Berichte

aus der Biologischen Bundesanstalt für Land- und Forstwirtschaft

Reports

from the Federal Biological Research Centre for Agriculture and Forestry

Heft 56
2000

Einführung in die Biometrie unter Berücksichtigung der Software SAS

Teil 4:

Korrelationsanalyse, Regressionsanalyse und Kovarianzanalyse

Zur Nutzung von SAS/INSIGHT® und der Analyst Application

Introduction to the Biometry in Regard to the Software SAS
Part 4: Correlation Analysis, Regression Analysis, and Covariance
Notes on SAS/INSIGHT® and the Analyst Application

Bearbeitet von
compiled by

Eckard Moll

Zentrale EDV-Gruppe, Außenstelle Kleinmachnow

Central Working Group Data Processing, Branch Office Kleinmachnow

Herausgeber

Biologische Bundesanstalt für Land- und Forstwirtschaft
Braunschweig, Deutschland

Verlag:
Eigenverlag

Vertrieb:
Saphir-Verlag, Gutsstraße 15, D-38551 Ribbesbüttel
Telefon +49/(0) 53 74-65 76
Telefax +49/(0) 53 74-65 77

ISSN-Nummer: 0947-8809

Kontaktadresse:
Dr. Eckard Moll
Biologische Bundesanstalt für Land- und Forstwirtschaft
Zentrale EDV-Gruppe, Außenstelle Kleinmachnow
Stahnsdorfer Damm 81
D-14532 Kleinmachnow
Telefon +49/(0) 3 32 03-48-331
Telefax +49/(0) 3 32 03-48-424
E-Mail E.Moll@BBA.de

Vorwort

Die bisher in loser Folge erschienenen Hefte zur „Einführung in die Biometrie unter Berücksichtigung der Software SAS" in der Reihe „Berichte aus der Biologischen Bundesanstalt für Land- und Forstwirtschaft"
Heft 23: Teil1: Grundbegriffe, beschreibende Statistik und Vergleich zweier Mittelwerte
Heft 31: Teil2: Vergleich von mehr als zwei Mittelwerten, ein- und zweifaktorielle Varianzanalyse mit festen und zufälligen Effekten
Heft 46: Teil3: Die Varianzanalyse im Feldversuchswesen
sind mit viel Interesse aufgenommen worden. Wiederholt wurde ich gebeten, diese Serie fortzusetzen.

Das vorliegende Heft beschäftigt sich vor allem mit der Korrelations-, Regressions- und Kovarianzanalyse und schließt damit die „Einführung in die Biometrie" ab.
Gestreift werden aber auch SAS/INSIGHT® und der Analyst. Das sind moderne Werkzeuge, die ohne tiefgehende SAS-Kenntnisse zum Erfolg führen, weil man sich durchclicken kann.

Ich bedanke mich bei allen für ihr Interesse und die vielfältige Unterstützung.

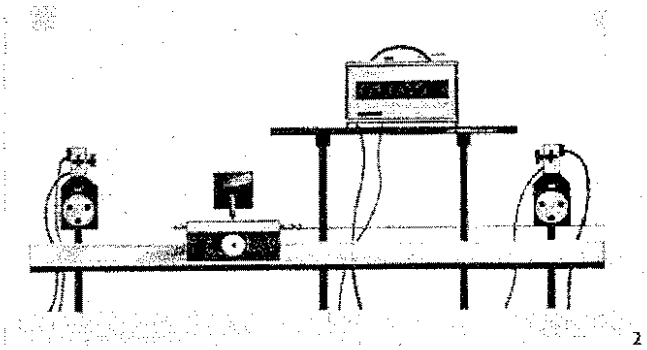
Weiterbildungskurse zu speziellen biometrischen Problemen werden im Auftrag des Senats der Bundesforschungsanstalten des Bundesministeriums für Ernährung, Landwirtschaft und Forsten seit 1979 für Mitarbeiter des Geschäftsbereiches durchgeführt. Die aktuellen Kursangebote können unserer Web-Seite „Biometrie und SAS®-Anwendung im Geschäftsbereich des BML" http://www.mol.shuttle.de/wspc/sas/biometrie.html entnommen werden.

Inhaltsverzeichnis

14 Korrelationsanalyse

14.1 Einführung in die Korrelations- und Regressionsanalyse

Der Blick in ein Schulbuch soll an Bekanntes anknüpfen. Es ist ein Physik-Buch[1], in dem das Kapitel „Gleichmäßig beschleunigte geradlinige Bewegung" behandelt wird. Ein Experiment wird betrachtet (Abb. 1), bei dem aus physikalischer Sicht auffällt, daß äußere, störende Einflüsse wie beispielsweise Reibung so gering wie möglich sein sollen (Schwebekörper auf der Luftkissenbahn).



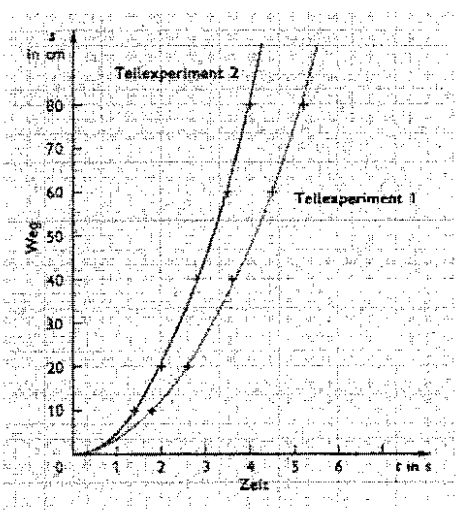
2

Experiment 2
Ein Schwebekörper auf der Luftkissenbahn wird durch eine konstante Kraft gleichmäßig beschleunigt. Gemessen werden die Zeiten, die für festgelegte Wege benötigt werden, und die Momentangeschwindigkeiten, die der Schwebekörper zu diesen Zeiten erreicht hat. Danach wird das Experiment bei größerer Beschleunigung wiederholt. Die Meßwerte werden in eine Tabelle eingetragen.

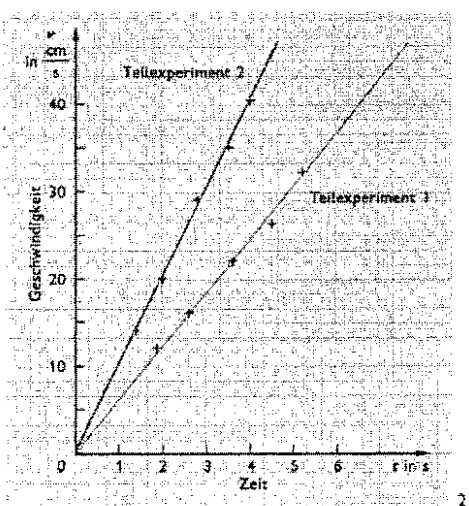
	s in cm	t in s	v in $\frac{cm}{s}$	$a = \frac{v}{t}$ in $\frac{cm}{s^2}$	$s = \frac{1}{2} a \cdot t^2$ in cm
Teil-experiment 1	10	1,85	12	6,5	11
	20	2,61	16	6,1	21
	40	3,58	22	6,1	40
	60	4,49	26	5,8	58
	80	5,22	32	6,1	83
Teil-experiment 2	10	1,43	14	9,8	10
	20	2,03	20	9,9	20
	40	2,80	29	10,4	41
	60	3,51	35	10,0	62
	80	4,04	40	9,9	81

Die tabellarisch erfaßten Meßwerte werden grafisch dargestellt.

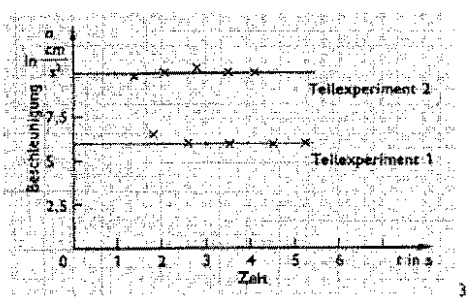
Die physikalischen Größen sind:
s: Weg
t: Zeit
v: Geschwindigkeit
a: Beschleunigung



1 Weg-Zeit-Diagramm



2 Geschwindigkeit-Zeit-Diagramm



3 Beschleunigung-Zeit-Diagramm

Abb. 1: Meßwertergebnisse eines physikalischen Experimentes

[1] Lehrbuch Physik, Sekundarstufe 1 (Klassen 9 und 10), Volk und Wissen Verlag GmbH, 1994, S. 15/16

Was läßt sich anhand der Meßwerte und besonders an ihrer Visualisierung erkennen?

1. Die Meßwerte schwanken um die als durchgezogene Kurve eingezeichneten physikalischen Gesetzmäßigkeiten.
2. Die Schwankungen um die jeweilige physikalische Gesetzmäßigkeit von Teilbild 1, 2 oder 3 der Abb. 1 sind gering, d. h. die physikalische Gesetzmäßigkeit kann durch die Meßwerte hinreichend gut beschrieben werden.
3. Zur Beschreibung der jeweiligen physikalischen Gesetzmäßigkeit kann entweder eine Geraden- oder eine quadratische Gleichung herangezogen werden.
4. Auch für einen Bereich, wo keine Messungen vorliegen, ist die Kurve als geltende physikalische Gesetzmäßigkeit weiter gezeichnet.

Was bedeuten diese Erkenntnisse hinsichtlich einer statistischen Analyse?

Davon ausgehend, daß die Meßwerte *zufälligen Schwankungen* unterliegen, kann eine statistische Analyse ins Spiel kommen. Diese hat (in der Hauptsache) zwei Zielstellung:

- die Untersuchung der Stärke eines Zusammenhanges zwischen zwei (zufälligen) Merkmalen und
- die funktionale Beschreibung der Abhängigkeit eines Merkmals von einem (oder mehreren) anderen Merkmalen.

Die statistischen Verfahren der zuerst genannten Zielstellung werden unter dem Begriff der Korrelationsanalyse zusammengefaßt. Die zweite Zielstellung findet sich in den Verfahren der Regressionsanalyse wieder.

Eine Bemerkung zur letzten der oben aufgeführten Erkenntnisse (Punkt 4) soll bereits an dieser Stelle gemacht werden:
Die Gültigkeit eines physikalischen Gesetzes kann auch über den beobachteten Bereich hinaus angenommen werden. Das ist so nicht automatisch auf biologische Merkmale übertragbar. Der Geltungsbereich für die Beschreibung eines Zusammenhanges zweier oder mehrerer biologischer Merkmale wird deshalb in der Regel auf den gemessenen Bereich beschränkt. Prognose- und Trendaussagen, die sich auch der Regressionsanalyse bedienen, weichen davon natürlich ab.

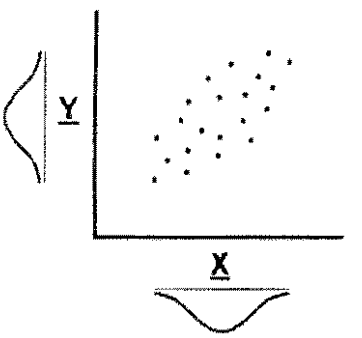
14.2 Maßzahlen der Korrelation

14.2.1 Der Produkt-Momenten-Korrelationskoeffizient

14.2.1.1 Die Schätzung des Produkt-Momenten-Korrelationskoeffizienten

Die bekannteste Maßzahl zur Beschreibung eines Zusammenhanges zwischen verschiedenen Merkmalen ist der Produkt-Momenten-Korrelationskoeffizient von Pearson bzw. nach Bravais. Es ist zwar das Zusammenhangsmaß, aber es gibt noch andere. Und wie immer sind an die Anwendung dieser Maßzahl einige Voraussetzungen geknüpft.

- Die beiden Variablen (Merkmale) X und Y sind metrisch.
- Die beiden Variablen X und Y sind Zufallsvariable.
- Die Zufallsvariablen X und Y sind zweidimensional normalverteilt (binormalverteilt).



Der (lineare) Korrelationskoeffizient ρ ist definiert durch $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$.

Dieser Parameter wird asymptotisch erwartungstreu geschätzt aus den Stichprobenpaaren ($\underline{x}_i$, $\underline{y}_i$), [i=1, ... ,n] der Zufallsvariablen $\underline{X}$ und $\underline{Y}$ durch

$$r = \frac{\sum_{i=1}^{n}(\underline{x}_i - \bar{\underline{x}})(\underline{y}_i - \bar{\underline{y}})}{\sqrt{\sum_{i=1}^{n}(\underline{x}_i - \bar{\underline{x}})^2 \sum_{i=1}^{n}(\underline{y}_i - \bar{\underline{y}})^2}} .$$

Als
Beispiel 14.1
sollen die obigen Meßwerte des Teilexperiments 1 für die Geschwindigkeit v und die Zeit t herangezogen werden. Es soll der Zusammenhang zwischen der Geschwindigkeit und der Zeit geschätzt werden.

Papier und Bleistift

Für die Handrechnung bietet sich folgende Rechenformel an:

$$r = \frac{\sum_{i=1}^{n}(\underline{x}_i - \bar{\underline{x}})(\underline{y}_i - \bar{\underline{y}})}{\sqrt{\sum_{i=1}^{n}(\underline{x}_i - \bar{\underline{x}})^2 \sum_{i=1}^{n}(\underline{y}_i - \bar{\underline{y}})^2}} = \frac{\sum_{i=1}^{n}\underline{x}_i\underline{y}_i - \frac{1}{n}\sum_{i=1}^{n}\underline{x}_i \sum_{i=1}^{n}\underline{y}_i}{\sqrt{\left(\sum_{i=1}^{n}\underline{x}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\underline{x}_i\right)^2\right)\left(\sum_{i=1}^{n}\underline{y}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\underline{y}_i\right)^2\right)}} .$$

t	v	t*v	t^2	v^2
1,85	12	22,20	3,4225	144
2,61	16	41,76	6,8121	256
3,58	22	78,76	12,8164	484
4,49	26	116,74	20,1601	676
5,22	32	167,04	27,2484	1024
17,75	108	426,50	70,4595	2584

$$r = \frac{\sum_{i=1}^{n}\underline{x}_i\underline{y}_i - \frac{1}{n}\sum_{i=1}^{n}\underline{x}_i \sum_{i=1}^{n}\underline{y}_i}{\sqrt{\left(\sum_{i=1}^{n}\underline{x}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\underline{x}_i\right)^2\right)\left(\sum_{i=1}^{n}\underline{y}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\underline{y}_i\right)^2\right)}}$$

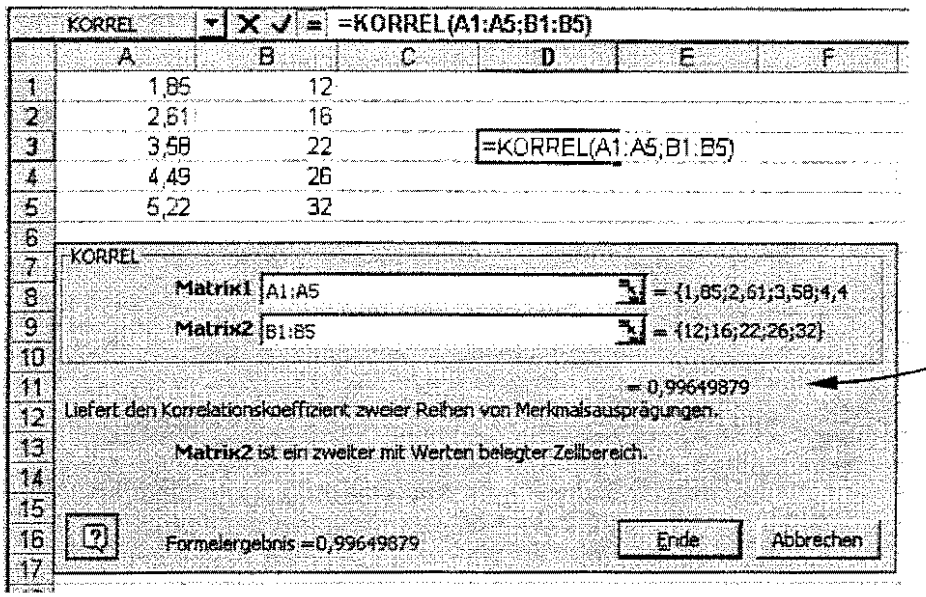
$$r = \frac{426{,}50 - 17{,}75 * 108/5}{\sqrt{(70{,}4595 - 17{,}75 * 17{,}75/5)(108 - 2584 * 2584/5)}}$$

r = 0,9965

EXCEL

Die Meßwerte der beiden Variablen (A: Zeit t ; B: Geschwindigkeit v) werden wie bekannt eingetragen. Der Funktionsaufruf, der den Schätzwert für den Korrelationskoeffizienten liefert, ist

KORREL(Matrix$_{Variable1}$; Matrix$_{Variable2}$):



Das Ergebnis ist:

0,9965

SAS

Die SAS-Prozedur ist PROC CORR:

```
data bsp141;
  input t v;
lines;
1.85      12
2.61      16
3.58      22
4.49      26
5.22      32
;
proc corr;
  var t v;
run;
```

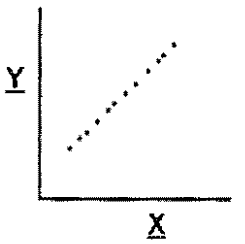
```
Correlation Analysis

   2 'VAR' Variables:  T        V

                        Simple Statistics

Variable          N       Mean    Std Dev        Sum    Minimum    Maximum
T                 5     3.5500     1.3645    17.7500     1.8500     5.2200
V                 5    21.6000     7.9246      108.0    12.0000    32.0000

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 5

                    T                    V

T             1.00000              0.99650
               0.0                  0.0002

V             0.99650              1.00000
               0.0002               0.0
```

Der Schätzwert des Korrelationskoeffizient für die beiden Variablen ist r = 0,9965.

Diese Zahl sagt nichts aus, wenn man sie nicht einordnen kann. Deshalb:
Wie ist diese Maßzahl hinsichtlich des Zusammenhanges zweier Zufallsvariabler einzuschätzen?

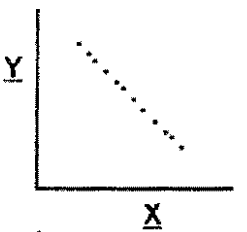
- Für einen strengen linearen Zusammenhang, d. h. alle Werte liegen auf einer Geraden, **und** es gilt, daß sich bei Zunahme einer Variablen um eine Einheit auch die andere Variable um eine Einheit zunimmt (positive Korrelation), d. h. die Werte liegen auf der Winkelhalbierenden, gilt

  $\rho = 1.$



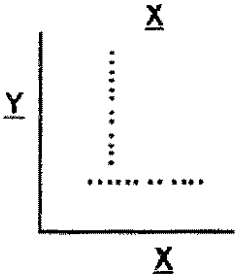
- Für einen strengen linearen Zusammenhang, d. h. alle Werte liegen auf einer Geraden, **und** es gilt, daß sich bei Zunahme einer Variablen um eine Einheit auch die andere Variable um eine Einheit abnimmt (negative Korrelation) gilt

  $\rho = -1.$



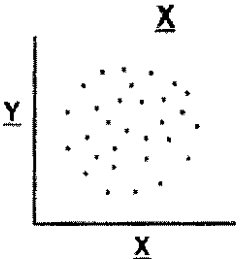
- Für einen strengen linearen Zusammenhang, d. h. alle Werte liegen auf einer Geraden, **und** es gilt, daß sich bei Zunahme einer Variablen um eine Einheit die andere Variable nicht verändert, d. h. die Werte liegen parallel zu einer Achse, gilt

  $\rho = 0.$



- Wird der Wertevorrat (fast) vollständig eingenommen und es ist keinerlei Zusammenhang erkennbar, gilt

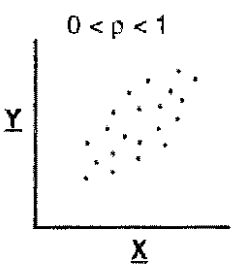
  $\rho = 0.$

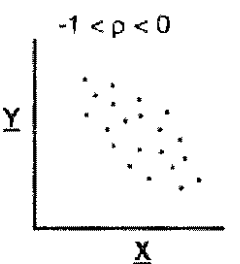


Der Korrelationskoeffizient kann also die Werte

$$\boxed{-1 \leq \rho \leq 1}$$

annehmen.

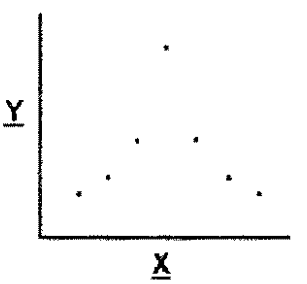




Das gilt für alle Maßzahlen, die einen Zusammenhang - eine Korrelation - zwischen Zufallsvariablen beschreiben.

Da der berechnete Schätzwert r = 0,9965 nahe +1 liegt, kann man von einer streng linearen, positiven Korrelation ausgehen.

Beachtet werden muß, daß auch ein Zusammenhang wie er nebenstehend abgebildet ist, zu einem Korrelationskoeffizienten $\rho = 0$ führt, weil der Zusammenhang zwischen den beiden Zufallsvariablen zwar funktional beschreibbar, aber nicht linear ist!



14.2.1.2 (1-α)-Konfidenzintervall für den Produkt-Momenten-Korrelationskoeffizienten

Unter Zuhilfenahme einer auf die $u_{1-\alpha/2}$-Quantile der standardisierten Normalverteilung basierenden Transformation, der z-Transformation nach Fisher,

$$z(u) = \frac{1}{2}\ln\left(\frac{1+u_{1-\alpha/2}}{1-u_{1-\alpha/2}}\right)$$

kann für den Produkt-Momenten-Korrelationskoeffizienten ρ mit dem Schätzwert r und den Hilfsgrößen

$$z_u = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{u_{1-\alpha/2}}{\sqrt{n-3}} \quad \text{und} \quad z_o = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}$$

ein (1-α)-Konfidenzintervall berechnet werden:

$$\left\langle \frac{e^{2z_u}-1}{e^{2z_u}+1} ; \frac{e^{2z_o}-1}{e^{2z_o}+1} \right\rangle \quad .$$

Dieser Zusammenhang gilt für große n. Es ist unschwer zu erkennen, daß (rein formal) $n \geq 4$ sein muß, um überhaupt ein Konfidenzintervall berechnen zu können – von seinen (approximativen) Eigenschaften einmal abgesehen.

Das 0,95-Konfidenzintervall wird berechnet für die Werte den obigen Beispiels:

$n = 5$
$r = 0{,}9965$
$\alpha = 0{,}05$

$u_{1-\alpha/2} = u_{0,975} = 1{,}96$ ←

```
data uquantil;
  u = probit(0.975);
proc print noobs;
  var u;
run;
```

U
1.95996

$$z_u = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{u_{1-\alpha/2}}{\sqrt{n-3}} = \frac{1}{2}\ln\left(\frac{1+0{,}9965}{1-0{,}9965}\right) - \frac{1{,}96}{\sqrt{5-3}} = 1{,}78726$$

$$z_o = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + \frac{u_{1-\alpha/2}}{\sqrt{n-3}} = \frac{1}{2}\ln\left(\frac{1+0{,}9965}{1-0{,}9965}\right) + \frac{1{,}96}{\sqrt{5-3}} = 4{,}55912$$

$$\left\langle \frac{e^{2z_u}-1}{e^{2z_u}+1} ; \frac{e^{2z_o}-1}{e^{2z_o}+1} \right\rangle = \left\langle \frac{e^{2\cdot 1{,}78726}-1}{e^{2\cdot 1{,}78726}+1} ; \frac{e^{2\cdot 4{,}55912}-1}{e^{2\cdot 4{,}55912}+1} \right\rangle = \langle 0{,}9455 ; 0{,}9998 \rangle$$

Das Konfidenzintervall $\langle 0{,}9455 ; 0{,}9998 \rangle$ ist bezüglich des geschätzten Korrelationskoeffizienzen nicht symmetrisch.

14.2.1.3 Test des Produkt-Momenten-Korrelationskoeffizienten

Für einen Test des Produkt-Momenten-Korrelationskoeffizienten kann die Nullhypothese nur lauten:

H_0: $\rho = 0$: es besteht keinerlei Zusammenhang zwischen den Zufallsvariablen .

Sie wird getestet gegen die Alternativhypothese

H_A^1: $\rho \neq 0$ (zweiseitig) bzw.

H_A^2: $\rho < 0$ (einseitig) oder H_A^3: $\rho > 0$ (einseitig) .

Die Testgröße unter der Nullhypothese ist mit n-2 Freiheitsgraden t-verteilt:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad .$$

Sie wird mit dem t-Quantil $t_{1-\alpha;\ n-2}$ für den zweiseitigen (H_A^1) und $t_{1-\alpha/2;\ n-2}$ für den einseitigen Test (H_A^2 oder H_A^3) verglichen.

Es soll mit $\alpha = 0.05$ (zweiseitig) getestet werden, ob der geschätzte Korrelationskoeffizient des obigen Beispiels r = 0,9965 signifikant von Null verschieden ist.

$r = 0{,}9965$
$n = 5$
$\alpha = 0{,}05$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}9965 * \sqrt{5-2}}{\sqrt{1-0{,}9965^2}} = 20{,}6476 \quad .$$

Dieser berechnete Wert ist so groß, daß sich eigentlich ein Vergleich mit dem t-Quantil erübrigt: der geschätzte Korrelationskoeffizient unterscheidet sich signifikant von Null.

$t_{berechnet} = 20{,}6476 > t_{0,95;\ 3} = 3{,}182$ (Tab. 5.4, Teil 1, S. 53)

Die mit der vorgegebenen Irrtumswahrscheinlichkeit α zu vergleichende Überschreitungswahrscheinlichkeit läßt sich mit einem kleinen SAS-Programm berechnen:

```
data a;
  p = 1-probt(20.6476,3);
proc print noobs; run;
```

P
.00012422

Da $p = 0.00012 < 0.05 = \alpha$ ist die Nullhypothese zu verwerfen: der geschätzte Korrelationskoeffizient ist signifikant von Null verschieden. Die Überschreitungswahrscheinlichkeiten werden im SAS-Output (aufgerundet auf die vierte Dezimalstelle) in der Zeile unterhalb der geschätzten Korrelationskoeffizienten (s. o.) angegeben.

Aufgabe 14.1: Die Daten des jährlichen Genußmittelverbrauchs je Einwohner[2] von Kaffee (in kg) und Tee (in g) sollen hinsichtlich einer (linearen) Abhängigkeit zwischen den Variablen untersucht werden. Der Korrelationskoeffizient und das 95%-Konfidenzintervall sind zu berechnen. Bei $\alpha = 0.05$ ist zu testen, ob für den geschätzten Korrelationskoeffizient angenommen werden kann, daß er von Null verschieden ist.

Jahr	Kaffee	Tee	Jahr	Kaffee	Tee	Jahr	Kaffee	Tee	Jahr	Kaffee	Tee
1965	3.72	139	1966	3.71	127	1967	3.64	129	1968	3.95	143
1969	4.07	148	1970	4.06	134	1971	4.32	154	1972	4.56	164
1973	4.47	173	1974	4.63	169	1975	4.80	171	1976	5.04	196
1977	4.62	204	1978	5.25	204	1979	5.59	239	1980	5.80	248
1981	5.89	259	1982	5.77	245	1983	5.76	249	1984	6.09	245
1985	6.20	246	1986	6.20	235	1987	6.40	240	1988	6.60	240
1989	6.70	226	1990	5.73	186	1991	6.23	202	1992	6.10	177

[2] Bundesministerium für Arbeits- und Sozialordnung (Herausgeber): Statistisches Taschenbuch 1994, Arbeits- und Sozialstatistik, Tab. 6.6

14.2.1.4 Scheinkorrelation und künstliche Korrelation

Der Schätzwert der Korrelationskoeffizienten für die beiden Variablen der Aufgabe 14.1 (s. o.) ist $r = 0{,}87$, sein 0,95-Konfidenzintervall $(0{,}73\,;0{,}94)$.

Zur Analyse des Zusammenhanges zweier Variabler gehört neben der Bewertung der Maßzahl immer eine Betrachtung, ob es für einen solchen Zusammenhang eine kausale, fachliche Begründung gibt, d. h. ob eine solche Analyse überhaupt sinnvoll ist.

Das klassische Beispiel der Analyse des Zusammenhanges zwischen der Anzahl nistender Störche und der Anzahl Neugeborener ist typisch dafür, daß die Korrelation zweier Ereignisse betrachtet wurde, für die es keinen fachlichen Zusammenhang gibt. In einem solchen Fall spricht man von Scheinkorrelation. Scheinkorrelationen sind durch die fachliche Begründung des Zusammenhanges zu vermeiden – sei denn, es wird ausschließlich auf den hohen Unterhaltungswert von Scheinkorrelationen orientiert.

Eine Scheinkorrelation zwischen dem Kaffee- und dem Teeverbrauch ist zwar nicht auszuschließen. Die zwischen beiden Variablen berechnete Korrelation ist aber eher auf eine künstliche Korrelation zurückzuführen, die dann vorliegt, wenn die betrachtete Korrelation enger erscheint, als sie es in Wirklichkeit ist.
Bei der Betrachtung des jährlichen Prokopfverbrauchs liegen dieselben Bevölkerungszahlen zugrunde, so daß ein zusätzlicher, verzerrender Einfluß nicht vorliegen dürfte.

Eine künstliche Korrelation wird vor allem dann erreicht, wenn durch Stichprobennahme oder Behandlung (im allgemeinen Sinne) ein weiterer, nicht in die Betrachtung eingehender Einfluß verstärkend wirkt. Das kann eine bestimmte Bodenart, Pflanzenart, Behandlung oder ein anderer exogener Faktor sein. Die Analyse der Versuchsfrage und die sinnvolle Auswahl des Stichprobenverfahrens können durch fachliche und sachliche Überlegungen die Scheinkorrelation und die künstliche Korrelation so gering wie möglich halten oder sogar vermeiden.

Formal berechnet werden können die Maßzahlen der Korrelation sowohl bei Schein- als auch künstlicher Korrelation. Wichtig ist die fachliche Interpretation.

14.2.1.5 Die Schätzung des Produkt-Momenten-Korrelationskoeffizienten mittels SAS-Prozedur PROC FREQ

Ausgehend vom physikalischen Beispiel 14.1 soll der Produkt-Momenten-Korrelationskoeffizienten mit Hilfe der SAS-Prozedur PROC FREQ geschätzt werden.

```
proc freq data=bsp141;
  tables t * v /measures noprint;
run;
```

Option `measures` : liefert eine Vielzahl von Schätzwerten für den Zusammenhang (Korrelation) zweier Variabler
Option `noprint`: unterdrückt die Ausgabe der Häufigkeitstafel und einiger Teststatistiken

Für die hier zu betrachtende Schätzung des Produkt-Momenten-Korrelationskoeffizienten soll nur auf die Ergebniszeile mit `Pearson Correlation` geachtet werden. Der Schätzwert ist wie oben 0.996.

STATISTICS FOR TABLE OF T BY V

Statistic	Value	ASE
Gamma	1.000	0.000
Kendall's Tau-b	1.000	0.000
Stuart's Tau-c	1.000	0.000
Somers' D C|R	1.000	0.000
Somers' D R|C	1.000	0.000
Pearson Correlation	0.996	0.002
Spearman Correlation	1.000	0.000
Lambda Asymmetric C|R	1.000	0.000
Lambda Asymmetric R|C	1.000	0.000
Lambda Symmetric	1.000	0.000
Uncertainty Coefficient C|R	1.000	0.000
Uncertainty Coefficient R|C	1.000	0.000
Uncertainty Coefficient Symmetric	1.000	0.000

Sample Size = 5

Zusätzlich zum Schätzwert $r_{Pearson}$ erfolgt eine zweite Angabe. Das ist der asymptotische Standardfehler ASE (asymptotic standard error). Der ASE ist gültig bei hinreichend großer Stichprobenzahl, denn nur dann ist die Maßzahl für die Korrelation approximativ normalverteilt und es kann ein (1-α)-Konfidenzintervall angegeben werden[3]. Für $\alpha = 0.05$ lautet es mit $u_{0,975} = 1,96$

$$\langle r_{Pearson} - 1,96 * ASE ; r_{Pearson} + 1,96 * ASE \rangle \quad .$$

Mit $r_{Pearson} = 0,996$ und $ASE = 0,002$ ergibt sich ein 0,95-Konfidenzintervall $\langle 0,992 ; 1 \rangle$. Es ist wesentlich enger als das zuvor berechnete (s. 14.2.1.2). Beide Berechnungen sind approximativ. Die Verwendung des asymptotischen Standardfehlers ASE unterstreicht sehr stark, daß die Berechnung dieses Konfidenzintervalls für große n gilt.

14.2.2 Der Rangkorrelationskoeffizient von SPEARMAN

Es gibt, wie bereits im obigen SAS-Output zu erkennen ist, mehrere Maßzahlen der Korrelation. Das ist vorteilhaft, denn nicht alle Variablen sind metrische Zufallsvariable mit binormalverteilten Grundgesamtheiten.

Die Korrelation zwischen zwei Zufallsvariablen, die nicht zweidimensional normalverteilt sind, wird mit Hilfe des Spearmanschen Rangkorrelationskoeffizienten geschätzt. Dabei kann der Zusammenhang zwischen den beiden Zufallsvariablen nichtlinear sein, muß aber monoton wachsend bzw. monoton fallend sein.

Für das Beispiel 14.2 wird von der Aufgabe 14.1 ausgegangen und anstelle der metrischen Werte werden Rangzahlen gebildet. Der gleiche Maßstab, d. h. entweder kg oder g, muß nicht hergestellt werden, da die Ränge für jede Variable einzeln gebildet werden.

```
proc rank data = aufg141 out = rank141;
  var kaffee tee;
  ranks rkaffee rtee;
run;
```

[3] STOKES, M. E., C. S. DAVIS and G. G. KOCH: Categorial Data Analysis Using the SAS System Cary, NC: SAS Institute Inc., 1995. 499 pp.

Diese nichtnormal verteilten Rangzahlen `rkaffee` und `rtee` sind die Daten, zu denen der Spearmansche Rangkorrelationskoeffizient berechnet werden soll.
Der Spearmansche Rangkorrelationskoeffizient wird berechnet nach

$$r_S = 1 - \frac{6\sum(R_X - R_Y)^2}{n(n^2-1)} = 1 - \frac{6\sum(D_i)^2}{n(n^2-1)} \quad ,$$

wobei $D = R_X - R_Y$ die Differenz der gepaarten Ränge ist.

Papier und Bleistift

Jahr	Kaffee	Tee	R_{Kaffee}	R_{Tee}	$D = (R_{Kaffee} - R_{Tee})$	D^2
1965	3,72	139	3,0	4,0	-1,0	1,00
1966	3,71	127	2,0	1,0	1,0	1,00
1967	3,64	129	1,0	2,0	-1,0	1,00
1968	3,95	143	4,0	5,0	-1,0	1,00
1969	4,07	148	6,0	6,0	0,0	0,00
1970	4,06	134	5,0	3,0	2,0	4,00
1971	4,32	154	7,0	7,0	0,0	0,00
1972	4,56	164	9,0	8,0	1,0	1,00
1973	4,47	173	8,0	11,0	-3,0	9,00
1974	4,63	169	11,0	9,0	2,0	4,00
1975	4,80	171	12,0	10,0	2,0	4,00
1976	5,04	196	13,0	14,0	-1,0	1,00
1977	4,62	204	10,0	16,5	-6,5	42,25
1978	5,25	204	14,0	16,5	-2,5	6,25
1979	5,59	239	15,0	20,0	-5,0	25,00
1980	5,80	248	19,0	26,0	-7,0	49,00
1981	5,89	259	20,0	28,0	-8,0	64,00
1982	5,77	245	18,0	23,5	-5,5	30,25
1983	5,76	249	17,0	27,0	-10,0	100,00
1984	6,09	245	21,0	23,5	-2,5	6,25
1985	6,20	246	23,5	25,0	-1,5	2,25
1986	6,20	235	23,5	19,0	4,5	20,25
1987	6,40	240	26,0	21,5	4,5	20,25
1988	6,60	240	27,0	21,5	5,5	30,25
1989	6,70	226	28,0	18,0	10,0	100,00
1990	5,73	186	16,0	13,0	3,0	9,00
1991	6,23	202	25,0	15,0	10,0	100,00
1992	6,10	177	22,0	12,0	10,0	100,00
						732,00

$$r_S = 1 - \frac{6\sum(D_i)^2}{n(n^2-1)} = 1 - \frac{6 * 732{,}00}{28(28^2-1)} = 1 - 0{,}2 = 0{,}8$$

Natürlich sind auch hier – wie bei Rangstatistiken üblich – die Bindungen zu berücksichtigen:

$$r_{S_{corr}} = 1 - \frac{6\sum(R_X - R_Y)^2}{n(n^2-1)/\left(\sum(t_{X_i}^3 - t_{X_i}) - \sum(t_{Y_i}^3 - t_{Y_i})\right)/2} = 1 - \frac{6\sum(D_i)^2}{n(n^2-1)/\left(\sum(t_{X_i}^3 - t_{X_i}) - \sum(t_{Y_i}^3 - t_{Y_i})\right)/2}$$

Rang 23,5 2 x

Rang 16,5 2 x
Rang 21,5 2 x
Rang 23,5 2 x

$$\sum(t_{X_i}^3 - t_{X_i}) - \sum(t_{Y_i}^3 - t_{Y_i}) = (2^3-2) - \left[(2^3-2)+(2^3-2)+(2^3-2)\right] = 12$$

$$r_{S_{corr}} = 1 - \frac{6\sum(D_i)^2}{n(n^2-1) - \left(\sum(t_{X_i}^3 - t_{X_i}) - \sum(t_{Y_i}^3 - t_{Y_i})\right)/2} = 1 - \frac{6 * 732{,}00}{28(28^2-1) - 12/2} = 1 - 0{,}2 = 0{,}8$$

Die Korrektur bewirkt (für dieses Beispiel) erst Änderungen in der 4. Dezimalstelle.

