikanztest, der die Mittelwertdifferenz ins Verhältnis zur umweltbedingten Streuung setzt: nur wenn der Mittelwertunterschied groß ist im Vergleich zur Streuung der Daten, kann auf einen echten Sortenunterschied geschlossen werden.

Die Wiederholungen müssen "**echte**" Wiederholungen sein. Das ist in der Regel gleichbedeutend mit Randomisationseinheiten, also denjenigen Einheiten, denen bei der Randomisation (Abschnitt 8.2) die Behandlungen zufällig zugeordnet werden.

Zahl der Wiederholungen

Die Zahl der Wiederholungen für eine Behandlung ist gleich der Zahl der Randomisationseinheiten, denen die Behandlung zufällig zugeordnet wurde

Beispiel (unechte Wiederholungen): Es wird ein Versuch mit Kopfsalat zum Vergleich dreier verschiedener Herbizide angelegt. Für jede Behandlung wird eine Parzelle angelegt. Je Parzelle werden 10 Salatpflanzen gepflanzt. Für jeden Salatkopf wird der Ertrag gemessen. Somit liegen je Behandlung 10 Werte vor. Der Salatkopf ist die Beobachtungseinheit.

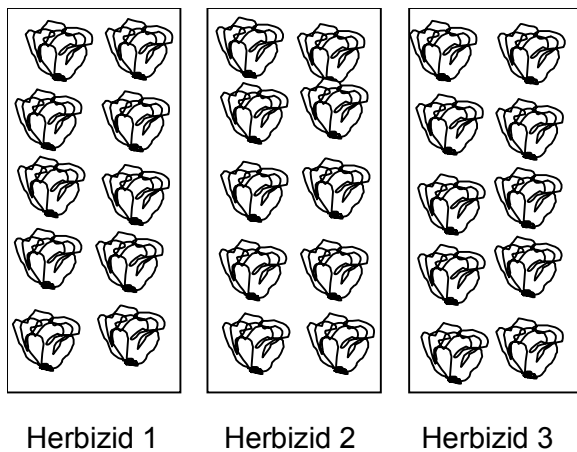
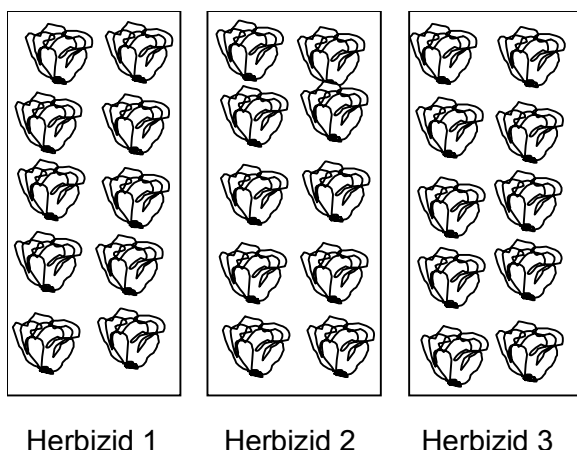


Abb. 8.1: Ein Plan ohne echte Wiederholung.

Wir könnten daher eine Varianzanalyse zum Vergleich der drei Behandlungen durchführen, wobei die Salatköpfe als Wiederholungen betrachtet werden. Wenn hierbei signifikante Unterschiede zwischen den Herbizid-Mittelwerten nachgewiesen werden, wissen wir aufgrund der inadäquaten Versuchsanlage allerdings nicht, ob diese durch Bodenunterschiede zwischen den drei Parzellen oder durch die Herbizide selbst verursacht sind. Bei dieser Versuchsanlage liegt eine Vermengung von Behandlungs- und Parzellenunterschieden vor. Das Problem besteht hier darin, dass die Salatköpfe keine Randomisationseinheiten sind, sondern nur Beobachtungseinheiten. **Die Randomisationseinheit ist hier die Parzelle, und daher liegt nur eine Wiederholung pro Behandlung vor.**

Um Parzellen- von Behandlungsunterschieden zu trennen, ist es notwendig, je Behandlung mehrere Parzellen zu prüfen (mindestens 2) und die Behandlungen zufällig den Parzellen zuzuordnen (Randomisation). Die Parzellen stellen dann Randomisationseinheiten und damit "echte" Wiederholungen dar, während die Salatköpfe je Parzelle "**unechte**" Wiederholungen sind, d.h. es handelt sich um Stichproben von den echten Wiederholungen.



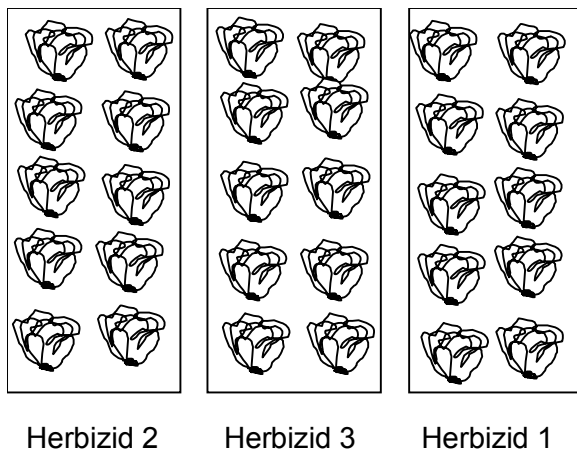


Abb. 8.2: Ein Plan mit zwei (!) echten Wiederholungen je Behandlung (Herbizid).

Für die Auswertung muss man zunächst die Erträge der Köpfe je Parzelle mitteln. Im zweiten Schritt werden dann die **Parzellenmittelwerte** varianzanalytisch ausgewertet. Eine Auswertung der Einzelwerte je Kopf mit einer einfachen Varianzanalyse ist **nicht** zulässig, weil dabei unechte Wiederholungen wie echte Wiederholungen behandelt würden.

Besonders muss man bei mehrstufigen Untersuchungen (z.B. Feldversuch und Laboranalysen) aufpassen, dass durch Bildung von Mischproben nicht wesentliche Fehlerquellen unter den Tisch gekehrt werden.

Beispiel (Falsche Mischprobenbildung): In einem Feldversuch werden drei verschiedene Bodenbearbeitungsmaßnahmen bei Weizen in einer vollständig randomisierten Anlage mit fünf Wiederholungen geprüft. Ziel des Versuches ist es, die Verfügbarkeit von Nährstoffen im Boden nach der Ernte in Abhängigkeit von der Bodenbearbeitung zu erfassen. Nach der Ernte wird auf jeder Parzelle eine Bodenprobe gezogen, um den Gehalt verschiedener Nährstoffe zu ermitteln. Die Versuchsansteller stellen fest, dass eine relativ große Streuung zwischen den Parzellen besteht. Aus diesem Grund fassen sie die Bodenproben aus den fünf Wiederholungen einer Behandlung zu einer Mischprobe zusammen, homogenisieren diese, „um den Versuchsfehler zu reduzieren“, und teilen sie dann für die anschließenden Laboruntersuchungen jeweils wieder in fünf Teilproben auf. Am Ende der Analysen liegen dann für jede Behandlung und jedes Merkmal fünf Analysewerte vor, die varianzanalytisch ausgewertet werden.

Um es vorweg zu nehmen: Das Vorgehen der Versuchsansteller beinhaltet einen gravierenden Denkfehler. Bei den fünf Labor-Werten je Behandlung handelt es sich um keine „echten“ Wiederholungen, weil der Versuchsfehler aus dem Feldversuch nicht mehr erfassbar ist. Es wird lediglich der Analysefehler aus dem Labor erfasst. In dem extremen, aber nicht unwahrscheinlichen Fall, dass der Versuchsfehler im Feld groß und der Analysefehler im Labor klein ist, wird man sehr leicht hoch signifikante Unterschiede zwischen den Mischproben nachweisen können. Diese können aber nicht mehr ursächlich auf die unterschiedlichen Behandlungen zurückgeführt werden. Um dies einzusehen, ist es hilfreich sich vorzustellen, dass die drei Behandlungen völlig identisch sind, d.h., auf allen Parzellen wird dieselbe Behandlung angewendet. Dann werden, wie bereits beschrieben, jeweils die Bodenproben von fünf Parzellen zu einer Mischprobe zusammengefasst. Insgesamt entstehen so drei Mischproben.

In diesen Mischproben die Heterogenität der Parzellen jetzt gewissermaßen eingefroren. Es gibt sehr wahrscheinlich Unterschiede im Gehalt an Nährstoffen, weil die Parzellen, aus denen die Mischproben gebildet worden sind, heterogen sind. Wenn nun die Analytik im Labor sehr präzise ist, wird man diese Unterschiede mit großer Wahrscheinlichkeit nachweisen können. Die Unterschiede haben in diesem Fall aber gar nichts mit Behandlungseffekten zu tun! Es handelt sich lediglich um Umwelteffekte, also Effekte durch Unterschiede der verschiedenen Parzellen.

Falls nun zusätzlich echte Behandlungseffekte vorliegen, so hat man das Problem, dass diese bei der oben beschriebenen Bildung von Mischproben mit den Umwelteffekten vermischt sind. Man kann sie statistisch nicht trennen. Daher ist der Nachweis, dass gefundene Unterschiede auf die Behandlungen zurückzuführen sind, nicht mehr möglich. Um eine solche Trennung zu ermöglichen, ist es zwingend erforderlich, dass die Bodenproben der verschiedenen Parzellen im Labor jeweils getrennt untersucht werden. Zur Überlegung der Versuchsansteller, man könne durch Bildung von Mischproben den Versuchsfehler im Feldversuch ausschalten, muss klar gesagt werden: Dies ist ein eklatanter Fehlschluss. Der Versuchsfehler bleibt vollständig in der Mischprobe enthalten. Durch die Mischung wird es aber unmöglich, bei der Auswertung den Fehler von Behandlungseffekten zu trennen.

8.2 Randomisation

Unter Randomisation versteht man die zufallsmäßige Zuordnung der Behandlungen zu den Randomisationseinheiten. Die Randomisation ist die Grundlage dafür, dass die Fehler e im linearen Modell als stochastisch unabhängig betrachtet werden können.

Beispiel (vollständig randomisierte Anlage): In Kap. 4 hatten wir einen Sortenversuch mit $t = 5$ Sorten (A, B, C, D, E) und $r = 4$ Wiederholungen varianzanalytisch ausgewertet. Dieser Versuch wurde in einer vollständig randomisierten Anlage mit 20 Parzellen (Versuchseinheiten, Feldstücke) durchgeführt. Im folgenden Plan ist für jede Parzelle eine laufende Nummer (oben links) sowie die Bezeichnung der jeweiligen Sorte eingetragen.

1 B	2 D	3 D	4 E
5 C	6 E	7 A	8 B
9 A	10 C	11 E	12 A
13 D	14 D	15 A	16 B
17 B	18 C	19 E	20 C

Die Sorten (Behandlungen) sind hier zufällig den Randomisationseinheiten (Parzellen, Teilstücken) zugeordnet. Diese Randomisation kann man auf verschiedene Weise erhalten. In einfachen Fällen kann man Lose verwenden. Gängiger ist es, zwischen 0 und 1 gleichverteilte Zufallszahlen zu verwenden, die mit

einem Computer erzeugt werden können. Dies soll am Beispiel des Sortenversuches erläutert werden.

Beispiel: $t = 5$ Sorten (A, B, C, D, E) und $r = 4$ Wiederholungen

Schritt 1: Zeichne einen Lageplan für den Versuch. Nummeriere die Randomisationseinheiten systematisch durch.

Beispiel:

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20

Schritt 2: Schreibe r mal die Bezeichnung der Behandlung 1, r mal die Bezeichnung der Behandlung 2, etc. in eine Spalte. Beschaffe eine Folge von $n = rt$ Zufallszahlen (PC, Tabellen) und schreibe diese in eine zweite Spalte. Die Zufallszahlen sollten genügend Stellen haben, um Bindungen zu vermeiden.

Beispiel:

Sorte Zufallszahl

a	0.35501
a	0.73998
a	0.18893
a	0.87083
b	0.91218
b	0.31559
b	0.87631
b	0.03742
c	0.99517
c	0.91288
c	0.16755
c	0.46718
d	0.06820
d	0.83801
d	0.82835
d	0.04390
e	0.18850
e	0.73515
e	0.13536

e 0.97068

Schritt 3: Ordne die Einträge der Liste nach der Größe der Zufallszahl und füge danach eine Spalte mit den geordneten Zahlen 1 bis rt für die Randomisationseinheiten hinzu. Diese Liste ordnet jeder Randomisationseinheit zufällig eine Behandlung zu.

Beispiel:

Sorte Zufallszahl Parzelle

b	0.03742	1
d	0.04390	2
d	0.06820	3
e	0.13536	4
c	0.16755	5
e	0.18850	6
a	0.18893	7
b	0.31559	8
a	0.35501	9
c	0.46718	10
e	0.73515	11
a	0.73998	12
d	0.82835	13
d	0.83801	14
a	0.87083	15
b	0.87631	16
b	0.91218	17
c	0.91288	18
e	0.97068	19
c	0.99517	20

Randomisationseinheit 1 erhält Behandlung B, Randomisationseinheit 2 Behandlung D, etc. □

Schritt 4: Trage die randomisierten Behandlungen in den Lageplan ein.

Beispiel:

1 B	2 D	3 D	4 E
5 C	6 E	7 A	8 B
9 A	10 C	11 E	12 A
13 D	14 D	15 A	16 B
17 B	18 C	19 E	20 C

□

Die obige Versuchsanlage wird als **vollständig randomisierte Anlage** bezeichnet.

8.3 Blockbildung

Anstatt die Behandlungen völlig zufällig auf die Randomisationseinheiten zu randomisieren, wie in 8.2 (vollständig randomisierte Anlage), ist es oft besser, eine **Blockbildung** vorzunehmen. Hierbei werden mehrere Randomisationseinheiten so zu einem Block zusammengefasst, dass jede Behandlung einmal in jedem Block vorkommt. Die Randomisation findet dann getrennt für jeden Block statt. Die resultierende Anlage wird als **Blockanlage** bezeichnet.

Beispiel: Blockanlage mit 6 Behandlungen je Block. Ausrichtung der Blöcke orthogonal zu einem Gradienten.

Block 1	5	4	6	2	1	3
Block 2	4	5	6	3	2	1
Block 3	2	6	4	3	1	5
Block 4	6	1	4	2	3	5

G
R
A
D
I
E
N
T

Die Blockbildung ist vor allem dann vorteilhaft, wenn ein Gradient in den Umweltbedingungen zwischen den Randomisationseinheiten zu erwarten ist. Die Blöcke sollten dann so angeordnet werden, dass innerhalb der Blöcke möglichst homogene Bedingungen herrschen. Dies ermöglicht einen relativ genauen (genauer als bei vollständiger Randomisation) Vergleich der Behandlungen **innerhalb** der Blöcke. Der Versuchsfehler der Varianzanalyse (MQ_{Fehler}) ist bei einer Blockanlage in der Regel kleiner als bei einer vollständig randomisierten Anlage, sofern Blockunterschiede bestehen. **Bei der überwiegenden Mehrzahl von Versuchen mit Tieren und Pflanzen wird das Prinzip der Blockbildung angewendet, weil es einen Gewinn an Genauigkeit ermöglicht.** Wenn ein Versuch geplant wird, sollte man prüfen, ob eine Blockbildung möglich und sinnvoll ist.

Das Prinzip der Blockbildung hängt eng mit dem Prinzip der gepaarte Beobachtung zusammen (Abschnitte 3.8, 3.10 und 3.12). Gepaarte Beobachtungen sind als Spezialfall einer Blockanlage für zwei Behandlungen anzusehen.

8.4 Randomisierte vollständige Blockanlage

Beispiel: In einem On-farm Versuch sollen fünf Sorten in sechs Wiederholungen geprüft werden. Die Versuchsfläche ist sehr heterogen, so dass ein hoher Versuchsfehler zu erwarten ist. Ein Gespräch mit dem Landwirt ergibt, dass sich die Betriebsfläche in drei relativ homogene Teilflächen aufgliedert, die als "sandig", "lehmig" und "feucht" charakterisiert werden können. Da die Betriebsfläche relativ

klein ist, müssen Parzellen in allen drei Teilflächen angelegt werden. Um den Versuchsfehler zu reduzieren, wird der Teilflächentyp als Blockungsfaktor verwendet. In jeder der drei Teilflächen werden zwei vollständige Blöcke angelegt (jeder Block hat fünf Parzellen, je eine pro Sorte).

Da benachbarte Parzellen meistens ähnlicher sind als weiter entfernte, umfasst ein Block meistens eine zusammenhängende Fläche. Es besteht allerdings kein Zwang, zusammenhängende Blöcke anzulegen. Es kann sogar sein, dass das Ziel der Blockbildung, nämlich möglichst homogene Versuchsbedingungen innerhalb der Blöcke zu erzielen, besser erreicht wird, wenn ein oder mehrere Blöcke aus verschiedenen, räumlich getrennten Flächen bestehen.

Beispiel: In einem Tierhaltungsversuch mit Rindern sollen drei verschiedene Fütterungsvarianten geprüft werden. Der Versuch soll in fünf verschiedenen Ställen durchgeführt werden. Da die Bedingungen (Stallklima, etc.) zwischen den Ställen schwanken, wird der Stall als Blockungsfaktor verwendet. Je Stall werden drei Gruppen von Rindern aufgestellt, je eine Gruppe pro Fütterungsvariante. Auf diese Weise werden Stallunterschiede vom Versuchsfehler getrennt.

Blockungsfaktoren müssen nicht an physische Versuchseinheiten gebunden sein. So kann beispielsweise der Faktor Zeit zur Blockbildung herangezogen werden.

Beispiel: Es wird ein Laborexperiment zum Wachstum von Bakterien auf zwanzig verschiedenen Nährmedien geplant. Die Nährmedien sollen in Petrischalen ausgegossen und mit Bakterien beimpft werden. Im Labor steht nur ein Klimaschrank zur Verfügung, in dem die Petrischalen aufgestellt werden sollen, und der Schrank fasst nur maximal 20 Petrischalen. Um nun eine Mindestzahl von Wiederholungen je Medium prüfen zu können, muss die Prüfung über mehrere Termine verteilt werden. Es bietet sich an, die Termine als Blockungsfaktor zu verwenden. Hierzu werden je Termin 20 Petrischalen vorbereitet, wobei für jedes Medium eine Petrischale verwendet wird. Durch diese Blockbildung werden Schwankungen der Wachstumsbedingungen zu den verschiedenen Terminen ausgeschaltet.

Beispiel (Prof. Wünsche, FG Obstbau, Uni Hohenheim): Bei Versuchen mit Obstbäumen erwartet man bei den meisten Ertragsmerkmalen einen Zusammenhang mit der Größe des Baumes. Die Größe des Baumes ist hoch korreliert mit dem Stammumfang, der leicht zu messen ist. Bei der Durchführung eines Blockversuches werden die Bäume so zu Blöcken gruppiert, dass die Bäume innerhalb eines Blocks relativ ähnliche Stammumfänge aufweisen. Dies bedingt zwangsläufig, dass die Bäume eines Blocks räumlich getrennt sind. Dies ist jedoch völlig legitim, solange das Ziel der Blockbildung erreicht wird, dass nämlich die Homogenität innerhalb der Blöcke maximiert wird.

8.4.1 Randomisation

Bei der randomisierten vollständigen Blockanlage hat jeder Block t Randomisationseinheiten, auf welche die Behandlungen vollständig randomisiert werden. Wichtig ist, dass die Randomisation für jeden Block getrennt durchgeführt wird, also nicht dieselbe Randomisation für jeden Block verwendet wird. Die Schritte

der Randomisation für eine Blockanlage mit r Blöcken unter Zuhilfenahme von Zufallszahlen sind wie folgt:

Schritt 1: Schreibe eine Tabelle mit rt Zeilen. Die erste Spalte für Blöcke wird wie folgt mit Zahlen gefüllt: t mal die Zahl 1, t mal die Zahl 2, und t mal die Zahl r . Eine zweite Spalte für Behandlungen wird wie folgt mit Zahlen gefüllt: Die Folge 1, 2, ..., t wird r mal untereinander geschrieben. Sodann beschaffe man eine Folge von rt Zufallszahlen und schreibe diese in eine dritte Spalte.

Beispiel: Ein Versuch mit $t = 6$ Behandlungen soll in $r = 4$ vollständigen Blöcken angelegt werden. Der systematischen Anordnung der Kodierung für Behandlungen und Blöcke fügen wir Zufallszahlen hinzu (RAND).

BLOCK	BEH	RAND
1	1	0.74545
1	2	0.73011
1	3	0.88049
1	4	0.60624
1	5	0.43252
1	6	0.67680
2	1	0.94808
2	2	0.92623
2	3	0.56710
2	4	0.33980
2	5	0.50406
2	6	0.55069
3	1	0.71713
3	2	0.34703
3	3	0.53382
3	4	0.36714
3	5	0.87127
3	6	0.35735
4	1	0.22315
4	2	0.50857
4	3	0.74479
4	4	0.24658
4	5	0.83590
4	6	0.09860

Schritt 2: Sortiere die Liste getrennt für jeden Block nach den Zufallszahlen (RAND).

Beispiel: Für den ersten Block finden wir nach Sortieren:

BLOCK	BEH	RAND
1	5	0.43252
1	4	0.60624
1	6	0.67680

1	2	0.73011
1	1	0.74545
1	3	0.88049

Für alle Blöcke zusammen sieht die sortierte Liste wie folgt aus:

BLOCK	BEH	RAND
1	5	0.43252
1	4	0.60624
1	6	0.67680
1	2	0.73011
1	1	0.74545
1	3	0.88049
2	4	0.33980
2	5	0.50406
2	6	0.55069
2	3	0.56710
2	2	0.92623
2	1	0.94808
3	2	0.34703
3	6	0.35735
3	4	0.36714
3	3	0.53382
3	1	0.71713
3	5	0.87127
4	6	0.09860
4	1	0.22315
4	4	0.24658
4	2	0.50857
4	3	0.74479
4	5	0.83590

Schritt 3: Erstelle einen Lageplan. Zeichne das Ergebnis der Randomisation in den Lageplan ein. Berücksichtige dabei Gradienten (Bodenfruchtbarkeit, Staunässe, etc).

Beispiel:

Block 1: 5 4 6 2 1 3

Block 2: 4 5 6 3 2 1

Block 3: 2 6 4 3 1 5

Block 4: 6 1 4 2 3 5

Block 1	5	4	6	2	1	3
Block 2	4	5	6	3	2	1
Block 3	2	6	4	3	1	5
Block 4	6	1	4	2	3	5

G
R
A
D
I
E
N
T

↑

Die Blöcke werden orthogonal zum Gradienten (falls vorhanden) ausgerichtet. Hierdurch erreichen wir maximale Homogenität der Parzellen innerhalb eines Blocks. Falls der Gradient z.B. Staunässe ist, so sind im Idealfall alle Parzellen in Block 1 gleich schwach und in Block 4 gleich stark von dem Problem betroffen. Vergleiche der Behandlungen innerhalb eines Blocks weisen maximale Genauigkeit auf.

SAS Anweisungen zur Erzeugung einer Blockanlage mit 4 Blöcken und 6 Behandlungen:

```
proc plan;
factor block=4 ordered behandlung=6;
run;
```

8.5 Lateinisches Quadrat

Die randomisierte vollständige Blockanlage erlaubt die Ausschaltung einer einzigen Störgröße. In manchen Anwendungen gibt es mehr als eine Störgröße, deren Ausschaltung durch Blockbildung sinnvoll ist. Anlagen, in denen zwei Blockungsfaktoren verwendet werden, die orthogonal zueinander (d.h. voneinander unabhängig) sind, heißen **Zeilen-Spalten Pläne** (Englisch: row-column designs). Der einfachste Typ eines Zeilen-Spalten-Planes ist das **Lateinische Quadrat**, auf den wir uns hier beschränken wollen. In Feldversuchen kann es sinnvoll sein, Bodenunterschiede sowohl durch Zeilen als auch durch Spalten auszuschalten. In einem Experiment zur Prüfung des Geschmacks verschiedener Nahrungsmittelprodukte können Zeilen verschiedene Prüfpersonen sein, während Spalten der Reihenfolge entsprechen, in der eine Prüfperson die verschiedenen Produkte prüft. In einem Fütterungsversuch mit Tieren zum Test verschiedener Futtermittel, bei dem jedes Tier alle Futtermittel prüfen soll, können die Zeilen den Tieren und die Spalten den Zeitpunkten der Fütterung entsprechen.

Beispiel (Mudra, 1958): Zur Kontrolle des Bakteriengehaltes von Milch aus fünf verschiedenen Betrieben (A, B, C, D, E) werden Milchproben entnommen. Von jedem Betrieb wurden fünf Proben genommen und getrennt untersucht. Somit standen je Betrieb fünf Wiederholungen zur Verfügung. Die Untersuchung einer Probe dauert längere Zeit. Pro Tag können fünf Proben hintereinander untersucht werden. Die Untersuchungen beginnen etwa um 8:30, 10:00, 11:30, 14:00 und 15:30. Die Untersuchungen werden über fünf Tage durchgeführt.

Da die Bakterienvermehrung stark vom Tag abhängen kann, bietet es sich an, den Tag als Blockungsfaktor zu verwenden. Daher wurde je Betrieb und Tag eine Probe genommen; die fünf Proben je Betrieb verteilen sich also über fünf Tage. Außerdem hängt der Bakteriengehalt wegen der natürlichen Vermehrung stark von der Tageszeit ab. Um diese Fehlerquelle auszuschalten, ist es sinnvoll, die Tageszeit als weiteren Blockungsfaktor zu verwenden. Als Versuchsanlage kommt ein Lateinisches Quadrat in Frage.

Tageszeit	Tag				
	1	2	3	4	5
08:30	A	B	C	D	E
10:00	D	C	E	B	A
11:30	C	A	D	E	B
14:00	B	E	A	C	D
15:30	E	D	B	A	C

Jeder Betrieb wird einmal zu jeder Tageszeit und einmal an jedem Tag geprüft. Außerdem bilden jeder Tag sowie jede Tageszeit jeweils einen vollständigen Block. Wir haben es also mit zwei orthogonalen (und vollständigen) Blockstrukturen zu tun.

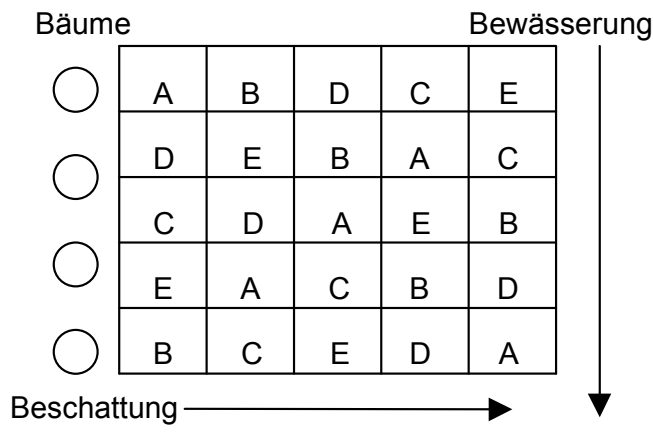
Beispiel (John und Quenouille, 1977): In einem Fütterungsversuch mit Mäusen wurden $t = 7$ Futtermittel (A, B, C, D, E, F, G) getestet. Es standen mehrere Würfe von Mäusen zur Verfügung. Da Mäuse eines Wurfes genetisch relativ ähnlich sind, bot es sich an, Würfe als Blockungsvariable zu verwenden.

Allerdings unterscheiden sich die Mäuse innerhalb eines Wurfes auch im Gewicht. Die Versuchsansteller ordneten die Mäuse eines Wurfes daher nach ihrem Gewicht. Die Variable "Rangfolge im Gewicht" wurde als zweite Blockungsvariable verwendet. Es wurde auch hier ein Lateinisches Quadrat verwendet. Der randomisierte Versuchsplan sah wie folgt aus:

Wurf	Rangfolge im Gewicht						
	1	2	3	4	5	6	7
1	E	A	G	B	F	D	C
2	F	D	A	G	B	C	E
3	A	G	C	F	E	B	D
4	C	E	B	D	A	F	G
5	G	F	E	C	D	A	B
6	D	B	F	E	C	G	A
7	B	C	D	A	G	E	F

Beispiel (Petersen, 1994): Ein Bakteriologe untersuchte den Einfluss vier verschiedener Quellen von Saatgut-Inokulum für Rhizobium-Bakterien (A, B, C, D) sowie einer Kontrollbehandlung (E) auf das Wachstum von Alfalfa. Die Parzellen des Versuchs wurden mit Furchenbewässerung bewässert. Dabei lagen aus versuchstechnischen Gründen immer mehrere Parzellen hintereinander. Es war damit zu rechnen, dass die Bewässerung der oberen Parzellen sich systematisch von derjenigen der unteren unterscheiden würde, so dass ein Gradient "oben-unten" zu erwarten war. Außerdem befand sich an einem der vier Ränder der rechteckigen Versuchsfläche eine Baumreihe, die eine Beschattung orthogonal zur

Bewässerungsrichtung verursachte. Eine Blockung in Zeilen und Spalten wurde verwendet, um beide Fehlerquellen gleichzeitig auszuschalten (Lateinisches Quadrat).



Beispiel (Landtechnik Hohenheim, Dr. Grimm, Prof. Jungbluth): Es soll die Staubentwicklung in drei charakteristischen Stallbereichen (BEREICH) verglichen werden. Der Staub wird in zwei Fraktionen gegliedert (FRAKTION), die verglichen werden sollen. Für die Messung stehen insgesamt sechs Geräte zur Verfügung, die jeweils in einem Stall aufgestellt werden sollen. Um die Staubfraktionen getrennt erfassen zu können, werden zwei verschiedene Filtertypen verwendet. Je Gerät kann nur ein Filter gleichzeitig eingesetzt werden. Die Untersuchung soll über 6 Wochen durchgeführt werden. Nach jeweils einer Woche werden die Staubfilter entnommen und gewechselt, wobei dann die Staubmenge ermittelt wird.

Es wird damit gerechnet, dass es zeitliche Änderungen der Staubmenge gibt. Daher soll die Messwoche (WOCHE) als Blockvariable verwendet werden. Ein weiteres Problem besteht darin, dass es systematische Messfehler der Geräte gibt, wie sich aus Voruntersuchungen ergeben hat. Daher soll ein Gerät an verschiedenen Orten im Stall zu Einsatz kommen, um diese systematischen Messfehler auszuschalten. Es bietet sich an, das Gerät (GERÄT) als zweite Blockvariable zu verwenden.

Die Behandlungsfaktoren des Versuches sind BEREICH mit drei Stufen und Staubfraktion (FRAKTION) mit zwei Stufen. Es ergeben sich insgesamt sechs Behandlungen. Die Konstellation ist insofern einfach, als jede der beiden Blockvariablen ebenfalls sechs Stufen hat. Daher liegt es nahe, den Versuch nach einem Lateinischen Quadrat anzulegen.

Beispiel: In der Tierernährung und Futtermittelkunde werden Versuchstiere häufig durch einen chirurgischen Eingriff mit einer künstlichen Magen- oder Darmöffnung (Fistel) versehen, welche die wiederholte Entnahme von Magen- oder Darminhalt erlauben. Der Eingriff ist relativ aufwändig und aus ethischen Gründen ist es geboten, die Zahl der Tiere, welche einem solchen Eingriff unterzogen werden, zu minimieren. Aus diesen Gründen ist die Zahl von fistulierten Tieren, die für Fütterungsversuche zur Verfügung stehen, in der Regel sehr begrenzt. Dies hat zur Folge, dass an ein und demselben Tier mehrere Fütterungsvarianten geprüft werden müssen. Dies lässt sich in einem lateinischen Quadrat (oder einem ähnlichen Zeilen-Spalten Plan) realisieren, bei dem die Tiere als der eine Blockfaktor und die Periode, in der eine Fütterung zur Anwendung kommt, als der zweite Blockfaktor verwendet wird. Zwischen den Perioden müssen Pausen eingeschaltet werden, in denen alle

Tiere dieselbe Fütterung erhalten, um Nachfolgeeffekte von einer zur anderen Periode auszuschalten (carry-over effects). Besser als ein einfaches lateinisches Quadrat sind sog. Cross-over Designs, die es erlauben, carry-over Effekte zu schätzen und optimal auszuschalten (siehe Abschnitt 8.5.2).

8.5.1 Randomisation

Schritt 1: Wähle einen Grundplan für ein Lateinisches Quadrat (z.B. Cochran und Cox, 1957).

Schritt 2: Randomisiere die Zeilen.

Schritt 3: Randomisiere die Spalten.

Schritt 4: Ordne die Behandlungsbezeichnungen zufällig den Symbolen im randomisierten Grundplan zu.

Beispiel: Es soll ein Versuch mit $t = 4$ Behandlungen N1, N2, N3 und N4 als Lateinisches Quadrat angelegt werden.

Schritt 1: Wir wählen folgenden Plan aus Cochran und Cox (1957):

```
a b c d
b a d c
c d a b
d c b a
```

Schritt 2: Wir schreiben die Zeilennummern 1 bis 4 untereinander in eine Spalte, erzeugen eine zweite Spalte mit einer Folge von Zufallszahlen.

ZEILE	RAND1
1	0.58876
2	0.70511
3	0.12879
4	0.99552

Sortieren nach der Zufallszahl liefert eine Randomisation der Zeilen:

ZEILE	RAND1
3	0.12879
1	0.58876
2	0.70511
4	0.99552

Schreibe Zeile 3 des Grundplans in Zeile 1, Zeile 1 in Zeile 2 etc.

1	a	b	c	d		3	c	d	a	b
2	b	a	d	c	⇒	1	a	b	c	d
3	c	d	a	b		2	b	a	d	c
4	d	c	b	a		4	d	c	b	a

Schritt 3: Wie in Schritt 2 erzeugen wir eine zufällige Folge der Spaltennummern 1 bis 4:

4, 2, 1, 3

Schreibe Spalte 4 des bezüglich der Zeilen randomisierten Grundplans (Schritt 2) in Spalte 1, Spalte 2 in Spalte 2, etc.

	1	2	3	4		4	2	1	3	
3	c	d	a	b		3	b	d	c	a
1	a	b	c	d	⇒	1	d	b	a	c
2	b	a	d	c		2	c	a	b	d
4	d	c	b	a		4	a	c	d	b

Schritt 4: Wir erzeugen eine zufällige Abfolge der Symbole a, b, c, d und schreiben diese in eine zweite Spalte.

SYMBOL	RAND1		SYMBOL	RAND1
a	0.80661		b	0.01157
b	0.01157	Sortieren	d	0.25140
c	0.33509	⇒	c	0.33509
d	0.25140		a	0.80661

Wir schreiben die Behandlungsbezeichnungen sowie die Zufallsfolge der Symbole in zwei Spalten nebeneinander. Hieraus ergibt sich eine zufällige Zuordnung der Symbole zu den Behandlungsbezeichnungen.

N1	b
N2	d
N3	c
N4	a

Der randomisierte Plan ist wie folgt:

	4	2	1	3		4	2	1	3	
3	b	d	c	a		3	N1	N2	N3	N4
1	d	b	a	c	⇒	1	N2	N1	N4	N3
2	c	a	b	d		2	N3	N4	N1	N2
4	a	c	d	b		4	N4	N3	N2	N1

□

Bemerkung: Ganz streng genommen könnte Schritt 4 der Randomisation bei Lateinischen Quadraten entfallen. Bei anderen Zeilen-Spalten-Plänen kann dieser

Schritt dagegen keinesfalls entfallen, weshalb er hier anhand der einfachen Falls eines Lateinischen Quadrates mit aufgenommen wurde. Standardsoftware für Versuchsdesigns wie z.B. CycDesign, randomisieren alle Zeilen-Spalten-Pläne, auch Lateinische Quadrate, so wie es hier beschrieben wurde.

8.5.2 Cross-Over Designs

Versuche mit Tieren sind häufig dadurch charakterisiert, dass die Zahl der Versuchstiere besonders eingeschränkt ist. Dies gilt z.B. für fistulierte Tiere, bei denen eine permanente Kanüle am Darm oder Pansen eingerichtet worden ist, um wiederholt Darm- bzw. Mageninhalt entnehmen zu können. Solche Tiere sind sehr aufwändig in der Haltung, so dass in der Regel nur wenige Tiere für einen Versuch zur Verfügung stehen. In diesen Fällen wird jedes Tier typischer Weise zur Prüfung mehrerer Behandlungen eingesetzt. Dies führt zu verbundenen Stichproben und somit in der Regel zu einer Erhöhung der Genauigkeit. Die Besonderheit der hier geschilderten Art von Versuchen ist, dass es wichtig ist, in welcher Reihenfolge die Behandlungen bei einem Tier geprüft werden. Es stellt sich die Frage, in welcher Reihenfolge die Behandlungen angewendet werden sollen. Designs für diese Art von Versuch heißen **Cross-over Designs**. Der einfachste Fall eines sog. Cross-over Designs liegt dann vor, wenn nur zwei Behandlungen A und B zu prüfen sind. Würde man bei jedem Tier zu erst A und dann B prüfen, könnte man nicht ausschließen, dass eventuelle signifikante Unterschiede zwischen A und B in Wirklichkeit Nachwirkungen der jeweils vorhergehenden Behandlung oder Carry-over Effekte sind. Daher ist es besser, bei der Hälfte der Tiere die Reihenfolge umzudrehen (**Cross-over**), also erst B und dann A zu prüfen. Hierher rührt die Bezeichnung Cross-over Design.

Da im allgemeinen mit Unterschieden zwischen den Zeitpunkten oder Perioden zu rechnen ist, in denen die Behandlungen angewendet werden, muss der Zeitpunktes als Blockungsfaktor berücksichtigt werden. Gleichzeitig muss offensichtlich das Tier selbst als Blockungsfaktor betrachtet werden. Sofern jedes Tier jede Behandlung prüft, kann man hier ein Lateinisches Quadrat mit Blockungsfaktoren Tier und Zeit einsetzen. Dabei ist zu gewährleisten, dass zwischen der Anwendung verschiedener Behandlungen (Futter, Medikamente) genügend Zeit zwischen den Behandlungen gelassen wird, um **Nachfolgeeffekte (Carry-over Effekte)** gering zu halten. Wenn die Zahl der Behandlungen so groß ist, dass ein Tier nur jeweils einen Teil der Behandlungen prüfen kann, so müssen wiederum unvollständige Blöcke für den Blockfaktor Tier verwendet werden (siehe Abschnitt 8.6).

Oft können Carry-over Effekte nicht ganz ausgeschlossen werden. In diesen Fällen ist wichtig, welche Behandlung unmittelbar vor und nach einer interessierenden Behandlung eingesetzt wurde, also in welcher Reihenfolge die Behandlungen angewendet wurden. Es ist dann wünschenswert, das Design so zu gestalten, dass möglichst jede Behandlung in gleicher Anzahl der interessierenden Behandlung voranging oder dieser folgte. Man kann dann Carry-over Effekte in gewissem Umfang modellseitig berücksichtigen und bei der Auswertung durch entsprechende Adjustierungen herausrechnen. Es gibt hierzu verschiedene Modellansätze und dementsprechend auch verschiedene jeweils optimale Cross-over Designs. Näheres findet sich bei Johnes B, Kenward MG 2003 Design and analysis of cross-over trials.

8.6 Anlagen mit unvollständigen oder ungleich großen Blöcken

Beim Lateinischen Quadrat sowie bei der randomisierten vollständigen Blockanlage müssen vollständige Blöcke gebildet werden, d.h. jede Behandlung kommt genau einmal in jedem Block vor. Oft ist es nicht möglich oder nicht sinnvoll, vollständige Blöcke zu bilden. In diesen Fällen können Versuchsanlagen mit unvollständigen Blöcken verwendet werden. Es kann außerdem dazu kommen, dass die Blockgröße variiert.

Beispiel (Mead, 1978): Sieben Medikamente, die das Wachstum von Ferkeln beeinflussen, sollen an 35 Ferkeln geprüft werden. Die Ferkel stammen aus Würfen à 5, 6, 7, 7 und 10 Tieren. Der Wurf soll als Blockungsfaktor verwendet werden.

In den Blöcken à 7 Ferkel (III und IV) prüfen wir jedes Medikament einmal. In den Blöcken à 5 und à 6 Ferkel (I und II) müssen wir ein bzw. zwei Medikamente weglassen (unvollständige Blöcke), im Block à 10 Ferkel (V) müssen drei Medikamente hinzugefügt werden. Damit jedes Medikament gleich oft geprüft wird, prüfen wir die Medikamente 5, 6 und 7 in Block V zweimal und lassen jedes dieser Medikamente in einem der Blöcke I und II einmal weg. Der Standardfehler einer Differenz variiert leicht zwischen den Behandlungspaaren, u.a. weil die Blockgrößen variieren (siehe Abschnitt 8.7), aber das Design ist effizient.

	Block				
	I	II	III	IV	V
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	4	4	4	4	4
5		6	5	5	5
		7	6	6	6
			7	7	7
					5
					6
					7

Dieser Grundplan muss natürlich noch randomisiert werden, d.h. die Behandlungen müssen zufällig den Tieren eines Blocks zugeordnet werden.

Beispiel: In einem On-farm Versuch sollen sechs verschiedene Reissorten geprüft werden. Fünf Betriebe beteiligen sich an dem Versuch. Der Betrieb soll als Blockungsfaktor verwendet werden. Die Randomisationseinheit soll jeweils ein ganzes Feld des Landwirtes sein. Die Betriebe sind unterschiedlich groß, so dass nicht jeder Bauer sechs Felder zur Verfügung stellen kann. Daher scheidet eine randomisierte vollständige Blockanlage aus. Die fünf Betriebe stellen 3, 4, 5, 5 und 6 Felder zur Verfügung. Wir gehen zunächst von fünf vollständigen Blöcken aus und lassen dann Behandlungen weg. Insgesamt benötigen wir $3+2+1+1=7$

Auslassungen. Fünf Sorten müssen einmal weggelassen werden und eine zweimal. Wir streichen systematisch beginnend beim Block I fortlaufend die Sorten 1 bis 6 und streichen dann in Block IV die Sorte 1 ein zweites mal.

Block				
I	II	III	IV	V
	1	1		1
	2	2	2	2
	3	3	3	3
4		4	4	4
5		5	5	5
6	6		6	6

Auch hier muss natürlich noch randomisiert werden.

Beispiel: Ein Versuch mit 151 Sorten soll in einer Blockanlage durchgeführt werden. Bei vollständigen Blöcken hätte ein Block 151 Versuchsglieder und würde somit sehr groß. Wegen der Blockgröße ist mit größerer Heterogenität innerhalb eines Blocks zu rechnen, was dem Ziel der Blockbildung entgegenläuft. Es ist sinnvoll, kleinere Blöcke zu bilden, um eine größere Homogenität zu erzielen. Kleinere Blöcke müssen aber notwendigerweise unvollständig sein, es kann also nicht jede Sorte in jedem Block stehen. Es gibt eine Reihe von Anlagen mit unvollständigen Blöcken, z.B. sog. **Zweisatz- und Dreisatzgitter** sowie sog. **α -Designs** (siehe Kempton, R.A., Fox, P.N. 1997. Statistical methods for plant variety evaluation. Chapman & Hall, London).

Beispiel: (Landtechnik Hohenheim, Dr. Grimm, Prof. Jungbluth): Es soll die Staubentwicklung in drei charakteristischen Stallbereichen (BEREICH) verglichen werden. Der Staub wird in zwei Fraktionen gegliedert (FRAKTION), die verglichen werden sollen. Für die Messung stehen insgesamt sechs Geräte zur Verfügung, die jeweils in einem Stall aufgestellt werden sollen. Um die Staubfraktionen getrennt erfassen zu können, werden zwei verschiedene Filtertypen verwendet. Je Gerät kann nur ein Filter gleichzeitig eingesetzt werden. Die Untersuchung soll über 4 Wochen durchgeführt werden. Nach jeweils einer Woche werden die Staubfilter entnommen und gewechselt, wobei dann die Staubmenge ermittelt wird.

Es wird damit gerechnet, dass es zeitliche Änderungen der Staubmenge gibt. Daher soll die Messwoche (WOCHE) als Blockvariable verwendet werden. Ein weiteres Problem besteht darin, dass es systematische Messfehler der Geräte gibt, wie sich aus Voruntersuchungen ergeben hat. Daher soll ein Gerät an verschiedenen Orten im Stall zu Einsatz kommen, um diese systematischen Messfehler auszuschalten. Es bietet sich an, das Gerät (GERAET) als zweite Blockvariable zu verwenden.

Die Behandlungsfaktoren des Versuches sind BEREICH mit drei Stufen und Staubfraktion (FRAKTION) mit zwei Stufen. Es ergeben sich insgesamt sechs Behandlungen. Da nur vier Messwochen zur Verfügung stehen und ein Gerät jeweils nur eine der sechs Behandlungen prüfen kann, wird im gesamten Versuch ein Gerät bei nur jeweils vier der sechs Behandlungen eingesetzt.

Es liegen zwei Blockfaktoren vor: WOCHE und GERAET. Die Blöcke, die durch die Geräte definiert sind, sind zwangsläufig unvollständig (vier von sechs

Behandlungen). Bei der Planung ist darauf zu achten, dass die sechs Behandlungen so "gerecht" wie möglich über die Wochen und Geräte verteilt werden. Der folgende **Zeilen-Spalten-Plan** wurde mit der Software GENDEX erhalten:

Geräte						Woche
A	B	C	D	E	F	
5	4	3	6	2	1	a
1	5	4	3	6	2	b
2	1	6	4	5	3	c
6	3	2	1	4	5	d

Die sechs Behandlungen (2 Fraktionen x 3 Bereiche) sind mit 1 bis 6 durchnummeriert. Das Design ist gerecht in dem Sinne, dass jede Behandlung genau einmal in jeder Woche und höchstens einmal mit jedem Gerät geprüft wird.

8.7 Auswertung einer Blockanlage

8.7.1 Varianzanalyse einer Blockanlage

Zur varianzanalytischen Auswertung einer Blockanlage kann ein lineares Modell zugrundegelegt werden, welches neben einem Behandlungs- auch einen Blockeffekt hat:

$$y_{ij} = \mu + b_j + \tau_i + e_{ij}, \quad i = 1, \dots, t; \quad j = 1, \dots, r,$$

wobei

μ = Konstante

τ_i = Effekt der i -ten Behandlung

b_j = Effekt des j -ten Blocks

Die Auswertung erfolgt völlig analog der Auswertung einer vollständig randomisierten Anlage (Kap. 4), mit dem Unterschied, dass in der Varianzanalyse-Tabelle als Streuungsursache die Blockeffekte hinzukommen. Da es sich um ein lineares Modell handelt, sind die allgemeinen Resultate in Abschnitt 6.8 und 6.9 anwendbar.

Die Streuungszerlegung beruht auf der folgenden Modell-Sequenz:

Modell	SQ_{Fehler}	FG_{Fehler}
(0) $y_{ij} = \mu + e_{ij}$	$SQ_{Fehler}^{(0)}$	$n-1$
(1) $y_{ij} = \mu + b_j + e_{ij}$	$SQ_{Fehler}^{(1)}$	$n-1-(r-1) = n-r$
(2) $y_{ij} = \mu + b_j + \tau_i + e_{ij}$	$SQ_{Fehler}^{(2)}$	$n-r-(t-1) = n-r-t+1$

Wir schreiben die Summen der Quadrate in eine Varianzanalyse-Tabelle:

Ursache	FG	SQ	MQ
Blöcke vor Beh.	$r-1$	$SQ(b_j \mu) = SQ_{Fehler}^{(0)} - SQ_{Fehler}^{(1)}$	$SQ(b_j \mu)/(r-1)$
Beh. nach Blöcken	$t-1$	$SQ(\tau_i b_j, \mu) = SQ_{Fehler}^{(1)} - SQ_{Fehler}^{(2)}$	$SQ(\tau_i b_j, \mu)/(t-1)$
Fehler	$n-t-r+1$	$SQ_{Fehler}^{(2)}$	$SQ_{Fehler}^{(2)}/(n-t-r+1)$

Hierbei ist r die Zahl der Blöcke, t die Zahl der Behandlungen und n die Gesamtzahl der Beobachtungen. Im Fall balancierter Daten ist die Zahl der Beobachtungen gleich $n = rt$. Es sei aber betont, dass die obige Tabelle auch im Fall unbalancierter Daten gültig ist ($n < rt$). Für unbalancierte Daten ist es wichtig, dass die Blockeffekte vor den Behandlungseffekten angepasst werden. Nur so erhält man eine um die Blockeffekte bereinigte Beurteilung von Behandlungseffekten. Andernfalls ist der resultierende F-Test durch Blockeffekte verzerrt. Im Fall balancierter Daten spielt die Reihenfolge der Anpassung dagegen keine Rolle, weil Blöcke und Behandlungen dann orthogonal, also unabhängig sind. Passen wir die Blockeffekte nach den Behandlungseffekten an, erhalten wir dieselbe Varianzanalyse wie im umgekehrten Fall. Bei unbalancierten Daten geht diese Übereinstimmung verloren.

Die Freiheitsgrade für eine Streuungsursache entsprechen wie üblich der entsprechenden Reduktion der Fehler-FG in der Modellsequenz. Modell (0) hat z.B. $n-1$ FG, weil ein Parameter (μ) im linearen Modell steht. Modell (1) hat zwar r Parameter mehr, davon sind aber nur $r-1$ freie Parameter, weil hier ein überparametrisiertes Modell vorliegt (siehe Abschnitt 6.8.1, Beispiel Varianzanalyse). Somit hat Modell (1) $n-1-(r-1) = n-r$ Fehler-FG. Dies kann man auch so erklären, dass Modell (1) für jeden Block einen eigenen Mittelwert anpasst, und davon gibt es r Stück. Die Mittelwerte sind hier die eigentlichen frei wählbaren Parameter. Also hat man $n-r$ Fehler-FG. Die Differenz der Fehler-FG von Modell (0) und Modell (1) beträgt $r-1$, und dies sind die FG für den Faktor Block in der Varianzanalyse-Tabelle. Die FG für die Behandlungen ergeben sich analog.

Die Entscheidungsregel für den Test auf Behandlungseffekte lautet:

Verwerfe $H_0: \tau_1 = \tau_2 = \dots = \tau_t$ falls

$$F_{Vers} = \frac{SQ(\tau_i | b_j, \mu)/(t-1)}{SQ_{Fehler}^{(2)}/(n-r-t+1)} = \frac{MQ_{Beh}}{MQ_{Fehler}} > F_{Tab}(1-\alpha, t-1, n-r-t+1) \quad (\text{Tab. VI})$$

Für **balancierte Daten** haben die SQ folgende einfache Form:

$$SQ_{Blöcke} = SQ(b_j | \mu) = t \sum_{j=1}^r (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 = \sum_{j=1}^r y_{\cdot j}^2/t - y_{\cdot\cdot}^2/(rt)$$

$$SQ_{Beh} = SQ(\tau_i | b_j, \mu) = r \sum_{i=1}^t (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^t y_{i\cdot}^2/r - y_{\cdot\cdot}^2/(rt)$$

$$SQ_{Fehler}^{(2)} = \sum_{i=1}^t \sum_{j=1}^r (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^t \sum_{j=1}^r y_{ij}^2 - \sum_{j=1}^r y_{\cdot j}^2/t - \sum_{i=1}^t y_{i\cdot}^2/r + y_{\cdot\cdot}^2/(rt)$$

Bei den SQ ist jeweils die mittlere Formel anschaulicher, was die Interpretation als Quadratsumme betrifft, während die Formel auf der rechten Seite rechenstechnisch zu bevorzugen ist. Das SQ_{Beh} beispielsweise ist die Summe der quadrierten Abweichungen der Behandlungsmittelwerte vom Gesamtmittelwert. Je größer die Behandlungsunterschiede, desto größer wird folglich das SQ_{Beh} . Analoges gilt für das $SQ_{Blöcke}$. Im unbalancierten Fall ist keine so anschauliche Ausdruck der SQ möglich. Aber die Interpretation bleibt dieselbe: **Je größer ein SQ (bzw. MQ), desto einflussreicher der betreffende Effekt.**

Beispiel (balanciert): Ein Blockversuch, bei dem der Ertrag (kg/ha) der Reissorte IR8 bei 6 verschiedenen Aussaatstärken gemessen wurde, lieferte folgendes Ergebnis (Gomez und Gomez, 1984):

Saatstärke (kg/ha)	Block			
	1	2	3	4
25	5113	5398	5307	4678
50	5346	5952	4719	4264
75	5272	5713	5483	4749
100	5164	4831	4986	4410
125	4804	4848	4432	4748
150	5254	4542	4919	4098

Die Einflussvariable Aussaatstärke ist quantitativ, so dass eine Regression für die Auswertung sinnvoll ist. Wir wollen hier sie zunächst wie einen qualitativen Faktor auswerten. Berechnung der Randsummen:

Saatstärke (kg/ha)	Block				$y_{i.}$
	1	2	3	4	
25	5113	5398	5307	4678	20496
50	5346	5952	4719	4264	20281
75	5272	5713	5483	4749	21217
100	5164	4831	4986	4410	19391
125	4804	4848	4432	4748	18832
150	5254	4542	4919	4098	18813

$$y_{.j} \quad 30953 \quad 31284 \quad 29846 \quad 26947 \quad 119030 = y_{..}$$

$$I = \sum_{j=1}^r y_{.j}^2 / t = (30953^2 + 31284^2 + 29846^2 + 26947^2) / 6 = 592283565$$

$$II = \sum_{i=1}^t y_{i.}^2 / r = (20496^2 + 20281^2 + 21217^2 + 19391^2 + 18832^2 + 18813^2) / 4 \\ = 591537535$$

$$III = y_{..}^2 / (rt) = 119030^2 / 24 = 590339204,17$$

$$IV = \sum_{i=1}^t \sum_{j=1}^r y_{ij}^2 = 5113^2 + \dots + 4098^2 = 595140272$$

$$SQ_{Bl\ddot{o}cke} = I - III = 592283565 - 590339204,17 = 1944360,83$$

$$SQ_{Beh} = II - III = 591537535 - 590339204,17 = 1198330,83$$

$$SQ_{Fehler} = IV - I - II + III = 595140272 - 592283565 - 591537535 + 590339204,17 = 165376,17$$

Zusammenfassung in Varianzanalyse-Tabelle:

Ursache	FG	SQ	MQ	F
Blöcke	3	1944361	648120	
Behandlung	5	1198331	239666	2,17
Fehler	15	1658376	110558	

Da die Daten orthogonal sind, haben wir in die Varianzanalyse-Tabelle als Ursachenbezeichnungen "Blöcke" statt "Blöcke vor Behandlungen" sowie "Behandlungen" statt "Behandlungen nach Blöcken" geschrieben. Der F-Test auf Behandlungsunterschiede ist nicht signifikant [$F_{Vers} = 2,17 < F_{Tab}(t-1 = 5, n-r-t+1 = 15; \alpha = 5\%) = 2,9$].

Beispiel (unbalanciert): Im vorangegangenen Beispiel seien zwei Beobachtungen ausgefallen:

Saatstärke (kg/ha)	Block			
	1	2	3	4
25	.	5398	5307	4678
50	5346	5952	4719	4264
75	5272	5713	5483	4749
100	5164	4831	4986	4410
125	4804	4848	.	4748
150	5254	4542	4919	4098

Modell	SQ_{Fehler}
(0) $y_{ij} = \mu + e_{ij}$	4492823
(1) $y_{ij} = \mu + b_j + e_{ij}$	2501236
(2) $y_{ij} = \mu + b_j + \tau_i + e_{ij}$	1464571

$$SQ(b_j | \mu) = SQ_{Fehler}^{(0)} - SQ_{Fehler}^{(1)} = 4492823 - 2501236 = 1991587$$

$$SQ(\tau_i | b_j, \mu) = SQ_{Fehler}^{(1)} - SQ_{Fehler}^{(2)} = 2501236 - 1464571 = 1036665$$

$$SQ_{Fehler}^{(2)} = 1464571$$

Ursache	FG	SQ	MQ	F_{vers}
Blöcke vor Beh.	3	1991587	663863	1,84
Beh. nach Blöcken	5	1036665	207333	
Fehler	13	1464571	112659	

$$F_{Vers} = 1,84 < F_{Tab}(t-1 = 5, n-r-t+1 = 13; \alpha = 5\%) = 3,03$$

Es bestehen keine signifikanten Behandlungsunterschiede.

8.7.2 Mittelwertvergleiche in einer Blockanlage

Bei **balancierten Daten** können einfache arithmetische Mittelwerte berechnet und mit dem LSD oder HSD (Tukey) Test verglichen werden. LSD und HSD berechnen sich analog zu der vollständig randomisierten Anlage (Abschnitte 4.4 und 4.5). Der einzige Unterschied besteht in der Zahl der Freiheitsgrade für den Fehler. Es gilt:

Vergleichsbezogene Irrtumswahrscheinlichkeit (t-Test; t_{tab} aus Tab. II):

$$\text{Verwerfe } H_0: \tau_s = \tau_u, \text{ wenn } |\bar{y}_{s.} - \bar{y}_{u.}| > t_{Tab}(n-r-t+1, \alpha) \sqrt{2MQ_{Fehler}/r} = LSD$$

Versuchsbezogene Irrtumswahrscheinlichkeit (Tukey; q_{tab} aus Tab. VII):

$$\text{Verwerfe } H_0: \tau_s = \tau_u, \text{ wenn } |\bar{y}_{s.} - \bar{y}_{u.}| > q_{Tab}(t, n-r-t+1, \alpha) \sqrt{MQ_{Fehler}/r} = HSD$$

Beispiel: Für den Aussaatstärkeversuch mit Reis sollen Mittelwertvergleiche bei Einhaltung der versuchsbezogenen Irrtumswahrscheinlichkeit durchgeführt werden ($\alpha = 5\%$). Wir finden $MQ_{Fehler} = 110558$, $r = 4$ und $q_{Tab}(t = 6, n-r-t+1, \alpha = 5\%) = 4,595$. Damit ist $HSD = 4,595 \times \sqrt{(110558/4)} = 763,9$. Die Mittelwerte der Behandlungen sind:

Saatstärke	Mittelwert
25	5124,0 ^a
50	5070,3 ^a
75	5304,4 ^a
100	4847,8 ^a
125	4708,0 ^a
150	4703,3 ^a
<i>HSD</i>	763,9

(Mittelwerte, die mit demselben Buchstaben versehen sind, sind nicht signifikant voneinander verschieden).

Es bestehen keine signifikanten Unterschiede.

Bei **unbalancierten Daten** sind arithmetische Mittelwerte nicht mehr aussagekräftig, wie im folgenden erläutert wird.

Beispiel: Drei Sorten werden in drei Blöcken geprüft:

		Sorte		
		1	2	3
Block	1	10	20	30
	2	20	30	40
	3	60	70	80
Sortenmittel $\bar{y}_{i.}$:		30	40	50

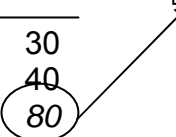
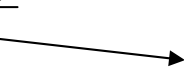
Sorte 3 hat den größten Mittelwert. Die Daten sind vereinfacht dahingehend, dass Abwesenheit von Versuchsfehlern und exakte Gültigkeit des additiven Modells ($\mu + b_j + \tau_i$) angenommen wird. In diesem Fall sind die Differenzen der Sorten in jedem Block gleich. In diesem Fall sind die arithmetischen Mittelwerte $\bar{y}_{i.}$ sinnvoll für die Beurteilung der Sorten.

Nun sei angenommen, dass die Beobachtung der Sorte 3 in Block 3 ausfällt:

		Sorte		
		1	2	3
Block	1	10	20	30
	2	20	30	40
	3	60	70	.
Sortenmittel $\bar{y}_{i.}$:		30	40	35

Wir haben zunächst arithmetische Mittel berechnet. Wegen des fehlenden Wertes hat sich jetzt der Mittelwert für Sorte 3 geändert. Er liegt nun zwischen denen für die Sorten 1 und 2. Falls die Auswertung auf Basis einfacher Mittelwerte erfolgt, ist die Beurteilung der Sorten somit nicht identisch für balancierte und für unbalancierte Daten. Hierbei sei betont, dass Abwesenheit von Versuchsfehlern angenommen wird. Die unterschiedlichen Beurteilungen sind also nicht durch Zufallsschwankungen bedingt. Tatsächlich ist die Beurteilung mittels einfacher Mittelwerte irreführend. Es muss berücksichtigt werden, dass Block 3 der beste Block ist und Sorte 3 nicht die Chance hatte, ihr Leistungspotential in diesem Block zu zeigen.

Die hypothetischen Daten sind so gewählt, dass die Differenz der Sorten 3 und 1 in den Blöcken 1 und 2 jeweils 20 dt/ha beträgt. Dieselbe Differenz würden wir in Block 3 erwarten. Da dort die Sorte 3 den Ertrag 60 dt/ha hat, erwarten wir $60 + 20 = 80$ dt/ha für die Sorte 3. Wenn wir diesen geschätzten Wert für den fehlenden Wert einsetzen, erhalten wir folgende Mittelwerte:

		Sorte			
		1	2	3	
Block	1	10	20	30	
	2	20	30	40	
	3	60	70	80	
Sortenmittel:		30	40	50	

Der sich für Sorte 3 ergebende Mittelwert über die drei Blöcke ist ein sog. **adjustierter Mittelwert (korrigierter Mittelwert)**. Dasselbe Ergebnis würden wir erhalten beim Vergleich der Sorten 2 und 3, weil die Daten fehlerfrei sind.

Beispiel: Das erste Beispiel war recht künstlich, weil Abwesenheit von Fehlern angenommen wurde. Nun fügen wir Fehlereffekte hinzu und erhalten folgende Daten:

		Sorte		
		1	2	3
Block	1	12	18	37
	2	18	32	33
	3	57	73	.
Sortenmittel $\bar{y}_{i.}$:		29	41	35

Die Sortendifferenzen sind nun nicht mehr dieselben von Block zu Block. Analysieren wir die ersten beiden Blöcke für sich und verwerfen den unvollständigen dritten Block, so erhalten wir:

		Sorte		
		1	2	3
Block	1	12	18	37
	2	18	32	33
Sortenmittel $\bar{y}_{i.}$:		15	25	35

Der plausibelste Wert für den fehlenden Wert in Block 3 ist derjenige, der in der besten Übereinstimmung ist mit den Sortendifferenzen, die sich für die ersten beiden Blöcke ergeben. Die Betrachtung wird dadurch kompliziert, dass die Differenzen sich von Block zu Block unterscheiden. Eine mögliche Forderung für den fehlenden Wert ist, dass die Differenz zu den Werten von Sorte 1 und 2 in Block 3 *im Durchschnitt* den jeweiligen Differenzen aus den ersten beiden Blöcken entsprechen soll. Wir verwenden den Durchschnitt, weil völlige Übereinstimmung für jede einzelne Differenz wegen des Versuchsfehlers nicht möglich ist. Wir bezeichnen im folgenden den fehlenden Wert mit m :

		Sorte		
		1	2	3
Block	1	12	18	37
	2	18	32	33
	3	57	73	m

Nun berechnen wir die Differenzen unter Verwendung des fehlenden Wertes (m):

	Sortendifferenzen		
	Differenz I 3 minus 1	Differenz II 3 minus 2	Durchschnitt I und II
Durchschnitt über Blöcke 1 & 2	20	10	15
Daten in Block 3	$m - 57$	$m - 73$	$m - 65$

Setzen wir nun die durchschnittliche Differenz von "3-1" und "3-2" in den beiden ersten Blöcken gleich den entsprechenden Durchschnitt in Block 3, so erhalten wir

$$15 = m - 65 \Leftrightarrow m = 80$$

Einsetzen des geschätzten fehlenden Wertes in die Tabelle liefert:

	Sorte			
	1	2	3	
Block	1	12	18	37
	2	18	32	33
	3	57	73	80
Sortenmittel:	29	41	50	

Geschätzter Wert

Adjustierter Mittelwert!

Der sich für Sorte 3 ergebende Mittelwert über die drei Blöcke ist wieder ein adjustierter Mittelwert.

Eine alternative Methode zur Bestimmung des fehlenden Wertes m verwendet die Methode der kleinsten Quadrate. Für die beiden Differenzen (3 minus 1 und 3 minus 2) haben wir einen Schätzwert aus den ersten beiden Blöcken. Diese sollen im Sinne der Kleinsten Quadrate möglichst nahe an den entsprechenden Differenzen im dritten Block liegen, was zu folgendem zu minimierenden Zielkriterium führt:

$$SQ = [20 - (m - 57)]^2 + [10 - (m - 73)]^2 = (77 - m)^2 + (83 - m)^2$$

Optimierung führt zu folgender Lösung (dieselbe wie oben):

$$\partial SQ / \partial m = 2(77 - m) + 2(83 - m) = 0 \Leftrightarrow 154 + 166 = 4m \Leftrightarrow m = 320 / 4 = 80$$

Noch eine Methode: Man kann für die unvollständige Tabelle Block x Sorte sog. Tetraden bilden, welche jeweils die leere Zelle umfassen. Eine Tetrade stellt eine 2 x 2 Tafel dar aus zwei Blöcken und zwei Sorten, wobei eine Block-Sorten-Kombination den fehlenden Wert repräsentiert. Im vorliegenden Beispiel ist eine solche Tetrade gegeben durch:

	Sorte	
	2	3
Block		
2	32	33
3	73	m

Der fehlende Wert m wird anhand dieser Tetrade geschätzt nach der Gleichung $32 - 33 = 73 - m$. Die Lösung ist $m = 74$. Es gibt vier solcher Tetraden, welche den fehlenden Wert m umfassen. Jede liefert einen anderen Schätzwert für m . Das Mittel dieser vier Werte beträgt, wie man leicht nachprüft, wiederum $m = 80$.