SAS

```
proc freq data=aufg141;
  tables kaffee * tee /measures noprint;
run;
```

```
STATISTICS FOR TABLE OF KAFFEE BY TEE

Statistic                                 Value          ASE
-------------------------------------------------------------
Gamma                                     0.631        0.099
Kendall's Tau-b                           0.628        0.099
Stuart's Tau-c                            0.627        0.100

Somers' D C|R                             0.626        0.099
Somers' D R|C                             0.629        0.099

Pearson Correlation                       0.869        0.045
Spearman Correlation                      0.800        0.094

Lambda Asymmetric C|R                     0.962        0.038
Lambda Asymmetric R|C                     0.885        0.063
Lambda Symmetric                          0.923        0.037

Uncertainty Coefficient C|R               0.984        0.011
Uncertainty Coefficient R|C               0.955        0.016
Uncertainty Coefficient Symmetric         0.969        0.009

Sample Size = 28
```

Produkt-Momenten-Korrelationskoeffizient

Rangkorrelationskoeffizient

Das SAS-Output liefert noch weitere Korrelations-Statistiken, auf die hier nicht eingegangen werden soll. Sie werden beispielsweise in STOKES, M. E., C. S. DAVIS and G. G. KOCH: Categorial Data Analysis Using the SAS System, Cary, NC: SAS Institute Inc., 1995, 499 pp. beschrieben.

15 Regressionsanalyse

Bei der Korrelationsanalyse geht es darum, ein Maß für die Stärke des Zusammenhanges von Zufallsvariablen zu finden. Soll die (stochastische) Abhängigkeit von (quantitativen) Variablen in Form einer analytischen Funktion beschrieben werden, bedient man sich der Regressionsanalyse. Während es bei der Korrelationsanalyse für die Stärke des Zusammenhanges uninteressant ist, welche Variable von welcher abhängt, ist das für die Formulierung der Regressionsfunktion notwendig. So wird zwischen dem Regressand, der unabhängigen Variablen (bzw. Variablenvektor) oder der „Zielgröße", und dem Regressor oder den Regressoren, den abhängigen Variablen oder den „Einflußgrößen" unterschieden. Die zu schätzenden Koeffizienten der Regressionsfunktion nennt man Regressionskoeffizienten. Ein besonderer Regressionskoeffizient ist das Absolutglied, der Intercept.

Ist die Regressionsfunktion eine Gerade, so spricht man von der linearen Regression. Als einfache lineare Regression wird eine lineare Regressionsfunktion mit einem Regressanden und einem Regressor bezeichnet; die zweifache lineare Regressionsfunktion ist eine lineare Funktion mit zwei Regressoren. Wird die lineare Regressionsfunktion um das quadratische Glied – bzw. die quadratischen Glieder – erweitert, spricht man von der quadratischen Regression, bei Gliedern dritten Grades von der kubischen und allgemein bei einem Polynom r-ten Grades von der polynomialen Regression.

Unter das Teilgebiet der nichtlinearen Regression fallen die quasilinearen Regressionsfunktionen, das sind Funktionen, die auf eine lineare Regressionsfunktion zurückgeführt werden können, und die eigentlich nichtlinearen Regressionsfunktionen, bei denen mindestens ein Regressor echt nichtlinear ist. Unter die letze Gruppe fallen unter anderem die Wachstumsfunktionen.

15.1 Regressionsfunktionen

Bezeichnen wir die (zufälligen) Beobachtungswerte des Regressanden mit $\underline{y}_i$ (i = 1, 2, ..., n), so lautet die allgemeine Modellgleichung $\underline{y}_i = \eta_i + \underline{e}_i$ (i = 1, 2, ..., n). Die Beobachtungswerte $\underline{y}_i$ des Regressanden werden bestimmt durch die Werte der Regressionsfunktion η_i und einen Zufallsfehler $\underline{e}_i$. Die Zufallsfehler $\underline{e}_i$ (i = 1, 2, ..., n) sind voneinander unabhängig, haben den Erwartungswert Null und die gleiche Varianz σ^2 .

Die einfachste Regressionsfunktion, die der einfachen linearen Regression, lautet $\eta = \beta_0 + \beta_1 x$, wobei β_0 und β_1 die Regressionskoeffizienten sind: β_0 das Absolutglied (Intercept) und β_1 der Anstieg der Regressionsgeraden. x ist der Regressor.

Andere Regressionsfunktionen mit s Regressoren sind:

lineare Regression: $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_s x_s = \beta_0 + \sum_{k=1}^{s} \beta_k x_k$

quadratische Regression: $\eta = \beta_0 + \beta_{11} x_1 + \beta_{12} x_2 + ... + \beta_{1s} x_s + \beta_{21} x_1^2 + \beta_{22} x_2^2 + ... + \beta_{2s} x_s^2$

$$= \beta_0 + \sum_{k=1}^{s} \beta_{1k} x_k + \sum_{k=1}^{s} \beta_{2k} x_k^2$$

polynominale Regression: $\eta = \beta_0 + \sum_{k=1}^{s} \beta_{1k} x_k + \sum_{k=1}^{s} \beta_{2k} x_k^2 + ... + \sum_{k=1}^{s} \beta_{rk} x_k^r$

quasilineare Regression: $\eta = \beta_0 + \beta_1 f_1(x_1, x_2, .., x_s) + \beta_2 f_2(x_1, x_2, .., x_s) + ... + \beta_q f_q(x_1, x_2, .., x_s)$

die $f_1(x_1, x_2, .., x_s)$ sind bekannte Funktionen

eigentlich nichtlineare Regression: Regressionskoeffizienten können auch nichtlinear in die Funktion der Regressoren eingehen

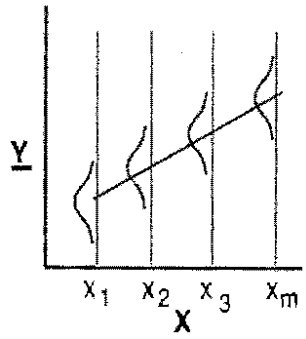
Der Charakter der x_i bestimmt das Modell der Regressionsanalyse.

15.2 Die Modelle der Regressionsanalyse

Die Werte x_i der Merkmale, die als Regressoren ausgewählt werden, können vor dem Versuch festgelegt werden; sie sind einstellbare Meßstellen. Dann spricht man vom

Modell I der Regressionsanalyse.

Im allgemeinen wird als Verteilung für die Zufallsfehler $\underline{e}_{ij}$ ($i = 1, 2, ..., n$; $j = 1, 2, ...m_i$) die Normalverteilung $N[0, \sigma^2]$ angenommen.
n: Anzahl der Meßstellen, m_i: Anzahl der Messungen an der Meßstelle i

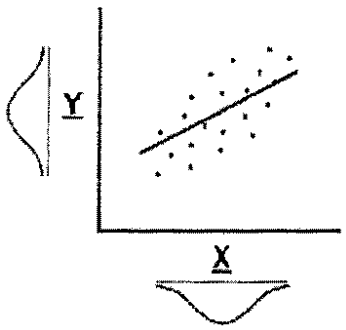


Sind die Werte $\underline{x}_i$ der Merkmale, die als Regressoren ausgewählt werden, zufällige Beobachtungswerte, die nicht vor dem Versuch festgelegt werden können, dann spricht man vom

Modell II der Regressionsanalyse.

Die Zufallsfehler $\underline{e}_i$ sind mit $N[0, \sigma^2]$ normalverteilt ($i = 1, 2, ..., n$).

Zusätzlich gilt: die Zufallsfehler $\underline{e}_i$ sind von den $\underline{x}_i$ unabhängig,
$E(\underline{x}_i) = \mu_x$, $VAR(x_i) = \sigma^2_x$ für alle i.



Da bei mehreren Regressoren diese sowohl einstellbare Meßstellen (s. Modell I der Regressionsanalyse) als auch zufällige Beobachtungswerte (s. Modell II der Regressionsanalyse) haben können, wird ein solches Modell **gemischtes Modell der Regressionsanalyse** genannt.

Bei der Auswertung sind die Modelle, was ihre numerische Seite anbelangt, gleich. Da die Meßstellen beim Modell II der Regressionsanalyse (und beim gemischten Modell) nicht einstellbar sind, unterscheiden sich die Modelle bei der Versuchsplanung wesentlich. Des weiteren ist die Korrelationsanalyse beim Modell I der Regressionsanalyse nicht gestattet, weil der Regressor nicht zufällig ist.

15.3 Berechnen der Statistiken für das einfache lineare Regressionsmodell

15.3.1 Die Regressionskoeffizienten

Von einfacher Regression wird gesprochen, wenn ein Regressor und ein Regressand betrachtet werden. Ist die Beziehung zwischen diesen beiden Variablen linear, so liegt eine einfache lineare Regression vor. Wenn beide Variable Zufallsvariable sind, wird durch die Aufgaben- und Zielstellung festgelegt, welche dieser beiden Variablen Regressand und welche Regressor ist.

Die Modellgleichung für die einfache lineare Regression (s.o.) lautet allgemein:

$$\underline{y}_i = \beta_0 + \beta_1 x_i + \underline{e}_i \quad (i = 1, 2, ..., n) \ .$$

Es geht also zunächst darum, die beiden Regressionskoeffizienten β_0 und β_1 (Linearität vorausgesetzt) zu schätzen. Diese Schätzung basiert auf der Methode der kleinsten Quadrate. Davon ausgehend, daß die Gerade in der Grundgesamtheit als Beziehung zwischen den beiden Variablen gilt, stellt sie die geschätzten Erwartungswerte für den Regressanden y an der Stelle i

des Regressors x dar. Die Koeffizienten der Regressionsgeraden werden nun so berechnet, daß die Quadrate der Abweichungen der geschätzten Erwartungswerte von jedem beobachteten Wert ein Minimum sind. Der Anstieg der Regressionsgeraden β_1 wird geschätzt mit

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_\cdot)(y_i - \bar{y}_\cdot)}{\sum_{i=1}^{n}(x_i - \bar{x}_\cdot)^2} = \frac{SP_{xy}}{SQ_x} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

und das Absolutglied der Regressionsgeraden β_0 mit

$$b_0 = \bar{y}_\cdot - b_1 * \bar{x}_\cdot = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 * \frac{1}{n}\sum_{i=1}^{n} x_i$$

Beispiel

Der Verbrauch von Kaffee und Tee soll betrachtet werden (s. Aufgabe 14.1). Angenommen wird, daß der Verbrauch von Kaffee in kg (Y) in Abhängigkeit von Verbrauch von Tee in g (X) linear ist. Die Regressionskoeffizienten sollen berechnet werden.

	Kaffee (Y)	Tee (X)	X * Y	X^2
	3,72	139	517,08	19321,00
	3,71	127	471,17	16129,00
	3,64	129	469,56	16641,00
	3,95	143	564,85	20449,00
	4,07	148	602,36	21904,00
	4,06	134	544,04	17956,00
	4,32	154	665,28	23716,00
	4,56	164	747,84	26896,00
	4,47	173	773,31	29929,00
	4,63	169	782,47	28561,00
	4,80	171	820,80	29241,00
	5,04	196	987,84	38416,00
	4,62	204	942,48	41616,00
	5,25	204	1071,00	41616,00
	5,59	239	1336,01	57121,00
	5,80	248	1438,40	61504,00
	5,89	259	1525,51	67081,00
	5,77	245	1413,65	60025,00
	5,76	249	1434,24	62001,00
	6,09	245	1492,05	60025,00
	6,20	246	1525,20	60516,00
	6,20	235	1457,00	55225,00
	6,40	240	1536,00	57600,00
	6,60	240	1584,00	57600,00
	6,70	226	1514,20	51076,00
	5,73	186	1065,78	34596,00
	6,23	202	1258,46	40804,00
	6,10	177	1079,70	31329,00
Summe	145,90	5492,00	29620,28	1128894,00

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{29620{,}28 - \frac{5492{,}00 * 145{,}90}{28}}{1128894{,}00 - \frac{5492{,}00^2}{28}}$$

$$= 0{,}019$$

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 * \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{145{,}90}{28} - 0{,}0194 * \frac{5492{,}00}{28}$$

$$= 1{,}405$$

Die geschätzte Regressionsgerade lautet:

Kaffeeverbrauch (in kg) = 1,405 + 0,019 * Teeverbrauch (in g)

Bei der Interpretation merkt man, daß die Beachtung der jeweiligen Maßeinheit wichtig ist. Bei der (statistischen) Zunahme des Teeverbrauchs von einem Gramm steigt der Kaffeeverbrauch um 0,019 kg = 19 g.

15.3.2 Das Bestimmtheitsmaß

Die statistische Maßzahl Bestimmtheitsmaß B liegt im Intervall $0 \leq B \leq 1$. Sie ist ein Maß für die Anpassung der beobachteten Werte an die Regressionsgerade. Gleichzeitig beschreibt das Bestimmtheitsmaß die Abhängigkeit der beiden Zufallsvariablen. Das bedeutet: die beobachteten Werte können (fast) auf der Regressionsgeraden liegen; wenn diese aber parallel zu einer der Koordinatenachsen verläuft, ist B gleich Null bzw. nahe Null.
Diese statistische Maßzahl Bestimmtheitsmaß ist nur für das Modell II der Regressionsanalyse definiert. Eine Interpretation für das Modell I ist also falsch! (Natürlich ist eine solche Größe auch für das Regressionsmodell I numerisch zu berechnen. Aber durch die fixe Wahl der Meß- oder Beobachtungsstellen x_i ist das Ergebnis in starkem Maße von der Wahl dieser x_i abhängig und damit manipulierbar.)

Es gibt viele Varianten, das Bestimmtheitsmaß zu berechnet. Einige wären:

$$B = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_\cdot)^2} = 1 - \frac{SQ}{SQ_y} = \frac{b_1 * SP_{xy}}{SQ_y} = \frac{SP_{xy}^2}{SQ_x * SQ_y} = \frac{\textit{durch das Modell erkärte Varianz}}{\textit{Gesamtvarianz}}$$

Es gilt $B = r^2$, wobei r der (Produkt-Momenten-)Korrelationskoeffizient ist.

Für das Beispiel ist $SQ_y = 25{,}767$ und folglich $B = 1 - \frac{SQ}{SQ_y} = 1 - \frac{6{,}314}{25{,}767} = 0{,}755$.

Das bedeutet, daß 75,5 % des Anteils an der Varianz von Y (Kaffeeverbrauch) durch Veränderung des Regressors X (Teeverbrauch) erklärt wird.

15.3.3 Die Varianz um die Regressionsgerade, die Restvarianz

Für das Beispiel werden zur Veranschaulichung die geschätzten Erwartungswerte, die die Punkte der Regressionsgeraden sind, nur für wenige Werte (Teeverbrauchs-Werte) berechnet:

```
data erwart;
  do tee = 130 to 250 by 20;
    kaffee = 1.405 + 0.019 * tee;
    output;
  end;
proc print data=erwart noobs;
run;
```

TEE	KAFFEE
130	3.875
150	4.255
170	4.635
190	5.015
210	5.395
230	5.775
250	6.155

Betrachtet werden die jeweiligen Abweichungen des Kaffeeverbrauchs von den geschätzten Erwartungswerten (s. u.).

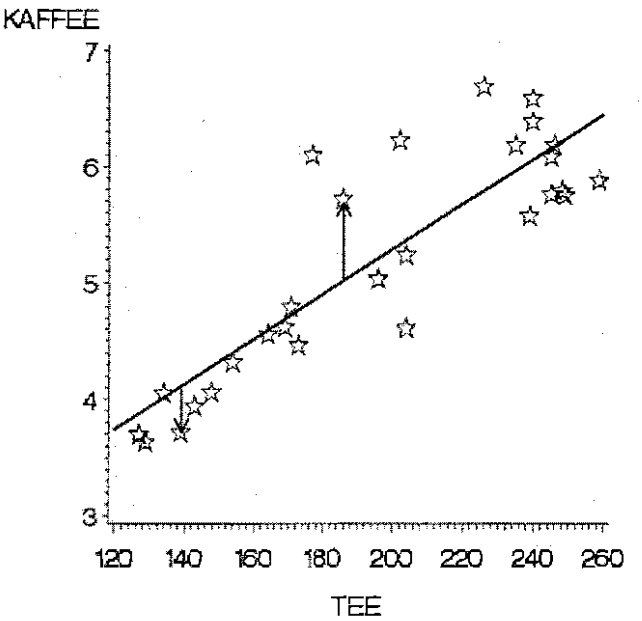
Der Mittelwert der Differenzen ist 0,078. Er müßte Null sein (und ist es auch !), was nur auf die Rechengenauigkeit zurück zu führen ist.

Die Summe der Quadrate dieser Differenzen, der Abweichungsquadrate,

$$SQ = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 + b_1 * x_i)^2$$

Tee	Kaffee	$\widehat{E(Kaffee)}$	Differenz
127	3.71	3.82	-0.11
129	3.64	3.86	-0.22
134	4.06	3.95	0.11
139	3.72	4.05	-0.33
143	3.95	4.12	-0.17
148	4.07	4.22	-0.15
154	4.32	4.33	-0.01
164	4.56	4.52	0.04
169	4.63	4.62	0.01
171	4.80	4.65	0.15
173	4.47	4.69	-0.22
177	6.10	4.77	1.33
186	5.73	4.94	0.79
196	5.04	5.13	-0.09
202	6.23	5.24	0.99
204	4.62	5.28	-0.66
204	5.25	5.28	-0.03
226	6.70	5.70	1.00
235	6.20	5.87	0.33
239	5.59	5.95	-0.36
240	6.40	5.97	0.43
240	6.60	5.97	0.63
245	5.77	6.06	-0.29
245	6.09	6.06	0.03
246	6.20	6.08	0.12
248	5.80	6.12	-0.32
249	5.76	6.14	-0.38
259	5.89	6.33	-0.44

Diese Abweichungen sind in der Grafik an zwei Punkten demonstriert:



berücksichtigt die Anzahl der Modellparameter, so daß die Varianz um die Regressionsgerade allgemein

$$s^2 = \frac{SQ}{n-s-1}$$

mit n - s - 1 Freiheitsgraden ist, wobei s die Anzahl der Regressoren (Einflußgrößen) ist. Im Falle der einfachen linearen Regression sind n – 2 Freiheitsgrade zu betrachten. Diese Varianz ist die Varianz der Regression, die Restvarianz. Ihr Zahlenwert ist mit SQ = 6,314 für obiges Beispiel

$$s^2 = \frac{6{,}314}{n-2} = 0{,}243 \quad .$$

Für das Modell der Regressionsanalyse hängen – wie für das Varianzanalysemodell – die Summen der Abweichungsquadrate additiv zusammen: $SQ_{Gesamt} = SQ_{Regression} + SQ_{Rest}$ bzw.

$$SQ_{Rest} = SQ_y - \frac{SP_{xy}^2}{SQ_x} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \quad .$$

$$SQ_{Gesamt} = SQ_y = 25{,}767$$

$$SP_{xy} = 29620{,}28 - \frac{5492{,}00 * 145{,}90}{28} = 1003{,}037$$

$$SQ_x = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 1128894 - \frac{5492^2}{28} = 51677{,}429$$

$$SQ_{Regression} = \frac{SP_{xy}^2}{SQ_x} = \frac{(1003{,}037)^2}{51677{,}429} = 19{,}469$$

$SQ_{Rest} = 25{,}767 - 19{,}469 = 6{,}298$

Berechnet werden nun unter Berücksichtigung der entsprechenden Freiheitsgrade die Varianzen (d.h. deren Schätzwerte):

$$s^2_{Regression} = \frac{SQ_{Regression}}{FG_{Regression}} = \frac{SQ_{Regression}}{s} = \frac{SQ_{Regression}}{1} = 19{,}469 \quad [s = 1 \text{ gilt nur für die einfache Regression}]$$

$$s^2_{Rest} = \frac{SQ_{Rest}}{FG_{Rest}} = \frac{SQ_{Rest}}{n - s - 1} = \frac{6{,}298}{28 - 2} = \frac{6{,}298}{26} = 0{,}242$$

[Diese Restvarianz und die aus den Differenzen der beobachteten Werte und den geschätzten Erwartungswerten berechnete Varianz sind dieselben und müssen natürlich übereinstimmen. Die numerischen Unterschiede resultieren ausschließlich aus Rundungsfehlern.]

15.3.4 Test auf Linearität

Aus diesen beiden Varianzen läßt sich eine F-verteilte Prüfgröße konstruieren, mit deren Hilfe getestet werden kann, ob die Nullhypothese, wonach die beobachteten Werte einer Geraden folgen, beibehalten werden kann oder zu verwerfen ist. Für das Beispiel ergibt sich:

$$F = \frac{s^2_{Regression}}{s^2_{Rest}} = \frac{19{,}469}{0{,}242} = 80{,}45 \; > \; F_{1-\alpha/2;\,1,\,26} = 4{,}225 \; \rightarrow \text{ die Nullhypothese ist abzulehnen!}$$

Dieses Ergebnis widerspricht der Anschauung. Das könnte daran liegen, daß die Irrtumswahrscheinlichkeit ungünstig gewählt wurde, denn es soll ja möglichst lange die Nullhypothese beibehalten werden (im Gegensatz zum „üblichen" Testziel). Die grafische Darstellung läßt keinen anderen Regressionsansatz außer dem linearen erkennen.

15.3.5 Berechnen der Statistiken mit SAS

In SAS wird die Berechnung dieser Statistiken mit Hilfe der Prozedur REG realisiert. Für das Beispiel sind die Programmzeilen:

```
data a;
  input jahr kaffee tee @@;
lines;
1965  3.72  139      1966  3.71  127      1967  3.64  129      1968  3.95  143
1969  4.07  148      1970  4.06  134      1971  4.32  154      1972  4.56  164
1973  4.47  173      1974  4.63  169      1975  4.80  171      1976  5.04  196
1977  4.62  204      1978  5.25  204      1979  5.59  239      1980  5.80  248
1981  5.89  259      1982  5.77  245      1983  5.76  249      1984  6.09  245
1985  6.20  246      1986  6.20  235      1987  6.40  240      1988  6.60  240
1989  6.70  226      1990  5.73  186      1991  6.23  202      1992  6.10  177
;
proc reg;
  model kaffee = tee;
run;
quit;
```

Innerhalb einer Prozedur REG können mehrere Modellansätze gerechnet werden, die, wenn sie nicht mit einem Label markiert sind, durchnummeriert werden. Das ist der Hinweis in der ersten Zeile des Output: Model1. Das Output ist dreigeteilt, in den Varianzanalyseteil, den Abschnitt, der einige Maßzahlen enthält, und den eigentlichen Regressionsteil.

Regressionsanalyse

```
Model: MODEL1
Dependent Variable: KAFFEE

Analysis of Variance

                          Sum of          Mean
Source            DF      Squares        Square       F Value       Prob>F

Model              1     19.46853      19.46853        80.369       0.0001
Error             26      6.29826       0.24224
C Total           27     25.76679

    Root MSE       0.49218     R-square       0.7556
    Dep Mean       5.21071     Adj R-sq       0.7462
    C.V.           9.44553

Parameter Estimates

                   Parameter       Standard    T for H0:
Variable  DF        Estimate          Error   Parameter=0    Prob > |T|

INTERCEP   1       1.403664     0.43473102         3.229        0.0034
TEE        1       0.019410     0.00216508         8.965        0.0001
```

In der Varianztabelle sind die Restvarianz [Mean SquareError = 0.24224] und der F-Wert [F-Value = 80.369] mit der dazugehörigen Überschreitungswahrscheinlichkeit [Prob>F] und den Freiheitsgraden [DF] zu finden.

In anschließenden Abschnitt sind links die Standardabweichung, d. h. die Wurzel aus der Restvarianz [Root MSE = 0.49218] , der Mittelwert für den Regressanden Kaffee-Verbrauch [Dep Mean = 5.21071] und der Variationskoeffizient [C.V. = 9.44553] zu finden. Der Variationskoeffizient ist der mit 100 multiplizierte Quotient aus s_{Rest} und dem mittleren Wert des Regressanden : $C.V. = \frac{s_{Rest}}{\bar{y}_.} * 100 = \frac{0{,}49218}{5{,}21071} = 9{,}44553$.

Rechts steht das Bestimmtheitsmaß [R-square = 0.7556] und darunter das adjustierte Bestimmtheitsmaß [Adj R-sq = 0.7462]. Das adjustierte Bestimmtheitsmaß wird berechnet

$$B_{adjustiert} = 1 - \frac{(1-B)(n-1)}{FG_{Rest}} = 1 - \frac{(1-0{,}7556)(28-1)}{26} = 0{,}7462$$

Im dritten Ausgabeteil sind die Schätzwerte für die Regressionskoeffizienten:

Absolutglied der Regressionsgeraden b_0 [INTERCEP] : 1.403664

Anstieg der Regressionsgeraden b_1 [TEE] : 0.019410

ihre Standardfehler:

für b_0: $s_{\bar{b}_0} = 0.43473102$

für b_1: $s_{\bar{b}_1} = 0.00216508$

die t-verteilten Prüfgrößen zum Test der Nullhypothesen: die Regressionskoeffizienten sind Null :

für b_0: $t = \frac{b_0}{s_{\bar{b}_0}} = 3{,}229$

für b_1: $t = \frac{b_1}{s_{\bar{b}_1}} = 8{,}965$

mit den dazugehörigen Überschreitungswahrscheinlichkeiten:

für b_0: $p = 0{,}0034$

für b_1: $p \leq 0{,}0001$

aufgelistet.

15.3.6 (1-α)-Konfidenzintervalle der Regressionskoeffizienzen

Die zweiseitigen (1-α)-Konfidenzintervalle für die Regressionskoeffizienten sind:

für β_0: $\left\langle b_0 - t_{1-\alpha/2;FG_{Rest}} * s_{b_0} \;;\; b_0 + t_{1-\alpha/2;FG_{Rest}} * s_{b_0} \right\rangle$

für β_1: $\left\langle b_1 - t_{1-\alpha/2;FG_{Rest}} * s_{b_1} \;;\; b_1 + t_{1-\alpha/2;FG_{Rest}} * s_{b_1} \right\rangle$

Ihre Eigenschaften sind allerdings im Modell I und Modell II der Regressionsanalyse verschieden.

Für das Beispiel Kaffeeverbrauch (in kg) in Abhängigkeit vom Teeverbrauch (in g) werden die Konfidenzintervalle für $\alpha = 0.05$ berechnet.

$t_{1-\alpha/2\,;\,FG_{Rest}} = t_{0,975;\,26} = 2{,}056$

für β_0: $\langle 1{,}404 - 2{,}056 * 0{,}435 \;;\; 1{,}404 + 2{,}056 * 0{,}435 \rangle = \langle 0{,}51 \;;\; 2{,}30 \rangle$

für β_1: $\langle 0{,}019 - 2{,}056 * 0{,}002 \;;\; 0{,}019 + 2{,}056 * 0{,}002 \rangle = \langle 0{,}015 \;;\; 0{,}023 \rangle$

15.3.7 Konfidenzintervalle der Erwartungswerte der Regressionsgeraden und Vertrauensintervalle

Da die Regressionsgerade aus den Erwartungswerten für den Regressanden $\underline{Y}$ gebildet wird, lassen sich auch um jeden der Schätzwerte zweiseitige (1-α)-Konfidenzintervalle berechnen, die mit einer Wahrscheinlichkeit (1-α) den wahren Parameter, d. h. den Punkt der Regressionsgeraden in der Grundgesamtheit, überdecken.

Der zu berücksichtigende Standardfehler ist abhängig vom jeweiligen x_i:

$$\left\langle \hat{y}_i - t_{1-\alpha/2;FG_{Rest}} * s_{\hat{y}_i} \;;\; \hat{y}_i + t_{1-\alpha/2;FG_{Rest}} * s_{\hat{y}_i} \right\rangle \qquad \text{mit} \qquad s_{\hat{y}_i} = s_{Rest} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x}_\bullet)^2}{\sum_{i=1}^{n} x_i^2}}$$

Das hat zur Folge, daß das (1-α)-Konfidenzintervall im Mittelpunkt $\bar{x}_\bullet$ am kleinsten ist. Je weiter man sich von diesem Mittelpunkt entfernt, desto größer wird s_{Y_i} und damit auch das dazugehörige (1-α)-Konfidenzintervall. Diese Konfidenzintervalle sollen für das obige Beispiel bei $\alpha = 0{,}05$ berechnet werden.

Bekannt oder einfach zu berechnen sind:

$t_{1-\alpha/2\,;\,FG_{Rest}} = t_{0,975;\,26} = 2{,}056$

$s_{Rest} = 0{,}49218$

$\bar{x}_\bullet = 196{,}143$

$\sum_{i=1}^{n} x_i^2 = 1128894$

Tee x_i	Kaffee y_i	$\widehat{E(\text{Kaffee})}$ $\hat{y}_i$	$s_{\hat{Y}_i}$	$t_{1-\alpha/2;FG_{Rest}} * s_{\hat{Y}_i}$	$\hat{y}_i - t * s_{\hat{Y}_i}$	$\hat{y}_i + t * s_{\hat{Y}_i}$
127	3.71	3.82	0,098	0,2023	3,618	4,022
129	3.64	3.86	0,098	0,2016	3,658	4,062
134	4.06	3.95	0,097	0,2002	3,750	4,150
139	3.72	4.05	0,097	0,1988	3,851	4,249
143	3.95	4.12	0,096	0,1978	3,922	4,318
148	4.07	4.22	0,096	0,1967	4,023	4,417
154	4.32	4.33	0,095	0,1954	4,135	4,525
164	4.56	4.52	0,094	0,1937	4,326	4,714
169	4.63	4.62	0,094	0,1930	4,427	4,813
171	4.80	4.65	0,094	0,1927	4,457	4,843
173	4.47	4.69	0,094	0,1925	4,497	4,883
177	6.10	4.77	0,093	0,1921	4,578	4,962
186	5.73	4.94	0,093	0,1915	4,749	5,131
196	5.04	5.13	0,093	0,1912	4,939	5,321
202	6.23	5.24	0,093	0,1913	5,049	5,431
204	4.62	5.28	0,093	0,1914	5,089	5,471
204	5.25	5.28	0,093	0,1914	5,089	5,471
226	6.70	5.70	0,094	0,1933	5,507	5,893
235	6.20	5.87	0,095	0,1948	5,675	6,065
239	5.59	5.95	0,095	0,1955	5,754	6,146
240	6.40	5.97	0,095	0,1957	5,774	6,166
240	6.60	5.97	0,095	0,1957	5,774	6,166
245	5.77	6.06	0,096	0,1968	5,863	6,257
245	6.09	6.06	0,096	0,1968	5,863	6,257
246	6.20	6.08	0,096	0,1970	5,883	6,277
248	5.80	6.12	0,096	0,1975	5,922	6,318
249	5.76	6.14	0,096	0,1978	5,942	6,338
259	5.89	6.33	0,097	0,2004	6,130	6,530

Die beiden letzten Tabellenspalten beinhalten die Grenzen der 0,95-Konfidenzintervalle für die Erwartungswerte. Will man ein Intervall konstruieren, in dem mit einer Wahrscheinlichkeit von 1-α ein einzelner Wert liegt, dann muß dieses Intervall größer sein, als das für die Erwartungswerte.

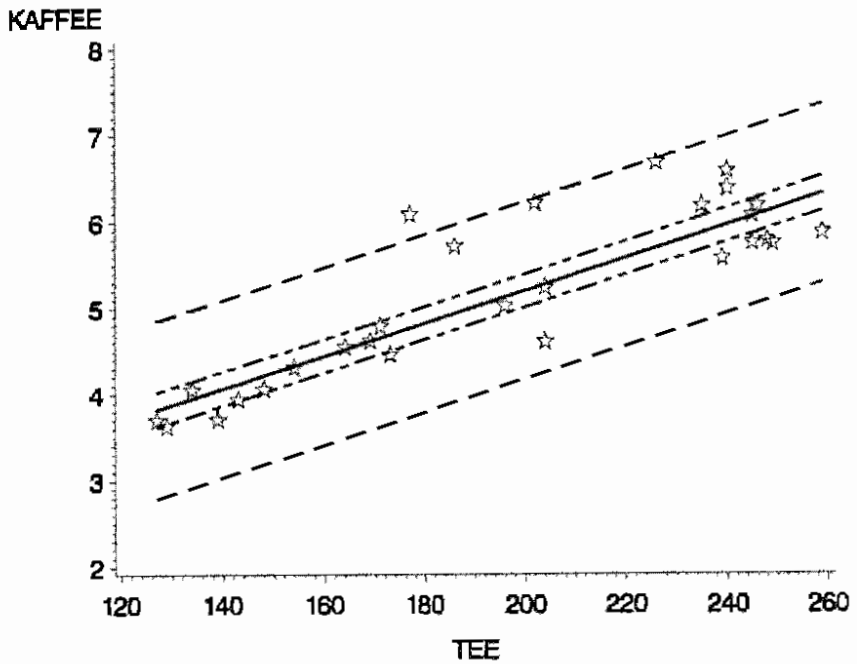
Dieses (1-α)-Vertrauensintervall basiert auf der Vergrößerung der Variabilität:

$$\left\langle \hat{y}_i - t_{1-\alpha/2;FG_{Rest}} * s_{Y_i} \;;\; \hat{y}_i + t_{1-\alpha/2;FG_{Rest}} * s_{Y_i} \right\rangle \qquad \text{mit} \qquad s_{Y_i} = s_{Rest} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x}_\bullet)^2}{\sum_{i=1}^{n} x_i^2}}$$

Die Berechnung der Intervallgrenzen des 0,95-Vertrauensintervalls ist nachstehend aufgeführt.

Tee x_i	Kaffee y_i	$\widehat{E(Kaffee)}$ $\hat{y}_i$	s_{Y_i}	$t_{1-\alpha/2;FG_{Rest}} * s_{Y_i}$	$\hat{y}_i - t * s_{Y_i}$	$\hat{y}_i + t * s_{Y_i}$
127	3,71	3,82	0,502	1,0319	2,788	4,852
129	3,64	3,86	0,502	1,0318	2,828	4,892
134	4,06	3,95	0,502	1,0315	2,918	4,982
139	3,72	4,05	0,502	1,0313	3,019	5,081
143	3,95	4,12	0,501	1,0311	3,089	5,151
148	4,07	4,22	0,501	1,0309	3,189	5,251
154	4,32	4,33	0,501	1,0306	3,299	5,361
164	4,56	4,52	0,501	1,0303	3,490	5,550
169	4,63	4,62	0,501	1,0302	3,590	5,650
171	4,8	4,65	0,501	1,0301	3,620	5,680
173	4,47	4,69	0,501	1,0301	3,660	5,720
177	6,1	4,77	0,501	1,0300	3,740	5,800
186	5,73	4,94	0,501	1,0299	3,910	5,970
196	5,04	5,13	0,501	1,0298	4,100	6,160
202	6,23	5,24	0,501	1,0298	4,210	6,270
204	4,62	5,28	0,501	1,0299	4,250	6,310
204	5,25	5,28	0,501	1,0299	4,250	6,310
226	6,7	5,7	0,501	1,0302	4,670	6,730
235	6,2	5,87	0,501	1,0305	4,840	6,900
239	5,59	5,95	0,501	1,0306	4,919	6,981
240	6,4	5,97	0,501	1,0307	4,939	7,001
240	6,6	5,97	0,501	1,0307	4,939	7,001
245	5,77	6,06	0,501	1,0309	5,029	7,091
245	6,09	6,06	0,501	1,0309	5,029	7,091
246	6,2	6,08	0,501	1,0309	5,049	7,111
248	5,8	6,12	0,501	1,0310	5,089	7,151
249	5,76	6,14	0,501	1,0311	5,109	7,171
259	5,89	6,33	0,502	1,0316	5,298	7,362

Die Grafik zeigt eindrucksvoll die Lage der 0,95-Konfidenzintervalle und der 0,95-Vertrauensintervalle.



Nur ein Wert von 28 liegt außerhalb des Vertrauensintervalls. Bei zwei Werten außerhalb dieses Intervalls wären die vereinbarten 5% bereits überschritten.

Die Realisierung in SAS erfolgt so, daß die Modellanweisung um die Optionen `clm` {Konfidenzintervalle für die Mittelwerte} für die 0,95-Konfidenzintervalle und/oder `cli` {Konfidenzintervalle für die Einzelwerte} für die 0,95-Vertrauensintervalle erweitert werden. In der Ausgabe kommen dann allerdings noch zusätzliche Informationen vor.

```
model kaffee = tee / clm cli;
```

Dem bereits bekannten Teil (s. o.) folgt in der Ausgabe:

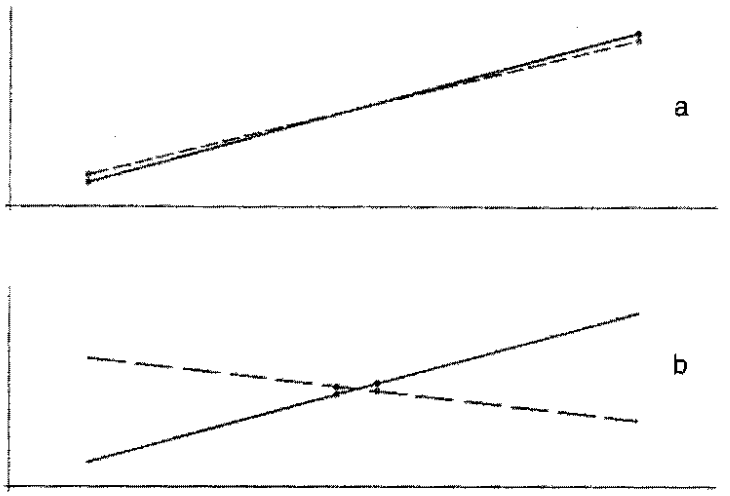
	Kaffee			0,95-Konfidenzintervall		0,95-Vertrauensintervall		
				untere Gr.	obere Gr.	untere Gr.	obere Gr.	
Obs	Dep Var KAFFEE	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict	Residual
1	3.7200	4.1016	0.155	3.7834	4.4198	3.0411	5.1621	-0.3816
2	3.7100	3.8687	0.176	3.5064	4.2310	2.7941	4.9433	-0.1587
3	3.6400	3.9075	0.173	3.5528	4.2622	2.8354	4.9796	-0.2675
4	3.9500	4.1792	0.148	3.8751	4.4834	3.1228	5.2356	-0.2292
5	4.0700	4.2763	0.140	3.9891	4.5634	3.2246	5.3279	-0.2063
6	4.0600	4.0045	0.164	3.6683	4.3408	2.9385	5.0706	0.0555
7	4.3200	4.3927	0.130	4.1249	4.6606	3.3462	5.4393	-0.0727
8	4.5600	4.5868	0.116	4.3481	4.8256	3.5473	5.6263	-0.0268
9	4.4700	4.7615	0.106	4.5444	4.9787	3.7268	5.7963	-0.2915
10	4.6300	4.6839	0.110	4.4577	4.9100	3.6472	5.7205	-0.0539
11	4.8000	4.7227	0.108	4.5012	4.9442	3.6870	5.7584	0.0773
12	5.0400	5.2079	0.093	5.0167	5.3991	4.1783	6.2375	-0.1679
13	4.6200	5.3632	0.095	5.1689	5.5576	4.3330	6.3934	-0.7432
14	5.2500	5.3632	0.095	5.1689	5.5576	4.3330	6.3934	-0.1132
15	5.5900	6.0426	0.131	5.7725	6.3126	4.9954	7.0897	-0.4526
16	5.8000	6.2172	0.146	5.9175	6.5169	5.1621	7.2724	-0.4172
17	5.8900	6.4307	0.165	6.0919	6.7696	5.3638	7.4977	-0.5407
18	5.7700	6.1590	0.141	5.8695	6.4485	5.1067	7.2113	-0.3890
19	5.7600	6.2366	0.147	5.9335	6.5398	5.1805	7.2928	-0.4766
20	6.0900	6.1590	0.141	5.8695	6.4485	5.1067	7.2113	-0.0690
21	6.2000	6.1784	0.142	5.8855	6.4713	5.1252	7.2317	0.0216
22	6.2000	5.9649	0.125	5.7071	6.2227	4.9209	7.0089	0.2351
23	6.4000	6.0620	0.133	5.7887	6.3352	5.0140	7.1099	0.3380
24	6.6000	6.0620	0.133	5.7887	6.3352	5.0140	7.1099	0.5380
25	6.7000	5.7902	0.113	5.5574	6.0231	4.7521	6.8284	0.9098
26	5.7300	5.0138	0.096	4.8174	5.2103	3.9833	6.0444	0.7162
27	6.2300	5.3244	0.094	5.1314	5.5174	4.2945	6.3543	0.9056
28	6.1000	4.8392	0.102	4.6298	5.0485	3.8060	5.8723	1.2608

Sum of Residuals	0
Sum of Squared Residuals	6.2983
Predicted Resid SS (Press)	7.0785

15.4 Zur Versuchsplanung für das einfache lineare Regressionsmodell

15.4.1 Versuchsplanung für das Modell I

Für die Versuchsplanung ist die lineare Regression eigentlich eine ganz „einfache" Angelegenheit, wenn die Meßstellen einstellbar sind (Modell I). Man mißt an den Randpunkten des Versuchsbereichs $[x_{min}\ ,\ x_{max}]$ und hat bereits aufgrund des vorausgesetzten linearen Zusammenhangs das Regressionsmodell mit Hilfe der Zweipunktegleichung. Wenn diese Werte (Schätzwerte) nicht weit von den erwarteten Werten in der Grundgesamtheit entfernt sind, sind nur sehr geringe Abweichungen zwischen der geschätzten und der tatsächlichen Regressionsfunktion zu verzeichnen. Die nachstehende Skizze (a), in der die gestrichelte Gerade die geschätzte Funktion ist, soll das veranschaulichen.
Anders sieht es aus, wenn die Meßpunkte nahe beieinander liegen (b). Bei gleicher Abweichung der Schätzwerte von den Erwartungswerten wie in (a) sind falsche Schlußfolgerungen unausweichlich.



Deshalb und um die Annahme der Linearität zu bestätigen erfolgt bei der Versuchsplanung zum Modell I der Regressionsanalyse neben der Stichprobenumfangsplanung häufig auch eine Planung der Meßpunkte innerhalb des Versuchsbereiches. Dieser Teil ist natürlich beim Modell II der Regressionsanalyse gegenstandslos, weil der Regressor zufällig ist. Die numerische Berechnung der Regressionskoeffizienten ist für beide Modelle gleich. Da aber die Verteilungen der Schätzungen für die Regressionskoeffizienten zwischen Modell I und Modell II und auch die erwarteten Längen der Konfidenzintervalle beider Regressionsmodelle verschieden sind, hat das Auswirkungen auf die Versuchsplanung.

Ein Versuchsplan läßt sich beschreiben durch die Meßstellen $x_1, x_2, \ldots, x_m$ und die Anzahl der an diesen Meßstellen durchzuführenden Messungen n_i ($i = 1, 2, \ldots, m$). Der Versuchsumfang ist folglich $N = \sum_{i=1}^{m} n_i$. Die Kurzform, um solch einen Versuchsplan V zu beschreiben, ist mit $m \geq 2$

$$V = \begin{pmatrix} x_1, x_2, \ldots, x_m \\ n_1, n_2, \ldots, n_m \end{pmatrix} \quad .$$

Optimal bei einem linearen Zusammenhang ist der Versuchsplan, bei dem an den Randpunkten des Versuchsbereiches mit jeweils dem halben Stichprobenumfang gemessen wird (s. o.):

N gerade: $$V_{optimal} = \begin{pmatrix} x_{min}, x_{max} \\ N/2, N/2 \end{pmatrix}$$

N ungerade: $$V_{optimal} = \begin{pmatrix} x_{min}, x_{max} \\ \frac{N-1}{2}, \frac{N+1}{2} \end{pmatrix} \quad \text{bzw.} \quad V_{optimal} = \begin{pmatrix} x_{min}, x_{max} \\ \frac{N+1}{2}, \frac{N-1}{2} \end{pmatrix} \quad \text{(Zweipunktpläne)}$$

oder $$V_{optimal} = \begin{pmatrix} x_{min}, & \frac{x_{max} + x_{min}}{2}, & x_{max} \\ \frac{N-1}{2}, & 1, & \frac{N-1}{2} \end{pmatrix} \quad \text{(Dreipunktplan)} \quad .$$

Diese optimalen Pläne werden in der Praxis jedoch selten eingesetzt, weil die Voraussetzung der strengen Linearität - wie obige Skizze zeigt - wesentlich ist. Im allgemeinen wird der Versuchsbereich in $m - 1$ gleichgroße Abschnitte geteilt werden, so daß m ($m > 2$) Meßstellen x_i ($i = 1, 2, \ldots, m$) in äquidistantem Abstand voneinander entstehen. Wenn nicht weitere Anforderungen wie beispielsweise Minimierung der Kosten oder dgl. auf den Versuchsplan Einfluß

haben, wird im allgemeinen der Versuchsumfang N zu gleichen Teilen n auf die m Meßpunkte aufgeteilt. Der Versuchsplan wäre dann

$$V = \begin{pmatrix} x_1, x_2, \ldots, x_m \\ n, n, \ldots, n \end{pmatrix} \quad \text{mit } N = m * n .$$

Für die Parameter des Regressionsmodells können die (1-α)-Konfidenzintervalle

für β_0: $\left(\underline{b}_0 - \underline{s}_{b_0}\, t_{1-\alpha/2;N-2} \,;\, \underline{b}_0 + \underline{s}_{b_0}\, t_{1-\alpha/2;N-2}\right)$

für β_1: $\left(\underline{b}_1 - \underline{s}_{b_1}\, t_{1-\alpha/2;N-2} \,;\, \underline{b}_1 + \underline{s}_{b_1}\, t_{1-\alpha/2;N-2}\right)$

betrachtet werden. Der Versuchsumfang N hängt von der halben erwarteten Breite des Konfidenzintervalls für β_1 d ab, die als Genauigkeitsforderung vorzugeben ist. Für den optimalen Versuchsplan (s. o.) ist der Versuchsumfang N näherungsweise (RASCH 1983)[1]

$$N = \frac{4\sigma^2}{d^2 (x_{max} - x_{min})^2}\, t^2_{1-\alpha/2;N-2} \quad .$$

Für m Meßstellen mit je n Messungen (N = m * n) kann die Anzahl der Beobachtungen nach RASCH (1983) berechnet werden:

$$n = \frac{\sigma^2}{d^2\, SQ^*_x}\, t^2_{1-\alpha/2;m*n-2} \quad \text{mit} \quad SQ^*_x = \sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2 \quad .$$

Zur Schätzung der Anzahl n der wiederholten Beobachtungen je Meßpunkt kann die Varianz σ^2 durch den Schätzwert s^2 ersetzt werden. Bei der iterativen Berechnung ist zu beachten, daß n ganzzahlig sein muß, was durch die Funktion *KleinsteGanzeZahl* erreicht wird:

$$\textit{KleinsteGanzeZahl}(x) = \lceil x \rceil = \begin{cases} x & \text{wenn x ganzzahlig} \\ 1 + \text{ganzzahliger Anteil von x} & \text{sonst} \end{cases}$$

In der Formel zur Berechnung des minimal notwendigen Stichprobenumfangs n ist auch erkennbar, daß eine relative, auf die Variabilität bezogene Genauigkeitsvorgabe $c = \frac{d}{\sigma}$ bzw. $c = \frac{d}{s}$ möglich ist.

Beispiel

Bezogen wird sich auf ein Beispielsdatensatz von GRIMM und RECKNAGEL[2] zur Abhängigkeit der Extinktion von Androsteron von seiner Konzentration. Die Konzentration wurde in Milligramm mit einem Pulfrich-Photometer gemessen. Die Autoren betonen, daß es sich um ein stark vereinfachtes Beispiel handelt. Die Konzentration ist die einstellbare Variable: die x_i sind also fix (Modell I der Regressionsanalyse).

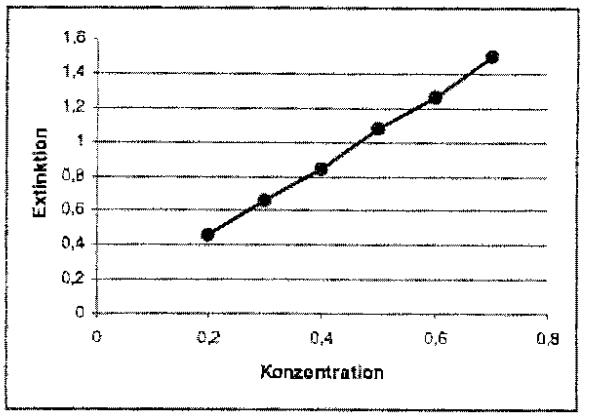
Die Daten sind:

Konzentration	Extinktion von Androsteron
0,2	0,46
0,3	0,66
0,4	0,84
0,5	1,08
0,6	1,26
0,7	1,50

[1] RASCH, D.: Biometrie. Einführung in die Biostatistik. Deutscher Landwirtschaftsverlag Berlin, 1983

[2] GRIMM, H. und RECKNAGEL, R.-D.: Grundkurs Biostatistik, Gustav Fischer Verlag, Jena, 1985, S. 29

Die Linearität soll nur anhand der Grafik gezeigt werden.



Unter der Annahme, daß dieser Versuch für die Schätzung der Variabilität repräsentativ ist (bei 6 Wertepaaren [!?]) soll von diesen Daten ausgehend ein neuer Versuch geplant werden, für
a) den optimalen Versuchsplan
b) die Konzentrationen 0,2 , 0,5 und 0,7.

optimaler Versuchsplan (a)

Die Linearität ist durch den (repräsentativen) Versuch gezeigt für den Bereich von x_{min} = 0,2 mg bis x_{max}=0,7 mg. Benötigt wird ein Schätzwert für σ^2. Die Auswertung des Versuches liefert:

```
data ra1;
  input konz exti;
lines;
0.2       0.46
0.3       0.66
0.4       0.84
0.5       1.08
0.6       1.26
0.7       1.50
;
proc reg;
  model exti = konz;
run;
quit;
```

```
Model: MODEL1
Dependent Variable: EXTI

Analysis of Variance

                 Sum of      Mean
Source    DF    Squares    Square    F Value   Prob>F

Model      1    0.74882   0.74882   2285.651   0.0001
Error      4    0.00131   0.00033
C Total    5    0.75013

  Root MSE      0.01810     R-square    0.9983
  Dep Mean      0.96667     Adj R-sq    0.9978
  C.V.          1.87244

Parameter Estimates

            Parameter    Standard   T for H0:
Variable DF  Estimate       Error  Parameter=0  Prob > |T|

INTERCEP  1  0.035810  0.02082559        1.719      0.1606
KONZ      1  2.068571  0.04326787       47.808      0.0001
```

Der Schätzwert für die (Rest-)Varianz ist 0,0003.

Das Konfidenzniveau wird mit α = 0,05 festgelegt. Die halbe erwartete Breite d wird in Einheiten der Standardabweichung angegeben, so daß $c = \frac{d}{s} = 3{,}5$.

Da beiderseitig der Gleichung $N \approx \frac{4\,\sigma^2\, t^2_{1-\alpha/2;N-2}}{d^2\,(x_{max}-x_{min})^2} = \frac{4\;\; t^2_{1-\alpha/2;N-2}}{c^2\,(x_{max}-x_{min})^2}$ der Gesamt-Stichprobenumfang N steht, kann sie nur iterativ gelöst werden.

Im ersten Schritt wird mit N = ∞ begonnen und in den nachfolgenden das Ergebnis der vorangegangenen Rechnung eingesetzt – bis Konvergenz erkennbar wird:

N	$t_{1-\alpha/2;\,N-2}$	$N_{berechnet}$	*KleinsteGanzeZahl*($N_{berechnet}$)
$N_0 = \infty$	1,960	$N \approx \frac{4 * 1{,}960^2}{3{,}5^2\,(0{,}7-0{,}2)^2} = 5{,}018$	$N_1 = 6$
$N_1 = 6$	2,776	$N \approx \frac{4 * 2{,}776^2}{3{,}5^2\,(0{,}7-0{,}2)^2} = 10{,}065$	$N_2 = 11$
$N_2 = 11$	2,262	$N \approx \frac{4 * 2{,}262^2}{3{,}5^2\,(0{,}7-0{,}2)^2} = 6{,}683$	$N_3 = 7$
$N_3 = 7$	2,571	$N \approx \frac{4 * 2{,}571^2}{3{,}5^2\,(0{,}7-0{,}2)^2} = 8{,}634$	$N_4 = 9$
$N_4 = 9$	2,365	$N \approx \frac{4 * 2{,}365^2}{3{,}5^2\,(0{,}7-0{,}2)^2} = 7{,}305$	$N_5 = 8$
$N_5 = 8$	2,447	$N \approx \frac{4 * 2{,}447^2}{3{,}5^2\,(0{,}7-0{,}2)^2} = 7{,}821$	$N = 8$

Die Lösung lautet N = 8 und damit ist der optimale Versuchsplan

$$V_{optimal} = \begin{pmatrix} x_{min}, x_{max} \\ N/2, N/2 \end{pmatrix} = \begin{pmatrix} 0{,}2 & 0{,}7 \\ 4 & 4 \end{pmatrix}.$$

Versuchsplan für drei Meßstellen (b)

$m = 3 \qquad c = \frac{d}{s} = 3{,}5 \qquad \alpha = 0{,}05$

Zu berechnen ist SQ_x^* :

$x_i = 0{,}2;\ 0{,}5;\ 0{,}7 \qquad \sum_{i=1}^{m} x_i = 1{,}4 \qquad \sum_{i=1}^{m} x_i^2 = 0{,}78 \qquad SQ_x^* = 0{,}78 - \frac{1}{3}(1{,}4)^2 = 0{,}1267$

Der minimal notwendige Stichprobenumfang $n \approx \frac{\sigma^2\, t^2_{1-\alpha/2;\,m*n-2}}{d^2\, SQ_x^*} = \frac{t^2_{1-\alpha/2;\,m*n-2}}{c^2\, SQ_x^*}$ muß ebenfalls iterativ berechnet werden:

n	m*n-2	$t_{1-\alpha/2;\,m*n-2}$	$n_{berechnet}$	*KleinsteGanzeZahl*($n_{berechnet}$)
$n_0 = \infty$	∞	1,960	$n \approx \frac{1{,}960^2}{3{,}5^2 * 0{,}1267} = 2{,}475$	$n_1 = 3$
$n_1 = 3$	4	2,776	$n \approx \frac{2{,}776^2}{3{,}5^2 * 0{,}1267} = 4{,}965$	$n_2 = 5$
$n_2 = 5$	13	2,160	$n \approx \frac{2{,}776^2}{3{,}5^2 * 0{,}1267} = 3{,}006$	$n_3 = 4$
$n_3 = 4$	10	2,228	$n \approx \frac{2{,}228^2}{3{,}5^2 * 0{,}1267} = 3{,}198$	$n = 4$

Das Ergebnis ist zwar wieder 4, aber pro Meßstelle. Der Versuchsplan lautet nunmehr:

$$V = \begin{pmatrix} 0{,}2 & 0{,}5 & 0{,}7 \\ 4 & 4 & 4 \end{pmatrix}.$$

15.4.2 Versuchsplanung für das Modell II

Beim Regressionsmodell II sind Regressor und Regressand Zufallsvariable. Die Frage nach festen Meßpunkten (Regressor) steht nicht mehr. Die Berechnung der Konfidenzintervalle und Testgrößen erfolgt für beide Modelle nach den gleichen Formeln. Allerdings unterscheiden sich die erwarteten Breiten der Konfidenzintervalle und bei Tests die Risiken 2. Art.

Wenn d die halbe erwartete Breite des (1-α)-Konfidenzintervalls von β_1 ist, dann ist der Stichprobenumfang n für diese Genauigkeitsvorgabe näherungsweise (RASCH 1983)

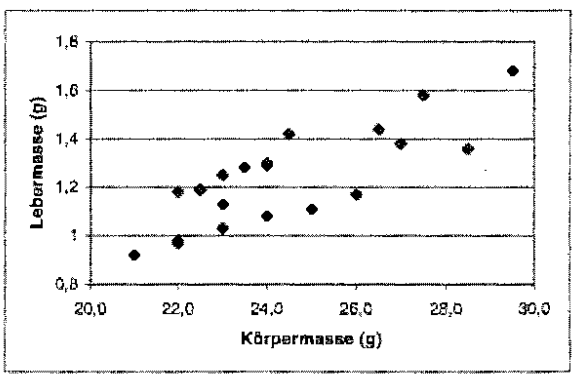
$$n-3 = \frac{\sigma^2}{d^2\,\sigma_x^2}\, t^2_{1-\alpha/2;n-2} \quad .$$

Für die Berechnung sind die Schätzwerte für die Varianz der x-Werte innerhalb des Versuchsbereichs s^2_x und die Varianz um die Regressionsgerade (Fehler- oder Restvarianz) s^2 notwendig.

Beispiel

Den Betrachtungen zu Versuchsplanung wird ein (zielgerichtet) verringerter Datensatz von GRIMM und RECKNAGEL (1985, S. 28) zugrunde gelegt. Gemessen wurden von 25 Mäusen eines bestimmten Stammes die Körpermasse und die Lebermasse in Gramm. Die ausgewählten Daten beider Zufallsvariablen, wobei die Lebermasse der Regressor sein soll, und deren grafische Darstellung sind nachstehend aufgeführt. Von einer linearen Regression kann ausgegangen werden.

Körper-masse	Leber-masse	Körper-masse	Leber-masse
28,5	1,36	23,0	1,13
27,5	1,58	22,0	1,18
24,5	1,42	27,0	1,38
26,5	1,44	29,5	1,68
21,0	0,92	24,0	1,08
24,0	1,30	23,0	1,03
22,0	0,98	26,0	1,17
24,0	1,29	22,5	1,19
22,0	0,97	23,5	1,28
23,0	1,25	25,0	1,11



Vorausgesetzt wird, daß diese Daten die Grundgesamtheit repräsentieren und die daraus gewonnenen Schätzwerte für die Varianzen anstelle der Varianzen der Grundgesamtheit gesetzt werden können.
Das Konfidenzniveau wird festgelegt mit $\alpha = 0{,}05$.

Der Schätzwert für die (Rest-)Varianz σ^2 ist $s^2 = 1{,}95$ und der für die Varianz des Regressors σ_x^2 ist $s_x^2 = 0.04$. Vorzugeben ist noch die halbe erwartete Breite des (1-α)-Konfidenzintervalls für den Regressionsparameter β_1 (Anstieg der Regressionsgeraden) : d = 4.

Die Berechnung des Mindeststichprobenumfangs n muß wieder iterativ erfolgen:

n	$t_{1-\alpha/2;\, n-2}$	$n_{berechnet}$	*KleinsteGanzeZahl*($n_{berechnet}$)
$n_0 = \infty$	1,960	$n \approx \frac{1{,}95 * 1{,}960^2}{4^2 * 0{,}04} + 3 = 14{,}705$	$n_1 = 15$
$n_1 = 15$	2,160	$n \approx \frac{1{,}95 * 2{,}160^2}{4^2 * 0{,}04} + 3 = 17{,}216$	$n_2 = 18$
$n_2 = 18$	2,120	$n \approx \frac{1{,}95 * 2{,}120^2}{4^2 * 0{,}04} + 3 = 16{,}694$	$n_3 = 17$
$n_3 = 17$	2,131	$n \approx \frac{1{,}95 * 2{,}131^2}{4^2 * 0{,}04} + 3 = 16{,}836$	$n = 17$

Unter den getroffenen Annahmen sind in einer neuen Versuchsreihe von (mindestens) 17 Mäusen jeweils Körper- und Lebermasse zu ermitteln.

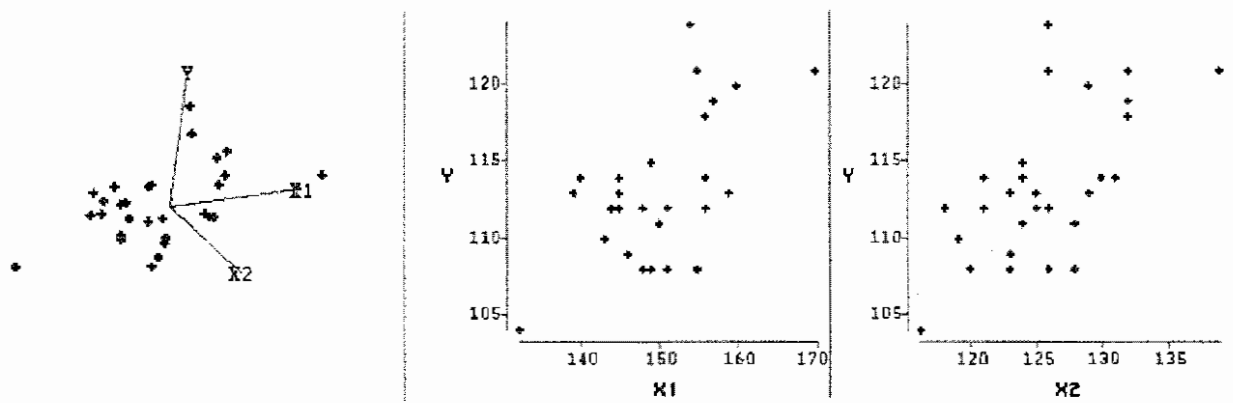
15.5 Berechnen der Statistiken für das multiple lineare Regressionsmodell

Das multiple lineare Regressionsmodell betrachtet den linearen Zusammenhang zwischen einem Regressanden und mehreren Regressoren. Ein Beispiel für das Regressionsmodell II ist:

„Es soll der lineare Zusammenhang bei einjährigen weiblichen Kälbern einer Rinderrasse zwischen den Merkmalen Widerristhöhe (y), Brustumfang (x_1) und Rumpflänge (x_2) untersucht werden." (RASCH 1983)[3] Die Daten (in mm) sind:

y	x_1	x_2	y	x_1	x_2	y	x_1	x_2
110	143	119	113	159	129	108	149	120
112	156	118	112	144	118	113	145	123
108	151	126	114	145	121	109	146	123
108	148	128	113	139	125	104	132	116
112	144	118	115	149	124	112	148	121
111	150	128	121	170	139	124	154	126
114	156	131	120	160	129	118	156	132
112	145	125	114	156	130	112	151	126
121	155	126	108	155	123	121	155	132
119	157	132	111	150	124	114	140	124

Um bei zwei Regressoren die Abhängigkeiten grafisch darzustellen, ist schon eine räumliche Darstellung erforderlich. Bei drei Regressoren ist schon eine vierte Dimension notwendig. Die Darstellung eines Regressors mit dem Regressanden ist die jeweilige Projektion in die Ebene. Die Beispielsdaten sind nachstehend mit Hilfe von SAS/INSIGHT (s. Kapitel 16) grafisch dargestellt.



[3] Rasch, D.: Biometrie. Einführung in die Biostatistik
Deutscher Landwirtschaftsverlag Berlin, 1983, S. 202/203

Die Regressionsfunktion für das Beispiel hat folgende allgemeine Form: $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
Zu schätzen sind das Absolutglied β_0 (Intercept) und die Regressionskoefizienten β_1 und β_2. Um die Lösungen aufzuschreiben, bedient man sich im allgemeinen der Matrizenschreibweise. Das soll hier nicht dargelegt werden. Die Umsetzung mit SAS ist eine Erweiterung der Modell-Anweisung um die weiteren Regressoren. Bei mehreren Regressoren kann es interessant sein zu wissen, wie stark die Abhängigkeit der Zufallsvariablen (!) untereinander ist. Die Berechnung der paarweisen Korrelationskoeffizienten wird durch die Option `corr` angewiesen.

```
data mreg1;
  input y x1 x2 @@;
lines;
110 143 119   113 159 129   108 149 120
112 156 118   112 144 118   113 145 123
108 151 126   114 145 121   109 146 123
108 148 128   113 139 125   104 132 116
112 144 118   115 149 124   112 148 121
111 150 128   121 170 139   124 154 126
114 156 131   120 160 129   118 156 132
112 145 125   114 156 130   112 151 126
121 155 126   108 155 123   121 155 132
119 157 132   111 150 124   114 140 124
;
proc reg corr;
  model y = x1 x2;
run;
quit;
```

```
Correlation

CORR                X1                X2                 Y

X1             1.0000            0.7468            0.5979
X2             0.7468            1.0000            0.6087
Y              0.5979            0.6087            1.0000


Model: MODEL1
Dependent Variable: Y

Analysis of Variance

                          Sum of           Mean
Source           DF      Squares         Square      F Value       Prob>F

Model             2    269.07507      134.53754        9.653       0.0007
Error            27    376.29160       13.93673
C Total          29    645.36667

    Root MSE        3.73319     R-square        0.4169
    Dep Mean      113.43333     Adj R-sq        0.3737
    C.V.            3.29109

Parameter Estimates

                   Parameter       Standard    T for H0:
Variable  DF        Estimate          Error   Parameter=0    Prob > |T|

INTERCEP   1      41.388771    16.76425667         2.469        0.0202
X1         1       0.203718     0.13895180         1.466        0.1542
X2         1       0.330931     0.19935387         1.660        0.1085
```

Die geschätzten Korrelationen zwischen den Zufallsvariablen sind:

$r_{y;x_1} = r_{x_1;y} = 0.60$ $\qquad$ $r_{y;x_2} = r_{x_2;y} = 0.61$ $\qquad$ $r_{x_1;x_2} = r_{x_2;x_1} = 0.75$

Die geschätzte Geradengleichung lautet:

$y = 41{,}4 + 0{,}2 * x_1 + 0{,}3 * x_2$, d.h. Widerristhöhe = 41,4 + 0,2*Brustumfang + 0,3*Rumpflänge

Das Bestimmtheitsmaß von $B = r^2 = 0{,}42$ ist ein multiples Bestimmtheitsmaß:

$$B_{mult.} = B_{y|x_1,x_2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_.)^2} = \frac{\sum_{j=1}^{s}(b_j * SP_{yx_j})}{SQ_y}$$

Die Nullhypothese, daß eine Gerade im Raum angenommen werden kann, ist abzulehnen, wenn eine Irrtumswahrscheinlichkeit $\alpha \geq 0{,}001$ vorgegeben worden wäre. Die Rotation der dreidimensionalen Darstellung (s. o.) zeigt aber, daß die Gerade die beste Anpassung liefern müßte. Um eine Einschätzung vornehmen zu können, werden die Residuen, die Abweichungen der Werte des Regressanden von den geschätzten Werten, berechnet und grafisch dargestellt.

```
data mreg1;
  input y x1 x2 @@;
lines;
110 143 119   113 159 129   108 149 120
112 156 118   112 144 118   113 145 123
108 151 126   114 145 121   109 146 123
108 148 128   113 139 125   104 132 116
112 144 118   115 149 124   112 148 121
111 150 128   121 170 139   124 154 126
114 156 131   120 160 129   118 156 132
112 145 125   114 156 130   112 151 126
121 155 126   108 155 123   121 155 132
119 157 132   111 150 124   114 140 124
;
proc reg corr outest=erg;
  model y = x1 x2;
run; quit;
data _null_;
  set erg (keep=intercep x1 x2);
  call symput ("b0",intercep);
  call symput ("b1",x1);
  call symput ("b2",x2);
proc means data=mreg1 noprint;
  var x1 x2;
  output out=minmax min=mix1 mix2 max=max1 max2;
run;
data _null_;
  set minmax (keep=mix1 mix2 max1 max2);
  call symput ("mix1",mix1);
  call symput ("mix2",mix2);
  call symput ("max1",max1);
  call symput ("max2",max2);
data schaetz;
  do x1 = &mix1 to &max1;
    do x2 = &mix2 to &max2;
      ydach = &b0 + &b1 * x1 + &b2 * x2;
      output;
    end;
  end;
proc sort data=mreg1;
  by x1 x2;
data zusam;
  merge mreg1 schaetz;
  if y ^= . ;
  diff = y - ydach;
  by x1 x2;
run;

proc print noobs;
run;

goptions ftext=swiss htext=1.4;
symbol c=black h=2 v=star i=none;
proc gplot;
  plot diff * y / vref=0;
run; quit;
```

die Schätzwerte für die Regressionskoeffizienten werden auf die Datei erg gespeichert ...

und als Macro-Variable vereinbart;

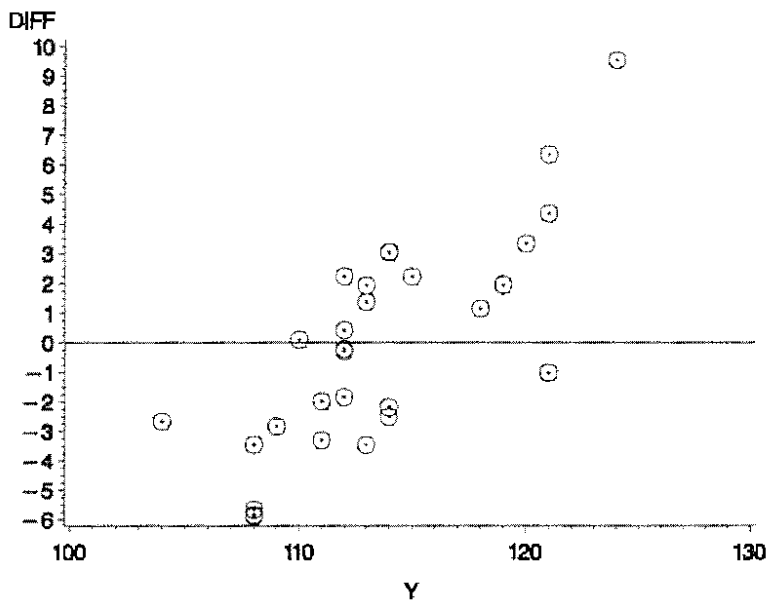
Min und Max der Regressoren-Werte werden gespeichert ...

und als Macro-Variable vereinbart.

berechnen der Schätzwerte der Geraden

berechnen der Differenzen zwischen gemessenen und geschätzten Werten

Y	X1	X2	YDACH	DIFF
104	132	116	106.658	-2.66752
113	139	125	111.072	1.92808
114	140	124	110.945	3.05529
110	143	119	109.901	0.09879
112	144	118	109.774	2.22600
112	144	118	109.774	2.22600
114	145	121	110.971	3.02949
113	145	123	111.632	1.36763
112	145	125	112.294	-0.29423
109	146	123	111.836	-2.83609
112	148	121	111.582	0.41834
108	148	128	113.898	-5.89818
108	149	120	111.454	-3.45445
115	149	124	112.778	2.22183
111	150	124	112.982	-1.98189
111	150	128	114.306	-3.30561
108	151	126	113.847	-5.84747
112	151	126	113.847	-1.84747
124	154	126	114.459	9.54137
108	155	123	113.670	-5.66955
121	155	126	114.662	6.33766
121	155	132	116.648	4.35207
112	156	118	112.219	-0.21862
114	156	130	116.190	-2.18978
114	156	131	116.521	-2.52072
118	156	132	116.852	1.14835
119	157	132	117.055	1.94464
113	159	129	116.470	-3.47001
120	160	129	116.674	3.32627
121	170	139	122.020	-1.02021



Bei einer guten Anpassung an eine Gerade müßten die Abweichungen zufällig um die Null-Linie schwanken.

In den Daten des Beispiels ist aber eine systematische Abweichung erkennbar: kleinere beobachtete y-Werte (Widerristhöhe) werden mit größerem Wert geschätzt als beobachtet, größere y-Werte werden mit kleinerem Wert geschätzt als beobachtet.

Des weiteren scheint das Datenmaterial geschichtet zu sein (beispielsweise könnten die Kälber aus unterschiedlichen Herkünften sein ...): eine Trennung ist bei 116/117 mm Widerristhöhe erkennbar.

In der Literatur wird dazu leider keine Erklärung gegeben. Zu überlegen ist aber, ob die Annahme der Linearität des Regressionsmodells für die gesamte Datenmenge des Beispiels aufrecht erhalten werden kann.

15.6 Variablenselektion im multiplen linearen Regressionsmodell

Häufig fängt man mit vielen Regressor-Variablen an, um zur Beschreibung der Beziehungen der Variablen ein multiples lineare Regressionsmodell aufzustellen. Ein mit Einflußgrößen „überfrachtetes" Modell ist fachlich schwer zu interpretieren. Um wenige aber dafür „effektive" Regressoren im Modell zu belassen, gibt es in der SAS Prozedur REG verschiedene automatisierte Verfahren zur Variablenselektion. Der Option `selection=` können `rsquare`, `forward`, `backward` oder `stepwise` folgen, wodurch diese Verfahren angesprochen werden.