Wir haben nun an zwei einfachen Beispielen intuitiv angedeutet, was adjustierte Mittelwerte sind. Eine allgemeine Methode, die für die beiden Beispiele zu denselben Ergebnissen führt, ist die Methode der Kleinsten Quadrate, wobei das additive Modell $\mu + b_j + \tau_i$ zugrunde gelegt wird. Adjustierte Mittelwerte lassen sich über dieses Modell definieren. Hierzu betrachten wir den Erwartungswert einer Beobachtung:

$$E(y_{ij}) = \eta_{ij} = \mu + b_j + \tau_i$$

Es ist naheliegend, den Erwartungswert als den Durchschnitt der η_{ij} über die Blöcke zu definieren, also

$$\bar{\eta}_{i.} = \mu + \bar{b}_{.} + \tau_i = \mu + \frac{b_1 + b_2 + \dots + b_r}{r} + \tau_i$$

Um diese Mittelwerte zu schätzen, setzen wir die Kleinst-Quadrat-Lösungen der Effekte ein, die hier wie üblich mit $\hat{\mu}$, \hat{b}_j , und $\hat{\tau}_i$ bezeichnet werden. Der sich ergebende Schätzer für den Mittelwert $\bar{\eta}_{i.}$, der sog. **adjustierte Mittelwert** oder **Kleinst-Quadrat-Mittelwert** (least square mean = LSMEAN), lautet

$$\hat{\bar{\eta}}_{i.} = \hat{\mu} + \hat{\bar{b}}_{.} + \hat{\tau}_i = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \dots + \hat{b}_r}{r} + \hat{\tau}_i$$

Im **Spezialfall balancierter Daten** stimmen adjustierte Mittelwerte und einfache (arithmetische) Mittelwerte überein:

$$\hat{\bar{\eta}}_{i.} = \bar{y}_{i.}$$

Beispiel: Für das zweite Beispiel lautet das lineare Modell in Matrixform

$$y = \begin{pmatrix} 12 \\ 18 \\ 37 \\ 18 \\ 32 \\ 33 \\ 57 \\ 73 \end{pmatrix} = X\beta + e = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ b_1 \\ b_2 \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{31} \\ e_{32} \end{pmatrix}$$

Hierbei haben wir die Restriktion $\tau_3 = b_3 = 0$ gewählt. Eine solche Restriktion erlaubt es, in einem überparametrisierten Modell wie dem vorliegenden eine Kleinst-Quadrat-Lösung zu erhalten (Abschnitt 6.8). Wir finden folgende adjustierten Mittelwerte oder Kleinst-Quadrat-Mittelwerte:

$$\hat{\beta} = (X'X)^{-1} X'y = \begin{pmatrix} \hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{pmatrix} = \begin{pmatrix} 80 \\ -47,66 \\ -42,33 \\ -21 \\ -9 \end{pmatrix}$$

$$\hat{\eta}_{1\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_1 = 80 + \frac{-47,66 - 42,33 + 0}{3} - 21 = 29$$

$$\hat{\eta}_{2\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_2 = 80 + \frac{-47,66 - 42,33 + 0}{3} - 9 = 41$$

$$\hat{\eta}_{3\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_3 = 80 + \frac{-47,66 - 42,33 + 0}{3} + 0 = 50$$

Diese Mittelwerte können mittels t-Tests verglichen werden. Hierzu beachte man, dass die Differenz zweier Mittelwerte nur von den Behandlungseffekten abhängt. So gilt beispielsweise

$$\bar{\eta}_{1\bullet} - \bar{\eta}_{2\bullet} = \tau_1 - \tau_2 = k'\beta = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ b_1 \\ b_2 \end{pmatrix}$$

Hierbei ist $k' = (0 \ 1 \ -1 \ 0 \ 0)$ ein Koeffizientenvektor (siehe Abschnitt 6.7). Da die Differenz zweier Mittelwerte die allgemeine Form $\lambda = k'\beta$ hat, können wir die allgemeinen Resultate aus Abschnitt 6.8 anwenden, um einen t-Test der Nullhypothese $H_0: \tau_1 - \tau_2 = 0$ durchzuführen. Wir zeigen einige Zwischenschritte der

Berechnungen exemplarisch für den Vergleich der Sorten 1 und 2. In der Praxis wird man die Berechnungen mit einem PC Programm für lineare Modelle durchführen.

$$SQ_{Fehler} = (y - Xb)'(y - Xb) = 93,33$$

n = Zahl der Beobachtungen in $y = 8$

$$FG_{Fehler} = n - Rang(X) = 3$$

$$s^2 = SQ_{Fehler}/FG_{Fehler} = 31,11$$

Man beachte, dass die Designmatrix X fünf unabhängige Spalten hat, wenn wir die Restriktionen $\tau_3 = 0$ und $b_2 = 0$ beachten, so dass $Rang(X) = 5$ und die Fehler-Freiheitsgrade gleich 3 sind [$FG_{Fehler} = n - Rang(X) = 3$].

$$H_0: \lambda = k'\beta = \tau_1 - \tau_2 = 0$$

$$\hat{\lambda} = \hat{\tau}_1 - \hat{\tau}_2 = -21 + 9 = -12$$

$$t_{Vers} = \frac{|\hat{\lambda}|}{\sqrt{k'(X'X)^{-1}ks^2}} = \frac{12}{\sqrt{0,666 * 31,11}} = 2,63 < t_{Tab}(FG = 3; \alpha = 5\%) = 3,182$$

Die beiden Sorten sind nicht signifikant verschieden. Dies Ergebnis erhalten wir auch mit folgenden SAS Anweisungen:

```
data;
input block cultivar yield;
datalines;
1 1 12
1 2 18
1 3 37
2 1 18
2 2 32
2 3 33
3 1 57
3 2 73
3 3 .
;
proc mixed;
class cultivar block;
model yield=block cultivar/solution;
lsmeans cultivar/pdiff; run;
```

Ergebnis:

Least Squares Means

Effect	cultivar	Estimate	Standard Error	DF	t Value	Pr > t
cultivar	1	29.0000	3.2203	3	9.01	0.0029
cultivar	2	41.0000	3.2203	3	12.73	0.0010
cultivar	3	50.0000	4.2601	3	11.74	0.0013

Differences of Least Squares Means

Effect	cultivar	_cultivar	Estimate	Standard Error	DF	t Value	Pr > t
cultivar	1	2	-12.0000	4.5542	3	-2.63	0.0780
cultivar	1	3	-21.0000	5.3403	3	-3.93	0.0293
cultivar	2	3	-9.0000	5.3403	3	-1.69	0.1905

Es besteht nach dem t-Test ein signifikanter Unterschied zwischen den Sorten 1 und 3. Man beachte, dass die Standardfehler der Differenzen nicht konstant sind: Der Fehler ist am kleinsten für den Vergleich zwischen Sorte 1 und 2, weil für diese kein Wert fehlt ("1-2": 4.5542, "1-3" und "2-3": 5,3403).

Um das versuchsbezogene Irrtumsniveau einzuhalten, kann das **Simulationsverfahren von Edwards und Berry** (1987; *Biometrics* **43**, 913-928) verwendet werden, welches im balancierten Fall dem Tukey-Test entspricht.

Beispiel: Die Anweisungen für das Reisbeispiel sind:

```
data;
input dichte block ertrag;
cards;
25      1      .
25      2      5398
25      3      5307
25      4      4678
50      1      5346
50      2      5952
50      3      4719
50      4      4264
75      1      5272
75      2      5713
75      3      5483
75      4      4749
100     1      5164
100     2      4831
100     3      4986
100     4      4410
125     1      4804
125     2      4848
125     3      .
125     4      4748
150     1      5254
150     2      4542
150     3      4919
150     4      4098
;
proc glm;
class block dichte;
model ertrag=block dichte;
lsmeans dichte/pdiff adjust=sim(nsamp=100000);
```

```
run;

proc mixed;
class block dichte;
model ertrag=block dichte;
lsmeans dichte/pdiff adjust=sim(nsamp=100000);
run;
```

Ergebnis:

Least Squares Means

Effect	dichte	Estimate	Standard Error	DF	t Value	Pr > t
dichte	25	5200.69	198.70	13	26.17	<.0001
dichte	50	5070.25	167.82	13	30.21	<.0001
dichte	75	5304.25	167.82	13	31.61	<.0001
dichte	100	4847.75	167.82	13	28.89	<.0001
dichte	125	4819.19	198.70	13	24.25	<.0001
dichte	150	4703.25	167.82	13	28.02	<.0001

Diese p-Werte
sind irrelevant!

Differences of Least Squares Means

Edwards -Berry p-Werte

Effect	dichte	_dichte	Estimate	Standard Error	DF	t Value	Pr > t	Adjustment	Adj P
dichte	25	50	130.44	260.09	13	0.50	0.6244	Simulate	0.9955
dichte	25	75	-103.56	260.09	13	-0.40	0.6970	Simulate	0.9986
dichte	25	100	352.94	260.09	13	1.36	0.1979	Simulate	0.7473
dichte	25	125	381.50	283.67	13	1.34	0.2017	Simulate	0.7540
dichte	25	150	497.44	260.09	13	1.91	0.0781	Simulate	0.4342
dichte	50	75	-234.00	237.34	13	-0.99	0.3422	Simulate	0.9139
dichte	50	100	222.50	237.34	13	0.94	0.3656	Simulate	0.9291
dichte	50	125	251.06	260.09	13	0.97	0.3520	Simulate	0.9206
dichte	50	150	367.00	237.34	13	1.55	0.1460	Simulate	0.6394
dichte	75	100	456.50	237.34	13	1.92	0.0766	Simulate	0.4286
dichte	75	125	485.06	260.09	13	1.86	0.0849	Simulate	0.4597
dichte	75	150	601.00	237.34	13	2.53	0.0250	Simulate	0.1831
dichte	100	125	28.5625	260.09	13	0.11	0.9142	Simulate	1.0000
dichte	100	150	144.50	237.34	13	0.61	0.5531	Simulate	0.9884
dichte	125	150	115.94	260.09	13	0.45	0.6631	Simulate	0.9974

Kompakte Darstellung:

Saatstärke Mittelwert

25	5200,7 ^a
50	5070,3 ^a
75	5304,4 ^a
100	4847,8 ^a
125	4819,2 ^a
150	4703,3 ^a

(Mittelwerte, die mit einem gemeinsamen Buchstaben versehen sind, sind nicht signifikant voneinander verschieden).

Die Buchstabendarstellung muss hier von Hand mit der in Abschnitt 4.4 beschriebenen Methode hergeleitet werden. Im vorliegenden Fall ist dies einfach, da alle Tests nicht signifikant sind. Im allgemeinen ist jedoch nicht garantiert, dass die Buchstabendarstellung mit der in 4.4 beschriebenen Methode möglich ist. Dies ist der Grund, warum die Buchstabendarstellung nicht mit der LSMEANS Anweisung erhalten werden kann. Für diesen Fall steht eine allgemeinere Methode zur Verfügung (H.P. Piepho, 2004, *Journal of Computational and Graphical Statistics*), die in einem SAS Makro implementiert ist (<http://www.uni-hohenheim.de/bioinformatik/>). Das Makro kann in jedem Fall benutzt werden, um die Buchstabendarstellung berechnen zu lassen.

8.8 Auswertung eines Lateinischen Quadrats

8.8.1 Varianzanalyse eines Lateinischen Quadrats

Die Auswertung von Lateinischen Quadraten erfolgt nach dem Modell

$$y_{ijh} = \mu + b_j + c_h + \tau_i + e_{ijh}$$

$$(i, j, h = 1, \dots, t)$$

y_{ijh} = Beobachtung der i -ten Behandlung in der j -ten Zeile und der h -ten Spalte

μ = Gesamteffekt

b_j = Effekt der j -ten Zeile

c_h = Effekt der h -ten Spalte

τ_i = Effekt der i -ten Behandlung

e_{ijh} = Fehler der Beobachtung y_{ijh}

Zur Auswertung verwenden wir wieder die allgemeinen Methoden aus Abschnitt 6.8 an. Die Varianzanalyse basiert auf der folgenden Sequenz von Modellen:

Modell	SQ_{Fehler}
(0) $y_{ijh} = \mu + e_{ijh}$	$SQ_{\text{Fehler}}^{(0)}$
(1) $y_{ijh} = \mu + b_j + e_{ijh}$	$SQ_{\text{Fehler}}^{(1)}$
(2) $y_{ijh} = \mu + b_j + c_h + e_{ijh}$	$SQ_{\text{Fehler}}^{(2)}$
(3) $y_{ijh} = \mu + b_j + c_h + \tau_i + e_{ijh}$	$SQ_{\text{Fehler}}^{(3)}$
Quadratsummen:	
$SQ(b_j \mu)$	$= SQ_{\text{Fehler}}^{(0)} - SQ_{\text{Fehler}}^{(1)}$
$SQ(c_h b_j, \mu)$	$= SQ_{\text{Fehler}}^{(1)} - SQ_{\text{Fehler}}^{(2)}$
$SQ(\tau_i c_h, b_j, \mu)$	$= SQ_{\text{Fehler}}^{(2)} - SQ_{\text{Fehler}}^{(3)}$

Varianzanalyse-Tabelle:

Ursache	FG	SQ	F_{Vers}
Zeilen	$t - 1$	$SQ(b_j \mu)$	$F_{Vers} = \frac{SQ(\tau_i c_h, b_j, \mu)/(t-1)}{SQ_{Fehler}^{(3)}/(n-3t+2)} = \frac{MQ_{Beh}}{MQ_{Fehler}}$
Spalten (adj.)	$t - 1$	$SQ(c_h b_j, \mu)$	
Behandlungen (adj.)	$t - 1$	$SQ(\tau_i c_h, b_j, \mu)$	
Fehler	$n - 3t + 2$	$SQ_{Fehler}^{(3)}$	

t = Zahl der Behandlungen = Zahl der Zeilen = Zahl der Spalten

n = Gesamtzahl der Beobachtungen

Bei **balancierten Daten** gelten die folgenden einfachen Rechenformeln:

Zeilensumme: $y_{\bullet j \bullet}$

Spaltensumme: $y_{\bullet \bullet h}$

Behandlungssumme: $y_{i \bullet \bullet}$

Gesamtsumme: $y_{\bullet \bullet \bullet}$

$$SQ_{Zeilen} = t \sum_{j=1}^t (\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet \bullet \bullet})^2 = \sum_{j=1}^t y_{\bullet j \bullet}^2 / t - y_{\bullet \bullet \bullet}^2 / t^2$$

$$SQ_{Spalten} = t \sum_{h=1}^t (\bar{y}_{\bullet \bullet h} - \bar{y}_{\bullet \bullet \bullet})^2 = \sum_{h=1}^t y_{\bullet \bullet h}^2 / t - y_{\bullet \bullet \bullet}^2 / t^2$$

$$SQ_{Behandlung} = t \sum_{i=1}^t (\bar{y}_{i \bullet \bullet} - \bar{y}_{\bullet \bullet \bullet})^2 = \sum_{i=1}^t y_{i \bullet \bullet}^2 / t - y_{\bullet \bullet \bullet}^2 / t^2$$

$$SQ_{Total} = \sum_{i,j,h} (y_{ijh} - \bar{y}_{\bullet \bullet \bullet})^2 = \sum_{i,j,h} y_{ijh}^2 - y_{\bullet \bullet \bullet}^2 / t^2$$

$$SQ_{Fehler} = SQ_{Total} - SQ_{Zeilen} - SQ_{Spalten} - SQ_{Behandlung}$$

Die Varianzanalyse-Tabelle hat im balancierten Fall die folgende Form:

Ursache	FG	SQ
Zeilen	$t-1$	SQ_{Zeilen}
Spalten	$t-1$	$SQ_{Spalten}$
Behandlungen	$t-1$	$SQ_{Behandlungen}$
Fehler	$(t-1)(t-2)$	SQ_{Fehler}

Beispiel (Mudra, 1958): Der Bakteriengehalt der Milch auf fünf Betrieben (Behandlungen) wurde in einem Lateinischen Quadrat untersucht mit Tageszeiten und Tagen als Zeilen- und Spaltenfaktoren. Die Logarithmen der Bakterienzahlen waren wie folgt:

Tageszeit	1	2	Tag 3	4	5	Summe
08:30	A 1,9	B 1,2	C 0,7	D 2,2	E 2,3	8,3
10:00	D 2,3	C 2,0	E 0,6	B 2,6	A 2,3	9,8
11:30	C 2,1	A 1,5	D 1,7	E 1,1	B 3,0	9,4
14:00	B 2,9	E 1,1	A 1,2	C 1,8	D 2,6	9,6
15:30	E 1,8	D 2,1	B 2,0	A 2,4	C 2,5	10,8
Summe:	11,0	7,9	6,2	10,1	12,7	47,9
Betrieb: Betriebssumme:	A 9,3	B 11,7	C 9,1	D 10,9	E 6,9	

Wir wollen eine Varianzanalyse für diese **balancierten Daten** durchführen.

$$SQ_{\text{Zeilen}} = SQ_{\text{Tageszeiten}} = (8,3^2 + 9,8^2 + 9,4^2 + 9,6^2 + 10,8^2)/5 - 47,9^2/25 = 0,64$$

$$SQ_{\text{Spalten}} = SQ_{\text{Tage}} = (11,0^2 + 7,9^2 + 6,2^2 + 10,1^2 + 12,7^2)/5 - 47,9^2/25 = 5,25$$

$$SQ_{\text{Behandlung}} = SQ_{\text{Betrieb}} = (9,3^2 + 11,7^2 + 9,1^2 + 10,9^2 + 6,9^2)/5 - 47,9^2/25 = 2,74$$

$$SQ_{\text{Total}} = (1,9^2 + 2,3^2 + \dots + 2,5^2) - 47,9^2/25 = 10,07$$

$$SQ_{\text{Fehler}} = 10,07 - 0,64 - 5,25 - 2,74 = 1,44$$

Varianzanalyse-Tabelle:

Ursache	FG	SQ	MQ	F_{Vers}
Zeilen	4	0,64	0,160	
Spalten	4	5,25	1,312	
Behandlungen	4	2,74	0,685	5,71
Fehler	12	1,44	0,120	

Da $F_{\text{Vers}} = 5,71 > F_{\text{Tab}}(4, 12, \alpha = 5\%) = 3,26$, liegen signifikante Unterschiede zwischen den Betrieben vor.

Nehmen wir nun an, dass zwei Beobachtungen fehlen (**unbalancierte Daten**):

Tageszeit	Tag				
	1	2	3	4	5
08:30	A 1,9	B 1,2	C .	D 2,2	E 2,3
10:00	D 2,3	C 2,0	E 0,6	B 2,6	A 2,3
11:30	C 2,1	A 1,5	D 1,7	E 1,1	B 3,0
14:00	B 2,9	E 1,1	A 1,2	C 1,8	D .
15:30	E 1,8	D 2,1	B 2,0	A 2,4	C 2,5

Modell	SQ_{Fehler}	FG_{Fehler}
(0) $y_{ijh} = \mu + e_{ijh}$	8,115	22
(1) $y_{ijh} = \mu + b_j + e_{ijh}$	7,702	18
(2) $y_{ijh} = \mu + b_j + c_h + e_{ijh}$	4,044	14
(3) $y_{ijh} = \mu + b_j + c_h + \tau_i + e_{ijh}$	1,292	10

$$SQ(b_j|\mu) = 8,115 - 7,702 = 0,413$$

$$SQ(c_h|b_j, \mu) = 7,702 - 4,044 = 3,658$$

$$SQ(\tau_i|c_h, b_j, \mu) = 4,044 - 1,292 = 2,752$$

$$SQ_{Fehler}^{(3)} = 1,292$$

$$t = 5$$

$$FG_{Fehler} = n - 3t + 2 = 23 - 15 + 2 = 10$$

Ursache	FG	SQ	MQ	F_{Vers}
Zeilen	4	0,413	0,1032	5,32
Spalten (adj.)	4	3,658	0,9145	
Behandlungen (adj.)	4	2,752	0,6879	
Fehler	10	1,292	0,1292	

$$F_{Tab}(\alpha = 5\%, FG_1 = 4, FG_2 = 10) = 3,48 < F_{Vers}$$

Es bestehen signifikante Behandlungsunterschiede.

8.8.2 Mittelwertvergleiche in einem Lateinischen Quadrat

Bei **balancierten Daten** können einfache arithmetische Mittelwerte berechnet und mit dem LSD oder HSD (Tukey) Test verglichen werden. LSD und HSD berechnen sich analog zu der vollständig randomisierten Anlage (Abschnitt 4.4). Der einzige Unterschied besteht in der Zahl der Freiheitsgrade für den Fehler. Es gilt:

Vergleichsbezogene Irrtumswahrscheinlichkeit (t-Test; t_{tab} aus Tab. II):

Verwerfe $H_0: \tau_s = \tau_u$, wenn $|\bar{y}_{s.} - \bar{y}_{u.}| > t_{Tab}(n - 3t + 2, \alpha) \sqrt{2MQ_{Fehler}/t} = LSD$

Versuchsbezogene Irrtumswahrscheinlichkeit (Tukey; q_{tab} aus Tab. VII):

Verwerfe $H_0: \tau_s = \tau_u$, wenn $|\bar{y}_{s.} - \bar{y}_{u.}| > q_{Tab}(t, n - 3t + 2, \alpha) \sqrt{MQ_{Fehler}/t} = HSD$

Beispiel: Für die balancierten Milch-Daten finden wir:

$$q_{Tab}(t = 5, 12, 5\%) = 4,51$$

$$MQ_{Fehler} = 0,120$$

$$t = 5$$

$$HSD = q_{Tab}[t, (n - 3t + 2, \alpha)] \sqrt{MQ_{Fehler}/t} = 4,51 \times \sqrt{0,120/5} = 0,70$$

Die Mittelwerte der Bakterienzahlen (log-Skala) sind:

Betriebe Mittelwert

A	1,86 ^{ab}
B	2,34 ^a
C	1,82 ^{ab}
D	2,18 ^a
E	1,38 ^b

Mittelwerte, die mit demselben Buchstaben versehen sind, sind nicht signifikant verschieden.

HSD 0,70

Die Herleitung der Buchstabendarstellung ist wie folgt (vgl. Abschnitt 4.4). Wir ordnen die Betriebe der Größe nach und tragen sie in eine Kreuztabelle der folgenden Form ein:

	B (2,34)	D (2,18)	A (1,86)	C (1,82)	E (1,38)
B (2,34)		0,16	0,48	0,52	0,96*
D (2,18)			0,32	0,36	0,80*
A (1,86)				0,04	0,48
C (1,82)					0,44
E (1,38)					

(*Signifikante Differenz: HSD = 0,70)

Zu beachten ist, dass die Behandlungen sowohl in den Zeilen als auch in den Spalten der Größe nach geordnet sind. In der Tabelle sind auch die Werte der

B D A C E

B D A C E

B D A C E

B D A C E

B D A C E

_____ a

 _____ b

Wir schreiben die Behandlungen und ihre Mittelwerte in eine Tabelle. Sodann werden die Buchstaben der Linien, von denen eine Behandlung unterstrichen ist, hinter den Mittelwert der betreffenden Behandlung geschrieben (siehe obige Tabelle).

Bei **unbalancierten Daten** betrachten wir wieder den Erwartungswert einer Beobachtung

$$E(y_{ijh}) = \eta_{ijh} = \mu + b_j + c_h + \tau_i$$

Die adjustierten Mittelwerte sind Kleinstquadratschätzungen der Funktion

$$\bar{\eta}_{i..} = \mu + \bar{b}_{.} + \bar{c}_{.} + \tau_i$$

Im unbalancierten Fall werden, wie bei der Blockanlage, die Kleinst-Quadrat-Schätzungen der Effekte eingesetzt, um adjustierte Mittelwerte zu erhalten. Für die Durchführung von Tests können die allgemeinen Methoden aus Abschnitt 6.8 verwendet werden. Die SAS Anweisungen für den unbalancierten Milchdatensatz lauten bei Verwendung des t-Tests (vergleichsbezogene Irrtumswahrscheinlichkeit):

```
data;
input zeile spalte betrieb $ y;
datalines;
1      1      A      1.9
1      2      B      1.2
1      3      C      .
1      4      D      2.2
1      5      E      2.3
2      1      D      2.3
2      2      C      2.0
2      3      E      0.6
2      4      B      2.6
2      5      A      2.3
3      1      C      2.1
3      2      A      1.5
3      3      D      1.7
3      4      E      1.1
3      5      B      3.0
4      1      B      2.9
4      2      E      1.1
4      3      A      1.2
4      4      C      1.8
4      5      D      .
5      1      E      1.8
5      2      D      2.1
5      3      B      2.0
5      4      A      2.4
5      5      C      2.5
;
proc mixed;
class zeile spalte betrieb;
model y=zeile spalte betrieb;
```

```
lsmeans betrieb/pdiff;
run;
```

Ergebnis:

Least Squares Means

Effect	betrieb	Estimate	Standard Error	DF	t Value	Pr > t
betrieb	A	1.8600	0.1608	10	11.57	<.0001
betrieb	B	2.3400	0.1608	10	14.56	<.0001
betrieb	C	1.8857	0.1922	10	9.81	<.0001
betrieb	D	2.2557	0.1922	10	11.74	<.0001
betrieb	E	1.3800	0.1608	10	8.58	<.0001

Differences of Least Squares Means

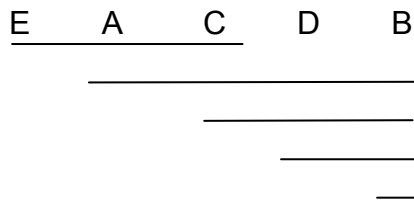
Effect	betrieb	_betrieb	Estimate	Standard Error	DF	t Value	Pr > t
betrieb	A	B	-0.4800	0.2274	10	-2.11	0.0609
betrieb	A	C	-0.02571	0.2505	10	-0.10	0.9203
betrieb	A	D	-0.3957	0.2505	10	-1.58	0.1453
betrieb	A	E	0.4800	0.2274	10	2.11	0.0609
betrieb	B	C	0.4543	0.2505	10	1.81	0.0999
betrieb	B	D	0.08429	0.2505	10	0.34	0.7435
betrieb	B	E	0.9600	0.2274	10	4.22	0.0018
betrieb	C	D	-0.3700	0.2785	10	-1.33	0.2134
betrieb	C	E	0.5057	0.2505	10	2.02	0.0712
betrieb	D	E	0.8757	0.2505	10	3.50	0.0058

Die Standardfehler der Differenzen sind nicht konstant, so dass keine gemeinsame Grenzdifferenz berechnet werden kann. Eine Buchstabendarstellung erhalten wir mit der Methode aus Abschnitt 4.4. Die Behandlungen werden nach der Größe der Mittelwerte sortiert und in eine Kreuztabelle wie folgt eingetragen:

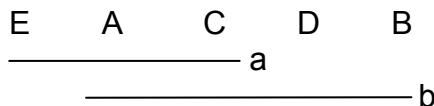
	E (1,38)	A (1,86)	C (1,89)	D (2,26)	B (2,34)
E		ns	ns	*	*
A			ns	ns	ns
C				ns	ns
D					ns
B					

(*Signifikante Differenz nach t-Test)

In die Tabelle wird eingetragen, ob ein Vergleich signifikant (*) oder nicht signifikant (ns) war. Sodann wird aus der Tabelle eine Liniendarstellung abgeleitet, wobei nicht signifikant verschiedene Behandlungen unterstrichen werden (siehe Abschnitt 4.4).



Die obige Darstellung enthält noch einige Redundanzen. Die Linien 3, 4 und 5 sind in der zweiten Linie "enthalten" und können daher wegfallen. Die Darstellung sieht dann wie folgt aus:



Betriebe	Mittelwert
A	1,86 ^{ab}
B	2,34 ^b
C	1,89 ^{ab}
D	2,26 ^b
E	1,38 ^a

Mittelwerte, die mit einem gemeinsamen Buchstaben versehen sind, sind nicht signifikant verschieden.

Es kann bei Unbalanciertheit der Daten manchmal passieren (auch bei der Blockanlage), dass eine Buchstabendarstellung nicht mit der in Abschnitt 4.4 besprochenen Methode möglich ist. Für diesen Fall steht eine allgemeinere Methode zur Verfügung, die in einem SAS Makro implementiert ist (<http://www.uni-hohenheim.de/bioinformatik/>). Das Makro kann in jedem Fall benutzt werden, um die Buchstabendarstellung berechnen zu lassen.

Abschließend sei noch bemerkt, dass wie bei der Blockanlage im Fall **balancierter Daten** die Kleinstquadratschätzung der Mittelwerte identisch mit dem arithmetischen Mittelwert sind: $\hat{\eta}_{i..} = \bar{y}_{i..}$

8.9 Regression in einer Blockanlage

Beispiel (balanciert): Ein Blockversuch, bei dem der Ertrag (kg/ha) der Reissorte IR8 bei 6 verschiedenen Aussaatstärken gemessen wurde, lieferte folgendes Ergebnis (Gomez und Gomez, 1984):

Saatstärke (kg/ha)	Block			
	1	2	3	4
25	5113	5398	5307	4678
50	5346	5952	4719	4264
75	5272	5713	5483	4749
100	5164	4831	4986	4410
125	4804	4848	4432	4748
150	5254	4542	4919	4098

Die Einflussvariable Aussaatstärke ist quantitativ, so dass eine Regression für die Auswertung sinnvoll ist. Wir hatten diesen Datensatz im Abschnitt 8.7 zunächst so ausgewertet, als sei der Faktor qualitativ. Nun soll berücksichtigt werden, dass eine Regression möglich ist. Wir wollen eine lineare Regression anpassen und hierbei einen Lack-of-fit Test durchführen, d.h. auf eine Abweichung von der Linearität prüfen, so wie dies in Abschnitt 6.6 bereits beschrieben worden ist. Hierzu wird folgende Modellsequenz betrachtet:

Modell	SQ_{Fehler}	FG_{Fehler}
(0) $y_{ij} = \mu + e_{ij}$	4801067,83	23
(1) $y_{ij} = \mu + b_j + e_{ij}$	2856707,00	20
(2) $y_{ij} = \mu + b_j + \beta_1 x_i + e_{ij}$	2096672,20	19
(3) $y_{ij} = \mu + b_j + \beta_1 x_i + \delta_i + e_{ij}$	1658376,17	15

Der Parameter δ_i ist der sog. Lack-of-fit Effekt. Falls eine lineare Beziehung zwischen y und x besteht, so gilt $H_0: \delta_1 = \delta_2 = \dots = \delta_t = 0$, und man kann diesen Effekt wegfällen lassen (vgl. Abschnitt 6.6).

Varianzanalyse:

Ursache	FG	SQ	MQ	F
$b_j \mu$	3	1944360,83	648120,28	5,86
$\beta_1 \mu, b_j$	1	760034,80	760034,80	6,87
$\delta_i \mu, b_j, \beta_1$	4	438296,03	109574,01	0,99
Fehler (e_{ij})	15	1658376,17	110558,41	

Die Steigung ist bei $\alpha = 0,05$ signifikant [$F_{Vers} = 6,87 > F_{Tab}(1-\alpha, 1, n-r-t+1 = 15) = 4,54$], während der Lack-of-Fit Test (δ_i) nicht signifikant ist [$F_{Vers} = 0,99 < F_{Tab}(1-\alpha, 4,$

$n-r-t+1 = 15) = 3,06]$. Das Modell passt somit und wir haben einen linearen Zusammenhang nachgewiesen. Damit haben wir gegenüber einem Mittelwertvergleich an Aussagekraft gewonnen.

Wir haben hier den linearen Term gegen den Rest in einem Modell mit Lack-of-fit-Effekt geprüft. Dies ist dann sinnvoll, wenn der Lack-of-fit-Effekt nicht signifikant ist. Allerdings würde hier im Prinzip auch die Möglichkeit bestehen, den nicht-signifikanten Lack-of-fit-Effekt aus dem Modell zu nehmen und basierend auf dem linearen Regressionsmodell (2) den Fehlerterm neu zu berechnen. Dies bedeutet, dass es in diesem Fall mehrere Möglichkeiten gibt, valide F-Tests zu konstruieren. Allerdings ist das oben angegebene Vorgehen aus folgendem Grunde zu bevorzugen: Bei der Auswertung jeweils mit dem Lack-of-fit-Effekt im Modell wird als Fehlerterm jeweils immer dasselbe MQ_{Fehler} verwendet, und zwar das des **saturierten Modells**, bei dem jede Behandlung einen eigenen, von den anderen Behandlungen unabhängigen Effekt bekommt. Dieses MQ_{Fehler} ist identisch mit demjenigen einer einfachen Varianzanalyse für eine Blockanlage. Der Fehlerterm ist somit unabhängig vom Grad des jeweils betrachteten Polynoms. Diese Unabhängigkeit wäre nicht mehr gegeben, wenn wir zum Test des Polynoms jeweils den Lack-of-fit-Effekt aus dem Modell nehmen.

Zur Berechnung der Freiheitsgrade für die Behandlungseffekte kurz folgende Erläuterung (siehe auch Exkurs weiter unten in diesem Abschnitt beim nächsten Beispiel). Wenn der Faktor Saatstärke hier als qualitativ betrachtet wird (wie im Abschnitt 8.7.1.), so liegen $t - 1 = 5$ Behandlungsfreiheitsgrade vor. Diese werden bei der linearen Regression mit Lack-of-fit Test verbraucht. Ein FG entfällt auf den Regressionskoeffizienten (β_1), der Rest (4 FG) verbleiben für den Lack-of-fit Term (δ_i). Analog werden auch die SQ aufgeteilt. Das gesamte Behandlungs-SQ (qualitativer Faktor) ist gleich 1198330,83 (vgl. Beispiel für balancierte Blockanlage in Abschnitt 8.7.1). Dies entspricht genau der Summe der SQs für den Regressionskoeffizienten (β_1) und den Lack-of-fit Term (δ_i).

Wir wollen nun noch die Regressionsgleichung schätzen. Dazu betrachten wir zunächst den Erwartungswert einer Beobachtung unter dem Modell

$$y_{ij} = \mu + b_j + \beta_1 x_i + e_{ij}$$

Der Erwartungswert ist

$$E(y_{ij}) = \eta_{ij} = \mu + b_j + \beta_1 x_i$$

Unsere Regressionsgleichung erhalten wir durch Mittelung dieses Erwartungswertes einer Beobachtung über die Blöcke:

$$\bar{\eta}_{i.} = \underbrace{\mu + \bar{b}_{.}}_{\text{Achsenabschnitt}} + \beta_1 x_i$$

Die Schätzung der Regressionsgleichung erhalten wir durch Einsetzen der Kleinstquadratlösungen für die Parameter. Wir finden

$$\hat{\mu} = 4855,9$$

$$\hat{b}_1 = 667,7$$

$$\hat{b}_2 = 772,8$$

$$\hat{b}_3 = 483,2$$

$$\hat{b}_4 = 0 \quad (\text{Restriktion!})$$

$$\hat{\beta}_1 = -4,168$$

Einsetzen in die Regressionsgleichung $\bar{\eta}_i = \mu + \bar{b} + \beta_1 x_i$ liefert

$$\begin{aligned} \text{ERTRAG} &= 4855,9 + (667,7 + 772,8 + 483,2 + 0)/4 - 4,168 \times \text{AUSSAATSTÄRKE} \\ &= 5324,3 - 4,168 \times \text{AUSSAATSTÄRKE} \end{aligned}$$

Diese Regressionsgleichung hat folgende Interpretation: Jedes zusätzliche kg Saatgut pro ha senkt den Ertrag um ca. 4,2 kg/ha. Diese Aussage gilt für Aussaatmengen zwischen 25 und 150 kg/ha. □

Beispiel: Zuckerrohrerträge in einer randomisierten vollständigen Blockanlage mit fünf Blöcken und sechs N-Stufen (Peterson, 1994).

Stickstoff (kg/ha)	Block				
	1	2	3	4	5
0	89,49	54,56	74,33	78,20	61,51
25	108,78	102,01	105,04	105,23	106,52
50	136,28	129,51	132,54	132,73	134,02
100	157,63	167,39	155,39	146,85	155,81
150	185,96	176,66	178,53	195,34	185,56
200	195,09	190,43	183,52	180,99	205,69

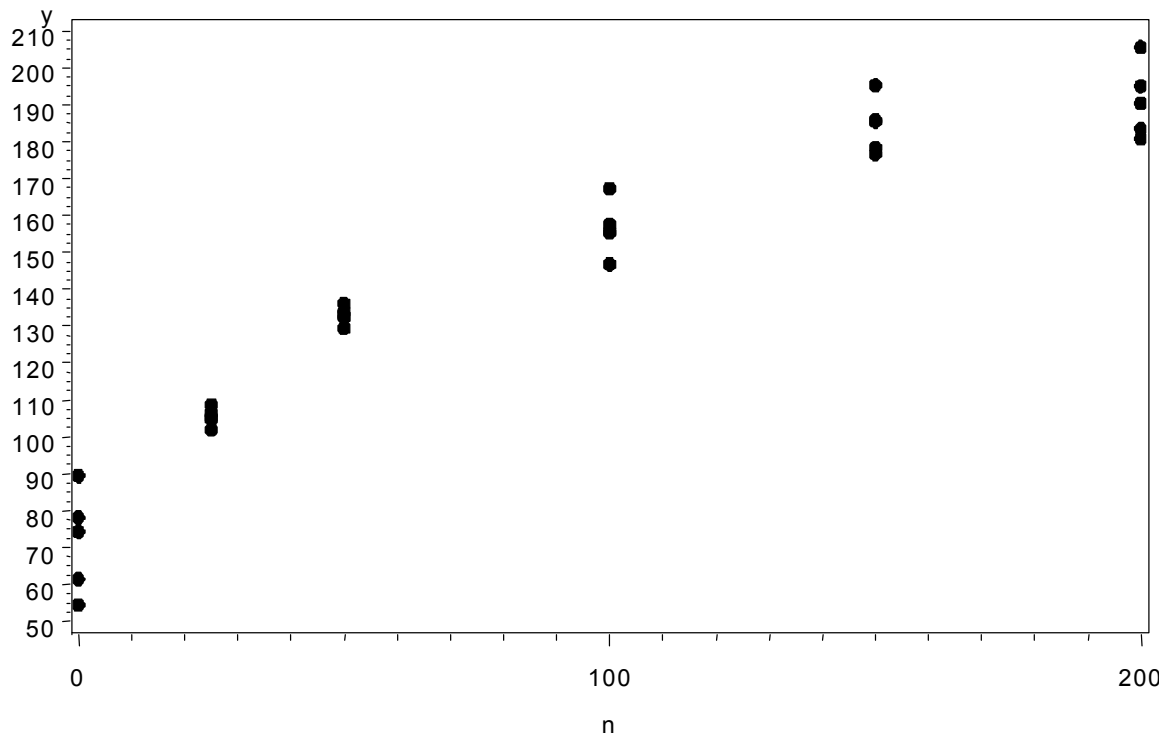


Abb. 8.3: Plot der Erträge gegen die N-Menge.

Der Plot der Daten in Abb. 8.3 legt ein Modell mit abnehmendem Ertragszuwachs nahe. Wir betrachten hier ein Polynom 2. Grades und eine Mitscherlich-Funktion. In beiden Fällen kann ein Lack-of-fit Test durchgeführt werden, da je N-Stufe mehrere Beobachtungen vorliegen (vgl. Abschnitt 6.6). Für das Polynom betrachten wir folgende Modellsequenz:

Modell	SQ_{Fehler}	FG_{Fehler}
(0) $y_{ij} = \mu + e_{ij}$	55518,6	29
(1) $y_{ij} = \mu + b_j + e_{ij}$	55243,3	25
(2) $y_{ij} = \mu + b_j + \beta_1 x_i + e_{ij}$	5778,6	24
(3) $y_{ij} = \mu + b_j + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$	1866,3	23
(4) $y_{ij} = \mu + b_j + \beta_1 x_i + \beta_2 x_i^2 + \delta_i + e_{ij}$	1357,8	20

Varianzanalyse:

Ursache	FG	SQ	MQ	F	p-Wert
$b_j \mu$	4	275,4	68,8	1,01	0,4238
$\beta_1 \mu, b_j$	1	49464,7	49464,7	728,62	<,0001
$\beta_2 \mu, b_j, \beta_1$	1	3912,3	3912,3	57,63	<,0001
$\delta_i \mu, b_j, \beta_1, \beta_2$	3	508,5	169,5	2,50	0,0891
Fehler (e_{ij})	20	1357,8	67,9		

Der Lack-of-fit (δ_i) ist nicht signifikant, daher passt das quadratische Modell. Linearer und quadratischer Term sind signifikant, wobei hier lediglich der signifikante quadratische Term relevant ist. Selbst wenn der lineare Term nicht signifikant wäre, würden wir diesen Term im Modell lassen. Generell sollten bei einem Polynom p -ten Grades immer alle Terme bis x^p im Modell sein. Der Lack-of-fit Term hat 3 FG, weil von den insgesamt $t - 1 = 5$ FG für die Behandlungen zwei für den linearen und den quadratischen Term verbraucht werden, so daß 3 FG für den Lack-of-fit übrig bleiben.

Exkurs zu Freiheitsgraden für Lack-of-fit: Eine andere Sichtweise auf die Freiheitsgrade des Lack-of-fit Terms ergibt sich, wenn man sich vergegenwärtigt, dass man durch 2 Punkte genau eine Gerade legen kann, durch 3 Punkte genau eine quadratische Gleichung, usw. Wenn wir t Behandlungen haben, so haben wir t Mittelwerte für eine Regression zur Verfügung. Durch diese t Punkte können wir genau ein Polynom $(t - 1)$ -ten Grades legen. Den Lack-of-fit können wir daher nur prüfen für Polynome, deren Grad kleiner als $(t - 1)$ ist.

Betrachten wir nun den Fall von $t = 6$ wie im vorliegenden Beispiel, und wählen nun willkürlich z.B. die letzten drei Punkte aus, um ein Polynom 2. Grades anzupassen (Abb. 8.4). Für diese Punkte gibt es eine perfekte Anpassung. Aber für die ersten drei gibt es eine Abweichung, also einen Lack-of-fit. Dies entspricht genau den drei Freiheitsgraden für den Lack-of-fit im quadratischen Modell.

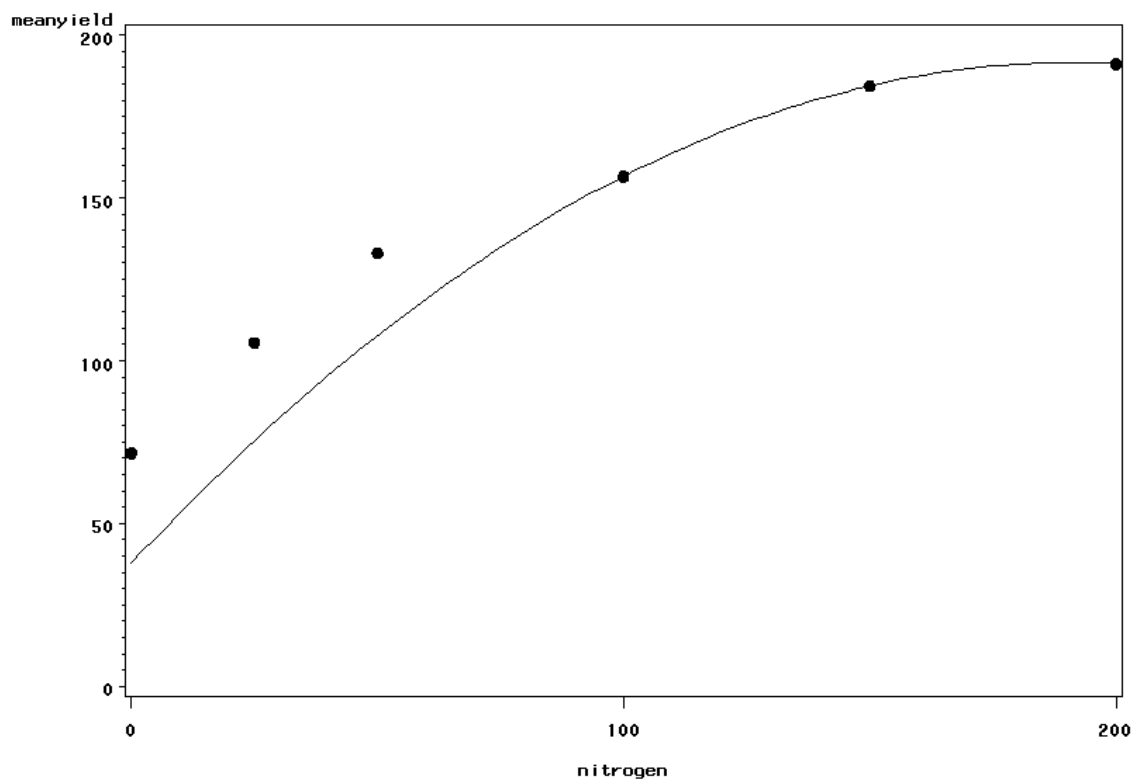


Abb. 8.4: Quadratische Regression durch die Mittelwerte für die drei höchsten N-Gaben. Der Lack-of-fit ergibt sich für die drei niedrigsten N-Gaben, was den drei Freiheitsgraden für den Lack-of-fit entspricht.

Generell haben wir bei einem Polynom p -ten Grades $(t - p - 1)$ Freiheitsgrade für den Lack-of-fit. Setzen wir $p = (t - 1)$, so bleiben keine Freiheitsgrade für den Lack-of-fit übrig.

Zurück zu unserem Beispiel mit sechs N-Düngerstufen. Mit der SAS Prozedur GLM erhält man folgende Kleinst-Quadrat-Lösung für die Effekte:

Parameter		Estimate		Standard Error	t Value	Pr > t
Intercept		38.96800000	B	38.58768810	1.01	0.3246
block	1	4.02000000	B	4.75702915	0.85	0.4081
block	2	-4.75833333	B	4.75702915	-1.00	0.3291
block	3	-3.29333333	B	4.75702915	-0.69	0.4967
block	4	-1.62833333	B	4.75702915	-0.34	0.7357
block	5	0.00000000	B	.	.	.
N		1.60902000	B	0.54405109	2.96	0.0078
N*N		-0.00421240	B	0.00180517	-2.33	0.0302
lackfit	1	33.78200000	B	38.64628751	0.87	0.3924
lackfit	2	30.08725000	B	26.37095912	1.14	0.2674
lackfit	3	25.26000000	B	16.47883236	1.53	0.1410
lackfit	4	0.00000000	B	.	.	.
lackfit	5	0.00000000	B	.	.	.
lackfit	6	0.00000000	B	.	.	.

Die Lack-of-fit Effekte für die höchsten drei N-Stufen sind gleich Null, was genau der Anpassung der quadratischen Regression mit diesen drei Stufen entspricht. Die drei ersten Stufen erhalten dagegen Schätzwerte für den Lack-of-fit, die von Null verschieden sind. Diese Schätzwerte entsprechen gerade genau den vertikalen Abständen der ersten drei Punkte von der Kurve in Abb. 8.4. Die Schätzwerte für den linearen und quadratischen Term entsprechen ebenfalls denjenigen in der Abb. 8.4. Die Quadratsummen für die Polynomialterme sowie den Lack-of-fit würden sich übrigens nicht ändern, wenn die Restriktion für die ersten drei statt der letzten drei N-Stufen angewendet würde. Lediglich die Parameterschätzwerte in diesem überparametrisierten Modell würden sich ändern, nicht jedoch die vorhergesagten Werte (**Ende des Exkurses**).

Für die Mitscherlich-Funktion (nichtlineare Regression) gehen wir analog vor wie bei der quadratischen Regression:

Modell	SQ_{Fehler}	FG_{Fehler}
(0) $y_{ij} = \alpha + e_{ij}$	55518,6	29
(1) $y_{ij} = \alpha + b_j + e_{ij}$	55243,3	25
(2) $y_{ij} = \alpha + b_j + (\beta - \alpha) \exp(-\gamma x_i) + e_{ij}$	1641,8	23
(3) $y_{ij} = \alpha + b_j + (\beta - \alpha) \exp(-\gamma x_i) + \delta_i + e_{ij}$	1357,8	20

Varianzanalyse:

Ursache	FG	SQ	MQ	F	p-Wert
$b_j \alpha$	4	275,4	68,8	1,01	0,4238
$\beta, \gamma \alpha, b_j$	2	53601,5	26800,8	394,7	<,0001
$\delta_i \beta, \gamma, \alpha, b_j$	3	284,0	94,7	1,39	0,2736
Fehler (e_{ij})	20	1357,8	67,9		

Auch hier ist der Lack-of-fit (δ_i) nicht signifikant, während die Regression signifikant ist. Die FG für Lack-of-fit sind hier ebenfalls drei. Zwar hat das Mitscherlich-Modell 3 Parameter, so daß man vermuten könnte, daß nur 2 FG für den Lack-of-fit bleiben. Zu berücksichtigen ist aber, dass der Parameter α dem Achsenabschnitt entspricht, und dieser ist nicht in den insgesamt zu verteilenden Behandlungs-FG enthalten (das ist auch nicht der Fall in der einfachen Varianzanalyse).

Das quadratische Modell und das Mitscherlich-Modell können nicht per F-Test verglichen werden, da sie nicht in einer hierarchischen Beziehung stehen. Stattdessen vergleichen wir die Restvarianz beider Modelle als Modellselektionskriterium (Abschnitt 6.10).

Modell	s^2
$y_{ij} = \mu + b_j + \beta_1 x_i + \beta_2 x_i^2 + e_{ij}$	81,1
$y_{ij} = \alpha + b_j + (\beta - \alpha) \exp(-\gamma x_i) + e_{ij}$	71,4

Das Mitscherlich-Modell ist das bessere wegen der kleineren Restvarianz.

Vergleich zum allgemeinen F-tests für den Vergleich geschachtelter Modelle

(Abschnitt 6.9): Im obigen Beispiel wurden alle F-Tests gegen das MQ_{Fehler} für das saturierte Modell, welches den Lack-of-fit Effekt umfaßt, durchgeführt. Das saturierte Modell schöpft alle Behandlungsfreiheitsgrade aus, so dass die Schätzung des Fehlers ausschließlich auf der Streuung innerhalb der Behandlung beruht. Diese Fehlerschätzung ist somit unabhängig vom gewählten nichtlinearen Regressionsmodell, was ein Vorteil ist. Wenn der Lack-of-fit Effekt nicht signifikant ist, wird in derselben Varianzanalyse-Tabelle eine Zeile weiter oben der nächste Effekt betrachtet, beim Polynom 2. Grades z.B. der quadratische Term, wobei der Fehlerterm des F-Tests unverändert bleibt.

Das hier vorgestellte Vorgehen stellt allerdings eine Modifikation des in Abschnitt 6.9 vorgestellten allgemeinen Verfahrens zum Vergleich zweier geschachtelter Modelle dar. Dieses Verfahren könnte im vorliegenden Fall so angewendet werden, dass beispielsweise bei nicht signifikantem F-Test für den Lack-of-fit das Modell um diesen Effekt reduziert wird und das MQ_{Fehler} für das nun volle Modell quadratische Modell ohne Lack-of-fit Effekt neu berechnet wird, wie in der folgenden Tabelle dargestellt.

Ursache	FG	SQ	MQ	F	p-Wert
$b_j \mu$	4	275,4	68,8	1,01	0,4238
$\beta_1 \mu, b_j$	1	49464,7	49464,7	728,62	<,0001
$\beta_2 \mu, b_j, \beta_1$	1	3912,3	3912,3	48,21	<,0001
Fehler (e_{ij})	23	1866,3	81,1		

Der Fehlerterm für den Test der quadratischen Terms im Modell hat sich durch die Modellreduktion geändert. Insbesondere ist bei diesem Vorgehen generell der Fehlerterm abhängig vom jeweils gerade betrachteten Modell für die Behandlungseffekte. Man hat also keinen vom Modell unabhängigen Schätzwert für die Fehlervarianz mehr. Es ist aber gerade ein großer Vorteil eines randomisierten Versuches mit Wiederholungen, dass immer ein von den Behandlungseffekten unabhängiger Schätzwert der Fehlervarianz möglich ist. Aus diesem Grunde wird in diesem Skript in allen Fällen, bei denen eine randomisierte Versuchsanlage mit Wiederholungen zugrunde liegt, ein einziger Fehlerterm verwendet. Sofern jedoch keine solche Versuchsanlage vorliegt, bleibt nur die Anwendung des in Abschnitt 6.9 vorgestellten F-Tests, bei dem der Fehlerterm immer abhängig ist vom gerade zu testenden Effekt abhängt.

Vergleich zur multiplen Regression (Abschnitt 6.10): Das hier vorgestellte Verfahren zur Wahl des Polynomgrades hat gewisse Ähnlichkeiten sowohl zur Vorwärts-Selektion und als auch zur Rückwärts-Elimination bei der multiplen Regression. Wir lesen eine sequentielle Varianzanalyse-Tabelle wie die hier verwendete immer von unten nach oben, bis wir auf einen signifikanten Effekt stoßen. So prüfen wir den Lack-of-fit Effekt immer zuerst. Ist dieser signifikant, können wir in der Varianzanalyse-Tabelle direkt weiter gehen zum höchsten angepaßten Polynomial-Term und schauen ob dieser signifikant ist. Oder aber wir eliminieren den Lack-of-fit Term und berechnen eine neue Varianzanalyse-Tabelle. Beides hat etwas von der Rückwärts-Elimination. Andererseits erhöhen wir den Grad des Polynoms solange, bis der Lack-of-fit Test nicht signifikant ist. Dies hat somit etwas von der Vorwärts-Selektion, wobei wir beim Polynom in einer festen Reihenfolge, aufsteigend vom Term niedrigsten Grades (Null) Polynomial-Terme aufsteigenden Grades hinzufügen (siehe Abschnitt 6.11).

Umsetzung in SAS

```
data;
input block n y;
z1=0; z2=0; z3=0; z4=0;
if block=1 then z1=1;
if block=2 then z2=1;
if block=3 then z3=1;
if block=4 then z4=1;
datalines;
1      0      89.49
2      0      54.56
3      0      74.33
4      0      78.20
```

5	0	61.51
1	25	108.78
2	25	102.01
3	25	105.04
4	25	105.23
5	25	106.52
1	50	136.28
2	50	129.51
3	50	132.54
4	50	132.73
5	50	134.02
1	100	157.63
2	100	167.39
3	100	155.39
4	100	146.85
5	100	155.81
1	150	185.96
2	150	176.66
3	150	178.53
4	150	195.34
5	150	185.56
1	200	195.09
2	200	190.43
3	200	183.52
4	200	180.99
5	200	205.69

```

;
symbol value=dot;
proc gplot;
plot y*n;
run;

proc glm;
class block;
model y=block n n*n;
run;

proc nlin;
parms alpha=200 beta=70 gamma=0.0136 b1=0 b2=0 b3=0 b4=0;
model y= alpha +(beta - alpha)*exp(-gamma*n)
      + b1*z1 + b2*z2 + b3*z3 + b4*z4;
run;

```

Um in NLIN für das Mitscherlich-Modell Blockeffekte anzupassen, muss eine Dummy-Kodierung vorgenommen werden, weil NLIN keine CLASS Anweisung hat wie GLM. Hierzu verwenden wir Dummy-Variablen z_1 , z_2 , z_3 und z_4 , die wie folgt definiert sind:

Block	z_1	z_2	z_3	z_4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	0	0	0	0

Das Modell lautet dann

$$y_{ij} = \alpha + (\beta - \alpha) \exp(-\gamma x_i) + z_{1ij} b_1 + z_{2ij} b_2 + z_{3ij} b_3 + z_{4ij} b_4 + e_{ij}$$

Hierbei haben wir die übliche Restriktion $b_5 = 0$ eingeführt. Die Dummy-Variablen entsprechen hier den Nullen und Einsen in der Designmatrix eines linearen Modells. Die Dummy-variablen "fischen" unter den Blockeffekten immer den jeweils zutreffenden Effekt heraus: Für Block 1 den Effekt b_1 , für Block 2 den Effekt b_2 , usw. Der 5. Blockeffekt ist gleich Null gesetzt und taucht daher nicht im Modell auf.

9. Zweistufige Stichproben

9.1 Modell und Auswertung

Beispiel: Bei Rübenblättern einer Pflanze wurde der Gehalt an Calcium (in % der Trockenmasse) bestimmt (Snedecor und Cochran, 1967, S. 281). Pro Blatt wurden 4 Analysen durchgeführt.

Blatt	% Calcium			
1	3,28	3,09	3,03	3,03
2	3,52	3,48	3,38	3,38
3	2,88	2,80	2,81	2,76
4	3,34	3,38	3,23	3,26

Ziel: Bestimmung des durchschnittlichen Calciumgehaltes des Blattapparates der Pflanze.

Modellierung: Wir können hier den einfachen Mittelwert der 16 Beobachtungen berechnen. Dies ist allerdings nur ein Schätzwert des tatsächlichen Calciumgehaltes des Blattapparates der Pflanze, weil wir nur eine Stichprobe von Blättern untersucht haben und weil die Messmethode einen Analysefehler hat, wie durch die Streuung der Messwiederholungen offenbar wird. Daher ist es sinnvoll, den Mittelwert mit einer Genauigkeitsangabe zu versehen, z.B. dem Standardfehler. Je größer der Standardfehler, desto ungenauer ist der Mittelwert. Der Standardfehler kann auch zur Berechnung eines Vertrauensintervalls herangezogen werden, welches mit vorgegebener Wahrscheinlichkeit den wahren Ca-Gehalt der Pflanze überdeckt. Auch für ein solches Intervall benötigen wir den Standardfehler.

Um den Standardfehler (Quadratwurzel aus der Varianz des Mittelwertes) zu berechnen, könnten wir die einfache Annahme machen, dass die 16 Messwerte eine einfache Zufallsstichprobe aus der Grundgesamtheit "Pflanze" darstellen und die Methoden aus Abschnitt 3.5 anwenden. Hierfür können wir das folgende lineare Modell aufstellen:

$$y_{ij} = \mu + e_{ij}, \quad i = 1, \dots, t; \quad j = 1, \dots, r$$

wobei

y_{ij} = Ca-Menge der j -ten Messwiederholung für das i -te Blatt

μ = theoretischer Mittelwert (Erwartungswert) der Pflanze; tatsächlicher Ca-Gehalt der gesamten Pflanze

e_{ij} = Zufallsabweichung der Beobachtung y_{ij} vom Erwartungswert

t = Zahl der Blätter

r = Zahl der Messwiederholungen je Blatt

Für die Zufallsabweichung e_{ij} machen wir folgende Verteilungsannahme:

$$e_{ij} \sim N(0, \sigma^2)$$

Nach diesem Modell hat der einfache Mittelwert

$$\hat{\mu} = \bar{y}_{..} = \frac{\sum_{i=1}^t \sum_{j=1}^r y_{ij}}{rt}$$

der ein Schätzer für μ ist, folgende Varianz:

$$\text{var}(\bar{y}_{..}) = \frac{\sigma^2}{rt}$$

Calzium-Gehalt

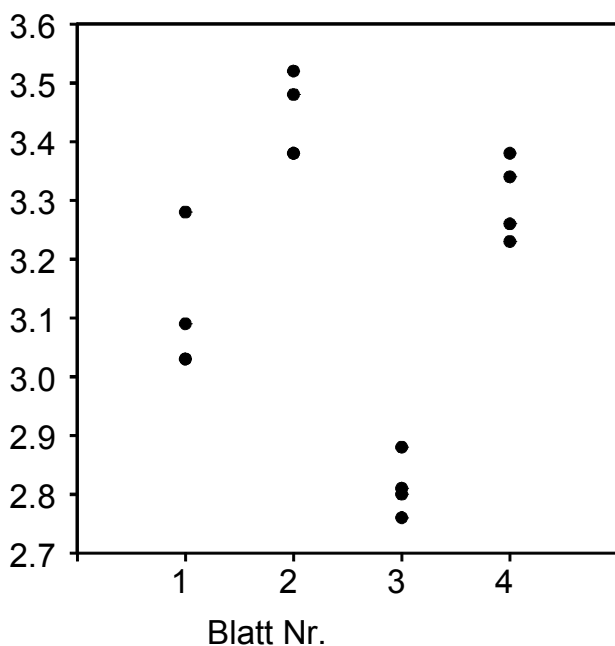


Abb. 9.1: Vier Messwiederholungen der Calziumgehalte von vier Blättern einer Pflanze.

Das Modell impliziert, dass alle 16 Messungen voneinander statistisch unabhängig sind, dass also die Korrelation bzw. die **Kovarianz** aller Beobachtungspaare Null ist. Diese Annahme ist aber nicht vernünftig, denn zwei Messwiederholungen, die von demselben Blatt stammen, sind sich tendenziell ähnlicher als zwei Messwerte von verschiedenen Blättern, wie Abb. 9.1 zeigt. In anderen Worten: zwei Messwerte vom selben Blatt sind positiv miteinander korreliert, die Kovarianz ist positiv. Die positive Kovarianz kann modellseitig einfach berücksichtigt werden, indem ein Blatteffekt aufgenommen wird:

$$y_{ij} = \mu + b_i + e_{ij} \quad , \quad i = 1, \dots, t; \quad j = 1, \dots, r$$

wobei

b_i = Effekt des i -ten Blatts ($i = 1, \dots, t$ = Zahl der Blätter)

e_{ij} = Messfehler der Beobachtung y_{ij}

Da eine Zufallsstichprobe von Blättern untersucht wurde, ist der Blatteffekt als zufällig zu betrachten:

$$b_i \sim N(0, \sigma_b^2)$$

Formal hat das lineare Modell dieselbe Form wie das einer einfachen Varianzanalyse (Kap. 4). Der Behandlungseffekt τ_i entspricht hier dem Blatteffekt b_i . Der Unterschied besteht darin, dass wir hier die Blatteffekte als zufällig und nicht als fest betrachten, weil eine Zufallsstichprobe von Blättern vorliegt. Auch sind wir nicht an Mittelwertunterschieden zwischen den Blättern interessiert. Vielmehr soll ja über die Blätter gemittelt werden. Da das Modell mehr als einen zufälligen Effekt hat, spricht man von einem **zufälligen Modell (random model)**.

Die Messwerte des Ca-Gehaltes sind durch den zufälligen Blatteffekt jetzt nicht mehr alle unabhängig. Zwar sind Messwerte von verschiedenen Blättern nach wie vor unabhängig, aber Messwerte vom selben Blatt sind korreliert bzw. sie weisen eine positive Kovarianz auf. Die Kovarianz zweier Messwerte vom i -ten Blatt ist gegeben durch:

$$\text{cov}(y_{ij}, y_{ij'}) = \sigma_b^2$$

Die Kovarianz entspricht hier der Varianz der Blatteffekte, weil die Messwerte y_{ij} und $y_{ij'}$ vom selben Blatt stammen und somit den zufälligen Blatteffekt b_i gemeinsam haben. Die Varianz einer Messung ist

$$\text{var}(y_{ij}) = \sigma_b^2 + \sigma^2$$

Die Korrelation zweier Messwerte vom selben Blatt ist

$$\text{corr}(y_{ij}, y_{ij'}) = \frac{\text{cov}(y_{ij}, y_{ij'})}{\sqrt{\text{var}(y_{ij}) \text{var}(y_{ij'})}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

Man spricht auch von der sog. **Intraclass-Korrelation**. Die positive Korrelation spiegelt die Tatsache wieder, dass sich Messwerte vom selben Blatt ähnlicher sind als Messwerte von verschiedenen Blättern. Nach dem um Blatteffekte erweiterten Modell hat der Mittelwert folgende Varianz:

$$\text{var}(\bar{y}_{..}) = \frac{\sigma_b^2}{t} + \frac{\sigma^2}{rt}$$

Die Varianz ist also um den Betrag σ_b^2/t größer als wenn wir Unabhängigkeit aller Beobachtungen annehmen würden! Der Grund ist wie folgt: Bei positiv korrelierten Messwerten bringt jeder zusätzliche Messwert (selbes Blatt) weniger zusätzliche Information als bei unabhängigen Messwerten (neues Blatt). In der extremen Situation einer perfekten Korrelation (Messfehlervarianz $\sigma^2 = 0$; $\text{corr}(y_{ij}, y_{ij'}) = 1$) liefert ein einziger Messwert von einem Blatt dieselbe Information wie 5, 10 oder

1000 Messwerte vom selben Blatt, denn die wiederholten Messwerte sind alle identisch. Wir können also noch so viele Messwiederholungen machen, dadurch würde sich die Genauigkeit des Mittelwertes kein bisschen erhöhen.

Auswertung:

Varianzanalyse-Tabelle:

Quelle	FG	MQ	E(MQ)
Pflanzen ("Behandlungen")	$t - 1$	$MQ_{Beh} = \frac{SQ_{Beh}}{t - 1}$	$\sigma^2 + r\sigma_b^2$
Fehler	$t(r - 1)$	$MQ_{Fehler} = \frac{SQ_{Fehler}}{t(r - 1)}$	σ^2

Quadratsummen (wie in einfacher Varianzanalyse; siehe Kap. 4):

$$SQ_{Beh} = r \sum_{i=1}^t (\bar{y}_{i\cdot} - \bar{y}_{..})^2 = \sum_{i=1}^t y_{i\cdot}^2 / r - y_{..}^2 / (rt);$$

$$SQ_{Fehler} = \sum_{i=1}^t \sum_{j=1}^r (y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{i=1}^t \sum_{j=1}^r y_{ij}^2 - \sum_{i=1}^t y_{i\cdot}^2 / r$$

wobei

Beobachtung: $y_{ij} = j\text{-ter Messwert der } i\text{-ten Beobachtungseinheit (Blatt)}$
 $(i = 1, \dots, t; j = 1, \dots, r)$

Summen der

Beobachtungseinheiten: $y_{i\cdot} = \sum_{j=1}^r y_{ij}$

Gesamtsumme: $y_{..} = \sum_{i=1}^t \sum_{j=1}^r y_{ij}$

Schätzung der Varianzkomponenten:

$$\hat{\sigma}^2 = MQ_{Fehler}$$

$$\hat{\sigma}_b^2 = \frac{MQ_{Beh} - MQ_{Fehler}}{r}$$

$$\wedge \quad var(\bar{y}_{..}) = \frac{\hat{\sigma}_b^2}{t} + \frac{\hat{\sigma}^2}{rt} = \frac{MQ_{Beh}}{rt}$$

$(1-\alpha)100\%$ Vertrauensintervall für μ :

$$\bar{y}_{..} \pm t_{Tab}(FG, \alpha) \sqrt{\hat{var}(\bar{y}_{..})}$$

wobei $t_{Tab}(FG, \alpha)$ der kritische Wert der t-Verteilung mit $FG = t-1$ Freiheitsgraden und Irrtumswahrscheinlichkeit α ist [siehe Tab. II (zweiseitig)].

Quelle	FG	MQ	E(MQ)
Blätter	3	0,296123	$\sigma^2 + 4\sigma_b^2$
Fehler	12	0,006602	σ^2

$$\hat{\sigma}^2 = 0,006602$$

$$\hat{\sigma}_b^2 = \frac{0,296123 - 0,006602}{4} = 0,07238$$

Vertrauensintervall:

$$\bar{y}_{..} = 3,1656$$

$$FG = t - 1 = 3$$

$$\alpha = 5\%$$

$$t_{Tab} = 3,182$$

$$3,1656 \pm 3,182 \sqrt{\frac{0,29612}{4 \cdot 4}}$$

$$\Rightarrow \text{Untere Grenze} = 2,7327, \text{ Obere Grenze} = 3,5986$$

Man beachte, dass hier die richtige Auswertung auch erhalten werden kann, indem Mittelwerte je Blatt berechnet werden ($\bar{y}_{i.}$) und diese dann wie im Fall einer einfachen Stichprobe ausgewertet werden (siehe Abschnitt 3.5: Vertrauensintervall für einen Mittelwert). Dies ist deshalb der Fall, weil die Mittelwerte $\bar{y}_{i.}$ untereinander statistisch unabhängig sind und somit als einfache Stichprobe aufgefasst werden können, und weil die Unterstichproben je Blatt gleich groß sind.

SAS Anweisungen:

```
data ruebe1;
input blatt messung calzium;
datalines;
1      1      3.28
1      2      3.09
1      3      3.03
1      4      3.03
2      1      3.52
2      2      3.48
```

2	3	3.38
2	4	3.38
3	1	2.88
3	2	2.80
3	3	2.81
3	4	2.76
4	1	3.34
4	2	3.38
4	3	3.23
4	4	3.26

```

;
proc mixed data=ruebe1;
class blatt;
model calzium=/solution cl;
random blatt;
run;

```

Output:

Covariance Parameter Estimates

Cov Parm	Estimate
blatt	0.07238
Residual	0.006602

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	Lower	Upper
Intercept	3.1656	0.1360	3	2.7327	3.5986

9.2 Optimale Allokation

Beispiel: Snedecor und Cochran (S. 529) beschreiben eine Studie zur Entwicklung eines optimalen Stichprobenplanes zur Ermittlung des Zuckergehaltes bei Zuckerrüben. Auf 100 Parzellen eines Blindversuches wurden je 10 Pflanzen analysiert. Um die Bedingungen eines Blockversuchs zu simulieren, wurde das MQ "zwischen den Parzellen" auf Basis der Streuung zwischen den Parzellen innerhalb von Blöcken à 5 Parzellen berechnet. Je Block gibt es hierbei 4 Freiheitsgrade für die Parzellen. Da 20 Blöcke vorliegen, beträgt die Zahl der Freiheitsgrade für Parzellen insgesamt 80. Die Varianzanalyse war wie folgt:

Ursache	Freiheitsgrade	MQ	E(MQ)
Zwischen den Parzellen	80	2,9254	$\sigma^2 + 10\sigma_b^2$
Innerhalb der Parzellen	900	2,1374	σ^2

Ziel: Bestimmung der optimalen Zahl von Pflanzen je Parzelle.

Modellierung: Da eine zweistufige Stichprobe vorliegt, kann analog zum vorangegangenen Beispiel folgendes Modell angesetzt werden:

$$y_{ij} = \mu + b_i + e_{ij}, \quad i = 1, \dots, t; \quad j = 1, \dots, r$$

wobei

y_{ij} = Zuckergehalt der j -ten Pflanze auf der i -ten Parzelle

μ = theoretischer Mittelwert (Erwartungswert); durchschnittlicher Zuckergehalt aller Pflanzen im Feld (Grundgesamtheit)

b_i = Effekt der i -ten Parzelle ($i = 1, \dots, t$)

e_{ij} = Fehler der Beobachtung y_{ij}

Verteilungsannahmen:

$$e_{ij} \sim N(0, \sigma^2)$$

$$b_i \sim N(0, \sigma_b^2)$$

Zur Ableitung eines optimalen Stichprobenplanes betrachten wir die Genauigkeit der Mittelwertschätzung (Varianz) sowie die Kosten der Untersuchung. Die Varianz des Mittelwertes ist

$$V = \text{var}(\bar{y}_{..}) = \frac{\sigma_b^2}{t} + \frac{\sigma^2}{rt}$$

wobei

t = Zahl der Parzellen

r = Zahl der Pflanzen je Parzelle

σ_b^2 = Varianz zwischen den Parzellen

σ^2 = Varianz der Pflanzen innerhalb der Parzellen

Die Gesamtkosten einer Erhebung sind

$$K = k_1 t + k_2 r t$$

wobei

k_1 = Kosten je Parzelle (ohne Pflanzen)

k_2 = Kosten je Pflanze

Bei der Planung des optimalen Stichprobenumfanges gibt es zwei operationalisierbare Ziele:

(1) Die Kosten K werden minimiert für gegebene Varianz V

(2) Die Varianz V wird minimiert für ein gegebenes Budget K

ad (1): Da die Varianz fest vorgegeben ist, gilt

$$t = \frac{1}{V} \left(\sigma_b^2 + \frac{\sigma^2}{r} \right)$$

und somit

$$K = t(k_1 + k_2 r) = \frac{1}{V} (k_1 + k_2 r) \left(\sigma_b^2 + \frac{\sigma^2}{r} \right)$$

ad (2): Da die Kosten fest vorgegeben sind, gilt

$$t = \frac{K}{k_1 + k_2 r}$$

und somit

$$V = \frac{1}{t} \left(\sigma_b^2 + \frac{\sigma^2}{r} \right) = \frac{1}{K} (k_1 + k_2 r) \left(\sigma_b^2 + \frac{\sigma^2}{r} \right)$$

In beiden Fällen ist also der Ausdruck

$$F = VK = (k_1 + k_2 r) \left(\sigma_b^2 + \frac{\sigma^2}{r} \right)$$

zu minimieren, wobei jeweils entweder K oder V fest vorgegeben ist. Setzt man die erste Ableitung von F nach r gleich Null und löst nach r auf, so erhält man

$$r = \sqrt{\frac{k_1 \sigma^2}{k_2 \sigma_b^2}}$$

Da die zweite Ableitung positiv ist, ist dies der Wert für r , der F minimiert. Dies ist die optimale Allokation bei Gültigkeit des angenommenen linearen Modells. Den Wert für t erhält man durch Einsetzen in die Gleichung für V (Varianz) bzw. K (Kosten), je nachdem welche der beiden Größen vorgegeben wurde.

Auswertung: Um diese Allokationsformel praktisch nutzen zu können, müssen Schätzwerte der Varianz eingesetzt werden. In unserem Beispiel ergibt die Varianzkomponentenschätzung mit der ANOVA Methode $\hat{\sigma}_b^2 = 0,0788$ und $\hat{\sigma}^2 = 2,1374$, so dass

$$r = \sqrt{\frac{2,1374}{0,0788}} \sqrt{\frac{k_1}{k_2}} \approx 5,2 \sqrt{\frac{k_1}{k_2}}$$

Die Kosten (k_1, k_2) sind hier nicht bekannt, so dass keine weiteren Berechnungen möglich sind. Bei etwa gleichen Kosten je Parzelle (k_1) und je Pflanze (k_2) ist eine Allokation von ca. $r = 5$ Pflanzen je Parzelle optimal.

10. Zweifaktorielle Varianzanalyse - Wechselwirkung

Bisher haben wir im wesentlichen Versuche und Erhebungen behandelt, bei denen lediglich ein Faktor variiert und in seiner Wirkung auf eine Zielvariable untersucht wurde. Die einzige Ausnahme war die multiple lineare Regression, bei der die Wirkung von zwei oder mehr quantitativen Einflussvariablen (Faktoren) auf eine Zielvariable untersucht wurde. Im folgenden wird der Fall betrachtet, bei dem zwei qualitative Einflussvariablen (Faktoren) untersucht werden. An diesem Beispiel wird das für faktorielle Versuche integrale Konzept der Wechselwirkung (Interaktion) mehrerer Faktoren erläutert. Wir gehen außerdem darauf ein, wie Wechselwirkungen bei der Auswertung und Interpretation zu berücksichtigen sind.

Beispiel: Es sollen sieben neue Weizensorten A, B, ..., G geprüft werden. Da die Leistungsfähigkeit der Sorten stark von der Düngerform abhängt, wird neben dem Faktor Sorte auch der Faktor Dünger untersucht. Dazu werden vier verschiedene N-Stufen N_1 , N_2 , N_3 und N_4 geprüft. Es ergeben sich somit 28 Faktorkombinationen = Behandlungen.

		Sorten						
		A	B	C	D	E	F	G
Düngerstufen	N_1							
	N_2							
	N_3							
	N_4							

□

Wenn die Zahl der Stufen je Faktor groß ist, kann die Zahl der zu prüfenden Behandlungen leicht sehr groß werden. Das Problem verschärft sich noch bei drei- und vierfaktoriellen Versuchen. Die Auswahl der Faktorstufen ist daher bei faktoriellen Versuchen besonders wichtig.

10.1 Wechselwirkungen (Interaktionen)

Ein Hauptgrund für die Durchführung von faktoriellen Versuchen ist die Untersuchung von Wechselwirkungen (Synonym: Interaktionen). Wechselwirkungen liegen dann vor, wenn die Differenzen der Zielvariable für die Stufen eines Faktors abhängen von den Stufen eines anderen Faktors.

Beispiel: In einem Fütterungsversuch werden zwei Futtermittel A und B an vier Rinderrassen geprüft. Es werden die täglichen Zunahmen als Zielvariable erfasst. Bei zwei Rinderrassen liefert Futter A die besseren täglichen Zunahmen, während bei den anderen beiden Futter B besser abschneidet. Die Differenz der täglichen Zunahmen zwischen den beiden Futtermitteln hängt hier von der Rinderrasse ab. Es liegen somit Wechselwirkungen vor. □

Beispiel: Die folgenden Abbildungen zeigen die Wirkung von 2 N-Stufen N0 und N1 auf den Ertrag zweier Weizensorten X und Y.

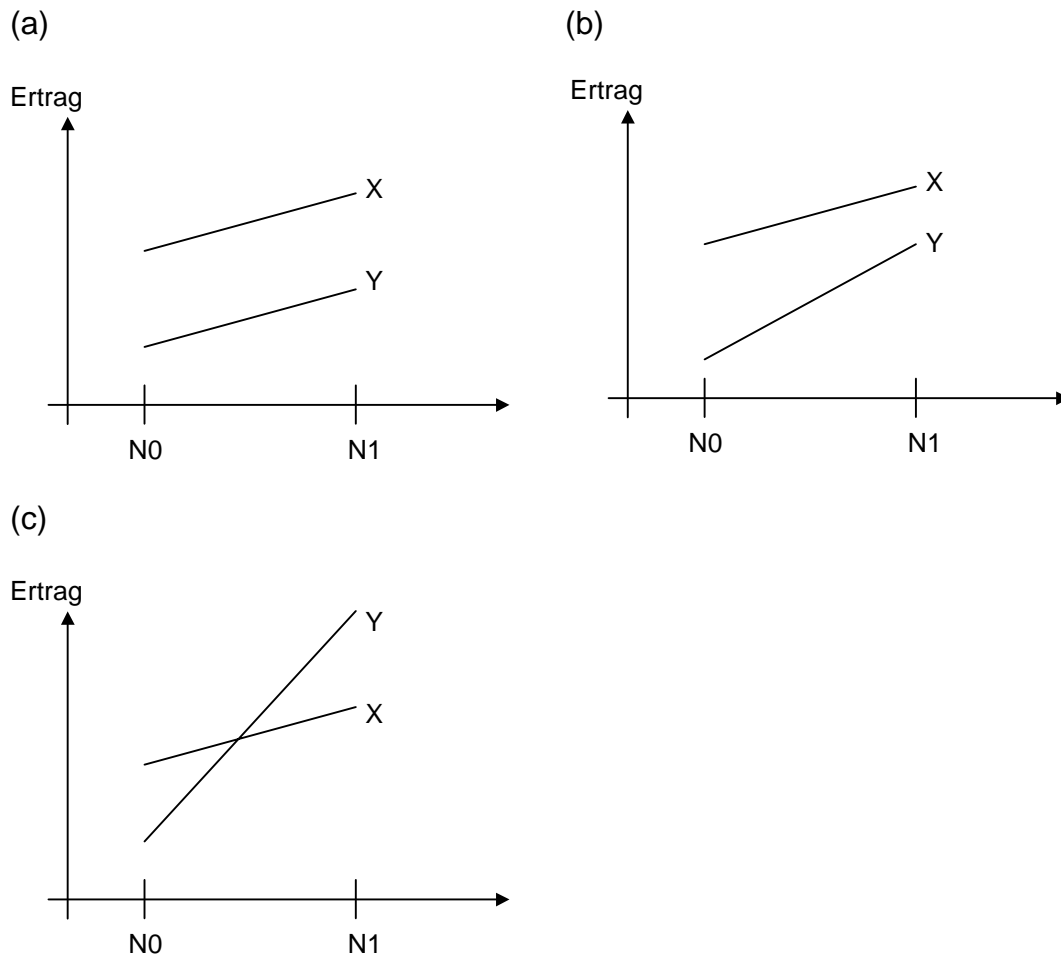


Abb. 10.1: Einige Beispiele für einen zweifaktoriellen Versuch mit zwei Sorten (X, Y) und zwei Düngerstufen (N0, N1).

In Abb. 10.1(a) verlaufen die Ertragslinien parallel. Der vertikale Abstand der Linien ist konstant. Die Ertragsdifferenz der Sorten ist unabhängig von der Stufe der N-Düngung. Analog ist der Ertragsunterschied der Dünger unabhängig von der Sorte. Man sagt, hier liegen keine Wechselwirkungen vor, weil die beiden Faktoren unabhängig voneinander also **additiv** wirken. In den Abbildungen 10.1(b) und 10.1(c) ist dagegen eine Wechselwirkung vorhanden. In Abb. 10.1(c) ist diese so stark, dass die Rangfolge der beiden Sorten sich zwischen den N-Stufen ändert. Man spricht hier auch von **Ranginteraktionen (Rangwechselwirkungen)**. Generell liegt immer dann eine Wechselwirkung vor, wenn das Reaktionsmuster von dem der Parallelität (bzw. Additivität) in 10.1(a) abweicht. □

10.2 Modellierung

Wir gehen im folgenden davon aus, dass der Versuch vollständig randomisiert wurde. In diesem Fall lautet das lineare Modell zur Auswertung eines zweifaktoriellen Versuches mit den Faktoren Sorte und Dünger:

$$y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\eta_{ij}} + e_{ijk}$$

wobei

y_{ijk} = Ertrag k -ten Wiederholung der i -ten Sorte bei der j -ten Düngermenge

μ = Gesamteffekt

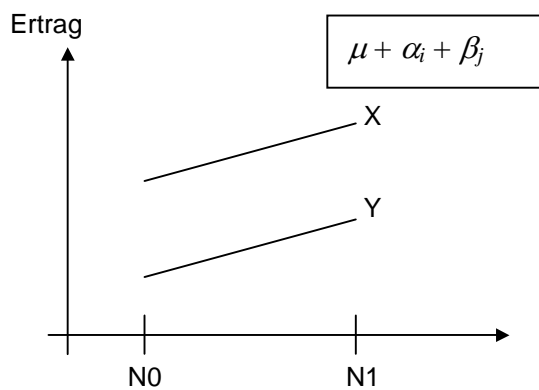
α_i = Haupteffekt der i -ten Sorte

β_j = Haupteffekt der j -ten Düngermenge

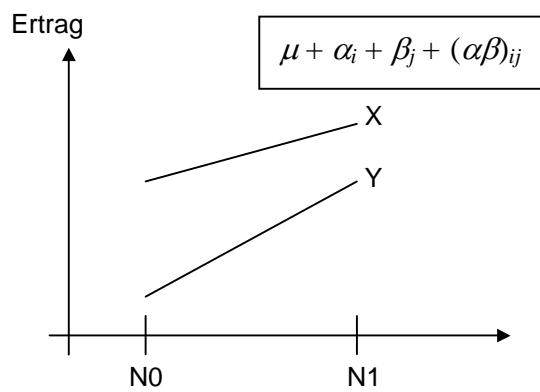
$(\alpha\beta)_{ij}$ = Wechselwirkung der i -ten Sorte und der j -ten Düngermenge

e_{ijk} = Fehler von y_{ijk}

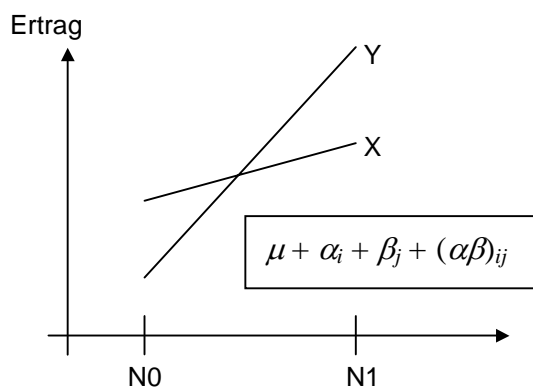
(a)



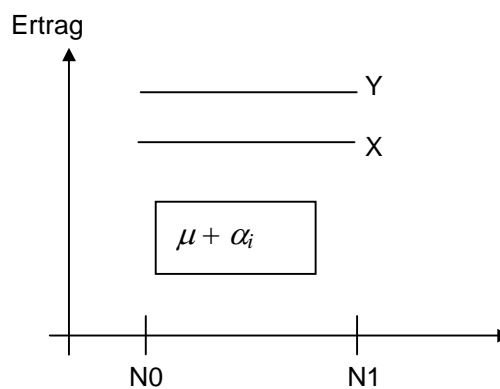
(b)



(c)



(d)



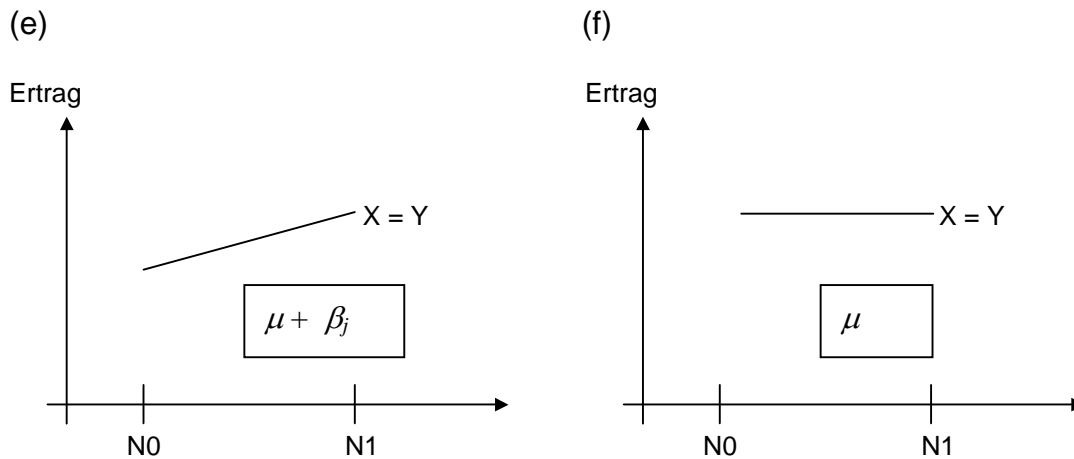
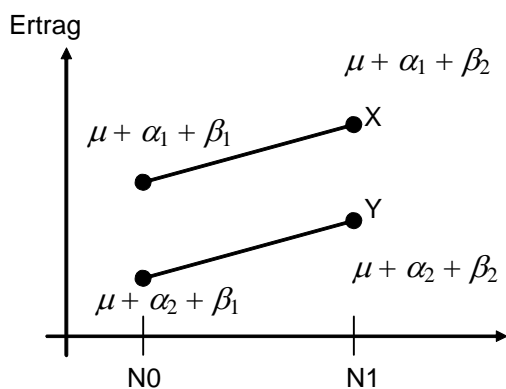


Abb. 10.2: Einige Beispiele für einen zweifaktoriellen Versuch mit zwei Sorten (X, Y) und zwei Düngerstufen (N0, N1) mit den dazugehörigen Modellen.

Je nach Reaktionsmuster der beteiligten beiden Faktoren werden nicht alle Effekte zur vollständigen Beschreibung der Ergebnisse benötigt. Wenn beispielsweise keine Wechselwirkungen vorliegen, kann der Wechselwirkungsterm wegfallen bzw. sind die Wechselwirkungseffekte alle gleich Null. In der Abb. 10.2 sind einige Beispiele für zweifaktorielle Versuche mit zwei Sorten und zwei Düngerstufen dargestellt, bei denen zum Teil Effekte gleich Null sind, weil das Reaktionsmuster einfach ist. Welcher der Effekte im zweifaktoriellen Modell benötigt wird, kann mit einer zweifaktoriellen Varianzanalyse festgestellt werden. Der Test der einzelnen Effekte im zweifaktoriellen Modell ist wichtig für die Modellwahl sowie die Wahl der sich anschließenden Mittelwertvergleiche. Wir betrachten im folgenden zunächst den balancierten Fall (Abschnitte 10.3 und 10.4). Im Anschluss wird der unbalancierte Fall behandelt (Abschnitte 10.5 und 10.6).

Es ist hilfreich, sich klar zu machen, dass das additive Modell für den Erwartungswert der i -ten Sorte und den j -ten Dünger, $\eta_{ij} = \mu + \alpha_i + \beta_j$, die Situation „Abwesenheit von Wechselwirkungen“ adäquat beschreibt. Hierzu betrachten wir wieder das Beispiel mit zwei Sorten und zwei Düngern, die im folgenden nochmals graphisch dargestellt ist.



Bei Abwesenheit von Wechselwirkungen erwarten wir, dass die Differenz der Erwartungswerte der beiden Dünger für beide Sorten gleich sind. Die Differenz der Erwartungswerte beider Dünger für die erste Sorte (X) ist

$$\eta_{11} - \eta_{12} = \mu + \alpha_1 + \beta_1 - (\mu + \alpha_1 + \beta_2) = \beta_1 - \beta_2.$$

Für die zweite Sorte (Y) ist diese Differenz

$$\eta_{21} - \eta_{22} = \mu + \alpha_2 + \beta_1 - (\mu + \alpha_2 + \beta_2) = \beta_1 - \beta_2.$$

Wir sehen, dass beide Differenzen gleich groß sind, wie gefordert. Analog können wir die Differenz der beiden Sorten getrennt für beide Dünger betrachten. Für den ersten Dünger (N0) ist die Differenz der beiden Sorten gleich

$$\eta_{11} - \eta_{21} = \mu + \alpha_1 + \beta_1 - (\mu + \alpha_2 + \beta_1) = \alpha_1 - \alpha_2,$$

und für den zweiten Dünger (N1) ist die Differenz

$$\eta_{12} - \eta_{22} = \mu + \alpha_1 + \beta_2 - (\mu + \alpha_2 + \beta_2) = \alpha_1 - \alpha_2.$$

Auch diese beiden Differenzen sind identisch. Dies zeigt, dass das additive Modell den Fall ohne Wechselwirkungen adäquat beschreibt.

10.3 Varianzanalyse bei balancierten Daten

Schema einer zweifaktoriellen Varianzanalyse für die Faktoren Sorte und Dünger im Fall balancierter Daten (gleiche Zahl von Wiederholungen für jede Behandlung).

Ursache/Effekt	FG	SQ	F
<hr/>			
Hauptwirkung			
Faktor A (α_i)	$a-1$	SQ_A	$F_{Vers}^{(A)} = \frac{SQ_A/(a-1)}{SQ_{Fehler}/[ab(r-1)]} = \frac{MQ_A}{MQ_{Fehler}}$
Hauptwirkung			
Faktor B (β_j)	$b-1$	SQ_B	$F_{Vers}^{(B)} = \frac{SQ_B/(b-1)}{SQ_{Fehler}/[ab(r-1)]} = \frac{MQ_B}{MQ_{Fehler}}$
Wechselwirkung			
A \times B [$(\alpha\beta_{ij})$]	$(a-1)(b-1)$	SQ_{AB}	$F_{Vers}^{(AB)} = \frac{SQ_{AB}/[(a-1)(b-1)]}{SQ_{Fehler}/[ab(r-1)]} = \frac{MQ_{AB}}{MQ_{Fehler}}$
Fehler	$ab(r-1)$	SQ_{Fehler}	

Die Rechenformeln für die SQ lauten:

$$SQ_A = \sum_{i=1}^a y_{i..}^2 / (br) - y_{...}^2 / n$$

$$SQ_B = \sum_{j=1}^b y_{.j.}^2 / (ar) - y_{...}^2 / n$$

$$SQ_{AB} = \sum_{i=1}^a \sum_{j=1}^b y_{ij\cdot}^2 / r - \sum_{i=1}^a y_{i\cdot\cdot}^2 / (br) - \sum_{j=1}^b y_{\cdot j\cdot}^2 / (ar) + y_{\cdot\cdot\cdot}^2 / n$$

$$SQ_{Fehler} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r y_{ijk}^2 - \sum_{i=1}^a \sum_{j=1}^b y_{ij\cdot}^2 / r$$

$$n = abr$$

a = Zahl der Stufen des Faktors A (z.B. Sorten)

b = Zahl der Stufen des Faktors B (z.B. Dünger)

r = Zahl der Wiederholungen je Faktorkombination

y_{ijk} = k -ter Messwert der i -ten Stufe des Faktors A und der j -ten Stufe des Faktors B

(1) Wechselwirkungen zuerst testen:

$H_0: (\alpha\beta)_{ij} = 0$ für alle $i, j \Rightarrow F_{Vers}^{(AB)}$ (Wechselwirkungen)

Falls signifikant, einfache Mittelwerte vergleichen.

(2) Hauptwirkungen nur testen, wenn Wechselwirkungen nicht signifikant:

$H_0: \alpha_i = 0$ für alle $i \Rightarrow F_{Vers}^{(A)}$ (Hauptwirkungen Faktor A)

Falls signifikant, bestehen Unterschiede der Randmittelwerte des Faktors A
 \Rightarrow Randmittelwerte A vergleichen.

$H_0: \beta_j = 0$ für alle $j \Rightarrow F_{Vers}^{(B)}$ (Hauptwirkungen Faktor B)

Falls signifikant, bestehen Unterschiede der Randmittelwerte des Faktor B.
 \Rightarrow Randmittelwerte B vergleichen.

Alle F-Tests werden mit F_{tab} aus Tab. VI mit den entsprechenden Freiheitsgraden für Zähler und Nenner von F_{Vers} durchgeführt. Wenn $F_{Vers} \geq F_{Tab}$, wird H_0 verworfen, andernfalls wird H_0 beibehalten.

Im obigen Kasten werden praktische Rechenformeln für die SQ angegeben. Für das Verständnis ist es hilfreich, die folgenden Formeln zu betrachten:

$$SQ_A = rb \sum_{i=1}^a (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$$

$$SQ_B = ra \sum_{j=1}^b (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$$

$$SQ_{AB} = r \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})^2$$

Die hierbei beteiligten Mittelwerte sind:

$\bar{y}_{ij\cdot}$ = Einfacher Mittelwert der i -ten Sorte (Faktor A) und j -ten Düngerstufe (Faktor B)

$\bar{y}_{i\cdot\cdot}$ = Randmittelwert der i -ten Sorte (gemittelt über Düngerstufe)

$\bar{y}_{\cdot j\cdot}$ = Randmittelwert der j -ten Düngerstufe (gemittelt über Sorten)

$\bar{y}_{\cdot\cdot\cdot}$ = Gesamtmittelwert

SQ_A : Wenn die Randmittelwerte des Faktors A (z.B. Sorte) sich stark unterscheiden, so sind die Abweichungen des Randmittelwerte $\bar{y}_{i\cdot\cdot}$ vom Gesamtmittel $\bar{y}_{\cdot\cdot\cdot}$ groß, und das SQ_A wird groß. Das SQ_A bzw. MQ_A ist daher ein Maß für die Größe der Unterschiede zwischen den Randmittelwerten des Faktors A.

SQ_B : Wenn die Randmittelwerte des Faktors B (z.B. Dünger) sich stark unterscheiden, so sind die Abweichungen des Randmittelwerte $\bar{y}_{\cdot j\cdot}$ vom Gesamtmittel $\bar{y}_{\cdot\cdot\cdot}$ groß, und das SQ_B wird groß. Das SQ_B bzw. MQ_B ist daher ein Maß für die Größe der Unterschiede zwischen den Randmittelwerten des Faktors B.

SQ_{AB} : Wechselwirkungen sind dann abwesend, wenn die Wirkungen des einen Faktors unabhängig sind von der Stufe des anderen Faktors. Betrachten wir zur Erläuterung die Wirkung des Düngers. Diese Wirkung können wir bei der i -ten Sorte z.B. durch folgende Differenz quantifizieren:

Wirkung Dünger j bei Sorte i : $\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot}$

Wenn nun keine Wechselwirkung vorliegt, dann muss diese Wirkung bei jeder Sorte gleich sein. Daher ist es sinnvoll, die sortenspezifischen Wirkungen zu mitteln:

Mittel der Wirkungen $\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot}$ über die Sorten = $\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot}$
= mittlere Wirkung des Dünger j .

Wenn nun keine Wechselwirkungen vorliegen, so ist die sortenspezifische Düngewirkung gleich der mittleren Düngewirkung, und es gilt

$$\begin{aligned}\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} &= \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot} \quad \text{bzw.} \\ \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot}) &= \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot} = 0\end{aligned}$$

Liegen dagegen Wechselwirkungen vor, so ist

$$\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot} \neq 0.$$

Offensichtlich ist also der Term $(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})$ ein Maß für die Größe der Wechselwirkung. Und genau dieser Term geht in quadrierter Form in das SQ_{AB} für die Wechselwirkungen ein. Je größer die Wechselwirkungen sind, um so stärker weichen die Terme $(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})$ von Null ab, und um so größer wird das SQ_{AB} .

Aufgrund des Ergebnisses der Varianzanalyse kann geschlossen werden, welcher der in Abb. 10.2 veranschaulichten Reaktionstypen vermutlich vorliegt. Dies wird in Tab. 10.1 erläutert.

Eine Bemerkung zu den Freiheitsgraden: Die Zahl der Behandlungen beträgt insgesamt $t = a \cdot b$. Somit sind die insgesamt zu vergebenden Behandlungs-FG gleich $(t-1) = (ab-1)$. Für die Hauptwirkung des Faktors A, der a Stufen hat, werden $(a-1)$ FG verbraucht. Analog ergeben sich für die Hauptwirkung des Faktors B $(b-1)$ FG. Die FG für die Wechselwirkung ergeben sich durch Differenzbildung als

$$(ab-1) - (a-1) - (b-1) = (a-1)(b-1)$$

Tab. 10.1: Erwartetes Ergebnis der Varianzanalyse in einem zweifaktorielle Versuch mit zwei Sorten und zwei Düngerstufen bei Reaktionsmustern wie in Abb. 10.2.

	Effekt		
	Wechselwirkung	Hauptwirkung Sorte	Hauptwirkung Dünger
Fall in Abb. 10.2	$(\alpha\beta)_{ij}$	α_i	β_j
(a)	nicht signifikant	signifikant	signifikant
(b)	signifikant	nicht testen	nicht testen
(c)	signifikant	nicht testen	nicht testen
(d)	nicht signifikant	signifikant	nicht signifikant
(e)	nicht signifikant	nicht signifikant	signifikant
(f)	nicht signifikant	nicht signifikant	nicht signifikant

Bemerkung: Die Rechenformeln für die SQ ergeben sich durch Anwendung der allgemeinen Resultate in den Abschnitten 6.8 und 6.9. Dies wird bei der Behandlung des Falls unbalancierter Daten (Abschnitt 10.5) näher erläutert.

10.4 Mittelwertvergleiche bei balancierten Daten

In einem zweifaktoriellen Versuch mit den Faktoren Sorte und Dünger können vier Arten von Mittelwertvergleichen durchgeführt werden:

Randmittelwerte:

- (1) Sortenmittelwerte, gemittelt über die Düngerstufen
- (2) Düngermittelwerte, gemittelt über die Sorten

Einfache Mittelwerte:

- (3) Sortenmittelwerte getrennt für jede Düngerstufe
- (4) Düngermittelwerte getrennt für jede Sorte

Die Vergleiche der einfachen Mittelwerte (3 und 4) sind immer dann sinnvoll, wenn die Wechselwirkungen signifikant sind, während die Vergleiche der Randmittelwerte

(1 und 2) nur sinnvoll sind, wenn die Wechselwirkungen nicht signifikant sind. Für Vergleich (1) muss zudem die Hauptwirkung der Sorten signifikant sein, während bei Vergleich (2) die Hauptwirkungen der Düngerstufen signifikant sein muss.

Bemerkung: Der F-Test der Hauptwirkungen des Faktors A (B) prüft die Gleichheit der Randmittelwerte für den Faktor A (B). Dies erklärt, warum ein Test der Hauptwirkungen nur sinnvoll ist, wenn keine Wechselwirkungen vorliegen.

Beispiel: Ein Versuch mit Wechselwirkungen habe folgendes Ergebnis (Mittelwerte über jeweils 3 Wiederholungen in t/ha; alle Mittelwerte seien signifikant verschieden) (Beachte: die Zahlen sind hypothetisch):

		Dünger		Sortenmittel über Düngerstufen (Randmittelwerte)
		1	2	
Sorte	1	1	2	1,5
	2	4	3	3,5
Düngermittel über Sorten (Randmittelwerte)		2,5	2,5	

einfache Mittelwerte

Die Düngermittelwerte über die Sorten (Randmittelwerte Dünger) zeigen hier keine Unterschiede. Hieraus kann allerdings nicht geschlossen werden, dass die Düngerauswahl keinen Effekt hat, weil starke Wechselwirkungen vorliegen. Das Gegenteil ist der Fall, wie eine Betrachtung der einfachen Mittelwerte zeigt: Bei Sorte 1 ist Dünger 2 um eine t/ha besser, bei Sorte 2 ist Dünger 1 um eine t/ha besser. Ähnliches gilt für die Sortenmittel über die Düngerstufen (Randmittelwerte der Sorten). Hier ist Sorte 1 um 2 t/ha unterlegen (1,5 t/ha gegenüber 3,5 t/ha). Allerdings gilt diese Differenz der Sorten für keine der beiden Düngerstufen: Bei Dünger 1 beträgt die Unterlegenheit 3 t/ha, bei Dünger 2 ist die Differenz nur 1 t/ha. Das Beispiel zeigt, dass bei Vorhandensein von Wechselwirkungen einfache Mittelwerte verglichen werden müssen, also hier Sortenmittel getrennt für jede Düngerstufe und Düngermittelwerte getrennt für jede Sorte.

Beispiel: Ein Versuch ohne Wechselwirkungen habe folgendes Ergebnis (Mittelwerte über jeweils 3 Wiederholungen in t/ha; alle Mittelwerte seien signifikant verschieden) (Beachte: die Zahlen sind hypothetisch):

		Dünger		Sortenmittel über Düngerstufen (Randmittelwerte)
		1	2	
Sorte	1	1	2	1,5
	2	3	4	3,5
Düngermittel über Sorten (Randmittelwerte)		2	3	

Die Düngermittelwerte über die Sorten (Randmittelwerte Dünger) zeigen eine Differenz von 1 t/ha zugunsten von Dünger 2. Diese Differenz gilt hier auch für jede der beiden Sorten, weil keine Wechselwirkungen vorliegen. Um die Differenz der Dünger zu prüfen, können daher die Randmittelwerte für die Dünger betrachtet werden. Diese sind genauer als die einfachen Mittelwerte, da sie sich aus mehr Einzelwerten zusammensetzen, nämlich aus 6 Einzelwerten im Vergleich zu 3 Einzelwerten für einfache Mittelwerte. Die Randmittelwerte für die Dünger sind deshalb interpretierbar, weil keine Wechselwirkungen vorliegen. Analoges gilt für die Randmittelwerte der Sorten. Wir brauchen aus den genannten Gründen in diesem Fall die einfachen Mittelwerte nicht statistisch zu vergleichen; es reicht die Betrachtung der Randmittelwerte.

Bemerkung: In den beiden hypothetischen Beispielen sind die einfachen Mittelwerte so gewählt, dass im Fall ohne Wechselwirkungen die Differenz der Sorten bei beiden Düngern exakt gleich ist. In der Praxis wird dieser Fall nie beobachtet werden, selbst wenn keine Wechselwirkungen vorliegen, da die Mittelwerte immer einer Zufallsschwankung unterliegen. Ob die „Differenz der Differenzen“ der einfachen Mittelwerte so groß ist, dass auf echte Wechselwirkungen geschlossen werden kann, kann ein F-Test der Wechselwirkungen klären. Welche Mittelwerte also zu vergleichen sind, hängt vom Ergebnis der Varianzanalyse ab.

In den in Abb. 10.2 dargestellten Fällen sind die in Tab. 10.2 angegebenen Vergleiche sinnvoll:

Tab. 10.2: Sinnvolle Mittelwertvergleiche in einem zweifaktoriellen Versuch mit zwei Sorten und zwei Düngerstufen bei Reaktionsmustern wie in Abb. 10.2.

Fall in Abb. 10.2	Sinnvolle Mittelwertvergleiche
(a)	Randmittelwerte Sorte und Dünger
(b)	Einfache Mittelwerte Sorte \times Dünger
(c)	Einfache Mittelwerte Sorte \times Dünger
(d)	Randmittelwerte Sorte
(e)	Randmittelwerte Dünger
(f)	Keine Vergleiche

Tab. 10.2(a): Mittelwertvergleiche bei balancierten Daten in Abhängigkeit vom Ergebnis der Varianzanalyse.

Test für Effekte ^{\$}			Selektiertes Modell (ohne Fehler)	Mittelwerte ^{&}	MEANS-Anweisung (SAS) [§]
$(\alpha\beta)_{ij}$	α_i	β_j			
*	kein Test	kein Test	$\eta_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	$\bar{y}_{ij\bullet}$	MEANS A*B;
ns	*	*	$\eta_{ij} = \mu + \alpha_i + \beta_j$	$\bar{y}_{i\bullet\bullet}$ und $\bar{y}_{\bullet j\bullet}$	MEANS A B/LSD;
ns	*	ns	$\eta_{ij} = \mu + \alpha_i$	$\bar{y}_{i\bullet\bullet}$	MEANS A/LSD;
ns	ns	*	$\eta_{ij} = \mu + \beta_j$	$\bar{y}_{\bullet j\bullet}$	MEANS B/LSD;

\$ ns: nicht signifikant; *: signifikant

§ PROC GLM; für A*B Mittelwerte kann GLM leider keine Grenzdifferenz berechnen.

& $\bar{y}_{ij\bullet}$ sind einfache Mittelwerte, $\bar{y}_{i\bullet\bullet}$ und $\bar{y}_{\bullet j\bullet}$ sind Randmittelwerte:

$$\bar{y}_{ij\bullet} = \frac{\sum_{k=1}^r y_{ijk}}{r}; \quad \bar{y}_{i\bullet\bullet} = \frac{\sum_{j=1}^b \bar{y}_{ij\bullet}}{b} = \frac{\sum_{j=1}^b \sum_{k=1}^r y_{ijk}}{br}; \quad \bar{y}_{\bullet j\bullet} = \frac{\sum_{i=1}^a \bar{y}_{ij\bullet}}{a} = \frac{\sum_{i=1}^a \sum_{k=1}^r y_{ijk}}{ar}$$

Wir können das Vorgehen bei der Auswertung eines zweifaktoriellen Versuches auch so betrachten, dass nach der Varianzanalyse eine Modellwahl erfolgt. In Abhängigkeit von der Modellwahl werden dann Mittelwerte berechnet. Das Vorgehen ist in Tab. 10.2(a) zusammengefasst. Die zu berechnenden Mittelwerte können im balancierten Fall mit der MEANS Anweisung der SAS Prozedur GLM erhalten werden. Die je nach Art der berechneten Mittelwerte zu verwendenden Grenzdifferenzen sind im folgenden Kasten wiedergegeben, wobei hier nur der t-Test berücksichtigt wird (vergleichsbezogene Irrtumswahrscheinlichkeit). Die Einhaltung der versuchsbezogenen Irrtumswahrscheinlichkeit in faktoriellen Versuchen ist ein schwieriges Problem, das hier nicht näher behandelt werden soll.

Berechnung von Grenzdifferenzen im zweifaktoriellen Versuch (vergleichsbezogene Irrtumswahrscheinlichkeit α):

Randmittelwerte A: A-Mittelwerte, gemittelt über die Stufen von B (nur wenn Wechselwirkungen nicht signifikant, Hauptwirkung A signifikant):

$$LSD = t(\alpha; FG_{Fehler}) \sqrt{\frac{2MQ_{Fehler}}{rb}}$$

Randmittelwerte B: B-Mittelwerte, gemittelt über die Stufen von A (nur wenn Wechselwirkungen nicht signifikant, Hauptwirkung B signifikant):

$$LSD = t(\alpha; FG_{Fehler}) \sqrt{\frac{2MQ_{Fehler}}{ra}}$$

Einfache Mittelwerte: A-Mittelwerte getrennt für jede B-Stufe oder B-Mittelwerte getrennt für jede A-Stufe (nur wenn Wechselwirkungen signifikant):

$$LSD = t(\alpha; FG_{Fehler}) \sqrt{\frac{2MQ_{Fehler}}{r}}$$

a = Zahl der Stufen des Faktors A (z.B. Sorten)

b = Zahl der Stufen des Faktors B (z.B. Dünger)

r = Zahl der Wiederholungen je Faktorkombination

Beispiel: Ein Versuch mit zwei Sorten, zwei Düngerstufen und fünf Wiederholungen lieferte folgende Ergebnisse.

Varianzanalyse-Tabelle:

Ursache	FG	SQ	MQ	F _{Vers}	§p-Wert
Sorte	1	3,2401	3,2401	16,95	0,0008
Dünger	1	2,4151	2,4151	12,63	0,0026
Sorte × Dünger	1	0,2311	0,2311	1,21	0,2879
Fehler	16	3,0594	0,1912		

§ siehe Anhang D

Sorte	Dünger	Einfacher Mittelwert ($\bar{y}_{ij\cdot}$)
1	1	2,33
1	2	2,81
2	1	2,92
2	2	3,83

Da die Wechselwirkungen nicht signifikant sind, schauen wir die Tests der Hauptwirkungen an. Beide sind signifikant, so dass wir sowohl die Sortenmittelwerte über die Düngerstufen als auch die Düngermittelwerte über die Sorten prüfen (Randmittelwerte).

Berechnung der Randmittelwerte aus den einfachen Mittelwerten:

		Dünger		Sortenmittel über Düngerstufen
		1	2	
Sorte	1	2,33	2,81	$(2,33+2,81)/2 = 2,57 = \bar{y}_{1\cdot\cdot}$
	2	2,92	3,83	$(2,92+3,83)/2 = 3,38 = \bar{y}_{2\cdot\cdot}$
Düngermittel über Sorten		2,63	3,32	$\bar{y}_{\cdot j\cdot}$

$\bar{y}_{\cdot 1\cdot} = (2,33+2,92)/2$ $(2,81+3,83)/2 = \bar{y}_{\cdot 2\cdot}$

$$t(\alpha = 5\%; FG_{Fehler} = 16) = 2,120$$

$$MQ_{Fehler} = 0,1912$$

$$a = b = 2, r = 5$$

$$LSD(\text{Sorten}): LSD = t(\alpha; FG_{Fehler}) \sqrt{\frac{2MQ_{Fehler}}{rb}} = 2,120 \sqrt{\frac{2 \times 0,1912}{5 \times 2}} = 0,4146$$

$$LSD(\text{Dünger}): LSD = t(\alpha; FG_{Fehler}) \sqrt{\frac{2MQ_{Fehler}}{ra}} = 2,120 \sqrt{\frac{2 \times 0,1912}{5 \times 2}} = 0,4146$$

Tab. 10.3: Vergleich der Sortenmittel über die Dünger.

Sorte	§Mittelwert
1	2,57 ^a
2	3,38 ^b
LSD	0,41

§Mittelwerte, die mit demselben Buchstaben versehen sind, sind nicht signifikant verschieden

Tab. 10.4: Vergleich der Düngermittel über die Sorten.

Dünger	§Mittelwert
1	2,63 ^a
2	3,32 ^b
LSD	0,41

§Mittelwerte, die mit demselben Buchstaben versehen sind, sind nicht signifikant verschieden

Beispiel (Köhler et al., 1984): In einem Gewächshausversuch wurde der Einfluss von zwei Faktoren auf den Ertrag einer Weinsorte untersucht. Faktor A: Düngeform (D1, D2); Faktor B: Pflanzenschutzbehandlung (P1, P2, P3).

		Pflanzenschutz		
		P1	P2	P3
Düngung	D1	21,3	22,3	23,8
		20,9	21,6	23,7
		20,4	21,0	22,6
	Mittelwert:	20,9	21,6	23,4
	D2	12,7	12,0	14,5
		14,9	14,2	16,7
		12,9	12,1	14,5
	Mittelwert:	13,5	12,8	15,7



●: P1; ○: P2; □: P3

Die obige graphische Abbildung der Behandlungsmittelwerte zeigt, dass die Düngerwirkung für die drei Pflanzenschutz-Behandlungen etwa gleich ist, was auf Abwesenheit von Wechselwirkungen schließen lässt. Genauen Aufschluss liefert allerdings erst die Varianzanalyse. Man beachte, dass die Abszisse dieser Darstellung keine quantitative Skala hat. Die Verbindung der Datenpunkte mit einer Linie dient nur zur Verdeutlichung, welche Punkte zur selben Pflanzenschutz-Stufe gehören.

Varianzanalyse-Tabelle:

Ursache	FG	SQ	F	§p-Wert
Dünger	1	296,87	312,49	0,0001
Pflanzenschutz	2	17,78	9,36	0,0036
Wechselwirkung	2	1,79	0,89	0,4371
Fehler	12	11,41		

§ siehe Anhang D

Die Hauptwirkungen sind signifikant (Dünger: $F_{Vers} = 312,49 > F_{Tab} = 4,75$; Pflanzenschutz: $F_{Vers} = 9,36 > F_{Tab} = 3,89$), während die Wechselwirkungen nicht signifikant sind ($F_{Vers} = 0,89 < F_{Tab} = 3,89$).

Da die Wechselwirkungen nicht signifikant sind, können wir die Randmittelwerte, also die Mittelwerte der Düngebehandlungen über die Pflanzenschutzbehandlungen sowie die Mittelwerte der Pflanzenschutzbehandlungen über die Düngerstufen vergleichen. Wir finden folgende Randmittelwerte und Grenzdifferenzen (t-Test; LSD):

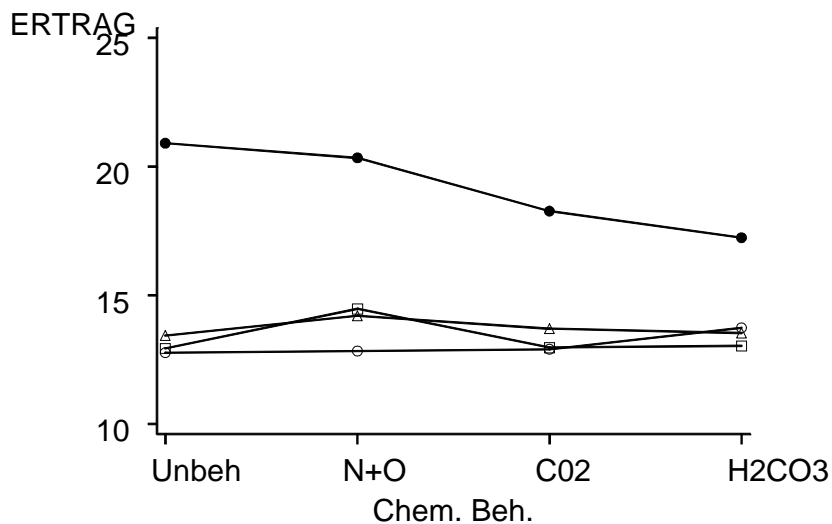
Pflanzenschutz- behandlung	Mittelwert	Dünger	Mittelwert
P1	17,18 ^b	D1	21,96 ^a
P2	17,20 ^b	D2	13,83 ^b
P3	19,30 ^a		
LSD($\alpha=5\%$)	1,23	LSD($\alpha=5\%$)	1,00

(Mittelwerte, die mit demselben Buchstaben versehen sind, sind nicht signifikant voneinander verschieden)

Die Pflanzenschutzbehandlungen P1 und P2 unterscheiden sich nicht signifikant, sind aber jeweils signifikant schlechter als die Behandlung P3. Diese Aussage gilt unabhängig von der Düngerbehandlung, da keine Wechselwirkungen vorliegen. Außerdem ist der Dünger D1 signifikant besser als D2, unabhängig von der Pflanzenschutzbehandlung. □

Beispiel (Weber, 1986): In einem zweifaktoriellen Gewächshausversuch wurde die Wirkung verschiedener Düngerformen (Faktor A) sowie verschiedener chemischer Behandlungen (Faktor B) auf das Wachstum von Weizenpflanzen untersucht.

Chemische Behandlung (Faktor B)				
Dünger (Faktor A)	Unbehandelt	N + O	CO ₂ -Gas	H ₂ CO ₃
Unbehandelt	21,4	20,9	19,6	17,6
	21,2	20,3	18,8	16,6
	20,1	19,8	16,4	17,5
	Mittelwert: 20,9	20,3	18,3	17,2
Strohdüngung	12,0	13,6	13,0	13,3
	14,2	13,3	13,7	14,0
	12,1	11,6	12,0	13,9
	Mittelwert: 12,8	12,8	12,9	13,7
Stroh- und Phosphatdüngung	13,5	14,0	12,9	12,4
	11,9	15,6	12,9	13,7
	13,4	13,8	13,1	13,0
	Mittelwert: 12,9	14,4	13,0	13,0
Stroh-, Phosphat- und Kalkdüngung	12,8	14,1	14,2	12,0
	13,8	13,2	13,6	14,6
	13,7	15,3	13,3	14,0
	Mittelwert: 13,4	14,2	13,7	13,5



•: Unbehandelt; □: Stroh; Δ: Stroh + P; ○: Stroh + P + Kalk.

Die graphische Darstellung der Mittelwerte zeigt, dass wahrscheinlich Wechselwirkungen vorliegen. In den ungedüngten Varianten hat die chemische Behandlung (Faktor B) einen deutlichen Einfluss auf den Ertrag, während bei den gedüngten Varianten kaum Unterschiede in den chemischen Behandlungsvarianten (Faktor B) bestehen. Allerdings können wir nicht sagen, ob die Unterschiede statistisch gesichert sind. Hierzu führen wir eine Varianzanalyse durch:

Ursache	<i>FG</i>	<i>SQ</i>	<i>F_{Vers}</i>	p-Wert
Dünger	3	306,24	102,08	0,0001
Chem. Beh.	3	9,17	3,06	0,0211
Wechselwirkung	9	25,46	2,83	0,0045
Fehler	32	26,28		

Es liegen signifikante Wechselwirkungen vor. Daher ignorieren wir die Tests der Hauptwirkungen und vergleichen die Mittelwerte der Düngung getrennt für jede chemische Behandlung und umgekehrt.

Tab. 10.5: Vergleich der Düngerbehandlungen getrennt für jede Stufe der chem. Behandlung (Mittelwerte in einer Spalte, die mit demselben Buchstaben versehen sind, sind nicht signifikant verschieden)

Dünger (Faktor A)	Chemische Behandlung (Faktor B)			
	Unbehandelt	N + O	CO ₂ -Gas	H ₂ CO ₃
Unbehandelt	20,9 ^a	20,3 ^a	18,3 ^a	17,2 ^a
Strohdüngung	12,8 ^b	12,8 ^c	12,9 ^b	13,7 ^b
Stroh- und Phosphatdüngung	12,9 ^b	14,4 ^b	13,0 ^b	13,0 ^b
Stroh-, Phosphat- und Kalkdüngung	13,4 ^b	14,2 ^{bc}	13,7 ^b	13,5 ^b

LSD($\alpha=5\%$) = 1,51

Tab. 10.6: Vergleich der chem. Behandlungen getrennt für jede Stufe der Düngerbehandlung (Mittelwerte in einer Zeile, die mit demselben Buchstaben versehen sind, sind nicht signifikant verschieden)

Dünger (Faktor A)	Chemische Behandlung (Faktor B)			
	Unbehandelt	N + O	CO ₂ -Gas	H ₂ CO ₃
Unbehandelt	20,9 ^a	20,3 ^a	18,3 ^b	17,2 ^b
Strohdüngung	12,8 ^a	12,8 ^a	12,9 ^a	13,7 ^a
Stroh- und Phosphatdüngung	12,9 ^a	14,4 ^a	13,0 ^a	13,0 ^a
Stroh-, Phosphat- und Kalkdüngung	13,4 ^a	14,2 ^a	13,7 ^a	13,5 ^a

LSD($\alpha=5\%$) = 1,51

Wir haben hier beide möglichen Tabellen dargestellt. In einer Veröffentlichung wird in der Regel eine Tabelle reichen. Wenn die Vergleiche in den Zeilen und in den Spalten in einer Tabelle dargestellt werden sollen, kann man auch verschiedene Symbole für die Zeilen- und die Spaltenvergleiche verwenden, z.B. kleine und große Buchstaben, obwohl die Darstellung dann etwas unübersichtlich werden kann. □

10.5 Varianzanalyse bei unbalancierten Daten

Wenn für einen balanciert geplanten Versuch (jede Behandlung hat dieselbe Zahl von Beobachtungen) Beobachtungen ausfallen und die Daten somit unbalanciert werden, ist die Auswertung etwas komplizierter als bei balancierten Daten. Wir setzen hier voraus, dass **für jede Faktorkombination mindestens eine Beobachtung** vorliegt.

Zur Erstellung einer Varianzanalyse benutzen wir das allgemeine Prinzip des sequentiellen Modellaufbaus. Hierzu betrachten wir die folgende Sequenz von Modellen:

Modell	SQ_{Fehler}
(0) $y_{ijk} = \mu + e_{ijk}$	$SQ_{Fehler}^{(0)}$
(1) $y_{ijk} = \mu + \alpha_i + e_{ijk}$	$SQ_{Fehler}^{(1)}$
(2) $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$	$SQ_{Fehler}^{(2)}$
(3) $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$	$SQ_{Fehler}^{(3)}$

Dies führt zu der folgenden Varianzanalyse-Tabelle:

Ursache	FG	$SQ(\text{Parameter})$
Faktor A	$a-1$	$SQ(\alpha_i / \mu) = SQ_{Fehler}^{(0)} - SQ_{Fehler}^{(1)}$
Faktor B, bereinigt um A	$b-1$	$SQ(\beta_j / \alpha_i, \mu) = SQ_{Fehler}^{(1)} - SQ_{Fehler}^{(2)}$
Wechselwirkungen	$(a-1)(b-1)$	$SQ[(\alpha\beta)_{ij} / \alpha_i, \beta_j, \mu] = SQ_{Fehler}^{(2)} - SQ_{Fehler}^{(3)}$
Fehler	$n - ab$	$SQ_{Fehler}^{(3)}$

a = Zahl der Stufen des Faktors A

b = Zahl der Stufen des Faktors B

n = Gesamtzahl der Beobachtungen

Die sich ergebenden MQ aller drei Behandlungseffekte werden gegen

$MQ_{Fehler} = SQ_{Fehler}^{(3)} / (n - ab)$ geprüft:

Hauptwirkung A (unbereinigt): $F_{vers} = \frac{SQ(\alpha_i / \mu) / (a-1)}{SQ_{Fehler}^{(3)} / (n - ab)}$ (falscher Test!)

Hauptwirkung B (bereinigt um A):
$$F_{\text{vers}} = \frac{SQ(\beta_j | \alpha_i, \mu) / (b-1)}{SQ_{\text{Fehler}}^{(3)} / (n-ab)}$$

Wechselwirkung:
$$F_{\text{vers}} = \frac{SQ[(\alpha\beta)_{ij} | \alpha_i, \beta_j, \mu] / [(a-1)(b-1)]}{SQ_{\text{Fehler}}^{(3)} / (n-ab)}$$

Wie im balancierten Fall testen wir zunächst die Wechselwirkungen. Sind diese signifikant, schließen wir, dass die Mittelwerte eines Faktors getrennt für die Stufen des jeweils anderen Faktors verglichen werden müssen. Die Tests der Hauptwirkungen haben bei Vorhandensein von Wechselwirkung dann keine weitere Bedeutung.

Falls die Wechselwirkungen nicht signifikant sind, können wir Hauptwirkungen interpretieren und betrachten daher die Tests der beiden Hauptwirkungen für Faktor A und Faktor B. Für unbalancierte Daten spielt die Reihenfolge der Anpassung der Effekte eine Rolle. Um den Faktor B zu testen, müssen wir zuvor den Faktor A angepasst haben. Nur dann ist der Test für den Faktor B um den Effekt des Faktors A bereinigt. Dies ist bei der obigen Modellsequenz der Fall. Andernfalls kann der Test für B davon beeinflusst werden, wie groß die Effekte von A sind. Das ergibt sich auch aus dem allgemeinen Vorgehen zum Test einer Nullhypothese im linearen Modell (Abschnitt 6.9) welches erfordert, ein volles und ein dazu reduziertes Modell zu spezifizieren. Im Fall des Tests für B (β_j) ist das volle Modell $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$ und das reduzierte $y_{ijk} = \mu + \alpha_i + e_{ijk}$. In Abweichung vom in Abschnitt 6.9 beschriebenen Verfahren wird das MQ der F-Statistik allerdings nicht auf Basis der SQ_{Fehler} des vollen Modells $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$, sondern auf Basis des gesättigten Modells mit Interaktionen berechnet. Dies hat vor allem den Vorteil, dass alle Tests im Rahmen einer Varianzanalyse-Tabelle aufgeführt werden können. Näheres zur zweifaktoriellen Varianzanalyse bei unbalancierten Daten findet sich bei S. R. Searle (1987: Linear models for unbalanced data. Wiley, New York).

In der obigen Sequenz haben wir A vor B angepasst, so dass wir den "richtigen" Test für B erhalten. Dagegen liefert die Sequenz nicht den richtigen Test für A. Dieser unbereinigte F-Test für A, der von Computerprogrammen automatisch mit ausgegeben wird, sollte daher ignoriert werden!

Um den richtigen Test für A zu erhalten, müssen wir B vor A anpassen, also folgende Sequenz betrachten:

Modell	SQ_{Fehler}
(0) $y_{ijk} = \mu + e_{ijk}$	$SQ_{\text{Fehler}}^{(0)}$
(1) $y_{ijk} = \mu + \beta_j + e_{ijk}$	$SQ_{\text{Fehler}}^{(1*)}$
(2) $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$	$SQ_{\text{Fehler}}^{(2)}$
(3) $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$	$SQ_{\text{Fehler}}^{(3)}$

Dies führt zu der folgenden Varianzanalyse-Tabelle:

Ursache	FG	$SQ(\text{Parameter})$
Faktor B	$b-1$	$SQ(\beta_j \mu) = SQ_{Fehler}^{(0)} - SQ_{Fehler}^{(1*)}$
Faktor A, bereinigt um B	$a-1$	$SQ(\alpha_i \beta_j, \mu) = SQ_{Fehler}^{(1*)} - SQ_{Fehler}^{(2)}$
Wechselwirkungen	$(a-1)(b-1)$	$SQ[(\alpha\beta)_{ij} \alpha_i, \beta_j, \mu] = SQ_{Fehler}^{(2)} - SQ_{Fehler}^{(3)}$
Fehler	$n - ab$	$SQ_{Fehler}^{(3)}$

Die sich ergebenden MQ aller drei Behandlungseffekte werden wiederum gegen $MQ_{Fehler} = SQ_{Fehler}^{(3)} / (n - ab)$ geprüft:

Hauptwirkung B (unbereinigt):
$$F_{vers} = \frac{SQ(\beta_j | \mu) / (b-1)}{SQ_{Fehler}^{(3)} / (n - ab)} \quad (\text{falscher Test!})$$

Hauptwirkung A (bereinigt um B):
$$F_{vers} = \frac{SQ(\alpha_i | \beta_j, \mu) / (a-1)}{SQ_{Fehler}^{(3)} / (n - ab)}$$

Wechselwirkung:
$$F_{vers} = \frac{SQ[(\alpha\beta)_{ij} | \alpha_i, \beta_j, \mu] / [(a-1)(b-1)]}{SQ_{Fehler}^{(3)} / (n - ab)}$$

Im Fall balancierter Daten liefern beide Sequenzen dieselbe Quadratsummenzerlegung, da dann die Faktoren A und B orthogonal sind. In diesem Fall ergeben sich die in Abschnitt 10.3 angegebenen einfachen Rechenformeln für die SQ . Es gilt:

$$\begin{aligned} SQ_A &= SQ(\alpha_i | \mu) = SQ(\alpha_i | \beta_j, \mu) \quad , \\ SQ_B &= SQ(\beta_j | \mu) = SQ(\beta_j | \alpha_i, \mu) \quad \text{und} \\ SQ_{AB} &= SQ[(\alpha\beta)_{ij} | \alpha_i, \beta_j, \mu] \end{aligned}$$

Die hier vorgestellte sequentielle Form der Varianzanalyse kann mit SAS mittels der Typ I Quadratsummen erhalten werden. Es gibt außerdem sog. Typ III Quadratsummen. Diese liefern bei unbalancierten Daten andere Quadratsummen als Typ I. Typ III Quadratsummen sind vor allem dann relevant, wenn man Hauptwirkungen trotz signifikanter Wechselwirkungen testen will. Dieser Fall ist aus meiner Sicht aber praktisch irrelevant, weshalb auf Typ III Quadratsummen hier nicht eingegangen wird. Details finden sich bei Dufner, Jensen, Schumacher: Statistik mit SAS. Teubner, Stuttgart.

10.6 Mittelwertvergleiche bei unbalancierten Daten

Bei unbalancierten Daten gibt es keine einfachen Formeln für Mittelwerte. Anstelle einfacher arithmetischer Mittelwerte müssen adjustierte Mittelwerte berechnet werden, was mit den allgemeinen Methoden aus Abschnitt 6.8 möglich ist (vgl. Abschnitte 8.7 und 8.8).

Basierend auf der Varianzanalyse wird zunächst das adäquate lineare Modell gewählt. Basierend auf diesem Modell werden sodann **adjustierte** Mittelwerte

berechnet. Die in Tab. 10.7 mittels der Modelleffekte definierten Mittelwerte werden geschätzt, indem die Kleinst-Quadrat-Lösungen für die Parameter eingesetzt werden die mit den Methoden aus Abschnitt 6.8 erhalten werden.

Beispiel (Köhler et al., 1984): In einem Gewächshausversuch wurde der Einfluss von zwei Faktoren auf den Ertrag einer Weinsorte untersucht. Faktor A: Düngeform (D1, D2); Faktor B: Pflanzenschutzbehandlung (P1, P2, P3). Dieses Beispiel hatten wir bereits bei der Auswertung balancierter Daten verwendet. Wir nehmen nun an, dass zwei Beobachtungen ausgefallen sind.

		Pflanzenschutz		
		P1	P2	P3
Düngung	D1	21,3	22,3	23,8
		20,9	21,6	*
		20,4	21,0	22,6
	D2	12,7	12,0	14,5
		14,9	14,2	16,7
		*	12,1	14,5

Wir verwenden die folgende Anweisung (Y = Ertrag, D = Dünger, P = Pflanzenschutz) in der SAS Prozedur GLM:

```
proc glm;
class d p;
model y= d p d*p;
run;
```

Hiermit finden wir folgende Varianzanalyse (im Standard-Output fehlt die letzte Zeile = Fehler):

Source	DF	Type I SS	Mean Square	F Value	Pr > F
d	1	242.5806250	242.5806250	226.57	<.0001
p	2	13.6037756	6.8018878	6.35	0.0166
d*p	2	2.1683077	1.0841538	1.01	0.3977

Tab. 10.7: Mittelwertvergleiche bei unbalancierten Daten in Abhängigkeit vom Ergebnis der Varianzanalyse.

Test für Effekte ^{\$}			Selektiertes Modell (ohne Fehler)	Mittelwerte ^{&}	LSMEANS-Anweisung (SAS) [§]
$(\alpha\beta)_{ij}$	α_i	β_j			
*	kein Test	kein Test	$\eta_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	$\hat{\eta}_{ij}$	LSMEANS A*B/PDIFF;
ns	*	*	$\eta_{ij} = \mu + \alpha_i + \beta_j$	$\hat{\eta}_{i\cdot}$ mit $\bar{\eta}_{i\cdot} = \mu + \alpha_i + \bar{\beta}_{\cdot}$ $\hat{\eta}_{\cdot j}$ mit $\bar{\eta}_{\cdot j} = \mu + \beta_j + \bar{\alpha}_{\cdot}$	LSMEANS A B/PDIFF;
ns	*	ns	$\eta_{ij} = \mu + \alpha_i$	$\hat{\eta}_{i\cdot}$ mit $\bar{\eta}_{i\cdot} = \mu + \alpha_i$	LSMEANS A/PDIFF;
ns	ns	*	$\eta_{ij} = \mu + \beta_j$	$\hat{\eta}_{\cdot j}$ mit $\bar{\eta}_{\cdot j} = \mu + \beta_j$	LSMEANS B/PDIFF;

\$ ns: nicht signifikant; *: signifikant

§ PROC GLM oder PROC MIXED

& Generell ist $\bar{\eta}_{i\cdot} = \frac{\sum_{j=1}^b \eta_{ij}}{b}$ und $\bar{\eta}_{\cdot j} = \frac{\sum_{i=1}^a \eta_{ij}}{a}$, wobei für η_{ij} das jeweils selektierte Modell einzusetzen ist. $\hat{\eta}_{ij}$ sind einfache adjustierte Mittelwerte, $\hat{\eta}_{i\cdot}$ und $\hat{\eta}_{\cdot j}$ sind adjustierte Randmittelwerte. Im balancierten Fall gilt: $\hat{\eta}_{ij} = \bar{y}_{ij\cdot}$, $\hat{\eta}_{i\cdot} = \bar{y}_{i\cdot\cdot}$ und $\hat{\eta}_{\cdot j} = \bar{y}_{\cdot j\cdot}$. [vgl. Tab. 10.2(a)]

Da die Wechselwirkungen nicht signifikant sind, prüfen wir nun die Hauptwirkungen. In obiger Analyse wird P nach D angepasst, weil P in der Modellanweisung nach D steht. Daher liefert diese Analyse den "richtigen" Test für die Hauptwirkung der Pflanzenschutzmaßnahmen (P). Diese Hauptwirkung ist signifikant. Für die Hauptwirkung der Düngung (D) passen wir D nach P an:

```
proc glm;
class d p;
model y= p d d*p;
run;
```

Dies liefert folgende Analyse (auch hier fehlt die letzte Zeile = Fehler):

Source	DF	Type I SS	Mean Square	F Value	Pr > F
p	2	4.3393750	2.1696875	2.03	0.1825
d	1	251.8450256	251.8450256	235.22	<.0001
d*p	2	2.1683077	1.0841538	1.01	0.3977

Auch die Hauptwirkung für D ist signifikant. Man beachte, dass der "falsche" Test für P hier nicht signifikant war, während der "richtige" Test mit der vorhergehenden Analyse signifikant war.

Da die Wechselwirkung nicht signifikant, beide Hauptwirkungen aber signifikant sind, wählen wir das folgende Modell für die Berechnung von Mittelwerten:

$$\eta_{ij} = \mu + \alpha_i + \beta_j$$

wobei α_i der Haupteffekt der Düngerbehandlung und β_j der Haupteffekt der Pflanzenschutzbehandlung ist. Basierend auf diesem Modell berechnen wir die **adjustierten Randmittelwerte** der beiden Faktoren A (Dünger) und B (Pflanzenschutz) ($\bar{\eta}_{i\cdot}$ und $\bar{\eta}_{\cdot j}$) sowie deren paarweise Vergleiche. Bei Gültigkeit des Modells $\eta_{ij} = \mu + \alpha_i + \beta_j$ haben die Mittelwerte folgende Form:

Adjustierte Randmittelwerte für Düngerbehandlung:

$$\bar{\eta}_{i\cdot} = \frac{\sum_{j=1}^b \eta_{ij}}{b} = \frac{\sum_{j=1}^b (\mu + \alpha_i + \beta_j)}{b} = \mu + \alpha_i + \bar{\beta}_{\cdot}$$

mit

$$\bar{\beta}_{\cdot} = \frac{\sum_{j=1}^b \beta_j}{b}$$

Adjustierte Randmittelwerte für Pflanzenschutzbehandlung:

$$\bar{\eta}_{\bullet j} = \frac{\sum_{i=1}^a n_{ij}}{a} = \frac{\sum_{i=1}^a (\mu + \alpha_i + \beta_j)}{a} = \mu + \beta_j + \bar{\alpha}_{\bullet}$$

mit

$$\bar{\alpha}_{\bullet} = \frac{\sum_{i=1}^a \alpha_i}{a}$$

Die **adjustierten Randmittelwerte** werden durch Einsetzen der Kleinst-Quadrat-Lösungen für die Effekte des Modells (μ, α_i, β_j) geschätzt. Standardfehler der adjustierten Randmittelwerte und Standardfehler von Mittelwertdifferenzen werden mit den Methoden in Abschnitt 6.8 berechnet. Zur Auswertung benutzen wir wegen der komfortableren Ausgabe der Ergebnisse die MIXED Prozedur (und nicht GLM):

```
proc mixed;
class d p;
model y= p d;
lsmeans p d/pdiff; run;
```

Ergebnis:

Least Squares Means

Effect	d	p	Estimate	Standard Error	DF	t Value	Pr > t
p		1	17.2364	0.4662	12	36.97	<.0001
p		2	17.2000	0.4229	12	40.67	<.0001
p		3	19.2236	0.4662	12	41.24	<.0001
d	1		21.9046	0.3692	12	59.33	<.0001
d	2		13.8687	0.3692	12	37.56	<.0001

Differences of Least Squares Means

Effect	d	p	_d	_p	Estimate	Standard Error	DF	t Value	Pr > t
p		1		2	0.03641	0.6294	12	0.06	0.9548
p		1		3	-1.9872	0.6635	12	-3.00	0.0112
p		2		3	-2.0236	0.6294	12	-3.22	0.0074
d	1		2		8.0359	0.5245	12	15.32	<.0001

Pflanzenschutz- Behandlung (Faktor B)	Adjustierter Randmittelwert ($\hat{\eta}_{j\cdot}$)	Dünger (Faktor A)	Adjustierter Randmittelwert ($\hat{\eta}_{i\cdot}$)
P1	17,24 ^b	D1	21,90 ^a
P2	17,20 ^b	D2	13,87 ^b
P3	19,22 ^a		

(Mittelwerte in einer Spalte, die mit demselben Buchstaben versehen sind, sind nicht signifikant voneinander verschieden)

Die Randmittelwerte für D2 und P1 berechnen sich aus den Kleinst-Quadrat-Schätzungen der Parameter

$$\hat{\mu} = 15,206 \quad (\text{Gesamteffekt})$$

$$\hat{\alpha}_1 = 8,036; \hat{\alpha}_2 = 0 \quad (\text{Dünger-Haupteffekte})$$

$$\hat{\beta}_1 = -1,987; \hat{\beta}_2 = -2,024; \hat{\beta}_3 = 0 \quad (\text{Pflanzenschutz-Haupteffekte})$$

wie folgt:

$$\hat{\eta}_{2\cdot} = \frac{\sum_{j=1}^b \hat{\eta}_{ij}}{b} = \hat{\mu} + \hat{\alpha}_2 + \hat{\beta}_{\cdot} = 15,206 + 0 + \frac{-1,987 - 2,024 + 0}{3} = 13,87 \quad (\text{D2})$$

$$\hat{\eta}_{\cdot 1} = \frac{\sum_{i=1}^a \hat{\eta}_{ij}}{a} = \hat{\mu} + \hat{\beta}_1 + \hat{\alpha}_{\cdot} = 15,206 - 1,987 + \frac{8,036 + 0}{2} = 17,24 \quad (\text{P1})$$

Die anderen Randmittelwerte berechnen sich analog. Das Ergebnis ist sehr ähnlich dem Resultat bei balancierten Daten. Man beachte, dass im allgemeinen keine gemeinsame Grenzdifferenz berechnet werden kann, da der Standardfehler einer Differenz nicht konstant ist. Für die Vergleiche P1 - P2 und P2 - P3 ist der Standardfehler 0,6294, für P1 - P3 ist der Standardfehler 0,6635. Aber man kann meistens, wie auch im vorliegenden Fall, eine Buchstabendarstellung erhalten. Allerdings kann es wegen der Unbalanciertheit passieren, dass eine solche Darstellung nicht mit der in Abschnitt 4.4 besprochenen Methode möglich ist. Für diesen Fall steht eine allgemeinere Methode zur Verfügung, die in einem SAS Makro implementiert ist (<http://www.uni-hohenheim.de/bioinformatik/>).