An einem Beispiel sollen diese Verfahren erläutert werden:

Für 35 Datensätze[4] (ERTRAG.DAT) aus vergleichbaren Regierungsbezirken Deutschlands soll für das Jahr 1992 eine möglichst einfache Ertragsfunktion aufgestellt werden. Die Variablen sind:

AWS	Bodenpunkte (Ackerwertzahl)
Ertrag	Winterweizenertrag in dt/ha
Temp	mittlere Temperatur in den Monaten April bis Juni in °C
Regen	mittlere Niederschlagsmenge in den Monaten April bis Juni in mm
Mineral	Mineraldünger-Kosten in DM/ha
PSM	Pflanzenschutzmittel-Kosten in DM/ha

AWS	Ertrag	Temp	Regen	Mineral	PSM
70	75	13.2	44	217	198
71	79	13.6	46	217	202
50	65	14.4	44	217	192
68	70	13.7	49	217	187
75	72	14.0	58	255	199
74	79	14.2	69	255	187
34	49	14.0	53	255	138
57	73	12.8	68	255	187
70	72	14.8	64	241	157
58	72	13.0	71	241	134
53	71	12.6	62	241	182
48	67	13.2	59	231	140
48	61	13.3	73	266	157
49	60	14.0	65	266	193
43	59	14.0	75	266	153
49	64	13.0	60	256	152
59	69	13.7	56	256	143
59	72	15.6	55	208	323
52	58	13.5	71	256	97
68	72	14.0	69	256	210
44	57	13.0	28	174	206
46	56	14.0	33	174	203
45	48	13.0	30	174	207
39	35	14.0	27	136	108
39	36	15.0	20	136	85
73	62	14.5	34	116	163
68	49	14.5	35	116	105
54	61	12.5	48	112	151
37	53	13.0	45	112	174
48	48	14.0	37	113	124
54	52	14.0	30	113	122
41	60	13.0	48	113	127
50	54	16.0	56	70	150
60	76	15.6	47	310	168
54	71	15.6	92	220	204

Die weiteren Optionen im nachstehenden SAS-Programm bedeuten:

`best = n`	nur die n besten Varianten jeder Selektionsstufe werden ausgegeben
`b`	die geschätzten Koeffizienten werden auf jeder Stufe ausgegeben

```
filename f 'ERTRAG.DAT';
data ertrag;
  infile f;
  input awz ertrag temp regen mineral psm;
proc reg data=ertrag;
  ohne:     model ertrag=awz regen mineral temp psm;
  rsquare:  model ertrag=awz regen mineral temp psm
                / selection= rsquare best = 1 b;
  forward:  model ertrag=awz regen mineral temp psm
                / selection= forward best = 1 b;
  backward: model ertrag=awz regen mineral temp psm
                / selection= backward best = 1 b;
  stepwise: model ertrag=awz regen mineral temp psm
                / selection= stepwise best = 1 b;
run; quit;
```

[4] MOLL, E.: Zur Umsetzung biometrischer Verfahren in SAS mit Beispielen aus dem Pflanzenschutz Berichte aus der Biologischen Bundesanstalt für Land- und Forstwirtschaft, Heft 10, 1996, S.120

Die Ausgabe jedes der sechs Modelle wird getrennt vorgenommen.

ohne

```
Model: OHNE
Dependent Variable: ERTRAG

Analysis of Variance

                           Sum of          Mean
Source                DF  Squares        Square     F Value     Prob>F

Model                  5  3608.86758   721.77352      28.182     0.0001
Error                 29   742.73242    25.61146
C Total               34  4351.60000

    Root MSE        5.06078     R-square       0.8293
    Dep Mean       62.20000     Adj R-sq       0.7999
    C.V.            8.13630

Parameter Estimates

                   Parameter       Standard    T for H0:
Variable  DF        Estimate          Error   Parameter=0    Prob > |T|

INTERCEP   1      30.512372    13.93293562         2.190        0.0367
AWZ        1       0.508281     0.08094824         6.279        0.0001
REGEN      1       0.209991     0.06771009         3.101        0.0043
MINERAL    1       0.035776     0.01838077         1.946        0.0614
TEMP       1      -1.910513     0.99202529        -1.926        0.0640
PSM        1       0.074442     0.02119075         3.513        0.0015
```

Diese Ausgabeform ist bekannt. Das multiple Bestimmtheitsmaß ist 0,83. Bei einem vorgegebenen $\alpha = 0{,}05$ ist zu erkennen, daß die Nullhypothese [der Parameter ist Null] für die Variablen `Mineral` und `Temp` nicht abgelehnt werden kann.

rsquare

```
N = 35      Regression Models for Dependent Variable: ERTRAG

                          Parameter
Number in    R-square     Estimates
  Model                   Intercept       AWZ     REGEN  MINERAL      TEMP       PSM

      1   0.47951731       25.3535    0.6763         .        .         .         .
-------------------------------------------------------------------------------------
      2   0.69328421       13.6391    0.5848    0.3210        .         .         .
-------------------------------------------------------------------------------------
      3   0.77725384        6.5467    0.5022    0.2930        .         .    0.0783
-------------------------------------------------------------------------------------
      4   0.80749039        5.4614    0.4836    0.2004   0.0412         .    0.0700
-------------------------------------------------------------------------------------
      5   0.82931969       30.5124    0.5083    0.2100   0.0358   -1.9105    0.0744
-------------------------------------------------------------------------------------
```

Das Kriterium für die Auswahl der Variablen ist das Bestimmtheitsmaß. Auf jeder Selektionsstufe werden für alle Variablenkombinationen die Bestimmtheitsmaße berechnet und der Größe nach geordnet ausgegeben. Mit `best = 1` wird diese Ausgabe auf das jeweils beste Bestimmtheitsmaß begrenzt. Zu erkennen ist, wie sich die Koeffizienten vom Regressionsmodell Ertrag = f(AWS) bis Ertrag = f(AWS, REGEN, MINERAL, TEMP, PSM) ändern. So verbessert die Hinzunahme der Variablen TEMP das Bestimmtheitsmaß um etwas mehr als 0,02.

forward

```
Forward Selection Procedure for Dependent Variable ERTRAG

Step 1   Variable AWZ Entered      R-square = 0.47951731   C(p) = 57.43432730

                 DF        Sum of Squares      Mean Square          F   Prob>F

Regression        1         2086.66752830   2086.66752830      30.40   0.0001
Error            33         2264.93247170     68.63431732
Total            34         4351.60000000
```

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	25.35349664	6.82766893	946.39434660	13.79	0.0008
AWZ	0.67625989	0.12264721	2086.66752830	30.40	0.0001

Bounds on condition number: 1, 1

Step 2 Variable REGEN Entered R-square = 0.69328421 C(p) = 23.11355829

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	3016.89554764	1508.44777382	36.17	0.0001
Error	32	1334.70445236	41.70951414		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	13.63910721	5.87217685	225.01359780	5.39	0.0267
AWZ	0.58477844	0.09755289	1498.77872871	35.93	0.0001
REGEN	0.32095484	0.06796210	930.22801933	22.30	0.0001

Bounds on condition number: 1.041049, 4.164196

Step 3 Variable PSM Entered R-square = 0.77725384 C(p) = 10.84642027

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	3382.29782034	1127.43260678	36.06	0.0001
Error	31	969.30217966	31.26781225		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	6.54673828	5.49129923	44.44237243	1.42	0.2422
AWZ	0.50220103	0.08785026	1021.80167936	32.68	0.0001
REGEN	0.29304464	0.05940713	760.83049926	24.33	0.0001
PSM	0.07833434	0.02291476	365.40227270	11.69	0.0018

Bounds on condition number: 1.126198, 9.926651

Step 4 Variable MINERAL Entered R-square = 0.80749039 C(p) = 7.70897948

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	3513.87519318	878.46879830	31.46	0.0001
Error	30	837.72480682	27.92416023		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	5.46138913	5.21342318	30.64364393	1.10	0.3032
AWZ	0.48362511	0.08346019	937.64626119	33.58	0.0001
REGEN	0.20043885	0.07051116	225.64627488	8.08	0.0080
MINERAL	0.04117447	0.01896826	131.57737284	4.71	0.0380
PSM	0.06998048	0.02199423	282.69389175	10.12	0.0034

Bounds on condition number: 1.762071, 22.92429

Step 5 Variable TEMP Entered R-square = 0.82931969 C(p) = 6.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	3608.86758277	721.77351655	28.18	0.0001
Error	29	742.73241723	25.61146266		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	30.51237249	13.93293562	122.82904522	4.80	0.0367
AWZ	0.50828149	0.08094824	1009.78323986	39.43	0.0001
REGEN	0.20999125	0.06771009	246.33719064	9.62	0.0043
MINERAL	0.03577640	0.01838077	97.02868201	3.79	0.0614
TEMP	-1.91051284	0.99202529	94.99238959	3.71	0.0640
PSM	0.07444155	0.02119075	316.06233884	12.34	0.0015

Bounds on condition number: 1.804021, 34.43631

All variables have been entered into the model.

Summary of Forward Selection Procedure for Dependent Variable ERTRAG

Step	Variable Entered	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	AWZ	1	0.4795	0.4795	57.4343	30.4027	0.0001
2	REGEN	2	0.2138	0.6933	23.1136	22.3025	0.0001
3	PSM	3	0.0840	0.7773	10.8464	11.6862	0.0018
4	MINERAL	4	0.0302	0.8075	7.7090	4.7120	0.0380
5	TEMP	5	0.0218	0.8293	6.0000	3.7090	0.0640

Im ersten Schritt wird vom bezüglich des Bestimmtheitsmaßes besten Modell mit nur einem Regressor ausgegangen. Im ersten Schritt ist das die Variable AWS. In den folgenden Schritten kommt jeweils die Variable hinzu (und bleibt im Modell), die die größte Verbesserung des Bestimmtheitsmaßes liefert. Das geht so lange, bis keine Variable übrig bleibt, die ein bestimmtes Signifikanzniveau unterschreitet. Standardseitig ist dieses Signifikanzniveau für `forward` relativ hoch auf 0,5 gesetzt. Mit der Option `slentry=` kann ein Wert vorgegeben werden.

backward

Backward Elimination Procedure for Dependent Variable ERTRAG

Step 0 All Variables Entered R-square = 0.82931969 C(p) = 6.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	3608.86758277	721.77351655	28.18	0.0001
Error	29	742.73241723	25.61146266		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	30.51237249	13.93293562	122.82904522	4.80	0.0367
AWZ	0.50828149	0.08094824	1009.78323986	39.43	0.0001
REGEN	0.20999125	0.06771009	246.33719064	9.62	0.0043
MINERAL	0.03577640	0.01838077	97.02868201	3.79	0.0614
TEMP	-1.91051284	0.99202529	94.99238959	3.71	0.0640
PSM	0.07444155	0.02119075	316.06233884	12.34	0.0015

Bounds on condition number: 1.804021, 34.43631

All variables left in the model are significant at the 0.1000 level.

Alle Variablen (Regressoren) werden im ersten Schritt einbezogen. Danach werden schrittweise diejenigen Variablen herausgenommen (und bleiben draußen), die am wenigsten zur Verbesserung des Bestimmtheitsmaßes beitragen: die die größte Überschreitungswahrscheinlichkeit (`Prob>F`) aufweisen. Das standardseitig eingestellte Signifikanzniveau für `backward` ist 0,1. Es kann mit der Option `slstay=` verändert werden. Da bereits beim ersten Schritt alle Variablen unterhalb dieser Grenzen liegen, bleibt es nur bei diesem Schritt.

stepwise

Stepwise Procedure for Dependent Variable ERTRAG

Step 1 Variable AWZ Entered R-square = 0.47951731 C(p) = 57.43432730

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	2086.66752830	2086.66752830	30.40	0.0001
Error	33	2264.93247170	68.63431732		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	25.35349664	6.82766893	946.39434660	13.79	0.0008
AWZ	0.67625989	0.12264721	2086.66752830	30.40	0.0001

Bounds on condition number: 1, 1

Step 2 Variable REGEN Entered R-square = 0.69328421 C(p) = 23.11355829

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	3016.89554764	1508.44777382	36.17	0.0001
Error	32	1334.70445236	41.70951414		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	13.63910721	5.87217685	225.01359780	5.39	0.0267
AWZ	0.58477844	0.09755289	1498.77872871	35.93	0.0001
REGEN	0.32095484	0.06796210	930.22801933	22.30	0.0001

Bounds on condition number: 1.041049, 4.164196

Step 3 Variable PSM Entered R-square = 0.77725384 C(p) = 10.84642027

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	3382.29782034	1127.43260678	36.06	0.0001
Error	31	969.30217966	31.26781225		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	6.54673828	5.49129923	44.44237243	1.42	0.2422
AWZ	0.50220103	0.08785026	1021.80167936	32.68	0.0001
REGEN	0.29304464	0.05940713	760.83049926	24.33	0.0001
PSM	0.07833434	0.02291476	365.40227270	11.69	0.0018

Bounds on condition number: 1.126198, 9.926651

Step 4 Variable MINERAL Entered R-square = 0.80749039 C(p) = 7.70897948

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	3513.87519318	878.46879830	31.46	0.0001
Error	30	837.72480682	27.92416023		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	5.46138913	5.21342318	30.64364393	1.10	0.3032
AWZ	0.48362511	0.08346019	937.64626119	33.58	0.0001
REGEN	0.20043885	0.07051116	225.64627488	8.08	0.0080
MINERAL	0.04117447	0.01896826	131.57737284	4.71	0.0380
PSM	0.06998048	0.02199423	282.69389175	10.12	0.0034

Bounds on condition number: 1.762071, 22.92429

Step 5 Variable TEMP Entered R-square = 0.82931969 C(p) = 6.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	3608.86758277	721.77351655	28.18	0.0001
Error	29	742.73241723	25.61146266		
Total	34	4351.60000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	30.51237249	13.93293562	122.82904522	4.80	0.0367
AWZ	0.50828149	0.08094824	1009.78323986	39.43	0.0001
REGEN	0.20999125	0.06771009	246.33719064	9.62	0.0043
MINERAL	0.03577640	0.01838077	97.02868201	3.79	0.0614
TEMP	-1.91051284	0.99202529	94.99238959	3.71	0.0640
PSM	0.07444155	0.02119075	316.06233884	12.34	0.0015

Bounds on condition number: 1.804021, 34.43631

All variables left in the model are significant at the 0.1500 level.
No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable ERTRAG

Step	Variable Entered	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	AWZ		1	0.4795	0.4795	57.4343	30.4027	0.0001
2	REGEN		2	0.2138	0.6933	23.1136	22.3025	0.0001
3	PSM		3	0.0840	0.7773	10.8464	11.6862	0.0018
4	MINERAL		4	0.0302	0.8075	7.7090	4.7120	0.0380
5	TEMP		5	0.0218	0.8293	6.0000	3.7090	0.0640

Das stepwise-Verfahren stellt eine Kombination aus forward und backward dar. Dementsprechend können auch zwei Signifikanzniveaus vorgegeben werden: *significance level entry* (sle) mit `slentry=` und *significance level stay* (sls) mit `slstay=` . Beide Grenzen sind gleich groß mit 0,15 voreingestellt.
Das Verfahren beginnt wie `forward` bis zu zwei Regressoren. In jedem nun folgenden Schritt wird geprüft, ob Variable, die bereits im Modell sind, mit ihrer Überschreitungswahrscheinlichkeit (`Prob>F`) oberhalb der sls-Schranke liegen. Dann werden sie wieder aus dem Modell heraus genommen. Es wird die Variable, die den größten Beitrag zur Varianzerklärung liefert in das Modell aufgenommen, wenn die Überschreitungswahrscheinlichkeit kleiner als die sle-Schranke ist. Das bedeutet, daß eine Variable aus dem Modell herausgenommen werden kann und in einem späteren Schritt wieder aufgenommen wird.

Weiter soll auf diese und andere Verfahren zur Variablenselektion in einem multiplen linearen Regressionsmodell nicht eingegangen werden.

Aufgabe 15.1: Berechnen Sie für die nachstehenden Daten eine geeignete (α=0,05) Regressionsfunktion und die 0,95-Konfidenzintervalle. Gehen Sie dabei von einem Polynom dritten Grades aus.

2.0	0	9.6	211
2.4	2	12.4	376
3.2	10	15.0	572
5.3	49	18.9	941
7.8	130	21.3	1214

15.7 Nichtlineare Regressionsfunktionen

Wachstumsfunktionen z. B. die Gompertzfunktion $y = f(x) = \alpha\, e^{\beta\, e^{\gamma x}}$ sind eigentlich nichtlineare Funktionen.

Eine Funktion wie beispielsweise $y = f(x) = 2{,}8 - 2{,}1*x + 3{,}75*x^2 + 0{,}08*x^3$ läßt sich durch Variablentransformation wieder in eine lineare Funktion überführen:

$x^2 \rightarrow z1 \qquad x^3 \rightarrow z2$

$y = f(x) = 2{,}8 - 2{,}1*x + 3{,}75*z1 + 0{,}08*z2$

Eigentlich nichtlineare Funktionen können nicht wie lineare behandelt werden. Für sie steht in SAS die Prozedur NLIN zur Verfügung. Polynominale und quasilineare Funktionen können sowohl nichtlinear als auch lineare – nach Variablentransformation – betrachtet werden. Anhand der oben aufgeführten Funktion dritten Grades sollen beide Wege demonstriert werden.

Zunächst werden Daten erzeugt, die dem Regressionsmodell II genügen:

```
data a;
  do x = 1, 2.37, 4.8, 6, 8.77, 12, 15.6, 20;
     y = +2.8 - 2.1*x + 3.75*x*x + 0.08*x*x*x;
     output;
  end;
proc print noobs;
run;
```

X	Y
1.00	4.53
2.37	19.95
4.80	87.97
6.00	142.48
8.77	326.77
12.00	655.84
15.60	1186.35
20.00	2100.80

Die y-Werte werden gerundet und sind nun die Ausgangsdaten. Ziel ist es, die gewählten Koeffizienten als Schätzwerte wieder zu erhalten.

```
data daten;
  input x y;
  z1 = x*x;
  z2 = x*x*x;
lines;
 1.00    3
 2.37   20
 4.80   90
 6.00  140
 8.77  327
12.00  655
15.60 1190
20.00 2100
;
proc reg corr;
  model y = x z1 z2 /clm;
run;  quit;
```

Mit den transformierten Variablen wird eine multiple lineare Regressionsanalyse durchgeführt. Gleichzeitig sollen die paarweisen Korrelationskoeffizienten geschätzt werden.

```
Correlation

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

                          Sum of          Mean
Source            DF      Squares        Square      F Value       Prob>F

Model              3 3830687.7695 1276895.9232   231054.876       0.0001
Error              4      22.10550       5.52638
C Total            7 3830709.875
```

```
     Root MSE          2.35082     R-square        1.0000
     Dep Mean        565.62500     Adj R-sq        1.0000
     C.V.              0.41562

 Parameter Estimates

                    Parameter       Standard     T for H0:
 Variable  DF        Estimate          Error    Parameter=0     Prob > |T|

 INTERCEP   1        2.786397     3.00769497          0.926         0.4067
 X          1       -2.587146     1.33745191         -1.934         0.1252
 Z1         1        3.836116     0.15304897         25.065         0.0001
 Z2         1        0.076884     0.00484281         15.876         0.0001


          Dep Var    Predict   Std Err  Lower95%  Upper95%
   Obs          Y      Value   Predict      Mean      Mean  Residual

     1     3.0000     4.1123     2.017   -1.4888    9.7133   -1.1123
     2    20.0000    19.2254     1.317   15.5676   22.8832    0.7746
     3    90.0000    87.2550     1.343   83.5263   90.9837    2.7450
     4      140.0      142.0     1.393     138.1     145.8   -1.9707
     5      327.0      327.0     1.305     323.4     330.6  -0.00449
     6      655.0      657.0     1.498     652.8     661.2   -1.9976
     7     1190.0     1187.9     1.799    1182.9    1192.9    2.1307
     8     2100.0     2100.6     2.318    2094.1    2107.0   -0.5651

 Sum of Residuals                           0
 Sum of Squared Residuals             22.1055
 Predicted Resid SS (Press)          492.7334
```

Die Variablen sind untereinander sehr hoch korreliert. Aufgrund der Konstruktion ist das für x, z1 und z2 klar.
Das multiple Bestimmtheitsmaß wird mit 1 geschätzt, d. h. die Gesamtvarianz wird (fast) völlig durch die Modellvarianz erklärt.
Die geschätzte Modellgleichung lautet: $y = 3{,}26 - 2{,}25*x + 3{,}76*z1 + 0{,}08*z2$.

Das Bestimmtheitsmaß ist nur für die lineare Regression definiert. Für eine lineare Regressionsfunktion bestehend aus transformierten Variablen ist es im transformierten Raum interpretierbar. Diese Maßzahl fehlt dementsprechend bei der nichtlinearen Betrachtung.

```
data daten;
  input x y;
  z1 = x*x;
  z2 = x*x*x;
lines;
 1.00    5
 2.37   20
 4.80   88
 6.00  142
 8.77  327
12.00  656
15.60 1186
20.00 2101
;
proc nlin;
  model y = a + b*x + c*z1 + d*z2;
  parms a = 1 to 5
        b = -1 to -5
        c = 1 to 5
        d = 0.05 to 0.1 by 0.0125
        ;
run;
```

SAS-Prozedur NLIN

die zu schätzende Funktion mit ihren Koeffizienten

aufgrund der iterativen Berechnung müssen Vorgaben für die Parameterschätzung vorgenommen werden!

Non-Linear Least Squares Grid Search Dependent Variable Y

A	B	C	D	Sum of Squares
1.000000	-1.000000	1.000000	0.050000	2588960
2.000000	-1.000000	1.000000	0.050000	2583107
3.000000	-1.000000	1.000000	0.050000	2577270
4.000000	-1.000000	1.000000	0.050000	2571449
5.000000	-1.000000	1.000000	0.050000	2565643
1.000000	-1.000000	2.000000	0.050000	1235851
2.000000	-1.000000	2.000000	0.050000	1231857
3.000000	-1.000000	2.000000	0.050000	1227880
4.000000	-1.000000	2.000000	0.050000	1223919
5.000000	-1.000000	2.000000	0.050000	1219973
1.000000	-1.000000	3.000000	0.050000	378211
2.000000	-1.000000	3.000000	0.050000	376078
3.000000	-1.000000	3.000000	0.050000	373960

• • • alle Parameterkombinationen • • •

4.000000	-1.000000	2.000000	0.100000	435248
5.000000	-1.000000	2.000000	0.100000	432756
1.000000	-1.000000	3.000000	0.100000	28679.211248
2.000000	-1.000000	3.000000	0.100000	27999.679285
3.000000	-1.000000	3.000000	0.100000	27336.147322
4.000000	-1.000000	3.000000	0.100000	26688.615359
5.000000	-1.000000	3.000000	0.100000	26057.083397
1.000000	-1.000000	4.000000	0.100000	110011
2.000000	-1.000000	4.000000	0.100000	111191
3.000000	-1.000000	4.000000	0.100000	112387
4.000000	-1.000000	4.000000	0.100000	113600
5.000000	-1.000000	4.000000	0.100000	114828
1.000000	-1.000000	5.000000	0.100000	686812
2.000000	-1.000000	5.000000	0.100000	689852
3.000000	-1.000000	5.000000	0.100000	692909
4.000000	-1.000000	5.000000	0.100000	695981
5.000000	-1.000000	5.000000	0.100000	699069

Non-Linear Least Squares Iterative Phase Dependent Variable Y Method: Gauss-Newton

Iter	A	B	C	D	Sum of Squares
0	1.000000	-1.000000	4.000000	0.062500	1137.487774
1	2.786397	-2.587146	3.836116	0.076884	22.105501
2	2.786397	-2.587146	3.836116	0.076884	22.105501

NOTE: Convergence criterion met.

Non-Linear Least Squares Summary Statistics Dependent Variable Y

Source	DF	Sum of Squares	Mean Square
Regression	4	6390140.8945	1597535.2236
Residual	4	22.1055	5.5264
Uncorrected Total	8	6390163.0000	
(Corrected Total)	7	3830709.8750	

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval Lower	Upper
A	2.786396951	3.0076949723	-5.5641900436	11.136983945
B	-2.587145908	1.3374519067	-6.3004574640	1.126165648
C	3.836116239	0.1530489704	3.4111899243	4.261042555
D	0.076884396	0.0048428122	0.0634387756	0.090330016

Asymptotic Correlation Matrix

Corr	A	B	C	D
A	1	-0.881658678	0.7721844035	-0.695120779
B	-0.881658678	1	-0.96889826	0.9198271722
C	0.7721844035	-0.96889826	1	-0.986954481
D	-0.695120779	0.9198271722	-0.986954481	1

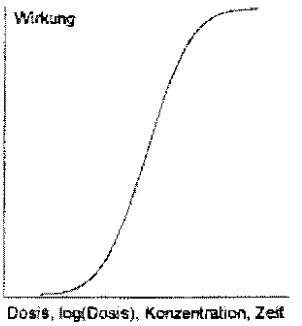
Die Algorithmen zur Berechnung der Koeffizienten einer nichtlinearen Regressionsfunktion unterscheiden sich von denen zur Berechnung der Koeffizienten einer linearen. Im ersten Teil der Ausgabe wird mit Hilfe der vorgegebenen Werte für die zu schätzenden Parameter ein „vernünftiger" Startwert für die anschließenden iterative Berechnung der Lösung gesucht. Die erste „interessante" Ausgabe ist für den Anwender die Varianztabelle. Die Ergebnisse gleichen sich bei linearer und nichtlinearer Betrachtung. Was fehlt (fehlen muß) sind einige Maßzahlen.

Die Schätzwerte für die Koeffizienten stimmen mit dem Ergebnis von PROC REG überein. Zusätzlich werden für alle Koeffizienten die 0,95-Konfidenzintervalle angegeben.

Sowohl für die lineare als auch die nichtlineare Regression gilt, daß die Anzahl der Werte-Tupel mindestens so groß sein muß wie die der zu schätzenden Parameter!

15.8 Bioassay auf der Grundlage von Probit-, Logit und ähnlichen Transformationen

Bioassay (biological Assay) ist ein eigenständiges Gebiet der Biometrie. Es basiert unter anderem auf speziellen nichtlinearen Transformationen oder die Zugrundelegung bestimmter Wachstumsfunktionen wie die logistische oder die Gompertzfunktion. Ausgangspunkt ist ein sigmoider Verlauf der Wirkung über der Dosis, dem Logarithmus der Dosis, der Konzentration oder über der Zeit.

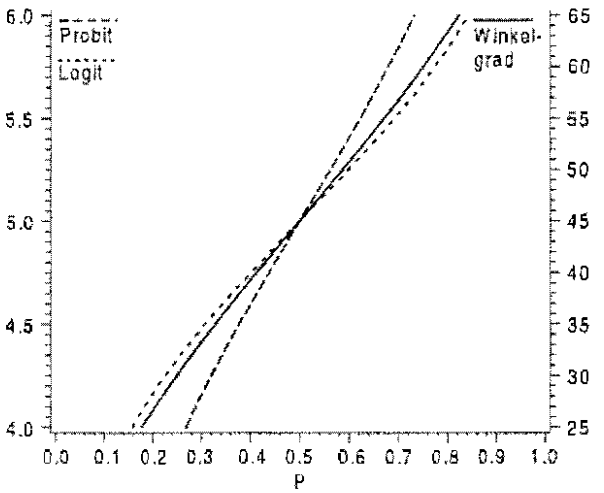


Das Ziel besteht darin, die sigmoide Funktion in eine Gerade zu überführen, um dann mit bekannten Elementen der linearen Regressionsanalyse arbeiten bzw. auf diese aufbauen zu können.

Solche Transformationen sind beispielsweise die Probit-, Logit, Loglog- oder auch Winkelgradtransformation:

Probit Y der Wirkung P (mit $Y = u + 5$)		$P = \frac{1}{2\pi} \int_{-\infty}^{Y-5} e^{-\frac{1}{2}u^2}\, du$
Logit Y der Wirkung P	$Y = \ln\frac{P}{1-P} + 5$	$\frac{P}{1-P} = e^{2(Y-5)}$
Loglog Y der Wirkung P		$P = e^{-e^{-Y}}$
Winkelgrad Y der Wirkung P	$Y = arc\sin\sqrt{P}$	$P = \sin^2 Y$

Im mittleren, dem annähernd linearen Bereich, ähneln sich die Transformationen Probit-, Logit- und Winkelgrad-. Die Wirkung P ist in nebenstehender Grafik als Wahrscheinlichkeit abgetragen. Die linke Ordinatenachse ist die für die Probits und Logits. Erkennbar ist, daß 5 addiert werden, damit die transformierten Werte im positiven Bereich liegen. Die rechte Ordinatenachse ist der Winkelgradtransformation zuzuordnen.

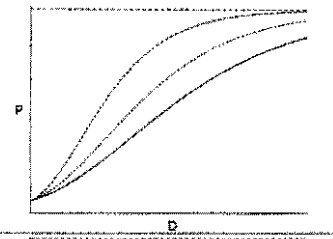


Die bereits angesprochenen speziellen Wachstumsfunktionen sind:

Gompertzfunktion

$$P = \alpha\, e^{\beta\, e^{\gamma D}}$$

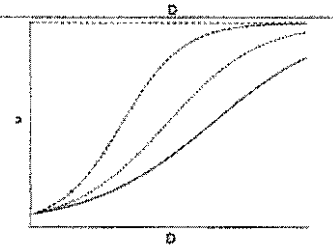
(D: Dosis , P: Wirkung)



Logistische Funktion

$$P = \frac{\alpha}{1 + \beta\, e^{\gamma D}}$$

(D: Dosis , P: Wirkung)



„Per Hand" werden die Transformationen wohl kaum noch realisiert, so daß gleich der Blick auf mögliche Realisierungen in SAS geworfen werden soll. Die Erläuterungen werden anhand eines Beispiels vorgenommen.

Beispiel:

Bei verschiedenen Dosen einer Substanz wurde ausgehend von einem bekannten Stichprobenumfang (Umfang) die Anzahl der Reagenten (z. B. tote Schadorganismen) gezählt. Die Probitgerade und die Erwartungswerte für 20%, 50% und 95% Wirkung sollen ausgegeben werden. Die Daten[5] sind:

PROBIT .DAT

Dosis	Umfang	Reagenten
1	10	1
2	12	2
30	10	4
40	10	5
500	12	8
600	10	8
7000	10	9

SAS-Programm:

```
data daten;
  infile 'probit.dat';
  input dosis umfang reagent;

proc probit data=daten
            log10                  Modellansatz mit dem Logarithmus der Dosis zur Basis 10
                                   [möglich auch LOG oder LN]
            inversecl;             (inverse) 95%-Konfidenzintervalle zur Dosis
  model reagent/umfang = dosis
run;
```

Als Option im Modell-Statement ist zum Beispiel die Angabe eines Verteilungstyps möglich; z.B.

```
  model reagent/umfang = dosis / D=normal
```

Die Normalverteilung (`D=normal`) ist die Standardeinstellung - Probitmodell.
Desweiteren kann mit `D=logistic` die logistische Verteilung - Logitmodell
und mit `D=gompertz` die Gompertzverteilung - Gompitmodell genutzt werden.

[5] MOLL, E.: Zur Umsetzung biometrischer Verfahren in SAS mit Beispielen aus dem Pflanzenschutz Berichte aus der Biologischen Bundesanstalt für Land- und Forstwirtschaft, Heft 10, 1996, S. 138

Im Modell-Statement darf keine Variable beispielsweise

```
Wirk = Reagent/Umfang;
...
   model wirk = dosis;
```

stehen. Die durch einen Schrägstrich getrennte Auflistung von zwei Variablen vor dem Gleichheitszeichen ist zwingend.

Ergebnis:

```
Probit Procedure

Data Set              =WORK.DATEN
Dependent Variable=REAGENT
Dependent Variable=UMFANG
Number of Observations=    7
Number of Events      =       37    Number of Trials =       74

Log Likelihood for NORMAL -38.48566707

Probit Procedure

Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value

INTERCPT   1 -1.1815424 0.305324  14.97534  0.0001 Intercept
Log10(DOS) 1 0.65638946 0.142638  21.17637  0.0001

Probit Model in Terms of Tolerance Distribution

          MU          SIGMA
    1.800063       1.523486

Estimated Covariance Matrix for Tolerance Parameters

                          MU             SIGMA

         MU          0.062566       -0.000336
      SIGMA         -0.000336        0.109604


Probit Procedure
Probit Analysis on Log10(DOSIS)

Probability    Log10(DOSIS) 95 Percent Fiducial Limits
                                    Lower       Upper

    0.01           -1.7441       -4.4549     -0.6062
    0.02           -1.3288       -3.7413     -0.3051
    0.03           -1.0653       -3.2899     -0.1128
    0.04           -0.8671       -2.9512      0.0329
    0.05           -0.7058       -2.6764      0.1520
    0.06           -0.5686       -2.4431      0.2540
    0.07           -0.4483       -2.2390      0.3439
    0.08           -0.3405       -2.0567      0.4248
    0.09           -0.2426       -1.8914      0.4989
    0.10           -0.1524       -1.7396      0.5674
    0.15            0.2211       -1.1161      0.8563
    0.20            0.5179       -0.6287      1.0940
    0.25            0.7725       -0.2192      1.3067
    0.30            1.0011        0.1386      1.5075
    0.35            1.2130        0.4587      1.7051
    0.40            1.4141        0.7488      1.9063
    0.45            1.6086        1.0138      2.1165
    0.50            1.8001        1.2572      2.3409
    0.55            1.9915        1.4824      2.5834
    0.60            2.1860        1.6935      2.8476
    0.65            2.3871        1.8953      3.1370
    0.70            2.5990        2.0935      3.4565
```

Geradengleichung:
Wirkung = -1.1815424 + 0.6564 lg(Dosis)

0.75	2.8276	2.2948	3.8139	
0.80	3.0823	2.5078	4.2230	
0.85	3.3791	2.7458	4.7101	
0.90	3.7525	3.0349	5.3333	
0.91	3.8427	3.1035	5.4851	
0.92	3.9407	3.1776	5.6504	
0.93	4.0484	3.2586	5.8327	
0.94	4.1687	3.3485	6.0367	
0.95	4.3060	3.4506	6.2700	
0.96	4.4672	3.5697	6.5448	
0.97	4.6654	3.7154	6.8834	
0.98	4.9289	3.9078	7.3348	
0.99	5.3442	4.2089	8.0483	

Probit Procedure
Probit Analysis on DOSIS

Probability	DOSIS	95 Percent Fiducial Limits		
		Lower	Upper	
0.01	0.01803	0.0000351	0.24763	
0.02	0.04690	0.0001814	0.49531	
0.03	0.08604	0.0005130	0.77134	
0.04	0.13581	0.00112	1.07859	
0.05	0.19686	0.00211	1.41902	
0.06	0.27001	0.00360	1.79456	
0.07	0.35622	0.00577	2.20732	
0.08	0.45652	0.00878	2.65962	
0.09	0.57206	0.01284	3.15401	
0.10	0.70411	0.01822	3.69330	
0.15	1.66369	0.07655	7.18307	
0.20	3.29507	0.23514	12.41766	20% Wirkung
0.25	5.92226	0.60366	20.26146	
0.30	10.02643	1.37607	32.17255	
0.35	16.33175	2.87535	50.71323	
0.40	25.94731	5.60781	80.58459	
0.45	40.60876	10.32221	130.78169	
0.50	63.10490	18.07842	219.22920	50% Wirkung
0.55	98.06326	30.36613	383.18491	
0.60	153.47363	49.37346	704.00946	
0.65	243.83346	78.58160	1371	
0.70	397.17305	124.03025	2861	
0.75	672.41686	197.15010	6514	
0.80	1209	321.94870	16710	
0.85	2394	556.92948	51297	
0.90	5656	1084	215450	
0.91	6961	1269	305578	
0.92	8723	1505	447135	
0.93	11179	1814	680265	
0.94	14748	2231	1088231	
0.95	20229	2822	1862128	95% Wirkung
0.96	29323	3713	3505678	
0.97	46283	5192	7645581	
0.98	84902	8087	21615712	
0.99	220913	16178	111768439	

95%-Konfidenzintervalle um die Wirkung

Es sollen zusätzlich zu den (inversen) Konfidenzintervallen um die Dosis (bzw. allgemein: Abszisse) die "normalen" 95%-Konfidenzintervalle um die Wirkung berechnet werden:

```
data daten;
  infile 'probit.dat';
  input dosis umfang reagent;

proc probit data=daten
            log10
            inversecl;
  model reagent/umfang = dosis;
  output out   =  ERGEBNIS
         p     =  prob
         std   =  std
         xbeta=   xbeta;

data konfi;
  set ergebnis;
  ug = xbeta - 1.96*std;
  og = xbeta + 1.96*std;

proc print noobs;
run;
```

Ausgabe auf Datei ERGEBNIS
p: erwartete Wahrscheinlichkeit
std: Standardabweichung
xbeta: $\alpha + \beta * \text{dosis}$ (bzw. log(dosis))

untere Grenze des Konfidenzintervalls
obere Grenze, wobei approximativ das 95%-Quantil der Normalverteilung verwendet wird

```
Probit Procedure

Data Set              =WORK.DATEN
Dependent Variable=REAGENT
Dependent Variable=UMFANG
Number of Observations=    7
Number of Events      =       37    Number of Trials =       74


Log Likelihood for NORMAL -38.48566707


Probit Procedure

Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value

INTERCPT   1 -1.1815424 0.305324  14.97534  0.0001 Intercept
Log10(DOS) 1 0.65638946 0.142638  21.17637  0.0001

Probit Model in Terms of Tolerance Distribution

          MU         SIGMA
    1.800063      1.523486


Estimated Covariance Matrix for Tolerance Parameters

                            MU             SIGMA

          MU          0.062566         -0.000336
       SIGMA         -0.000336          0.109604


Probit Procedure
Probit Analysis on Log10(DOSIS)

Probability    Log10(DOSIS) 95 Percent Fiducial Limits
                                    Lower       Upper

    0.01            -1.7441       -4.4549     -0.6062
    0.02            -1.3288       -3.7413     -0.3051
```

... siehe oben

```
    0.97             4.6654        3.7154      6.8834
    0.98             4.9289        3.9078      7.3348
    0.99             5.3442        4.2089      8.0483
```

```
Probit Procedure
Probit Analysis on DOSIS

Probability          DOSIS 95 Percent Fiducial Limits
                                    Lower       Upper

    0.01            0.01803      0.0000351      0.24763
    0.02            0.04690      0.0001814      0.49531

  ... siehe oben

    0.99             220913          16178     111768439


DOSIS UMFANG REAGENT    PROB     STD      XBETA      UG        OG

    1    10       1    0.11869 0.30532 -1.18154 -1.77998 -0.58311
    2    12       2    0.16257 0.27011 -0.98395 -1.51337 -0.45453
   30    10       4    0.41606 0.17070 -0.21198 -0.54655  0.12260
   40    10       5    0.44830 0.16671 -0.12997 -0.45672  0.19678
  500    12       8    0.72242 0.20791  0.59003  0.18253  0.99753
  600    10       8    0.73957 0.21502  0.64201  0.22056  1.06345
 7000    10       9    0.91026 0.33415  1.34234  0.68740  1.99728
```

PROB: erwartete Wirkung (PROB)
UG untere und OG obere Grenze des 95%-Konfidenzintervalls um die mittlere Wirkung.