Warum muss der Test für α um β bereinigt sein und umgekehrt?

Nachdem abschließend die Mittelwerte betrachtet wurden, ist es instruktiv, noch einmal auf die Varianzanalyse zurückzukommen, und zwar auf die Frage der Reihenfolge der Anpassung der Effekte.

Beispiel: Ein Versuch ohne Wechselwirkungen habe folgendes Ergebnis (Einzelwerte in t/ha; alle Mittelwerte seien signifikant verschieden) (Beachte: die Zahlen sind hypothetisch, der Einfachheit wird Abwesenheit von Fehlern angenommen).

	Dünger		Sortenmittel über Düngerstufen (naive Randmittelwerte)
	1	2	
Sorte 1	1	2 2 2 2 2 2 2 2 2	1,9
Einfache Mittelwerte:	1	2	
Sorte 2	3 3 3 3 3 3 3 3 3	4	3,1
Einfache Mittelwerte:	3	4	
Düngermittel über Sorten (naive Randmittelwerte)	2,8	2,2	

In diesem Beispiel gilt ein additives Modell ohne Wechselwirkungen, also

$$\eta_{ij} = \mu + \alpha_i + \beta_j$$

wobei α_i Haupteffekt der Sorte und β_j Haupteffekt des Düngers. Da die hypothetischen Daten frei von Versuchsfehlern sind, ist anhand der Einzelwerte bzw. der einfachen Mittelwerte offensichtlich, welche Behandlungswirkungen vorliegen. Sorte 2 ist um 2 t/ha besser als Sorte 1, und zwar unabhängig von der Düngervariante. Umgekehrt ist Dünger 2 um 1 t/ha besser als Dünger 1, und zwar unabhängig von der Sorte. Ein anderes Bild ergibt sich bei den "naiven" Randmittelwerten, die sich durch einfache Mittelung über alle Beobachtungen einer Zeile bzw. einer Spalte ergeben. Hier ist Sorte 2 nur um 1,2 t/ha besser als Sorte 1. Bei den Düngern ist der erste um 0,6 t/ha besser als der zweite. Hier verkehrt sich also sogar das Bild. Die Betrachtung zeigt, dass der Vergleich der naiven Randmittelwerte irreführend ist, wenn die Daten unbalanciert sind. Der Grund für diese Eigenschaft der naiven Randmittelwerte ist die ungleiche Gewichtung der vier Zellen/einfachen Mittelwerte der Kreuztabelle. Wir können die Randmittelwerte als

gewichtete Mittel der einfachen Mittelwerte betrachten, wobei der Stichprobenumfang als Gewicht eingeht. Der naive Randmittelwert für Sorte 1 ist

$$\frac{1+2+2+2+2+2+2+2+2+2}{10} = \frac{1+9 \cdot 2}{10} = 1,9$$

Der einfache Mittelwert der Sorte 1 bei Dünger 2 bekommt hier das neunfache Gewicht im Vergleich zum einfachen Mittelwert bei Dünger 1, weil bei Dünger 1 der Stichprobenumfang gleich 9 ist und bei Dünger 1 gleich 1. Das Randmittel für Sorte 1 wird also von Dünger 2 dominiert. Bei Sorte 2 ist es umgekehrt: Hier dominiert Dünger 1:

$$\frac{3+3+3+3+3+3+3+3+3+4}{10} = \frac{9 \cdot 3 + 4}{10} = 3,1$$

Wegen dieser ungleichen Gewichtung liefern die naiven Randmittelwerte ein verzerrtes Bild. "Gerechte" Randmittelwerte ergeben sich, indem das ungewichtete Mittel der jeweils relevanten Zellmittelwerte berechnet wird:

	Dünger		Sortenmittel über Düngerstufen (adjustierte Randmittelwerte)
	1	2	
Sorte 1	1	2 2 2 2 2 2 2 2 2 2	
Einfache Mittelwerte:	1	2	$(1+2)/2 = 1,5$
Sorte 2	3 3 3 3 3 3 3 3 3 3	4	
Einfache Mittelwerte:	3	4	$(3+4)/2 = 3,5$
Düngermittel über Sorten (adjustierte Randmittelwerte)	$(1+3)/2=2$	$(2+4)/2=3$	

[Diese einfache Analyse setzt genau genommen ein Modell mit Interaktionen voraus. Die Analyse nach dem Modell ohne Interaktionen ist etwas aufwendiger herzuleiten, weswegen dies hier nicht betrachtet wird. Wegen der Abwesenheit von Fehlern in den Daten und der Besetzungszahlen der Zellen sind die Ergebnisse beider Analysen hier jedoch identisch.]

Man erhält diese Mittelwerte im obigen Fall auch, wenn die Parameter des additiven Modells mit Hilfe der Methode der kleinsten Quadrate geschätzt werden (mit dem Computer):

$$\hat{\mu} = 4 \quad (\text{Gesamteffekt})$$

$$\hat{\alpha}_1 = -2; \hat{\alpha}_2 = 0 \quad (\text{Sorten-Haupteffekte})$$

$$\hat{\beta}_1 = -1; \hat{\beta}_2 = 0 \quad (\text{Dünger-Haupteffekte})$$

Einsetzen in die Gleichungen $\bar{\eta}_{i\cdot} = \mu + \alpha_i + \bar{\beta}_{\cdot}$ und $\bar{\eta}_{\cdot j} = \mu + \beta_j + \bar{\alpha}_{\cdot}$ liefert die Kleinst-Quadrat-Randmittelwerte.]

Betrachten wir nun eine Modellsequenz, bei der zuerst α_i und dann β_j angepasst wird:

$$(0) y_{ijk} = \mu + e_{ijk}$$

$$(1) y_{ijk} = \mu + \alpha_i + e_{ijk}$$

$$(2) y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

Das Modell $y_{ijk} = \mu + \alpha_i + e_{ijk}$ ist das einer einfachen Varianzanalyse für den Faktor A (Sorte). In anderen Worten, **das $SQ(\alpha_i|\mu)$ erfasst Unterschiede der beiden naiven Randmittelwerte für die Sorten!** Und diese ist offensichtlich nicht der richtige Vergleich für die beiden Sorten. Denn die naiven Randmittelwerte berücksichtigen nicht die ungleiche Zellbesetzung, die ja eine Verzerrung der Differenz der naiven Randmittelwerte gegenüber der Differenz der einfachen Mittelwerte verursacht. Daher ergibt sich der falsche F-Test für die Hauptwirkung A, wenn diese als erster Effekt angepasst wird. Dagegen ergibt sich der richtige Fest für B, weil der Effekt β_j nach α_i angepasst wird. Dies folgt aus dem generellen Ansatz zum Test einer Nullhypothese im linearen Modell (siehe Abschnitt 6.9).

10.7 Zweifaktorielle Versuche in anderen Versuchsanlagen (Blockanlage etc.)

Bisher haben wir angenommen, dass der zweifaktorielle Versuch in einer vollständig randomisierten Anlage durchgeführt wurde. Alternativ kommen andere Anlagen in Frage. Zunächst kann jede der für einfaktorielle Versuche besprochenen Anlagen verwendet werden, also insbesondere auch die **randomisierte vollständige Blockanlage**. Bei der Randomisation werden einfach die $t = a*b$ Faktorkombinationen innerhalb der Blöcke vollständig randomisiert. Die Freiheitsgrade für Fehler und Blöcke in der Varianzanalyse ergeben sich dann völlig analog zum einfaktoriellen Fall mit der Ersetzung $t = a*b$. Somit sind

$$FG_{\text{Fehler}} = n - ab - r + 1 \quad \text{und} \quad FG_{\text{Blöcke}} = r - 1$$

Die FG für die Haupteffekte und Interaktionen der beiden Behandlungsfaktoren sind unverändert, also $(a-1)$, $(b-1)$ und $(a-1)(b-1)$. In der Modellsequenz passen wir wie immer Blockeffekte vor Behandlungseffekten an. Alles oben bezüglich der Testung von Haupteffekten und Interaktionen sowie der Schätzung von Mittelwerten gesagte gilt auch bei der Blockanlage sowie bei jeder anderen Versuchsanlage. Das lineare Modell lautet:

$$y_{ijk} = \mu + b_k + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

wobei

- y_{ijk} = Ertrag k -ten Wiederholung der i -ten Sorte bei der j -ten Düngermenge
- μ = Gesamteffekt
- b_k = Effekt des k -ten Blocks
- α_i = Haupteffekt der i -ten Sorte
- β_j = Haupteffekt der j -ten Düngermenge
- $(\alpha\beta)_{ij}$ = Wechselwirkung der i -ten Sorte und der j -ten Düngermenge
- e_{ijk} = Fehler von y_{ijk}

Schema einer zweifaktoriellen Varianzanalyse für die Faktoren Sorte und Dünger im Fall balancierter Daten (gleiche Zahl von Wiederholungen für jede Behandlung), wenn der Versuch als **Blockanlage** mit r Blöcken angelegt wurde.

Ursache/Effekt	Freiheitsgrade (FG)
Blöcke (b_k)	$r-1$
Hauptwirkung Faktor A (α_i)	$a-1$
Hauptwirkung Faktor B (β_j)	$b-1$
Wechselwirkung A \times B [$(\alpha\beta)_{ij}$]	$(a-1)(b-1)$
Fehler	$(ab-1)(r-1)$

Neben den Anlagen, die auch für einfaktorielle Versuche in Frage kommen, gibt es verschiedene spezielle Versuchsanlagen für zweifaktorielle Versuche, bei denen die Randomisation für die beiden Faktoren getrennt erfolgt. Die bekanntesten sind die **Spaltenlage** und die **Streifenanlage**. Diese speziellen Anlagen werden an anderer Stelle behandelt.

11. Elementare nichtparametrische Verfahren

Die statistische Auswertung auf der Basis von linearen Modellen der Form $y = X\beta + e$ macht eine Reihe von Annahmen, die Voraussetzung für die Gültigkeit von Verfahren der schließenden Statistik (Tests, Vertrauensintervalle) sind, nämlich:

- Linearität/Additivität
- Unabhängigkeit der Fehler e
- Varianzhomogenität der Fehler e
- Normalverteilung der Fehler e

Bei Verletzung mindestens einer dieser Voraussetzungen verlieren Verfahren wie die Varianzanalyse, der t-Test oder die Regressionsanalyse strenggenommen ihre Gültigkeit. Wegen der Art der Modellvoraussetzungen werden diese Verfahren auch als **parametrische Verfahren** bezeichnet. Insbesondere die Annahme einer Normalverteilung ist eine "parametrische" Annahme in dem Sinne, dass diese Verteilung durch Parameter (Erwartungswert $X\beta$, Varianz σ^2) charakterisiert ist.

Bei Verletzung der Voraussetzungen stellt sich die Frage, welche alternativen Auswertungsstrategien zur Verfügung stehen. Wir haben in Kapitel 7 gesehen, dass eine Option in der Datentransformation liegt. Hierbei werden für die transformierten Daten dann die selben starken Annahmen getroffen wie sonst für die untransformierten Daten. Eine Alternative besteht in der Verwendung sog. **nichtparametrischer** oder **verteilungsfreier Verfahren**. Die Bezeichnung rührt daher, dass diese Verfahren schwächere Annahmen machen als parametrische Verfahren. Insbesondere setzen sie keine Normalverteilung voraus. Viele dieser Verfahren beruhen auf einer Umwandlung der Originaldaten in **Ränge**. Die jeweilige Teststatistik wird dann aus den Rängen berechnet. Ein solches Verfahren haben wir bereits kennengelernt, die sog. Rangkorrelation (Abschnitt 6.5: Korrelation bei nichtlinearen Zusammenhängen). In diesem Kapitel werden einige weitere elementare nichtparametrische Verfahren vorgestellt. Eine Übersicht findet sich in Tab. 11.1.

Literaturhinweise

- Bortz J., Lienert, G.A. 1998 Kurzgefasste Statistik für die klinische Forschung. Springer, Berlin.
- Bortz, J., Lienert, G.A., Boehnke, K. 1990 Verteilungsfreie Methoden in der Biostatistik. Springer, Berlin.
- Brunner, E., und Langer, F., 1999 Nichtparametrische Analyse longitudinaler Daten. Oldenbourg-Verlag, München.
- Brunner, E., Munzel, U. 2002 Nichtparametrische Datenanalyse. Springer, Berlin.
- Köhler, W., Schachtel, G., Voleske, P. 2001 Biostatistik. 3. Auflage. Springer, Berlin

Tab. 11.1: Einige nichtparametrische Verfahren und ihre parametrischen Analoga.

Design/Stichprobe/Problemstellung	Nichtparametrisch	Parametrisch
2 Gruppen, unverbundene Stichprobe	1. Median-Test 2. Mann-Whitney-Test	Unverbundener t-Test
>2 Gruppen, unverbundene Stichprobe (vollständig randomisierte Anlage)	Kruskal-Wallis-Test (H-Test)	Einfache Varianzanalyse
2 Gruppen, verbundene Stichprobe	1. Vorzeichen-Test 2. Wilcoxon-Test	Verbundener t-Test
> 2 Gruppen, verbundene Stichprobe (Blockanlage)	Friedman-Test	Varianzanalyse für Blockanlage
Korrelation	Spearman'sche Rangkorrelation	Pearson'sche Produkt-Moment Korrelation

11.1 Vergleich zweier unabhängiger Stichproben

Beispiel (Brunner und Langer, 1999, S. 136): In einer randomisierten, kontrollierten Studie (gesunde Nichtraucher, 21-30 Jahre alt) wurden nach einem Aderlaß von 750 ml Blut 1000 ml einer physiologischen Elektrolytlösung zusammen mit vier Testsubstanzen gegeben:

Propanolol: $n_1 = 10$ Probanden
Dobutamin: $n_2 = 13$ Probanden
Fenoterol: $n_3 = 13$ Probanden
Placebo: $n_4 = 13$ Probanden

In der Propanol-Gruppe erkrankten drei der ursprünglich 13 Probanden und konnten am Versuch nicht teilnehmen. Die Plasma-Renin Aktivität (PRA) [ng/ml/h] wurde zu 5 Zeitpunkten (0, 2, 6, 8 und 12 h) nach Aderlaß bestimmt. Es sollte u.a. untersucht werden, inwiefern die PRA durch die Testsubstanzen gesteigert bzw. reduziert würde.

Für jeden Patienten kann man die PRA-Werte gegen die Zeit plotten und die Punkte durch Geraden verbinden. Ein integrales Maß für die PRA ist die Fläche unter der sich ergebenden Kurve (Polygonzug) (area under the curve - AUC; siehe Kap. 13). Die Fläche unter der PRA-Kurve wird als äquivalent für die Menge des innerhalb von 12 Stunden endogen freigesetzten Angiotensin I angesehen. Somit lautet die einfache Fragestellung hier, ob sich die Gruppen in den AUC Werten unterscheiden.

Hier werden zunächst nur die AUC-Werte der zweiten und dritten Gruppe verglichen.

Dobutamin	Fenoterol
18,80	34,16
41,70	67,82
41,24	60,68
49,98	39,04
63,90	37,72
29,86	46,78
25,92	50,80
36,92	128,60
54,90	31,18
37,04	35,04
28,94	56,40
33,80	22,62
18,08	26,09

Es sollen die Mittelwerte der beiden Gruppen verglichen werden.

11.1.1 Median-Test

Median-Test

Fragestellung: Es sollen zwei unverbundene Gruppen verglichen werden. Die Nullhypothese H_0 fordert, dass beide Gruppen denselben Median haben.

Voraussetzung: Die Daten sind mindestens ordinalskaliert.

Rechenweg:

- (1) Bilde eine gemeinsame Stichprobe aus den Daten beider Gruppen
- (2) Bestimme den Median der gemeinsamen Stichprobe
- (3) Bestimme für jede Gruppe die Zahl der Beobachtungen, die kleiner als der Median sind sowie die Zahl der Beobachtungen, die größer oder gleich dem Median sind. Stelle aus diesen Anzahlen eine 2 x 2 Felder-Tafel auf.
- (4) Führe den Fisher-Exakt Test durch (vgl. Abschnitt 5.6).

Falls der Test nicht signifikant ist, kann die Nullhypothese identischer Mediane nicht verworfen werden. Andernfalls schließt man auf verschiedene Mediane.

Beispiel: Für die Plasma-Renin Studie ergibt sich aus der gemeinsamen Stichprobe der beiden Gruppen ein Median von 37,38. Für die Dobutamin-Gruppe sind 8 von 13 Beobachtungen kleiner als der Median, in der Fenoterol-Gruppe sind es 5 von 13. Hieraus ergibt sich folgende 2 x 2 Häufigkeits-Tafel:

	< Median	>=Median
Dobutamin	8	5
Fenoterol	5	8

Anwendung des **Fisher-Exakt-Test** liefert einen p -Wert (zweiseitig!) von $p = 0,4338$. Somit wird die Nullhypothese identischer Mediane nicht verworfen. Dobutamin und Fenoterol haben keine signifikant verschiedenen AUC-Werte.

11.1.2 Mann-Whitney Test

Mann-Whitney Test (teilweise auch als Wilcoxon-Test bezeichnet)

Fragestellung: Sind die Mediane zweier unabhängiger Stichproben signifikant verschieden? H_0 : die Mediane sind identisch.

Voraussetzung: Die beiden Grundgesamtheiten haben unter H_0 Verteilungen von gleicher Form, die Daten sind mindestens ordinalskaliert.

Rechenweg:

- (1) Bringe die $n_1 + n_2$ Stichprobenwerte in eine gemeinsame Rangfolge und berechne die Summen R_1 und R_2 der Rangzahlen der beiden Gruppen, wobei n_1 und n_2 die Stichprobenumfänge der beiden Gruppen sind.

$$(2) \text{ Berechne: } U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 - U_1$$

und wähle U_{vers} als den kleineren der beiden U-Werte.

$$(3) \text{ Berechne: } z_{vers} = \frac{\left| U_{vers} - \frac{n_1 n_2}{2} \right|}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

(4) Bestimme z_{tab} als das $(1-\alpha/2)$ -Quantil der Standardnormalverteilung ($z_{tab} = 1,96$ für $\alpha = 0,05$).

(5) Falls $z_{vers} > z_{tab}$, verwirfe H_0 . Andernfalls behalte H_0 bei.

Die Voraussetzung für diesen asymptotischen Test ist, dass $n_1, n_2 > 7$ und dass keine Rangbindungen auftreten (Zwei Beobachtungen haben denselben Rang). Andernfalls muss exakt getestet werden. Der exakte test erfordert Berechnungen, die praktikablerweise mit dem PC durchgeführt werden, z.B. mit der SAS-Prozedur NPAR1WAY. Wenn keine Bindungen vorliegen, können bei kleinen Stichproben alternativ tabellierte kritische Werte der exakten Verteilung von U_{vers} verwendet werden (Bortz et al., 1990).

Beispiel: Für die Plasma-Renin-Daten ergeben sich folgende Ränge und Rangsummen (beachte: es gibt keine Rangbindungen):

Dobutamin		Fenoterol	
Daten	Rang	Daten	Rang
18,80	2	34,16	10
41,70	17	67,82	25
41,24	16	60,68	23
49,98	19	39,04	15
63,90	24	37,72	14
29,86	7	46,78	18
25,92	4	50,80	20
36,92	12	128,60	26
54,90	21	31,18	8
37,04	13	35,04	11
28,94	6	56,40	22
33,80	9	22,62	3
18,08	1	26,09	5

Rangsummen: $R_1 = 151$ $R_2 = 200$
 $n_1 = 13$ $n_2 = 13$
Rangmittel: $R_1/n_1 = 11,6$ $R_2/n_2 = 15,4$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 13 \cdot 13 + \frac{13(13+1)}{2} - 151 = 169 + 91 - 151 = 109$$

$$U_2 = n_1 n_2 - U_1 = 169 - 109 = 60$$

$$U_{\text{vers}} = 60$$

$$z_{\text{vers}} = \frac{\left| U_{\text{vers}} - \frac{n_1 n_2}{2} \right|}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{|60 - 169/2|}{\sqrt{169 \cdot 27/12}} = 1,26 < z_{\text{tab}}(\alpha = 5\%) = 1,96$$

Die Mediane der beiden Behandlungen unterscheiden sich nicht signifikant. An den Rohdaten sowie an den Rangmitten sieht man, dass die Dobutamin-Gruppe den niedrigeren Median hat, aber dieser numerische Unterschied lässt sich statistisch nicht absichern.

Der mit dem PC berechnete asymptotische p-Wert ist $p = 0,2090$. Zum Vergleich der exakte Test mit der SAS Prozedur NPAR1WAY. Der exakte p-Wert ist $p = 0,2226$. Auch dieses Ergebnis lässt auf nicht-signifikante Unterschiede der Mediane schließen. Der asymptotische p-Wert ist etwas zu klein gegenüber dem exakten Wert.

Bemerkung: Oft hat der Median-Test eine geringere Teststärke als der Mann-Whitney-Test. Allerdings kann der Median-Test dann im Vorteil sein, wenn Ausreisser auftreten. Außerdem ist er zu bevorzugen, wenn die Form der Verteilungen nicht übereinstimmt.

Exakter Test

Der exakte Test beruht auf der Berechnung aller möglichen Werte, welche die Teststatistik U_{vers} annehmen kann. Für jeden der möglichen Werte muss die Wahrscheinlichkeit berechnet werden. Hierbei ist zu berücksichtigen, dass jede der möglichen Verteilungen der Ränge auf die Beobachtungen gleich wahrscheinlich ist. Dies soll an einem einfachen Beispiel erläutert werden. Angenommen, zwei Gruppen A und B werden in je drei Wiederholungen geprüft. Dann sind für den Mann-Whitney-Test die Ränge 1 bis 6 zu vergeben. Für die eindeutige Bestimmung des Wertes der Teststatistik ist es ausreichend, die für die Gruppe A vergebenen Ränge zu kennen. Die Reihenfolge innerhalb der Gruppe ist irrelevant für die Rangsumme und somit für

den Wert der Teststatistik. Daher gibt es $\binom{6}{3} = 20$ Konstellationen, von denen jede dieselbe Wahrscheinlichkeit von 5% hat:

Nr.	Ränge	R_1	U_1	U_2	U_{vers}	p
1	1 2 3	6	9	0	0	0,05
2	1 2 4	7	8	1	1	0,05
3	1 2 5	8	7	2	2	0,05
4	1 2 6	9	6	3	3	0,05
5	1 3 4	8	7	2	2	0,05
6	1 3 5	9	6	3	3	0,05
7	1 3 6	10	5	4	4	0,05
8	1 4 5	10	5	4	4	0,05
9	1 4 6	11	4	5	4	0,05
10	1 5 6	12	3	6	3	0,05
11	2 3 4	9	6	3	3	0,05
12	2 3 5	10	5	4	4	0,05
13	2 3 6	11	4	5	4	0,05
14	2 4 5	11	4	5	4	0,05
15	2 4 6	12	3	6	3	0,05
16	2 5 6	13	2	7	2	0,05
17	3 4 5	12	3	6	3	0,05
18	3 4 6	13	2	7	2	0,05
19	3 5 6	14	1	8	1	0,05
20	4 5 6	15	0	9	0	0,05

Um die exakte Verteilung von U_{vers} zu erhalten, sortiert man nach der Größe von U_{vers} :

Nr.	Ränge	R_1	U_1	U_2	U_{vers}	p	
1	1 2 3	6	9	0	0	0,05	} 0,10
20	4 5 6	15	0	9	0	0,05	
2	1 2 4	7	8	1	1	0,05	} 0,10
19	3 5 6	14	1	8	1	0,05	
3	1 2 5	8	7	2	2	0,05	} 0,20
5	1 3 4	8	7	2	2	0,05	
16	2 5 6	13	2	7	2	0,05	
18	3 4 6	13	2	7	2	0,05	} 0,30
4	1 2 6	9	6	3	3	0,05	
6	1 3 5	9	6	3	3	0,05	
10	1 5 6	12	3	6	3	0,05	
11	2 3 4	9	6	3	3	0,05	
15	2 4 6	12	3	6	3	0,05	} 0,30
17	3 4 5	12	3	6	3	0,05	
7	1 3 6	10	5	4	4	0,05	
8	1 4 5	10	5	4	4	0,05	
9	1 4 6	11	4	5	4	0,05	
12	2 3 5	10	5	4	4	0,05	
13	2 3 6	11	4	5	4	0,05	
14	2 4 5	11	4	5	4	0,05	

Es gibt nur fünf verschiedene Werte, welche die Teststatistik U_{vers} annehmen kann: 0, 1, 2, 3 und 4. Um die Wahrscheinlichkeitsfunktion zu erhalten, werden jeweils die Wahrscheinlichkeiten der Fälle addiert, welche zum selben Wert der Teststatistik führen. Der p -Wert für einen U_{vers} -Wert ist die Summe der Wahrscheinlichkeiten für Werte die so extrem sind wie der aktuelle oder noch extremer. Je kleiner der Wert, umso extremer.

U_{vers}	Wahrscheinlichkeit	P-Wert (exakt)
0	0,10	0,10
1	0,10	0,20
2	0,20	0,40
3	0,30	0,70
4	0,30	1,00

Zum Vergleich die p -Werte, die sich ergeben, wenn der oben beschriebene asymptotische Test basierend auf z_{vers} verwendet wird:

U_{vers}	P-Wert (asymptotisch)
0	0,0495
1	0,1266
2	0,2752
3	0,5127
4	0,8273

Diese p -Werte sind weit von den exakten entfernt. So würde ein Wert $U_{vers} = 0$ nach dem exakten Test also auf dem 5% Niveau nicht signifikant sein ($p = 0,10$), während er nach dem asymptotischen Test gerade signifikant wäre ($p = 0,0495$). Der asymptotische Test ist hier nicht gültig, weil die exakte Verteilung sehr diskret ist (nur fünf mögliche Werte), während die Normalverteilungsapproximation von einer stetigen Verteilung ausgeht. Je größer der Stichprobenumfang, umso besser wird die Normalverteilungsapproximation. Hieraus ergibt sich die oben genannte Faustregel, dass der asymptotische Test verwendet werden darf, falls $n_1, n_2 > 7$.

Rangbindungen

Wenn derselbe Messwert mehrmals vorkommt, ergeben sich Rangbindungen. Dies wird am einfachsten an einem kleinen Beispiel deutlich. Angenommen, es liegen folgende Messwerte vor:

Gruppe 1: 10, 34, 21
Gruppe 2: 10, 9, 25

Der Wert 9 erhält den Rang 1. Die Ränge 2 und 3 sind für die beiden Beobachtungen mit Wert 10 zu vergeben. Man vergibt hier den mittleren Rang 2,5. Die Werte 21, 25 und 34 erhalten die Ränge 4, 5 und 6.

Gruppe	Messwert	Rang
1	10	2,5
1	34	6
1	21	4
2	10	2,5
2	9	1
2	25	5

$$U_{\text{vers}} = 2,5, \quad p\text{-Wert (exakt)} = 0,50$$

Man überlegt sich leicht, dass die exakte Verteilung im Fall von Rangbindungen anders ist als ohne Rangbindungen (siehe Abschnitt "exakter Test"). Zum einen ergeben sich andere Werte für die Teststatistik U_{vers} . Ohne Rangbindungen sind bei diesem Design die möglichen Werte von U_{vers} gegeben durch 0, 1, 2, 3 und 4. Hier haben wir dagegen $U_{\text{vers}} = 2,5$. Zum anderen sind die p -Werte anders. Hier findet man $p = 0,50$, ein Wert, der im Fall ohne Bindungen nicht auftritt.

11.2 Kruskal-Wallis-Test (H-Test) für mehr als zwei unverbundene Stichproben

Beispiel: Die AUC-Werte aller vier Behandlungen der Plasma-Renin-Studie sind wie folgt:

Propanolol	Dobutamin	Fenoterol	Placebo
13,00	18,80	34,16	20,94
10,00	41,70	67,82	25,56
12,86	41,24	60,68	19,64
11,38	49,98	39,04	9,26
9,94	63,90	37,72	16,42
13,24	29,86	46,78	16,22
7,32	25,92	50,80	16,90
1,10	36,92	128,60	26,46
12,56	54,90	31,18	15,26
14,36	37,04	35,04	10,90
	28,94	56,40	18,16
	33,80	22,62	21,50
	18,08	26,09	26,58

Es sollen die Mittelwerte der vier Gruppen verglichen werden. Die zu prüfende Nullhypothese besagt, dass die Mediane der vier Gruppen übereinstimmen.

Kruskal-Wallis Test

Fragestellung: Sind die Mediane von t unabhängigen Stichproben signifikant verschieden? H_0 : die Mediane sind identisch.

Voraussetzung: Die Grundgesamtheiten haben unter H_0 Verteilungen von gleicher Form, die Daten sind mindestens ordinalskaliert.

Rechenweg:

(1) Bringe die $n_1 + n_2 + \dots + n_t$ Stichprobenwerte in eine gemeinsame Rangfolge und berechne die Summen R_1, R_2, \dots, R_t der Rangzahlen der Gruppen, wobei n_i der Stichprobenumfang der i -ten Gruppe ist.

(2) Berechne:
$$H_{vers} = \frac{12}{N(N+1)} \sum_{i=1}^t \frac{R_i^2}{n_i} - 3(N+1)$$

wobei $N = n_1 + n_2 + \dots + n_t$

(3) Bestimme H_{tab} als das $(1-\alpha)$ -Quantil der Chi-Quadrat-Verteilung mit $t-1$ Freiheitsgraden (Tab. VIII).

(4) Falls $H_{vers} > H_{tab}$, verwirfe H_0 . Andernfalls behalte H_0 bei.

Die Voraussetzung für diesen asymptotischen Test ist dass alle $n_i > 4$ und $t > 3$ ist und dass keine Rangbindungen auftreten (Zwei Beobachtungen haben denselben Rang). Andernfalls muss exakt getestet werden. Details können hier nicht besprochen werden. Mit der SAS-Prozedur NPAR1WAY kann der exakte Test durchgeführt werden. Wenn keine Bindungen vorliegen, können bei kleinen Stichproben alternativ tabellierte kritische Werte der exakten Verteilung von H_{vers} verwendet werden (Bortz et al., 1990).

Beispiel: Plasma-Renin-Studie (beachte: hier liegen keine Rangbindungen vor).

Propanolol		Dobutamin		Fenoterol		Placebo	
Daten	Rang	Daten	Rang	Daten	Rang	Daten	Rang
13,00	10	18,80	19	34,16	33	20,94	21
10,00	5	41,70	40	67,82	48	25,56	24
12,86	9	41,24	39	60,68	46	19,64	20
11,38	7	49,98	42	39,04	38	9,26	3
9,94	4	63,90	47	37,72	37	16,42	15
13,24	11	29,86	30	46,78	41	16,22	14
7,32	2	25,92	25	50,80	43	16,90	16
1,10	1	36,92	35	128,60	49	26,46	27
12,56	8	54,90	44	31,18	31	15,26	13
14,36	12	37,04	36	35,04	34	10,90	6
		28,94	29	56,40	45	18,16	18
		33,80	32	22,62	23	21,50	22
		18,08	17	26,09	26	26,58	28

Rangsummen:

$$R_1 = 69$$

$$R_2 = 435$$

$$R_3 = 494$$

$$R_4 = 227$$

Rangmittel:

$$R_1/n_1 = 6,9$$

$$R_2/n_2 = 33,5$$

$$R_3/n_3 = 38,0$$

$$R_4/n_4 = 17,5$$

$$\begin{aligned}
H_{vers} &= \frac{12}{N(N+1)} \sum_{i=1}^t \frac{R_i^2}{n_i} - 3(N+1) \\
&= \frac{12}{49 \cdot 50} \left(\frac{69^2}{10} + \frac{435^2}{13} + \frac{494^2}{13} + \frac{227^2}{13} \right) - 3 \cdot 50 \\
&= 34,984
\end{aligned}$$

$$H_{tab} = 7,815 < H_{vers}$$

Die Nullhypothese gleicher Mediane wird verworfen. Die Rangmittel zeigen, dass Propanolol und der Placebo die kleinsten Mediane haben. Eine weitere Analyse umfasst paarweise Vergleiche mit dem Mann-Whitney-Test aus Abschnitt 11.1.1.

11.3 Vergleich zweier verbundener Stichproben

Beispiel (Brunner und Langer, 1999, S. 14, z.T. wörtlich übernommen): Bei Patienten, die wegen eines Karzinoms eine Chemotherapie erhalten sollen, werden vor Therapiebeginn aus dem peripheren Blut sog. Stammzell-Konzentrate gewonnen, die nach der Therapie den Patienten zur Regeneration des hämolytischen Systems wieder re-infundiert werden. Das Stammzell-Konzentrat wird zur Konservierung bei -196 °C in flüssigem Stickstoff eingefroren. Es ist nun wichtig zu wissen, ob durch den Vorgang des Einfrierens und Wiederauftauens wesentliche Eigenschaften der Stammzellen verloren gehen.

Eine wichtige Messgröße ist die Anzahl der koloniebildenden Einheiten der Granulozyten-Makrophagen Linie (colony forming units, CFU-GM). Diese wurde jeweils vor dem Einfrieren und nach dem Wiederauftauen bestimmt. Als wesentliche Einflussfaktoren auf die Anzahl der CFU-GM ist u.a. die Belastung des Blutbildenden Systems durch Art und Ausmaß der möglichen Vorbelastungen durch Chemotherapien (niedrig/hoch) vermerkt worden. Ferner ist wegen der verschiedenen Tumorarten das Geschlecht der Patienten zu berücksichtigen. Wir betrachten hier lediglich die Gruppe der männlichen Patienten mit niedriger Vorbelastung. Die CFU-GM-Werte [$10^5/\text{kg}$] für diese Gruppe waren wie folgt:

CFU-GM [10^5 /kg]

Probe Nr.	vor	nach
90	1,4580	1,4160
104	1,6550	0,8020
105	3,0160	1,8210
117	0,8110	0,7880
118	1,6120	1,5040
145	0,1580	0,1450
146	0,1660	0,1120
147	0,2870	0,3180
149	0,7460	0,4930
150	1,1900	1,8090
151	2,4150	1,0610
163	157,4910	220,3630
179	7,1520	4,5660
186	1,5650	1,3240
188	9,2190	5,3350
189	3,0970	2,1090
190	0,9720	0,4000
191	0,8110	1,0690
194	1,8430	1,8090
199	2,5290	1,9260
200	7,3290	4,7940
201	0,9550	0,7850
217	4,2000	1,7240
218	1,9720	0,6820
219	1,7070	1,2520
241	4,0680	1,1470
240	3,8100	2,4112
260	12,7812	8,1911
89	0,3831	0,1406
203	3,3303	1,2813

Man beachte, dass für Probe Nr. 163 sehr hohe Werte gemessen wurden, so dass es sich um einen Ausreißer handelt. Außerdem ist die Verteilung der Messwerte auch ohne den Ausreißer relativ schief, so dass die Normalverteilungsannahme verletzt ist. Der Ausreißer führt dazu, dass ein verbundener t-Test, der wegen der Abweichung von der Normalverteilungsannahme problematisch ist, nicht signifikant ist ($p = 0,6132$).

11.3.1 Vorzeichen-Test

Vorzeichen-Test

Fragestellung: Ist der Median der Differenzen in einer verbundenen Stichprobe signifikant von Null verschieden? H_0 : der Median der Differenzen ist gleich Null.

Voraussetzung: Die gemeinsame Verteilung der beiden Behandlungen erfüllt die Austauschbarkeitsbedingung, d.h., die Form der Verteilung ändert sich nicht, wenn die beiden Behandlungen vertauscht werden. Die Daten sind mindestens ordinalskaliert.

Rechenweg:

(1) Berechne die paarweisen Differenzen d_i der Beobachtungen. Falls Differenzen $d_i = 0$ auftreten, verwerfe die betreffenden Beobachtungen (Bei stetigen Verteilungen tritt dieser Fall nicht ein). Der Stichprobenumfang nach Verwerfen der Null Differenzen wird mit n bezeichnet.

(2) Unter der Nullhypothese treten negative und positive Differenzen mit gleicher Wahrscheinlichkeit auf. Die Zahl der positiven (und die der negativen) Differenzen folgt einer Binomialverteilung mit Wahrscheinlichkeit $\pi = 0,5$ und Konstante n . Hiervon ausgehend lässt sich ein exakter p-Wert wie folgt berechnen:

$$p = 2 \times 0.5^n \sum_{i=0}^m \binom{n}{i}$$

wobei m die kleinere der beiden Anzahlen ist (Anzahl negativer und positiver Differenzen). Der p -Wert gibt die Wahrscheinlichkeit, einen Wert für m zu erhalten, der so klein ist wie der beobachtete oder noch extremer (beachte die Analogie zum Fisher-Exakt-Test). Der p -Wert entspricht der Wahrscheinlichkeit

$$p = P(Y \leq m) + P(Y \geq n - m)$$

wobei Y einer Binomialverteilung mit Wahrscheinlichkeit $\pi = 0,5$ und Konstante n folgt.

(3) Falls $p < \alpha$, verwerfe H_0 . Andernfalls behalte H_0 bei.

Beispiel:

Probe Nr.	CFU-GM [10^5 /kg]		Differenz	Vorzeichen
	vor	nach		
90	1,4580	1,4160	0,0420	+
104	1,6550	0,8020	0,8530	+
105	3,0160	1,8210	1,1950	+
117	0,8110	0,7880	0,0230	+
118	1,6120	1,5040	0,1080	+
145	0,1580	0,1450	0,0130	+
146	0,1660	0,1120	0,0540	+
147	0,2870	0,3180	-0,0310	-
149	0,7460	0,4930	0,2530	+
150	1,1900	1,8090	-0,6190	-
151	2,4150	1,0610	1,3540	+
163	157,4910	220,3630	-62,8720	-
179	7,1520	4,5660	2,5860	+
186	1,5650	1,3240	0,2410	+
188	9,2190	5,3350	3,8840	+
189	3,0970	2,1090	0,9880	+
190	0,9720	0,4000	0,5720	+
191	0,8110	1,0690	-0,2580	-
194	1,8430	1,8090	0,0340	+
199	2,5290	1,9260	0,6030	+
200	7,3290	4,7940	2,5350	+
201	0,9550	0,7850	0,1700	+
217	4,2000	1,7240	2,4760	+
218	1,9720	0,6820	1,2900	+
219	1,7070	1,2520	0,4550	+
241	4,0680	1,1470	2,9210	+
240	3,8100	2,4112	1,3988	+
260	12,7812	8,1911	4,5901	+
89	0,3831	0,1406	0,2425	+
203	3,3303	1,2813	2,0490	+

Die Zahl der negativen Vorzeichen ist $m = 4$ bei einem Stichprobenumfang von $n = 30$, was darauf hindeutet, dass eine Reduktion des CFU Wertes stattgefunden hat. Der exakte p-Wert ergibt sich als

$$p = 2 \times 0,5^{30} \sum_{i=0}^4 \binom{30}{i} = 0,000059476$$

Somit besteht ein deutlich signifikanter Unterschied zwischen den CFU Werten vor und nach der Therapie. Dieser exakte p-Wert kann mit der SAS Prozedur UNIVARIATE berechnet werden (siehe 11.3.3). Man beachte, dass im Gegensatz zum t-Test mit dem Vorzeichen-Test ein signifikanter Unterschied gefunden wird. Der Vorzeichen Test ist also unempfindlich gegen Ausreißer.

11.3.2 Vorzeichen-Rangtest (Wilcoxon-Test)

Die Berechnung der Teststatistik für diesen Test ist einfach nachzuvollziehen. Selbst bei einem Stichprobenumfang von 50 Messwertpaaren muss allerdings ein exakter p-Wert mittels des sog. Shift-Algorithmus berechnet werden (Brunner und Langer, 1999), was nur mittels eines Computers möglich ist. Trotz der Notwendigkeit der Nutzung eines Computers werden hier die Rechenschritte zur Bestimmung der Teststatistik beschrieben.

Vorzeichen-Rangtest

Fragestellung: Sind die Mediane zweier verbundener Stichproben signifikant verschieden? H_0 : Die Mediane sind gleich.

Voraussetzung: Die gemeinsame Verteilung der beiden Behandlungen erfüllt die Austauschbarkeitsbedingung, d.h., die Form der Verteilung ändert sich nicht, wenn die beiden Behandlungen vertauscht werden. Die Daten sind mindestens ordinalskaliert.

Rechenweg:

(1) Berechne die paarweisen Differenzen d_i der Beobachtungen. Falls Differenzen $d_i = 0$ auftreten, verwerfe die betreffenden Beobachtungen (Bei stetigen Verteilungen tritt dieser Fall nicht ein). Der Stichprobenumfang nach Verwerfen der Null Differenzen wird mit n bezeichnet.

(2) Bringe die Absolutbeträge (!) der n Differenzen in eine Rangfolge.

(3) Berechne

W^+ = Summe der Ränge von positiven Differenzen und

W^- = Summe der Ränge von negativen Differenzen

$$W^+ + W^- = \frac{n(n+1)}{2}$$

und nimm die kleinere der beiden Größen W^+ und W^- als W_{Vers} .

(4) Berechne den exakten p-Wert (geht nur mittels Computerprogramm). Falls $p < \alpha$, verwerfe H_0 . Andernfalls behalte H_0 bei.

Probe Nr.	vor	nach	Betrag (!) Differenz	Vorz.	Rang
90	1,4580	1,4160	0,0420	+	5
104	1,6550	0,8020	0,8530	+	17
105	3,0160	1,8210	1,1950	+	19
117	0,8110	0,7880	0,0230	+	2
118	1,6120	1,5040	0,1080	+	7
145	0,1580	0,1450	0,0130	+	1
146	0,1660	0,1120	0,0540	+	6
147	0,2870	0,3180	0,0310	-	3
149	0,7460	0,4930	0,2530	+	11
150	1,1900	1,8090	0,6190	-	16
151	2,4150	1,0610	1,3540	+	21
163	157,4910	220,3630	62,8720	-	30
179	7,1520	4,5660	2,5860	+	26
186	1,5650	1,3240	0,2410	+	9
188	9,2190	5,3350	3,8840	+	28
189	3,0970	2,1090	0,9880	+	18
190	0,9720	0,4000	0,5720	+	14
191	0,8110	1,0690	0,2580	-	12
194	1,8430	1,8090	0,0340	+	4
199	2,5290	1,9260	0,6030	+	15
200	7,3290	4,7940	2,5350	+	25
201	0,9550	0,7850	0,1700	+	8
217	4,2000	1,7240	2,4760	+	24
218	1,9720	0,6820	1,2900	+	20
219	1,7070	1,2520	0,4550	+	13
241	4,0680	1,1470	2,9210	+	27
240	3,8100	2,4112	1,3988	+	22
260	12,7812	8,1911	4,5901	+	29
89	0,3831	0,1406	0,2425	+	10
203	3,3303	1,2813	2,0490	+	23

$$W^- = 3+18+30+12 = 63$$

$$W^+ = 30 \cdot 31/2 - W^- = 465 - 63 = 402$$

$$W_{\text{vers}} = 63$$

Der exakte p-Wert beträgt $p = 0,000092543$ (berechnet mit der SAS Prozedur UNIVARIATE). Auch der Vorzeichen-Rangtest zeigt eine signifikante Differenz an. Abschließend nochmals die p-Werte der drei angewendeten Tests:

Test	p-Wert (exakt)
t-Test	0,61324
Vorzeichen-Test	0,000059476
Vorzeichen-Rangtest	0,000092543

Bemerkung: Der Vorzeichen-Rangtest nutzt die Information über die Größe der Differenzen, während der Vorzeichen-Test nur das Vorzeichen der Differenzen berücksichtigt. Daher hat der Vorzeichen-Rangtest in der Regel eine bessere

Teststärke. Allerdings weist der Vorzeichen-Test eine größere Robustheit gegenüber Ausreißern auf.

SAS Anweisungen für Vorzeichen-Test und Vorzeichen-Rangtest

```
data stamm;  
input probe vor nach;  
    diff=vor-nach;  
datalines;  
    90    1.4580    1.4160  
    104    1.6550    0.8020  
    105    3.0160    1.8210  
    117    0.8110    0.7880  
    118    1.6120    1.5040  
    145    0.1580    0.1450  
    146    0.1660    0.1120  
    147    0.2870    0.3180  
    149    0.7460    0.4930  
    150    1.1900    1.8090  
    151    2.4150    1.0610  
    163    157.4910    220.3630  
    179    7.1520    4.5660  
    186    1.5650    1.3240  
    188    9.2190    5.3350  
    189    3.0970    2.1090  
    190    0.9720    0.4000  
    191    0.8110    1.0690  
    194    1.8430    1.8090  
    199    2.5290    1.9260  
    200    7.3290    4.7940  
    201    0.9550    0.7850  
    217    4.2000    1.7240  
    218    1.9720    0.6820  
    219    1.7070    1.2520  
    241    4.0680    1.1470  
    240    3.8100    2.4112  
    260    12.7812    8.1911  
    89    0.3831    0.1406  
    203    3.3303    1.2813  
;  
proc univariate data=stamm;  
var diff;  
run;
```

11.4 Friedman-Test für mehr als zwei verbundene Stichproben

Beispiel (Köhler et al., S. 197): Die Wirkung zweier Insektizide (DDT und Malathion) sowie einer Kontrolle wurde in einem Blockversuch untersucht. Dazu hat man 6 zufällig ausgewählte, verschiedene Felder zu je einem Drittel mit DDT, mit Malathion

bzw. nicht (Kontrolle) besprüht und eine Woche später stichprobenartig nach Insektenlarven abgesucht. Jedes Feld entspricht einem Block.

Feld (Block)	Anzahl Larven		
	Kontrolle	DDT	Malathion
1	10	4	3
2	14	2	6
3	17	0	8
4	8	3	0
5	9	2	3
6	31	11	16

Friedman-Test

Fragestellung: Bestehen signifikante Unterschiede zwischen den Medianen von t verbundenen Stichproben? H_0 : Die t Mediane sind gleich.

Voraussetzung: Die gemeinsame Verteilung der Behandlungen erfüllt die paarweise Austauschbarkeitsbedingung. Es liegen keine Rangbindungen vor.

Rechenweg:

(1) Bringe die Beobachtung *getrennt für jeden Block (!)* in eine Rangfolge.

(2) Berechne die Rangsumme R_i für jede Behandlung.

(3) Berechne

$$Q_{vers} = \left(\frac{12}{n \cdot t \cdot (t+1)} \sum_{i=1}^t R_i^2 \right) - 3n \cdot (t+1)$$

(n = Zahl der Wiederholungen (Blöcke), t = Zahl der Behandlungen)

(4) Falls $t = 3$ und $3 \leq n \leq 9$ oder
 $t = 4$ und $3 \leq n \leq 4$

berechne exakte p-Werte mittels der unten stehenden Tabelle (Bortz und Lienert, 1999). Falls $p < \alpha$, verwirfe H_0 . Andernfalls behalte H_0 bei.

Andernfalls teste asymptotisch. Bestimme hierzu Q_{tab} als $(1-\alpha)$ -Quantil der Chi-Quadrat-Verteilung mit $(t-1)$ Freiheitsgraden (Tab. VIII). Falls $Q_{vers} > Q_{tab}$, verwirfe H_0 . Andernfalls behalte H_0 bei.

Tabelle: Exakte p-Werte für den Friedman-Test.

$t = 3$ (Zahl der Behandlungen)

$n = 3$		$n = 4$		$n = 5$			
Q_{vers}	P-Wert	Q_{vers}	P-Wert	Q_{vers}	P-Wert		
2,0	0,528	3,5	0,273	2,8	0,367		
2,7	0,361	4,5	0,125	3,6	0,182		
4,7	0,194	6,0	0,069	4,8	0,124		
6,0	0,028	6,5	0,042	5,2	0,093		
		8,0	0,0046	6,4	0,039		
				7,6	0,024		
				8,4	0,0085		
				10,0	0,00077		
$n = 6$		$n = 7$		$n = 8$		$n = 9$	
Q_{vers}	P-Wert	Q_{vers}	P-Wert	Q_{vers}	P-Wert	Q_{vers}	P-Wert
2,3	0,430	3,7	0,192	3,3	0,236	2,9	0,278
3,0	0,252	4,6	0,112	4,0	0,149	4,2	0,154
4,0	0,184	5,4	0,085	4,8	0,120	4,7	0,107
4,3	0,142	6,0	0,052	5,3	0,079	5,6	0,069
5,3	0,072	7,1	0,027	6,3	0,047	6,2	0,048
6,3	0,052	7,7	0,021	7,0	0,030	8,0	0,019
7,0	0,029	8,0	0,016	9,0	0,0099	8,7	0,010
8,3	0,012	8,9	0,0084	9,8	0,0048	9,6	0,0060
9,0	0,0081	10,3	0,0036	10,8	0,0024	10,7	0,0035
9,3	0,0055	10,6	0,0027	12,0	0,0011	11,6	0,0013
10,3	0,0017	11,1	0,0012	12,3	0,00086	12,7	0,00066
12,0	0,00013	12,3	0,00032	13,0	0,00026	14,0	0,00020

$t = 4$ (Zahl der Behandlungen)

$n = 3$		$n = 4$					
Q_{vers}	P-Wert	Q_{vers}	P-Wert	Q_{vers}	P-Wert	Q_{vers}	P-Wert
5,0	0,207	3,6	0,355	6,3	0,094	9,3	0,012
5,8	0,148	3,9	0,324	6,6	0,077	9,9	0,0062
6,6	0,075	4,5	0,242	6,9	0,068	10,2	0,0027
7,4	0,033	4,8	0,200	7,5	0,052	10,8	0,0016
8,2	0,017	5,4	0,158	7,8	0,036	11,1	0,00094
9,0	0,0017	5,7	0,141	8,4	0,019	12,0	0,00007

Beispiel:

Feld (Block)	Rang Anzahl Larven innerhalb Block		
	Kontrolle	DDT	Malathion
1	3	2	1
2	3	1	2
3	3	1	2
4	3	2	1
5	3	1	2
6	3	1	2
R_i	18	8	10

$t = 3, n = 6 \Rightarrow$ exakter Test

$$Q_{\text{vers}} = \frac{12}{6 \cdot 3 \cdot 4} (18^2 + 8^2 + 10^2) - 3 \cdot 6 \cdot 4 = 7,0$$

p -Wert = 0,029

Es bestehen signifikante Unterschiede zwischen den drei Behandlungen.

Bemerkung 1: Der exakte Friedman-Test ist in vielen Software-Paketen nicht implementiert, weil die Berechnung exakter p -Werte sehr rechenaufwendig ist. Man muss hier auf in Lehrbüchern tabellierte Werte zurückgreifen oder den asymptotischen Test durchführen. Bei Rangbindungen stehen keine tabellierten Werte zur Verfügung. Die einfachste Lösung für diesen Fall besteht in einer Varianzanalyse der Rangdaten nach dem Modell für eine Blockanlage (Abschnitt 8.7). Diese ist im Fall, dass keine Rangbindungen vorliegen, asymptotisch äquivalent zum oben beschriebenen Chi-Quadrat-Test. In manchen Büchern werden Adjustierungsformeln für den Fall von Rangbindungen angegeben.

Bemerkung 2: Multiple paarweise Vergleiche werden am besten mit dem Vorzeichen-Rangtest durchgeführt. Hier stellt sich allerdings oft das Problem des geringen Stichprobenumfanges und damit der geringen Teststärke.

11.5 Vor- und Nachteile nichtparametrischer Verfahren

(1) **Es stehen vor allem einfache nichtparametrische Verfahren zur Verfügung:** Nichtparametrische Verfahren, die analoge parametrische Verfahren ersetzen können, gibt es vor allem für einfache Verfahren (t-Test, Varianzanalyse). Bei komplexeren Designs und Fragestellungen ist die nichtparametrische Methodik nicht so flexibel und weit entwickelt wie das auf linearen Modellen basierende Instrumentarium, obschon es einige vielversprechende Entwicklungen gibt (Brunner, E., und Langer, F., 1999, Nichtparametrische Analyse longitudinaler Daten. Oldenbourg-Verlag, München).

(2) Nichtparametrische Verfahren sind nicht frei von statistischen Annahmen:

Diese Tatsache wird oft übersehen. Viele Verfahren treffen die Annahme, dass die Fehler identisch und unabhängig verteilt sind. Somit wird eigentlich nur die Annahme der Normalverteilung fallengelassen. Selbst diese schwächeren Annahmen können verletzt sein, z.B. bei Varianzheterogenität.

(3) Robustheit parametrischer Verfahren: Viele parametrische Verfahren, z.B. die Varianzanalyse und der t-Test sind sehr **robust** gegen Abweichungen von der Normalverteilung. Dies bedeutet, dass diese Verfahren das vorgegebene Signifikanzniveau α auch dann noch "einigermaßen" einhalten, wenn die Fehler nicht normalverteilt sind. Die Robustheit ist um so größer, je größer der Stichprobenumfang ist. Diese Eigenschaft ist im wesentlichen auf die Wirkung des zentralen Grenzwertsatzes zurückzuführen (vgl. Abschnitt 3.4).

(4) Modifikationen parametrischer Verfahren oft zu bevorzugen: Desweiteren gibt es Modifikationen parametrischer Verfahren, die dem Problem der Varianzheterogenität begegnen (z.B. Welch-Test als Modifikation des t-Tests) und es ermöglichen, weiterhin im sehr flexiblen Rahmen linearer Modelle auszuwerten. Solche Erweiterungen sind unter dem Dach der sog. **gemischten Modelle** zu fassen, auf die an späterer Stelle noch einzugehen ist. Außerdem führt eine Transformation oft zu Daten, welche die üblichen Voraussetzungen besser erfüllen.

(5) Informationsverlust: Durch die Rangtransformation geht Information verloren. Bei zumindest annähernder Normalverteilung der Daten haben nichtparametrische Tests daher oft eine geringere Teststärke (einen höheren β -Fehler) als ihre parametrischen Analoga.

Trotz dieser Nachteile erfreuen sich nichtparametrische Verfahren großer Beliebtheit, und diese Beliebtheit ist berechtigt. Sie sind vor allem dann angezeigt, wenn die Stichprobenumfänge klein sind, weil dann parametrische Verfahren weniger robust sind, und wenn eine starke Verletzung der Normalverteilungsannahme vermutet wird. Ein Dilemma besteht darin, dass es gerade bei kleinen Stichproben schwer ist, die Annahme der Normalverteilung zu überprüfen (Residuenplots, Tests).

(6) Bei Rangbindungen exakt testen: Abschließend sei nochmals betont, dass bei Rangbindungen exakt getestet werden sollte, wobei ein Computer verwendet werden muss, da die exakte Verteilung nur für den Fall tabelliert ist, dass keine Bindungen vorliegen. Die Rechenzeit kann bei Tests zum Vergleich von mehr als zwei Gruppen (Kruskal-Wallis- oder Friedman-Test) allerdings bei großen Stichproben sehr lang werden. In vielen Lehrbüchern finden sich Adjustierungsformeln für den Fall von Rangbindungen, die zu asymptotisch gültigen Test führen und bei großen Stichproben anwendbar sind. Hier werden diese Adjustierungen der Kürze halber nicht behandelt.

(7) Bei kleinen Stichproben ist generell ein exakter Test vorzuziehen.

12. Kovarianzanalyse

In vielen Versuchen ist damit zu rechnen, dass Störgrößen einen Einfluss auf das Versuchsergebnis haben. Solche Störgrößen sollten bei der Planung wie bei der Auswertung berücksichtigt werden. Kann die Störgröße vor Beginn des Versuches gemessen werden, so kann eine Blockbildung benutzt werden, um den Einfluss der Störgröße zu reduzieren (Kap. 8) (Blöcke kann man natürlich auch bilden, wenn die Störgröße zwar bekannt ist, aber nicht gemessen wird/werden kann, sondern nur abgeschätzt werden kann). Eine andere Möglichkeit besteht in der Verwendung der Kovarianzanalyse. Die Idee hierbei ist, Beobachtungen von Versuchseinheiten mit verschiedenen Versuchsbedingungen, gemessen mit Hilfe einer Kovariable, so zu korrigieren, als hätten sie dieselben Bedingungen gehabt.

Beispiel: In einem Fütterungsexperiment wurde der Einfluss vier verschiedener Fütterungsarten auf die täglichen Zunahmen von Schweinen untersucht. Es ist bekannt, dass die täglichen Zunahmen im Durchschnitt um so höher sind, je größer das Anfangsgewicht der Schweine ist. Es lässt sich nicht vermeiden, dass in einem Fütterungsversuch die Anfangsgewichte von Tier zu Tier variieren. Das Anfangsgewicht ist in diesem Versuch eine Störgröße, deren Einfluss ausgeschaltet oder zumindest minimiert werden sollte. Eine Möglichkeit besteht darin, Gruppen von Tieren zu bilden, so dass innerhalb der Gruppen die Anfangsgewichte möglichst ähnlich sind (Blockbildung; siehe Kapitel 8). Eine zweite Möglichkeit ist die Anwendung der Kovarianzanalyse, wobei die Anfangsgewichte als Kovariable verwendet werden.

Grundlage der Kovarianzanalyse ist eine Regression der Zielvariable (z.B. tägliche Zunahme) auf die Kovariable (z.B. Anfangsgewicht). Für jede Behandlung wird eine solche Regression durchgeführt. Falls die Regressionsgeraden parallel verlaufen, können die vertikalen Abstände der Regressionsgeraden zur Beurteilung von Behandlungsunterschieden herangezogen werden.

Beispiel: Das Prinzip der Kovarianzanalyse soll an folgendem hypothetischen (extremen) Beispiel verdeutlicht werden. Die nachstehenden Daten sind für zwei verschiedene Futter 1 und 2, welche an je sechs Schweine verabreicht wurden.

Futter 1		Futter 2	
Anfangsgewicht x (Pfund)	Zunahme y (Pfund)	Anfangsgewicht x (Pfund)	Zunahme y (Pfund)
41	0,89	67	0,94
23	0,76	73	1,01
18	0,75	55	0,88
33	0,89	67	0,92
25	0,71	72	0,99
30	0,88	53	0,81
Mittel:	0,81		0,93

Die Gruppen für die zwei Futterarten unterscheiden sich stark im Anfangsgewicht. In der Praxis würde man solch starke Unterschiede durch eine randomisierte und somit

gleichmäßigere Verteilung der Tiere vermeiden. Trotzdem lassen sich gewisse Unterschiede im Anfangsgewicht nicht ausschalten. Die Wirkung solcher Unterschiede soll durch das vorliegende Beispiel deutlich werden.

In der folgenden Abbildung sind die Zunahmen gegen die Anfangsgewichte abgetragen. Wir beobachten einen positiven Zusammenhang von Zunahme und Anfangsgewicht (durch Regressionsgeraden hervorgehoben).

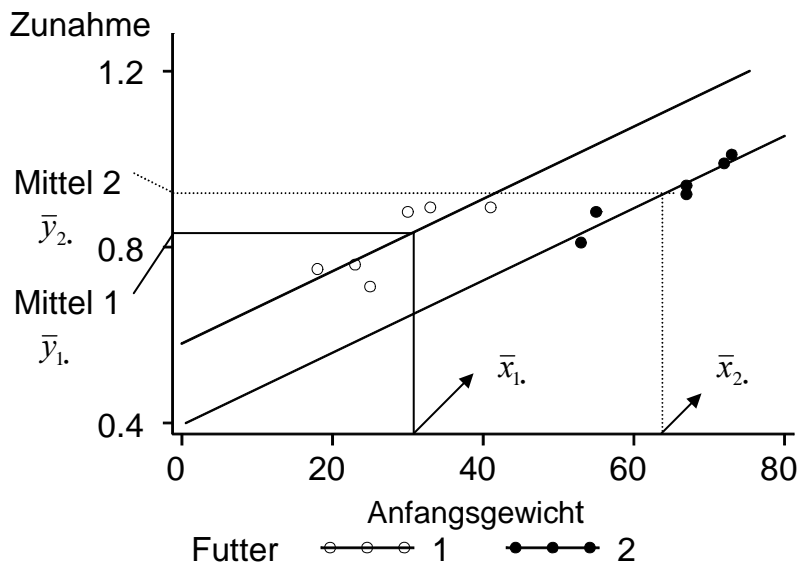


Abb. 12.1: Anfangsgewichte (Kovariablen) und Zunahmen (Zielvariable) für zwei Futtermittel - einfache Mittelwerte.

Futter 1 hat kleinere tägliche Zunahmen (Mittelwert: 0,81; Futter 2: 0,93), aber auch kleinere Anfangsgewichte. Die Regressionsgerade für Futter 1 liegt über der von Futter 2. Dies ist so zu interpretieren, dass bei gleichem Anfangsgewicht Futter 1 die besseren Zunahmen aufgewiesen hätte als Futter 2, und das obwohl im Versuch Futter 1 einen kleineren Mittelwert der täglichen Zunahmen aufweist als Futter 2.

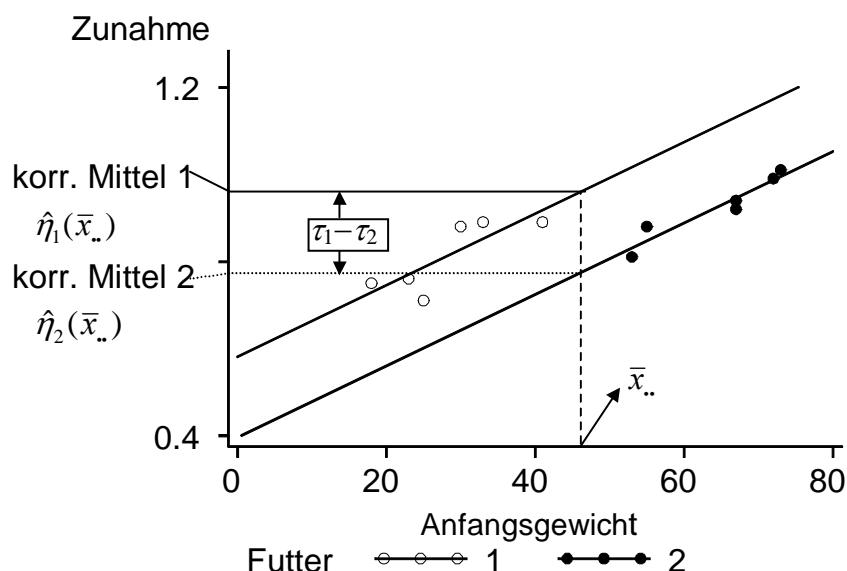


Abb. 12.2: Kovarianzanalyse für Anfangsgewichte (Kovariablen) und Zunahmen (Zielvariable) bei zwei Futtermitteln - adjustierte und einfache Mittelwerte.

Die Kovarianzanalyse korrigiert nun die Behandlungsmittelwerte mittels der Regressionslinien auf den gleichen Wert der Störgröße (Anfangsgewicht). In der Regel wird als Bezugsgröße der Mittelwert der Kovariable verwendet. Der Mittelwert des Anfangsgewichts beträgt 46,4 Pfund. Wir ermitteln nun den Wert der täglichen Zunahme, der nach der Regression für ein Anfangsgewicht von 46,4 Pfund zu erwarten wäre. Dies ist in obiger Graphik veranschaulicht. Der korrigierte Mittelwert für Futter 1 liegt oberhalb des korrigierten Mittelwertes für Futter 2 (während der unkorrigierte Mittelwert für Futter 1 kleiner ist als der von Futter 2!). Die Differenz der korrigierten Mittelwerte entspricht dem vertikalen Abstand der beiden Regressionslinien. Dieses Beispiel macht deutlich, dass wir beim Ignorieren der Kovariablen ein verzerrtes Bild bekommen hätten. Die Verzerrung wäre so extrem, dass wir schließen würden, Futter 1 ist schlechter als Futter 2, während wir bei Berücksichtigung der unterschiedlichen Anfangsgewichte zum entgegengesetzten Ergebnis kommen. □

Natürlich ist die Kovarianzanalyse nur zulässig und sinnvoll, wenn die Regressionslinien parallel verlaufen. Diese Annahme muss jeweils überprüft werden.

Neben der Korrektur auf denselben Wert der Kovariable führt die Kovarianzanalyse meist auch zu einem Gewinn an Genauigkeit, da der Teil der Versuchsstreuung, der durch die Kovariable zu erklären ist, ausgeschaltet wird. Dieser Gewinn an Genauigkeit wird später an einem anderen Beispiel erläutert.

Neben der Möglichkeit der Ausschaltung einer Störgröße kann die Kovarianzanalyse außerdem eingesetzt werden, um Daten auszuwerten, bei denen ein quantitativer und ein qualitativer Einflussfaktor zu berücksichtigen sind. Hierauf wird in Abschnitt 12.4 an einen soziologischen Beispiel eingegangen. Zunächst wird das Vorgehen bei einer Kovarianzanalyse zur Ausschaltung einer Störgröße detailliert am Beispiel eines Fütterungsversuches erläutert (Abschnitte 12.1 bis 12.3).

12.1 Modellbildung

Das kovarianzanalytische Modell kann als Erweiterung des einfachen varianzanalytischen Modells

$$y_{ij} = \mu + \tau_i + e_{ij}$$

(μ = Gesamtmittel; τ_i = Behandlungseffekt) um einen Regressionsterm angesehen werden. Das erweiterte Modell lautet:

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + e_{ij} \quad (i = 1, \dots, t; j = 1, \dots, r_i)$$

wobei x_{ij} der Wert der Kovariable für die Beobachtung y_{ij} ist und β die gemeinsame Steigung. Wir lassen hier zu, dass die Zahl der Wiederholungen nicht für jede Behandlung dieselbe ist (r_i = Zahl der Wiederholungen der i -ten Behandlung). Die Fehler e_{ij} werden wie üblich als normalverteilt mit Mittelwert Null und Varianz σ^2 angenommen [$e_{ij} \sim N(0, \sigma^2)$].

In diesem Modell werden Behandlungsunterschiede gemessen, indem die Erwartungswerte bei konstantem Wert der Kovariablen verglichen werden. Der Erwartungswert der i -ten Behandlung an der Stelle x_{ij} ist

$$\eta_i(x_{ij}) = E(y_{ij}|x_{ij}) = \mu + \tau_i + \beta x_{ij}$$

Die Differenz der Erwartungswerte an der Stelle x_0 für zwei Behandlungen u und s ist:

$$\eta_u(x_0) - \eta_s(x_0) = E(y_{uj}|x_0) - E(y_{sj}|x_0) = \mu + \tau_u + \beta x_0 - (\mu + \tau_s + \beta x_0) = \tau_u - \tau_s$$

Diese Differenz ist nur abhängig von den Behandlungseffekten. Sie entspricht dem vertikalen Abstand der Regressionsgeraden für die beiden Behandlungen.

Unter der Nullhypothese lautet das (reduzierte) Modell:

$$y_{ij} = \mu + \beta x_{ij} + e_{ij}$$

Nach diesem Modell liegen die Regressionslinien aller Behandlungen übereinander.

Die Kovarianzanalyse basiert auf der Annahme, dass die Regressionslinien parallel verlaufen. Um diese Annahme zu prüfen, müssen wir das Kovarianzanalyse-Modell so erweitern, dass jede Behandlung eine eigene Steigung haben kann. Das erweiterte Modell lautet:

$$y_{ij} = \mu + \beta_i x_{ij} + \tau_i + e_{ij}$$

wobei β_i die Steigung der i -ten Behandlung ist. Die Nullhypothese der Parallelität der Regressionsgeraden lautet:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_t.$$

Für die Prüfung der Nullhypothese ist es sinnvoll, die Steigung β_i zu reparametrisieren, wie wir auch im varianzanalytischen Modell $y_{ij} = \mu + \tau_i + e_{ij}$ den Erwartungswert μ_i als Summe von Gesamtmittelwert und Behandlungseffekt ausgedrückt haben ($\mu_i = \mu + \tau_i$). Analog schreiben wir

$$\beta_i = \beta + \delta_i$$

wobei β eine gemeinsame Steigung ist und δ_i die Abweichung von der gemeinsamen Steigung. Das Modell lautet mit dieser Reparametrisierung

$$y_{ij} = \mu + \beta x_{ij} + \tau_i + \delta_i x_{ij} + e_{ij}$$

Die Nullhypothese der Parallelität kann formuliert werden als

$$H_0: \delta_1 = \delta_2 = \dots = \delta_t = 0$$

Man beachte, dass die δ -Effekte die Rolle eines Lack-of-fit Effektes spielen.

12.2 Varianzanalyse

Für die weitere Entwicklung ist es hilfreich, die folgende hierarchische Sequenz von Modellen zu betrachten:

Modell	SQ_{Fehler}	FG_{Fehler}
(0) $y_{ij} = \mu + e_{ij}$	$SQ_{Fehler}^{(0)}$	$n - 1$
(1) $y_{ij} = \mu + \beta x_{ij} + e_{ij}$	$SQ_{Fehler}^{(1)}$	$n - 2$
(2) $y_{ij} = \mu + \beta x_{ij} + \tau_i + e_{ij}$	$SQ_{Fehler}^{(2)}$	$n - t - 1$
(3) $y_{ij} = \mu + \beta x_{ij} + \tau_i + \delta_i x_{ij} + e_{ij}$	$SQ_{Fehler}^{(3)}$	$n - 2t$

$n = \sum_i r_i$ = Gesamtzahl der Beobachtungen;

r_i = Zahl der Wiederholungen je Behandlung

Wichtig ist, dass wir in der Sequenz den Term für die Kovariable βx_{ij} vor dem Behandlungseffekt τ_i anpassen. Nur so erhalten wir eine um den Einfluss der Kovariable bereinigte Summe der Quadrate für die Behandlungseffekte.

Für die Berechnung der SQ gibt es explizite Formeln (siehe z.B. Dean und Voss, 1998), die aber hier nicht angegeben werden. Wir gehen davon aus, dass ein Computerprogramm zur Verfügung steht, dass diese Berechnungen durchführt.

Wir verwenden wieder das Prinzip des Modellaufbaus. Modell (1) ist ein Sonderfall von Modell (2) mit $\tau_1 = \tau_2 = \dots = \tau_t = 0$. Daher können wir

$$H_0: \tau_1 = \tau_2 = \dots = \tau_t = 0$$

durch einen Vergleich der SQ_{Fehler} der beiden Modelle (1) und (2) prüfen.

Die Nullhypothese paralleler Regressionslinien

$$H_0: \delta_1 = \delta_2 = \dots = \delta_t = 0$$

kann durch den Vergleich der Modelle (2) und (3) geprüft werden. Dieser Test ist als erster durchzuführen, um zu prüfen, ob die Voraussetzung für die Kovarianzanalyse gegeben ist: Parallelität der Regressionsgeraden. Die Varianzanalyse-Tabelle hat folgende Form:

Ursache	FG	SQ	MQ
β (Kovariable)	1	$SQ(\beta \mu) = SQ_{Fehler}^{(0)} - SQ_{Fehler}^{(1)}$	$SQ(\beta \mu)/1$
τ_i (Behandlungen)	$t - 1$	$SQ(\tau_i \beta, \mu) = SQ_{Fehler}^{(1)} - SQ_{Fehler}^{(2)}$	$SQ(\tau_i \beta, \mu)/(t-1)$
δ_i (Lack of fit)	$t - 1$	$SQ(\delta_i \tau_i, \beta, \mu) = SQ_{Fehler}^{(2)} - SQ_{Fehler}^{(3)}$	$SQ(\delta_i \tau_i, \beta, \mu)/(t-1)$
Fehler	$n - 2t$	$SQ_{Fehler}^{(3)}$	$SQ_{Fehler}^{(3)}/(n - 2t)$

Man beachte, dass sich die Freiheitsgrade für eine Streuungsursache wie üblich aus der Differenz der Fehler-FG der jeweiligen beiden Modelle in der oben gezeigten Modell-Sequenz ergeben. Die Zahl der Fehler-FG eines Modells entspricht dabei der Zahl der Beobachtungen, abzüglich der Zahl der freien Parameter des Modells.

Modell	FG_{gesamt}	FG_{Modell}	FG_{Fehler}	Reduktion [§] FG_{Fehler}
(0) $y_{ij} = \mu + e_{ij}$	n	1	$n-1$	
(1) $y_{ij} = \mu + \beta x_{ij} + e_{ij}$	n	2	$n-2$	1
(2) $y_{ij} = \mu + \beta x_{ij} + \tau_i + e_{ij}$	n	$t+1$	$n-t-1$	$t-1$
(3) $y_{ij} = \mu + \beta x_{ij} + \tau_i + \delta_i x_{ij} + e_{ij}$	n	$2t$	$n-2t$	$t-1$

§ Reduktion von FG_{Fehler} in Modellsequenz. Entspricht FG des in der Modellsequenz hinzugefügten Modelleffektes.

So hat Modell (2) zwar $(t+2)$ Parameter. Allerdings ist für den Behandlungseffekt τ_i wie generell in varianzanalytischen Modellen eine Parameterrestriktion einzuführen, so dass ein FG abzuziehen ist; die Fehler-FG sind daher $(n-t-1)$, und nicht $(n-t-2)$. Die Differenz der FG zum vorangegangenen Modell (1) beträgt demnach $(t-1)$, und dies sind die Behandlungs-FG. Modell (3) hat im wesentlichen t Steigungen und t Achsenanschnitte, also $2t$ Parameter, so dass die Fehler-FG $(n-2t)$ betragen. Die Differenz der FG zu Modell (2) beträgt

$$(n - 2t) - (n - t - 1) = (t - 1)$$

Alle Modelleffekte werden gegen den Restfehler getestet, also gegen

$$MQ_{Fehler} = SQ_{Fehler}^{(3)} / (n - 2t):$$

Kovariablen:
$$F_{vers} = \frac{SQ(\beta | \mu)}{SQ_{Fehler}^{(3)} / (n - 2t)}$$

Behandlungen:
$$F_{vers} = \frac{SQ(\tau_i | \beta, \mu) / (t - 1)}{SQ_{Fehler}^{(3)} / (n - 2t)}$$

Lack of fit:
$$F_{vers} = \frac{SQ(\delta_i | \tau_i, \beta, \mu) / (t - 1)}{SQ_{Fehler}^{(3)} / (n - 2t)}$$

Beispiel: In einem Fütterungsversuch mit Schweinen wurden vier Futterarten an je zehn Tiere verfüttert, um den Einfluss auf die tägliche Zunahme (y) zu untersuchen (Snedecor und Cochran, 1967, S.440). Außerdem wurde das Anfangsgewicht (x) festgehalten. Es soll ein F-Test auf Behandlungsunterschiede durchgeführt werden, wobei das Anfangsgewicht als Kovariable benutzt wird. Die Daten sind in der folgenden Tabelle wiedergegeben.

Futter 1		Futter 2		Futter 3		Futter 4	
<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>
61	1,40	74	1,61	80	1,67	62	1,40
59	1,79	75	1,31	61	1,41	55	1,47
76	1,72	64	1,12	62	1,73	62	1,37
50	1,47	48	1,35	47	1,23	43	1,15
61	1,26	62	1,29	59	1,49	57	1,22
54	1,28	42	1,24	42	1,22	51	1,48
57	1,34	52	1,29	47	1,39	41	1,31
45	1,55	43	1,43	42	1,39	40	1,27
41	1,57	50	1,29	40	1,56	45	1,22
40	1,26	40	1,26	40	1,36	39	1,36

Wir finden wir folgende Sequenz:

Modell	SQ_{Fehler}	$SQ(\text{Parameter})$
(0) $y_{ij} = \mu + e_{ij}$	$SQ_{Fehler}^{(0)} = 1,0228$	
(1) $y_{ij} = \mu + \beta x_{ij} + e_{ij}$	$SQ_{Fehler}^{(1)} = 0,8731$	$SQ(\beta \mu) = 0,1497$ $SQ(\tau_i \beta, \mu) = 0,1688$ $SQ(\delta_i \tau_i, \beta, \mu) = 0,0253$
(2) $y_{ij} = \mu + \beta x_{ij} + \tau_i + e_{ij}$	$SQ_{Fehler}^{(2)} = 0,7043$	
(3) $y_{ij} = \mu + \beta x_{ij} + \tau_i + \delta_i x_{ij} + e_{ij}$	$SQ_{Fehler}^{(3)} = 0,6790$	

Damit ergibt sich folgende Varianzanalyse-Tabelle:

Ursache	<i>FG</i>	<i>SQ</i>	<i>MQ</i>	F_{Vers}	§p-Wert
β (Kovariable)	1	0,1497	0,1497	7,06	0,0122
τ_i (Behandlungen)	3	0,1688	0,0563	2,65	0,0654
δ_i (Lack-of-fit)	3	0,0253	0,0084	0,40	0,7556
Fehler	32	0,6790	0,0212		

§ Siehe Anhang D

Der Term δ_i ist nicht signifikant [$F_{Vers} = 0,40 < F_{Tab}(1-\alpha; t-1 = 3, n-2t = 32) = 2,90$ bei $\alpha = 5\%$], also wird die Nullhypothese der Parallelität der Regressionslinien beibehalten. Daher können wir nun den Test der Behandlungen betrachten. Da die Kovariable vor den Behandlungen angepasst wird, testet der F-Test für τ_i die adjustierten Mittelwerte. Die Behandlungseffekte sind bei $\alpha = 5\%$ nicht signifikant [$F_{Vers} = 2,65 < F_{Tab}(1-\alpha; t-1 = 3, n-2t = 32) = 2,90$], obwohl die Signifikanz mit einem p -Wert von 0,0654 fast erreicht wird. Wir werten dieses Ergebnis als leichten Hinweis auf Behandlungsunterschiede.

12.3 Mittelwertvergleiche

Adjustierte Mittelwerte können berechnet werden, wenn der Lack-of-fit (δ_i) nicht signifikant ist. In diesem Fall können wir mit dem reduzierten Modell

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + e_{ij}$$

rechnen. Adjustierte Mittelwerte können mit den allgemeinen Methoden aus Abschnitt 6.8 verglichen werden. Hiermit ergeben sich die folgenden Resultate.

Die sog. korrigierten Mittelwerte (adjustierte Mittelwerte, Kleinst-Quadrat-Mittelwerte) sind Schätzwerte des Erwartungswertes an der Stelle $\bar{x}_{..}$:

$$\eta_i(\bar{x}_{..}) = \mu + \tau_i + \beta \bar{x}_{..}$$

Man beachte, dass Differenzen zwischen Erwartungswerten nur von den Behandlungseffekten abhängen. So gilt z.B.

$$\eta_1(\bar{x}_{..}) - \eta_2(\bar{x}_{..}) = \mu + \tau_1 + \beta \bar{x}_{..} - (\mu + \tau_2 + \beta \bar{x}_{..}) = \tau_1 - \tau_2$$

Die Kleinstquadratschätzung der Parameter erhält man wie üblich durch Minimierung der Summe der Abweichungsquadrate

$$SQ_{Fehler}^{(2)} = \sum_{i=1}^t \sum_{j=1}^{r_i} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta} x_{ij})^2$$

was mit den allgemeinen Methoden aus Abschnitt 6.8 erfolgen kann. Im vorliegenden Fall ergeben sich relativ einfache Rechenformeln.

Es kann gezeigt werden, dass Einsetzen der Kleinstquadratlösungen für die Parameter zu der folgenden Formel für die korrigierten Mittelwerte führt:

$$\hat{\eta}_i(\bar{x}_{..}) = \hat{\mu} + \hat{\tau}_i + \hat{\beta} \bar{x}_{..} = \bar{y}_{i.} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..})$$

$$\text{mit } \hat{\beta} = CP_{xy} / SQ_x, \quad SQ_x = \sum_{i=1}^t \sum_{j=1}^{r_i} (x_{ij} - \bar{x}_{i.})^2 \quad \text{und} \quad CP_{xy} = \sum_{i=1}^t \sum_{j=1}^{r_i} (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.})$$

CP = Summe der Kreuzprodukte (cross products)

Beispiel: Für das hypothetische Beispiel finden wir die unkorrigierten Mittelwerte

Futter 1: $\bar{y}_{1.} = 0,81$

Futter 2: $\bar{y}_{2.} = 0,93$

Außerdem ist

$$\bar{x}_{1\bullet} = 28,333 ; \bar{x}_{2\bullet} = 64,500 ; \bar{x}_{\bullet\bullet} = 46,417$$

Die KleinstquadratLösung für den Regressionskoeffizienten ist

$$\hat{\beta} = 0,0082154$$

Damit sind die korrigierten Mittelwerte gegeben durch

$$\hat{\eta}_1(\bar{x}_{\bullet\bullet}) = \bar{y}_{1\bullet} - \hat{\beta}(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet}) = 0,81 - 0,0082154 \times (28,333 - 46,417) = 0,96$$

$$\hat{\eta}_2(\bar{x}_{\bullet\bullet}) = \bar{y}_{2\bullet} - \hat{\beta}(\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet}) = 0,93 - 0,0082154 \times (64,500 - 46,417) = 0,78$$

Man beachte, dass Futter 2 den besseren unkorrigierten, aber den schlechteren korrigierten Mittelwert hat.

Eine Differenz zwischen den Behandlungen u und s wird geschätzt durch

$$\hat{\eta}_u(\bar{x}_{\bullet\bullet}) - \hat{\eta}_s(\bar{x}_{\bullet\bullet}) = \hat{\tau}_u - \hat{\tau}_s$$

Für eine paarweise Differenz gilt

$$\text{var}(\hat{\eta}_u(\bar{x}_{\bullet\bullet}) - \hat{\eta}_s(\bar{x}_{\bullet\bullet})) = \text{var}(\hat{\tau}_u - \hat{\tau}_s) = \sigma^2 \left(\frac{1}{r_u} + \frac{1}{r_s} + \frac{(\bar{x}_{u\bullet} - \bar{x}_{s\bullet})^2}{SQ_x} \right)$$

$$\text{mit } SQ_x = \sum_{i=1}^t \sum_{j=1}^{r_i} (x_{ij} - \bar{x}_{i\bullet})^2.$$

Wir schätzen σ^2 durch

$$\hat{\sigma}^2 = MQ_{Fehler}^{(2)} = \frac{SQ_{Fehler}^{(2)}}{(n-t-1)}$$

Diese Formel zeigt, dass die Varianz einer Differenz am kleinsten ist, wenn die Behandlungen den selben Mittelwert für die Kovariable aufweisen, weil dann

$(\bar{x}_{u\bullet} - \bar{x}_{s\bullet})^2 = 0$ ist. Dies bedeutet für die Versuchsplanung, dass man möglichst ähnliche Mittelwerte der Kovariablen für die Behandlungen anstreben sollte, falls dies möglich ist.

Beispiel: In einem Fütterungsversuch ist anzustreben, dass die Tiergruppen für die verschiedenen Fütterungsvarianten möglichst ähnliche durchschnittliche Anfangsgewichte haben.

Test zum Vergleich zweier adjustierter Mittelwerte

$$H_0: \eta_u(\bar{x}_{\bullet\bullet}) - \eta_s(\bar{x}_{\bullet\bullet}) = \tau_u - \tau_s$$

(1) Berechne

$$t_{Vers} = \frac{|\hat{\eta}_u(\bar{x}_{..}) - \hat{\eta}_s(\bar{x}_{..})|}{\sqrt{\text{var}(\hat{\eta}_u(\bar{x}_{..}) - \hat{\eta}_s(\bar{x}_{..}))}}$$

(2) Bestimme $t_{Tab}(n-t-1, \alpha)$

(3) Falls $t_{Vers} > t_{Tab}$, verwirfe H_0 .

Da die Varianz einer Differenz nicht konstant ist, kann keine gemeinsame Grenzdifferenz berechnet werden.

Dieser Test hält die vergleichsbezogene Irrtumswahrscheinlichkeit ein. Vom Tukey-Verfahren glaubt man, dass es die versuchsbezogene Irrtumswahrscheinlichkeit einhält, aber dafür gibt es bisher keinen Beweis.

Beispiel: Für den Fütterungsversuch wollen wir alle paarweisen Vergleiche der vier Fütterungsvarianten durchführen und dabei eine vergleichsbezogene Irrtumswahrscheinlichkeit von $\alpha = 5\%$ einhalten.

Die unkorrigierten Mittelwerte sind:

$$\bar{y}_{1.} = 1,464; \quad \bar{y}_{2.} = 1,319; \quad \bar{y}_{3.} = 1,445; \quad \bar{y}_{4.} = 1,325;$$

Außerdem finden wir:

$$\bar{x}_{1.} = 54,4; \quad \bar{x}_{2.} = 55,0; \quad \bar{x}_{3.} = 52,0; \quad \bar{x}_{4.} = 49,5;$$

$$\bar{x}_{..} = 52,725; \quad \hat{\beta} = 0,005376$$

Hiermit berechnen wir die korrigierten Mittelwerte:

$$\hat{\eta}_1(\bar{x}_{..}) = \bar{y}_{1.} - \hat{\beta}(\bar{x}_{1.} - \bar{x}_{..}) = 1,464 - 0,005376 \times (54,4 - 52,725) = 1,455$$

$$\hat{\eta}_2(\bar{x}_{..}) = 1,306$$

$$\hat{\eta}_3(\bar{x}_{..}) = 1,449$$

$$\hat{\eta}_4(\bar{x}_{..}) = 1,342$$

Diese unterscheiden sich nur geringfügig von den unkorrigierten Mittelwerten. Mit $SQ_x = 5065,975$ und $MQ_{Fehler}^{(2)} = 0,0201$ können wir die Standardfehler der Differenzen berechnen. Für den Vergleich der Behandlungen 1 und 2 erhalten wir z.B.

$$\text{var}(\hat{\tau}_1 - \hat{\tau}_2) = MQ_{Fehler}^{(2)} \left(\frac{1}{r_1} + \frac{1}{r_2} + \frac{(\bar{x}_{1.} - \bar{x}_{2.})^2}{SQ_x} \right) = 0,0201 \times \left(\frac{1}{10} + \frac{1}{10} + \frac{(54,4 - 55,0)^2}{5065,975} \right) = 0,00402$$

und einen Standardfehler von $\sqrt{0,00402} = 0,0634$. Hiermit wird

$$t_{Vers} = \frac{1,455 - 1,306}{\sqrt{0,00402}} = 2,34$$

Da $t_{Vers} > t_{Tab}(35, \alpha = 5\%) = 2,030$, ist die Differenz signifikant von Null verschieden.

Analog berechnet man die Vertrauensintervalle für die anderen Differenzen. Die mit der SAS Prozedur erhaltenen Vertrauensintervalle für alle paarweisen Vergleiche sind im folgenden aufgeführt (SAS Output):

Least Squares Means

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	1.4550	0.04499	35	32.34	<.0001
trt	2	1.3068	0.04510	35	28.98	<.0001
trt	3	1.4489	0.04488	35	32.28	<.0001
trt	4	1.3423	0.04533	35	29.61	<.0001

Differences of Least Squares Means

Effect	trt	_trt	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	2	0.1482	0.06345	35	2.34	0.0253
trt	1	3	0.006097	0.06363	35	0.10	0.9242
trt	1	4	0.1127	0.06421	35	1.75	0.0881
trt	2	3	-0.1421	0.06373	35	-2.23	0.0322
trt	2	4	-0.03557	0.06441	35	-0.55	0.5843
trt	3	4	0.1066	0.06364	35	1.67	0.1030

Kovariablen darf nicht von den Behandlungen beeinflusst sein

Eine ganz wichtige Voraussetzung der Kovarianzanalyse zur Ausschaltung von Störgrößen ist, dass die Kovariable selbst nicht von den Behandlungsfaktoren beeinflusst ist. Dies setzt in der Regel voraus, dass die Kovariable gemessen werden kann, bevor die Behandlungen angewendet werden. Bei einem Fütterungsversuch wird zum Beispiel das Anfangsgewicht der Tiere gemessen, bevor die verschiedenen Fütterungsvarianten angewendet werden. Bei einem Feldversuch mit Mais zum Vergleich von verschiedenen Maislinien kann die Zahl der aufgelaufenen Pflanzen pro Parzelle als Kovariable verwendet werden. Allerdings setzt dies voraus, dass diese Zahl nicht vom Genotyp beeinflusst ist, dass also das Nicht-Auflaufen ausgesäter Körner nur von Faktoren abhängt, die nicht vom Genotyp der Pflanze beeinflusst werden. Dies gilt zum Beispiel für Umwelteinflüsse wie Insektenfrass oder die Tiefe der Ablage der Körner. Es kann aber auch sein, dass das Auflaufverhalten stark genetisch bedingt ist. In einem solchen Fall ist es nicht sinnvoll, die Zahl der aufgelaufenen Pflanzen als Kovariable zu verwenden, weil man dadurch genetische Unterschiede in der Zielvariable (z.B. Ertrag) „wegrechnen“ würde.

SAS Anweisungen

```
data;
input trt x y;
cards;
1 61 1.40
1 59 1.79
1 76 1.72
1 50 1.47
1 61 1.26
1 54 1.28
1 57 1.34
1 45 1.55
1 41 1.57
1 40 1.26
2 74 1.61
2 75 1.31
2 64 1.12
2 48 1.35
2 62 1.29
2 42 1.24
2 52 1.29
2 43 1.43
2 50 1.29
2 40 1.26
3 80 1.67
3 61 1.41
3 62 1.73
3 47 1.23
3 59 1.49
3 42 1.22
3 47 1.39
3 42 1.39
3 40 1.56
3 40 1.36
4 62 1.40
4 55 1.47
4 62 1.37
4 43 1.15
4 57 1.22
4 51 1.48
4 41 1.31
4 40 1.27
4 45 1.22
4 39 1.36
;
/*Kovarianzanalyse mit paarweisen Mittelwertvergleichen*/
proc mixed;
class trt;
model y= x trt/solution;
```

```
lsmeans trt/pdiff;
run;

/*Test auf Parallelitaet*/
proc glm;
class trt;
model y= x trt x*trt/solution;
run;
```

12.4 Ein soziologisches Beispiel

Beispiel (Nolan and Speed, 2000, Chapter 10): Die Child Care Health and Development Studies (CHDS) sind eine groß angelegte Studie in den USA zur Aufklärung von Faktoren, die den Gesundheitszustand von Babies beeinflussen. Wir betrachten hier zwei Faktoren, die das Geburtsgewicht beeinflussen. Das Geburtsgewicht (BWT, in Unzen = ounces, 1 Unze = 28,35 g) kann als Indikator für den Gesundheitszustand angesehen werden. Die betrachteten Einflussfaktoren sind das Raucherverhalten der Mutter (SMOKE=0: Nichtraucher; SMOKE=1: Raucher) sowie das Gewicht der Mutter in Pfund (WEIGHT). Wir verwenden einen Teildatensatz von 1236 männlichen Babys, die mindestens 28 Tage überlebt haben.

Man hatte im Vorfeld Hinweise darauf, dass Babys von Raucherinnen bei Geburt weniger wiegen als Babys von Nichtraucherinnen. Allerdings können Raucherinnen sich von Nichtraucherinnen durch andere Faktoren unterscheiden, die ebenfalls einen Einfluss auf das Geburtsgewicht haben, so z.B. das Gewicht der Mutter. Das Gesundheitsministerium hat 1989 folgende Feststellung getroffen: "Rauchen scheint einen größeren Einfluss auf das Geburtsgewicht zu haben als das Gewicht der Mutter". Mit Hilfe einer Kovarianzanalyse kann geprüft werden, ob das Gewicht und das Raucherverhalten der Mutter einen Einfluss auf das Geburtsgewicht haben. Außerdem kann, mit gewissen Einschränkungen, quantifiziert werden, welcher der beiden Faktoren wichtiger ist.

Tab. 12.1: Ausschnitt aus dem Datensatz über 1236 männliche Babys der Child Health and Development Studie.

	Baby						
	1	2	3	4	5	6	7
Geburtsgewicht (BWT)	120	113	128	123	108	136	138
Gewicht der Mutter (WEIGHT)	100	135	115	190	125	93	178
Raucherverhalten (SMOKE)	0	0	1	1	1	1	0

Die Daten sind in Abb. 12.1 graphisch dargestellt. Dort ist für die Babys rauchender Mütter und für Babys nichtrauchender Mütter die Regression des Geburtsgewichtes gegen das Gewicht der Mutter dargestellt. Eine Kleinst-Quadrat-Schätzung ergibt folgendes Resultat:

SMOKE=0 (Mutter raucht nicht): $BWT = 109,02 + 0,107 \times WEIGHT$
 SMOKE=1 (Mutter raucht): $BWT = 95,13 + 0,148 \times WEIGHT$

Es stellt sich nun die Frage, ob sich die Steigungen der Regressionen signifikant unterscheiden. Falls nicht, liegen keine **Wechselwirkungen** zwischen dem **qualitativen Einflussfaktor** "Rauchverhalten" und dem **quantitativen Einflussfaktor** "Gewicht der Mutter" vor. Es ist dann möglich, Differenzen im Geburtsgewicht bei Raucherinnen und Nichtraucherinnen unabhängig vom Gewicht der Mutter zu beurteilen. Ebenso kann der Einfluss des Gewichts der Mutter unabhängig vom Rauchverhalten interpretiert werden.

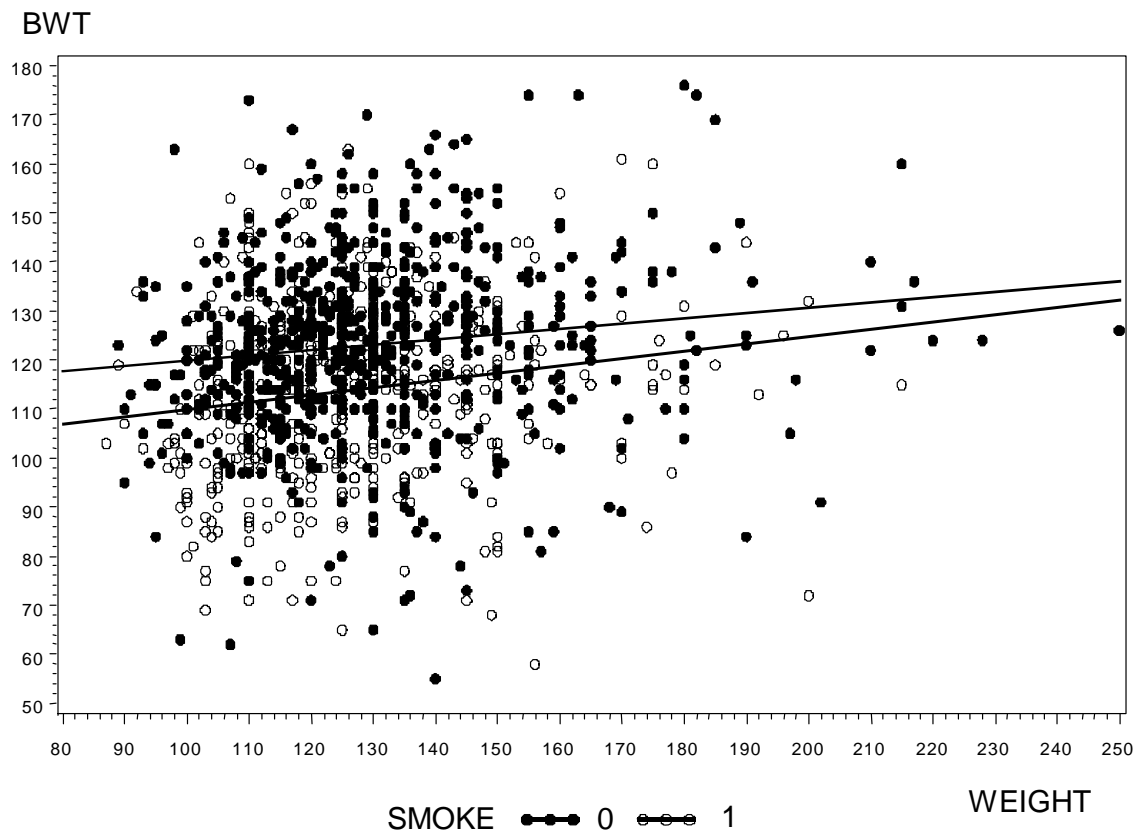


Abb. 12.1: Zwei lineare Regressionen des Geburtsgewichtes (BWT) gegen das Gewicht der Mutter (WEIGHT).

12.4.1 Modellierung

Wenn die Steigungen sich bei Raucherinnen und bei Nichtraucherinnen unterscheiden, lautet das Regressionsmodell:

$$y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij} \quad (12.1)$$

wobei

y_{ij} = Geburtsgewicht des j -ten Babys in der i -ten Gruppe
($i = 1$ für SMOKE = 0; $i = 2$ für SMOKE = 1)

β_i = Steigung für i -te Gruppe

x_{ij} = Gewicht der Mutter bei j -ten Baby in der i -ten Gruppe

e_{ij} = Zufallsabweichung von der Regression, $e_{ij} \sim N(0, \sigma^2)$

Wenn sich die Steigungen nicht unterscheiden, während die Achsenabschnitte verschieden sind, lautet das **reduzierte** Modell

$$y_{ij} = \alpha_i + \beta x_{ij} + e_{ij},$$

wobei

β = gemeinsame Steigung

Die Nullhypothese gleicher Steigungen lautet

$$H_0: \beta_1 = \beta_2 = \beta$$

Das reduzierte Modell ist ein Spezialfall des **vollen** Modells (12.1). Dies wird deutlicher durch folgende Reparametrisierung:

$$\beta_i = \beta + \delta_i$$

wobei

β = gemeinsame Steigung

δ_i = Abweichung von der gemeinsamen Steigung in der i -ten Gruppe ($i = 1, 2$),
Interaktionseffekt

Die Nullhypothese hat damit die äquivalente Form

$$H_0: \delta_1 = \delta_2 = 0$$

Außerdem ist folgende Reparametrisierung möglich:

$$\alpha_i = \mu + \tau_i$$

Hiermit kann das volle Modell reparametrisiert werden:

$$\begin{aligned} y_{ij} &= \mu + \tau_i + (\beta + \delta_i)x_{ij} + e_{ij} \\ &= \mu + \tau_i + \beta x_{ij} + \delta_i x_{ij} + e_{ij} \end{aligned}$$

12.4.2 Varianzanalyse und Modellselektion

Zum Test der Nullhypothese betrachten wir wie üblich eine Modellsequenz, an deren Ende das volle Modell steht (siehe 12.2):

Modell	SQ_{Fehler}	$SQ(\text{Parameter})$
(0) $y_{ij} = \mu + e_{ij}$	$SQ_{Fehler}^{(0)} = 399356$	$\begin{matrix} \nearrow \\ \nearrow \\ \nearrow \\ \longrightarrow \end{matrix} \begin{matrix} SQ(\beta \mu) = 9516 \\ SQ(\tau_i \beta, \mu) = 21204 \\ SQ(\delta_i \tau_i, \beta, \mu) = 198 \end{matrix}$
(1) $y_{ij} = \mu + \beta x_{ij} + e_{ij}$	$SQ_{Fehler}^{(1)} = 389840$	
(2) $y_{ij} = \mu + \beta x_{ij} + \tau_i + e_{ij}$	$SQ_{Fehler}^{(2)} = 368636$	
(3) $y_{ij} = \mu + \beta x_{ij} + \tau_i + \delta_i x_{ij} + e_{ij}$	$SQ_{Fehler}^{(3)} = 368438$	

Damit ergibt sich folgende Varianzanalyse-Tabelle:

Ursache	FG	SQ	MQ	F_{Vers}	$\S p\text{-Wert}$
β (WEIGHT)	1	9516	9516	30,63	<0,0001
τ_i (SMOKE)	1	21204	21204	68,26	<0,0001
δ_i (WEIGHT*SMOKE)	1	198	198	0,64	0,4250
Fehler	1186	368438	311		

\S siehe Anhang D

Der **Wechselwirkungsterm** ist nicht signifikant. Somit können wir auf Parallelität der Regressionsgeraden schließen. Außerdem können wir nun die **Hauptwirkungen** betrachten. Der F-Test für τ_i (SMOKE) ist signifikant und zeigt Unterschiede zwischen Raucherinnen und Nichtraucherinnen. Dieser Test ist um die Wirkung des Gewichts der Mutter (β) bereinigt, da dieser Effekt vor τ_i angepasst wurde. Der Test für das Gewicht der Mutter ist dagegen nicht um den Effekt des Raucherverhaltens bereinigt. Hierzu müssen wir SMOKE vor WEIGHT anpassen. Die Varianzanalyse ist wie folgt:

Modell	SQ_{Fehler}	$SQ(\text{Parameter})$
(0) $y_{ij} = \mu + e_{ij}$	$SQ_{Fehler}^{(0)} = 399356$	$\begin{matrix} \nearrow \\ \nearrow \\ \nearrow \\ \longrightarrow \end{matrix} \begin{matrix} SQ(\tau_i \mu) = 23014 \\ SQ(\beta \tau_i, \mu) = 7706 \\ SQ(\delta_i \tau_i, \beta, \mu) = 198 \end{matrix}$
(1) $y_{ij} = \mu + \tau_i + e_{ij}$	$SQ_{Fehler}^{(1)} = 376342$	
(2) $y_{ij} = \mu + \beta x_{ij} + \tau_i + e_{ij}$	$SQ_{Fehler}^{(2)} = 368636$	
(3) $y_{ij} = \mu + \beta x_{ij} + \tau_i + \delta_i x_{ij} + e_{ij}$	$SQ_{Fehler}^{(3)} = 368438$	

Damit ergibt sich folgende Varianzanalyse-Tabelle:

Ursache	FG	SQ	MQ	F	$\S p\text{-Wert}$
τ_i (SMOKE)	1	23014	23014	74,08	<0,0001
β (WEIGHT)	1	7706	7706	24,81	<0,0001
δ_i (WEIGHT*SMOKE)	1	198	198	0,64	0,4250
Fehler	1186	368438	311		

\S siehe Anhang D

Der Einfluss des Gewichts (Hauptwirkung) ist ebenfalls signifikant.

12.4.3 Schätzen des selektierten Modells

Aus der Varianzanalyse ergibt sich, dass folgendes **reduzierte** Modell die Daten adäquat beschreibt:

$$y_{ij} = \alpha_i + \beta x_{ij} + e_{ij} ,$$

Dies Modell umfasst nur noch die Hauptwirkungen, die Wechselwirkung ist weggelassen, da sie nicht signifikant war. Dies Modell ist in Abb. 12.2 dargestellt. Die Kleinst-Quadrat-Schätzung der Parameter ergibt:

$$\hat{\alpha}_1 = 107,1 \quad (\text{SMOKE}=0)$$

$$\hat{\alpha}_2 = 98,5 \quad (\text{SMOKE}=1)$$

$$\hat{\beta} = 0,122 \quad (\text{WEIGHT})$$

Die geschätzten Modelle für beide Gruppen lauten somit:

$$\text{SMOKE}=0 \text{ (Mutter raucht nicht):} \quad \text{BWT} = 107,1 + 0,122 \times \text{WEIGHT}$$

$$\text{SMOKE}=1 \text{ (Mutter raucht):} \quad \text{BWT} = 98,5 + 0,122 \times \text{WEIGHT}$$

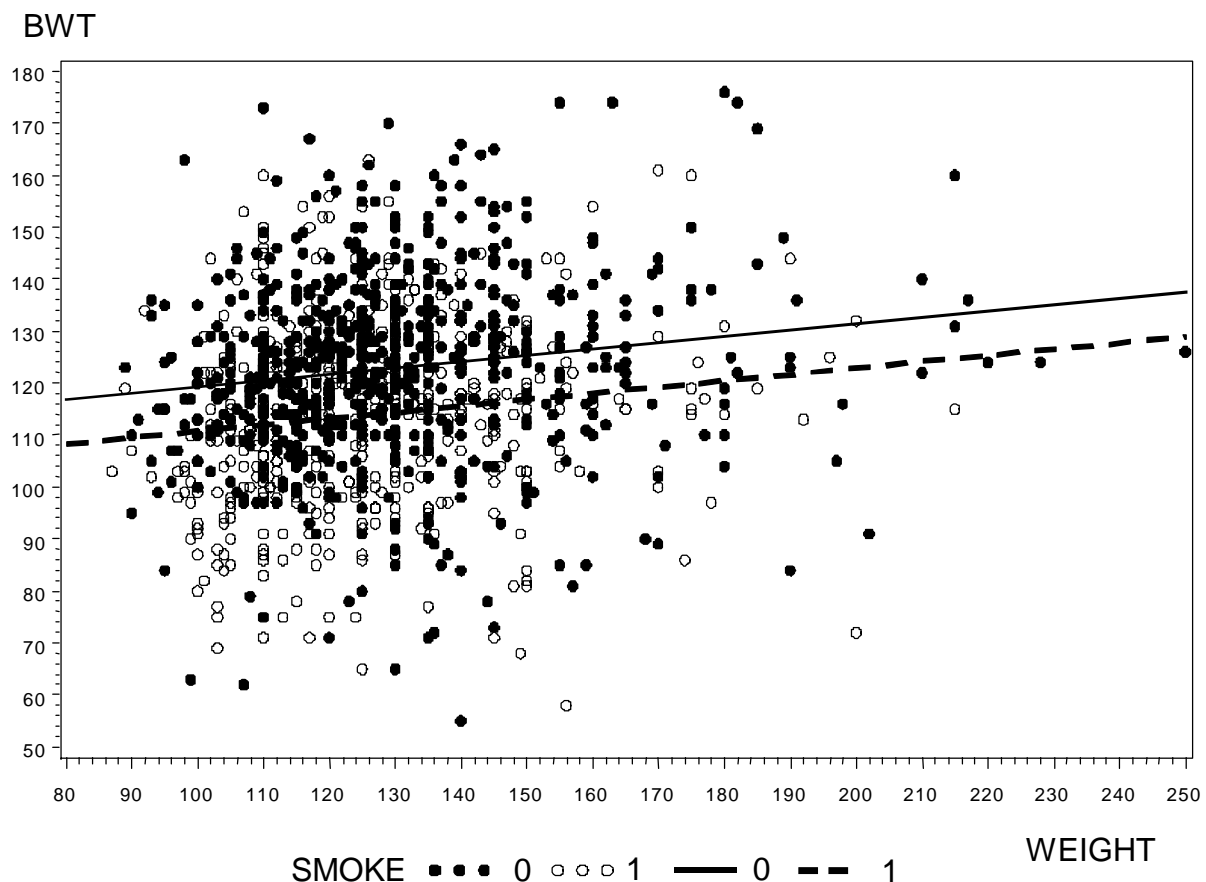


Abb. 12.2: Zwei lineare Regressionen des Geburtsgewichtes (BWT) gegen das Gewicht der Mutter (WEIGHT) bei gleicher Steigung.

Der Achsenabschnitt für Raucherinnen ist niedriger, also reduziert Rauchen das Geburtsgewicht der Babys. In anderen Worten: **wenn das Gewicht der Mutter**

konstant gehalten wird, ist das erwartete Geburtsgewicht bei Nichtraucherinnen 8,6 Unzen (ounces) über dem von Babys von Raucherinnen. Wegen der Abwesenheit von Wechselwirkungen gilt diese Aussage bei jedem Gewicht der Mutter.

Die gemeinsame Steigung von 0,122 bedeutet, dass **unabhängig vom Rauchverhalten der Mutter** eine Steigerung des Gewichts der Mutter um ein Pfund (pound) eine Erhöhung des erwarteten Geburtsgewichtes um 0,122 Unzen (ounces) zur Folge hat.

12.4.4 Welcher Faktor hat den größeren Einfluss?

Die Beantwortung dieser Frage ist nicht einwandfrei möglich. Wir haben hier dasselbe Problem wie bei der multiplen Regression, wo es ebenfalls unmöglich ist, zu entscheiden, welcher Faktor der wichtigste ist, sofern eine Korrelation unter den Einflussvariablen besteht (Multikollinearität). Man könnte an den größten F-Wert denken, um den wichtigsten Faktor zu bestimmen. Der "richtige" F-Wert für SMOKE (68,26) ist größer als der für WEIGHT (24,81). Aber der F-Wert alleine sagt nichts über die Wichtigkeit eines Faktors, sondern über die statistische Signifikanz. Dies sind zwei verschiedene Dinge. Für den qualitativen Faktor ist die Wirkung einfach über die Differenz der Achsenabschnitte zu quantifizieren. Diese Differenz beträgt 8,6 Unzen. Beim Gewicht der Mutter könnte man die Differenz des geschätzten Wertes für das Geburtsgewicht (BWT) beim niedrigsten Gewicht einer Mutter (etwa 85 Pfund) und beim höchsten Gewicht (etwa 250 Pfund) betrachten. Die Differenz ist

$$0,122 \cdot (250 - 85) = 20,13 \text{ Unzen}$$

Dies scheint darauf hinzudeuten, dass das Gewicht der Mutter wichtiger ist. Allerdings ist die Betrachtung nicht ganz fair, weil die beiden extremsten Gewichte herangezogen wurden. Alternativ könnten wir z.B. die Stichprobe am Median des Gewichtes der Mütter (125 Pfund) in zwei Gruppen ("leicht" und "schwer") teilen und die beiden Gruppenmediane errechnen.

Schwer: Median = 140 Pfund
Leicht: Median = 112 Pfund

Werten wir die Regression für diese beiden Werte aus und berechnen die Differenz, so finden wir:

$$0,122 \cdot (140 - 112) = 3,416 \text{ Unzen}$$

Dies wiederum deutet an, dass das Gewicht weniger wichtig ist als das Rauchverhalten.

Eine eindeutige Beurteilung der Wichtigkeit der beiden Faktoren ist nicht möglich.

Noch ein Beispiel

Abschließend noch ein Beispiel für die Anwendung der Kovarianzanalyse.

Beispiel (Prof. Wünsche, FG Obstbau, Uni Hohenheim): In Abschnitt 8.4 hatten wir die Blockbildung in Versuchen mit Obstbäumen betrachtet. Hierbei werden Bäume zu Blöcken gruppiert, so dass die Stammumfänge innerhalb eines Blocks möglichst ähnlich sind. Völlige Homogenität wird man aber nicht herstellen können. Daher lohnt es sich häufig, zusätzlich den Stammumfang als Kovariable ins Modell zu nehmen. Hiermit kann dann in der Regel noch ein Teil der Restheterogenität innerhalb der Blöcke erklärt und damit vom Restfehler abgetrennt werden, so dass die Genauigkeit der Auswertung steigt.

13. Messwiederholungen

In vielen Untersuchungen werden mehrere, oft viele Messungen, an ein und demselben Objekt gemacht. Die **wiederholten Messungen** (engl. *repeated measurements*) sind meistens über verschiedene Zeitpunkte verteilt. Bei der Auswertung von Messwiederholungen ist zu beachten, dass Messungen am selben Objekt zu verschiedenen Zeitpunkten nicht statistisch unabhängig sind.

Messwiederholungen verletzen daher die bisher fast durchweg gemachte Annahme der statistischen Unabhängigkeit der Fehler (Ausnahme: Abschnitt 9). Somit sind viele der bisher vorgestellten Verfahren nicht direkt auf die Rohdaten anwendbar.

Es gibt verschiedene statistische Verfahren, die für die Auswertung von Messwiederholungen geeignet sind. Die flexibelste Klasse von Verfahren fällt unter die Kategorie gemischter Modelle (*mixed models*) und wird in einer späteren Vorlesung behandelt.

Der einfachste Ansatz zur Auswertung von Messwiederholungen besteht in der Berechnung einer zusammenfassenden Maßzahl je Messreihe und Untersuchungseinheit (Randomisationseinheit). Hierdurch wird erreicht, dass je Untersuchungseinheit nur ein Wert vorliegt. Für die so erhaltenen Werte ist dann die Unabhängigkeitsannahme gültig, so dass eine Auswertung mit Standardverfahren erfolgen kann. Dieser Ansatz ist zwar nicht der effizienteste, aber der mit Abstand einfachste und daher in der Praxis am häufigsten verwendete. Zusätzlich kann für jeden Zeitpunkt getrennt eine varianzanalytische Auswertung erfolgen, da Messungen zu einem Zeitpunkt, die von verschiedenen Randomisationseinheiten stammen, statistisch unabhängig sind.

Beispiel: Drei Behandlungen (eine Kontrolle und zwei Chemikalien) wurden jeweils 10, 7 und 10 Ratten mit dem Futter verabreicht. Die Anlage wurde vollständig randomisiert, d.h. die Ratten wurden zufällig den Behandlungen zugewiesen. Die Gewichte der Tiere wurden zu fünf aufeinander folgenden Zeitpunkten gemessen (zu Beginn sowie 1, 2, 3 und 4 Wochen nach Beginn). Ziel der Untersuchung ist es, mögliche Unterschiede der Behandlungen hinsichtlich der Gewichtsentwicklung der Ratten zu ermitteln (aus Mead et al.: *Statistical methods for agriculture and biology*).

Tab. 13.1: Gewichtsdaten für Ratten bei 3 verschiedenen Behandlungen

	Woche					Differenz "4-0"	Regression <i>b</i>
	0	1	2	3	4		
Kontrolle							
Ratte 1	57	86	114	139	172	115	28,3
Ratte 2	60	93	123	146	177	117	28,7
Ratte 3	52	77	111	144	185	133	33,3
Ratte 4	49	67	100	129	164	115	29,2
Ratte 5	56	81	104	121	151	95	23,0
Ratte 6	46	70	102	131	153	107	27,5
Ratte 7	51	71	94	110	141	90	21,9
Ratte 8	63	91	112	130	154	91	22,1
Ratte 9	49	67	90	112	140	91	22,7
Ratte 10	57	82	110	139	169	112	28,1

	Woche					Differenz "4-0"	Regression <i>b</i>
	0	1	2	3	4		
Tyroxin							
Ratte 1	59	85	121	156	191	132	33,5
Ratte 2	54	71	90	110	138	84	20,7
Ratte 3	56	75	108	151	189	133	34,2
Ratte 4	59	85	116	148	177	118	29,9
Ratte 5	57	72	97	120	144	87	22,2
Ratte 6	52	73	97	116	140	88	21,9
Ratte 7	52	70	105	138	171	119	30,6

Thiouracil

Ratte 1	61	86	109	120	129	68	17,0
Ratte 2	59	80	101	111	126	67	16,5
Ratte 3	53	79	100	106	133	80	18,7
Ratte 4	59	88	100	111	122	63	14,9
Ratte 5	51	75	101	123	140	89	22,6
Ratte 6	51	75	92	100	119	68	16,1
Ratte 7	56	78	95	103	108	52	12,9
Ratte 8	58	69	93	114	138	80	20,5
Ratte 9	46	61	78	90	107	61	15,1
Ratte 10	53	72	89	104	122	69	17,0

Gewicht

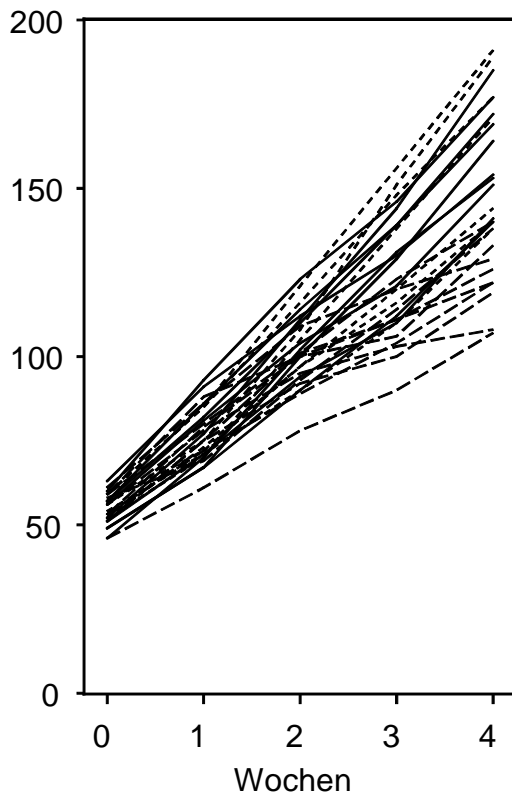


Abb. 13.1: Zeitliche Gewichtsverläufe für 27 Ratten bei drei verschiedenen Behandlungen. — Kontrolle Tyroxin - - - Thiouracil

Die Verlaufskurven in Abb. 13.1 zeigen, dass zu Beginn keine Unterschiede zu verzeichnen sind, während im späteren Verlauf deutlichere Unterschiede zu Tage treten.

Man könnte bei der Auswertung an eine Varianzanalyse mit den Faktoren Woche und Behandlung denken, bzw. an eine Kovarianzanalyse mit dem Faktor Woche als Kovariable, aber diese sind nicht zulässig, da Messungen am selben Tier Messwiederholungen sind, die miteinander korreliert sind. Die Daten verletzen die Annahme der Unabhängigkeit und Varianzhomogenität, wie im folgenden deutlich gemacht wird.

Ein zweifaktorielles Modell hat die Form

$$y_{ijk} = \eta_{ij} + e_{ijk}$$

wobei

η_{ij} = Erwartungswert der i -ten Behandlung in der j -ten Woche und

e_{ijk} = Zufallsabweichung des k -ten Tiers bei der i -ten Behandlung in der j -ten Woche

Üblicherweise wird man ein zweifaktorielles Modell für η_{ij} wählen, also

$$\eta_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

wobei

μ = allgemeiner Effekt

α_i = Haupteffekt der i -ten Behandlung

β_j = Haupteffekt der j -ten Woche

$(\alpha\beta)_{ij}$ = Wechselwirkung zwischen i -ter Behandlung und j -ter Woche

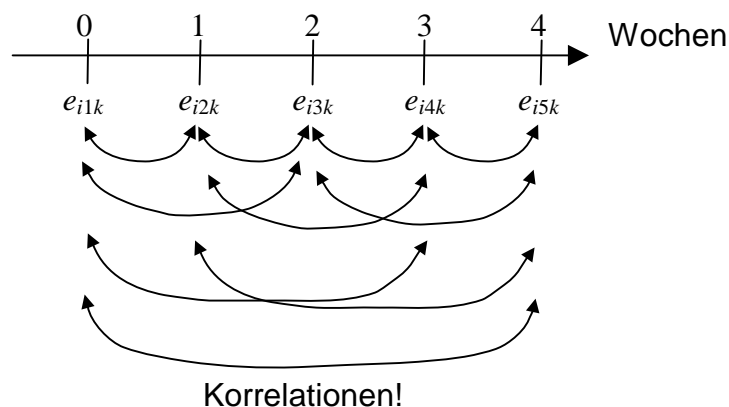


Abb. 13.2: Schematische Darstellung der Korrelationen zwischen Fehlern wiederholter Messungen am selben Tier.

Das Problem der Messwiederholungen besteht darin, dass die Fehler e_{ijk} beim selben Tier für verschiedene Zeitpunkte (j) korreliert sind. Für eine einfache

Varianzanalyse müsste für die Fehler des ik -ten Tieres zu den fünf verschiedenen Zeitpunkten ($e_{i1k}, e_{i2k}, e_{i3k}, e_{i4k}, e_{i5k}$) Unabhängigkeit und Varianzhomogenität gelten, also:

$$\text{var} \begin{pmatrix} e_{i1k} \\ e_{i2k} \\ e_{i3k} \\ e_{i4k} \\ e_{i5k} \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \sigma^2 \mathbf{I}$$

Obenstehender Ausdruck ist die Varianz-Kovarianz-Matrix der Fehler. Auf der Diagonalen stehen die Varianzen, jenseits der Diagonalen die Kovarianzen. Bei Unabhängigkeit sind die Kovarianzen Null. Bei Varianzhomogenität sind die Varianzen auf der Diagonale alle gleich, was bedeutet, dass zu jedem Zeitpunkt dieselbe Varianz gilt. Um zu untersuchen, wie sinnvoll diese Annahme ist, lassen wir nun von Null verschiedene Kovarianzen sowie Varianzheterogenität zu und betrachten das erweiterte Modell

$$\text{var} \begin{pmatrix} e_{i1k} \\ e_{i2k} \\ e_{i3k} \\ e_{i4k} \\ e_{i5k} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} & \sigma_{51} \\ \sigma_{12} & \sigma_2^2 & \sigma_{32} & \sigma_{42} & \sigma_{52} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{43} & \sigma_{53} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 & \sigma_{54} \\ \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_5^2 \end{pmatrix}$$

wobei

$$\sigma_j^2 = \text{var}(e_{ijk}) = E[e_{ijk}^2] = \text{Varianz zum } j\text{-ten Zeitpunkt}$$

$$\sigma_{jj'} = \text{cov}(e_{ijk}, e_{ij'k}) = E[e_{ijk} e_{ij'k}] = \text{Kovarianz der Zeitpunkte } j \text{ und } j'$$

Bei Unabhängigkeit gilt $\sigma_{jj'} = 0$ für alle $j \neq j'$. Bei Varianzhomogenität gilt $\sigma_j^2 = \sigma^2$ für alle j . Die Korrelation ist definiert als standardisierte Kovarianz (vgl. Abschnitt 6.2):

$$\rho_{jj'} = \text{corr}(e_{ijk}, e_{ij'k}) = \frac{\sigma_{jj'}}{\sqrt{\sigma_j^2 \sigma_{j'}^2}} = \frac{\sigma_{jj'}}{\sigma_j \sigma_{j'}}$$

Im Zusammenhang mit Messwiederholungen und Zeitreihendaten spricht man auch von **Autokorrelation**, weil hier ein Merkmal gewissermaßen mit sich selber korreliert ist, nämlich zu verschiedenen Zeitpunkten. Eine andere gebräuchliche Bezeichnung ist die **serielle Korrelation**. Mit der obigen Definition der Korrelation kann die Kovarianz wie folgt ausgedrückt werden:

$$\sigma_{jj'} = \rho_{jj'} \sigma_j \sigma_{j'}$$

Anstelle der Varianz-Kovarianz-Matrix können wir auch die Korrelationsmatrix berechnen:

$$\text{corr} \begin{pmatrix} e_{i1k} \\ e_{i2k} \\ e_{i3k} \\ e_{i4k} \\ e_{i5k} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{21} & \rho_{31} & \rho_{41} & \rho_{51} \\ \rho_{12} & 1 & \rho_{32} & \rho_{42} & \rho_{52} \\ \rho_{13} & \rho_{23} & 1 & \rho_{43} & \rho_{53} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 & \rho_{54} \\ \rho_{15} & \rho_{25} & \rho_{35} & \rho_{45} & 1 \end{pmatrix}$$

Die Parameter dieses Modells lassen sich leicht mit Hilfe einer Prozedur für gemischte Modelle schätzen, z.B. SAS PROC MIXED. Details sollen hier nicht besprochen werden. Die geschätzte Korrelation zwischen zwei Zeitpunkten j und j' , $\rho_{jj'} = \text{corr}(e_{ijk}, e_{ij'k})$, sowie die Varianzen, $\sigma_j^2 = \text{var}(e_{ijk})$, wurde wie folgt geschätzt (Korrelationsmatrix und Varianzen):

Woche	Woche					
	0	1	2	3	4	
0	1,0000	0,8575	0,6984	0,4694	0,3335	} Korrelationen
1	0,8575	1,0000	0,8557	0,5778	0,4222	
2	0,6984	0,8557	1,0000	0,8871	0,7760	
3	0,4694	0,5778	0,8871	1,0000	0,9409	
4	0,3335	0,4222	0,7760	0,9409	1,0000	
Varianzen	21,58	68,73	94,77	181,56	268,34	

Es bestehen deutliche Korrelationen zwischen den verschiedenen Zeitpunkten. Offensichtlich ist die Annahme der statistischen Unabhängigkeit der Fehler verletzt. Man sieht außerdem, dass die Korrelation umso enger ist, je näher die Zeitpunkte benachbart sind. Dieses Muster ist typisch für Zeitreihendaten. Außerdem steigt die Varianz von den frühen zu den späteren Zeitpunkten deutlich an, so dass auch die Annahme homogener Varianzen verletzt ist.

Zur Auswertung kann man zunächst für jeden Zeitpunkt eine einfache Varianzanalyse durchführen. Diese Auswertung ist gültig, da verschiedene Beobachtungen, die zum selben Zeitpunkt, aber bei verschiedenen Tieren gemacht wurden, unabhängig sind (Allerdings sind die Tests nicht unabhängig). Wir finden:

Woche=0:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	10.1857143	5.0928571	0.24	0.7916
Error	24	517.8142857	21.5755952		

Woche=1:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	36.542857	18.271429	0.27	0.7688
Error	24	1649.457143	68.727381		

Woche=2:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	601.394709	300.697354	3.17	0.0599
Error	24	2274.457143	94.769048		

Woche=3:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3563.605820	1781.802910	9.81	0.0008
Error	24	4357.357143	181.556548		

Woche=4:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9013.17884	4506.58942	16.79	<.0001
Error	24	6440.22857	268.34286		

Nach dieser Analyse sind ab der 3. Woche signifikante Unterschiede zu verzeichnen. Paarweise t-Tests liefern folgendes Ergebnis:

Behandlung	Woche				
	0	1	2	3	4
Kontrolle	54.0 a	78.5 a	106.0 a	130.1 b	160.6 b
Tyroxin	55.6 a	75.9 a	104.9 ab	134.1 b	164.3 b
Thiouracil	54.7 a	76.3 a	95.8 b	108.2 a	124.4 a

(Mittelwerte in einer Spalte, die mit demselben Buchstaben versehen sind, sind nicht signifikant verschieden; t-Test bei $\alpha = 5\%$)

Kontrolle und Tyrosin unterscheiden sich nicht, aber Thiouracil zeigt eine deutliche und signifikante Reduzierung des Wachstums ab der 2. Woche.

Nun zu zwei zusammenfassende Maßzahlen, die für die Fragestellung relevant sind. Der Anstieg des Gewichtes (siehe Abb. 13.1) kann einfach durch die Differenz zwischen der letzten und der ersten Woche quantifiziert werden (Tab. 13.1). Für diese Differenz führen wir eine einfache Varianzanalyse durch und finden:

DIFF "4-0"

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9000.81217	4500.40608	18.82	<.0001
Error	24	5739.92857	239.16369		

Der F-Test ist signifikant. Die Differenzen von Thiouracil zu den beiden anderen Behandlungen sind signifikant. Auch nach dieser Analyse ist diese Chemikalie als schädlich einzustufen.

Behandlung	Diff. "4-0"	Steigung	(Mittelwerte in einer Spalte, die mit demselben Buchstaben versehen sind, sind nicht signifikant verschieden; t-Test bei $\alpha = 5\%$)
Kontrolle	106.6 b	26.5 b	
Tyroxin	108.7 b	27.6 b	
Thiouracil	69.7 a	17.1 a	

Die Betrachtung der Daten in Abb.13.1 legt einen nahezu linearen Anstieg des Gewichtes nahe. Daher bietet sich die Steigung einer einfachen linearen Regression je Tier als weiteres zusammenfassendes Maß an. Wir berechnen also für jedes Tier eine Regression auf die Wochennummer (0, 1, 2, 3, 4). Die Regressionskoeffizienten sind in Tab. 13.1 wiedergegeben. Die Varianzanalyse der Steigungen ist wie folgt:

STEIGUNG:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	609.528714	304.764357	17.99	<.0001
Error	24	406.511286	16.937970		

Die Mittelwerte der Behandlungen unterscheiden sich signifikant, wie auch in der obigen Mittelwert-Tabelle ausgewiesen wird.

Die F-Werte für die Differenzen und für die Steigungen sind größer als bei getrennter Auswertung der Gewichte für jeden der fünf Zeitpunkte. Dies zeigt einen Gewinn an Genauigkeit an. Im Fall der Steigungen liegt dies in der Tatsache begründet, dass alle Daten berücksichtigt werden. Bei den Differenzen liegt der Gewinn an Genauigkeit in der unterschiedlichen Anfangsgewichte der Tiere begründet. Zusätzlich kann man eine Kovarianzanalyse durchführen, bei der die Differenz zwischen letzter und erster Woche als Zielvariable und das Anfangsgewicht als Kovariable verwendet wird (vgl. Kap. 12). Eine solche Analyse führt zu ähnlichen Ergebnissen:

WOCHE 4, adjustiert für Woche 0 (Kovarianzanalyse):

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Woche 0	1	16.146675	16.146675	0.06	0.8012
Behandlung	2	9000.743187	4500.371593	18.08	<.0001
Fehler	23	5723.85088	248.86308		

Behandlung	Adjustierter Mittelwert
Kontrolle	106.7 b
Tyroxin	108.6 b
Thiouracil	69.7 a

(Mittelwerte in einer Spalte, die mit demselben Buchstaben versehen sind, sind nicht signifikant verschieden; t-Test bei $\alpha = 5\%$)

Die Kovariable führt hier zu keinem Gewinn an Genauigkeit, die adjustierten Mittelwerte sind fast identisch mit den einfachen Mittelwerten für die Differenzen (DIFF).

Abschließend ist festzustellen, dass hier verschiedene Ansätze basierend auf zusammenfassenden Maßzahlen zu ähnlichen Ergebnissen führen. Solche Ansätze sind einer getrennten Auswertung für jeden Zeitpunkt vorzuziehen.

Nun noch zur Möglichkeit, über ein **gemischtes Modell** die serielle Korrelation sowie die Varianzheterogenität zwischen den Zeitpunkten zu berücksichtigen. Das Modell lautet, wie oben bereits erläutert:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

wobei die Fehler e_{ijk} beim selben Tier für verschiedene Zeitpunkte (j) korreliert sind nach einem sog. **unstrukturierten Modell** mit

$$\text{var} \begin{pmatrix} e_{i1k} \\ e_{i2k} \\ e_{i3k} \\ e_{i4k} \\ e_{i5k} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} & \rho_{51} \\ \sigma_{12} & \sigma_2^2 & \sigma_{32} & \sigma_{42} & \sigma_{52} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{43} & \sigma_{53} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 & \sigma_{54} \\ \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_5^2 \end{pmatrix}$$

Die Anweisungen für dieses Modell in der SAS Prozedur MIXED lauten:

```
proc mixed data=a;
class trt week;
model y=trt|week/ddfm=kr;
repeated week/sub=trtrat type=unr;
lsmeans trt/pdiff;
run;
```

Exemplarisch betrachten wir hier die paarweisen Vergleiche der Mittelwerte der Behandlungen über die Zeitpunkte.

Differences of Least Squares Means

Effect	trt	_trt	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	2	-1.1029	4.5590	24	-0.24	0.8109
trt	1	3	13.9600	4.1373	24	3.37	0.0025
trt	2	3	15.0629	4.5590	24	3.30	0.0030

Zum Vergleich passen wir ein Modell mit unabhängigen und varianz-homogenen Fehlern an:

Differences of Least Squares Means							
Effect	trt	_trt	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	2	-1.1029	2.4836	120	-0.44	0.6578
trt	1	3	13.9600	2.2538	120	6.19	<.0001
trt	2	3	15.0629	2.4836	120	6.06	<.0001

Zwar ist in beiden Fällen der Vergleich der Behandlungen 1 und 2 zur Behandlung 3 signifikant, wir sehen aber auch, dass die t-Werte (t_{Vers}) deutlich größer werden wenn die Korrelationen sowie die Varianzheterogenität ignoriert werden. Dies deutet darauf hin, dass bei Ignorieren der seriellen Korrelation für Messwiederholungen tendenziell zu viele Signifikanzen erzeugt werden, also der Fehler 1. Art über das nominelle (angestrebte) Niveau steigt. Dieses unterstreicht die Notwendigkeit, die serielle Korrelation auf jeden Fall zu berücksichtigen.

Beispiel: Bei der Prüfung von Pflanzenschutzmitteln wird die Wirkung der verschiedenen Mittel zu verschiedenen Zeitpunkten in Vegetationsverlauf erfasst, z.B. in Form des Anteils der befallenen Blattfläche oder des Anteils befallener Pflanzen. Der Befall wird in wiederholten Messungen auf derselben Versuchseinheit (Parzelle, Gefäß) durchgeführt, es liegen also Messwiederholungen vor. Ziel der Auswertung ist ein Vergleich der Wirksamkeit der verschiedenen Mittel und ein Vergleich der zeitlichen Verlaufskurven. Man könnte hier zur Auswertung an eine zweifaktorielle Varianzanalyse denken mit den Faktoren "Zeit" und "Mittel". Allerdings sind die in Kap. 10 vorgestellten Verfahren nicht anwendbar, weil die Unabhängigkeitsannahme verletzt ist. Stattdessen wird häufig die Zeitreihe einer Parzelle zu einer Maßzahl zusammengefasst, so dass je Parzelle nur noch ein einziger Wert vorliegt. Die resultierenden Daten können dann mittels einfaktorieller Varianzanalyse für den Faktor "Mittel" ausgewertet werden. Als zusammenfassende Maßzahl wird häufig die Fläche unter der Verlaufskurve berechnet als integrales Maß für die Befallsintensität bzw. -stärke (*engl. AUDPC = area under the disease progress curve*). Die AUDPC und der mittlere Befall wird wie folgt berechnet:

$$AUDPC = \sum_{j=1}^{w-1} \frac{1}{2} (B_j + B_{j+1}) * D_j$$

$$\text{Mittlerer Befall} = \frac{AUDPC}{\sum_{j=1}^{w-1} D_j}$$

B_j = Befall zum j -ten Messzeitpunkt ($j = 1, \dots, w$)

D_j = Zeitdauer vom j -ten zum $(j+1)$ -ten Messzeitpunkt

[vgl. L.V.Madden, G. Hughes, F. van den Bosch 2007 The study of plant disease epidemics. APS Press, St. Paul].

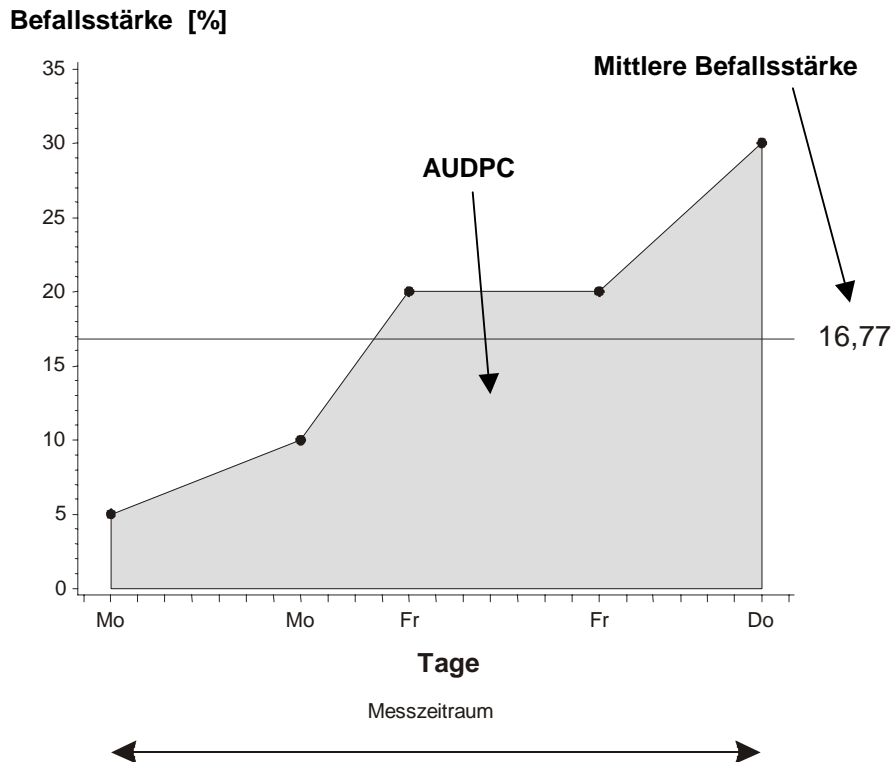


Abb. 13.3: Verlaufskurve für Befall (%) mit einem Pilz bei fünf Messzeitpunkten sowie der "area under the disease progress curve" (AUDPC).

Ein Beispiel zur Berechnung der AUDPC ist in Abb. 13.3 gegeben. Die Berechnung ist wie folgt:

Messung (j)	1	2	3	4	5
Tage	0	7	11	18	24
Zeitspanne (D_j)	7	4	7	6	
Befall (B_j)	5	10	20	20	30

$$AUDPC = \frac{5+10}{2} \cdot 7 + \frac{10+20}{2} \cdot 4 + \frac{20+20}{2} \cdot 7 + \frac{20+30}{2} \cdot 6 = 402,5$$

$$\text{Mittlerer Befall} = \frac{AUDPC}{\sum_{j=1}^{w-1} D_j} = \frac{402,5}{7+4+7+6} = 16,77$$

In einem Versuch zur Mittelprüfung können diese beiden Maßzahlen je Parzelle oder Gefäß ermittelt und dann varianzanalytisch verrechnet werden.