Berücksichtigung der natürlichen Mortalität

Beispiel:

Für die Beispielsdaten liegt in der Kontrollgruppe (Dosis = 0) eine natürliche Mortalität (ein Spontaneffekt) vor, der Stichprobenumfang auf jeder Dosisstufe ist 10:

VERSUCH .DAT

Dosis	Reagenten
0	1
0.1	4
0.2	6
0.3	7
0.4	8
0.5	9
0.6	10

Zunächst wird die Dosisstufe Null wie eine beliebige Dosis betrachtet.

```
data daten;
  n = 10;
  infile 'versuch.dat';
  input dosis r;
proc probit data=daten;
  model r/n=dosis;
run;
```

```
Probit Procedure

Data Set              =WORK.DATEN
Dependent Variable=R
Dependent Variable=N
Number of Observations=    7
Number of Events      =       45    Number of Trials =       70

Log Likelihood for NORMAL -31.94215979

Probit Procedure

Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value

INTERCPT   1 -0.9246883 0.317778  8.467266  0.0036 Intercept
DOSIS      1 4.86284315 1.079592  20.28903  0.0001

Probit Model in Terms of Tolerance Distribution

          MU          SIGMA
    0.190154       0.205641

Estimated Covariance Matrix for Tolerance Parameters

                            MU             SIGMA

          MU          0.001500       -0.000534
       SIGMA         -0.000534        0.002084
```

Achtung:

Im logarithmischen Modell (Option LOG , LOG10 oder LN ; s. o.) wird die Dosisstufe Null nicht berücksichtigt, weil kein Logarithmus gebildet werden kann!

Nun soll die natürliche Mortalität, die Reaktion in der Kontrollgruppe, Berücksichtigung finden.

```
proc probit data=daten
            optc ;
  model r/n=dosis;
run;
```

```
Probit Procedure

Data Set              =WORK.DATEN
Dependent Variable=R
Dependent Variable=N
Number of Observations=    7
Number of Events      =       45    Number of Trials =       70

Log Likelihood for NORMAL -32.27238278

Probit Procedure

Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value

INTERCPT   1 -0.7087181 0.474883  2.227275  0.1356 Intercept
DOSIS      1 4.52333382 1.172551  14.88174  0.0001
C          1          0 0.252757                 Lower threshold

Probit Model in Terms of Tolerance Distribution

          MU          SIGMA
     0.15668       0.221075

Estimated Covariance Matrix for Tolerance Parameters
```

	MU	SIGMA	C
MU	0.005880	-0.002464	0.015911
SIGMA	-0.002464	0.003284	-0.005975
C	0.015911	-0.005975	0.063886

Bemerkung:

Mit der zusätzlichen Option OPTC wird auch im logarithmischen Modell die Dosisstufe Null berücksichtigt.
Die natürliche Mortalität (Kontrollrate, Spontanrate, Spontaneffekt) wird bei allen Dosisstufen als additive Größe im Modell berücksichtigt. Die Kurve liegt als etwas höher als ohne Beachtung der Kontrollgruppe. Die Geradengleichungen lauten:

Wirkung = -0.925 + 4.863 Dosis ohne Berücksichtigung der natürlichen Mortalität
Wirkung = -0.709 + 4.523 Dosis mit Berücksichtigung der natürlichen Mortalität.

Ist die natürliche Mortalität ein feststehender Wert, kann sie auch als Option (c =) vorgegeben werden:

```
data daten;
  n = 10;
  infile 'versuch.dat';
  input dosis r;

proc probit data=daten
            (where=(dosis>0))

            c=0.1;
  model r/n=dosis;
run;
```

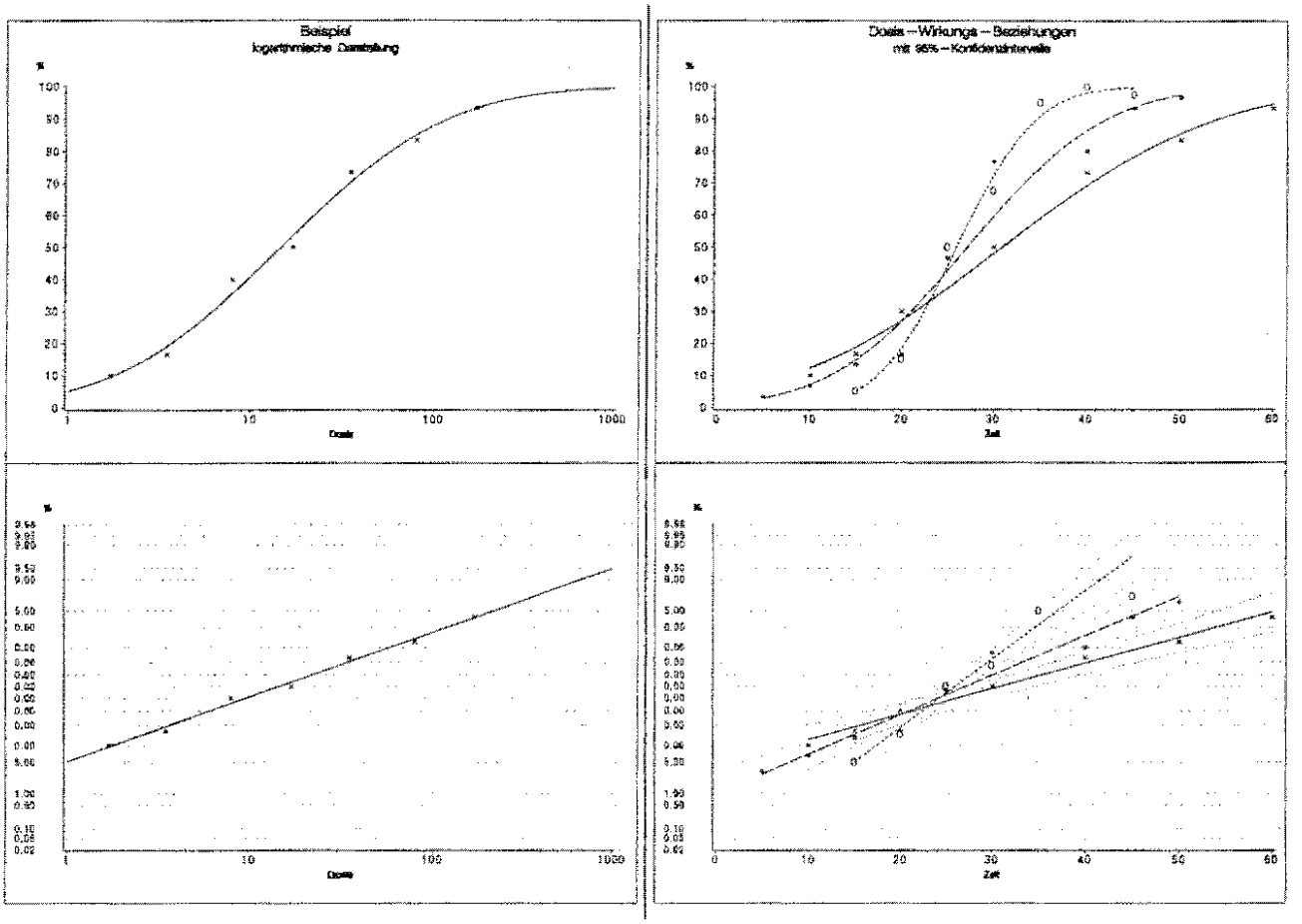
Dosisstufe Null soll nicht einbezogen werden, da c=0.1 natürliche Mortalität von 0.1

```
Probit Procedure

Data Set              =WORK.DATEN
Dependent Variable=R
Dependent Variable=N
Number of Observations=    6
Number of Events      =       44    Number of Trials =        60

Log Likelihood for NORMAL  -28.2675015

Probit Procedure

Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value

INTERCPT   1 -0.8750129 0.468392  3.489877  0.0617 Intercept
DOSIS      1 4.48147499 1.397845  10.27836  0.0013

C=0.1000

Probit Model in Terms of Tolerance Distribution

           MU          SIGMA
    0.195251       0.223141

Estimated Covariance Matrix for Tolerance Parameters

                            MU              SIGMA

          MU          0.003218          -0.002284
       SIGMA          -0.002284          0.004844
```

Die Geradengleichung lautet: Wirkung = -0.875 + 4.481 Dosis.
Der Unterschied zum obigen Ergebnis kommt dadurch zustande, daß die natürliche Mortalität in diesem Beispiel nicht geschätzt und auf die anderen Dosiswerte übertragen wird, sondern fest vorgegeben wurde. Allerdings dürfte die natürliche Mortalität in der Grundgesamtheit selten bekannt sein.

Darstellung im Wahrscheinlichkeitsnetz

Die traditionelle Auswertung basiert auf der Probittransformation. Aber zunehmend werden auch andere (s. o.) Transformationen und Verteilungen zugrunde gelegt. Um eine sigmoide Dosis-Wirkungs-Beziehung im Wahrscheinlichkeitsnetz (Probittransformation – Normalverteilung) darzustellen, bedarf es schon etwas Aufwand. Hilfe geben die dafür erarbeiteten SAS-Macros[6], mit denen auch nachstehende Grafiken erstellt wurden.



[6] MOLL, E.: Zur Umsetzung biometrischer Verfahren in SAS mit Beispielen aus dem Pflanzenschutz Berichte aus der Biologischen Bundesanstalt für Land- und Forstwirtschaft, Heft 10, 1996, S. 147-150 und 178-185

16 SAS Tools zur Datenanalyse am Beispiel des ANSCOMBE´ Quartett

Mit nur geringen SAS-Kenntnissen, die sich vor allem auf die Struktur einer SAS-Datei beziehen, können ohne Programmierung grafische Darstellungen zur beschreibenden Datenanalyse erstellt, statistische Maßzahlen berechnet aber auch Datenanalysen wie beispielsweise Tests durchgeführt werden.

16.1 Das ANSCOMBE´ Quartett

ANSCOMBE[7] betrachtet vier verschiedene Fälle zu jeweils zwei paarweisen Zufallsvariablen:

X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Sein Anliegen ist es anhand dieses Quartetts mit numerisch gleichen Ergebnissen zu zeigen:
- Daten ansehen

und erst wenn sie der jeweiligen Aufgabenstellung entsprechen
- Daten auswerten.

Zur Arbeit mit SAS/INSIGHT und der Analyst Application wird eine SAS-Datei erstellt.

```
data anscombe;
  input x1 y1 x2 y2 x3 y3 x4 y4;
  quartett = "Anscombe1" ; x = x1 ; y = y1 ; output;
  quartett = "Anscombe2" ; x = x2 ; y = y2 ; output;
  quartett = "Anscombe3" ; x = x3 ; y = y3 ; output;
  quartett = "Anscombe4" ; x = x4 ; y = y4 ; output;
  keep quartett x y;
lines;
10.0  8.04   10.0  9.14   10.0  7.46    8.0  6.58
 8.0  6.95    8.0  8.14    8.0  6.77    8.0  5.76
13.0  7.58   13.0  8.74   13.0 12.74    8.0  7.71
 9.0  8.81    9.0  8.77    9.0  7.11    8.0  8.84
11.0  8.33   11.0  9.26   11.0  7.81    8.0  8.47
14.0  9.96   14.0  8.10   14.0  8.84    8.0  7.04
 6.0  7.24    6.0  6.13    6.0  6.08    8.0  5.25
 4.0  4.26    4.0  3.10    4.0  5.39   19.0 12.50
12.0 10.84   12.0  9.13   12.0  8.15    8.0  5.56
 7.0  4.82    7.0  7.26    7.0  6.42    8.0  7.91
 5.0  5.68    5.0  4.74    5.0  5.73    8.0  6.89
;
run;
```

Klassifikationsvariable/Faktor: `quartett`
auszuwertende Variable: `x und y`

Die entstehende SAS-Datei anscombe befindet sich im SAS-Ordner WORK .

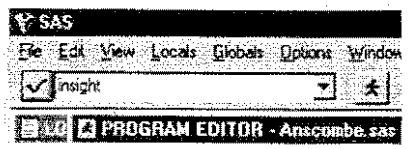
[7] F. J. ANSCOMBE: Graphs in Statistical Analysis, The American Statistican, Febr. 1973, p. 17-21

16.2 Zu SAS/INSIGHT

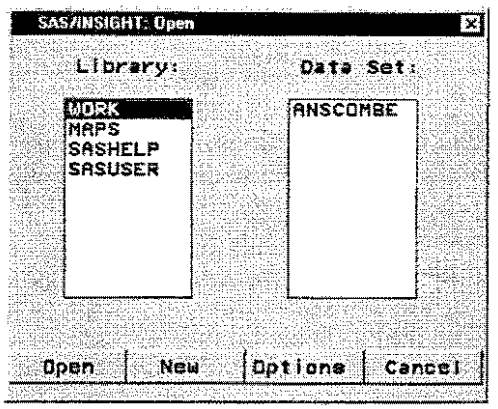
SAS/INSIGHT ist ein interaktives, sehr mächtiges, grafisch orientiertes Paket von SAS, zu dem es zwei Wege gibt:

1. als Befehl in der Command-Line

In die Command-Line wird INSIGHT geschrieben und damit SAS/INSIGHT gestartet:



Im neu eröffneten wird der Ordner (Library) ausgewählt, in dem die auszuwertende SAS-Datei steht. Diese Datei (Data Set) wird angeklickt und geöffnet.

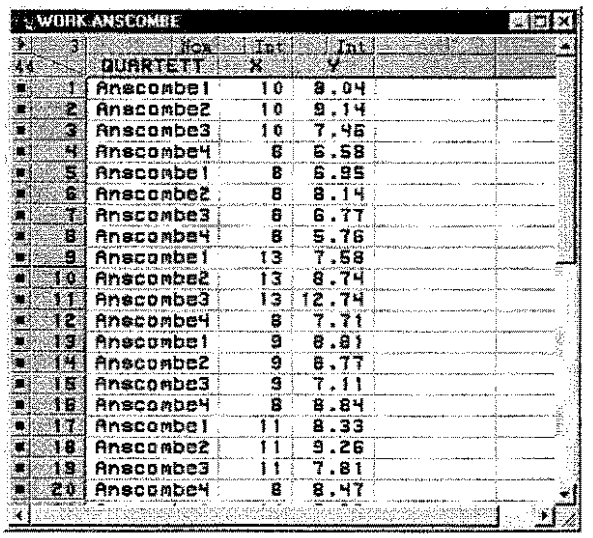


2. als Programm im Programm-Editor-Fenster

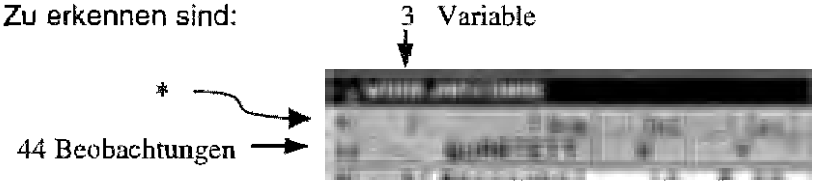
```
proc insight data=anscombe;
run;
```

Von nun an gleichen sich beide Wege.

Die SAS-Datei wird mit Lese- und Schreibrechten geöffnet:

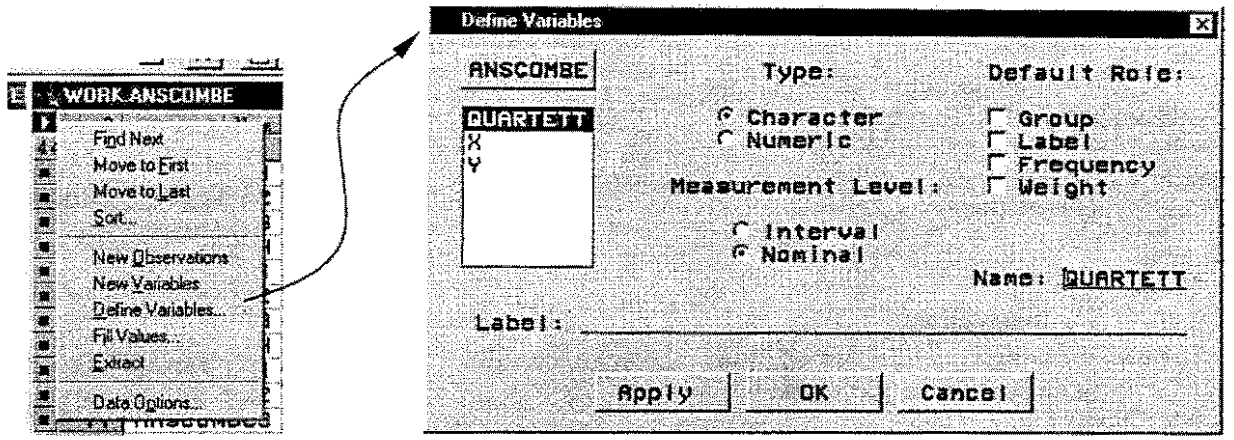


Zu erkennen sind:



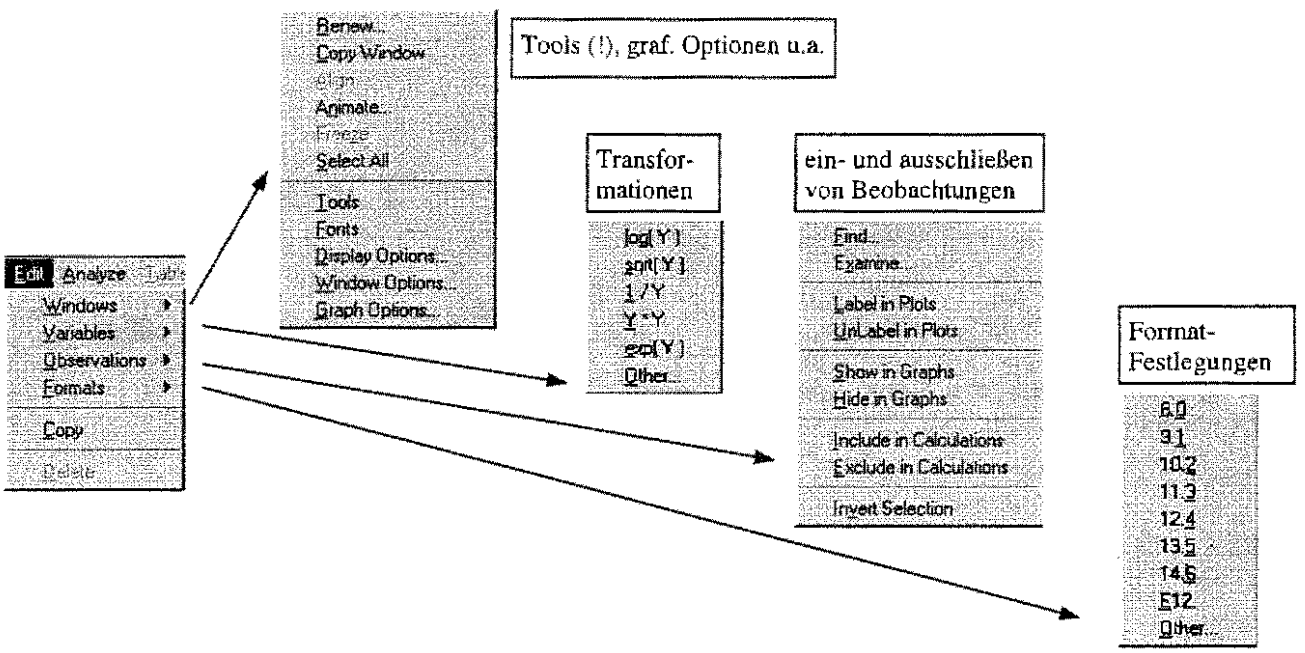
Voreingestellt sind nur zwei Skalenniveaus:
Nom: nominalskaliert
Int: intervallskaliert
Geändert werden diese mittels (s. *)

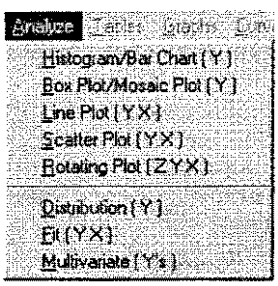
Nach dem Anklicken des Dreiecks erscheint ein neues Fenster, mit dessen Hilfe die Variablen neu definiert werden können:



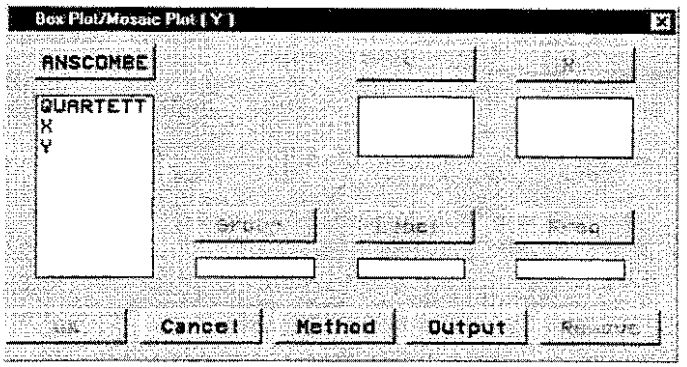
Das ist besonders dann wichtig, wenn es sich um Klassifikationsvariable (Faktoren) handelt, deren Stufen numerisch sind und automatisch eine Intervallskalierung festgestellt wird.

In der Menü-Leiste sind für die Analyse im Moment Edit und Analyze von Interesse.

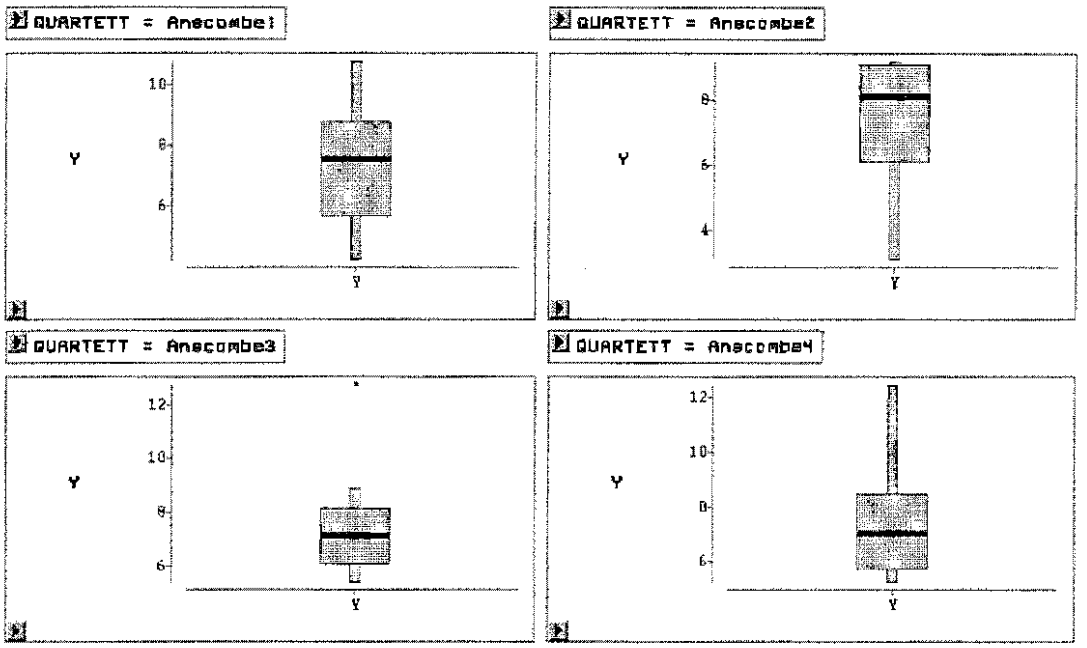




Fangen wir mit der (grafischen) Analyse an und wählen die Box-Plot-Darstellung:

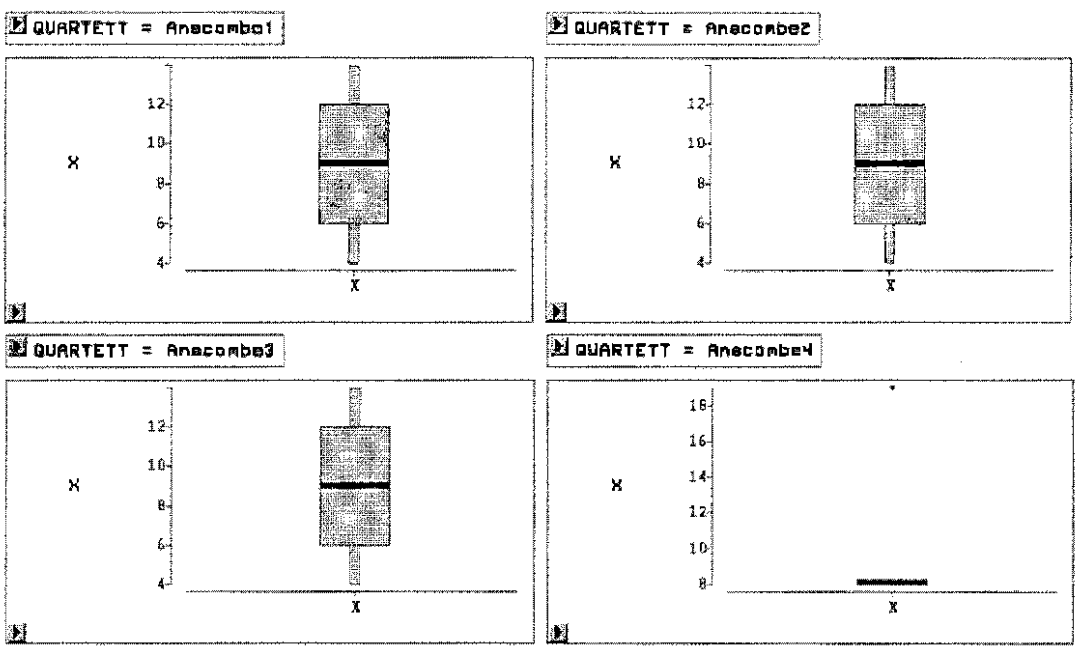


Variable Y → Y Variable QUARTETT → Group

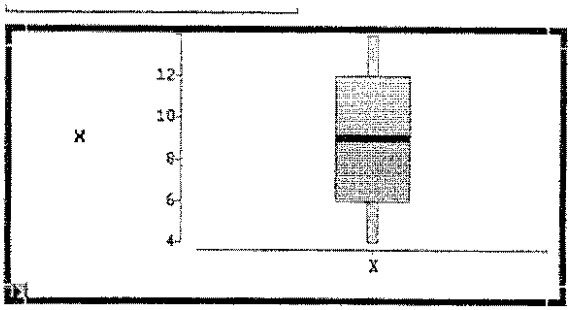


analog:

Variable X → Y Variable QUARTETT → Group

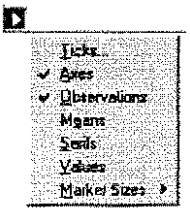


Wird auf die Umrandung der Grafik (oder der Überschrift) geklickt, so wird diese dicker gezeichnet: dieser Ausschnitt kann nun verschoben, vergrößert und verkleinert werden.

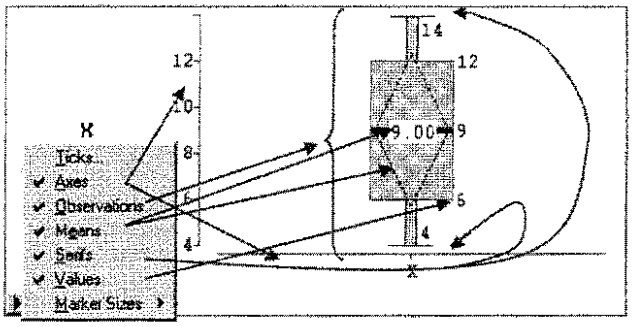


Weiterhin fallen sowohl in der (grafischen) Überschrift als auch in der Grafik selbst die Dreiecke in der linken unteren Ecke auf. Klicken Sie darauf. Für die Überschrift könnten Sie – wenn sinnvoll – die Formate ändern:

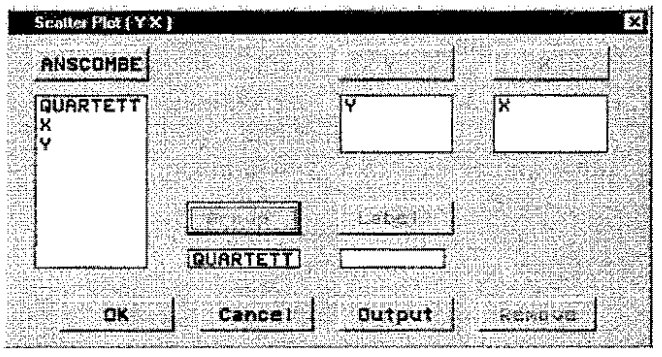
Bei der Grafik kann diese mittels eines sich öffnenden Fensters verändert werden:

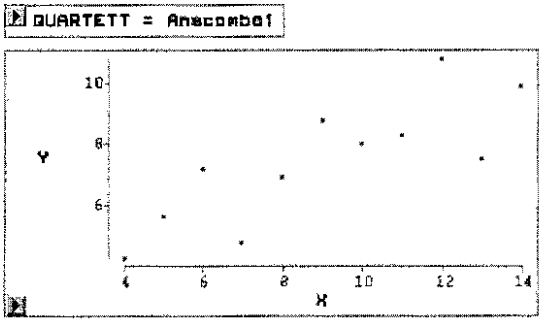


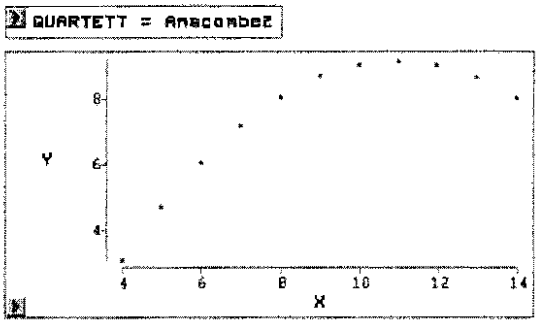
Das selbe Fenster wird auch geöffnet, wenn auf die Grafikfläche mit der rechten Mouse-Taste geklickt wird. Informationen können entfernt oder hinzugefügt werden:

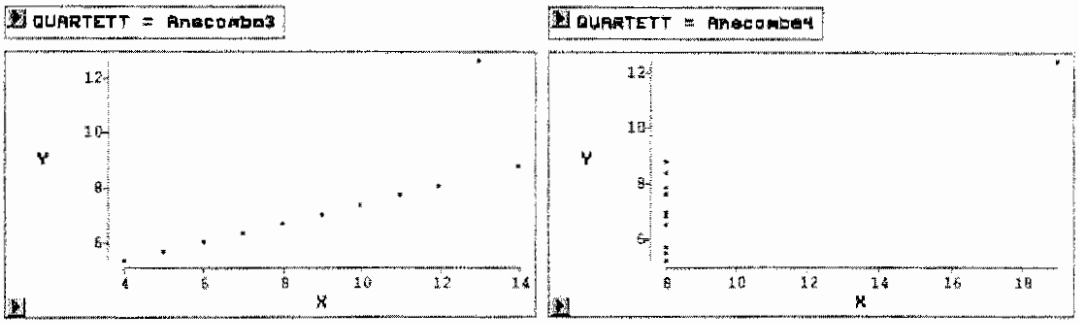


Aus dem Analyse-Fenster (s. o.) ist erkennbar, daß mehrere grafische Darstellungen möglich sind. Das Scatter-Plot wird nun ausgewählt:

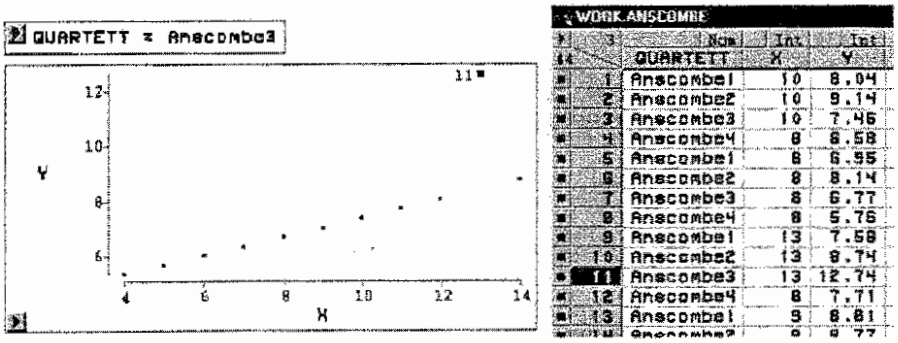








Wird eine einzelne Beobachtung, ein Punkt, auf der Grafik angeklickt, so wird die Nummer der Beobachtung an den Punkt geschrieben und gleichzeitig wird sie im Daten-Fenster markiert.



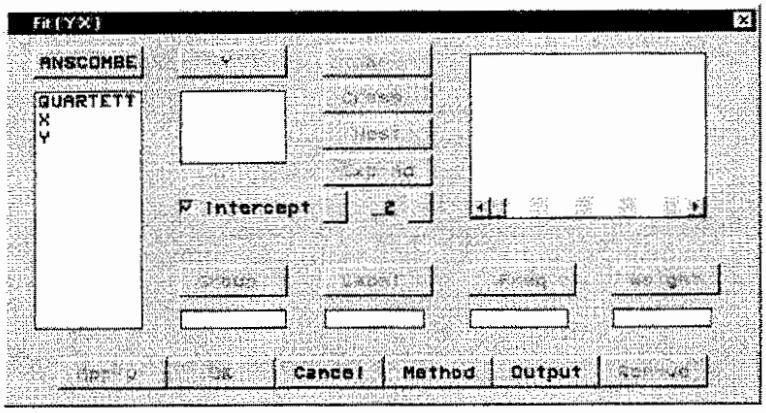
Das geht auch umgekehrt: wird im Datei-Fenster eine Beobachtung (keine Variable) ausgewählt, so wird in der Grafik diese Information hervorgehoben.

Auf ein interessantes Fenster soll nun eingegangen werden. Geöffnet wird es im Edit-Menü-Fenster unter Tools.