Beispiel: In Abschnitt 11.1 haben wir ein analoges Beispiel aus dem medizinischen Bereich besprochen. Dort wurde die Fläche unter der Verlaufskurve (AUC) für einen Blutparameter als integrales Maß für die Plasma-Renin Aktivität (PRA) analysiert.

Beispiel: Im Grünlandbereich werden oft Versuche mit mehreren Schnittzeitpunkten durchgeführt (meist 2 oder 3). Die Schnitte zu verschiedenen Zeitpunkten auf dersel-

ben Parzelle stellen Messwiederholungen dar, die korreliert sind. Mittelwerte und Summen über verschiedene Schnitte sind geeignete zusammenfassende Maßzahlen.

Beispiel: Im pflanzenbaulichen Bereich werden mehrjährige Versuche auf derselben Versuchsfläche durchgeführt, um verschiedene Fruchtfolgen und/oder Monokultursysteme (100-jähriger Roggenanbau in Halle) zu vergleichen. Hier liegen vieljährige Messwiederholungen auf derselben Parzelle vor, die untereinander korreliert sind. Die Untersuchung von zeitlichen Trends erfordert spezielle Verfahren, von denen die meisten unter die Rubrik der gemischten Modelle fallen.

Beispiel: Im ökonomischen Bereich sind oft große Zeitreihen zu analysieren mit dem Ziel, eine Prognose in die Zukunft zu machen (*forecasting*). In diesem Zusammenhang spricht man nicht mehr von Messwiederholungen, sondern von **Zeitreihenanalyse**. Für diesen Zweck gibt es eine Fülle verschiedener Verfahren, deren Gemeinsamkeit darin besteht, dass die Autokorrelation modelliert wird, um verlässlichere Prognosen zu erhalten. Ein hervorragendes Buch zum Einstieg in dieses Thema ist: Makridakis S, Wheelwright SC, Hyndman RJ 1998 Forecasting. Methods and applications. 3rd Edition. Wiley, New York.

Bisher haben wir Messwiederholungen in der Zeit betrachtet. Messwiederholungen können aber auch räumlich benachbarte Messungen auf derselben Randomisationseinheit sein.

Beispiel: Verschiedene Anbausysteme werden in einem Feldversuch verglichen hinsichtlich des Austrags von Stickstoff. Hierzu werden auf jeder Parzelle in drei verschiedenen Tiefen Bodenproben gezogen, in denen der N_{\min} -Gehalt bestimmt wird. Somit liegen je Parzelle drei Messwerte vor. Diese sind nicht unkorreliert, da sie auf derselben Randomisationseinheit erhalten wurden, und da verschiedene Bodentiefen nicht randomisiert werden können. Hier liegen die Messwiederholungen also im Raum vor und nicht in der Zeit. Es ist zu erwarten, dass benachbarte Bodentiefen höher korreliert sind als weiter entfernte (räumliche Autokorrelation). Eine statistische Auswertung muss der sich hieraus ergebenden räumlichen Korrelationsstruktur Rechnung tragen. Um Trendprofile im Boden zu vergleichen, bieten sich verschiedene zusammenfassende Maßzahlen an, z.B. lineare Trends oder einfache Differenzen ("oben minus unten"). Weiterführende Auswertung benötigen spezielle Verfahren, die in den Bereich der **räumlichen Statistik** oder **Geostatistik** fallen (Isaaks EH, Srivastava RM 1990 An introduction to geostatistics. OUP, Oxford; Webster R, Oliver MA 1990 Statistical Methods in Soil and Land Resource Survey. OUP, Oxford).

Abschließende Bemerkungen: Im Skript "Statistik" wird das Problem **verbundener Stichproben** behandelt. Hierbei werden zwei verschiedene Behandlungen oder Bedingungen jeweils auf derselben Versuchseinheit beobachtet. Zwei Beobachtungen von derselben Einheit sind miteinander korreliert. Es handelt sich bei verbundenen Stichproben daher um Messwiederholungen! Der "verbundene t-Test" berücksichtigt dies, indem Differenzen je Einheit berechnet werden. Diese Differenzen können als zusammenfassende Maßzahl aufgefasst werden. Unter der Nullhypothese erwarten wir eine Differenz von Null, und auf dieser Tatsache beruht dann der t-Test. Der verbundene t-Test ist auf zwei Messwiederholungen beschränkt, während wir in diesem Kapitel den Fall von mehr als zwei Messwiederholungen je Einheit betrachtet haben.

Es gibt natürlich auch Versuche, in denen die Zeit zwar ein Prüffaktor ist, aber keine wiederholten Messungen vorliegen. Das ist immer dann der Fall, wenn für jeden zu untersuchenden Zeitpunkt andere Versuchseinheiten (Randomisationseinheiten) verwendet werden. Um beispielsweise den Einfluß des Erntezeitpunktes auf die Qualität von Backweizen zu untersuchen, kann man einen Versuch mit vier verschiedenen Erntezeitpunkten als Blockanlage mit drei Wiederholungen anlegen. Zu jedem Zeitpunkt werden dann die drei Parzellen für den jeweiligen Zeitpunkt geerntet und auf ihre Qualität untersucht. Hier liegen keine Messwiederholungen vor, weil zu jedem Zeitpunkt andere Randomisationseinheiten (hier: Parzellen) untersucht werden.

14. Einführung in multivariate Verfahren

14.1. Zwei einführende Beispiele

Beispiel 1: Margarinendaten "semi-quantitativ" (Backhaus et al. 2000).

Ziele:

- Graphische Darstellung von Ähnlichkeiten zwischen Margarinemarken
- Gruppierung der Produkte
- Identifizierung produktspezifischer Eigenschaften
- Identifizierung wichtiger "Faktoren"

Datenstruktur:

- 30 Personen befragt
- 10 Merkmale, erfasst auf Linien-Skala (1-7)
- 11 Margarinemarken

Fragebogen (Ausschnitt, schematisch):

Beurteilen Sie bitte die Margarinemarke Rama anhand folgender Eigenschaften:							
Streichfähigkeit							
	1	2	3	4	5	6	7
Preis							
	1	2	3	4	5	6	7
usw. für die anderen acht Merkmale							

Personen markieren ihre Beurteilung auf der obigen 7er-Skala, wobei auch Zwischenstufen gewählt werden können. Die Daten können als "quantitativ" betrachtet werden.

Mittelwertbildung über die 30 Personen → Matrix: 11 Marken × 10 Eigenschaften.

Marke	STREI CHFÄH	PREIS	HALT BARK	UNGES FETTS	BACK EIGNG	GESCH MACK	KALOR IENGE	TIER FETT	VITAM INGEH	NATÜR LICHK
	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
Sanella	4.500	4.000	4.375	3.875	3.250	3.750	4.000	2.000	4.625	4.125
Homa	5.167	4.250	3.833	3.833	2.167	3.750	3.273	1.857	3.750	3.417
SB	5.069	3.824	4.765	3.438	4.235	4.471	3.765	1.923	3.529	3.529
Delicado	3.800	5.400	3.800	2.400	5.000	5.000	5.000	4.000	4.000	4.600
Hollbutt	3.444	5.056	3.778	3.765	3.944	5.389	5.056	5.615	4.222	5.278
Weihbutt	3.500	3.500	3.875	4.000	4.625	5.250	5.500	6.000	4.750	5.375
DuDarfst	5.250	3.417	4.583	3.917	4.333	4.417	4.667	3.250	4.500	3.583
Becel	5.857	4.429	4.929	3.857	4.071	5.071	2.929	2.091	4.571	3.786
Botteram	5.083	4.083	4.667	4.000	4.000	4.250	3.818	1.545	3.750	4.167
Flora	5.273	3.600	3.909	4.091	4.091	4.091	4.545	1.600	3.909	3.818
Rama	4.500	4.000	4.200	3.900	3.700	3.900	3.600	1.500	3.500	3.700

Beispiel 2: Margarinedaten qualitativ

Ziele wie in Beispiel 1

Die Daten sind **dichotom** oder **binär** (0-1). Ist das betreffende Merkmal vorhanden, wird eine 1 vergeben, andernfalls eine 0.

Marke	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
Sanella	1	0	1	1	1	0	1	1	1	0
Homa	1	0	0	1	1	1	0	1	0	1
SB	1	1	0	1	1	1	0	0	1	0
Delicado	0	0	1	1	0	0	1	0	1	0
Hollbutt	0	0	0	0	0	1	0	0	0	0
Weihbutt	0	0	0	0	1	0	1	0	0	1
DuDarfst	1	1	0	1	0	1	0	1	0	1
Becel	1	1	1	1	0	0	1	0	0	0
Botteram	0	0	1	1	1	1	0	0	0	1
Flora	1	1	1	1	1	0	1	0	1	0
Rama	1	0	1	1	1	1	1	1	1	0

V1 = Lagerzeit mehr als 1 Monat

V2 = Diätprodukt

V3 = Nationale Werbung

V4 = Becherverpackung

V5 = Pfundgröße

V6 = Verkaufshilfen

V7 = Eignung für Sonderangebote

V8 = Direktbezug vom Hersteller

V9 = Handelsspanne mehr als 20%

V10 = Beanstandungen im letzten Jahr

Bei Sanella beträgt z.B. die Lagerzeit mehr als 1 Monat (V1=1), aber es handelt sich nicht um ein Diätprodukt (V2=0).

14.2 Distanzen und Ähnlichkeiten

In beiden Beispielen ist ein Ziel die Gruppierung der Marken (allgemein: "Objekte") nach Ähnlichkeiten bzw. Distanzen, wobei die Distanzen/Ähnlichkeiten über die erhobenen Merkmale zu quantifizieren sind. Hier werden exemplarisch einige wichtige solche Maße beschrieben.

14.2.1 Euklidische Distanz

Das gängigste Distanzmaß ist die Euklidische Distanz. Diese lässt sich am einfachsten über den Satz des Pythagoras verstehen. Die Merkmale werden als Koordinaten in einem kartesischen Koordinatensystem aufgefasst. Jedes Produkt lässt sich dann als Punkt in diesem Koordinatensystem abbilden. Distanzen zwischen zwei Produkten (Margarinemarken) lassen sich als Abstand der Punkte der beiden Produkte darstellen. Falls nur zwei Merkmale vorliegen, haben wir zwei Koordinaten (x und y).

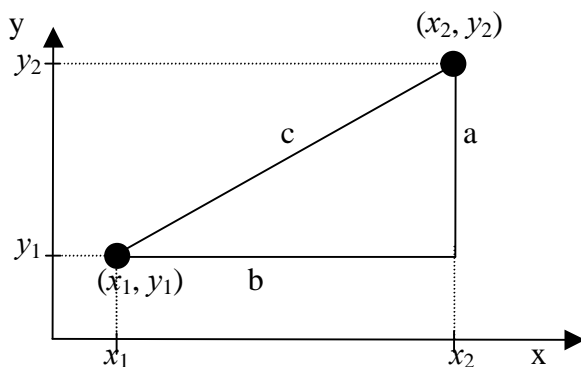


Abb. 14.2: Euklidischer Abstand zweier Produkte 1 und 2 bezüglich zweier Merkmale x und y .

Satz des Pythagoras: $c^2 = a^2 + b^2$

Euklidische Distanz (ED): $ED = c = \sqrt{a^2 + b^2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Diese Definition lässt sich auf beliebig viele Merkmale ausdehnen. An dieser Stelle ist es hilfreich, eine doppelte Indizierung einzuführen und nur einen Buchstaben (x) für verschiedene Merkmale zu verwenden. Ein Datensatz für n Objekte und p Merkmale hat folgende $n \times p$ Struktur:

	Merkmale			
	x_{11}	x_{12}	x_{1p}
	x_{21}	x_{22}	x_{2p}
Objekte (Produkte)	x_{31}	x_{32}	x_{3p}

	x_{n1}	x_{n2}	x_{np}

Hierbei ist

x_{ik} = Wert des k -ten Merkmals für das i -te Objekt (Produkt)

Es gilt dann für die Euklidische Distanz zwischen dem i -ten und dem j -ten Objekt:

$$ED = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

Oft wird auch die quadrierte Euklidische Distanz angegeben:

$$ED^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2 = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2$$

Beispiel 1: Der Euklidische Abstand zwischen Sanella und Homa ist

$$ED = \sqrt{(4,500 - 5,167)^2 + (4,000 - 4,250)^2 + \dots + (4,125 - 3,417)^2} = \sqrt{3,792} = 1,947$$

Die Euklidische Distanz ist ein Maß, das häufig für metrische Daten angewendet wird. Wenn die verschiedenen Merkmale ganz unterschiedliche Messniveaus haben, ist es sinnvoll, die Daten zu standardisieren, bevor die Distanz berechnet wird. Dies ist in Beispiel 1 deswegen nicht dringend angezeigt, weil alle Merkmale auf derselben Skala erhoben worden sind. Die üblichste Standardisierung ist

$$z_{ik} = \frac{x_{ik} - \bar{x}_{\bullet k}}{s_k},$$

wobei s_k die Standardabweichung und $\bar{x}_{\bullet k}$ der Mittelwert des k -ten Merkmals über alle Objekte ist. Die standardisierten Daten haben u.a. den großen Vorteil, dass sie dimensionslos sind.

Beispiel 3:

Unstandardisierte Daten:

Betrieb	Fläche (ha)	Einkommen (1.000 EUR)	Milchmenge (1.000 l)
1	120	75	12
2	110	44	0
3	90	34	44
4	150	39	178

Quadrierte Euklidische Distanzen (unstandardisiert):

Betrieb	2	3	4
1	1205	3605	29752
2		2436	33309
3			21581

Standardisierte Daten:

Betrieb	Fläche (ha)	Einkommen (1,000 EUR)	Milchmenge (1,000 l)
1	0,1	1,46285	-0,56844
2	-0,3	-0,21672	-0,71514
3	-1,1	-0,75851	-0,17726
4	1,3	-0,48762	1,46083

Quadrierte Euklidische Distanzen (standardisiert):

Betrieb	2	3	4
1	3,00	6,53	9,36
2		1,22	7,37
3			8,52

Wenn hier Euklidische Distanzen zwischen den Betrieben zu berechnen sind, ist eine Standardisierung dringend angezeigt. Dies überlegt man sich leicht, wenn man betrachtet, was eine Änderung der Maßeinheit zur Folge hat. So könnten wir die Betriebsfläche in m² anstatt in ha (1 ha = 10.000m²) ausdrücken, und die Milchmenge in Einheiten von 1.000.000 l:

Betrieb	Fläche (m ²)	Einkommen (1.000 EUR)	Milchmenge (1.000.000 l)
1	1200000	75	0,0012
2	1100000	44	0,0000
3	900000	34	0,0044
4	1500000	39	0,0178

Quadrierte Euklidische Distanzen ($\times 10^{-7}$) (unstandardisiert):

Betrieb	2	3	4
1	100	900	900
2		400	1600
3			3600

Standardisierte Daten:

Betrieb	Fläche (ha)	Einkommen (1,000 EUR)	Milchmenge (1,000 l)
1	0,1	1,46285	-0,56844
2	-0,3	-0,21672	-0,71514
3	-1,1	-0,75851	-0,17726
4	1,3	-0,48762	1,46083

Quadrierte Euklidische Distanzen (standardisiert):

Betrieb	2	3	4
1	3,00	6,53	9,36
2		1,22	7,37
3			8,52

Die standardisierten Daten bleiben dieselben. Für die unstandardisierten Daten werden jetzt die Distanzen völlig dominiert von den Betriebsflächen, da diese das mit Abstand größte Messniveau haben. Man beachte insbesondere, dass jetzt die Betriebe 1 und 2 sich am ähnlichsten sind, da sie die ähnlichsten Flächengrößen haben. Bei Standardisierung sind dagegen die Betriebe 2 und 3 am ähnlichsten.

Beispiel 1: Für die Margarinedaten sind die unstandardisierten Distanzen ausreichend, da auf derselben Messskala von 1 bis 7 gemessen wurde. Die Distanzen sind in Tab. 1 wiedergegeben. SB, Botteram, Flora und Rama weisen untereinander relativ geringe Distanzen auf. Dagegen ist Holländische Butter von diesen vier Marken sehr weit entfernt. Dies deutet auf eine Gruppierung der Marken hin, die allerdings nicht leicht aufzuspüren ist, wenn nur die Distanzmatrix herangezogen wird.

Tab. 14.1: Matrix der quadrierte Euklidischen Distanzen für elf Margarinemarken; Beispiel 1 (unstandardisierte Daten).

	1	2	3	4	5	6	7	8	9	10	11
1: Sanella	0	3,792	3,806	15,198	21,442	25,484	4,882	6,025	2,268	2,909	2,112
2: Homa		0	6,322	23,871	30,458	38,621	10,881	8,063	5,325	6,194	3,396
3: SB			0	14,151	24,971	28,933	3,998	3,471	1,099	2,361	1,725
4: Delicado				0	6,496	11,882	11,692	18,362	15,929	16,520	17,030
5: Holl. Butter					0	3,606	16,410	26,957	25,334	25,906	26,768
6: Weih. Butter						0	15,887	32,336	29,999	28,195	32,272
7: Dudarfst							0	6,422	5,156	3,825	6,932
8: Becel								0	3,395	6,376	6,022
9: Botteram									0	1,564	1,118
10: Flora										0	2,152
11: Rama											0

Tab. 14.2: Matrix der Simple-Matching Distanzen ($D = 1 - S$, wobei S der Simple-Matching-Koeffizient ist) für elf Margarinemarken; Beispiel 2.

	1	2	3	4	5	6	7	8	9	10	11
1: Sanella	0	0,5	0,5	0,7	0,2	0,4	0,3	0,6	0,4	0,8	0,9
2: Homa		0	0,6	0,2	0,5	0,5	0,8	0,3	0,7	0,3	0,6
3: SB			0	0,4	0,5	0,3	0,6	0,5	0,5	0,7	0,6
4: Delicado				0	0,5	0,5	0,2	0,7	0,5	0,7	0,6
5: Holl. Butter					0	0,6	0,5	0,4	0,6	0,2	0,3
6: Weih. Butter						0	0,3	0,4	0,6	0,4	0,3
7: Dudarfst							0	0,5	0,5	0,3	0,4
8: Becel								0	0,4	0,8	0,5
9: Botteram									0	0,4	0,5
10: Flora										0	0,7
11: Rama											0

14.2.2 Binäre Daten (0-1)

Man kann für binäre Daten ebenfalls eine Euklidische Distanz berechnen. Hier gibt es nur die Ausprägungen 0 und 1. Das folgende Beispiel erläutert den Rechengang für Beispiel 2 (Margarinemarken Sanella und Homa).

x_{1k} Sanella	1	0	1	1	1	0	1	1	1	0
x_{2k} Homa	1	0	0	1	1	1	0	1	0	1
$x_{1k}-x_{2k}$	0	0	1	0	0	-1	1	0	1	1
$(x_{1k}-x_{2k})^2$ §	0	0	1	0	0	1	1	0	1	1

$$ED^2 = 0 + 0 + 1 + 0 + 0 + 1 + 1 + 0 + 1 + 1 = 5$$

§ 0 : Merkmalsausprägung stimmt überein (*match*)

1 : Merkmalsausprägung stimmt nicht überein (*mismatch*)

Die quadrierte Euklidische Distanz entspricht bei binären Daten einfach der Zahl der nicht übereinstimmenden Merkmalsausprägungen (*mismatches*). Die Berechnung lässt sich auf alternative Weise vornehmen. Hierzu wird für ein Paar von Objekten (hier: Margarinemarken) eine 2×2 Kreuztabelle aufgestellt.

		Objekt j	
		1	0
Objekt i	1	a	b
	0	c	d

Hierbei ist

a = Anzahl der Merkmale, die sowohl für Objekt i als auch für Objekt j die Ausprägung 1 haben

b = Anzahl der Merkmale, die für Objekt i die Ausprägung 1 und für Objekt j die Ausprägung 0 haben.

c = Anzahl der Merkmale, die für Objekt i die Ausprägung 0 und für Objekt j die Ausprägung 1 haben.

d = Anzahl der Merkmale, die sowohl für Objekt i als auch für Objekt j die Ausprägung 0 haben.

Die quadrierte Euklidische Distanz lässt sich berechnen als

$$ED^2 = b + c$$

Dies entspricht der Zahl der Merkmale, für welche die beiden Objekte nicht übereinstimmen (*Mismatches*). Wenn man anstatt eines Distanzmaßes, also eines Maßes für die Unähnlichkeit, ein Ähnlichkeitsmaß berechnen will, so kann man beispielsweise die Zahl der Übereinstimmungen (*Matches*)

$$M = a + d$$

zugrundelegen. Dividiert man noch durch die Zahl der Merkmale ($a+b+c+d$), so erhält man den *Simple Matching* Koeffizienten:

$$S = \frac{a+d}{a+b+c+d} \quad (\text{Simple-Matching-Koeffizient})$$

Beispiel 2: Für die Distanzberechnung zwischen Homa und Sanella finden wir

		Sanella	
		1	0
Homa	1	$a = 4$	$b = 3$
	0	$c = 2$	$d = 1$

Somit ist

$$S = \frac{a+d}{a+b+c+d} = \frac{4+1}{4+3+2+1} = 0,5$$

und

$$ED^2 = b + c = 3 + 2 = 5$$

Basierend auf den Häufigkeiten a , b , c und d kann man eine ganze Reihe weiterer Ähnlichkeitsmaße berechnen. Die beiden neben dem Simple-Matching-Koeffizienten wichtigsten sind der *Jaccard-Koeffizient* und der *Dice-Koeffizient*:

$$S = \frac{a}{a+b+c} \quad (\text{Jaccard-Koeffizient})$$

$$S = \frac{2a}{2a+b+c} \quad (\text{Dice-Koeffizient})$$

Diese beiden Ähnlichkeitsmaße ignorieren die Zahl der 0-0 Matches (d), der sog. **negative matches**. Das kann sinnvoll sein, wenn eine Übereinstimmung im Fehlen eines Merkmals (0) nicht dieselbe Aussagekraft hat wie die Übereinstimmung im Vorhandensein eines Merkmals. Lautet in einer Erhebung mit einer Reihe von Personen das Merkmal "weiblich", so ist ein 0-0 Match zwischen zwei Personen (beide sind Männer) genau so aussagekräftig wie ein 1-1 Match (beide sind Frauen), ein sog. **positive match**. Lautet das Merkmal einer europaweiten Erhebung dagegen "Schwabe", so sind 1-1 Matches ein viel deutlicher Hinweis auf Ähnlichkeit als 0-0 Matches.

Viele der Ähnlichkeitsmaße liegen zwischen 0 und 1, wobei die größte Ähnlichkeit auftritt, wenn $S = 1$ ist, während die Ähnlichkeit bei $S = 0$ am geringsten ist. Aus den Ähnlichkeiten lassen sich Distanzen durch eine einfache Transformation berechnen, z.B.

$$D = 1 - S$$

Im Falle des Simple Matching Koeffizienten ist D gleich der standardisierten quadrierten Euklidischen Distanz, denn

$$1 - S = 1 - \frac{a+d}{a+b+c+d} = \frac{b+c}{a+b+c+d} = \frac{ED^2}{a+b+c+d}$$

Die Umwandlung von Ähnlichkeiten in Distanzen ist vor allem dann nötig, wenn eine graphische Darstellung der Distanzmatrix erstellt werden soll, wie bei der im folgenden beschriebenen Clusteranalyse.

14.2.3 Gemischte Daten

In vielen Fällen hat man sowohl kategoriale Daten als auch metrische Daten, für die man eine gemeinsame Distanz oder Ähnlichkeit berechnen möchte. Gower (1971: Biometrics 27, 857-872) hat vorgeschlagen, dass in solchen Fällen ein Ähnlichkeitsmaß wie folgt definiert werden kann:

$$\bar{S} = \frac{\sum_{k=1}^p w_k S_k}{\sum_{k=1}^p w_k}$$

wobei S_k ein Ähnlichkeitsmaß für das k -te Merkmal ist und w_k vom Nutzer zu wählende Gewichte. S_k ist dabei für verschiedene Datentypen wie folgt zu berechnen:

Binäre Daten: $S_k = 1$ bei einem *positive match* (1-1 match), sonst $S_k = 0$.

Kategoriale Daten: $S_k = 1$ bei einer Übereinstimmung der Ausprägungskategorie der beiden Objekte, sonst $S_k = 0$.

Metrische Daten: $S_k = 1 - |x_{ik} - x_{jk}| / R_k$,

wobei R_k die Variationsbreite des k -ten Merkmals ist und x_{ik} , x_{jk} die Merkmalswerte der zu vergleichenden Objekte i und j .

Die Gewichte w_k sind gleich 1, wenn das betreffende Merkmal bei beiden Objekten gemessen wurde, andernfalls ist $w_k = 0$. Im Fall fehlender Werte muss man für S_k einen beliebigen Wert einsetzen, z.B. $S_k = 0$, um die Ähnlichkeit berechnen zu können. Welcher Wert eingesetzt wird, hat keinen Einfluss auf die Ähnlichkeit, weil das Gewicht gleich 0 ist. Andere Gewichtungen sind möglich, wenn man den verschiedenen Merkmalen unterschiedlichen Einfluss geben will.

Beispiel: Eine Genbank hat für 4 Gersten-Genotypen (sog. Akzessionen) 3 verschiedene phänotypische Merkmale erfasst und will nun die Ähnlichkeit zwischen den Akzessionen bestimmen (Dies Beispiel ist zur Erläuterung: in Wirklichkeit handelt

es sich in der Regel um mehrere 100 oder 1000 Akzessionen und sehr viel mehr Merkmale).

Merkmalstyp	Merkmal	Akzession			
		1	2	3	4
Binär	Begrannung	ja	ja	nein	nein
Kategorial	Kornfarbe	braun	rot	beige	-
Metrisch	Tausendkorngewicht	30	24	56	31

Die Distanz der Akzessionen 1 und 2 berechnet sich wie folgt:

Merkmalstyp	Merkmal	S_k	w_k
Binär	Begrannung	1	1
Kategorial	Kornfarbe	0	1
Metrisch	Tausendkorngewicht	$1-6/32 = 0,8125$	1

$$\bar{S} = \frac{1 \times 1 + 1 \times 0 + 1 \times 0,8125}{1 + 1 + 1} = 0,6042$$

Die Distanz der Akzessionen 1 und 3 berechnet sich wie folgt:

Merkmalstyp	Merkmal	S_k	w_k
Binär	Begrannung	0	1
Kategorial	Kornfarbe	0	1
Metrisch	Tausendkorngewicht	$1-26/32 = 0,1875$	1

$$\bar{S} = \frac{1 \times 0 + 1 \times 0 + 1 \times 0,1875}{1 + 1 + 1} = 0,0625$$

Die Distanz der Akzessionen 1 und 4 berechnet sich wie folgt:

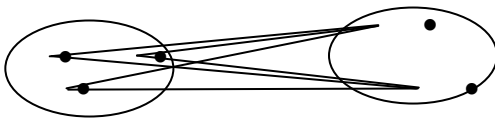
Merkmalstyp	Merkmal	S_k	w_k
Binär	Begrannung	0	1
Kategorial	Kornfarbe	0 (willkürlich)	0
Metrisch	Tausendkorngewicht	$1-1/32 = 0,96875$	1

$$\bar{S} = \frac{1 \times 0 + 0 \times 0 + 1 \times 0,96875}{1 + 0 + 1} = 0,48438$$

14.3 Clusteranalyse

Die Clusteranalyse ist ein Verfahren, mit dem eine Distanzmatrix in eine graphische Darstellung umgesetzt wird. Ziel dabei ist es, Gruppierungen der Objekte sichtbar zu machen. Die einfachste Klasse von Verfahren sind die hierarchischen agglomerativen Verfahren, bei denen zunächst jedes Objekt als eine eigene Gruppe (**Cluster**) betrachtet wird. Dann werden iterativ bestehende Gruppen fusioniert. Es werden bei jedem Fusionierungsschritt diejenigen beiden Gruppen zusammengefasst, die sich am ähnlichsten sind. Die Definition der Gruppenähnlichkeit kann auf verschiedene Weise erfolgen:

Average Linkage: Durchschnitt aller paarweisen Distanzen.



Single Linkage: Abstand der nächsten Objekte aus beiden Clustern.



Complete Linkage: Abstand der am weitesten entfernten Objekte aus beiden Clustern.



14.3.1 Average Linkage

Beim Average Linkage Verfahren wird das arithmetische Mittel aller paarweisen Distanzen zur Berechnung der Abstandes zweier Cluster herangezogen, wobei die Distanzen für solche Paare gemittelt werden, bei denen eine Objekt aus dem 1. Cluster kommt und das andere aus dem 2. Cluster.

Tab. 14.1: Hypothetische Datenmatrix für Margarinemarken 1 bis 4 und 10 dichotome Merkmale (a bis j). 0 = abwesend, 1 = anwesend.

	Merkmal									
	a	b	c	d	e	f	g	h	i	j
Marke										
1	1	1	1	0	1	1	1	1	0	1
2	0	1	0	1	1	1	1	1	0	0
3	0	1	1	1	1	0	1	0	1	0
4	1	1	1	0	1	1	0	0	1	1

Da hier binäre Daten vorliegen, kann z.B. der Dice-Koeffizient berechnet werden.

Dice Koeffizient S_{jk} :

	1	2	3	4
1	1,00			
2	0,71	1,00		
3	0,57	0,67	1,00	
4	0,80	0,46	0,62	1,00

Dice Distanz $D_{jk} = 1 - S_{jk}$:

	1	2	3	4
1	0			
2	0,29	0		
3	0,43	0,33	0	
4	0,20	0,54	0,38	0

Die kürzeste Distanz ist die zwischen Marke 1 und 4 ($D_{14} = 0,20$), also werden die Marken 1 und 4 zu einem Cluster zusammengefasst. Nach diesem ersten Schritt gibt es drei Cluster: (14), 2 und 3. Für diese Cluster wird eine neue Distanzmatrix berechnet. Die Distanz $D_{23} (= 0,33)$ ist unabhängig vom neuen Cluster (14) und bleibt daher unverändert. Die neue Distanz zwischen (14) und 2 ($D_{2(14)}$) wird aus dem Durchschnitt der Distanzen zwischen Marke 2 und jedem der Objekte des neuen Clusters (Marke 1 und Marke 4) berechnet:

$$D_{2(14)} = (D_{12} + D_{24})/2 = (0,29 + 0,54)/2 \sim 0,42$$

Ebenso wird die Distanz zwischen Marke 3 und dem Cluster (14) berechnet:

$$D_{3(14)} = (D_{13} + D_{34})/2 = (0,43 + 0,38)/2 \sim 0,41$$

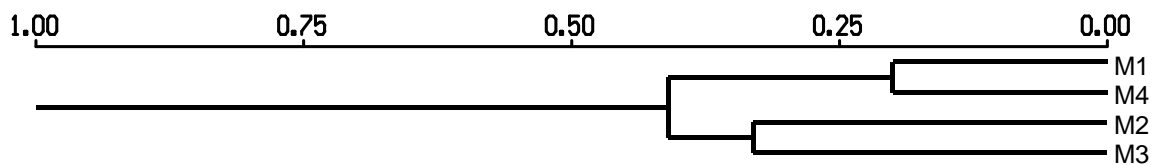
Die Dice-Distanzen nach diesen Berechnungen sind:

	2	3	(14)
2	0		
3	0,33	0	
(14)	0,42	0,41	0

Als nächstes werden die Marken 2 und 3 in einem Cluster verschmolzen, da die Distanz zwischen beiden die kürzeste in der neuen Distanzmatrix ist ($D_{23} = 0,33$). Im letzten Schritt werden dann die Cluster (14) und (23) zusammengeführt. Die durchschnittliche Distanz dieser beiden Cluster ist:

$$D_{(14)(23)} = (D_{12} + D_{13} + D_{24} + D_{34})/4 = (0,29 + 0,43 + 0,54 + 0,38)/4 = 0,41$$

Das Ergebnis der Clusterprozesses ist graphisch als Dendrogramm zusammengefasst. Die Skala oberhalb des Dendrogramms zeigt an, bei welcher Distanz die Cluster vereinigt wurden.



14.3.2 Single Linkage (nearest neighbor)

Wie bei Average Linkage werden die Marken 1 und 4 zuerst zusammengeführt, aber die neue Distanzmatrix wird anders berechnet. Die Single Linkage Distanz zwischen dem Cluster (14) und der Marke 2 ist die kürzere der beiden Distanzen D_{12} und D_{24} :

$$D_{2(14)} = \min(D_{12}, D_{24}) = \min(0,29; 0,54) = 0,29$$

Ebenso berechnet sich die neue Distanz zwischen (14) und 2:

$$D_{3(14)} = \min(D_{13}, D_{34}) = \min(0,43; 0,38) = 0,38$$

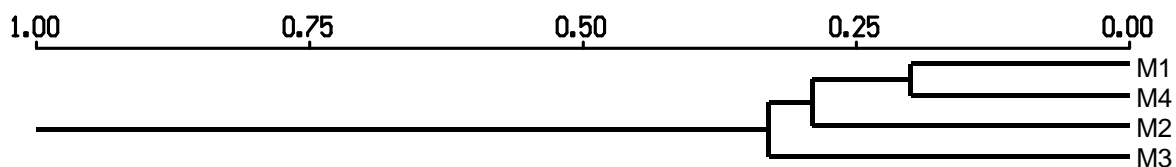
Die Dice-Distanzen nach dem ersten Durchgang sind:

	2	3	(14)
2	0		
3	0,33	0	
(14)	0,29	0,38	0

Im nächsten Schritt werden (14) und Marke 2 zusammengefasst, da $D_{2(14)} = 0,29$ die kürzeste Distanz in der neuen Distanzmatrix ist. Man sieht hier einen Fall von Kettenbildung (chaining), der für das Single-Linkage Clusterverfahren typisch ist. Wir müssen nun die Distanz zwischen dem neuen Cluster (124) und der verbleibenden Marke 3 berechnen:

$$D_{3(124)} = \min(D_{13}, D_{23}, D_{34}) = \min(0,43; 0,33; 0,38) = 0,33$$

Das Dendrogramm hat folgende Form:



14.3.3 Complete Linkage (furthest neighbor)

Wie bei Average Linkage und Complete Linkage werden die Marken 1 und 4 im ersten Schritt zusammengeführt. Die Complete-Linkage Distanz zwischen dem neuen Cluster (14) und dem Marke 2 ist die größere der beiden Distanzen D_{12} und D_{24} :

$$D_{2(14)} = \max(D_{12}, D_{24}) = \max(0,29; 0,54) = 0,54$$

Analog berechnen wir die neue Distanz $D_{3(14)}$:

$$D_{3(14)} = \max(D_{13}, D_{34}) = \max(0,43; 0,38) = 0,43$$

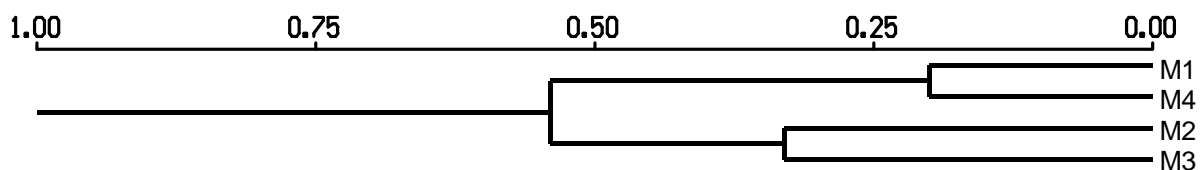
Die Dice-Distanzen nach dem ersten Schritt sind:

	2	3	(14)
2	0		
3	0,33	0	
(14)	0,54	0,43	0

Als nächstes werden die Marken 2 und 3 in ein Cluster gruppiert, da ihre Distanz die kürzeste ist ($D_{23} = 0,33$). Die Complete-Linkage Distanz zwischen den Clustern (23) und (14) ist das Maximum aller paarweisen Distanzen, in denen eine Marke aus Cluster (14) und eine Marke aus Cluster (23) stammt:

$$D_{(14)(23)} = \max(D_{12}, D_{13}, D_{24}, D_{34}) = \max(0,29; 0,43; 0,54; 0,38) = 0,54$$

Das Dendrogramm ist im folgenden abgebildet:



14.3.4 Anwendung auf Margarinedaten

Das Ergebnis der drei Verfahren ist in den Abbildungen 14.1 bis 14.3 für das Beispiel 1 dargestellt. An der Ordinate wird der Abstand dargestellt, bei dem zwei Cluster fusioniert wurden. Will man die Ähnlichkeit von zwei Produkten rekonstruieren, so muss man den Weg von einem zum anderen über das Dendrogramm suchen. Entscheidend für die Ähnlichkeit zweier Objekte ist der Ordinatenwert, bei dem beide Objekte zusammenlaufen. In Abb. 14.1 laufen SB und Botteram bei einer Distanz von ca. 1,0 zusammen, sind sich also ähnlich. Dagegen treffen SB und Delicado erst bei ca. 5,0 zusammen, sind also relativ verschieden. Die Distanzen von 5,0 ist aber nicht exakt identisch mit der Euklidischen Distanz zwischen den beiden Objekten "SB" und "Delicado", sondern dieser Wert entspricht der Distanzen der beiden im letzten zusammengeführten Cluster. Diese Unschärfe ist der Preis für die Projektion der aus p Variablen (Dimensionen) berechneten Distanzmatrix in zwei Dimensionen. Diese geht einher mit einem unvermeidbaren Informationsverlust, was typisch ist für viele multivariate Verfahren.

Alle drei Analysen sind recht ähnlich. Die relative Ähnlichkeit von SB, Botteram, Rama und Flora wird in allen dreien deutlich, und ebenso der Cluster der drei Produkte "Holl. Butter", "Weihnachtsbutter" und "Delicado".

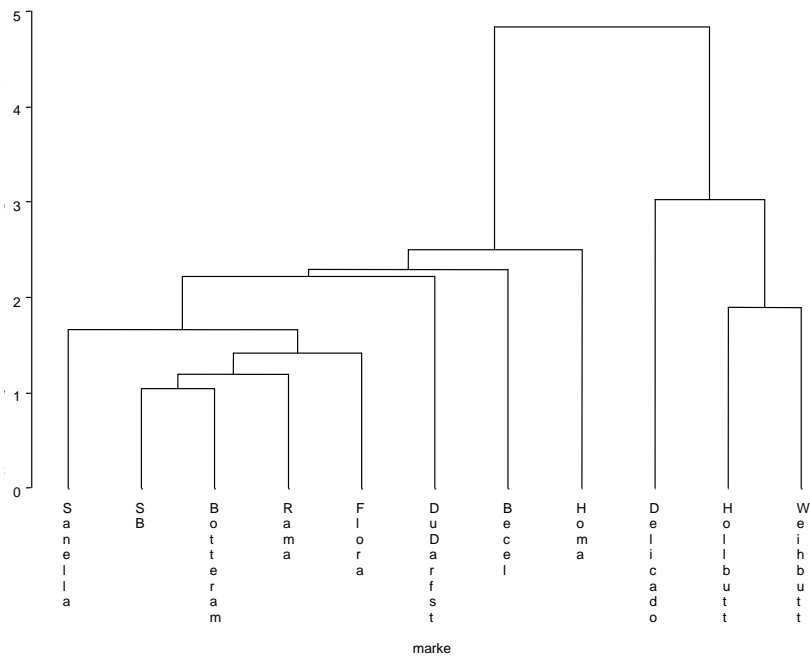


Abb. 14.1: Dendrogram für Margarine-Daten aus Beispiel 1 (unstandardisierte Daten, Euklidische Distanzen, Average Linkage).

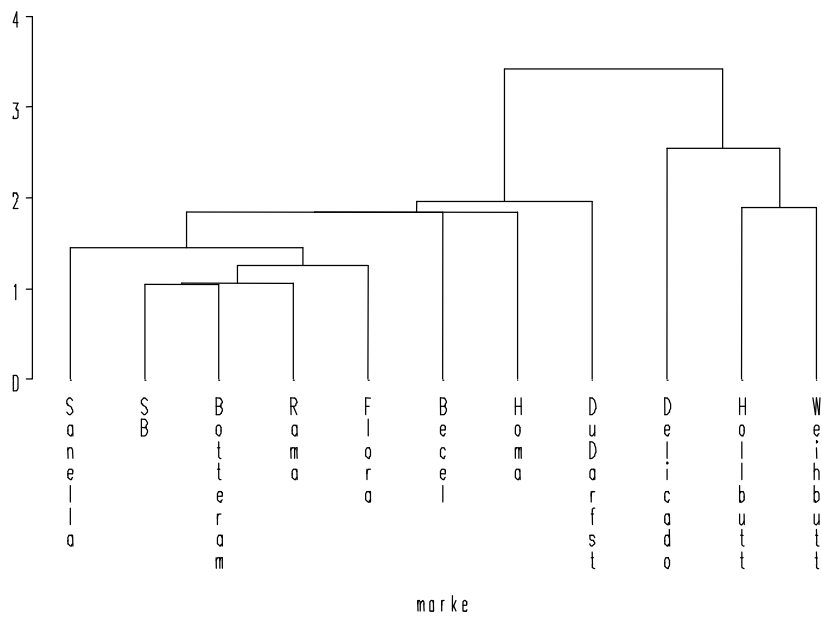


Abb. 14.2: Dendrogram für Margarine-Daten aus Beispiel 1 (unstandardisierte Daten, Euklidische Distanzen, Single Linkage).

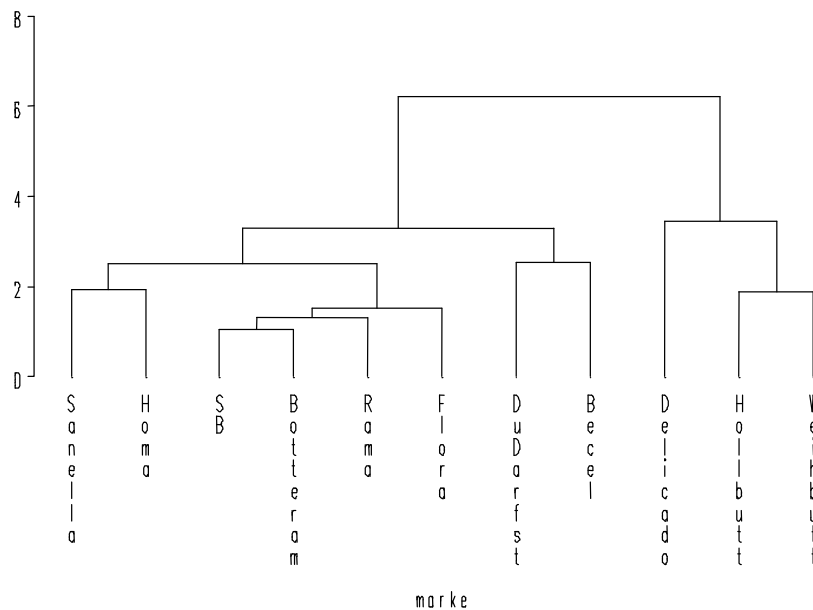


Abb. 14.3: Dendrogram für Margarine-Daten aus Beispiel 1 (unstandardisierte Daten, Euklidische Distanzen, Complete Linkage).

Übrigens ist der horizontale Abstand der Objekte (Produkte) in den Abbildungen 14.1 bis 14.3 kein Maß für die Ähnlichkeit. Das macht man sich am besten deutlich, in dem man sich das Dendrogramm als ein Mobile denkt, dessen Bügel frei drehbar sind. In Abb. 14.3 sind z.B. Becel und Delicado benachbart, aber keineswegs sehr ähnlich. Eine Drehung der beiden Bügel direkt unterhalb des obersten Bügels würde Becel ganz nach links und Delicado ganz nach rechts bringen.

Zum Vergleich noch das Dendrogramm für die Simple-Matching Koeffizienten aus Beispiel 2 (Complete Linkage). Die Produkte werden hier etwas anders gruppiert. Dies zeigt, dass die Wahl der Variablen wichtig für das Ergebnis der Analyse ist.

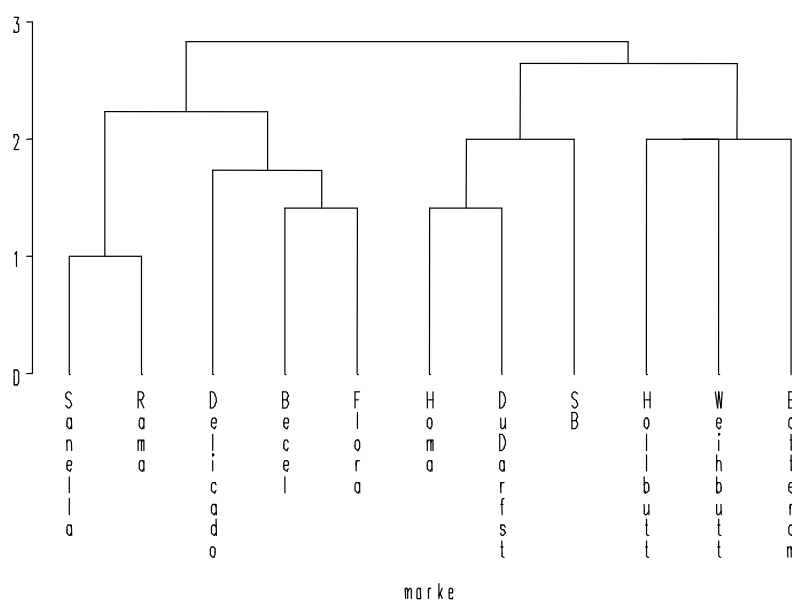


Abb. 14.4: Dendrogram für Margarine-Daten aus Beispiel 2 (Euklidische Distanzen = Distanz berechnet aus Simple-Matching Koeffizienten, Complete Linkage).

Kopphenische Korrelation: Dies ist eine Maßzahl für die Übereinstimmung der ursprünglichen Distanzmatrix und den durch das Dendrogramm implizierten Distanzen.

Mantel-Test: Dieser Test ist geeignet für den Vergleich von zwei Distanzmatrizen für dieselben Objekte. In vegetationsökologischen Untersuchungen werden an verschiedenen Orten (Objekte) oft einerseits die Abundanzen verschiedener Pflanzenarten und andererseits verschiedene Umweltparameter erhoben. Man kann nun zunächst basieren auf den Abundanzen eine Distanzmatrix der Orte berechnen. Dasselbe tut man für die Umweltparameter. Sodann stellt sich die Frage, ob die Distanzen für die Umweltparameter mit den Distanzen basierend auf den Abundanzen korreliert sind. Hierzu berechnet man einfach die Korrelation (Pearson oder Spearman) der Distanzen in den beiden Distanzmatrizen, wobei als Wertepaare jeweils die beiden Distanzen eines Ortpaares verwendet werden. Der Mantel-Test prüft, ob die Korrelation der Distanzmatrizen signifikant ist. Dies kann nicht mit Standardverfahren erfolgen, da die Distanzen einer Matrix nicht statistisch unabhängig sind. Daher ist das spezielle Verfahren des Mantel-Tests erforderlich.

SAS Anweisungen

Mit quantitativen Daten (Beispiel 1):

```
data daten;  
input marke $ v1-v10;  
datalines;  
Sanella 4.500 4.000 4.375 3.875 3.250 3.750 4.000 2.000 4.625 4.125  
Homa 5.167 4.250 3.833 3.833 2.167 3.750 3.273 1.857 3.750 3.417  
SB 5.069 3.824 4.765 3.438 4.235 4.471 3.765 1.923 3.529 3.529  
Delicado 3.800 5.400 3.800 2.400 5.000 5.000 5.000 4.000 4.000 4.600  
Hollbutt 3.444 5.056 3.778 3.765 3.944 5.389 5.056 5.615 4.222 5.278  
Weihbutt 3.500 3.500 3.875 4.000 4.625 5.250 5.500 6.000 4.750 5.375  
DuDarfst 5.250 3.417 4.583 3.917 4.333 4.417 4.667 3.250 4.500 3.583  
Becel 5.857 4.429 4.929 3.857 4.071 5.071 2.929 2.091 4.571 3.786  
Botteram 5.083 4.083 4.667 4.000 4.000 4.250 3.818 1.545 3.750 4.167  
Flora 5.273 3.600 3.909 4.091 4.091 4.091 4.545 1.600 3.909 3.818  
Rama 4.500 4.000 4.200 3.900 3.700 3.900 3.600 1.500 3.500 3.700  
;  
proc cluster data=daten method=complete outtree=dendro nonorm;  
var v1-v10;  
id marke;  
run;  
  
proc tree data=dendro;  
id marke;  
run; quit;
```

Mit binären Daten (Beispiel 2) dasselbe:

```
data daten;  
input marke $ v1-v10;  
datalines;  
Sanella 1 0 1 1 1 0 1 1 1 0  
Homa 1 0 0 1 1 1 0 1 0 1  
SB 1 1 0 1 1 1 0 0 1 0  
Delicado 0 0 1 1 0 0 1 0 1 0  
Hollbutt 0 0 0 0 0 1 0 0 0 0  
Weihbutt 0 0 0 0 1 0 1 0 0 1  
DuDarfst 1 1 0 1 0 1 0 1 0 1  
Becel 1 1 1 1 0 0 1 0 0 0  
Botteram 0 0 1 1 1 1 0 0 0 1  
Flora 1 1 1 1 1 0 1 0 1 0  
Rama 1 0 1 1 1 1 1 1 1 0  
;  
proc cluster data=daten method=complete outtree=dendro nonorm;  
var v1-v10;  
id marke;  
run;  
  
proc tree data=dendro;  
id marke;  
run; quit;
```

14.4 Hauptkomponentenanalyse

Die Hauptkomponentenanalyse (Principal Component Analysis - PCA) ist eine weitverbreitete explorative multivariate Methode. Sie untersucht die Struktur von n Objekten mit Messungen für je p Variablen und liefert informative graphische Darstellungen, welche sich z.B. zur Gruppierung von Objekten eignet. Ähnlich wie die Clusteranalyse hat die PCA zum Ziel, die Distanzen von Objekten im p -dimensionalen Merkmalsraum abzubilden.

Während die wichtigen Prinzipien der PCA recht klar und einfach sind, ist die mathematische Herleitung und Berechnung dieses Verfahrens etwas komplizierter. Daher wird zunächst in Abschnitt 14.4.1 ein erster Überblick gegeben, in dem die aus praktischer Sicht wesentlichen Punkte angesprochen werden. Mathematische Details werden in 14.4.2 beschrieben und sind zum Verständnis der wesentlichen Aspekte nicht zwingend erforderlich.

14.4.1 Ein erster Überblick

Die Euklidische Distanz für p Merkmale bezieht sich auf einen p -dimensionalen Raum. Bis zu drei Dimensionen kann man die Abstände noch graphisch darstellen, aber ab der vierten Dimension ist dies unmöglich. Allerdings kann man die Punkte aus dem p -dimensionalen Raum in einen 2- oder 3-dimensionalen Raum projizieren und dann die projizierten Punkte graphisch darstellen. Das Prinzip kann man am Beispiel einer Projektion von zwei Dimensionen auf eine Dimension erläutern. In Abb. 14.5 sind acht Punkte in einem 2-dimensionalen Koordinatensystem dargestellt. Zur Projektion in eine Dimension muss eine Gerade durch die Punkte gelegt werden. Die Punkte werden senkrecht auf die Gerade projiziert. Hierbei wird die Lage der Geraden so gewählt, dass die Summe der orthogonalen Abweichungsquadrate minimal wird. Dies bedingt gleichzeitig, dass die Summe der Abweichungsquadrate der projizierten Punkte auf der Gerade maximal wird. Die Gerade stellt die erste *Hauptkomponentenachse* (principal component axis - PCA) oder kurz Hauptachse der Punktwolke dar. Die zweite Hauptkomponentenachse steht senkrecht auf der ersten Hauptkomponentenachse. Insofern können wir auch sagen, dass das ursprüngliche Koordinatensystem so gedreht wird, dass im neuen Koordinatensystem die Werte auf der ersten Achse maximale Varianz haben, d.h., dass mit dieser Achse so viele Unterschiede wie möglich erfasst werden.

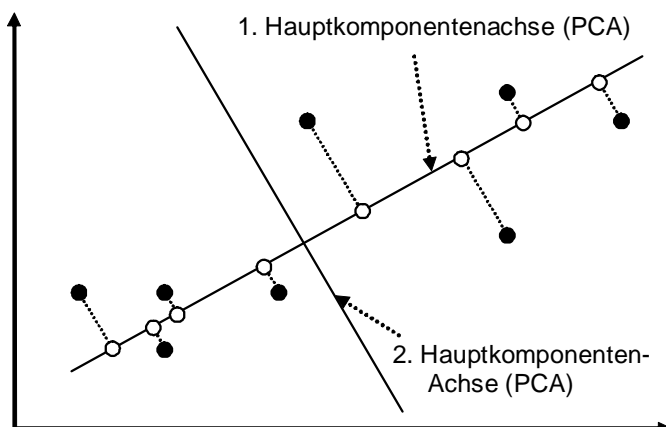


Abb. 14.5: Orthogonale Projektion auf die erste Hauptkomponentenachse PCA). Hypothetische Datenpunkte.

Eine Bemerkung zur Bedeutung der Varianz. In den meisten Fällen hatten wir es bisher mit Verfahren zu tun, bei denen es wünschenswert war, die (Rest-)Varianz zu minimieren. Hier war dagegen eben die Rede davon, dass eine Varianz maximiert werden soll, nämlich die Varianz, welche durch die erste Hauptkomponente erklärt werden kann. Die Gesamtvarianz ist allerdings bei gegebenen Daten eine Konstante. Es geht bei der PCA darum, die gesamte Varianz so auf die Hauptkomponenten zu verteilen, dass die ersten Achsen ein Maximum der gesamten Varianz erfassen. Dies ermöglicht es, die ursprünglichen Distanzen zwischen Objekten im p -dimensionalen Raum in nur zwei Dimensionen möglichst wahrheitsgetreu darzustellen.

Die soeben für die Projektion von zwei Dimensionen auf eine Dimension erläuterte Idee lässt sich auf beliebig viele Dimensionen erweitern. Im p -dimensionalen Raum wird die erste Hauptkomponente so gelegt, dass sie eine maximale Varianz aufweist. Die zweite Hauptkomponente muss zur ersten orthogonal (senkrecht) sein und wird wiederum so gelegt, dass die Varianz entlang dieser zweiten Hauptkomponente maximal wird. Die dritte muss orthogonal zu den beiden ersten Hauptkomponenten sein und dabei die Varianz maximieren, usw. Die ersten beiden Hauptkomponenten definieren eine Ebene, welche die Eigenschaft hat, dass die Varianz der orthogonalen Abstände der Punkte im p -dimensionalen Raum von dieser Ebene minimal wird, während die Varianz der Abstände in der Ebene maximal wird. In anderen Worten: Ein Maximum der gesamten Varianz wird in der Ebene dargestellt. Wegen der Projektion können die Relationen zwischen den paarweisen Abständen im p -dimensionalen Raum nicht exakt auf den 2-dimensionalen Raum abgebildet werden, jedoch ist die Projektion optimal in dem Sinne, dass die Verzerrung der tatsächlichen Verhältnisse im p -dimensionalen Raum minimal ist.

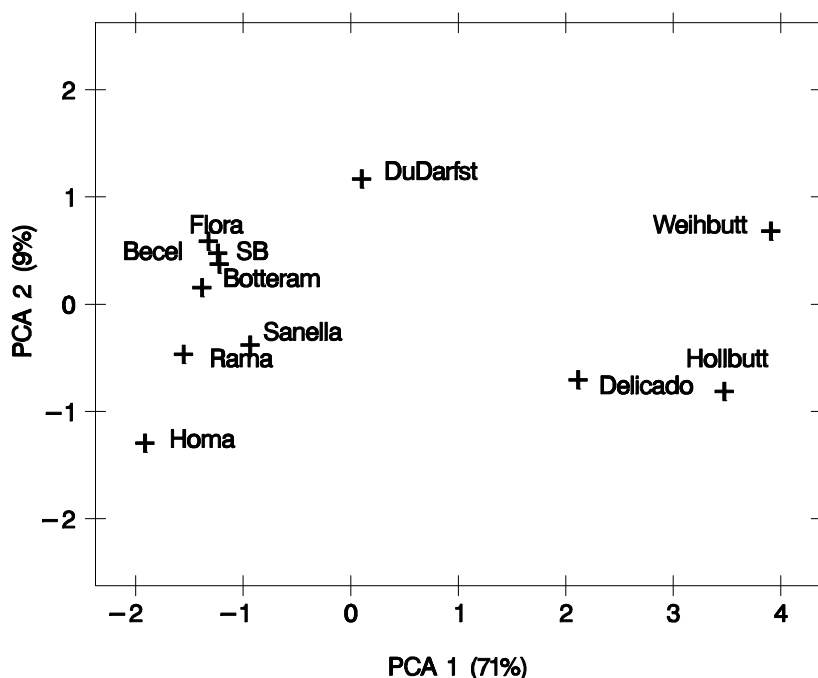


Abb. 14.6: Repräsentation der ersten beiden Hauptkomponenten (PCA1 und PCA2) für die Margarine-Daten aus Beispiel 1 (unstandardisierte Daten). PCA1 erklärt 71% der Varianz, PCA2 erklärt 9%.

Abb. 14.6 zeigt die Darstellung der ersten beiden Hauptkomponenten für die Margarine-Daten aus Beispiel 1. Ähnlich wie bei der Clusteranalyse zeigt sich auch

hier eine Untergliederung in zwei Gruppen. Die Hauptkomponentenanalyse macht jedoch deutlicher, in welchem Maße die Gruppen homogen sind oder weiter streuen.

Skalierung der Variablen

Ziel der PCA ist es, Abstände in einem p -dimensionalen Merkmalsraum auf 2 oder 3 Dimensionen zu projizieren. Bevor wir zu dieser Projektion kommen, wollen wir uns mit der Skalierung der Merkmalsvariablen beschäftigen. Hierzu betrachten wir den ganz einfachen Fall von nur 2 Variablen. Hier ist zur graphischen Darstellung gar keine Projektion erforderlich. Dies vereinfacht die Betrachtung der Skalierung der Variablen, auf die wir uns jetzt konzentrieren können, ohne dass eine PCA durchgeführt wird.

Beispiel 4 (Siba Hassan - 320 -): In einer ökologischen Untersuchung wurden Verschiedene Schwermetalle (Cd, Pb, Co, ... = Variablen) in Bodenproben an verschiedenen Orten (Objekte) aus einem Salzseegebiet untersucht. Dabei waren folgende Fragen von Interesse: Haben manche Metalle ähnliche Verteilung? Sind manche Metalle höher korreliert als andere? Welche Metalle dominieren an welchem Ort (Objekt)?

Diese Fragen lassen sich mit einer PCA untersuchen. Eine wichtiger Aspekt bei der PCA ist die Skalierung der Variablen. Dies soll im Folgenden an einem kleinen hypothetischen Datensatz mit nur 2 Variablen erläutert werden, welcher dem obigen Beispiel 4 entlehnt ist.

Ort	pH	Phosphor
1	8.2	60
2	7.5	75
3	6.7	66
4	6.2	76
5	6.0	83
6	5.8	74
7	5.3	99
8	4.7	86

Man beachte, dass die beiden Variablen ganz unterschiedliche Maßeinheiten haben und sehr verschiedene Wertebereiche aufweisen. Im Folgenden werden nun die Messwertpaare für die 8 Orte graphisch gegeneinander abgetragen (Abb. 14.7 bis 14.9). Alle Werte werden dabei zentriert, indem der Stichprobenmittelwert subtrahiert wird. Außerdem wird der Wert für Phosphor in den beiden letzten Graphiken (Abb. 14.8 und 14.9) durch 10 bzw. durch 100 dividiert. Die verwendeten Skalierungen für Phosphor sind also:

- (1) (Wert – Mittel)
- (2) (Wert – Mittel)/10
- (3) (Wert – Mittel)/100

pH Original-Skala (zentriert) (1)

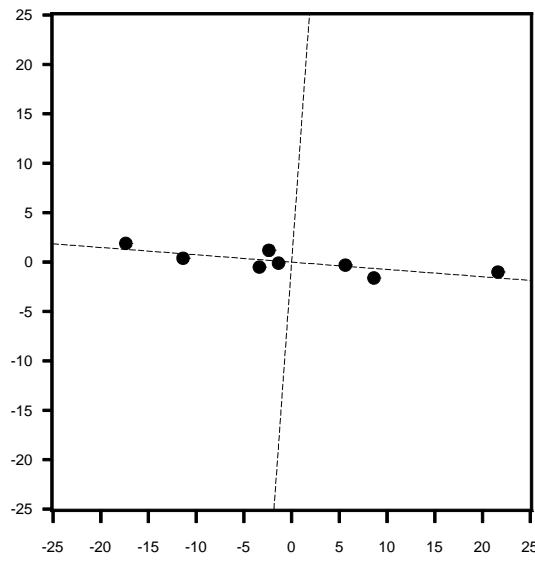


Abb. 14.7

pH Transformierte Skala (zentriert) (2)

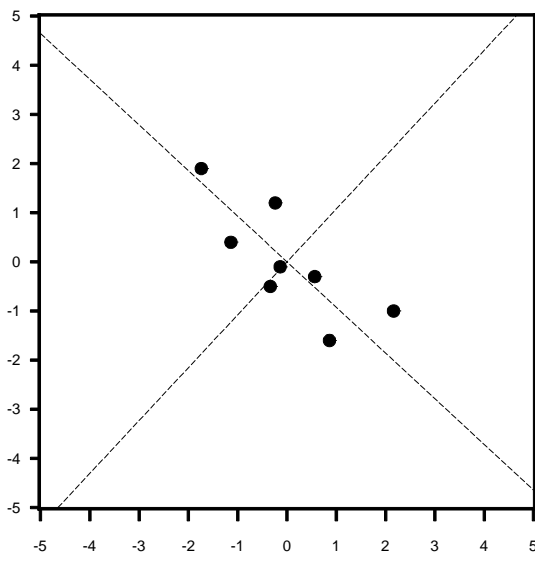


Abb. 14.8

pH Transformierte Skala (zentriert) (3)

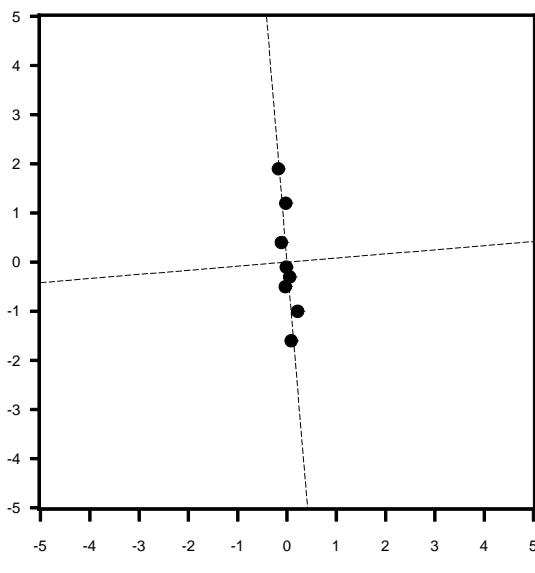


Abb. 14.9

Man sieht, dass der Einfluss auf die Gruppierung und die Abstände der Objekte (Orte) von der Skalierung abhängt. Je größer der Faktor, durch den der Phosphor-Wert dividiert wird, umso geringer der Einfluss von Phosphor. Im ersten Fall (Abb. 14.7) hängt der Abstand vor allem vom Phosphor ab, im dritten (Abb. 14.9) dagegen hauptsächlich vom pH-Wert. Dies bedeutet, dass der Einfluss einer Variablen davon abhängt, in welchen Maßeinheiten sie erfasst wird und wie eventuell umskaliert wird. Eine solche Abhängigkeit ist nicht sinnvoll. Besser ist es, die Daten so zu **standardisieren**, dass sie unabhängig sind von den Maßeinheiten. Hierzu verwendet man die z-Transformation:

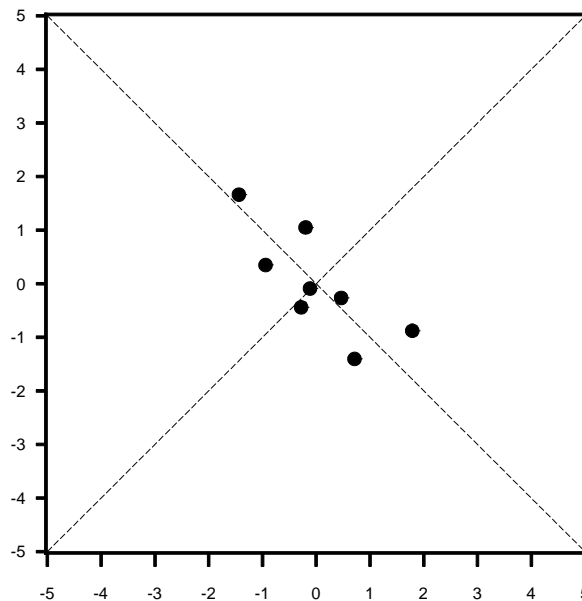
$$z_{ij} = \frac{x_{ij} - \bar{x}_{\bullet j}}{s_j} \quad , \quad \text{wobei}$$

x_{ij} = ursprünglicher Messwert für j -te Variable bei i -tem Objekt und

$$s_j^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_{\bullet j})^2}{n-1} \quad .$$

Man beachte, dass die standardisierten Daten unabhängig von Maßeinheiten sind. Ein Wechsel der Maßeinheiten bleibt ohne Folgen für die standardisierten Daten.

pH standardisiert



Phosphor standardisiert

Abb. 14.10

Skalierung der Achsen

Wichtig bei der graphischen Darstellung ist, dass die beiden Achsen **gleich skaliert** werden, dass also abgebildete Abstände für eine Einheit an beiden Achsen die gleiche Länge in cm oder mm haben. Ansonsten kommt es zu einer Verzerrung, so dass Euklidische Distanzen der Objekte nicht mehr interpretierbar sind, wie in Abb. 14.11 sichtbar wird.

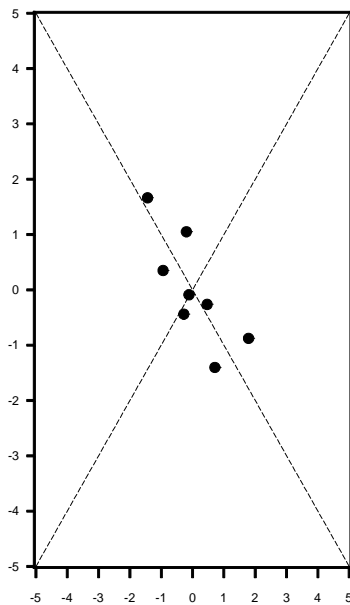
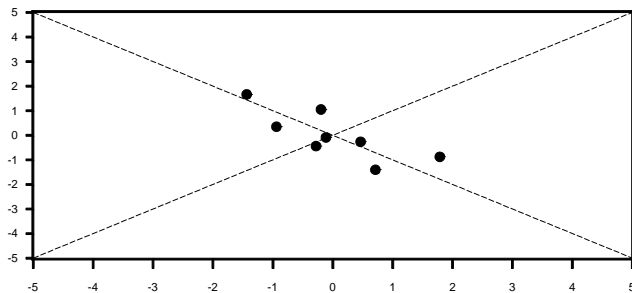


Abb. 14.11: Verzerrung der Achsen der Abb. 14.10 mit ursprünglich gleich skalierten Achsen.



Man muss daher insbesondere bei der Einbindung von PCA-Graphiken in Textverarbeitungsprogramme darauf achten, dass eine ursprünglich gleiche Achsenskalierung bei der Einbindung nicht verzerrt wird.

Berechnung der Hauptkomponenten

Die PCA basiert auf einer Analyse der standardisierten (z-transformierten) Datenmatrix

$$\mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & \cdot & \cdot & \cdot & z_{1p} \\ z_{21} & \cdot & \cdot & \cdot & \cdot & \cdot \\ z_{31} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ z_{n1} & z_{n2} & \cdot & \cdot & \cdot & z_{np} \end{pmatrix} \quad (\text{standardisiert auf Mittel}=0 \text{ und Varianz}=1!)$$

Die p Variablen werden als Koordinaten im p -dimensionalen Raum betrachtet. Es gibt dann n Punkte, korrespondierend zu n Objekten. Aus den ursprünglichen Daten werden neue Variablen berechnet. Dies entspricht einer **Rotation** des ursprünglichen Koordinatensystems. Von diesem werden dann oftmals nur die beiden ersten näher betrachtet, sofern diese den Großteil der Varianz erklären. Wie der Anteil der erklärten Varianz erfasst wird, betrachten wir später. Die neuen (rotierten) Koordinaten berechnen sich nach

$$a_{ik} = g_{k1}z_{i1} + g_{k2}z_{i2} + \dots + g_{kp}z_{ip} = \mathbf{z}_i^T \mathbf{g}_k \Rightarrow k\text{-te Hauptkomponente}$$

$$\mathbf{A} = \mathbf{Z}\mathbf{\Gamma}$$

$$\Gamma = \begin{pmatrix} g_{11} & g_{21} & \cdot & \cdot & \cdot & g_{1p} \\ g_{21} & \cdot & \cdot & \cdot & \cdot & \cdot \\ g_{31} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ g_{p1} & \cdot & \cdot & \cdot & \cdot & g_{pp} \end{pmatrix} = (g_1 \quad g_2 \quad \cdot \quad \cdot \quad \cdot \quad g_p)$$

g_1, g_2, \dots sind die **Eigenvektoren** \Rightarrow Gewichte! Das hochgestellte T bedeutet „transponiert“. Dies hat dieselbe Bedeutung wie ein Strich, also $z_i^T = z_i'$.

Die neuen Koordinaten a_{ik} sind Linearkombinationen der ursprünglichen Koordinaten (Variablen), wobei die Koeffizienten g_{ik} die Gewichte sind. Die Matrix $\Gamma = \{g_{ik}\}$ enthält die optimalen Gewichte, während die Matrix $A = \{a_{ik}\}$ die Koordinaten der Objekte im neuen rotierten Koordinatensystem enthält.

Optimale Gewichte g_{ik} : Die Gewichte werden so bestimmt, dass die erste Hauptkomponente maximale Varianz (Eigenwert) hat. Die zweite Hauptkomponente ist unabhängig von der ersten und erklärt die maximale Varianz vom Rest etc.

Optimale Gewichte bestimmt man durch eine sog. **Spektralzerlegung** (wird erst in 14.4.2 näher erklärt) von

$$C = \frac{1}{n-1} Z^T Z .$$

Hierbei ist C eine **Kovarianz-Matrix** bei zentrierten Daten. Dagegen ist C eine **Korrelations-Matrix** bei standardisierten Daten (Varianz = 1). Eine PCA sollte vorzugsweise mit der Korrelationsmatrix, also mit standardisierten Daten durchgeführt werden, insbesondere, wenn die verwendeten Variablen unterschiedliche Messniveaus und Maßeinheiten haben.

Ergebnis der Spektralzerlegung: Die Korrelationsmatrix C wird zerlegt nach der Gleichung

$$C = \Gamma A \Gamma^T, \text{ wobei}$$

$$\Gamma = (g_1 \quad g_2 \quad \cdot \quad \cdot \quad \cdot \quad g_p) \Rightarrow \text{Eigenvektoren}$$

$$A = \begin{pmatrix} \lambda_1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \lambda_2 & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \lambda_p \end{pmatrix} \quad \text{Eigenwerte } \lambda_1 > \lambda_2 > \dots > \lambda_p$$

Diese Zerlegung liefert die optimalen Gewichte (Γ), welche den Eigenvektoren entsprechen. Mit Hilfe der Eigenvektoren (Γ) werden die Hauptkomponenten aus den Daten (Z) nach $A = Z\Gamma$ berechnet. Die Eigenwerte messen die Bedeutung der verschiedenen Hauptkomponenten (siehe unten).

Erklärte Varianz

Für einen PCA Plot stellt man die ersten beiden Hauptkomponenten für die Objekte in einem Koordinatensystem dar. Die Hoffnung ist dabei, dass diese möglichst viel der gesamten Varianz erklären. Die erklärte Varianz kann aus den Eigenwerten berechnet werden.

$$\text{Durch 1. Hauptkomponente erklärte Varianz} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \dots + \lambda_p} 100\%$$

$$\text{Durch 2. Hauptkomponente erklärte Varianz} = \frac{\lambda_2}{\lambda_1 + \lambda_2 + \dots + \lambda_p} 100\%$$

Beispiel 4: Schwermetalldaten (Siba Hassan, Institut 320)

Orte	P	Fe	Mo	OC	Ni	CO	Cr	Cu	Pb	Zn	Mn	Ca	Cd
61	3843	15606	2.10	27.98	74	13.0	55	64	22.0	94	250	81278	1.40
62	5089	20977	2.30	4.41	49	12.0	59	27	12.0	86	386	147172	0.50
63	1319	20738	2.20	0.18	41	12.0	45	20	8.1	39	362	163445	0.40
68	996	92871	2.20	2.77	97	98.0	147	59	9.4	90	1910	22461	0.30
3	998	19400	0.39	2.02	41	13.0	47	15	6.0	73	345	216000	0.33
5	1793	2430	0.46	3.38	47	13.0	55	14	7.3	105	431	170000	0.40
51	1793	2430	0.46	3.38	47	13.0	55	14	7.3	105	431	170000	0.40
8	1637	24800	0.42	2.83	46	15.0	56	16	7.9	96	407	221000	0.40
9	900	18700	0.05	0.92	42	10.0	52	15	5.1	73	301	229000	0.34
10	912	65100	2.90	2.40	109	35.0	132	41	16.0	111	777	45700	0.60
12	1415	16300	0.29	2.48	36	9.7	46	17	5.7	44	292	260000	0.62
18	1621	24700	0.59	2.70	46	14.0	52	18	6.8	61	395	193000	0.63
100	3525	16003	0.65	20.00	52	24.0	49	60	17.0	68	415	12188	0.98

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	6.40726442	2.38798312	0.4929	0.4929
2	4.01928131	3.02163208	0.3092	0.8020
3	0.99764922	0.28270446	0.0767	0.8788
4	0.71494476	0.17603478	0.0550	0.9338
5	0.53890998	0.35547412	0.0415	0.9752
6	0.18343586	0.10121159	0.0141	0.9893
7	0.08222427	0.04032703	0.0063	0.9957
8	0.04189724	0.03122877	0.0032	0.9989
9	0.01066847	0.00726236	0.0008	0.9997
10	0.00340611	0.00308775	0.0003	1.0000
11	0.00031836	0.00031836	0.0000	1.0000
12	0.00000000	0.00000000	0.0000	1.0000
13	0.00000000		0.0000	1.0000

Die ersten beiden Hauptkomponenten erklären 80% der Varianz. Im PCA-Plot in Abb. 14.12 fallen einige Orte auf, die relativ weit von dem Gros der Orte liegen und somit ein abweichendes Schwermetallprofil aufweisen müssen.

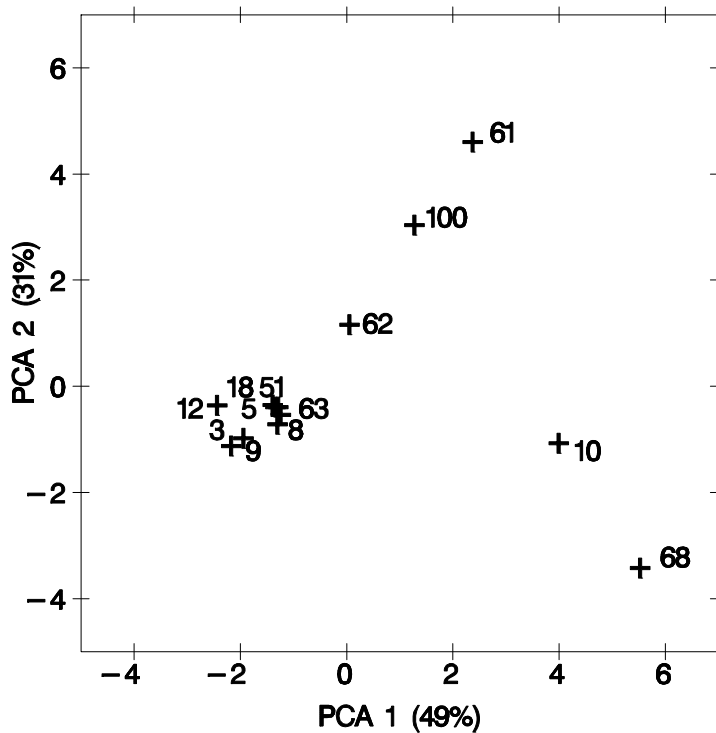


Abb. 14.12: PCA-Plot für standardisierte Daten (PCA1 erklärt 49% der Varianz, PCA2 erklärt 31%)

Im Folgenden wird für das Beispiel 4 gezeigt, dass bei PCA mit unstandardisierten Daten der Plot von Fe und Ca dominiert wird, weil diese die größten Messwerte aufweisen (Abb. 14.13 und 14.14). Man kann alle anderen Daten weglassen und bekommt fast dasselbe Bild!

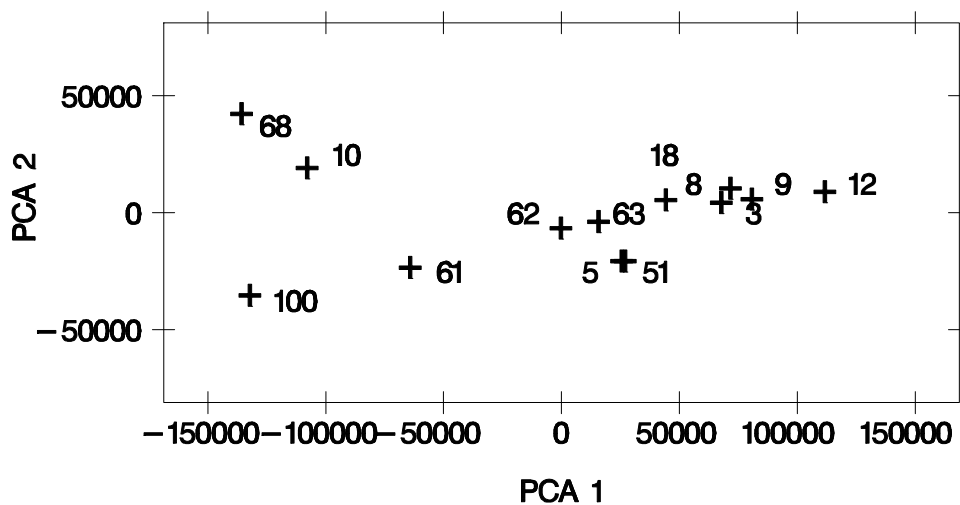


Abb. 14.13: PCA Plot für unstandardisierte Daten (alle Variablen) - Beispiel 4

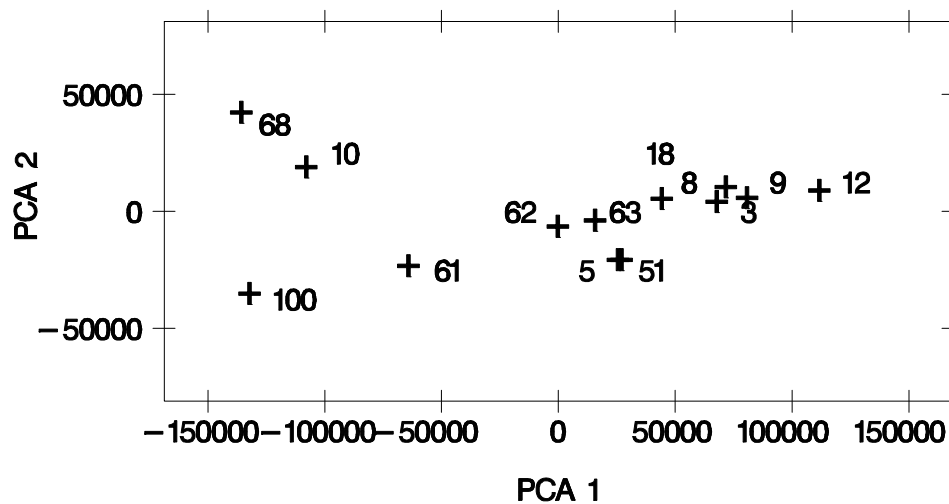
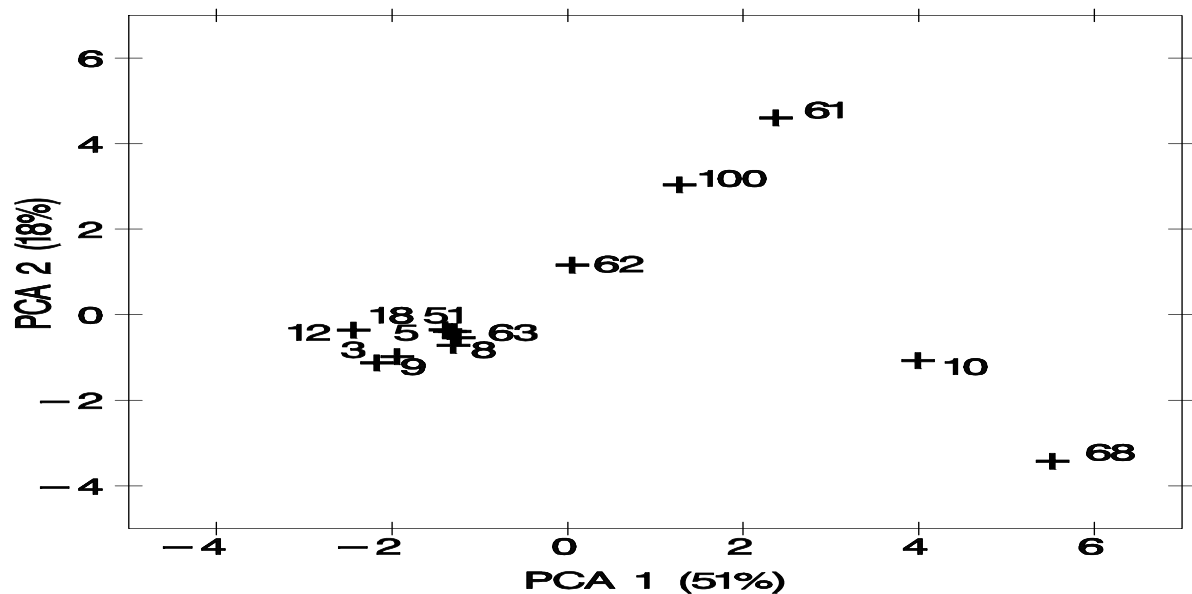


Abb. 14.14: PCA Plot für unstandardisierte Daten (nur Fe und Ca) - Beispiel 4

Außerdem kann man an diesem Beispiel zeigen, dass eine Streckung und Stauchung der Achsen, was einer Umskalierung entspricht, zu einer Verzerrung der Abstände der Objekte (Orte) führt (siehe Abb. 14.15). Bei gleicher Skalierung der Achsen kann man durch die Distanz der Punkte für die Objekte im PCA Plot die Euklidische Distanz der Objekte im ursprünglichen p -dimensionalen Raum approximieren. Man kann dann die PCA z.B. zur Clusterbildung nutzen. Diese Option geht verloren bei ungleicher Skalierung.



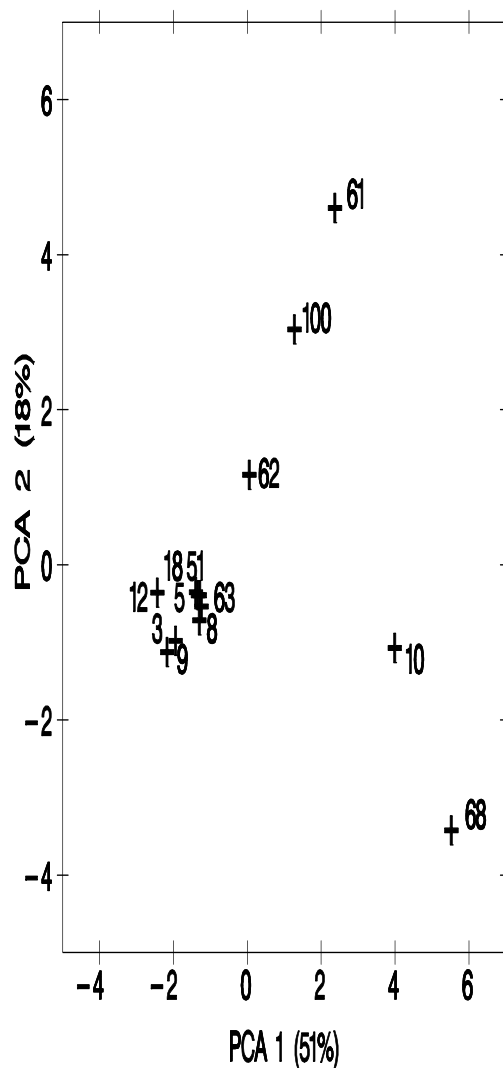


Abb. 14.15: PCA Plot für standardisierte Daten (ungleiche Achsenskalierung) - Beispiel 4

Korrelation der Hauptkomponenten zu den ursprünglichen Variablenwerten

Die Korrelation der Hauptkomponenten zu den ursprünglichen Variablenwerten ist gegeben durch

$$b_{ij} = g_{ij} \sqrt{\lambda_j / \sigma_i^2} \quad , \text{ wobei}$$

g_{ij} = i -tes Element des j -ten Eigenvektor

λ_j = j -ten Eigenwert

σ_i^2 = Varianz der i -ten Variable

Tab. 14.3 und Abb. 14.16 zeigen, dass z.B. Phosphor (P) stark mit der zweiten HK korreliert, während Zink (Zn) und Molybdän (Mo) besonders stark mit der ersten HK korrelieren.

Tab. 14.3: Korrelation der Hauptkomponenten (HK) zu den Variablen-Werten

Element	Eigenvektor		Korrelation der HK zu Variablen	
	g_{i1}	g_{i2}	b_{i1}	b_{i2}
P	0.047	0.388	0.12	0.78
Fe	0.318	-0.260	0.80	-0.52
Mo	0.294	0.035	0.74	0.07
OC	0.138	0.440	0.35	0.88
Ni	0.371	-0.065	0.94	-0.13
CO	0.317	-0.238	0.80	-0.48
Cr	0.339	-0.240	0.86	-0.48
Cu	0.337	0.213	0.85	0.43
Pb	0.260	0.364	0.66	0.73
Zn	0.171	-0.006	0.43	-0.01
Mn	0.299	-0.283	0.76	-0.57
Ca	-0.355	-0.119	-0.90	-0.24
Cd	0.115	0.444	0.29	0.89

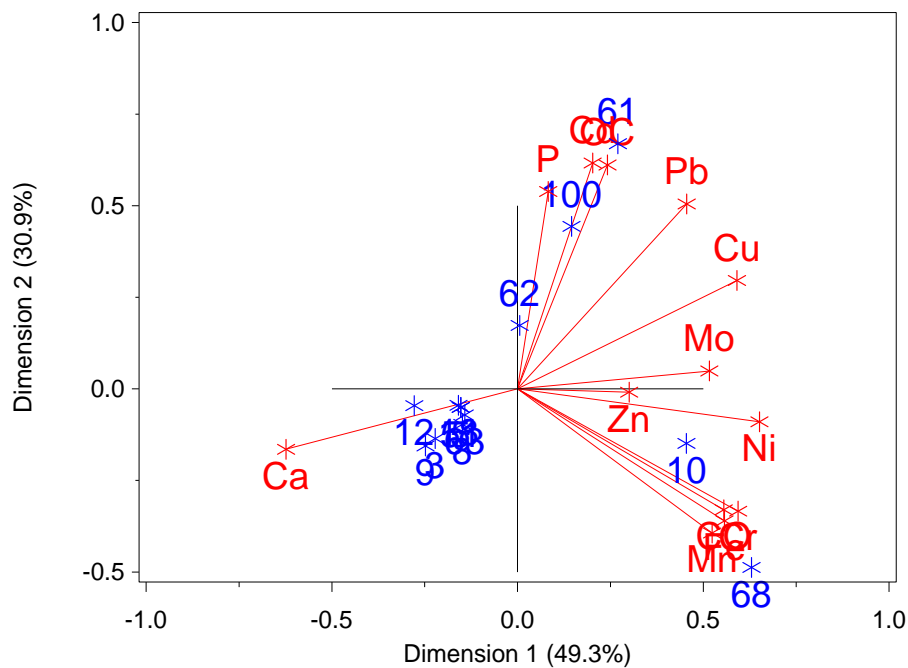


Abb. 14.16: Plot der b_{ij} (Biplot mit Option `factype=gh` in SAS-Makro `%biplot`)

Ladungen

Als "Ladungen" werden je nach Lehrbuch und Autor zweierlei Dinge bezeichnet:

- Eigenvektoren g_{ij}
- Korrelationen b_{ij}

Man muss genau im Handbuch nachschauen, was das Computerprogramm macht!

Ladungen (g_{ij}) für Schwermetalldaten (mit SAS PROC PRINCOMP):

The PRINCOMP Procedure							
Eigenvectors							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
P	0.046942	0.387978	0.086555	0.308727	0.721772	0.296520	-.307384
Fe	0.317541	-.259594	-.171797	0.079525	-.087753	0.344173	-.379458
Mo	0.294259	0.035168	0.007489	0.759250	-.093202	-.055706	0.525822
OC	0.137916	0.439976	-.081495	-.331748	-.023139	0.090946	0.346143
Ni	0.371144	-.064766	0.183150	0.048075	-.332724	0.092546	-.140357
CO	0.316505	-.237807	-.203564	-.235456	0.283620	0.088261	0.243451
Cr	0.338545	-.240163	0.110811	0.031756	-.100913	0.138181	-.277267
Cu	0.336819	0.212984	-.239475	-.190439	0.050884	-.018730	-.005333
Pb	0.259550	0.363542	0.081841	0.090153	-.175388	-.089601	-.114074
Zn	0.171165	-.006342	0.875612	-.233338	0.095429	0.060992	0.151621
Mn	0.298578	-.283288	-.136358	-.176612	0.311017	0.093943	0.328291
Ca	-.355331	-.118765	0.044861	0.062158	-.136931	0.779538	0.256064
Cd	0.115322	0.443758	-.132993	-.138526	-.320710	0.343793	-.001890

Eigenvectors							
	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	
P	-.112678	0.056735	-.108091	0.107119	0.024242	0.037966	
Fe	0.450100	-.514726	-.155613	-.099990	0.017800	-.154055	
Mo	0.016197	-.045636	-.072891	-.133551	-.118017	-.089203	
OC	0.252755	0.057285	-.327979	0.331535	0.072608	-.502152	
Ni	-.183480	0.199910	-.451936	0.411572	0.195513	0.444680	
CO	-.034673	-.060156	0.168911	0.335075	-.631591	0.233154	
Cr	-.256040	0.496035	0.220429	-.035012	-.125574	-.575392	
Cu	0.304721	0.449989	-.088621	-.590682	-.029980	0.302663	
Pb	0.301942	-.026101	0.706472	0.298674	0.183480	0.141489	
Zn	0.105615	-.177239	-.000804	-.243198	-.107254	0.022061	
Mn	-.245884	-.116389	0.173157	-.074656	0.676697	0.000000	
Ca	0.201320	0.274921	0.155102	0.034216	0.047338	0.134832	
Cd	-.569189	-.336344	0.094095	-.252749	-.144104	0.035293	

Biplots

In Biplots werden Objekte und Variablen gleichzeitig dargestellt (daher der Name). Die Grundlage von Biplots ist die Berechnung der Hauptkomponenten aus den ursprünglichen Variablen:

$$A_2 = Z\Gamma_2 \quad \text{mit} \quad \Gamma_2 = (g_1 \quad g_2)$$

Nun kann man umgekehrt die ursprüngliche Variablen aus den Hauptkomponenten berechnen, da

$$\Gamma\Gamma^T = \Gamma^T\Gamma = I$$

$$A_2\Gamma_2^T = Z\Gamma_2\Gamma_2^T \approx Z$$

In Biplots werden Objekte und Variablen wie folgt dargestellt:

Objekte: Punkte = Zeilen von A_2
 Variablen: Pfeile (Vektoren) = Zeilen von Γ_2

Die Daten (Variablen-Werte der verschiedenen Objekte) können aus dieser Darstellung rekonstruiert werden, indem die Objekt-Punkte auf die Variablen-Vektoren projiziert werden. Der Abstand eines Projektionspunktes vom Ursprung ist proportional zum ursprünglichen Variablen-Wert für das betreffende Objekt. Ein Beispiel ist in Abb. 14.17 dargestellt. Grundlage dieser Analyse ist die Tatsache, dass die Daten aus dem inneren Produkt von A_2 und Γ_2 rekonstruiert werden können: $Z \approx A_2 \Gamma_2^T$.

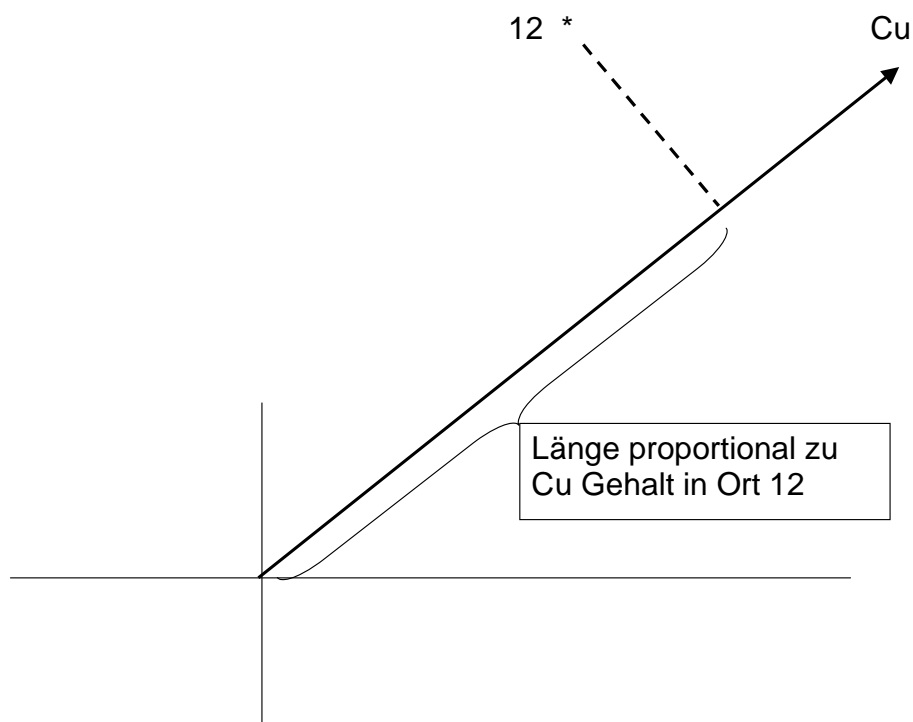


Abb. 14.17: Projektion des Ortes 12 auf den Vektor für Kupfer (Cu).

Drei Versionen des Biplot: Es gibt eine alternative Herleitung der PCA, die hier kurz erwähnt wird. Sie basiert auf einer **Singulärwertzerlegung** (singular value decomposition = SVD):

$$Z = USV^T$$

Z = standardisierte Daten-Matrix

S = Diagonal-Matrix mit Singulärwerten (= $\sqrt{\text{Eigenwerten}}$)

V und U sind orthogonale Matrizen

Basierend auf der SVD können drei verschiedene Versionen eines Biplot unterschieden werden:

- (1) $A_2 = U_2 S_2$ und $B_2 = V_2$
 - \Rightarrow Euklidische Distanzen der Objekte
 - \Rightarrow Cosinus zwischen Pfeilen der Objekte = Korrelation der Objekte

(2) $A_2 = U_2 S_2^{1/2}$ und $B_2 = V_2 S_2^{1/2} \Rightarrow$ keine Euklidischen Distanzen oder Korrelationen

(3) $A_2 = U_2$ und $B_2 = V_2 S_2 \Rightarrow$ Euklidische Distanzen der Variablen
 \Rightarrow Cosinus des Winkel zwischen Variablen Vektor = Korrelation der Variablen

Approximation der ursprünglichen Daten: Mit allen 3 Skalierungen können die Daten durch das innere Produkt approximiert werden:

$$Z \approx A_2 B_2^T = U_2 S_2 V_2^T$$

\Rightarrow Orthogonale Projektion der Objekt-Punkte auf Pfeile der Variablen (alle 3 Skalierungen)!

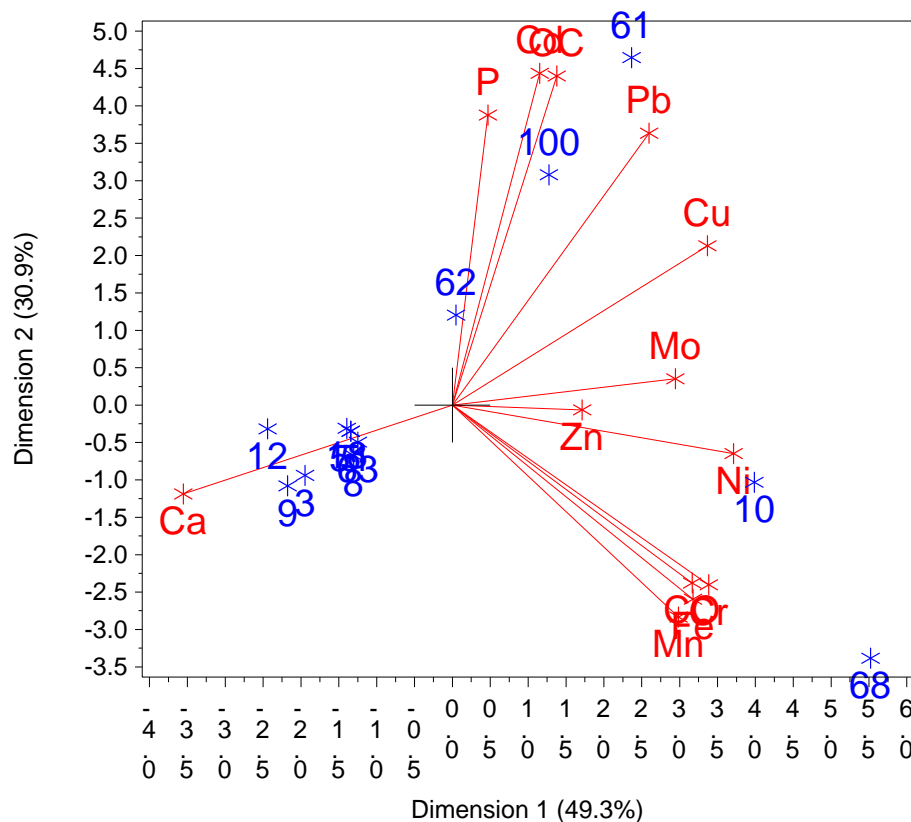


Abb. 14.18: (1) $A_2 = U_2 S_2$ und $B_2 = V_2$ [factype=JK in SAS-Makro %biplot]

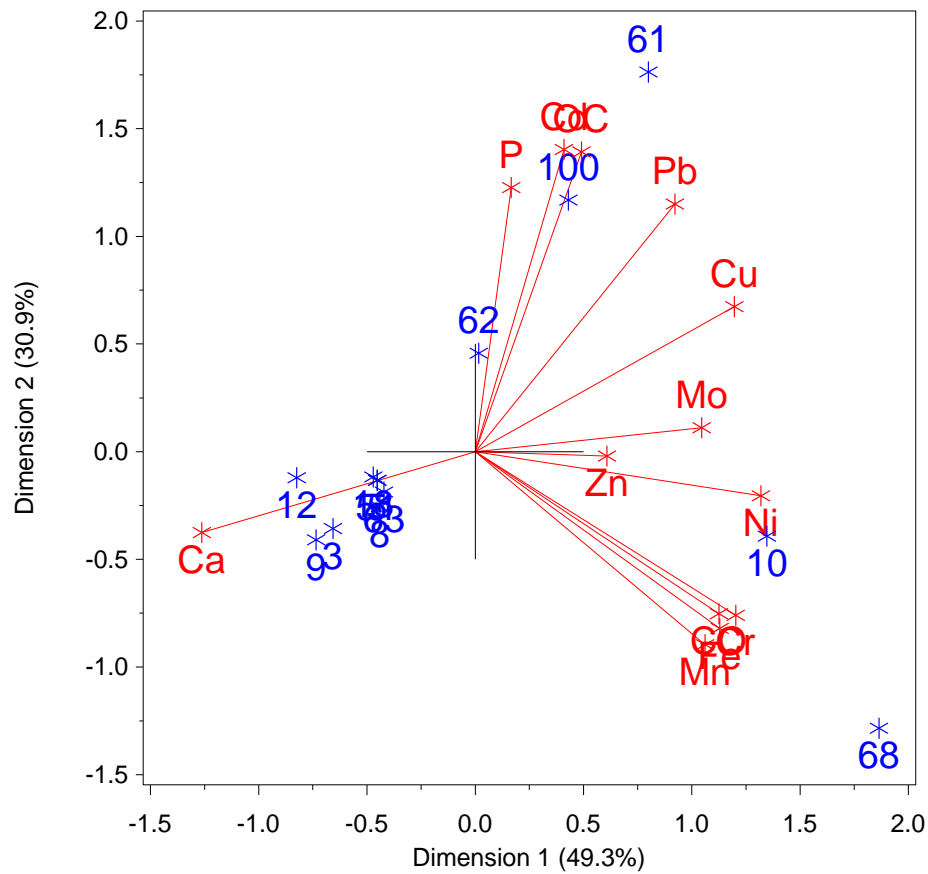


Abb. 14.18: (2) $A_2 = U_2 S_2^{1/2}$ und $B_2 = V_2 S_2^{1/2}$ [facttype=SYM in SAS-Makro %biplot]

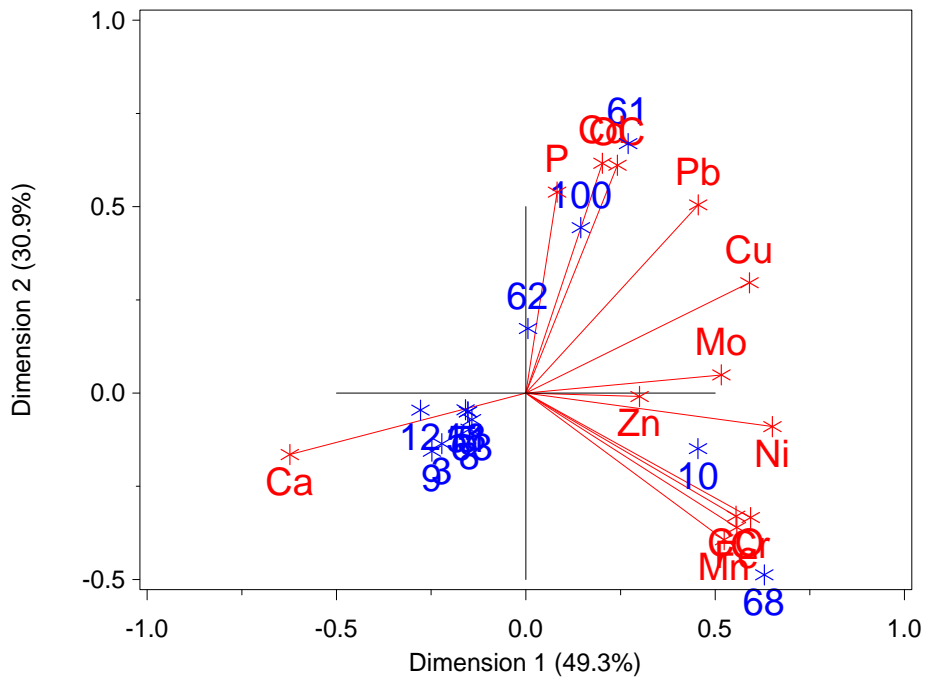


Abb. 14.19: (3) $A_2 = U_2$ und $B_2 = V_2 S_2$ [facttype=GH in SAS-Makro %biplot]

Man sieht, dass z.B. Ort 68 relativ hohe Gehalte an Mn und Co aufweist. Die Orte 100 und 61 haben dagegen relativ hohe P, Cd und OC (organic carbon)-Werte.

Sonstiges

PCA für Zähldaten: Mit Zähldaten sollte man keine PCA machen, v.a. wenn viele Nullen vorkommen. Dies führt oft zu Artefakten wie dem Hufeiseneffekt, bei dem die Punkte der Objekte auf einer stark gekrümmten Linie liegen. Besser ist eine **Korrespondenzanalyse** (correspondence analysis - CA).

Alternativen bei gegebener Distanzmatrix: Die PCA ist eine multivariate Methode, welche Euklidische Distanzen darstellen kann. Hat man eine andere Distanzmetrik, so kann man die sog. Hauptkoordinatenanalyse (Principal Coordinate Analysis - PCO, PCoA) verwenden, um eine graphische Darstellung wie bei der PCA zu erhalten. Eine PCO ist z.B. für die Visualisierung genetischer Distanzen interessant, die keine Euklidische Distanzen sind.

Weitere Alternativen: Es gibt viele der PCA verwandte Methoden, z.B.
- Multidimensional scaling (MDS) \Rightarrow PCA, PCO und CA sind Spezialfälle von MDS
- Kanonische Korrespondenzanalyse (Canonical correspondence analysis)

Zusammenfassung

- PCA und verwandte Methoden eignen sich zur Exploration und Visualisierung multivariater Daten
- Im Zweifelsfall sollten immer alle Variablen standardisiert werden (Korrelationsmatrix statt Kovarianzmatrix)
- Achtung bei der Skalierung von Punkten und Pfeilen in Biplots – Es gibt verschiedene Optionen mit verschiedenen Optionen
- Beide Hauptachsen sollten immer gleich skaliert werden
- Es sollte immer angegeben werden, welcher Teil der Varianz durch die dargestellten Hauptkomponenten erklärt wird

*14.4.2 Mathematische Details zur Berechnung der Hauptkomponenten

Die Berechnung der Koordinaten der Hauptkomponenten erfordert eine sog. Eigenwertanalyse, die im folgenden näher erläutert wird, und die auf zwei äquivalenten Wegen durchgeführt werden kann (siehe Digby und Kempton, 1987), die unten als *Methode 1* und *Methode 2* angesprochen werden.

Beispiel 1: Zur Erläuterung betrachten wir den einfachen Fall von nur zwei Merkmalen. Hierzu werden nur die Merkmale v1 (Streichfähigkeit) und v10 (Natürlichkeit) herangezogen, die besonders hoch korreliert sind. Abb. 14.20 zeigt den Plot von v10 gegen v1, wobei jede Margarine durch einen Punkt repräsentiert wird. Eine Rotation des Koordinatensystems mittels einer Hauptkomponentenanalyse führt dazu, dass die Varianz entlang der Abszisse (Hauptkomponente 1) maximiert wird (Abb. 14.21).

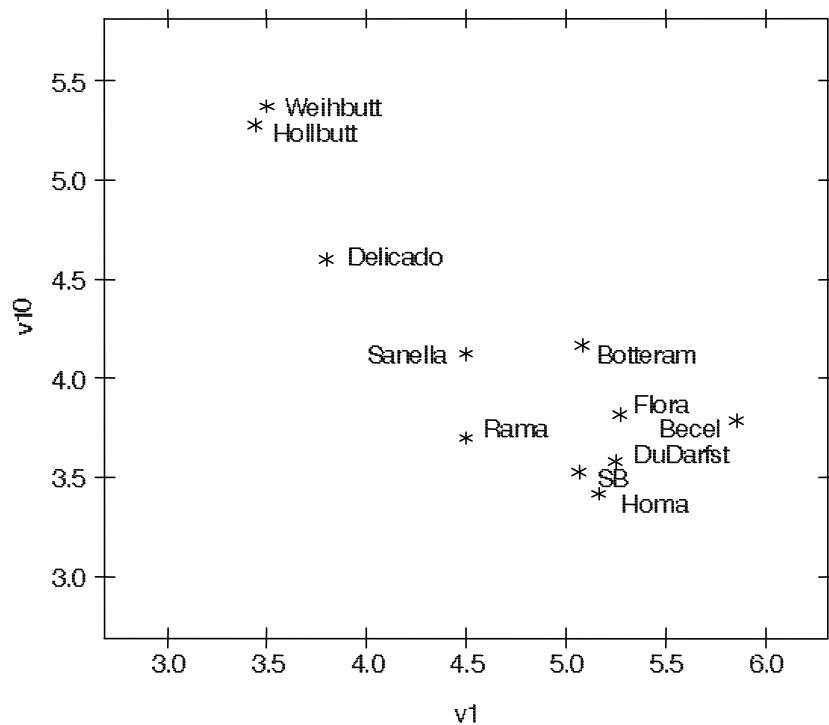


Abb. 14.20: Streudiagramm von v10 gegen v1 bei Margarine-Daten (Beispiel 1).

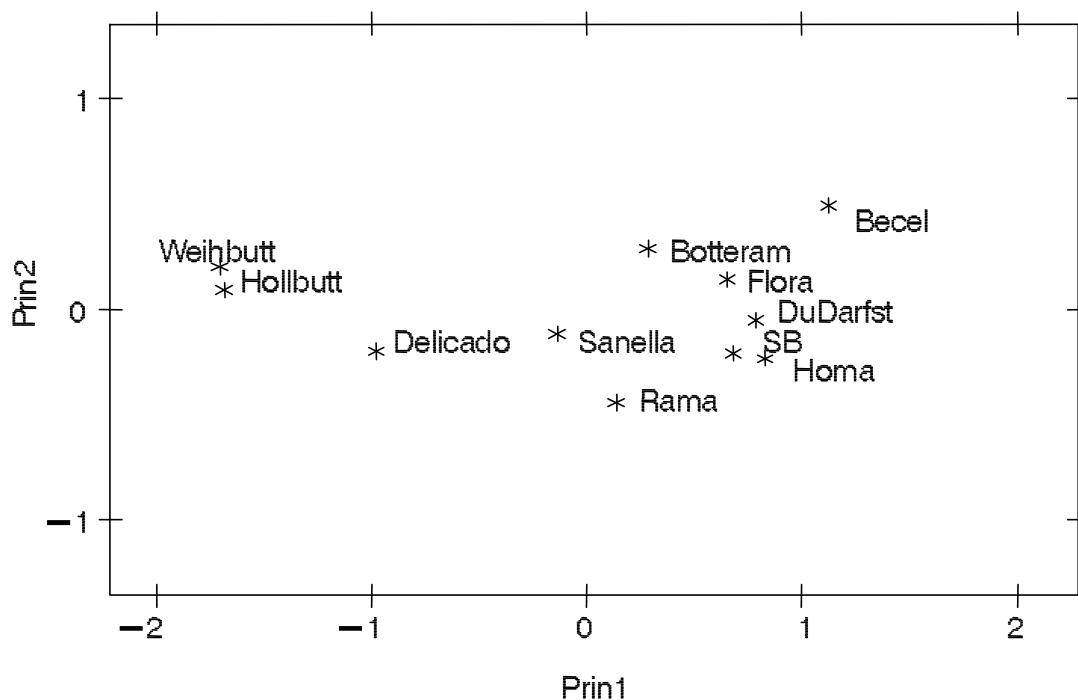


Abb. 14.21: Streudiagramm der beiden Hauptkomponenten (Prin1 und Prin2) basierend auf v10 und v1 bei Margarine-Daten (Beispiel 1).

Nun zur Berechnung der Hauptkomponenten. Üblicherweise werden die Daten zunächst standardisiert, entweder durch Subtraktion des Merkmalsmittelwertes oder zusätzlich durch Division durch die Standardabweichung. Hier subtrahieren wir nur den Merkmalsmittelwert, da die Merkmale auf derselben Skala erfasst wurden. Subtraktion des Mittelwertes ist sinnvoll, da andernfalls die erste Hauptkomponente

vor allem Mittelwertunterschiede abbilden würde. Die Standardisierung in Skalarform ist

$$z_{ik} = y_{ik} - \bar{y}_{\bullet k}$$

In anderen Fällen würden wir hier außerdem durch die Standardabweichung teilen (siehe Beispiel 3). Die Beobachtungen y_{ik} der 11 Margarine für die beiden Merkmale können wir in eine Matrix schreiben, ebenso die standardisierten Werte z_{ik} :

$$Y = \begin{pmatrix} 4.500 & 4.125 \\ 5.167 & 3.417 \\ 5.069 & 3.529 \\ 3.800 & 4.600 \\ 3.444 & 5.278 \\ 3.500 & 5.375 \\ 5.250 & 3.583 \\ 5.857 & 3.786 \\ 5.083 & 4.167 \\ 5.287 & 3.818 \\ 4.500 & 3.700 \end{pmatrix} \xrightarrow{\text{Standardisierung}} Z = \begin{pmatrix} -0.177 & +0.000 \\ +0.490 & -0.708 \\ +0.392 & -0.596 \\ -0.877 & +0.475 \\ -1.233 & +1.153 \\ -1.177 & +1.250 \\ +0.573 & -0.542 \\ +1.180 & -0.339 \\ +0.406 & +0.042 \\ +0.596 & -0.307 \\ -0.177 & -0.425 \end{pmatrix}$$

Nun zum Problem der optimalen Rotation des Koordinatensystems. Die Lösung erläutern wir der Einfachheit halber für den Fall von $p = 2$ Variablen.

Methode 1: Wir betrachten einen Punkt im Merkmalsraum, der durch die beiden Merkmale z_1 und z_2 gegeben ist (Abb. 14.22; siehe auch den konkreten Fall in Abb. 14.20). Der Punkt $z_i = (z_{i1}, z_{i2})$ repräsentiert das i -te Produkt. Zusätzlich ist jedes andere Produkt durch einen Punkt repräsentiert. Das ursprüngliche Koordinatensystem hat die Achsen z_1 und z_2 . Wir suchen nun ein neues Koordinatensystem mit den orthogonalen Hauptachsen g_1 und g_2 , so dass die Punkte entlang der ersten Achse eine maximale Varianz aufweisen. Für das Margarine-Beispiel ist das Ergebnis der Projektion in Abb. 14.21 dargestellt, wobei dort das Bild bereits so gedreht ist, dass die 1. Hauptachse in x -Richtung und die 2. Hauptachse in y -Richtung zeigt. Nun zu Abb. 14.22. Wir haben hier die beiden neuen Achsen als Vektoren im alten Koordinatensystem dargestellt und daher fettgeschriebene Symbole verwendet. Wir wollen annehmen, dass g_1 und g_2 Einheitsvektoren sind, also die Länge 1 haben. Ebenso wird der Datenpunkt (z_{i1}, z_{i2}) für das i -te Produkt durch den Vektor z_i repräsentiert, der auf diesen Punkt zeigt. z_i entspricht der i -ten Zeile der Datenmatrix Z . Den Wert auf der g_1 -Achse für einen Punkt erhalten wir einfach durch orthogonale Projektion des Punktes auf diese Achse. Der Projektionspunkt auf den Einheitsvektor g_1 lässt sich durch einen Vektor b_{i1} repräsentieren.

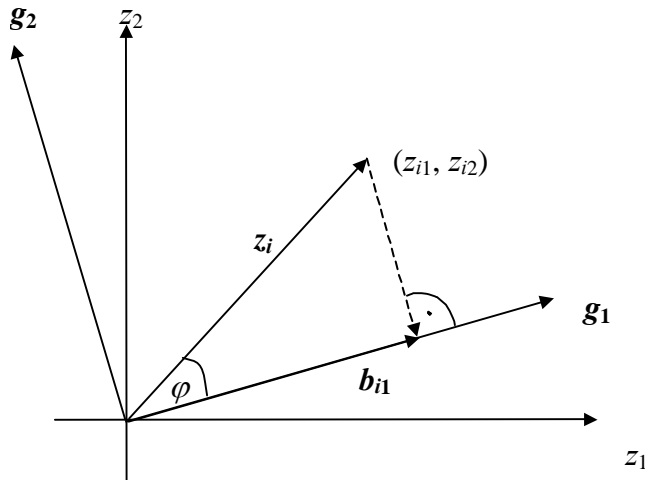


Abb. 14.22: Projektion des Punktes für das i -te Objekt (Produkt) auf die erste Hauptachse (\mathbf{g}_1) des rotierten Koordinatensystems.

Für die Länge $a_{i1} = |\mathbf{b}_{i1}|$ des Vektors \mathbf{b}_{i1} gilt nach den Regeln der linearen Algebra (Kosinus-Satz):

$$|z_i| |\mathbf{g}_1| \cos(\varphi) = |\mathbf{g}_1| |\mathbf{b}_{i1}| = |\mathbf{b}_{i1}| = a_{i1} = \mathbf{z}'_i \mathbf{g}_1 = \sum_{k=1}^p z_{ik} g_{1k}$$

Dies ergibt sich unter Beachtung der Tatsache, dass der Einheitsvektor \mathbf{g}_1 die Länge 1 hat. Die Länge $|\mathbf{b}_{i1}|$ entspricht dem Wert für den Punkt auf der ersten Hauptachse des neuen Koordinatensystems. Ziel der Hauptkomponentenanalyse ist es nun, die Varianz der Punkte entlang der ersten Hauptachse zu maximieren. Wegen der Zentrierung der Daten ist der Mittelwert aller Punkte auf jeder Hauptachse gleich Null. Die zu maximierende Varianz ist daher proportional zu

$$Q = \sum_{i=1}^n |\mathbf{b}_{i1}|^2 = \sum_{i=1}^n (\mathbf{z}_i^T \mathbf{g}_1)(\mathbf{z}_i^T \mathbf{g}_1) = \sum_{i=1}^n \mathbf{g}_1^T \mathbf{z}_i \mathbf{z}_i^T \mathbf{g}_1 = \mathbf{g}_1^T \mathbf{Z}^T \mathbf{Z} \mathbf{g}_1 = \mathbf{g}_1^T \boldsymbol{\Sigma} \mathbf{g}_1$$

wobei $\boldsymbol{\Sigma} = \mathbf{Z}^T \mathbf{Z}$ ist. Wir suchen also nun denjenigen Vektor \mathbf{g}_1 , für den Q maximal wird, wobei \mathbf{g}_1 ein Einheitsvektor sein soll, also ein Vektor der Länge Eins. Mathematisch gesehen haben wir hier ein Maximierungsproblem unter einer Nebenbedingung, die wie folgt formuliert werden kann:

$$\mathbf{g}_1^T \mathbf{g}_1 = 1 \quad (\text{Vektor } \mathbf{g}_1 \text{ hat Länge } 1)$$

Ein solches Problem kann mit Hilfe eines Tricks gelöst werden, der sog. Methode von Lagrange (siehe Anhang C). Hierzu maximieren wir

$$Q^* = \mathbf{g}_1^T \boldsymbol{\Sigma} \mathbf{g}_1 - \lambda (\mathbf{g}_1^T \mathbf{g}_1 - 1)$$

bezüglich der Parameter \mathbf{g}_1 und λ , dem sog. Lagrange-Multiplikator. Die Maximierung erfolgt durch Berechnen der 1. Ableitung nach den Parametern, Nullsetzen und Auflösen. Der Trick besteht darin, dass die 1. Ableitung nach λ und Nullsetzen

gerade die Nebenbedingung $\mathbf{g}_1^T \mathbf{g}_1 = 1$ liefert. Im Ergebnis wird die Maximierung also so vorgenommen, dass die Erfüllung der Nebenbedingung garantiert ist. Die 1. Ableitung nach \mathbf{g}_1 liefert:

$$(\boldsymbol{\Sigma} - \lambda \mathbf{I})\mathbf{g}_1 = \mathbf{0} \quad (14.1)$$

Diese Beziehung heißt **charakteristische Gleichung**. Es handelt sich um ein homogenes Gleichungssystem. Eine triviale Lösung dieses Gleichungssystems ist $\mathbf{g}_1 = \mathbf{0}$; an der sind wir aber nicht interessiert. Damit ein homogenes Gleichungssystem nichttriviale Lösungen hat, muss die Matrix $\mathbf{M} = \boldsymbol{\Sigma} - \lambda \mathbf{I}$ die Determinante Null haben. Andernfalls wäre \mathbf{M} invertierbar, so dass das Gleichungssystem die eindeutige (aber triviale) Lösung $\mathbf{g}_1 = \mathbf{0}$ hätte (siehe Lehrbücher zur linearen Algebra und Matrizenrechnung, z.B. Luh, 1982):

$$\det(\boldsymbol{\Sigma} - \lambda \mathbf{I}) = 0 \quad (14.2)$$

Die Determinante ist ein Polynom p -ten Grades in λ . Die Lösungen $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ der Gleichung heißen **Eigenwerte**. Einsetzen des größten Eigenwertes λ_1 in (14.1) und Auflösen nach \mathbf{g}_1 liefert die entsprechende Lösung für \mathbf{g}_1 , die als **Eigenvektor** bezeichnet wird (siehe Mathematik-Vorlesung!). Der erste Eigenvektor ist die gesuchte erste Hauptachse.

Beispiel: Für die Magarine-Daten finden wir für die Variablen v1 und v10 die folgende Varianz-Kovarianz-Matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0,63719 & -0,46913 \\ -0,46913 & 0,46593 \end{pmatrix}$$

$$\begin{aligned} \det(\boldsymbol{\Sigma} - \lambda \mathbf{I}) &= \det \begin{pmatrix} 0,63719 - \lambda & -0,46913 \\ -0,46913 & 0,46593 - \lambda \end{pmatrix} = (0,63719 - \lambda)(0,46593 - \lambda) - (-0,46913)^2 \\ &= \lambda^2 - 1,10313\lambda + 0,076807 = \lambda^2 + p\lambda + q = 0 \end{aligned}$$

Diese quadratische Gleichung hat folgende 2 Lösungen:

$$\lambda_{1,2} = -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q} \Rightarrow$$

$$\lambda_1 = 1,02845 \quad \text{und} \quad \lambda_2 = 0,074683$$

Der erste Eigenvektor ergibt sich wie folgt aus der Bedingung $(\boldsymbol{\Sigma} - \lambda \mathbf{I})\mathbf{g}_1 = \mathbf{0}$:

$$(\Sigma - \lambda I) \mathbf{g}_1 = \begin{pmatrix} 0,63719 - 1,02845 & -0,46913 \\ -0,46913 & 0,46593 - 1,02845 \end{pmatrix} \begin{pmatrix} g_{11} \\ g_{12} \end{pmatrix} = \begin{pmatrix} -0,39125 & -0,46913 \\ -0,46913 & -0,56251 \end{pmatrix} \begin{pmatrix} g_{11} \\ g_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Jedes Wertepaar (g_{11}, g_{12}) , welches z.B. die Gleichung $-0,39125g_{11} - 0,46913g_{12} = 0$ erfüllt, ist eine Lösung. Mit der Restriktion, dass \mathbf{g}_1 die Länge Eins hat, dass also $g_{11}^2 + g_{12}^2 = 1$ ist, findet man

$$\mathbf{g}_1 = \begin{pmatrix} 0,76797 \\ -0,64048 \end{pmatrix}.$$

Die Streuung entlang der ersten Hauptachse ist gegeben durch

$$Q = \mathbf{g}_1^T \Sigma \mathbf{g}_1 = \mathbf{g}_1^T \lambda_1 I \mathbf{g}_1 = \lambda_1$$

Der Eigenwert einer Achse ist also ein Maß für die Varianz entlang dieser Achse. In anderen Worten: der Eigenwert erfasst den durch die betreffende Hauptachse erklärte Varianz.

Einsetzen der anderen Eigenwerte liefert weitere Eigenvektoren $\mathbf{g}_2, \dots, \mathbf{g}_p$, die die Eigenschaft haben, zueinander orthogonal zu sein. Im Fall von nur zwei Merkmalen ($p = 2$) ist \mathbf{g}_2 die zweite der beiden gesuchten Hauptachsen (Abb. 14.22). Liegen mehr als zwei Merkmale vor, so gibt es entsprechend weitere Eigenwerte und Eigenvektoren (Hauptachsen).

Nun können wir die Koordinaten der Produkte bezüglich der ersten Hauptachse für das i -te Produkt berechnen nach

$$a_{i1} = \mathbf{z}_i^T \mathbf{g}_1$$

Eine kompakte Berechnung der 1. Hauptkomponente für alle Produkte ergibt sich durch

$$\mathbf{a}_1 = \mathbf{Z} \mathbf{g}_1$$

Analog können wir die anderen Hauptkomponenten durch Projektion auf die anderen Hauptachsen erhalten. Schließlich können wir alle Hauptkomponenten in eine Matrix schreiben:

$$\mathbf{A} = (\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_p) \quad (\text{Scores})$$

Ebenso sammeln wir die Eigenvektoren in einer Matrix:

$$\mathbf{\Gamma} = (\mathbf{g}_1 \mathbf{g}_2 \dots \mathbf{g}_p) \quad (\text{Ladungen})$$

Die Matrix Γ enthält die Koordinaten der Einheitsvektoren für das neue Koordinatensystem im alten Koordinatensystem. Hiermit finden wir alle Hauptkomponenten im auf die Hauptachsen rotierten Koordinatensystem durch

$$A = Z\Gamma$$

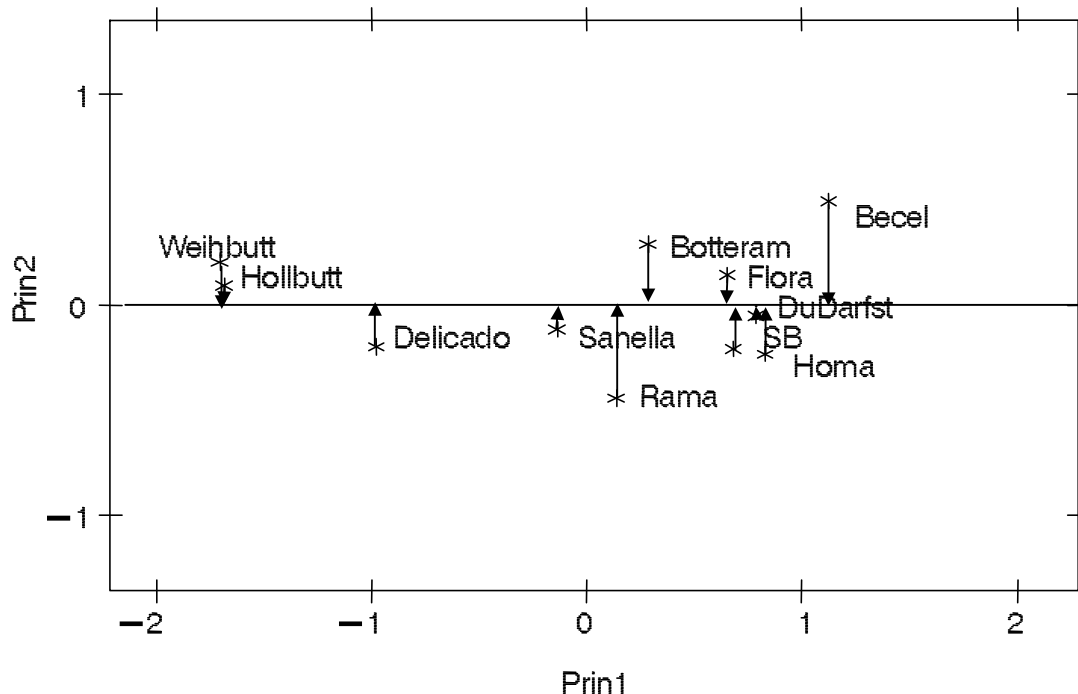


Abb. 14.23: Streudiagramm der beiden Hauptkomponenten für PCA basierend auf v10 und v1 bei Margarine-Daten (Beispiel 1); Projektion der Punkte auf die erste Hauptachse (Prin1; g_1).

Durch die Rotation des Koordinatensystems allein wird noch keine Reduktion der Dimensionalität erzielt. Dies ist im vorliegenden einfachen Beispiel auch eigentlich nicht notwendig, da nur zwei Dimensionen vorliegen (Merkmale v1 und v10) und diese problemlos graphisch dargestellt werden können. Zur Erläuterung können wir trotzdem eine Reduktion auf eine Dimension betrachten. Hierzu brauchen wir nur die Punkte auf die Abszisse (1. Hauptkomponente) zu projizieren (Abb. 14.23). Da der größte Anteil der gesamten Streuung entlang der Abszisse verläuft, geht durch die Projektion auf eine Dimension kaum Information verloren. Der Anteil der erklärten Varianz lässt sich mit den Eigenwerten quantifizieren (siehe oben). Wenn von den insgesamt p Koordinaten die ersten $q \leq p$ Koordinaten zur Projektion verwendet werden, so ist der Anteil der erklärten Varianz gleich

$$I = \frac{\sum_{k=1}^q \lambda_k}{\sum_{k=1}^p \lambda_k} 100\%$$

Im vorliegenden Fall mit nur zwei Variablen/Merkmalen (Abb. 14.22) ist $I = 3.207^2 / (3.207^2 + 0.864^2) 100\% = 93,23\%$. Der Informationsverlust durch Projektion auf die erste Hauptachse ist also gering.

Im allgemeinen ist die beste Projektion auf q Dimensionen gegeben durch die ersten q Spalten von $A = Z\Gamma$. Diese verwenden wir für eine graphische Darstellung. In der Regel wird $q = 2$ gewählt. Für $q = 3$ ist die Darstellung etwas komplexer und daher nicht gebräuchlich. Für $q > 3$ ist keine graphische Darstellung in einer Abbildung mehr möglich. Die Projektion in q Dimensionen mittels Hauptkomponentenanalyse ist die beste Darstellung in dem Sinne, dass ein Maximum an Information aus den p Dimensionen erhalten bleibt. Im allgemeinen Fall können wir die Koordinaten der Projektion auf die q ersten Hauptkomponenten auch ausdrücken als

$$A_q = Z\Gamma_q,$$

wobei Γ_q die Sub-Matrix mit den ersten q Spalten von Γ , also den ersten q Eigenvektoren ist. Die Koordinaten im rotierten Koordinatensystem heißen auch **Scores**. Für die kompletten Margarinedaten finden wir für die beiden ersten Hauptkomponenten:

$$\Gamma_2 = \begin{pmatrix} -0,319 & +0,380 \\ +0,091 & -0,547 \\ -0,109 & +0,320 \\ -0,063 & +0,212 \\ +0,189 & +0,540 \\ +0,226 & +0,168 \\ +0,330 & +0,160 \\ -0,763 & -0,037 \\ +0,103 & +0,232 \\ +0,294 & -0,087 \end{pmatrix} \begin{matrix} v1 \\ v2 \\ v3 \\ v4 \\ v5 \\ v6 \\ v7 \\ v8 \\ v9 \\ v10 \end{matrix} \quad A_2 = \begin{pmatrix} -0,937 & -0,360 \\ -1,918 & -1,274 \\ -1,226 & +0,395 \\ +2,111 & -0,685 \\ +3,471 & -0,794 \\ +3,906 & +0,704 \\ +0,101 & +1,188 \\ -1,327 & +0,608 \\ -1,386 & +0,173 \\ -1,237 & +0,495 \\ -1,556 & -0,449 \end{pmatrix} \begin{matrix} Sanella \\ Homa \\ SB \\ Delicado \\ Hollbutt \\ Weihbutt \\ DuDarfst \\ Becel \\ Botteram \\ Flora \\ Rama \end{matrix}$$

$$\lambda_1 = 4,547$$

$$\lambda_2 = 0,576$$

$$\lambda_3 = 0,561$$

$$\lambda_4 = 0,381$$

$$\lambda_5 = 0,137$$

$$\lambda_6 = 0,106$$

$$\lambda_7 = 0,075$$

$$\lambda_8 = 0,024$$

$$\lambda_9 = 0,012$$

$$\lambda_{10} = 0,002$$

$$\sum_{m=1}^{10} \lambda_m = \lambda_1 + \lambda_2 + \dots + \lambda_{10} = 6,422$$

Die beiden ersten Hauptkomponenten erklären 79,8% der Variation, denn

$$I = (4,547 + 0,576) / 6,422 = 0,798 = 79,8\%.$$

Dies stellt noch einen relativ guten Informationserhalt dar. Die Matrix A_2 enthält die beiden Koordinaten der in Abb. 14.6 dargestellten Punkte. Die Einträge beider Eigenvektoren (Matrix Γ) sind die **Ladungen** der beiden neuen Koordinaten für die

verschiedenen Variablen (Merkmale). Die Ladungen messen den Einfluss einer Variable auf die jeweilige neue Koordinate. Dabei kommt es auf den Betrag an, nicht auf das Vorzeichen. Bei der ersten Hauptkomponente haben die Merkmale v_1 , v_7 und v_8 die höchsten Ladungen, dominieren also den Wert der ersten Hauptkomponente. Dagegen laden v_2 und v_5 besonders hoch auf der 2. Hauptkomponente.

Spektralzerlegung: Die hier beschriebene Eigenwertanalyse ist eng verknüpft mit einer sog. *Spektralzerlegung* der Matrix der Summe der Kreuzprodukte

$$\Sigma = Z^T Z$$

Diese ist gegeben durch

$$\Sigma = \Gamma \Lambda \Gamma^T$$

wobei Γ die $(p \times p)$ *orthogonale* Matrix ($\Gamma \Gamma^T = \Gamma^T \Gamma = I$) der Eigenvektoren ist und Λ eine Diagonal-Matrix p -ter Ordnung mit Diagonalelementen $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ (Eigenwerte) ist.

Methode 2: Zur Berechnung der Hauptkomponenten kann eine sog. *Singulärwertzerlegung* der standardisierten Datenmatrix herangezogen werden:

$$Z = USV^T$$

wobei U eine $(n \times p)$ *orthonormale* Matrix, S eine Diagonalmatrix p -ter Ordnung, und V' die transponierte einer $(p \times p)$ *orthogonalen* Matrix. Hierbei wird $n \geq p$ angenommen. U und V haben folgende Eigenschaften:

$U^T U = I =$ Einheitsmatrix p -ter Ordnung (U ist orthonormal)

$V V^T = V^T V = I =$ Einheitsmatrix p -ter Ordnung (V ist orthogonal)

Die Diagonalelemente von S heißen *Singulärwerte* von Z und sind per Konvention immer so sortiert, dass

$$s_1 \geq s_2 \geq \dots \geq s_p \geq 0$$

Betrachten wir nur die beiden Merkmale v_1 und v_{10} , so finden wir

$$\begin{pmatrix} -0.177 & \pm 0.000 \\ +0.490 & -0.708 \\ +0.392 & -0.596 \\ -0.877 & +0.475 \\ -1.233 & +1.153 \\ -1.177 & +1.250 \\ +0.573 & -0.542 \\ +1.180 & -0.339 \\ +0.406 & +0.042 \\ +0.596 & -0.307 \\ -0.177 & -0.425 \end{pmatrix} = \begin{pmatrix} -0.042 & -0.131 \\ +0.259 & -0.266 \\ +0.213 & -0.239 \\ -0.305 & -0.228 \\ -0.525 & +0.111 \\ -0.531 & +0.239 \\ +0.246 & -0.057 \\ +0.350 & +0.573 \\ +0.089 & +0.338 \\ +0.204 & +0.169 \\ +0.043 & -0.509 \end{pmatrix} \begin{pmatrix} 3.207 & 0 \\ 0 & 0.864 \end{pmatrix} \begin{pmatrix} +0.768 & +0.640 \\ -0.640 & +0.768 \end{pmatrix}$$

$$\mathbf{Z} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$

Die Matrix \mathbf{V} definiert eine Rotation der ursprünglichen Koordinatenachsen zu einem neuen Koordinatensystem. Die Rotation wird durchgeführt, indem die Datenmatrix \mathbf{Z} mit der Rotationsmatrix \mathbf{V} multipliziert wird. Dies liefert die Koordinaten der Datenpunkte im neuen Koordinatensystem, $\mathbf{A} = \mathbf{ZV}$. In unserem Beispiel ist

$$\mathbf{A} = \begin{pmatrix} -0.135 & -0.113 \\ +0.830 & -0.230 \\ +0.683 & -0.207 \\ -0.977 & -0.197 \\ -1.685 & +0.096 \\ -1.704 & +0.206 \\ +0.788 & -0.049 \\ +1.124 & +0.495 \\ +0.285 & +0.292 \\ +0.655 & +0.146 \\ +0.137 & -0.440 \end{pmatrix}$$

Die Koordinaten im neuen Koordinatensystem werden auch als **Scores** bezeichnet. In skalarer Form haben die Scores folgende Form:

$$a_{ik} = z_{i1}v_{1k} + z_{i2}v_{2k} + \dots + z_{ip}v_{pk} = \sum_{m=1}^p z_{im}v_{mk}$$

Auf der ersten Hauptachse hat das sechste Produkt (Weihnachtsbutter) den niedrigsten Wert (-1,704), während das achte Produkt (Becel) den höchsten Wert hat (1,124). Diese beiden Produkte liegen daher am weitesten entfernt auf der ersten Hauptachse, wie man in Abb. 14.23 sehen kann. Übrigens gilt

$$A = ZV = USV^T V = US,$$

da $V^T V = I$ ist (V ist orthogonal). Somit können die Scores anhand der Singulärwertzerlegung auf zweierlei Weise berechnet werden. Die zweite, nämlich $A = US$, ist manchmal vorteilhafter.