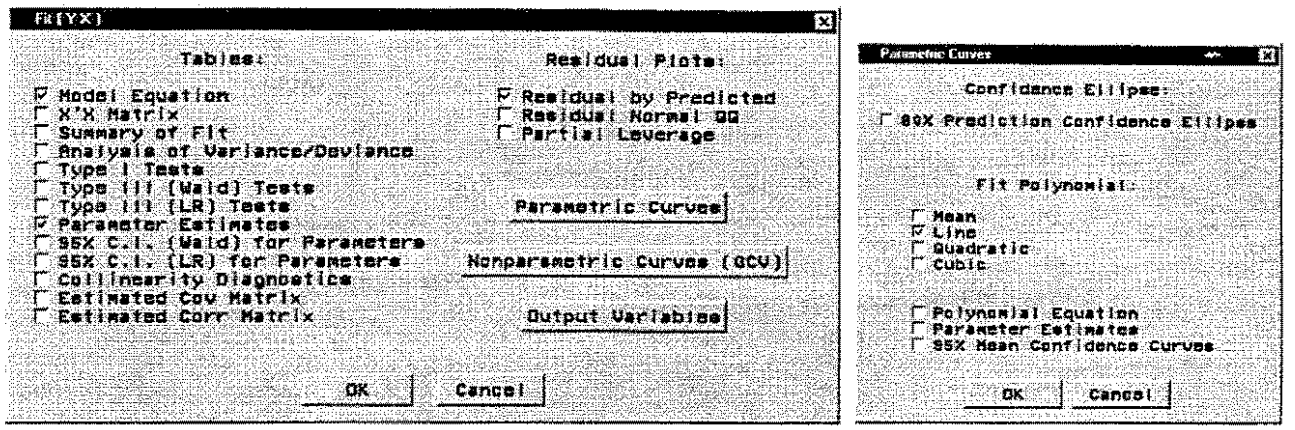
Werden im Datei-Fenster alle Beobachtungen markiert - das wird erreicht, indem auf die Anzahl der Beobachtungen (hier: 44) geklickt wird - können für alle Beobachtungen Symbol und Symbolfarbe mit Hilfe des Tools-Fensters neu gewählt werden. Jede dieser Schritte ist parallel im Datei-Fenster und in den grafischen Fenstern zu erkennen.

Auch nur für einzelne Beobachtungen oder eine Auswahl von Beobachtungen - die Strg-Taste unterstützt mehrere Auswahlen - können spezielle Symbole und Farben bestimmt werden.

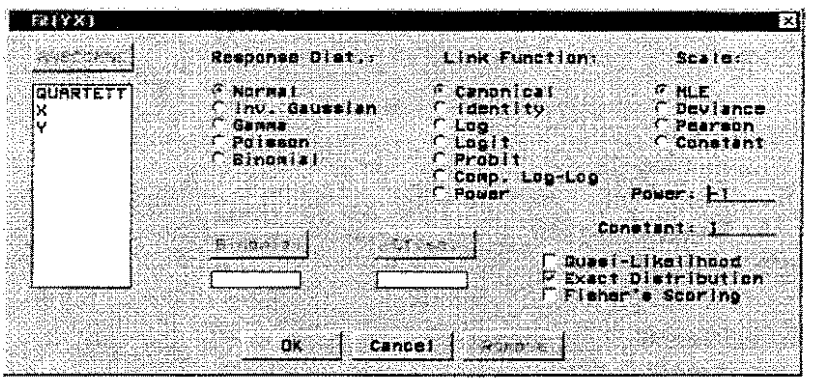
Im Analyse-Menü-Fenster sind einige Auswertungsmöglichkeiten zu finden: die Regressionsanalyse beispielsweise unter dem Menüpunkt Fit (X Y):



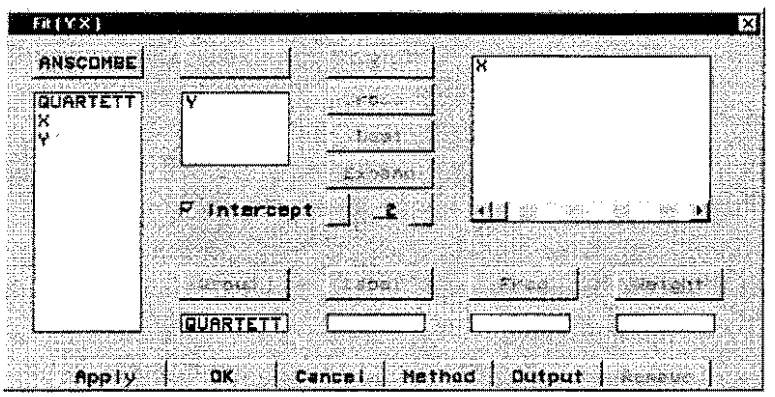
Betätigen Sie den Button Output und wählen nur die Modell-Gleichung, die Parameterschätzung und die Darstellung der Residuen aus:



Die lineare Anpassung (Parametric Curves) wird akzeptiert und zusätzliche Variable (Output Variables) sollen nicht ausgegeben werden.
Die Methode (Method) soll in ihrer Standardeinstellung belassen werden:



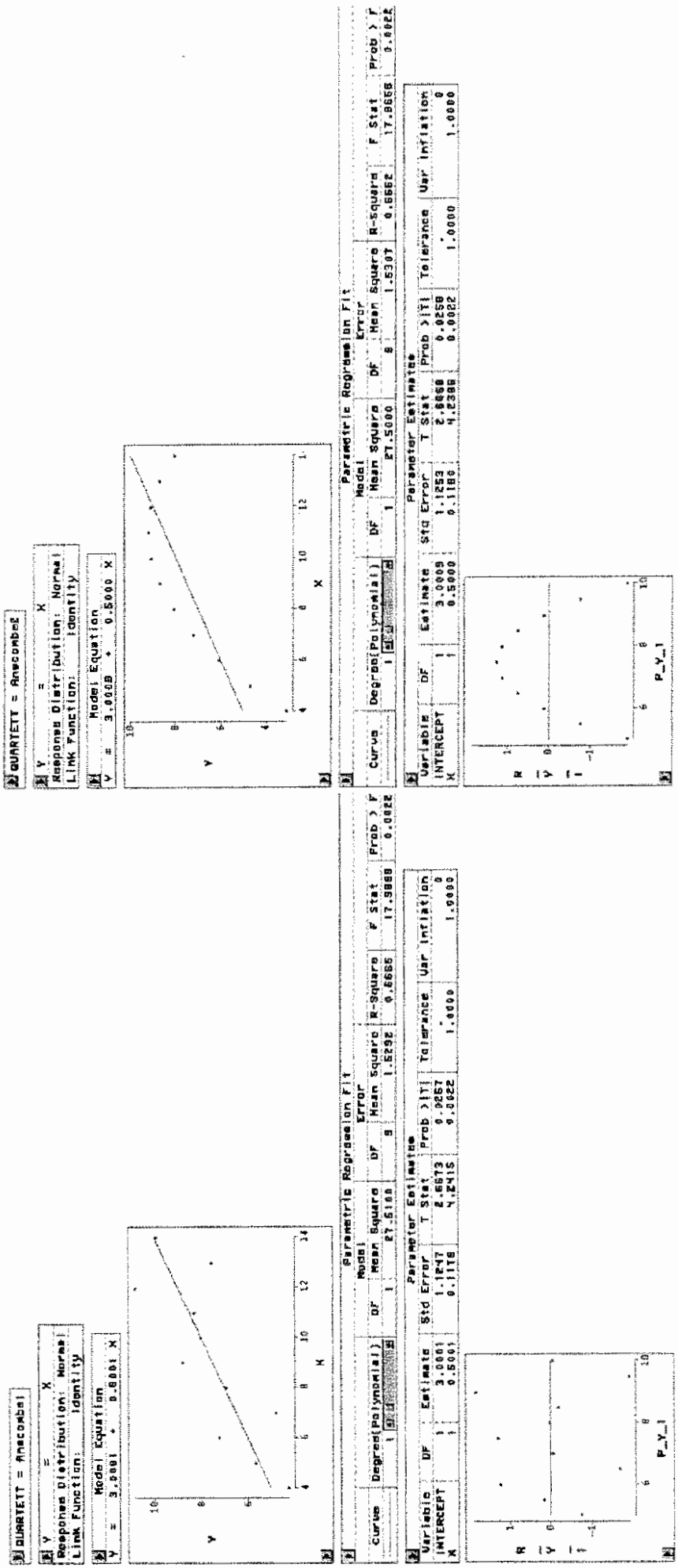
Die Zuordnung der auszuwertenden Variablen wird wie folgt vorgenommen

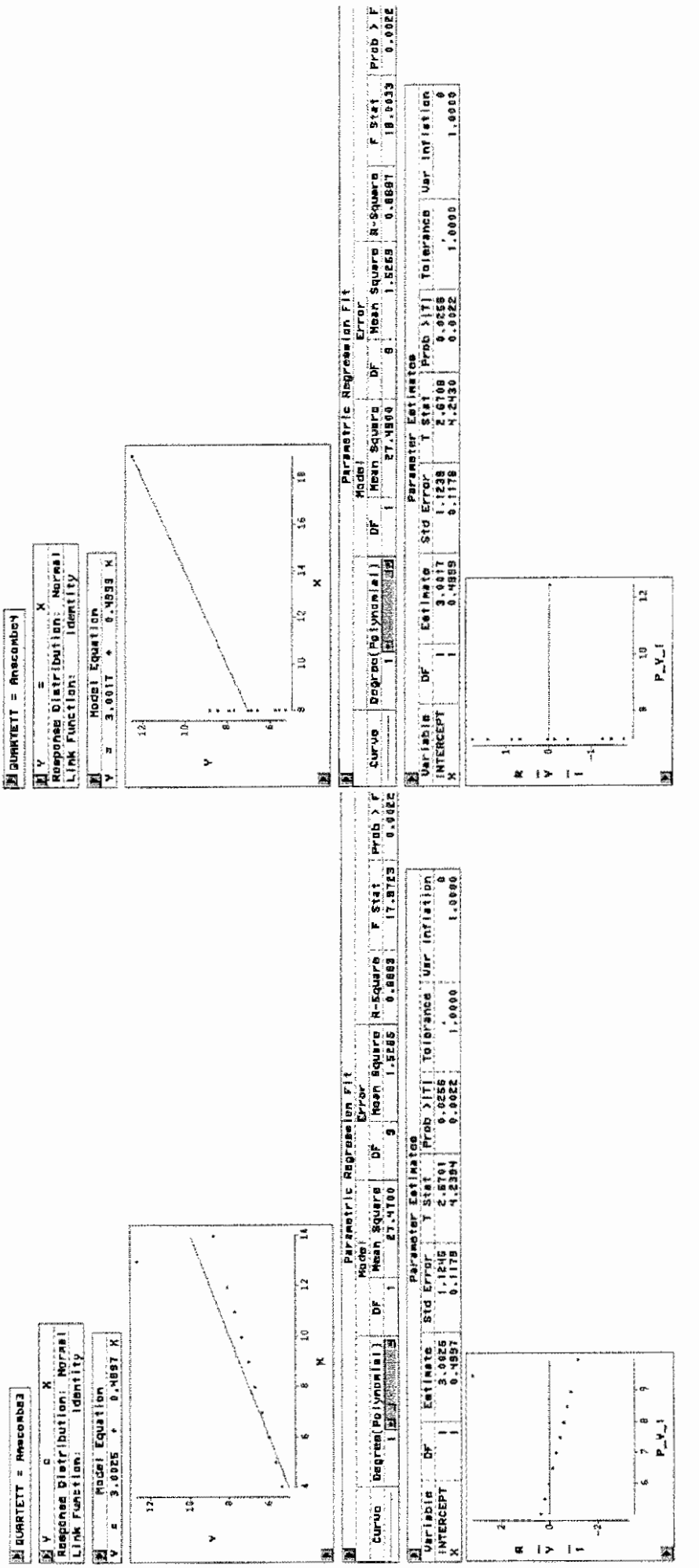


Die Ergebnisse sind:

	Modell-Gleichung	Bestimmtheitsmaß
Anscombe1	Y = 3.0001 + 0.5001 * X	0,6665
Anscombe2	Y = 3.0009 + 0.5000 * X	0,6662
Anscombe3	Y = 3.0025 + 0.4997 * X	0,6663
Anscombe4	Y = 3.0017 + 0.4999 * X	0,6667

Bis auf ganz geringe Abweichungen gleichen sich die Schätzungen.


QUARTETT = Anscombe1
Y = X
Response Distribution: Normal
Link Function: Identity
Model Equation
Y = 3.0001 + 0.5001 X
Parametric Regression Fit
Model
Curve Degree(Polynomial) DF Mean Square F Stat Prob > F
1 1 27.5100 17.9988 0.0022
Error
DF Mean Square R-Square
9 1.5282 0.6665
Parameter Estimate
Variable DF Estimate Std Error T Stat Prob >|t| Tolerance Var Inflation
INTERCEPT 1 3.0001 1.1247 2.6673 0.0257 . 0
X 1 0.5001 0.1178 4.2415 0.0022 1.0000 1.0000
QUARTETT = Anscombe2
Y = X
Response Distribution: Normal
Link Function: Identity
Model Equation
Y = 3.0009 + 0.5000 X
Parametric Regression Fit
Model
Curve Degree(Polynomial) DF Mean Square F Stat Prob > F
1 1 27.5000 17.8668 0.0022
Error
DF Mean Square R-Square
8 1.5307 0.6662
Parameter Estimate
Variable DF Estimate Std Error T Stat Prob >|t| Tolerance Var Inflation
INTERCEPT 1 3.0009 1.1253 2.6658 0.0258 . 0
X 1 0.5000 0.1180 4.2386 0.0022 1.0000 1.0000

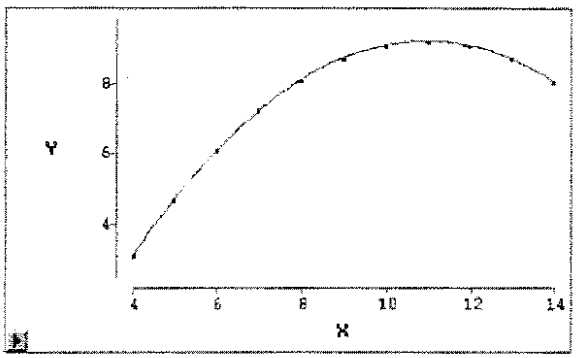

QUARTETT = Anscombe3
Y = X
Response Distribution: Normal
Link Function: Identity
Model Equation
Y = 3.0025 + 0.4997 X
Parametric Regression Fit
Model
Error
Curve Degree(Polynomial) DF Mean Square DF Mean Square R-Square F Stat Prob > F
1 1 27.4700 9 1.5265 0.8883 17.9725 0.0022
Parameter Estimates
Variable DF Estimate Std Error T Stat Prob >|T| Tolerance Var Inflation
INTERCEPT 1 3.0025 1.1245 2.6701 0.0256 . 0
X 1 0.4997 0.1179 4.2384 0.0022 1.0000 1.0000
QUARTETT = Anscombe4
Y = X
Response Distribution: Normal
Link Function: Identity
Model Equation
Y = 3.0017 + 0.4999 X
Parametric Regression Fit
Model
Error
Curve Degree(Polynomial) DF Mean Square DF Mean Square R-Square F Stat Prob > F
1 1 27.4900 9 1.5265 0.8887 18.0039 0.0022
Parameter Estimates
Variable DF Estimate Std Error T Stat Prob >|T| Tolerance Var Inflation
INTERCEPT 1 3.0017 1.1238 2.6708 0.0256 . 0
X 1 0.4999 0.1178 4.2430 0.0022 1.0000 1.0000

SAS/INSIGHT bietet zur Datenvisualisierung und –analyse viel, viel mehr als hier an diesem Beispiel gezeigt werden kann. Dazu gehören Verteilungsanpassungen oder dreidimensionale Darstellungen mit Rotation.

Auf eine Möglichkeit soll aber noch eingegangen werden. In den Grafiken befindet sich ein Slider [Degree[Polynominal)] . Für Anscomb2 wird der betätigt und Degree = 2 gesetzt:

Model Equation

Y = 3.0009 + 0.5000 X



Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
———	2	2	20.6381	8	2.79TE-06	1.0000	7376132.92	0.0001

Die Modell-Gleichung und die Darstellung der Residuen ändern sich nicht! – können auch nicht, da *linear* vorgegeben war.
Die Grafik, das Bestimmtheitsmaß und auch die Überschreitungswahrscheinlichkeit des F-Testes lassen eine wesentlich bessere Anpassung durch eine quadratische Funktion erkennen.

Die Möglichkeit des Nachvollziehens ist in SAS/INSIGHT gegeben.
Mit filename wird eine Datei zugeordnet, die dann alle Aktivitäten aufnimmt und ggf. wie ein „normales" SAS-Programm abgearbeitet werden kann. Die Zuordnung dieser Datei erfolgt im Aufruf von SAS/INSIGHT:

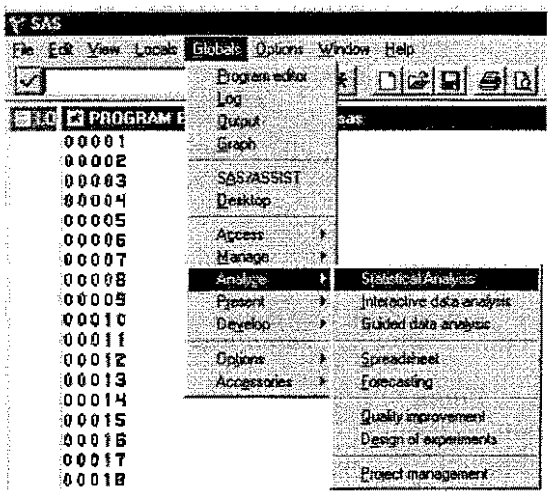
```
filename f_ins "<Dateiname>";
proc insight data=<SAS-Datei> file=f_ins;
run;
```

16.3 Zur Analyst Application

Wenn die SAS 6.12 Software mit SAS/STAT, SAS/GRAPH, SAS/ASSIST installiert ist und zusätzlich mit spezieller Zielstellung SAS/FSP [zum Editieren der Daten], SAS/ACCESS [zum Importieren von Dateien], SAS/QC [zur Nutzung des vollen Umfangs der Analyst Application] und SAS/IML [aus dem gleichen Grunde wie SAS/QC] , kann die Analyst Application von der SAS-Web-Seite

http://ww.sas.com/rnd/app/analyst.html

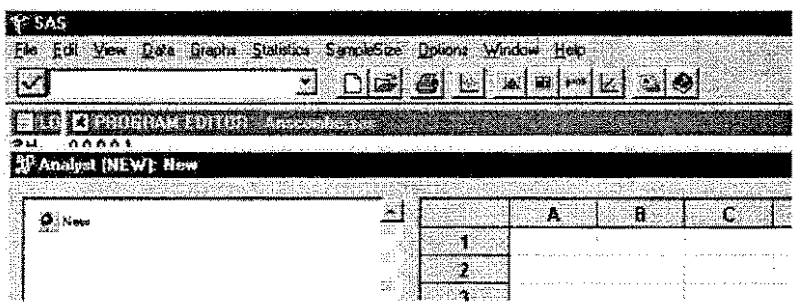
heruntergeladen werden. Das Paket *Analyst.exe* wird durch Starten ausgepackt. Im SAS 6.12 Ordner entsteht ein Ordner *Addon*. Die Dateien dieses Ordners sind wie dort vorgezeichnet in den Ordner STAT bzw. in die entsprechenden Unterverzeichnisse dieses Ordners zu kopieren.



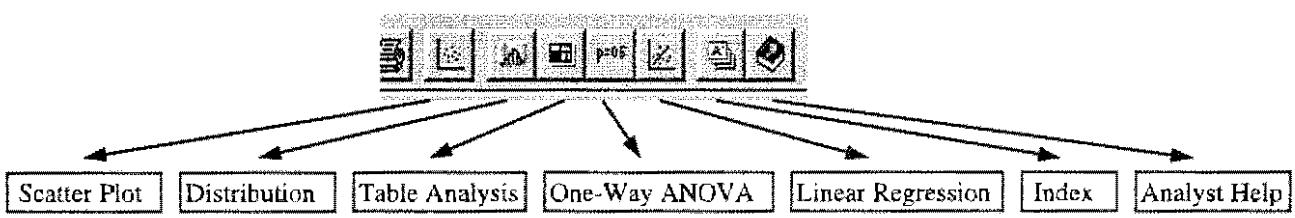
Über das PullDown-Menü `Globals` kommt man zu der Wahlmöglichkeit „`Statistische Analyse`".

In SAS 8 ist neben den bekannten Menüpunkten ein zusätzlicher: `Solution` zu finden, unter dem die Analyst Application integriert ist.

Es entsteht ein bisher ungewohnter Bildschirm mit Tabellenstruktur:

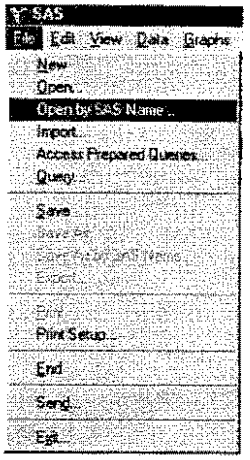


Auch ein Blick auf die Toolbox zeigt Neuerungen:

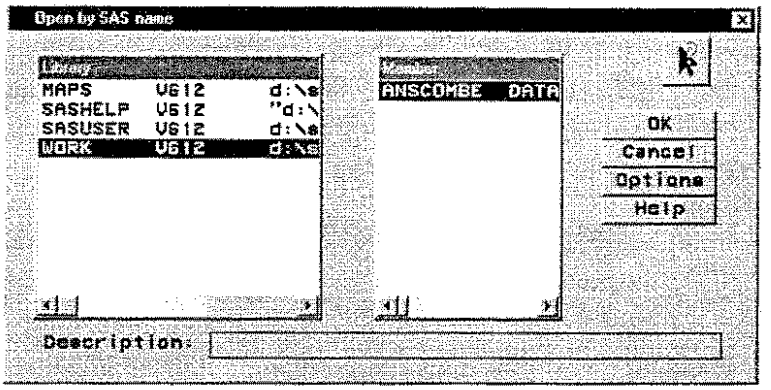


Die Tabelle ist leer: es fehlen die Daten.

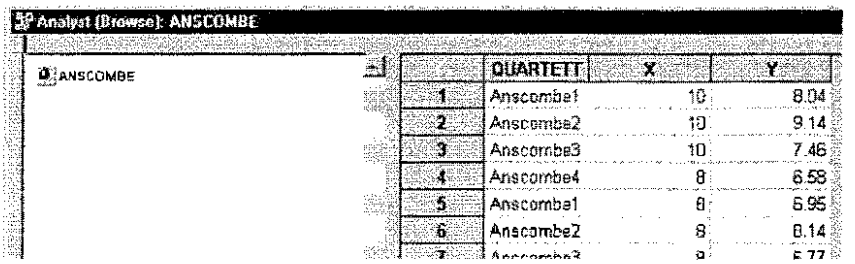
Es kann eine SAS-Datei geöffnet werden:



welche, erfolgt in einem speziellen Fenster:



Der Dateiinhalt wird angezeigt:



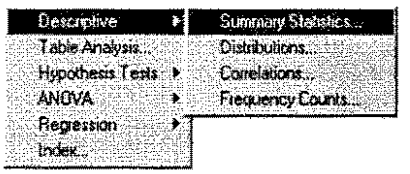
	QUARTETT	X	Y
1	Anscombe1	10	8.04
2	Anscombe2	10	9.14
3	Anscombe3	10	7.46
4	Anscombe4	8	6.58
5	Anscombe1	8	6.95
6	Anscombe2	8	8.14

Auch der Inhalt dieser Datei kann editiert werden.

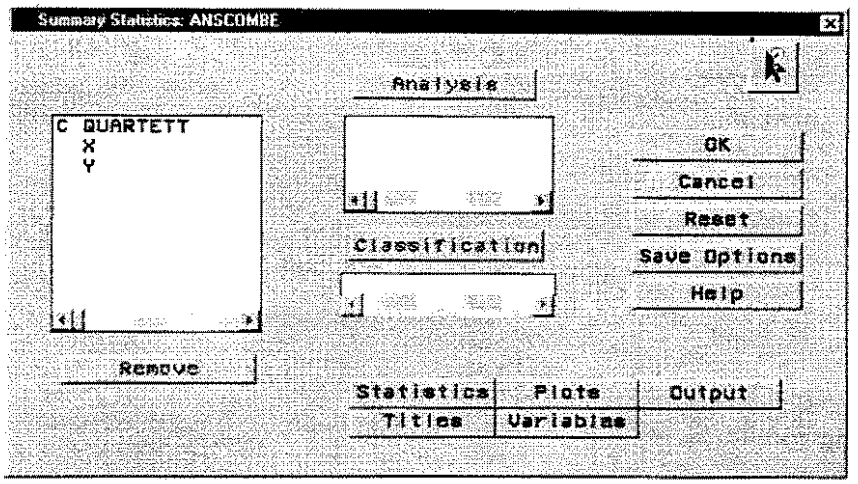
PullDown-Menü `Edit` oder
rechte Mousetaste / `Edit`

Statistische Maßzahlen

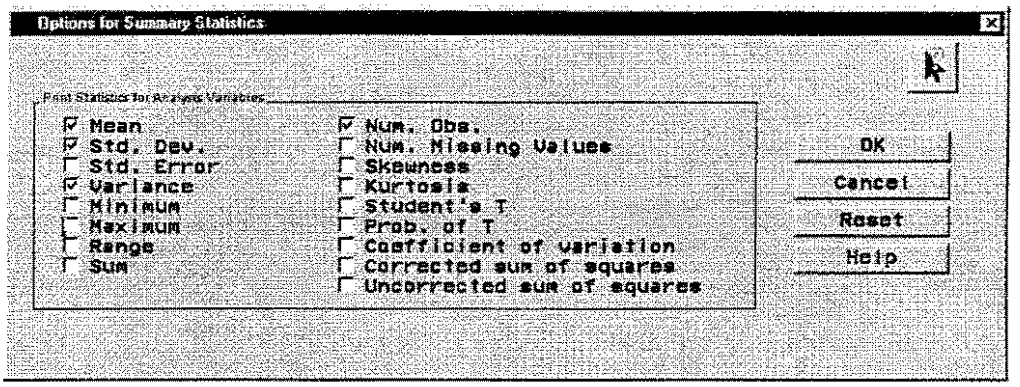
Zunächst sollen statistische Maßzahlen berechnet werden. Dazu wird vom PullDown-Menü oder rechte Mousetaste `Statistics` ausgegangen:



In dem folgenden Fenster werden die auszuwertenden Variablen und die zu berechnenden statistischen Maßzahlen festgelegt.



Die Variable, die ausgewertet werden sollen, werden markiert. Mit der Strg-Taste ist eine Mehrfachauswahl möglich. Für das Beispiel sind das die Variablen X und Y. Nun wird der inzwischen freigegebenen Button `Analysis` betätigt, wodurch die ausgewählten Variablen übernommen werden. Die Character-Variable QUARTETT soll als Klassifikationsvariable (Faktor mit 4 Stufen) dienen. Sie wird markiert und durch Betätigen des Button `Classifikation` entsprechend festgelegt. Mit Hilfe des Button `Statistics` werden die zu berechnenden statistischen Maßzahlen ausgewählt: Mittelwert, Standardabweichung, Varianz und Anzahl der Beobachtungen:

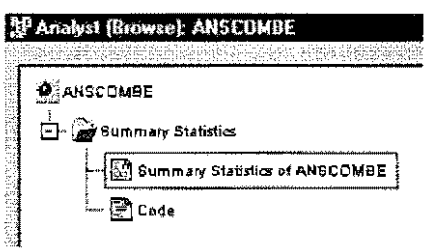


Es erscheint ein dem üblichen SAS-Output entsprechendes Ergebnis in einem speziellen Fenster.

QUARTETT	N Obs	Variable	Mean	Std Dev	Variance	N
Anscombe1	11	X	9.0000000	3.3166248	11.0000000	11
		Y	7.5009091	2.0315681	4.1272691	11
Anscombe2	11	X	9.0000000	3.3166248	11.0000000	11
		Y	7.5009091	2.0316567	4.1276291	11
Anscombe3	11	X	9.0000000	3.3166248	11.0000000	11
		Y	7.5000000	2.0304236	4.1226200	11
Anscombe4	11	X	9.0000000	3.3166248	11.0000000	11
		Y	7.5009091	2.0305785	4.1232491	11

Zu erkennen ist, daß die Anzahl der Beobachtungen identisch 11, die X-Mittelwert für alle 4 Fälle 9,0 , die Y-Mitelwerte 7,5 und auch die jeweiligen X- und Y-Varianzen und -Standardabweichungen gleich sind!

Wird das Fenster geschlossen, erscheint ein interessantes Bild:



Das, was bisher gemacht wurde, wird demonstriert. Und es ist ein Icon Code erkennbar.

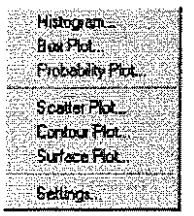
Der Inhalt dieses Icon ist:

```
*** Sort data by BY variables ***;
proc sort data=WORK.ANSCOMBE out=work._stsrt_;
   by QUARTETT;
run;
*** Print descriptive statistics for analysis variables ***;
options number;
options date;
/*  Remove any existing titles */
title;
proc means data=work._stsrt_
   mean std var n  ;
   var X Y;
   class QUARTETT;
```

Die entscheidende Prozedur ist PROC MEANS.
Mit `File / Save as` kann das Programm abgespeichert und damit wiederholt genutzt werden. Vor der Abarbeitung ist es um `run` zu ergänzen.

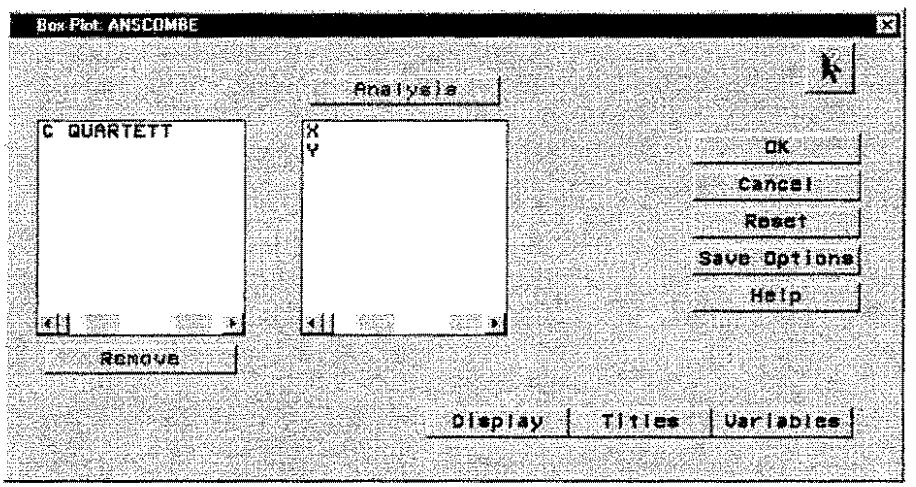
Grafische Darstellungen

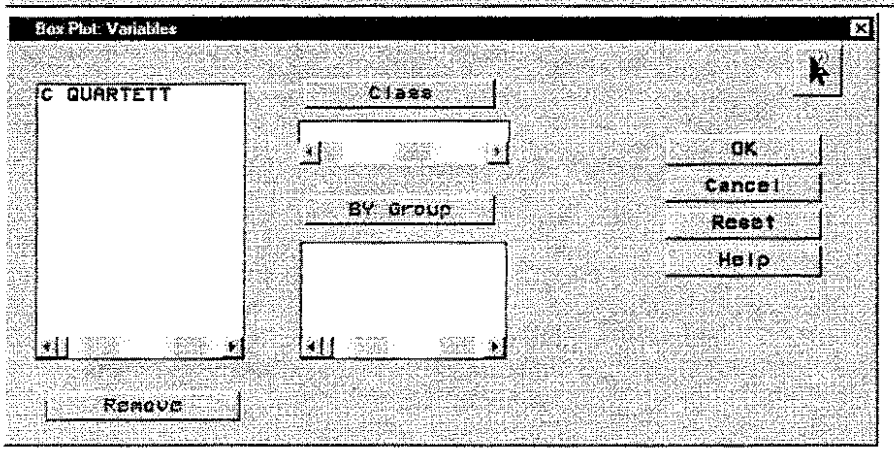
Doch zurück zu den Werten. Es dürfte von Interesse sein, sich die Werte grafisch anzusehen. Das kann bereits mit dem Button `Plots` im Summary Statistics-Fenster mit den Wahlmöglichkeiten o Histogram und o Box and whisker erfolgen oder separat



Das PullDown-Menü oder rechte Mousetaste `Graphs` liefert die nebenstehenden Möglichkeiten, von denen `Box Plot` ausgewählt wird.

Die auszuwertenden Variablen werden wieder entsprechend eingetragen:

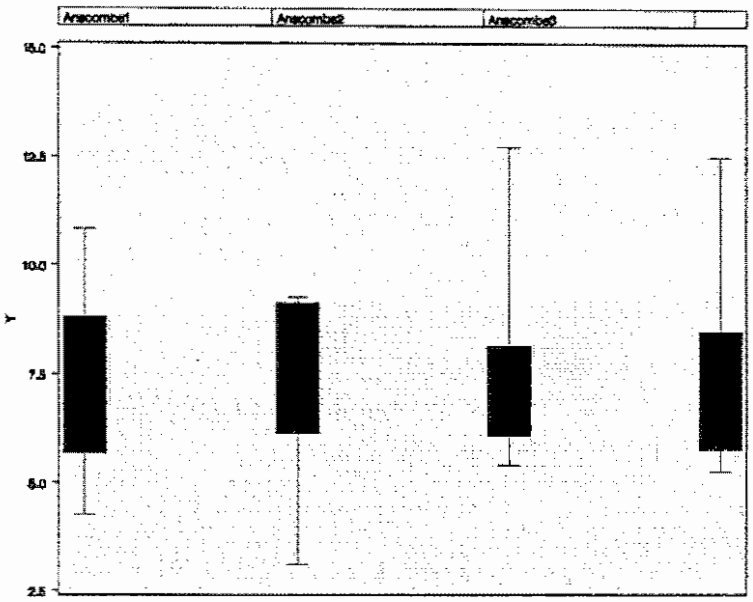




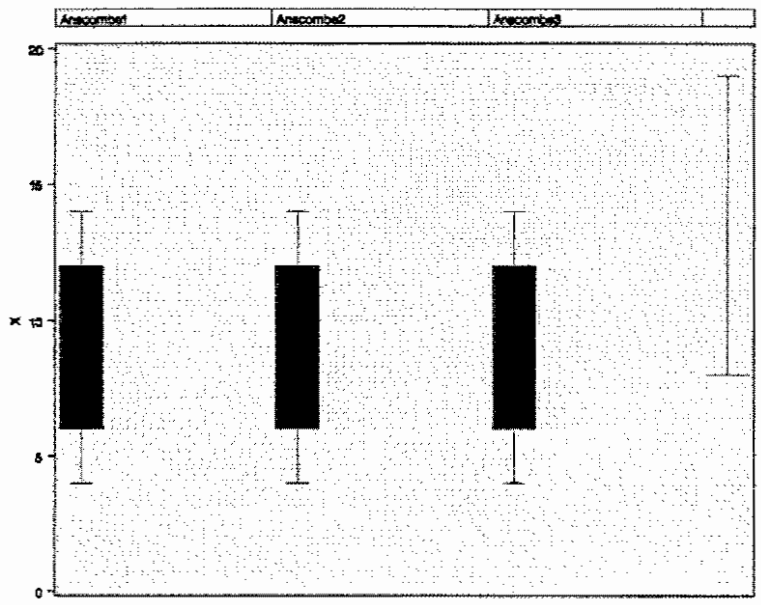
Da keine Darstellung der Variablen X und Y unabhängig von der Variablen ' QUARTETT erfolgen soll, wird der Button `Variables` betätigt:

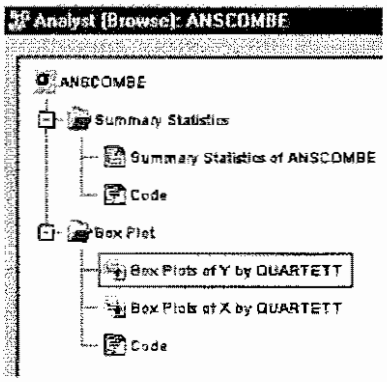
QUARTETT als Class-Variable liefert jeweils eine Grafik für X und Y; d. h. zwei Grafiken.
QUARTETT als BY-Variable führt zu einzelnen Grafiken für jeden Wert von QUARTETT sowohl für X als auch für Y , folglich zu
2 * 4 = 8 Grafiken.

Die Variable QUARTETT wird als Class-Variable gewählt. Das geöffnete grafische Fenster hat folgenden Inhalt.



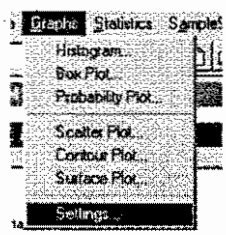
Neben dieser Grafik kann man sich auch die zweite Grafik
[Box Plots of X by QUARTETT]



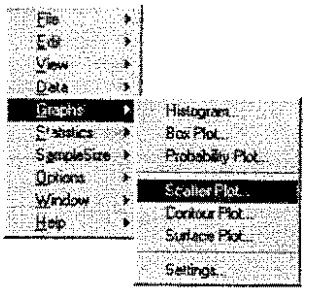


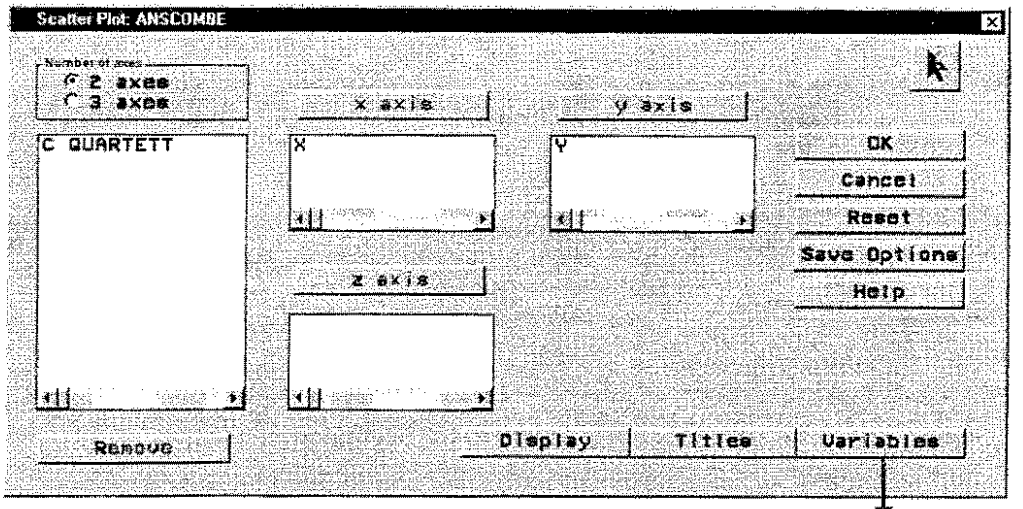
und den SAS-Code ansehen
(und abspeichern).

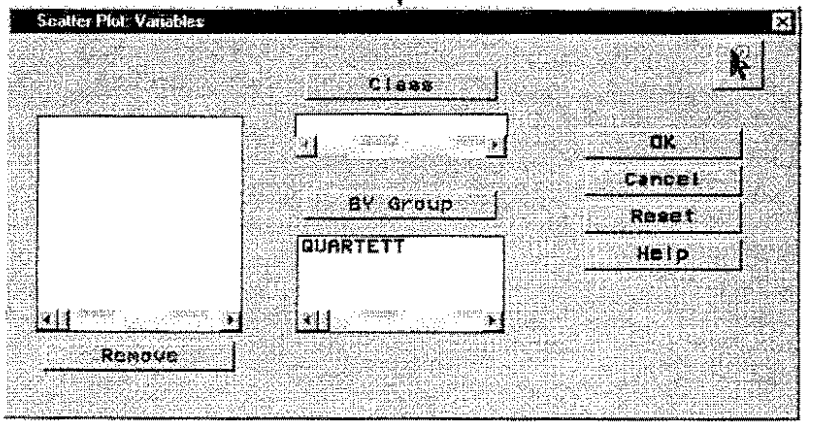
Die Einstellungen (Größen, Farben und dgl.) können mit `Settings` verändert werden:

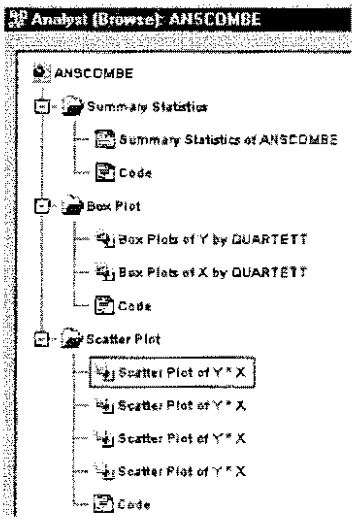


Die Gleichheit der statistischen Maßzahlen und diese Box-Plots „schreien danach", sich die einzelnen Grafiken anzusehen:



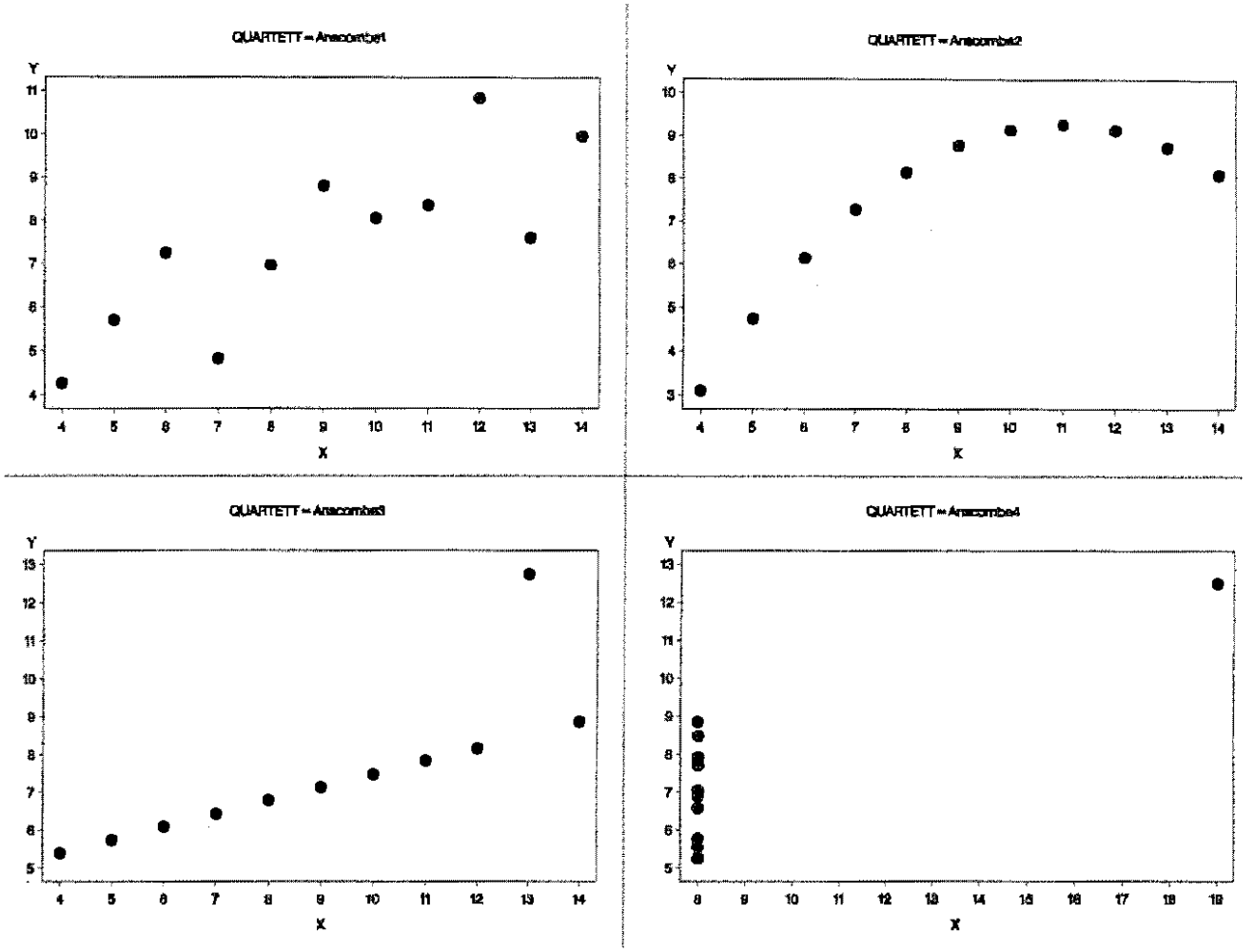






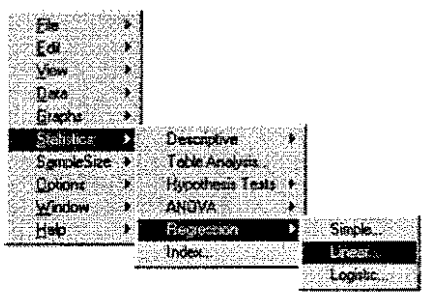
Die vier entstehenden Abbildungen sind:

das Anscombe-Quartett

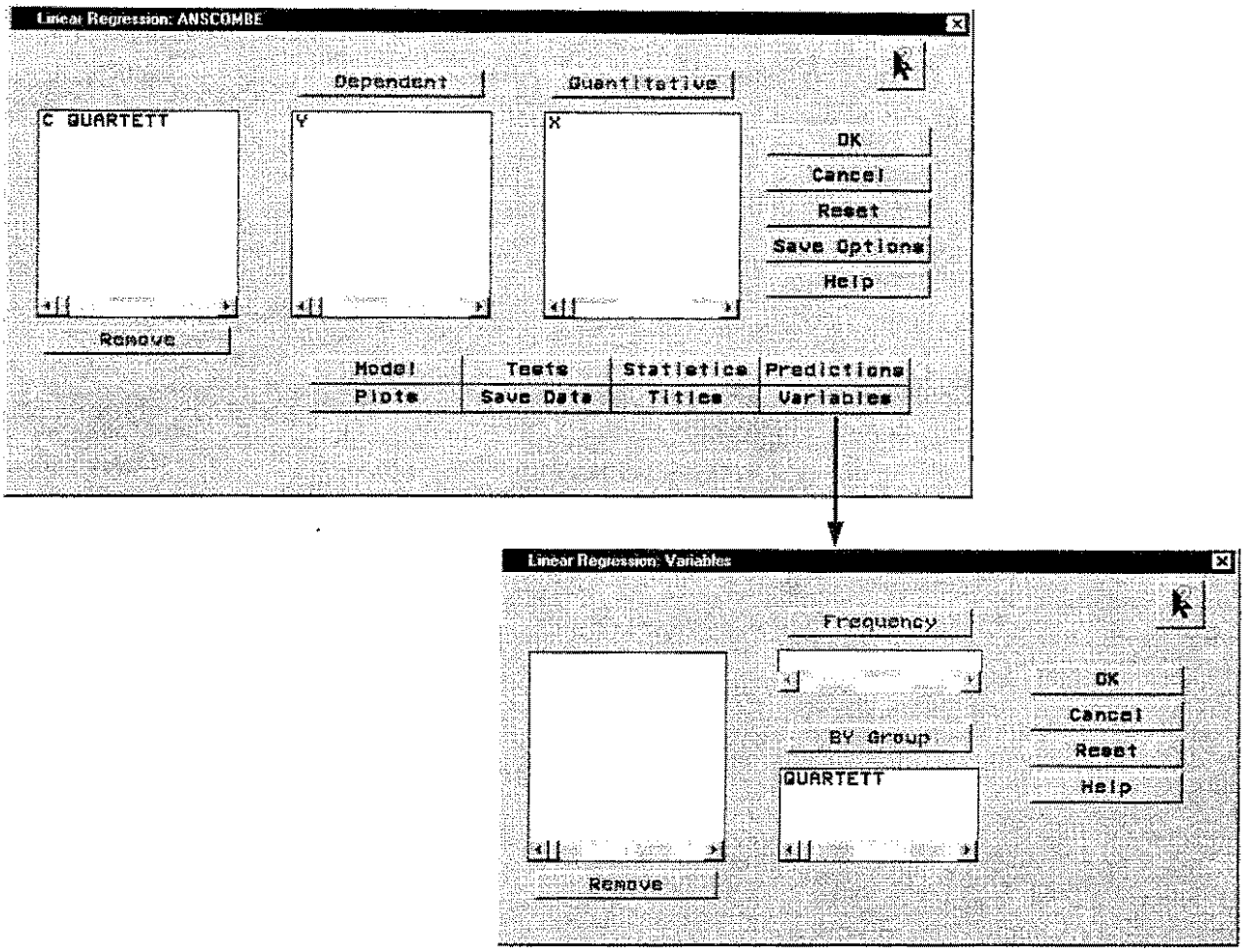


Regression

Die Analyst Application bietet mehr als nur statistische Maßzahlen.
Als Beispiel soll hier die Regression heran gezogen werden.



Mit Blick auf die grafische Darstellung, ist die lineare Regression nur in einem Fall angebracht.
Wird sie trotzdem für alle vier Fälle (formal) gerechnet,



ergibt sich folgendes Ergebnis.

```
QUARTETT=Anscombe1

Model: MODEL1
Dependent Variable: Y
Analysis of Variance

                          Sum of          Mean
Source                DF  Squares        Square    F Value    Prob>F
Model                  1  27.51000      27.51000      17.99    0.0022
Error                  9  13.76269       1.52919
C Total               10  41.27269

    Root MSE       1.23660    R-square     0.6665
    Dep Mean       7.50091    Adj R-sq     0.6295
    C.V.          16.48605

Parameter Estimates
                   Parameter     Standard    T for H0:
Variable      DF    Estimate        Error   Parameter=0   Prob > |T|
INTERCEP       1    3.000091   1.12474679         2.667       0.0257
X              1    0.500091   0.11790550         4.241       0.0022
```

```
QUARTETT=Anscombe2

Model: MODEL1
Dependent Variable: Y
Analysis of Variance

                          Sum of          Mean
Source                DF  Squares        Square     F Value    Prob>F
Model                  1  27.50000      27.50000       17.966    0.0022
Error                  9  13.77629       1.53070
C Total               10  41.27629

    Root MSE       1.23721    R-square    0.6662
    Dep Mean       7.50091    Adj R-sq    0.6292
    C.V.          16.49419

Parameter Estimates
                   Parameter      Standard    T for H0:
Variable  DF       Estimate         Error   Parameter=0   Prob > |T|
INTERCEP   1       3.000909    1.12530242         2.667       0.0258
X          1       0.500000    0.11796375         4.239       0.0022

QUARTETT=Anscombe3

Model: MODEL1
Dependent Variable: Y
Analysis of Variance

                          Sum of          Mean
Source                DF  Squares        Square     F Value    Prob>F
Model                  1  27.47001      27.47001       17.972    0.0022
Error                  9  13.75619       1.52847
C Total               10  41.22620

    Root MSE       1.23631    R-square    0.6663
    Dep Mean       7.50000    Adj R-sq    0.6292
    C.V.          16.48415

Parameter Estimates
                   Parameter      Standard    T for H0:
Variable  DF       Estimate         Error   Parameter=0   Prob > |T|
INTERCEP   1       3.002455    1.12448123         2.670       0.0256
X          1       0.499727    0.11787766         4.239       0.0022

QUARTETT=Anscombe4

Model: MODEL1
Dependent Variable: Y
Analysis of Variance

                          Sum of          Mean
Source                DF  Squares        Square     F Value    Prob>F
Model                  1  27.49000      27.49000       18.003    0.0022
Error                  9  13.74249       1.52694
C Total               10  41.23249

    Root MSE       1.23570    R-square    0.6667
    Dep Mean       7.50091    Adj R-sq    0.6297
    C.V.          16.47394

Parameter Estimates
                   Parameter      Standard    T for H0:
Variable  DF       Estimate         Error   Parameter=0   Prob > |T|
INTERCEP   1       3.001727    1.12392107         2.671       0.0256
X          1       0.499909    0.11781894         4.243       0.0022
```

Vergleichen Sie die Ergebnisse und Sie erkennen das Ziel von ANSCOMBE.

17 Kovarianzanalyse

17.1 Einführung in die Kovarianzanalyse

Mit Hilfe der Kovarianzanalyse wird wie bei der Varianzanalyse die mittlere Wirkung der Effekte der Prüffaktoren auf ein metrisches Merkmal untersucht. Hinzu kommen eine oder auch mehrere Variable, die einen (linearen) Einfluß auf das Prüfmerkmal haben. Mit der (linearen) Abhängigkeit des Merkmals von diesen Variablen werden folglich gleichzeitig Aspekte der Regressionsanalyse in die Auswertung einbezogen. Die Kovarianzanalyse verbindet die Varianz- und die Regressionsanalyse miteinander. Die Variablen, von denen das Merkmal abhängen, heißen Kovariable (von konkomitante Variable) oder Kovariate. Von einer einfachen Kovarianzanalyse wird gesprochen, wenn das Varianzanalysemodell um einen Regressor, eine (lineare) Einflußvariable, erweitert wird. Bei mehreren hinzukommenden Regressoren handelt es sich dann um eine mehrfache oder multiple Kovarianzanalyse.

Die Anwendung der Kovarianzanalyse ist nicht so verbreitet wie die der Varianzanalyse oder der Regressionsanalyse. Wenn man sich einige der nachfolgenden Anwendungen ansieht, erkennt man schnell den Vorteil.

MANHART[8] vergleicht beispielsweise die Lernergebnisse von Nutzern seiner Software unter Berücksichtigung von den das Vorwissen beschreibenden Kovariablen Vorwissen in Windows, Vorwissen in Statistik, Anzahl Tage der Computernutzung im Monat und Anzahl Tage der Windows-Nutzung im Monat.

RASCH u. a.[9] führen als Beispiele für die Kovarianzanalyse die varianzanalytische Auswertung von Feldversuchen unter Berücksichtigung der Fehlstellen als Kovariable und den Vergleich von Behandlungen bei unterschiedlichen Anfangsmassen im Tierversuch an. Sie nennen auch die Möglichkeit, Klassifikationsvariable wie unterschiedliche Jahres- oder Standortniveaus als Kovariable heranzuziehen.

Je nach Zielstellung kann bei der Kovarianzanalyse die Waage mehr zur Varianz- oder zur Regressionsanalyse ausschlagen. Beschreiben die Kovariablen einen störender Einfluß (z. B. unterschiedliche Ausgangssituation), den es auszugleichen gilt, dann könnte man sie als „bereinigte" Varianzanalyse ansehen. Eine „bereinigte" Regressionsanalyse wäre dann die Eliminierung des Einflusses von Kovariblen, die Klassifikationsfaktoren sind, um eine gemeinsame Regression auf gleichem Niveau zu rechnen. RASCH u. a. (1973, S. 244)

Eine gute Übersicht über die Anwendung der SAS-Prozedur GLM für varianzanalytische Auswertungen unbalanzierter Daten einschließlich der Kovarianzanalyse geben beispielsweise SEARLE und YEREX[10].

Im Sinne eines Grundkurses soll auch die Kovarianzanalyse nicht in aller Breite und Tiefe dargelegt werden, sondern anhand einfacher Modelle und Beispiele ein Zugang gegeben werden.

Das varianzanalytische Modell einer einfaktoriellen randomisierten Versuchsanlage mit Wiederholung ist $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$.

Die Stufen des Faktors A seien für diese Betrachtung fix. α_i (i = 1, ..., a) ist die Wirkung (Effekt) der i-ten Stufe des Faktors A und ε_{ij} (i = 1, ..., a; j = 1, ..., n) sind unabhängige, zufällige Versuchsfehler mit $N(0,\sigma^2)$. Die Beobachtungswerte y_{ij} sind bekanntlich:

---

[8] MANHART, P. A.: Gestaltung und Evaluation eines computerbasierten Lernprogramms für SPSS aus pädagigisch-psychologischer und software-ergonomischer Perspektive (http://members.aol.com/PManhart1/zulass/praesent1/sld001.htm) , 1998

[9] RASCH, D., G. ENDERLEIN und G. HERRENDÖRFER.: Biometrie. Verfahren, Tabellen, angewandte Statistik Deutscher Landwirtschaftsverlag, Berlin, 1973, S. 244

[10] SEARLE, S. R. and R. P. YEREX: ACO2: SAS GLM. Annotated Computer Output for Analysis of Variance of Unbalanced Data, Cornell University, Ithaca, New York, 1987

Stufen des Faktors A \ Wiederholung	1	2	...	j	...	n
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1n}
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2n}
...	...	...	...	...	...	...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{in}
...	...	...	...	...	...	...
a	y_{a1}	y_{a2}	...	y_{aj}	...	y_{an}

Wird dieses Modell um einen zusätzlichen Einflußfaktor Z , Regressor, erweitert, dann lautet das Modell

$$y_{ij} = \mu + \alpha_i + \beta_i(z_{ij} - \bar{z}_{..}) + \varepsilon_{ij} \qquad \text{(Regressor Z fix)}$$

$$y_{ij} = \mu + \alpha_i + \beta_i(\underline{z}_{ij} - \mu_z) + \underline{\varepsilon}_{ij} \qquad \text{(Regressor } \underline{Z} \text{ zufällig)}$$

β_i ist der Regressionskoeffizient zwischen Y und Z bei der i-ten Behandlung.
Häufig wird bei Kovarianzanalysen davon ausgegangen, daß $\beta_i = \beta$ (für alle i, $i = 1, ..., a$). Das muß aber nicht so sein. Die SAS-Prozeduren GLM und MIXED können auch mit verschiedenen β_i eine Auswertung vornehmen. (Für solche Fälle soll hier nur auf LITTELL, MILLIKEN, STROUP und WOLFINGER: SAS® System for Mixed Models, SAS Institute Inc., Cary, NC, USA, 1996 verwiesen werden.)

Als Versuchsergebnisse sind also immer die Paare (y_{ij} , z_{ij}) zu ermitteln. Für das eingangs betrachtete einfaktorielle Modell sind das nunmehr die Paare

Stufen des Faktors A \ Wiederholung	1	2	...	j	...	n
1	(y_{11} , z_{11})	(y_{12} , z_{12})	...	(y_{1j} , z_{1j})	...	(y_{1n} , z_{1n})
2	(y_{21} , z_{21})	(y_{22} , z_{22})	...	(y_{2j} , z_{2j})	...	(y_{2n} , z_{2n})
...	...	...	...	...	...	...
i	(y_{i1} , z_{i1})	(y_{i2} , z_{i2})	...	(y_{ij} , z_{ij})	...	(y_{in} , z_{in})
...	...	...	...	...	...	...
a	(y_{a1} , z_{a1})	(y_{a2} , z_{a2})	...	(y_{aj} , z_{aj})	...	(y_{an} , z_{an})

Ausgangspunkt der Betrachtungen war das einfaktorielle Varianzanalysemodell mit Wiederholung, zu dem ein Regressor hinzugefügt wurde. Zum gleichen Ergebnis kommt man, wenn ausgehend von einem Regressionsmodell Behandlungseffekte zusätzlich zu berücksichtigen sind.

17.2 Beispiel – Einfaktorielle randomisierte Anlage mit einer Kovariablen

Bei der Auswertung eines Fütterungsversuches mit Schweinen (HEINISCH, O.: Biometrie 2, 1974, S. 106 ff) in a = 5 Gruppen (i = 1, ...,5) zu je 6 Tieren (j = 1, ..., 6) wird das Anfangsgewicht der Tiere unter der Annahme, daß die Gewichtszunahme und das Anfangsgewicht korreliert sind, als Kovariable berücksichtigt. Der Fehler 1. Art α wurde mit 0.05 festgesetzt. Die Daten sind:

Tier	y_I	z_I	y_{II}	z_{II}	y_{III}	z_{III}	y_{IV}	z_{IV}	y_V	z_V
1	529	26	505	22	537	22	579	20	565	21
2	508	21	521	25	569	24	619	23	596	22
3	501	20	560	28	499	20	632	25	631	25
4	534	23	516	20	501	21	644	26	667	27
5	510	22	598	25	506	22	638	21	613	22
6	504	25	552	22	600	27	623	21	580	20

y_i: Gewichtszunahme in der i-ten Gruppe

z_i: Anfangsgewicht

Für die Handrechnung sind einige Summen der Abweichungsquadrate SQ und Summen der Abweichungsprodukte SP zu bilden.

Gruppe	Tier	Gewichtszunahme y	Anfangsgewicht z	y^2	z^2	y*z
I	1	529	26	279841	676	13754
	2	508	21	258064	441	10668
	3	501	20	251001	400	10020
	4	534	23	285156	529	12282
	5	510	22	260100	484	11220
	6	504	25	254016	625	12600
II	1	505	22	255025	484	11110
	2	521	25	271441	625	13025
	3	560	28	313600	784	15680
	4	516	20	266256	400	10320
	5	598	25	357604	625	14950
	6	552	22	304704	484	12144
III	1	537	22	288369	484	11814
	2	569	24	323761	576	13656
	3	499	20	249001	400	9980
	4	501	21	251001	441	10521
	5	506	22	256036	484	11132
	6	600	27	360000	729	16200
IV	1	579	20	335241	400	11580
	2	619	23	383161	529	14237
	3	632	25	399424	625	15800
	4	644	26	414736	676	16744
	5	638	21	407044	441	13398
	6	623	21	388129	441	13083
V	1	565	21	319225	441	11865
	2	596	22	355216	484	13112
	3	631	25	398161	625	15775
	4	667	27	444889	729	18009
	5	613	22	375769	484	13486
	6	580	20	336400	400	11600
Summe		16937	688	9642371	15946	389765

Gruppe	$\sum_{j=1}^{6} y_{ij}$	$\sum_{j=1}^{6} z_{ij}$	$\sum_{j=1}^{6} y_{ij}^2$	$\sum_{j=1}^{6} z_{ij}^2$	$\sum_{j=1}^{6} y_{ij} * z_{ij}$	$\sum_{j=1}^{6} y_{ij} * \sum_{j=1}^{6} z_{ij}$
I	3086	137	1588178	3155	70544	422782
II	3252	142	1768630	3402	77229	461784
III	3212	136	1728168	3114	73303	436832
IV	3735	136	2327735	3112	84842	507960
V	3652	137	2229660	3163	83847	500324
Summe	16937	688	9642371	15946	389765	2329682

Subtraktionsglied $Sgl_y = \frac{1}{N}\left(\sum_{i=1}^{5}\sum_{j=1}^{6} y_{ij}\right)^2 = \frac{16937^2}{30} = 9562065{,}633$

$SQ_y(\text{Gesamt}) = \sum_{i=1}^{5}\sum_{j=1}^{6} y_{ij}^2 - Sgl_y = 9642371 - 9562035{,}633 = 80305{,}367$

$SQ_y(A) = \frac{1}{n}\sum_{i=1}^{5}\left(\sum_{j=1}^{6} y_{ij}\right)^2 - Sgl_y = \frac{3086^2 + 3252^2 + 3212^2 + 3735^2 + 3652^2}{6} - 9562065{,}633 = 55129{,}867$

$SQ_y(\text{Rest}) = SQ_y(\text{Gesamt}) - SQ_y(A) = 80305{,}367 - 55129{,}867 = 25175{,}5$

Subtraktionsglied $Sgl_z = \frac{1}{N}\left(\sum_{i=1}^{5}\sum_{j=1}^{6} z_{ij}\right)^2 = \frac{688^2}{30} = 15778{,}133$

$SQ_z(\text{Gesamt}) = \sum_{i=1}^{5}\sum_{j=1}^{6} z_{ij}^2 - Sgl_z = 15946 - 15778{,}133 = 167{,}867$

$SQ_z(A) = \frac{1}{n}\sum_{i=1}^{5}\left(\sum_{j=1}^{6} z_{ij}\right)^2 - Sgl_z = \frac{137^2 + 142^2 + 136^2 + 136^2 + 137^2}{6} - 15778{,}133 = 4{,}200$

$SQ_z(\text{Rest}) = SQ_z(\text{Gesamt}) - SQ_z(A) = 167{,}887 - 4{,}200 = 163{,}667$

Subtraktionsglied $Sgl_{yz} = \frac{1}{N} * \sum_{i=1}^{5}\sum_{j=1}^{6} y_{ij} * \sum_{i=1}^{5}\sum_{j=1}^{6} z_{ij} = 388421{,}867$

$SP_{yz}(\text{Gesamt}) = \sum_{i=1}^{5}\sum_{j=1}^{6} y_{ij} * z_{ij} - Sgl_{yz} = 389765 - 388421{,}867 = 1343{,}133$

$SP_{yz}(A) = \frac{1}{n}\sum_{i=1}^{5}\left(\sum_{j=1}^{6} y_{ij} * \sum_{j=1}^{6} z_{ij}\right) - Sgl_{yz} = \frac{2329682}{6} - 388421{,}867 = -141{,}533$

$SP_{yz}(\text{Rest}) = SP_{yz}(\text{Gesamt}) - SP_{yz}(A) = 1343{,}133 - 141{,}533 = 1484{,}667$

Der bisherige Stand ist:

	SQ_y	SQ_z	SP_{xy}	FG
Gesamt	80305,367	167,867	1343,133	N - 1 = 30 - 1 = 29
A (zwischen den Gruppen)	55129,867	4,200	-141,533	a - 1 = 5 - 1 = 4
Rest (innerhalb der Gruppen)	25175,500	163,667	1484,667	N - a = 30 - 5 = 25

Der Regressionsanteil ist als Streuungsursache zu berücksichtigen. Dementsprechend sind die SQ-Werte anzupassen. Des weiteren entfällt ein Freiheitsgrad auf die Regression, weil für das Beispiel gilt: $\beta_i = \beta$ (i = 1, ..., a).

$SQ_{korr}(A) = SQ_y(A) - \frac{(SP_{yz}(A))^2}{SQ_z(A)} = 55129{,}867 - \frac{(-141{,}533)^2}{4{,}200} = 50360{,}418$

$SQ_{korr}(\text{Rest}) = SQ_y(\text{Rest}) - \frac{(SP_{yz}(\text{Rest}))^2}{SQ_z(\text{Rest})} = 25175{,}500 - \frac{(1484{,}667)^2}{163{,}667} = 11707{,}668$

$FG_{korr}(\text{Rest}) = 25 - 1 = 24$ (verringert um einen Freiheitsgrad)

Mit den beiden Schätzwerten für die Varianzen MQ(A) und MQ(Rest)

$MQ(A) = \frac{SQ_{korr}(A)}{FG(A)} = \frac{50360{,}418}{4} = 12590{,}105$

$MQ(\text{Rest}) = \frac{SQ_{korr}(\text{Rest})}{FG_{korr}(\text{Rest})} = \frac{11707{,}668}{24} = 487{,}820$

ist Testgröße des F-Testes $F(A) = \frac{MQ(A)}{MQ(\text{Rest})} = \frac{12590{,}105}{487{,}820} = 25{,}809$.

Wegen $F_{1-0,05;\ 4,\ 24} = 2,776 < F(A) = 25,809$ muß die Nullhypothese verworfen werden. Es kann folglich angenommen werden, daß zwischen den Gewichtszunahme der Gruppen signifikante Unterschiede bestehen.

Wenn die Mittelwerte miteinander verglichen werden sollen, müssen sie um den Regressionseffekt, der durch die Kovariable hervorgerufen wird, „bereinigt" werden.

Der Regressionskoeffizient ist:

$$b = \frac{SP_{yz}(Rest)}{SQ_z(Rest)} = \frac{1484,667}{163,667} = 9,071$$

Mit der Annahme, daß die Regressionsgerade durch den Schwerpunkt $(\bar{y}_{..}\ ;\bar{z}_{..})$ geht, lassen sich die bisherigen Mittelwerte $\bar{y}_{i.}$ um den Regressionseinfluß korrigieren:

$$\bar{y}_i^* = \bar{y}_{i.} - b*(\bar{z}_{i.} - \bar{z}_{..}) \qquad \text{mit } \bar{z}_{..} = \frac{1}{N}\sum_{i=1}^{5}\sum_{j=1}^{6} z_{ij} = \frac{688}{30} = 22,933$$

Gruppe	$\sum_{j=1}^{6} y_{ij}$	$\bar{y}_{i.} = \frac{1}{6}\sum_{j=1}^{6} y_{ij}$	$\sum_{j=1}^{6} z_{ij}$	$\bar{z}_{i.} = \frac{1}{6}\sum_{j=1}^{6} z_{ij}$	$\bar{y}_i^* = \bar{y}_{i.} - b*(\bar{z}_{i.} - \bar{z}_{..})$
I	3086	**514,33**	137	22,83	**515,24**
II	3252	**542,00**	142	23,67	**535,34**
III	3212	**535,33**	136	22,67	**537,75**
IV	3735	**622,50**	136	22,67	**624,92**
V	3652	**608,67**	137	22,83	**609,57**

Mit diesen Mittelwerten $\bar{y}_i^*$ können nun die bekannten multiplen Mittelwertvergleiche durchgeführt werden.

Die Kovarianztabelle, den/die Regressionskoeffizienten und die korrigierten Mittelwerte lassen sich ganz einfach mit Hilfe der bekannten SAS-Prozeduren GLM oder MIXED berechnen:

```
data heinisch;
  do tier = 1 to 6;
    do gruppe = 1 to 5;
      input gewicht anfang @@;
      output;
    end;
  end;
lines;
529 26  505 22  537 22  579 20  565 21
508 21  521 25  569 24  619 23  596 22
501 20  560 28  499 20  632 25  631 25
534 23  516 20  501 21  644 26  667 27
510 22  598 25  506 22  638 21  613 22
504 25  552 22  600 27  623 21  580 20
;

proc glm;
  class gruppe;
  model gewicht = gruppe anfang
                / ss3
                  solution ;
  means gruppe;
  lsmeans gruppe;
run;
quit;
```

nur die Prüffaktoren sind Klassifikationsfaktoren
zu den Prüffaktoren kommt die **Kovariable** hinzu
(Ausgabe der Quadratsummen Typ III)
zusätzliche Ausgaben wie z. B. Regressionskoeffizient
Ausgabe und Vergleich der **nicht korrigierten** Mittelwerte
Ausgabe und Vergleich der **korrigierten** Mittelwerte

```
General Linear Models Procedure
Class Level Information
Class     Levels     Values
GRUPPE         5     1 2 3 4 5

Number of observations in data set = 30

General Linear Models Procedure
Dependent Variable: GEWICHT

Source                      DF     Sum of Squares    F Value     Pr > F
Model                        5     68597.6983028      28.12      0.0001
Error                       24     11707.6683639
Corrected Total             29     80305.3666667

                    R-Square                 C.V.          GEWICHT Mean
                    0.854211             3.912140            564.566667

Source                      DF       Type III SS    F Value     Pr > F
GRUPPE                       4     57851.0313978      29.65      0.0001
ANFANG                       1     13467.8316361      27.61      0.0001

                                  T for H0:    Pr > |T|   Std Error of
Parameter          Estimate    Parameter=0                  Estimate

INTERCEPT      401.5390360 B          9.93      0.0001     40.43828796
GRUPPE     1   -94.3333333 B         -7.40      0.0001     12.75172557
           2   -74.2260692 B         -5.78      0.0001     12.83262808
           3   -71.8214528 B         -5.63      0.0001     12.75497152
           4    15.3452138 B          1.20      0.2407     12.75497152
           5     0.0000000 B             .           .              .
ANFANG          9.0712831             5.25      0.0001      1.72643196

NOTE: The X'X matrix has been found to be singular and a
      generalized inverse was used to solve the normal
      equations.   Estimates followed by the letter 'B' are
      biased, and are not unique estimators of the parameters.General Linear
Models Procedure

Level of      ---------GEWICHT---------     ----------ANFANG---------
GRUPPE     N      Mean          SD              Mean          SD
1          6   514.333333   13.7501515      22.8333333   2.31660671
2          6   542.000000   34.7735532      23.6666667   2.87518115
3          6   535.333333   41.6589326      22.6666667   2.50333111
4          6   622.500000   23.2271393      22.6666667   2.42212028
5          6   608.666667   36.9034777      22.8333333   2.63944439

General Linear Models Procedure
Least Squares Means

GRUPPE        GEWICHT
               LSMEAN
1          515.240462
2          535.347726
3          537.752342
4          624.919009
5          609.573795
```

$\bar{y}_{..}$

signifikanter Unterschied zwischen den Gruppen

signifikanter Regressionsanteil

Ausgabeteil der Option `solution`

b

Überschreitungswahrscheinlichkeit zum Testen des Regressionskoeffizienten

die nicht korrigierten Mittelwerte

die korrigierten Mittelwerte

„Automatisch" getestet wird der Regressionskoeffizient (bzw. bei mehreren: die Regressionskoeffizienten) mit der Nullhypothese H_0: $\beta = 0$.

Das ist besonders dann interessant, wenn das Hauptaugenmerk auf dem Teil der Regression bei der Kovarianzanalyse liegt.

Mittelwertvergleiche werden – wie bekannt – über die `means`- bzw. `lsmeans`-Anweisung realisiert.