Im allgemeinen ist die beste Projektion auf q Dimensionen gegeben durch die ersten q Spalten von $A = ZV = US$. Diese verwenden wir für eine graphische Darstellung. Die Projektion ist die beste Darstellung in dem Sinne, dass ein Maximum an Information aus den p Dimensionen erhalten bleibt. Im allgemeinen Fall können wir die Koordinaten der Projektion auf die q ersten Hauptkomponenten auch ausdrücken als

$$A_q = U_q S_q,$$

wobei U_q die Sub-Matrix mit den ersten q Spalten von U und S_q die Diagonalmatrix mit den ersten q Singulärwerten ist. Für die kompletten Margarinedaten finden wir:

$$U_2 = \begin{pmatrix} -0,139 & -0,150 \\ -0,284 & -0,531 \\ -0,182 & +0,165 \\ +0,313 & -0,286 \\ +0,515 & -0,331 \\ +0,579 & +0,293 \\ +0,015 & +0,495 \\ -0,197 & +0,253 \\ -0,206 & +0,072 \\ -0,184 & +0,206 \\ -0,231 & -0,187 \end{pmatrix}, S_2 = \begin{pmatrix} 6,743 & 0 \\ 0 & 2,401 \end{pmatrix}, A_2 = \begin{pmatrix} -0,937 & -0,360 \\ -1,918 & -1,274 \\ -1,226 & +0,395 \\ +2,111 & -0,685 \\ +3,471 & -0,794 \\ +3,906 & +0,704 \\ +0,101 & +1,188 \\ -1,327 & +0,608 \\ -1,386 & +0,173 \\ -1,237 & +0,495 \\ -1,556 & -0,449 \end{pmatrix} \begin{matrix} \textit{Sanella} \\ \textit{Homa} \\ \textit{SB} \\ \textit{Delicado} \\ \textit{Hollbutt} \\ \textit{Weihbutt} \\ \textit{DuDarfst} \\ \textit{Becel} \\ \textit{Botteram} \\ \textit{Flora} \\ \textit{Rama} \end{matrix}$$

Die Matrix A_2 enthält die beiden Koordinaten der in Abb. 14.6 dargestellten Punkte.

Abschließende Bemerkung: Die beiden Methoden (Methode 1, Spektralzerlegung, und Methode 2, Singulärwertzerlegung,) führen zu identischen Ergebnissen. So gilt z.B. $\lambda_m = s_m^2$, $A = S^2$ und $\Gamma = V$, was sich leicht aus folgender Ableitung ergibt:

$$\Sigma = Z^T Z = \underset{\substack{\downarrow \\ \text{Singulärwertzerlegung}}}{V} \underset{\substack{\downarrow \\ \text{Singulärwertzerlegung}}}{S} \underset{\substack{\downarrow \\ \text{Singulärwertzerlegung}}}{U} \underset{\substack{\downarrow \\ \text{Singulärwertzerlegung}}}{V}^T = \underset{\substack{\downarrow \\ \text{Singulärwertzerlegung}}}{V} S^2 \underset{\substack{\downarrow \\ \text{Singulärwertzerlegung}}}{V}^T = \underset{\substack{\downarrow \\ \text{Singulärwertzerlegung}}}{\Gamma} A \underset{\substack{\downarrow \\ \text{Singulärwertzerlegung}}}{\Gamma}^T.$$

Näheres findet sich bei Digby und Kempton (1987).

*14.4.3 Biplots

Betrachten wir nun eine weitere Darstellungsform für die Hauptkomponentenanalyse in einer Ebene (Projektion in $q = 2$ Dimensionen), den sog. Biplot. Die Bezeichnung rührt daher, dass in einem Biplot die Objekte durch Punkte und die Variablen durch Vektoren repräsentiert sind. Die Punkte für die Variablen werden außerdem durch Linien (Vektoren) mit dem Ursprung verbunden. Durch orthogonale Projektion der Objekte auf den Vektor einer Variable kann man den Messwert für diese Variable bei dem betreffenden Projekt abschätzen: Je weiter von Ursprung der Projektionspunkt, desto höher der Messwert. Liegt der Projektionspunkt in der Verlängerung des Vektors auf der entgegengesetzten Seite des Ursprungs, so ist der Messwert um so geringer, je weiter vom Ursprung der Projektionspunkt ist.

Beispiel 1: In Abb. 14.24 ist ein Biplot für die Variablen v1 (Streichfähigkeit) und v10 (Natürlichkeit) bei Margarine (Beispiel 1) dargestellt. Weihnachtsbutter hat unter allen Objekten einen Projektionspunkt auf den Vektor für v10, der am weitesten von allen vom Ursprung entfernt ist. Für v10 (Natürlichkeit) hatte diese Marke tatsächlich den höchsten Wert von allen (5,375). Becel hat dagegen einen Projektionspunkt auf den v10-Vektor, der auf der Verlängerung des Vektors auf der entgegengesetzten Seite des Ursprungs liegt. Tatsächlich hat Becel einen relativ niedrigen Wert für Natürlichkeit (3,786). Dafür hat Becel die "höchste" Projektion auf den v1-Vektor, also den höchsten Wert für Streichfähigkeit (5,857). Dagegen hat Weihnachtsbutter den zweitniedrigsten Wert (3,500), das an der "tiefen" Projektion auf den v1-Vektor ablesbar ist.

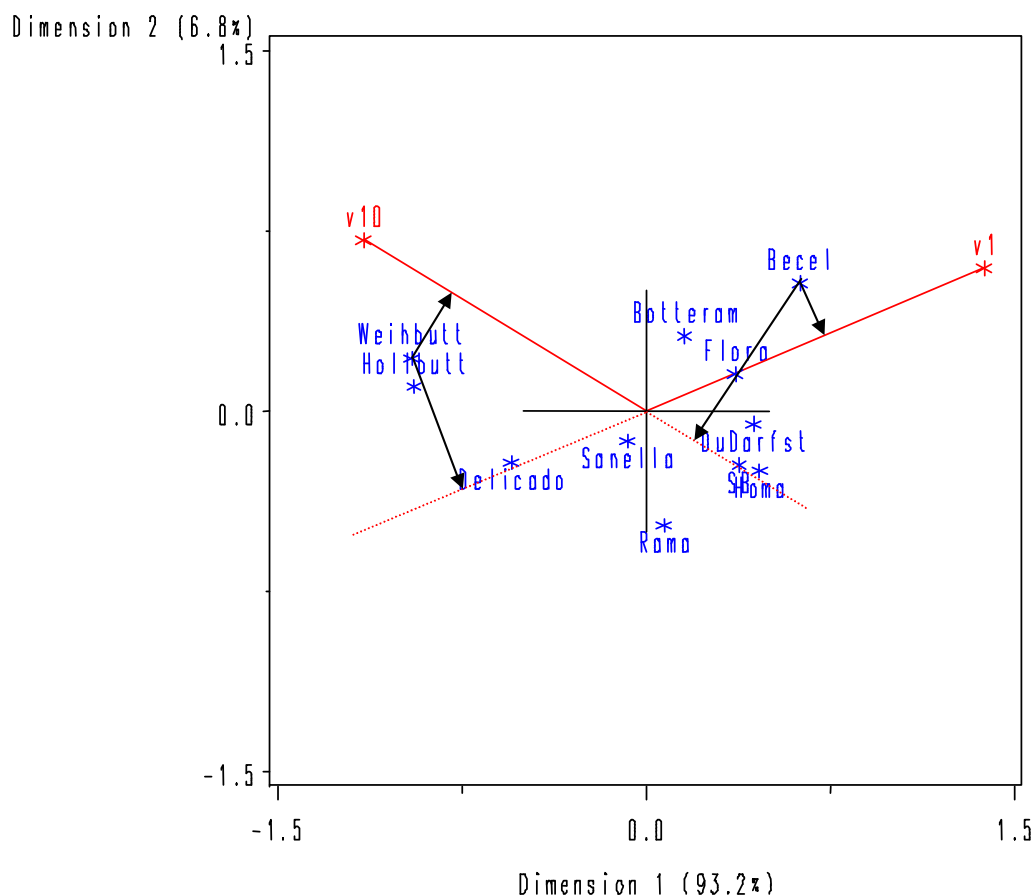


Abb. 14.24: Biplot für die Variablen v1 (Streichfähigkeit) und v10 (Natürlichkeit) bei Margarine (Beispiel 1).

In diesem Beispiel mit nur zwei Variablen hat der Biplot keinen Informationsverlust zur Folge, da alle $p = 2$ Dimensionen dargestellt werden können. Bei mehr als zwei Variablen ist mit einem gewissen Informationsverlust zu rechnen, so dass die Projektionen der Punkte auf die Vektoren die tatsächlichen Messwerte nicht exakt wiedergeben.

Abb. 14.25 zeigt den Biplot für alle 10 Variablen. Wir sehen dort, dass Weihnachtsbutter auch bei v_5 = Backeignung sehr gut abschneidet und für v_8 = Tierfett einen hohen Wert hat. Dies deckt sich mit den Rohdaten. Das Beispiel zeigt, dass ein Biplot einen sehr schnellen Blick auf die Rohdaten erlaubt und Besonderheiten direkt sichtbar macht, die einem bei der direkten Betrachtung der Rohdaten nicht so leicht auffallen würden. Dies gilt umso mehr, je größer der Datensatz ist.

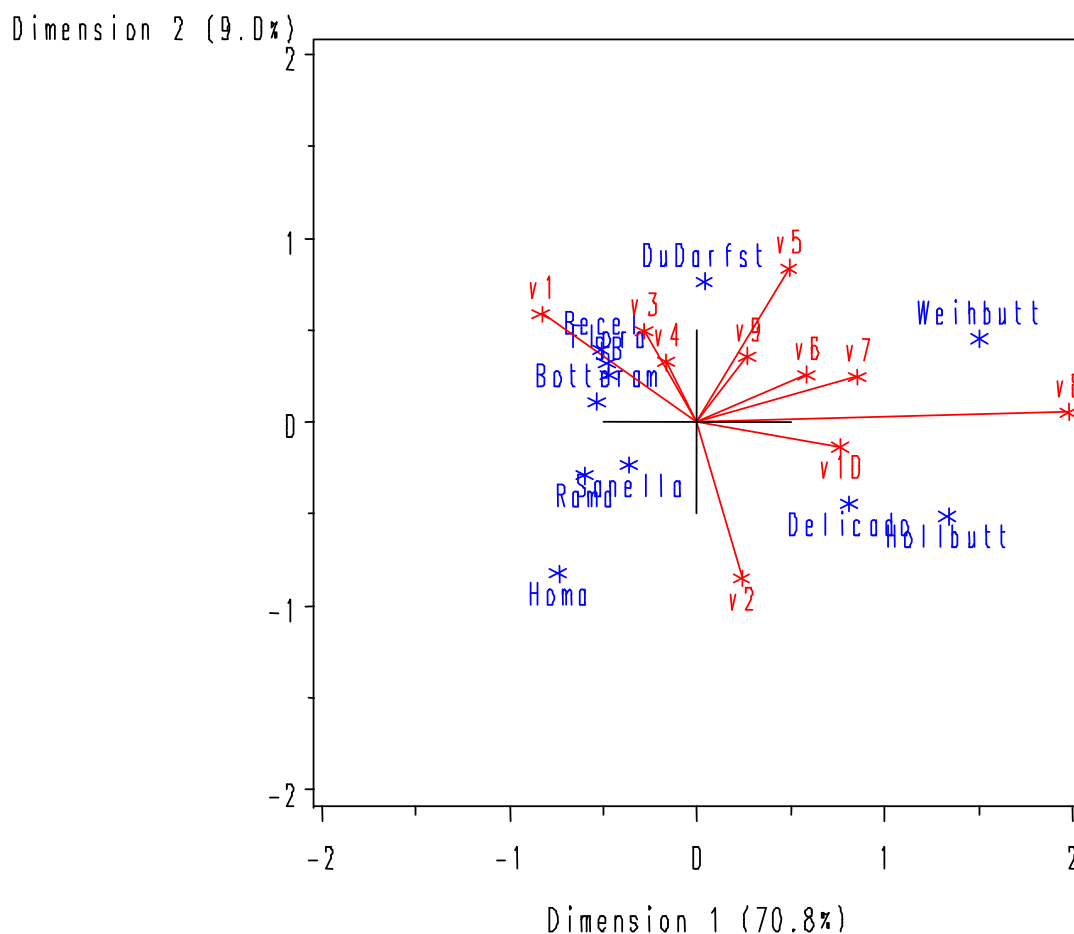


Abb. 14.25: Biplot für alle Variablen bei Margarine (Beispiel 1).

Nun zum theoretischen Hintergrund der Projektion. Die Punkte für die Objekte sind weiterhin gegeben durch

$$A_2 = U_2 S_2$$

während die Punkte für die Variablen gegeben sind durch

$$B_2 = V_2$$

Nun gilt nach der Singulärwertzerlegung

$$\mathbf{Z} = \mathbf{USV}^T .$$

Mit $\mathbf{A} = \mathbf{US}$ und $\mathbf{B} = \mathbf{V}$ können wir auch schreiben

$$\mathbf{Z} = \mathbf{AB}^T$$

Für einen einzelnen Wert gilt

$$z_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + \dots + a_{ip}b_{pk} = \sum_{m=1}^p a_{im}b_{mk} ,$$

wobei a_{im} und b_{mk} die Elemente von \mathbf{A} und \mathbf{B} sind. Diese Beziehung zeigt, dass die Daten aus \mathbf{A} und \mathbf{B} reproduziert werden können. Die Bedeutung dieser Tatsache liegt darin, dass man die Daten näherungsweise reproduzieren kann, wenn nur 2 Dimensionen verwendet werden:

$$\mathbf{Z} \approx \mathbf{Z}_2 = \mathbf{A}_2\mathbf{B}_2^T$$

Diese Approximation ist umso besser, je größer die ersten beiden Singulärwerte im Verhältnis zu den restlichen Singulärwerten sind (Im Extremfall, dass $s_3 = s_4 = \dots = s_p = 0$ ist, gilt sogar $\mathbf{Z} = \mathbf{A}_2\mathbf{B}_2^T$). Die Approximation für einen einzelnen Wert ist

$$z_{ik} \approx z_{ik}^{(2)} = a_{i1}b_{1k} + a_{i2}b_{2k}$$

Diese Approximation kann man als Skalarprodukt der Vektoren $\mathbf{a}_i = \begin{pmatrix} a_{i1} \\ a_{i2} \end{pmatrix}$ und

$\mathbf{b}_k = \begin{pmatrix} b_{k1} \\ b_{k2} \end{pmatrix}$ betrachten. Für dieses gilt:

$$\mathbf{a}_i^T \mathbf{b}_k = a_{i1}b_{1k} + a_{i2}b_{2k} = |\mathbf{a}_i| |\mathbf{b}_k| \cos(\theta)$$

wobei θ der Winkel der beiden Vektoren ist und " $|\cdot|$ " die Länge eines Vektors bezeichnet.

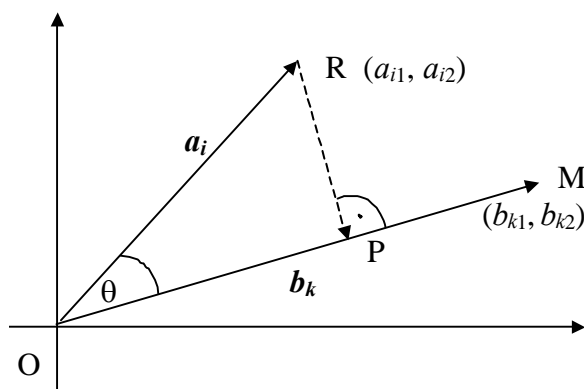


Abb. 14.26: Orthogonale Projektion des Punktes R (i -tes Objekt) auf die Gerade OM (Vektor zum Punkt für die k -te Variable).

In Abb. 14.26 ist nun die Projektion des Punktes für das i -te Produkt auf den k -ten Variablenvektor dargestellt. Anwendung der Regel für das Skalarprodukt liefert

$$a_{i1}b_{1k} + a_{i2}b_{2k} = |a_i| |b_k| \cos(\theta) = OM \cdot OR \cdot \cos(\theta) = OM \cdot OP$$

wobei wir die Beziehung

$$\cos(\theta) = \frac{OP}{OR}$$

genutzt haben. Nun ist für alle Produkte die Länge des Variablenvektors, OM, dieselbe. Also unterscheiden sich die Projektionen verschiedener Produkte nur durch die Länge der Strecke OP. OP ist aber der Abstand der Projektion des Produktpunktes auf den Variablenvektor vom Ursprung, und dieser ist nun proportional zu $z_{ik}^{(2)} = a_{i1}b_{1k} + a_{i2}b_{2k}$. Dies zeigt, dass der Abstand sich eignet, um den Messwert z_{ik} abzuschätzen (Digby und Kempton, 1987).

Zur graphischen Skalierung werden verschiedene Faktorisierungen verwendet:

- (1) $A_2 = U_2 S_2$ und $B_2 = V_2$ (Euklidische Distanzen für Produkte dargestellt)
- (2) $A_2 = U_2 S_2^{1/2}$ und $B_2 = V_2 S_2^{1/2}$ (keine Euklidischen Distanzen dargestellt)
- (3) $A_2 = U_2$ und $B_2 = V_2 S_2$ (Euklidische Distanzen für Variablen dargestellt)

Bei allen drei Faktorisierungen bleibt das oben bezüglich der Approximation von z_{ik} gesagte gültig. Allerdings sind im Biplot nur mit (1) Euklidische Distanzen der Produkte abgebildet, während (3) Distanzen für die Variablen darstellt. Mit (2) werden keine Distanzen dargestellt, nur die Projektion ist direkt interpretierbar. In den Abbildungen haben wir (2) verwendet, da sich hierbei Punkte für Produkte und Variablen in ähnlichen Abständen vom Ursprung bewegen.

Das eben gesagte wollen wir uns abschließend für den Fall (1) klar machen. Die Daten werden mit den beiden ersten Dimensionen approximiert nach

$$z_{ik}^{(2)} = u_{i1}s_1v_{1k} + u_{i2}s_2v_{2k}$$

Die quadrierte Euklidische Distanz zwischen den Produkten i und i' ist

$$\begin{aligned} ED^2 &= \sum_{k=1}^p (z_{ik} - z_{i'k})^2 \\ &= \sum_{k=1}^p (u_{i1}s_1v_{1k} + u_{i2}s_2v_{2k} - (u_{i'1}s_1v_{1k} + u_{i'2}s_2v_{2k}))^2 \\ &= \sum_{k=1}^p (u_{i1}s_1v_{1k} - u_{i'1}s_1v_{1k})^2 + \sum_{k=1}^p (u_{i2}s_2v_{2k} - u_{i'2}s_2v_{2k})^2 + 2 \sum_{k=1}^p (u_{i1}s_1v_{1k} - u_{i'1}s_1v_{1k})(u_{i2}s_2v_{2k} - u_{i'2}s_2v_{2k}) \\ &= (u_{i1}s_1 - u_{i'1}s_1)^2 \sum_{k=1}^p (v_{1k})^2 + (u_{i2}s_2 - u_{i'2}s_2)^2 \sum_{k=1}^p (v_{2k})^2 + 2(u_{i1}s_1 - u_{i'1}s_1)(u_{i2}s_2 - u_{i'2}s_2) \sum_{k=1}^p v_{1k}v_{2k} \\ &= (u_{i1}s_1 - u_{i'1}s_1)^2 + (u_{i2}s_2 - u_{i'2}s_2)^2 \end{aligned}$$

Hierbei haben wir die Tatsache genutzt, dass $\mathbf{v}_m = (v_{m1}, v_{m2}, \dots, v_{mp})^T$ ($m = 1, 2, \dots$) Eigenvektoren sind. Daher gilt:

$$\sum_{k=1}^p (v_{mk})^2 = 1 \quad \text{und} \quad \sum_{k=1}^p v_{1k} v_{2k} = 0.$$

Die Umformung zeigt im Rückschluss, dass sich bei der Faktorisierung $\mathbf{A}_2 = \mathbf{U}_2 \mathbf{S}_2$ (liefert Koordinaten $u_{1i} s_1$ und $u_{2i} s_2$ für das i -te Produkt!) Euklidische Abstände für die Produkte ergeben, was zu zeigen war.

Wichtig für alle geometrischen Interpretationen, die hier angesprochen wurden, ist, dass beide Achsen gleich skaliert sind. Andernfalls gelten diese Interpretationen nicht.

SAS Anweisungen

```
data daten;
input marke $ v1-v10;
datalines;
Sanella 4.500 4.000 4.375 3.875 3.250 3.750 4.000 2.000 4.625 4.125
Homa 5.167 4.250 3.833 3.833 2.167 3.750 3.273 1.857 3.750 3.417
SB 5.069 3.824 4.765 3.438 4.235 4.471 3.765 1.923 3.529 3.529
Delicado 3.800 5.400 3.800 2.400 5.000 5.000 5.000 4.000 4.000 4.600
Hollbutt 3.444 5.056 3.778 3.765 3.944 5.389 5.056 5.615 4.222 5.278
Weihbutt 3.500 3.500 3.875 4.000 4.625 5.250 5.500 6.000 4.750 5.375
DuDarfst 5.250 3.417 4.583 3.917 4.333 4.417 4.667 3.250 4.500 3.583
Becel 5.857 4.429 4.929 3.857 4.071 5.071 2.929 2.091 4.571 3.786
Botteram 5.083 4.083 4.667 4.000 4.000 4.250 3.818 1.545 3.750 4.167
Flora 5.273 3.600 3.909 4.091 4.091 4.091 4.545 1.600 3.909 3.818
Rama 4.500 4.000 4.200 3.900 3.700 3.900 3.600 1.500 3.500 3.700
;
proc princomp data=daten out=pca cov;
var v1-v10;
run;

%plotit(data=pca, labelvar=marke, plotvars=prin2 prin1);

%biplot(var=v1-v10, id=marke, data=daten, factype=sym, symbols=star);
```

Die Art der Faktorisierung wird im Makro %biplot mit der Option FACTYPE= festgelegt. Die Optionen sind wie folgt:

Faktorisierung	Option	Eigenschaften
(1) $\mathbf{A}_2 = \mathbf{U}_2 \mathbf{S}_2$ und $\mathbf{B}_2 = \mathbf{V}_2$	FACTYPE=JK	Euklidische Distanzen Produkte
(2) $\mathbf{A}_2 = \mathbf{U}_2 \mathbf{S}_2^{1/2}$ und $\mathbf{B}_2 = \mathbf{V}_2 \mathbf{S}_2^{1/2}$	FACTYPE=SYM	Keine Euklidischen Distanzen
(3) $\mathbf{A}_2 = \mathbf{U}_2$ und $\mathbf{B}_2 = \mathbf{V}_2 \mathbf{S}_2$	FACTYPE=GH	Euklidische Distanzen Variablen

Das %biplot Makro erhält man unter:

<http://euclid.psych.yorku.ca/ftp/sas/vcd/macros/biplot.sas> .

Zusammen mit diesem muss man außerdem das %equate Makro aktivieren, um gleiche Skalierung der Achsen zu erhalten.

Das %plotit Makro ist in der Standardinstallation von SAS enthalten. Um ein Makro in einer Sitzung verfügbar zu machen, muss man es in den Editor laden und "abschicken".

Literatur

Backhaus K, Erichson B, Plinke W, Weiber R 2000 Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. Springer, Berlin.

Digby PGN, Kempton RA 1987 Multivariate analysis of ecological communities. Chapman and Hall, London.

Gower JC 1971 A general coefficient of similarity and some of its properties. Biometrics 27, 857-872

Luh W 1982 Mathematik für Naturwissenschaftler. Akademieverlag Wiesbaden.

Ein sehr schönes Buch, welches die Anwendung statistischer Methoden in der Ernährungswissenschaft und Konsumentenforschung im Fokus hat, und dabei verschiedene Verfahren (multivariate Verfahren, lineare Modelle, nichtparametrische Verfahren) vorstellt und erläutert:

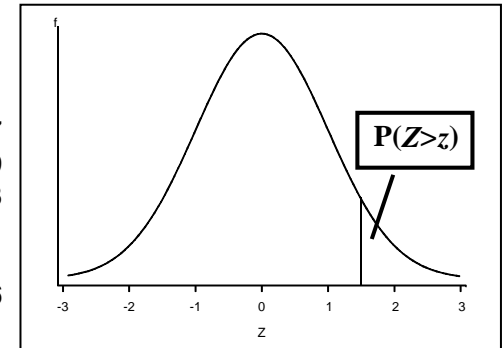
Gacula M Jr, Dingh J, Bi J, Atlan S 2009 Food and consumer research. Second edition. Wiley, New York.

Anhang A: Tabellen wichtiger Verteilungen

I.	Standardnormalverteilung - $P(Z > z)$	384
II.	t-Verteilung (zweiseitig)	385
II(b).	t-Verteilung (einseitig)	386
III.	Standardnormalverteilung - Quantile	387
IV.	Chi-Quadrat-Verteilung	388
V.	F-Verteilung	389
VI.	F-Verteilung	390
VII.	Studentisierte Variationsbreiten	391
VIII.	Chi-Quadrat-Verteilung	392

Tab. I: Überschreitungswahrscheinlichkeiten $P(Z > z)$ für die Standardnormalverteilung, Beispiel: $P(Z > 1,96) = 0,025$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014



Tab. II: Kritische Werte der t-Verteilung (zweiseitig)

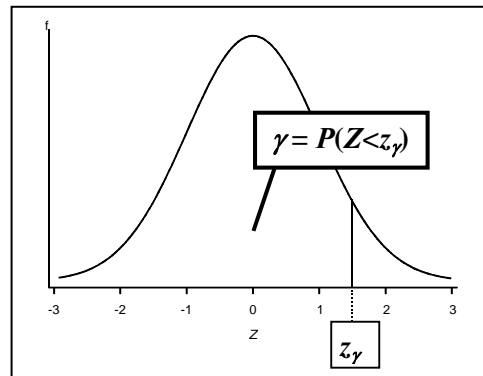
Irrtumswahrscheinlichkeit α				Irrtumswahrscheinlichkeit α			
α FG	0,05	0,01	0,001	α FG	0,05	0,01	0,001
1	12,706	63,657	636,619	51	2,008	2,676	3,492
2	4,303	9,925	31,599	52	2,007	2,674	3,488
3	3,182	5,841	12,924	53	2,006	2,672	3,484
4	2,776	4,604	8,610	54	2,005	2,670	3,480
5	2,571	4,032	6,869	55	2,004	2,668	3,476
6	2,447	3,707	5,959	56	2,003	2,667	3,473
7	2,365	3,499	5,408	57	2,002	2,665	3,470
8	2,306	3,355	5,041	58	2,002	2,663	3,466
9	2,262	3,250	4,781	59	2,001	2,662	3,463
10	2,228	3,169	4,587	60	2,000	2,660	3,460
11	2,201	3,106	4,437	61	2,000	2,659	3,457
12	2,179	3,055	4,318	62	1,999	2,657	3,454
13	2,160	3,012	4,221	63	1,998	2,656	3,452
14	2,145	2,977	4,140	64	1,998	2,655	3,449
15	2,131	2,947	4,073	65	1,997	2,654	3,447
16	2,120	2,921	4,015	66	1,997	2,652	3,444
17	2,110	2,898	3,965	67	1,996	2,651	3,442
18	2,101	2,878	3,922	68	1,995	2,650	3,439
19	2,093	2,861	3,883	69	1,995	2,649	3,437
20	2,086	2,845	3,850	70	1,994	2,648	3,435
21	2,080	2,831	3,819	71	1,994	2,647	3,433
22	2,074	2,819	3,792	72	1,993	2,646	3,431
23	2,069	2,807	3,768	73	1,993	2,645	3,429
24	2,064	2,797	3,745	74	1,993	2,644	3,427
25	2,060	2,787	3,725	75	1,992	2,643	3,425
26	2,056	2,779	3,707	76	1,992	2,642	3,423
27	2,052	2,771	3,690	77	1,991	2,641	3,421
28	2,048	2,763	3,674	78	1,991	2,640	3,420
29	2,045	2,756	3,659	79	1,990	2,640	3,418
30	2,042	2,750	3,646	80	1,990	2,639	3,416
31	2,040	2,744	3,633	81	1,990	2,638	3,415
32	2,037	2,738	3,622	82	1,989	2,637	3,413
33	2,035	2,733	3,611	83	1,989	2,636	3,412
34	2,032	2,728	3,601	84	1,989	2,636	3,410
35	2,030	2,724	3,591	85	1,988	2,635	3,409
36	2,028	2,719	3,582	86	1,988	2,634	3,407
37	2,026	2,715	3,574	87	1,988	2,634	3,406
38	2,024	2,712	3,566	88	1,987	2,633	3,405
39	2,023	2,708	3,558	89	1,987	2,632	3,403
40	2,021	2,704	3,551	90	1,987	2,632	3,402
41	2,020	2,701	3,544	91	1,986	2,631	3,401
42	2,018	2,698	3,538	92	1,986	2,630	3,399
43	2,017	2,695	3,532	93	1,986	2,630	3,398
44	2,015	2,692	3,526	94	1,986	2,629	3,397
45	2,014	2,690	3,520	95	1,985	2,629	3,396
46	2,013	2,687	3,515	96	1,985	2,628	3,395
47	2,012	2,685	3,510	97	1,985	2,627	3,394
48	2,011	2,682	3,505	98	1,984	2,627	3,393
49	2,010	2,680	3,500	99	1,984	2,626	3,392
50	2,009	2,678	3,496	∞	1,960	2,576	3,291

Tab. II(b): Kritische Werte der t-Verteilung (einseitig)

Irrtumswahrscheinlichkeit α				Irrtumswahrscheinlichkeit α			
FG \ α	0,05	0,01	0,001	FG \ α	0,05	0,01	0,001
1	6.314	31.821	318.309	51	1.675	2.402	3.258
2	2.920	6.965	22.327	52	1.675	2.400	3.255
3	2.353	4.541	10.215	53	1.674	2.399	3.251
4	2.132	3.747	7.173	54	1.674	2.397	3.248
5	2.015	3.365	5.893	55	1.673	2.396	3.245
6	1.943	3.143	5.208	56	1.673	2.395	3.242
7	1.895	2.998	4.785	57	1.672	2.394	3.239
8	1.860	2.896	4.501	58	1.672	2.392	3.237
9	1.833	2.821	4.297	59	1.671	2.391	3.234
10	1.812	2.764	4.144	60	1.671	2.390	3.232
11	1.796	2.718	4.025	61	1.670	2.389	3.229
12	1.782	2.681	3.930	62	1.670	2.388	3.227
13	1.771	2.650	3.852	63	1.669	2.387	3.225
14	1.761	2.624	3.787	64	1.669	2.386	3.223
15	1.753	2.602	3.733	65	1.669	2.385	3.220
16	1.746	2.583	3.686	66	1.668	2.384	3.218
17	1.740	2.567	3.646	67	1.668	2.383	3.216
18	1.734	2.552	3.610	68	1.668	2.382	3.214
19	1.729	2.539	3.579	69	1.667	2.382	3.213
20	1.725	2.528	3.552	70	1.667	2.381	3.211
21	1.721	2.518	3.527	71	1.667	2.380	3.209
22	1.717	2.508	3.505	72	1.666	2.379	3.207
23	1.714	2.500	3.485	73	1.666	2.379	3.206
24	1.711	2.492	3.467	74	1.666	2.378	3.204
25	1.708	2.485	3.450	75	1.665	2.377	3.202
26	1.706	2.479	3.435	76	1.665	2.376	3.201
27	1.703	2.473	3.421	77	1.665	2.376	3.199
28	1.701	2.467	3.408	78	1.665	2.375	3.198
29	1.699	2.462	3.396	79	1.664	2.374	3.197
30	1.697	2.457	3.385	80	1.664	2.374	3.195
31	1.696	2.453	3.375	81	1.664	2.373	3.194
32	1.694	2.449	3.365	82	1.664	2.373	3.193
33	1.692	2.445	3.356	83	1.663	2.372	3.191
34	1.691	2.441	3.348	84	1.663	2.372	3.190
35	1.690	2.438	3.340	85	1.663	2.371	3.189
36	1.688	2.434	3.333	86	1.663	2.370	3.188
37	1.687	2.431	3.326	87	1.663	2.370	3.187
38	1.686	2.429	3.319	88	1.662	2.369	3.185
39	1.685	2.426	3.313	89	1.662	2.369	3.184
40	1.684	2.423	3.307	90	1.662	2.368	3.183
41	1.683	2.421	3.301	91	1.662	2.368	3.182
42	1.682	2.418	3.296	92	1.662	2.368	3.181
43	1.681	2.416	3.291	93	1.661	2.367	3.180
44	1.680	2.414	3.286	94	1.661	2.367	3.179
45	1.679	2.412	3.281	95	1.661	2.366	3.178
46	1.679	2.410	3.277	96	1.661	2.366	3.177
47	1.678	2.408	3.273	97	1.661	2.365	3.176
48	1.677	2.407	3.269	98	1.661	2.365	3.175
49	1.677	2.405	3.265	99	1.660	2.365	3.175
50	1.676	2.403	3.261	∞	1.645	2.327	3.091

Tab. III: Quantile der Standardnormalverteilung. Werte für z_γ , so dass $P(Z < z_\gamma) = \gamma$, wobei Z einer Standardnormalverteilung folgt.

γ	z_γ
0,600	0,25335
0,700	0,52440
0,800	0,84162
0,900	1,28155
0,950	1,64485
0,975	1,95996
0,990	2,32635
0,995	2,57583



Ablese-Beispiel:

Suche $z_{1-\alpha/2}$ für $\alpha = 5\% = 0,05$

$$\Rightarrow \gamma = 1 - \alpha/2 = 1 - 0,05/2 = 0,975$$

$$\Rightarrow z_{1-\alpha/2} = z_\gamma = z_{0,975} = 1,95996 \approx 1,96$$

Tab. IV: Kritische Werte $\chi^2_{\gamma;v}$ der χ^2 -Verteilung mit v Freiheitsgraden, so dass $P(\chi^2 < \chi^2_{\gamma;v}) = \gamma$.

$\gamma \backslash v$	0,0005	0,005	0,025	0,975	0,995	0,9995
1	0,000	0,000	0,001	5,024	7,879	12,116
2	0,001	0,010	0,051	7,378	10,597	15,202
3	0,015	0,072	0,216	9,348	12,838	17,730
4	0,064	0,207	0,484	11,143	14,860	19,997
5	0,158	0,412	0,831	12,833	16,750	22,105
6	0,299	0,676	1,237	14,449	18,548	24,103
7	0,485	0,989	1,690	16,013	20,278	26,018
8	0,710	1,344	2,180	17,535	21,955	27,868
9	0,972	1,735	2,700	19,023	23,589	29,666
10	1,265	2,156	3,247	20,483	25,188	31,420
11	1,587	2,603	3,816	21,920	26,757	33,137
12	1,934	3,074	4,404	23,337	28,300	34,821
13	2,305	3,565	5,009	24,736	29,819	36,478
14	2,697	4,075	5,629	26,119	31,319	38,109
15	3,108	4,601	6,262	27,488	32,801	39,719
16	3,536	5,142	6,908	28,845	34,267	41,308
17	3,980	5,697	7,564	30,191	35,718	42,879
18	4,439	6,265	8,231	31,526	37,156	44,434
19	4,912	6,844	8,907	32,852	38,582	45,973
20	5,398	7,434	9,591	34,170	39,997	47,498
21	5,896	8,034	10,283	35,479	41,401	49,011
22	6,404	8,643	10,982	36,781	42,796	50,511
23	6,924	9,260	11,689	38,076	44,181	52,000
24	7,453	9,886	12,401	39,364	45,559	53,479
25	7,991	10,520	13,120	40,646	46,928	54,947
26	8,538	11,160	13,844	41,923	48,290	56,407
27	9,093	11,808	14,573	43,195	49,645	57,858
28	9,656	12,461	15,308	44,461	50,993	59,300
29	10,227	13,121	16,047	45,722	52,336	60,735
30	10,804	13,787	16,791	46,979	53,672	62,162
31	11,389	14,458	17,539	48,232	55,003	63,582
32	11,979	15,134	18,291	49,480	56,328	64,995
33	12,576	15,815	19,047	50,725	57,648	66,403
34	13,179	16,501	19,806	51,966	58,964	67,803
35	13,787	17,192	20,569	53,203	60,275	69,199
36	14,401	17,887	21,336	54,437	61,581	70,588
37	15,020	18,586	22,106	55,668	62,883	71,972
38	15,644	19,289	22,878	56,896	64,181	73,351
39	16,273	19,996	23,654	58,120	65,476	74,725
40	16,906	20,707	24,433	59,342	66,766	76,095
41	17,544	21,421	25,215	60,561	68,053	77,459
42	18,186	22,138	25,999	61,777	69,336	78,820
43	18,832	22,859	26,785	62,990	70,616	80,176
44	19,483	23,584	27,575	64,201	71,893	81,528
45	20,137	24,311	28,366	65,410	73,166	82,876
46	20,794	25,041	29,160	66,617	74,437	84,220
47	21,456	25,775	29,956	67,821	75,704	85,560
48	22,121	26,511	30,755	69,023	76,969	86,897
49	22,789	27,249	31,555	70,222	78,231	88,231
50	23,461	27,991	32,357	71,420	79,490	89,561

Tab. V: Kritische Werte $F_{(0,975; v_1, v_2)}$ der F -Verteilung mit v_1 und v_2 Freiheitsgraden, so dass $P(F < F_{(0,975; v_1, v_2)}) = 0,975$.

v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	50	∞
v_2																
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	998.1	1008	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.48	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.01	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.50	8.38	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.27	6.14	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.11	4.98	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.40	4.28	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.94	3.81	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.60	3.47	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.35	3.22	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.16	3.03	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.01	2.87	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.88	2.74	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.78	2.64	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.69	2.55	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.61	2.47	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.55	2.41	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.49	2.35	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.44	2.30	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.40	2.25	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.36	2.21	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.32	2.17	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.29	2.14	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.26	2.11	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.23	2.08	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.21	2.05	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.18	2.03	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.16	2.01	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.14	1.99	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.12	1.97	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	1.99	1.83	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.87	1.70	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.75	1.56	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.63	1.43	1.00

Tab. VI: Kritische Werte $F_{(0,95; v_1, v_2)}$ der F -Verteilung mit v_1 und v_2 Freiheitsgraden, so dass $P(F < F_{(0,95; v_1, v_2)}) = 0,95$.

v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	50	∞
v_2																
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.3	251.8	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.48	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.58	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.70	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.44	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.75	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.32	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.02	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.80	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.64	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.51	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.40	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.31	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.24	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.18	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.12	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.08	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.04	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.00	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	1.97	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	1.94	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.91	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.88	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.86	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.84	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.94	1.82	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.92	1.81	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.79	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.89	1.77	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.76	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.66	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.56	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.60	1.46	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.51	1.35	1.00

Tab. VII: Studentisierte Variationsbreiten $q_{Tab}(t, FG, \alpha = 5\%)$.

FG/t	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07	50.59	51.96	53.19	54.32	55.36	56.32	57.21	58.04	58.82	59.55
2	6.08	8.33	9.80	10.88	11.73	12.43	13.03	13.54	13.99	14.39	14.75	15.08	15.37	15.65	15.91	16.14	16.36	16.57	16.77
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	19.72	9.95	10.15	10.35	10.52	10.69	10.84	10.98	11.11	11.24
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.66	8.79	8.91	9.03	9.13	9.23
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.10	7.17
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11	6.19	6.27	6.34	6.40	6.47
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.13	6.20	6.27	6.33
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	5.99	6.05	6.11
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.71	5.79	5.85	5.91	5.97	6.03
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.78	5.85	5.90	5.96
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.54	5.61	5.67	5.73	5.79	5.84
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.31	5.39	5.46	5.53	5.59	5.65	5.70	5.75
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71
21	2.94	3.56	3.94	4.21	4.42	4.60	4.74	4.87	4.98	5.08	5.17	5.25	5.33	5.40	5.46	5.52	5.58	5.63	5.68
22	2.93	3.55	3.93	4.20	4.41	4.58	4.72	4.85	4.96	5.06	5.14	5.23	5.30	5.37	5.43	5.49	5.55	5.60	5.65
23	2.93	3.54	3.91	4.18	4.39	4.56	4.70	4.83	4.94	5.03	5.12	5.20	5.27	5.34	5.41	5.46	5.52	5.57	5.62
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.49	5.55	5.59
25	2.91	3.52	3.89	4.15	4.36	4.53	4.67	4.79	4.90	4.99	5.08	5.16	5.23	5.30	5.36	5.42	5.47	5.52	5.57
26	2.91	3.51	3.88	4.14	4.35	4.51	4.65	4.77	4.88	4.98	5.06	5.14	5.21	5.28	5.34	5.40	5.45	5.50	5.55
27	2.90	3.51	3.87	4.13	4.33	4.50	4.64	4.76	4.86	4.96	5.04	5.12	5.19	5.26	5.32	5.38	5.43	5.48	5.53
28	2.90	3.50	3.86	4.12	4.32	4.49	4.62	4.74	4.85	4.94	5.03	5.11	5.18	5.24	5.30	5.36	5.41	5.46	5.51
29	2.89	3.49	3.85	4.11	4.31	4.47	4.61	4.73	4.84	4.93	5.01	5.09	5.16	5.23	5.29	5.34	5.40	5.44	5.49
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.47
31	2.88	3.48	3.84	4.09	4.29	4.45	4.59	4.71	4.81	4.90	4.99	5.06	5.13	5.20	5.26	5.31	5.36	5.41	5.46
32	2.88	3.48	3.83	4.09	4.28	4.45	4.58	4.70	4.80	4.89	4.98	5.05	5.12	5.18	5.24	5.30	5.35	5.40	5.45
33	2.88	3.47	3.83	4.08	4.28	4.44	4.57	4.69	4.79	4.88	4.97	5.04	5.11	5.17	5.23	5.29	5.34	5.39	5.43
34	2.87	3.47	3.82	4.07	4.27	4.43	4.56	4.68	4.78	4.87	4.96	5.03	5.10	5.16	5.22	5.27	5.33	5.37	5.42
35	2.87	3.46	3.81	4.07	4.26	4.42	4.56	4.67	4.77	4.86	4.95	5.02	5.09	5.15	5.21	5.26	5.31	5.36	5.41
36	2.87	3.46	3.81	4.06	4.25	4.41	4.55	4.66	4.76	4.85	4.94	5.01	5.08	5.14	5.20	5.25	5.30	5.35	5.40
37	2.87	3.45	3.80	4.05	4.25	4.41	4.54	4.66	4.76	4.85	4.93	5.00	5.07	5.13	5.19	5.24	5.29	5.34	5.39
38	2.86	3.45	3.80	4.05	4.24	4.40	4.53	4.65	4.75	4.84	4.92	4.99	5.06	5.12	5.18	5.23	5.28	5.33	5.38
39	2.86	3.45	3.79	4.04	4.24	4.39	4.53	4.64	4.74	4.83	4.91	4.98	5.05	5.11	5.17	5.22	5.27	5.32	5.37
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90	4.98	5.04	5.11	5.16	5.22	5.27	5.31	5.36
50	2.84	3.42	3.76	4.00	4.19	4.34	4.47	4.58	4.68	4.77	4.85	4.92	4.98	5.04	5.10	5.15	5.20	5.24	5.29
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01

Tab. VIII: Kritische Werte $\chi^2_{\gamma;v}$ der χ^2 -Verteilung mit v Freiheitsgraden, so dass $P(\chi^2 < \chi^2_{\gamma;v}) = \gamma$.

Siehe auch
Tab. IV, wo
andere Werte
für γ tabelliert
sind.

$\gamma \backslash v$	0,90	0,95	0,99
1	2,706	3,841	6,635
2	4,605	5,991	9,210
3	6,251	7,815	11,345
4	7,779	9,488	13,277
5	9,236	11,070	15,086
6	10,645	12,592	16,812
7	12,017	14,067	18,475
8	13,362	15,507	20,090
9	14,684	16,919	21,666
10	15,987	18,307	23,209
11	17,275	19,675	24,725
12	18,549	21,026	26,217
13	19,812	22,362	27,688
14	21,064	23,685	29,141
15	22,307	24,996	30,578
16	23,542	26,296	32,000
17	24,769	27,587	33,409
18	25,989	28,869	34,805
19	27,204	30,144	36,191
20	28,412	31,410	37,566
21	29,615	32,671	38,932
22	30,813	33,924	40,289
23	32,007	35,172	41,638
24	33,196	36,415	42,980
25	34,382	37,652	44,314
26	35,563	38,885	45,642
27	36,741	40,113	46,963
28	37,916	41,337	48,278
29	39,087	42,557	49,588
30	40,256	43,773	50,892
31	41,422	44,985	52,191
32	42,585	46,194	53,486
33	43,745	47,400	54,776
34	44,903	48,602	56,061
35	46,059	49,802	57,342
36	47,212	50,998	58,619
37	48,363	52,192	59,893
38	49,513	53,384	61,162
39	50,660	54,572	62,428
40	51,805	55,758	63,691
41	52,949	56,942	64,950
42	54,090	58,124	66,206
43	55,230	59,304	67,459
44	56,369	60,481	68,710
45	57,505	61,656	69,957
46	58,641	62,830	71,201
47	59,774	64,001	72,443
48	60,907	65,171	73,683
49	62,038	66,339	74,919
50	63,167	67,505	76,154

Anhang B: Was ist eigentlich SAS?

SAS = Statistical Analysis System ist eines von mehreren großen Statistikpaketen (R, SPSS, BMDP, GENSTAT, ASREML, etc.), die auf PCs laufen und verschiedene statistische Verfahren anbieten. Alle in dieser Vorlesung vorgestellten Methoden können mit SAS implementiert werden. Das Programm verfügt einen Programm-Editor, in welchem Befehle und Anweisungen in Textform eingegeben werden müssen. An manchem Stellen im Skript sind solche Anweisungen exemplarisch aufgeführt. Die Anweisungen sind natürlich nicht prüfungsrelevant, sondern sie dienen nur zur Veranschaulichung, wie das eine oder andere Verfahren am PC umzusetzen ist. SAS ist in allen CIP-Pools des Rechenzentrums der Universität Hohenheim verfügbar. Es gibt verschiedene SAS-Kurse, die vom Rechenzentrum sowie von Herrn Dr. Schumacher (Institut 110) angeboten werden. Näheres findet man z.B. in Dufner, J., Jensen, U., Schumacher, E. 2002. Statistik mit SAS. Teubner, 2. Aufl. Stuttgart, und unter www.sas.com.

Anhang C: Die Methode von Lagrange

Die Lagrange-Methode ist ein Verfahren zur Berechnung von Stellen einer Funktion, an denen Extrema (Minima oder Maxima) auftreten können und die einer Nebenbedingung folgen. Ob eine mit der Lagrange-Methode berechnete Stelle tatsächlich ein Extremum darstellt, erfordert im allgemeinen eine komplizierte Berechnung. Die Lagrange-Methode selbst ist aber sehr einfach anzuwenden, wie im folgenden an einem Beispiel gezeigt wird. Für den Fall einer Funktion f zweier Veränderlicher x und y lautet die Lagrange-Regel wie folgt:

Zur Bestimmung der Extremwerte einer Funktion $z = f(x, y)$ mit der Nebenbedingung $\varphi(x, y) = 0$ bildet man mit dem unbestimmten Multiplikator λ eine Hilfsfunktion $F(x, y) = f(x, y) + \lambda \varphi(x, y)$ und die ersten partiellen Ableitungen dieser Hilfsfunktion nach x , y und λ . Die ersten Ableitungen werden gleich Null gesetzt. Aus dem resultierenden Gleichungssystem werden die Koordinaten x und y des möglichen Extremwertes und der Multiplikator λ berechnet.

Beispiel: Unter allen rechtwinkligen Dreiecken mit gegebener Hypotenuse c soll dasjenige mit dem größten Flächeninhalt gefunden werden.

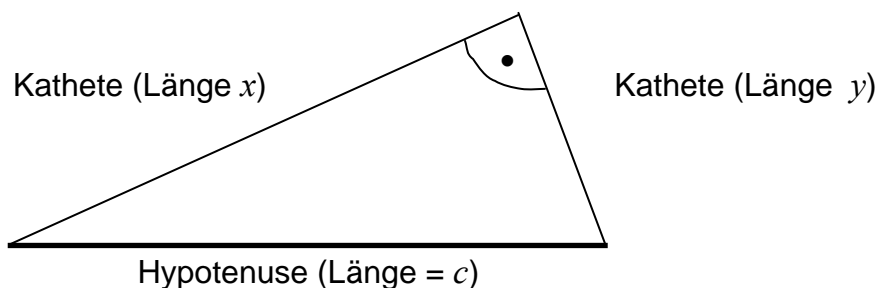


Abb. C1. Ein rechtwinkliges Dreieck.

Bezeichnet man die Katheten des Dreiecks mit x und y , so soll also der Flächeninhalt

$$f(x, y) = xy/2$$

ein Maximum werden. Da es sich um ein rechtwinkliges Dreieck handelt, lautet die Nebenbedingung

$$x^2 + y^2 = c^2$$

oder

$$\varphi(x, y) = x^2 + y^2 - c^2 = 0$$

Damit bildet man die Hilfsfunktion

$$F(x, y) = xy/2 + \lambda(x^2 + y^2 - c^2).$$

Die Ableitungen nach x und y ergeben

$$\frac{\partial F}{\partial x} = \frac{y}{2} + 2\lambda x = 0 \Rightarrow \lambda = -\frac{y}{4x} \quad \text{und hiermit}$$

$$\frac{\partial F}{\partial y} = \frac{x}{2} + 2\lambda y = \frac{x}{2} - \frac{y^2}{2x} = 0 \Rightarrow x^2 = y^2 \quad \text{sowie}$$

$$\frac{\partial F}{\partial \lambda} = \varphi(x, y) = x^2 + y^2 - c^2 = 0 \Rightarrow x^2 = \frac{c^2}{2}$$

Ein rechtwinkliges Dreieck mit Kathetenlängen $x = y = c/\sqrt{2}$ ist ein gleichschenkliges Dreieck. Dieses hat unter allen rechtwinkligen den größten Flächeninhalt.

Die Multiplikatorenregel kann für die Bestimmung der Extremwerte einer Funktion von mehr als zwei Veränderlichen mit Nebenbedingungen entsprechend erweitert werden (Mathematik: Kleine Enzyklopädie. Verlag Harri Deutsch, Frankfurt/M und Zürich, 1972).

Anhang D: Was besagt der p -Wert?

Wenn man statistische Auswertungen mit dem Computer durchführt, werden Testergebnisse generell in Form von p -Werten präsentiert. Im Skript zur Vorlesung „Statistik“ wird der p -Wert am Beispiel des t -Tests erläutert (Kap. 2). Hier rekapitulieren wir den Begriff kurz anhand der einfaktoriellen Varianzanalyse zum Vergleich von t Mittelwerten, der in Kap. 3 im Skript „Statistik“ bereits behandelt wurde. Es sei aber betont, dass hier vorgestellte Interpretation des p -Wertes für jeden Test gilt: Kleine p -Werte führen zur Ablehnung der Nullhypothese.

Beispiel: Im folgenden sind die Erträge (dt/ha) von 5 Sorten (A, B, C, D, E) eines Feldversuches in einem Lageplan eingezeichnet. Der Versuch war in 4 Wiederholungen angelegt. Die Behandlungen wurden zufällig den Parzellen (Feldstücken) zugeordnet (randomisiert).

B 21	D 34	D 32	E 24
C 27	E 23	A 31	B 23
A 32	C 29	E 27	A 37
D 31	D 27	A 32	B 25
B 19	C 34	E 26	C 34

Für eine Varianzanalyse können wir das folgende Modell ansetzen:

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, \dots, t; \quad j = 1, \dots, r$$

wobei

y_{ij} = Messwert der j -ten Wiederholung der i -ten Behandlung

μ_i = theoretischer Mittelwert (Erwartungswert) der i -ten Behandlung

e_{ij} = Zufallsabweichung der Beobachtung y_{ij} vom Behandlungsmittel

Eine alternative Schreibweise ergibt sich, wenn wir den Behandlungsmittelwert μ_i ersetzen durch $\mu + \tau_i$, so dass das Modell lautet:

$$y_{ij} = \mu + \tau_i + e_{ij}, \quad i = 1, \dots, t; \quad j = 1, \dots, r,$$

wobei

μ = Konstante

τ_i = Effekt der i -ten Behandlung

In dieser Schreibweise ist $\mu + \tau_i$ der wahre Mittelwert der i -ten Behandlung.

Die zu prüfende Nullhypothese lautet:

$$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 \quad (\text{alle Behandlungen sind gleich})$$

Die Varianzanalyse zum Test dieser Nullhypothese wird wie folgt durchgeführt:

Ursache	FG	SQ	MQ	$E(MQ)$	F_{Vers}
Behandlungen	$t - 1$	SQ_{Beh}	$MQ_{Beh} = \frac{SQ_{Beh}}{t - 1}$	$\sigma^2 + Q(\tau)$	$F_{Vers} = \frac{MQ_{Beh}}{MQ_{Fehler}}$
Fehler	$t(r - 1)$	SQ_{Fehler}	$MQ_{Fehler} = \frac{SQ_{Fehler}}{t(r - 1)}$	σ^2	

Quadratsummen:

$$SQ_{Beh} = r \sum_{i=1}^t (\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^t y_{i.}^2 / r - y_{..}^2 / (rt);$$

$$SQ_{Fehler} = \sum_{i=1}^t \sum_{j=1}^r (y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^t \sum_{j=1}^r y_{ij}^2 - \sum_{i=1}^t y_{i.}^2 / r$$

wobei

Beobachtung: y_{ij} = Messwert j -te Wiederholung der i -ten Behandlung

$$(i = 1, \dots, t; j = 1, \dots, r)$$

Behandlungssummen: $y_{i.} = \sum_{j=1}^r y_{ij}$

Gesamtsumme: $y_{..} = \sum_{i=1}^t \sum_{j=1}^r y_{ij}$

Verwerfe die Nullhypothese $H_0: \tau_1 = \tau_2 = \dots = \tau_t$ wenn gilt:

$$F_{Vers} = MQ_{Beh} / MQ_{Fehler} > F_{Tab} = F_{(1-\alpha; t-1, t(r-1))} \quad (\text{Tab. VI})$$

Zum Verständnis der Varianzanalyse ist es hilfreich, die Erwartungswerte der beiden MQ zu betrachten $[E(MQ)]$. Der Erwartungswert ergibt sich, wenn man sich vorstellt, dass derselbe Versuch sehr oft wiederholt wird. Die MQ werden aufgrund von Zufallseinflüssen (Randomisation, Versuchsfehler) von Versuch zu Versuch schwanken. Über viele Wiederholungen des Versuches würde man einen langfristigen Mittelwert, den Erwartungswert, erhalten.

Im Erwartungswert des MQ_{Beh} erscheint eine quadratische Form der Behandlungseffekte:

$$Q(\tau) = \frac{r \sum_{i=1}^t (\tau_i - \bar{\tau})^2}{t - 1}$$

Je größer die Behandlungsunterschiede sind, desto größer wird $Q(\tau)$. Unter der Nullhypothese gilt $Q(\tau) = 0$. In diesem Fall haben beide MQ denselben Erwartungswert. Für

den F-Wert (F_{vers}), also den Quotienten der beiden MQ , können wir daher einen Wert nahe Eins erwarten, wenn die Nullhypothese zutrifft. Falls dagegen die Nullhypothese verletzt ist, erwarten wir einen F-Wert, der deutlich größer als Eins ist. Offensichtlich sind große F-Werte (F_{vers}) ein Hinweis auf Gültigkeit der Alternativ-Hypothese (Behandlungsunterschiede). **Daher verwirft man H_0 für große F-Werte.**

Es ist nun ein kritischer F-Wert (F_{tab}) zu bestimmen, dessen Überschreiten zur Verwerfung der H_0 führt. Die Entscheidungsregel des Tests lautet dann:

Verwerfe H_0 , wenn $F_{vers} > F_{tab}$; andernfalls behalte H_0 bei.

Die Konstruktion des F-Tests beruht nun auf der Betrachtung der Verteilung von F_{vers} . Unter der Nullhypothese hat diese Teststatistik eine F-Verteilung mit $(t-1)$ und $t(r-1)$ Freiheitsgraden (Abb. D.1). Werte von F nahe Eins haben die größte Wahrscheinlichkeitsdichte. Selbst bei Gültigkeit der H_0 kann es jedoch mit einer gewissen Wahrscheinlichkeit passieren, dass sehr große F-Werte auftreten. Daher kann es vorkommen, dass trotz Gültigkeit der Nullhypothese $F_{vers} > F_{tab}$ wird und man folglich fälschlicherweise die H_0 verwirft.

Anhand der Dichtefunktion der F-Verteilung kann man nun Überschreitungswahrscheinlichkeiten für gegebene F-Werte berechnen. Hierzu berechnet man die Fläche unter der Dichtefunktion rechts vom betrachteten F-Wert. Der kritische Wert F_{tab} wird so gelegt, dass die Überschreitungswahrscheinlichkeit einen vorgegebenen Wert α annimmt, üblicherweise $\alpha = 5\%$. In Abb. D.1 ist α als die Summe der grau und schwarz schattierten Flächen rechts von F_{vers} dargestellt. Wenn man nun die Testvorschrift

„verwerfe H_0 , wenn $F_{vers} > F_{tab}$ ist“

befolgt, so beträgt die Wahrscheinlichkeit, fälschlicherweise H_0 zu verwerfen, genau α (Fehler 1. Art, α -Fehler). Somit kontrolliert der Test die Irrtumswahrscheinlichkeit bei α .

Wird nun ein Test mit dem Computer berechnet, so wird der Test mit Hilfe eines p -Wertes durchgeführt. Der p -Wert wird anhand von F_{vers} bestimmt.

Der p -Wert entspricht der (Überschreitungs-) Wahrscheinlichkeit, dass bei Gültigkeit der Nullhypothese ein F-Wert beobachtet wird, der größer ist, als der für die vorliegenden Daten berechnete Wert (F_{vers}).

Diese Wahrscheinlichkeit ist in Abb. D.1 als schwarz schattierte Fläche dargestellt. Wenn nun der p -Wert sehr klein ist, so ist der beobachtete F-Wert unter der Gültigkeit der Nullhypothese offensichtlich nicht sehr plausibel. **Somit wird H_0 für kleine p -Werte verworfen.** Auch für den p -Wert muss eine kritische Schwelle festgelegt werden, unterhalb derer H_0 verworfen wird. Betrachtung von Abb. D.1 zeigt, dass diese Schwelle genau bei α liegen muss. Denn er wird $F_{vers} > F_{tab}$ genau dann wenn $p < \alpha$ ist, und umgekehrt. Die Entscheidungsregel für die Interpretation des p -Wertes lautet:

Wenn $p < \alpha$, verwerfe H_0 ; andernfalls behalte H_0 bei.

Diese Entscheidungsregel gilt übrigens für jeden beliebigen Test und ist nicht auf den F-Test beschränkt.

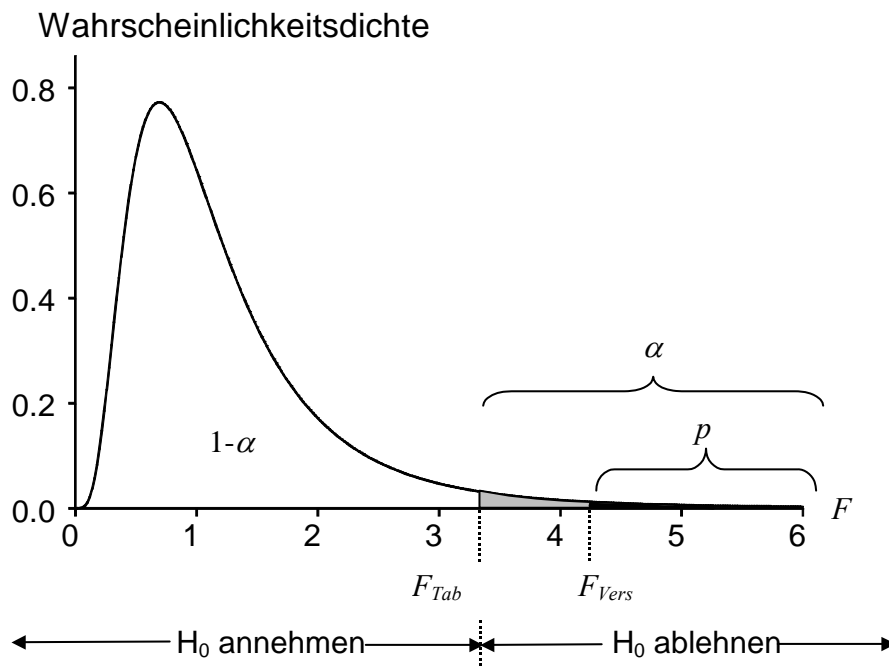


Abb. A. 1: Dichte einer F-Verteilung. p -Wert = schwarz schattierte Fläche. α -Fehler = graue und schwarz schattierte Fläche zusammen.

Beispiel: Für den Sortenversuch liefert eine Auswertung mit der SAS Prozedur GLM folgendes Ergebnis:

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model (Sorte)	4	348.8000000	87.2000000	11.28	0.0002
Error	15	116.0000000	7.7333333		
Corrected Total	19	464.8000000			

Der p -Wert beträgt $p = 0,0002 < \alpha = 0,05 = 5\%$. Somit bestehen signifikante Unterschiede zwischen den Sorten.

Anhang E: Freiheitsgrade

Der Begriff der Freiheitsgrade bereitet immer wieder Schwierigkeiten. Es gibt eine Reihe einfacher heuristischer Erklärungen, von denen einige im Skript Statistik in Abschnitt 5.8 beschrieben sind. Im Zusammenhang mit linearen Modellen ist die mathematisch klarste Definition diejenige, die in Abschnitt 6.8.2 gegeben wurde: $FG_{Fehler} = n - \text{Rang}(X)$, wobei n die Zahl der Beobachtungen und X die Designmatrix des linearen Modells ist.

Eine einfache heuristische Erklärung sei hier wiederholt. Wenn wir eine einfache Stichprobe von n Werten y_i ($i = 1, \dots, n$) betrachten und den Mittelwert \bar{y}_\cdot berechnen, so haben wir einen Parameter geschätzt. X hat eine Spalte von Einsen und somit $\text{Rang}(X) = 1$, so dass $FG_{Fehler} = n - 1$. Wir können auch sagen, dass, wenn wir den Mittelwert \bar{y}_\cdot beim beobachteten Wert fixieren, wir $n - 1$ der Werte y_i frei wählen können, dass wir also $n - 1$ Freiheitsgrade haben, während der n -te Wert dann automatisch festliegt.

Ähnlich ist es in der einfaktoriellen Varianzanalyse (Statistikskript, Kap. 4). Bei t Mittelwerten und r Wiederholungen können wir für jede Behandlung $r - 1$ der Werte frei wählen, während der r -te festliegt. Insgesamt haben wir also $t(r - 1)$ Freiheitsgrade für den Fehler.

Bei anderen Modellen ist es etwas komplizierter. Generell müssen wir oft eine Überparametrisierung des Modells beachten. Das gilt z.B. für die vollständige Blockanlage mit t Behandlungen und r Wiederholungen. In solchen Fällen ist es nicht mehr möglich (jedenfalls mir nicht), einfache heuristische Erklärungen zu geben. Die exakte Definition $FG_{Fehler} = n - \text{Rang}(X)$ funktioniert jedoch immer. Bei der Blockanlage ist eine Restriktion für die Behandlungseffekte sowie eine Restriktion für die Blockeffekte zu beachten. Beispielsweise kann man den letzten Behandlungseffekt gleich null setzen, so dass $t - 1$ Effekte zu schätzen sind. Analoges gilt für die Blöcke, bei denen nur $r - 1$ Effekte zu schätzen sind. Hinzu kommt der allgemeine Effekt μ , so dass insgesamt $1 + t - 1 + r - 1 = r + t - 1$ Effekte zu schätzen sind. Die Design-Matrix hat somit $r + t - 1$ unabhängige Spalten und $FG_{Fehler} = n - \text{Rang}(X) = n - r - t + 1$.

Die Freiheitsgrade für einen Effekt im Modell ergeben sich immer einfach als Differenz der Fehler-Freiheitsgrade für die Modelle jeweils mit und ohne diesen Effekt. Bei der einfaktoriellen Varianzanalyse haben wir z.B. im Modell ohne Behandlungseffekt nur den Gesamteffekt μ und somit $rt - 1$ Fehler-Freiheitsgrade. Im Modell mit Behandlungseffekten haben wir $t(r - 1)$ Freiheitsgrade für den Fehler. Die Differenz der Fehler-Freiheitsgrade beträgt $t - 1$, und dies sind die Behandlungsfreiheitsgrade. Dies entspricht genau der Zahl der Spalten in der Designmatrix für die Behandlungseffekte, wenn wir die Überparametrisierung beachten und den letzten Behandlungseffekt gleich Null setzen und somit aus der Designmatrix nehmen.

Anhang F: Fehlerfortpflanzung (Delta-Methode)

Gegeben sei eine Zufallsvariable Z , die als Funktion der Zufallsvariablen Y_1, Y_2, \dots, Y_n zu berechnen ist:

$$Z = f(Y_1, Y_2, \dots, Y_n).$$

Für die Varianz der Zufallsvariablen Z gilt näherungsweise das Fehlerfortpflanzungsgesetz:

$$\text{var}(Z) \approx \sum_{i=1}^n \left(\frac{\partial Z}{\partial Y_i} \right)_{Y_i=E(Y_i)}^2 \text{var}(Y_i) + \sum_{i=1}^n \sum_{j \neq i}^n \left(\frac{\partial Z}{\partial Y_i} \right)_{Y_i=E(Y_i)} \left(\frac{\partial Z}{\partial Y_j} \right)_{Y_j=E(Y_j)} \text{cov}(Y_i, Y_j).$$

Falls Z eine lineare Funktion der Zufallsvariablen Y_1, Y_2, \dots, Y_n ist, so ist dies Ergebnis exakt, wie man leicht nachprüft. Eine solche lineare Funktion sei gegeben durch $z = L'y$, wobei L eine Matrix mit festen Konstanten ist, $z = Z$ und $y' = (Y_1, Y_2, \dots, Y_n)$. Dann gilt

$$\text{var}(z) = L' \text{var}(y) L.$$

Die Gleichung gilt darüber hinaus auch dann, wenn z ein Vektor ist.

Einfaches Beispiel: Sie haben Ihre Daten nach $Z = \sqrt[3]{Y}$ transformiert und für die transformierten Daten einen Mittelwert berechnet. Unter Annahme der Normalverteilung auf der transformierten Skala ist dieser Mittelwert eine Schätzung des Medians.

Rücktransformation des Mittelwertes ist daher eine Medianschätzung (siehe Abschnitt 7.1). Sie haben den Standardfehler $s.f.(\bar{Z})$ des Mittelwertes \bar{Z} auf der transformierten Skala und wollen den Standardfehler $s.f.(M_Y)$ für den rücktransformierten Mittelwert (Medianschätzer) $M_Y = \bar{Z}^3$. Anwendung der Delta-Methode liefert:

$$\text{var}(M_Y) \approx \left(\frac{\partial M_Y}{\partial \bar{Z}} \right)^2 \text{var}(\bar{Z}) \approx (3\bar{Z}^2)^2 \text{var}(\bar{Z}), \text{ und somit}$$

$$s.f.(M_Y) \approx 3\bar{Z}^2 s.f.(\bar{Z}).$$

Rechenbeispiel: $\bar{Z} = 2$, $s.f.(\bar{Z}) = 0.2 \Rightarrow M_Y = 2^3 = 8$, $s.f.(M_Y) \approx 3 \cdot 2^2 \cdot 0.2 = 2.4$