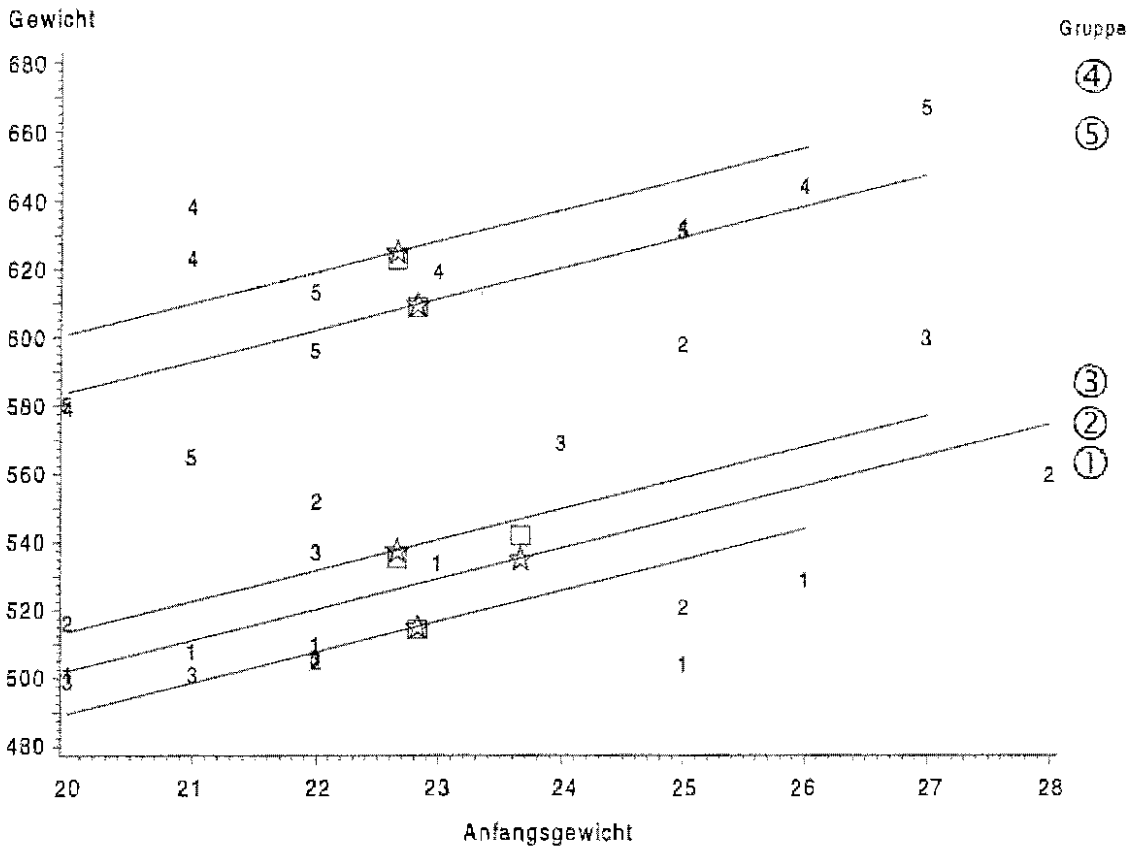
Beide Anweisungen

```
means gruppe / tukey;
lsmeans gruppe /pdiff adjust=tukey;
```

führen beispielsweise die Tukey-Testprozedur durch.

Aber: einmal mit den nicht korrigierten (`means`) und zum anderen mit den korrigierten (`lsmeans`) Mittelwerten!

Die folgende Darstellung veranschaulicht die Einzelwerte mit ihrer Gruppennummer (1, 2, ..., 5), die fünf Regressionsgeraden, die aufgrund der Annahme $\beta_i = \beta$ (für alle i, i = 1, ..., a) denselben Anstieg haben, die unkorrigierten $(\bar{z}_{i\bullet}\,;\bar{y}_{i\bullet})$ [`means` - Symbol: Quadrat] und die korrigierten (bereinigten) Mittelwerte $(\bar{z}_{i\bullet}\,;\bar{y}_{i}^{*})$ [`lsmeans` - Symbol: Stern]. Die jeweiligen Einzelwerte sind mit ihren Gruppennummern dargestellt.



Es gibt seitens der Aufgabenstellung kein Hinweis dafür, daß die Modellannahme $\beta_i = \beta$ (für alle i, i = 1, ..., a) nicht zugrunde gelegt werden kann, obwohl die linearen Regressionsgleichungen für die einzelnen Gruppen hinsichtlich des Anstieges Unterschiede aufweisen:

Gruppe 1:	Gewicht =	445,975 +	2,994	* Anfangsgewicht
Gruppe 2:	Gewicht =	390,266 +	6,411	* Anfangsgewicht
Gruppe 3:	Gewicht =	175,319 +	15,883	* Anfangsgewicht
Gruppe 4:	Gewicht =	481,864 +	6,204	* Anfangsgewicht
Gruppe 5:	Gewicht =	307,354 +	13,196	* Anfangsgewicht

Sollte in einer Aufgabenstellung von unterschiedlichen Regressionskoefizienten ausgegangen werden müssen, so kann das auch in den Modellanweisungen der SAS-Prozeduren GLM und MIXED realisiert werden.

17.3 Beispiel – Einfaktorielle Blockanlage mit Fehlstellenanzahl als Kovariable

Die Kovariable Fehlstellenanzahl beeinflußt die Erträge. Das Ziel ist das Ausschalten dieses Störeinflusses. Betrachtet wird ein Sortenversuch (Sorten: fix) mit Zuckerrüben (58 Normalrüben je Parzelle). Die Versuchsanlage ist eine einfaktorielle Blockanlage mit a = 5 Sorten und b = 4 Blocks[11]. Die Erträge (y) in dt/ha und die Fehlstellenanzahl (z) (Regressor: zufällig) sind:

Sorte	1		2		3		4		5	
Block	y	z	y	z	y	z	y	z	y	z
a	345	16	224	25	411	15	389	15	312	12
b	401*	26	257	13	378	24	290	14	312	15
c	344	6	312	5	488	3	488	2	356	7
d	323	11	312	2	356	18	345	12	290	11

* Das Beispiel enthält im Buch leider neben dem gekennzeichneten Schreibfehler weitere und zusätzlich Rechenfehler.

Das Ergebnis mit SAS lautet:

```
data fehl;
  do block =1 to 4;
    do sorte = 1 to 5;
      input ertrag fehlstel @@;
      output;
    end;
  end;
lines;
345   16   224   25   411   15   389   15   312   12
401   26   257   13   378   24   290   14   312   15
344    6   312    5   488    3   488    2   356    7
323   11   312    2   356   18   345   12   290   11
;
```

① PROC GLM

```
proc glm data=fehl;
  class sorte block;
  model ertrag = block sorte fehlstel
                 / ss3 solution;
  means sorte;
  lsmeans sorte / pdiff adjust=tukey;
run;
```

② PROC MIXED

```
proc mixed data=fehl;
  class sorte block;
  model ertrag = block sorte fehlstel / solution;
  lsmeans sorte / pdiff adjust=tukey;
run;
```

[11] RASCH, D., G. ENDERLEIN und G. HERRENDÖRFER.: Biometrie. Verfahren, Tabellen, angewandte Statistik Deutscher Landwirtschaftsverlag, Berlin, 1973, S. 244 ff

① PROC GLM

```
General Linear Models Procedure
Dependent Variable: ERTRAG
Source                      DF      Sum of Squares    F Value     Pr > F
Model                        8       62066.1457794       3.74     0.0233
Error                       11       22812.4042206
Corrected Total             19       84878.5500000

                     R-Square                 C.V.          ERTRAG Mean
                     0.731235            13.13705           346.650000

Source                      DF       Type III SS    F Value     Pr > F
BLOCK                        3       6875.2151414       1.11     0.3883
SORTE                        4      44356.9011547       5.35     0.0122
FEHLSTEL                     1       1918.2957794       0.92     0.3568

                                   T for H0:    Pr > |T|   Std Error of
Parameter            Estimate   Parameter=0                 Estimate

INTERCEPT        319.0228743 B         8.53      0.0001    37.41779875
BLOCK      1      25.0997535 B         0.78      0.4537    32.31820589
           2      20.8755391 B         0.60      0.5587    34.62034137
           3      57.3278497 B         1.75      0.1082    32.78922716
           4       0.0000000 B            .         .             .
SORTE      1      44.2584720 B         1.33      0.2119    33.39449309
           2     -41.2500000 B        -1.28      0.2265    32.20135812
           3      99.8662200 B         2.98      0.0126    33.56742692
           4      59.2845040 B         1.84      0.0929    32.22614937
           5       0.0000000 B            .         .             .
FEHLSTEL          -2.4309920           -0.96      0.3568     2.52763831

NOTE: The X'X matrix has been found to be singular and a
      generalized inverse was used to solve the normal
      equations.  Estimates followed by the letter 'B' are
      biased, and are not unique estimators of the parameters.
```

b

die nicht korrigierten Mittelwerte

```
General Linear Models Procedure
Level of      ----------ERTRAG---------    ---------FEHLSTEL--------
SORTE      N       Mean            SD            Mean            SD

1          4    353.250000    33.4103277     14.7500000     8.5391256
2          4    276.250000    43.4233040     11.2500000    10.2753751
3          4    408.250000    57.7718213     15.0000000     8.8317609
4          4    378.000000    83.7735042     10.7500000     5.9651767
5          4    317.500000    27.6827263     11.2500000     3.3040379
```

die korrigierten Mittelwerte mit den Überschreitungswahrscheinlichkeiten der Tukey-Prozedur

```
General Linear Models Procedure
Least Squares Means
Adjustment for multiple comparisons: Tukey-Kramer

SORTE          ERTRAG     Pr > |T| H0: LSMEAN(i)=LSMEAN(j)
               LSMEAN     i/j      1        2        3        4        5

1          358.476633     1     .        0.1461   0.4577   0.9907   0.6825
2          272.968161     2   0.1461     .        0.0103   0.0602   0.7074
3          414.084381     3   0.4577     0.0103   .        0.7540   0.0761
4          373.502665     4   0.9907     0.0602   0.7540   .        0.4003
5          314.218161     5   0.6825     0.7074   0.0761   0.4003   .
```

② PROC MIXED

The MIXED Procedure

Class Level Information

Class	Levels	Values
SORTE	5	1 2 3 4 5
BLOCK	4	1 2 3 4

Covariance Parameter Estimates (REML)

Cov Parm	Estimate
Residual	2073.8549291

Model Fitting Information for ERTRAG

Description	Value
Observations	20.0000
Res Log Likelihood	-65.6908
Akaike's Information Criterion	-66.6908
Schwarz's Bayesian Criterion	-66.8897
-2 Res Log Likelihood	131.3815

Ausgabeteil der Option solution

Solution for Fixed Effects

Effect	SORTE	BLOCK	Estimate	Std Error	DF	t	Pr > |t|
INTERCEPT			319.02287431	37.41779875	11	8.53	0.0001
BLOCK		1	25.09975354	32.31820589	11	0.78	0.4537
BLOCK		2	20.87553913	34.62034137	11	0.60	0.5587
BLOCK		3	57.32784966	32.78922716	11	1.75	0.1082
BLOCK		4	0.00000000	.	.	.	.
SORTE	1		44.25847197	33.39449309	11	1.33	0.2119
SORTE	2		-41.25000000	32.20135812	11	-1.28	0.2265
SORTE	3		99.86621996	33.56742692	11	2.98	0.0126
SORTE	4		59.28450400	32.22614937	11	1.84	0.0929
SORTE	5		0.00000000	.	.	.	.
FEHLSTEL			-2.43099199	2.52763831	11	-0.96	0.3568

b

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
BLOCK	3	11	1.11	0.3883
SORTE	4	11	5.35	0.0122
FEHLSTEL	1	11	0.92	0.3568

um die Regression bereinigte Mittelwerte

Least Squares Means

Effect	SORTE	LSMEAN	Std Error	DF	t	Pr > |t|
SORTE	1	358.47663278	23.40932888	11	15.31	0.0001
SORTE	2	272.96816081	23.02406575	11	11.86	0.0001
SORTE	3	414.08438078	23.56404285	11	17.57	0.0001
SORTE	4	373.50266482	23.24499800	11	16.07	0.0001
SORTE	5	314.21816081	23.02406575	11	13.65	0.0001

Überschreitungswahrscheinlichkeit des t-Testes

Überschreitungswahrscheinlichkeit der Tukey-Prozedur

Differences of Least Squares Means

Effect	SORTE	_SORTE	Difference	Std Error	DF	t	Pr > |t|	Adjustment	Adj P
SORTE	1	2	85.50847197	33.39449309	11	2.56	0.0265	Tukey-Kramer	0.1461
SORTE	1	3	-55.60774800	32.20755772	11	-1.73	0.1122	Tukey-Kramer	0.4577
SORTE	1	4	-15.02603204	33.75130740	11	-0.45	0.6648	Tukey-Kramer	0.9907
SORTE	1	5	44.25847197	33.39449309	11	1.33	0.2119	Tukey-Kramer	0.6825
SORTE	2	3	-141.1162200	33.56742692	11	-4.20	0.0015	Tukey-Kramer	0.0103
SORTE	2	4	-100.5345040	32.22614937	11	-3.12	0.0098	Tukey-Kramer	0.0602
SORTE	2	5	-41.25000000	32.20135812	11	-1.28	0.2265	Tukey-Kramer	0.7074
SORTE	3	4	40.58171596	33.94595663	11	1.20	0.2570	Tukey-Kramer	0.7540
SORTE	3	5	99.86621996	33.56742692	11	2.98	0.0126	Tukey-Kramer	0.0761
SORTE	4	5	59.28450400	32.22614937	11	1.84	0.0929	Tukey-Kramer	0.4003

17.4 Beispiel – randomisierte Anlage mit zwei Kovariablen

HAYS (1988[12], S. 757 ff) wertet sehr geringe Wirkungen offensichtlich nicht betrunken machender Alkoholmengen auf das Lernverhalten aus. 40 Erstsemestler wurden zufällig ausgewählt und zufällig auf 4 Gruppen aufgeteilt. Jede dieser Gruppe hatte eine Liste von 100 seltenen technischen Begriffen und deren Definitionen zu lernen. Motiviert wurden die Teilnehmer durch eine leistungsabhängige Bezahlung ihrer Lernergebnisse. Eine Nahrungsaufnahme drei Stunden vor und während der Untersuchung wurde unterbunden.

Gruppe A_1: kein Alkohol,
Gruppe A_2: eine Einheit Alkohol,
Gruppe A_3: zwei Einheiten Alkohol,
Gruppe A_4: drei Einheiten Alkohol.

Die beiden Kovariablen waren der IQ und das Gewicht in Pfund, wobei beide Variablen linear transformiert wurden:

X_1 = IQ – 120
X_2 = Gewicht – 125.

Die Kovariablen X_1, X_2 und die Fehler-Scores Y sind:

A_1			A_2			A_3			A_4		
X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	Y
11	13	8	-3	31	17	2	36	21	7	17	19
5	15	11	7	29	12	6	34	17	4	26	31
-4	19	9	21	12	19	8	18	18	4	35	15
6	25	7	-15	25	13	12	28	22	6	43	24
18	26	10	6	17	29	-4	26	33	-8	36	27
-2	23	6	8	16	24	2	51	14	10	22	23
5	33	14	11	42	14	-6	30	26	14	19	21
-3	18	12	-2	30	23	5	23	25	-6	21	20
4	25	13	-8	26	16	11	20	23	4	20	25
9	38	17	13	35	20	12	15	21	5	11	27

SAS-DataSet

```
data hays;
  do a = 1 to 4;
    input x1 x2 y @@;
    output;
  end;
lines;
11  13   8  -3  31  17   2  36  21   7  17  19
 5  15  11   7  29  12   6  34  17   4  26  31
-4  19   9  21  12  19   8  18  18   4  35  15
 6  25   7 -15  25  13  12  28  22   6  43  24
18  26  10   6  17  29  -4  26  33  -8  36  27
-2  23   6   8  16  24   2  51  14  10  22  23
 5  33  14  11  42  14  -6  30  26  14  19  21
-3  18  12  -2  30  23   5  23  25  -6  21  20
 4  25  13  -8  26  16  11  20  23   4  20  25
 9  38  17  13  35  20  12  15  21   5  11  27
;


proc glm;
  class a;
  model y = a x1 x2 / ss3 solution;
  means a;
  lsmeans a;
run;
```

[12] HAYS, W. L.: Statistics, Holt, Rinchart and Winston, Inc., 1988

```
General Linear Models Procedure
Dependent Variable: Y
Source                      DF     Sum of Squares    F Value     Pr > F
Model                        5      980.46379234       8.48     0.0001
Error                       34      786.63620766
Corrected Total             39     1767.10000000

                   R-Square             C.V.              Y Mean
                   0.554843         25.79103          18.6500000

Source                      DF       Type III SS    F Value     Pr > F
A                            3      976.62233731      14.07     0.0001
X1                           1        1.44801703       0.06     0.8040
X2                           1       29.07645000       1.26     0.2701

                                  T for H0:    Pr > |T|   Std Error of
Parameter            Estimate     Parameter=0               Estimate
INTERCEPT        25.72343959 B          9.43      0.0001    2.72899487
A          1    -12.62181315 B         -5.85      0.0001    2.15607802
           2     -4.37937240 B         -2.03      0.0499    2.15395242
           3     -0.87920196 B         -0.40      0.6880    2.17101514
           4      0.00000000 B             .           .             .
X1               -0.02595897          -0.25      0.8040    0.10376433
X2               -0.09678415          -1.12      0.2701    0.08633384

NOTE: The X'X matrix has been found to be singular and a
      generalized inverse was used to solve the normal
      equations.   Estimates followed by the letter 'B' are
      biased, and are not unique estimators of the parameters.

General Linear Models Procedure
Level of      ----------Y----------  ----------X1----------  ----------X2----------
A          N      Mean          SD        Mean          SD        Mean          SD
1         10  10.7000000  3.40098025  4.90000000  6.8060431  23.5000000  7.7781746
2         10  18.7000000  5.41705127  3.80000000 10.7372664  26.3000000  9.2141196
3         10  22.0000000  5.31245915  4.80000000  6.3560994  28.1000000 10.5140118
4         10  23.2000000  4.63800723  4.00000000  6.6164777  25.0000000  9.9554563

General Linear Models Procedure
Least Squares Means

A              Y
           LSMEAN
1     10.4982837
2     18.7407245
3     22.2408949
4     23.1200969
```

Die signifikanten Unterschiede zwischen den Gruppen sind klar erkennbar. Der (lineare) Regressionseinfluß beider Kovariablen ist nicht signifikant (s. S. 23). Die Überschreitungswahrscheinlichkeiten sind 0,80 bzw. 0,27. Das hat zur Folge, daß der Unterschied zwischen den nicht um den Einfluß der Regression mit den beiden Kovariablen X_1 und X_2 bereinigten und den bereinigten Mittelwerten nicht so groß ist.

Die korrigierten (besser: adjustierten) Mittelwerte ergeben sich aus:

$$\bar{y}_i^* = \bar{y}_{i\bullet} - b_1 * (\bar{x}_{1i\bullet} - \bar{x}_{1\bullet\bullet}) - b_2 * (\bar{x}_{2i\bullet} - \bar{x}_{2\bullet\bullet})$$

mit

$b_1 = -0{,}026$

und $b_2 = -0{,}097$.

Lösungen

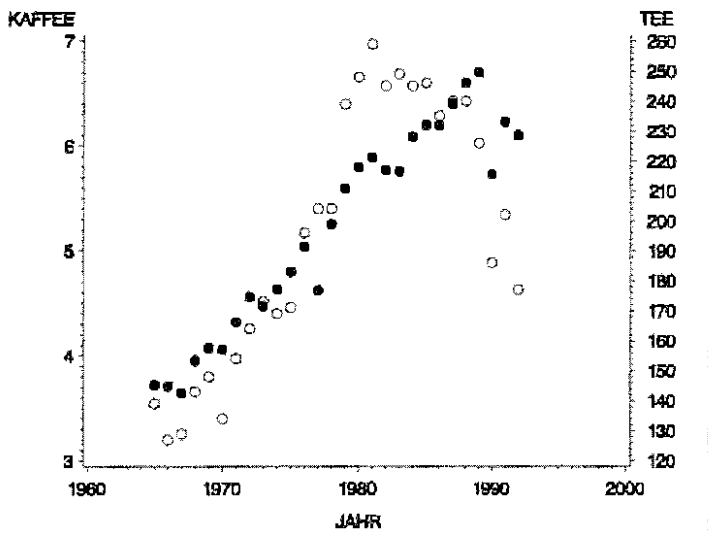
14 Korrelationsanalyse

Aufgabe 14.1: Die Daten des jährlichen Genußmittelverbrauchs je Einwohner[13] von Kaffe (in kg) und Tee (in g) sollen hinsichtlich einer (linearen) Abhängigkeit zwischen den Variablen untersucht werden. Der Korrelationskoeffizient und das 95%-Konfidenzintervall sind zu berechnen. Bei $\alpha = 0.05$ ist zu testen, ob für den geschätzten Korrelationskoeffizient angenommen werden kann, daß er von Null verschieden ist.

Jahr	Kaffee	Tee	Jahr	Kaffee	Tee	Jahr	Kaffee	Tee	Jahr	Kaffee	Tee
1965	3.72	139	1966	3.71	127	1967	3.64	129	1968	3.95	143
1969	4.07	148	1970	4.06	134	1971	4.32	154	1972	4.56	164
1973	4.47	173	1974	4.63	169	1975	4.80	171	1976	5.04	196
1977	4.62	204	1978	5.25	204	1979	5.59	239	1980	5.80	248
1981	5.89	259	1982	5.77	245	1983	5.76	249	1984	6.09	245
1985	6.20	246	1986	6.20	235	1987	6.40	240	1988	6.60	240
1989	6.70	226	1990	5.73	186	1991	6.23	202	1992	6.10	177

Lösung:

Die Berechnung des Korrelationskoeffizienten setzt Linearität voraus. Aus diesem Grunde werden die Daten zuerst grafisch dargestellt. In der Abb. L14.1 ist der Jahresverbrauch beider Genußmittel dargestellt. Um den Zusammenhang der beiden Größen zu erfassen, ist dieses Bild

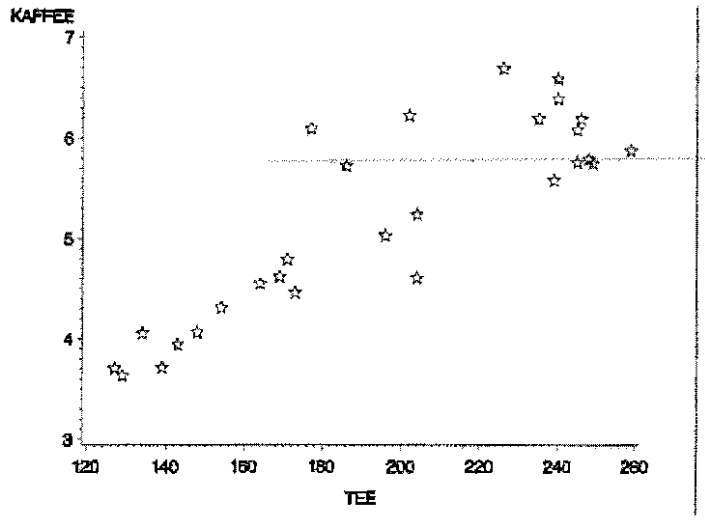



```
goptions htext=1.6 ftext=swiss;
symbol1 c=black h=1.8 v=dot
        i=none;
symbol2 c=black h=1.8 v=circle
        i=none;
proc gplot data=bsp141;
  plot  kaffee*jahr = 1 ;
  plot2 tee    *jahr = 2 ;
run;
quit;
```

Abb. L14.1: Jahresverbrauch von Kaffee und Tee

allerdings ungeeignet. Das kann nur die Darstellung der entsprechenden Wertepaare (Kaffeverbrauch , Teeverbrauch) bzw. (Teeverbrauch , Kaffeeverbrauch) erreichen. Die Abb. L14.2 läßt zwischen den beiden Variablen einen linearen Zusammenhang vermuten, so daß der (lineare) Korrelationskoeffizient als Maß für die Stärke des Zusammenhanges dieser beider (Zufalls-)Variablen berechnet werden kann.

[13] Bundesministerium für Arbeits- und Sozialordnung (Herausgeber): Statistisches Taschenbuch 1994, Arbeits- und Sozialstatistik, Tab. 6.6




```
goptions htext=1.6 ftext=swiss;
symbol1 c=black h=2.2 v="="
        i=none;
proc gplot data=bsp141;
  plot  kaffee * tee = 1 ;
run;
quit;
```

Abb. L14.2: Der (jährliche) Verbrauch von Kaffee und Tee

Schätzwert des Produkt-Momenten-Korrelationskoeffizienten

$$\sum_{i=1}^{n} Kaffee_i = 145{,}9 \qquad \sum_{i=1}^{n} Kaffee_i^2 = 786{,}01$$

$$\sum_{i=1}^{n} Tee_i = 5492 \qquad \sum_{i=1}^{n} Tee_i^2 = 1128894 \qquad \sum_{i=1}^{n} Kaffee_i * Tee_i = 29620{,}28$$

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

$$r = \frac{29620{,}28 - 145{,}9 * 5492/28}{\sqrt{(786{,}01 - 145{,}9 * 145{,}9/28)\,(1128894 - 5492 * 5492/28)}} = 0{,}8692$$

SAS-Programm:

```
proc corr data=bsp141;
  var kaffee tee;
run;
```

```
Correlation Analysis

   2 'VAR' Variables:  KAFFEE   TEE

                        Simple Statistics

Variable     N        Mean      Std Dev          Sum      Minimum      Maximum

KAFFEE      28    5.210714     0.976896   145.900000     3.640000     6.700000
TEE         28  196.142857    43.749044  5492.000000   127.000000   259.000000

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 28

                   KAFFEE              TEE

KAFFEE            1.00000           0.86923
                   0.0               0.0001

TEE               0.86923           1.00000
                   0.0001            0.0
```

Achtung: die Daten wurden mit ihren ungleichen (kg und g) Maßstäben gemeinsam ausgewertet.

Das ist allgemein nicht sinnvoll.

Wird der Kaffeeverbrauch in g umgerechnet, verändern sich einige Maßzahlen, der Korrelationskoeffizient bleibt derselbe.

Correlation Analysis

2 'VAR' Variables: KAFFEE TEE

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
KAFFEE	28	5210.714286	976.895796	145900	3640.000000	6700.000000
TEE	28	196.142857	43.749044	5492.000000	127.000000	259.000000

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 28

	KAFFEE	TEE
KAFFEE	1.00000 0.0	0.86923 0.0001
TEE	0.86923 0.0001	1.00000 0.0

0,95-Konfidenzintervall des Produkt-Momenten-Korrelationskoeffizienten

$n = 28$
$r = 0{,}8692$
$\alpha = 0{,}05$
$u_{1-\alpha/2} = u_{0{,}975} = 1{,}96$

$$z_u = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{u_{1-\alpha/2}}{\sqrt{n-3}} = \frac{1}{2}\ln\left(\frac{1+0{,}8692}{1-0{,}8692}\right) - \frac{1{,}96}{\sqrt{28-3}} = 0{,}9378$$

$$z_o = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + \frac{u_{1-\alpha/2}}{\sqrt{n-3}} = \frac{1}{2}\ln\left(\frac{1+0{,}8692}{1-0{,}8692}\right) + \frac{1{,}96}{\sqrt{28-3}} = 1{,}7218$$

$$\left\langle \frac{e^{2z_u}-1}{e^{2z_u}+1} ; \frac{e^{2z_o}-1}{e^{2z_o}+1} \right\rangle = \left\langle \frac{e^{2*0{,}9378}-1}{e^{2*0{,}9378}+1} ; \frac{e^{2*1{,}7218}-1}{e^{2*1{,}7218}+1} \right\rangle = \langle 0{,}7342 ; 0{,}9381 \rangle$$

Test des Produkt-Momenten-Korrelationskoeffizienten

$r = 0{,}8692$
$n = 28$
$\alpha = 0{,}05$ (zweiseitig)

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}8692 * \sqrt{28-2}}{\sqrt{1-0{,}8692^2}} = 8{,}963 \quad .$$

Da

$t_{berechnet} = 8{,}963 > t_{0{,}05;26} = 2{,}056$ (Tab. 5.4, Teil 1, S. 53)

muß die Nullhypothese (H_0: $\rho = 0$) verworfen werden. Es kann folglich davon ausgegangen werden, daß der Zusammenhang zwischen dem Kaffee- und dem Teeverbrauch signifikant von Null verschieden ist.

Die im SAS-Output angegebene Überschreitungswahrscheinlichkeit von 0,0001 (kleinere Überschreitungswahrscheinlichkeiten rundet SAS auf diesen Wert auf !) belegt ebenfalls den signifikanten Unterschied zu $\rho = 0$.

Aufgabe 15.1: Berechnen Sie für die nachstehenden Daten eine geeignete (α=0,05) Regressionsfunktion und die 0,95-Konfidenzintervalle. Gehen Sie dabei von einem Polynom dritten Grades aus.

2.0	0
2.4	2
3.2	10
5.3	49
7.8	130
9.6	211
12.4	376
15.0	572
18.9	941
21.3	1214

Lösung:

Es soll von einem Polynomansatz dritten Grades ausgegangen werden:

$$y = b_0 + b_1*x + b_2*x^2 + b_3*x^3$$

```
data aufg151;
  input x y;
  x2 = x*x;
  x3 = x*x*x;
lines;
 2.0      0
 2.4      2
 3.2     10
 5.3     49
 7.8    130
 9.6    211
12.4    376
15.0    572
18.9    941
21.3   1214
;
proc reg;
  model y = x x2 x3 ;
  model y = x x2 x3 / selection=rsquare best=1 b;
run;
quit;
```

```
Model: MODEL1
Dependent Variable: Y

Analysis of Variance

                          Sum of          Mean
Source            DF      Squares        Square       F Value       Prob>F

Model              3 1663260.1428 554420.04759  9311723.485       0.0001
Error              6      0.35724      0.05954
C Total            9    1663260.5

    Root MSE          0.24401     R-square        1.0000
    Dep Mean        350.50000     Adj R-sq        1.0000
    C.V.              0.06962

Parameter Estimates

                 Parameter       Standard    T for H0:
Variable  DF      Estimate          Error   Parameter=0    Prob > |T|

INTERCEP   1      1.572791     0.38762822         4.057        0.0067
X          1     -6.900087     0.15899839       -43.397        0.0001
X2         1      2.988895     0.01645089       181.686        0.0001
X3         1      0.000343     0.00047658         0.720        0.4984

  N = 10      Regression Models for Dependent Variable: Y

                          Parameter
Number in      R-square   Estimates
  Model                   Intercept          X          X2          X3

      1     0.99930983     -25.4800          .      2.7034           .
-------------------------------------------------------------------------
      2     0.99999977       1.8013    -7.0087      3.0006           .
-------------------------------------------------------------------------
      3     0.99999979       1.5728    -6.9001      2.9889    0.000343
-------------------------------------------------------------------------
```

Im ersten Teil werden die Koeffizienten für das Polynom dritten Grades geschätzt. Mit $\alpha = 0{,}05$ ist das kubische Glied nicht signifikant. Im zweiten Teil der Ausgabe wird die Verbesserung des multiplen Bestimmtheitsmaßes gezeigt: das kubische Glied trägt nur unwesentlich zur Verbesserung der Anpassung bei. Es ist also hinreichend, von einem Polynom zweiten Grades auszugehen.

```
proc reg data=aufg151;
  model y = x x2 / clm;
run; quit;
```

```
Model: MODEL1
Dependent Variable: Y

Analysis of Variance

                          Sum of          Mean
Source            DF      Squares        Square       F Value       Prob>F

Model              2 1663260.1119 831630.05594 14998654.105       0.0001
Error              7      0.38813      0.05545
C Total            9    1663260.5

    Root MSE          0.23547     R-square        1.0000
    Dep Mean        350.50000     Adj R-sq        1.0000
    C.V.              0.06718
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	1.801268	0.21499236	8.378	0.0001
X	1	-7.008683	0.04871803	-143.862	0.0001
X2	1	3.000638	0.00212409	1412.672	0.0001

Obs	Dep Var Y	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Residual
1	0	-0.2135	0.142	-0.5487	0.1216	0.2135
2	2.0000	2.2641	0.131	1.9555	2.5727	-0.2641
3	10.0000	10.1000	0.112	9.8345	10.3655	-0.1000
4	49.0000	48.9432	0.093	48.7231	49.1632	0.0568
5	130.0	129.7	0.106	129.4	129.9	0.3076
6	211.0	211.1	0.117	210.8	211.3	-0.0567
7	376.0	376.3	0.122	376.0	376.6	-0.2717
8	572.0	571.8	0.115	571.5	572.1	0.1854
9	941.0	941.2	0.131	940.9	941.5	-0.1951
10	1214.0	1213.9	0.194	1213.4	1214.3	0.1242

Sum of Residuals	1.42E-13
Sum of Squared Residuals	0.3881
Predicted Resid SS (Press)	0.8603

Die 0,95-Konfidenzintervalle liegen sehr eng um die Regressionsfunktion. Grafisch wären die Unterschiede kaum sichtbar. Zu erkennen ist aber, daß die Breite der Konfidenzintervalle in den Anfangs- und Endbereichen des Wertevorrates am größten ist:

x	y	untere Grenze	obere Grenze	Differenz
2.0	0	-0.5487	0.1216	0.6703
2.4	2	1.9555	2.5727	0.6172
3.2	10	9.8345	10.3655	0.5310
5.3	49	48.7231	49.1632	0.4401
7.8	130	129.4	129.9	0.5000
9.6	211	210.8	211.3	0.5000
12.4	376	376.0	376.6	0.6000
15.0	572	571.5	572.1	0.6000
18.9	941	940.9	941.5	0.6000
21.3	1214	1213.4	1214.3	0.9000

Korrektur

zu Teil 2 (Heft 31)

S. 103 unteren Teil der Tab. 10.2

Vergleich der Mittelwerte	g	m	c
A	a	a(a-1)/2	$1/\sqrt{bn}$
B	b	b(b-1)/2	$1/\sqrt{an}$
AB auf gleicher Stufe von A	ab	ab(ab-1)/2	$1/\sqrt{n}$
AB auf gleicher Stufe von B			
AB (allgemein)			

ersetzen durch

Vergleich der Mittelwerte	g	m	c
A	a	a(a-1)/2	$1/\sqrt{bn}$
B	b	b(b-1)/2	$1/\sqrt{an}$
AB auf gleicher Stufe von A	b	b(b-1)/2	$1/\sqrt{n}$
AB auf gleicher Stufe von B	a	a(a-1)/2	
AB (allgemein)	ab	ab(ab-1)/2	

Da sich der Fehler auch im Beispiel niedergeschlagen hat, sind die auf den Seiten 104-106 berechneten Grenzdifferenzen und Signifikanzaussagen zu korrigieren.

Ersetzen:

$$HSD_{\alpha;ab}^{\overline{AB}\text{ auf gleicher Stufe von B}} = q_{1-\alpha;ab,FG_{Rest}} * s_{Rest} / \sqrt{n}$$

zu berechnen. Aus der Varianztabelle wird s^2_{Rest} abgelesen: $s^2_{Rest} = MQ_{Rest} = 14{,}167$. Folglich ist $s_{Rest} = 3{,}764$. Mit $a*b = 4*6 = 24$ und $FG_{Rest} = 48$ ist $q_{1-\alpha;ab,FG_{Rest}} = q_{0,95;24,48} = 5{,}451$ (Tab. 8.6 ; SAS-Funktion: `probmc("RANGE",.,0.95,48,24)`). Somit ist

$$HSD_{\alpha;ab}^{\overline{AB}\text{ auf gleicher Stufe von B}} = q_{1-\alpha;ab,FG_{Rest}} * s_{Rest} / \sqrt{n} = 5{,}451*3{,}764/\sqrt{3} = 11{,}85 \quad .$$

durch

$$HSD_{\alpha;b}^{\overline{AB}\text{ auf gleicher Stufe von B}} = q_{1-\alpha;\,b,FG_{Rest}} * s_{Rest}/\sqrt{n}$$

zu berechnen. Aus der Varianztabelle wird s^2_{Rest} abgelesen: $s^2_{Rest} = MQ_{Rest} = 14{,}167$. Folglich ist $s_{Rest} = 3{,}764$. Mit $a*b = 4*6 = 24$ und $FG_{Rest} = 48$ ist $q_{1-\alpha;\,b,FG_{Rest}} = q_{0,95;6,48} = 4{,}197$ (Tab. 8.6 ; SAS-Funktion: `probmc("RANGE",.,0.95,48,6)`). Somit ist

$$HSD_{\alpha;b}^{\overline{AB}\text{ auf gleicher Stufe von B}} = q_{1-\alpha;\,b,FG_{Rest}} * s_{Rest}/\sqrt{n} = 4{,}197 * 3{,}764 /\sqrt{3} = 9{,}12 \quad .$$

Die Konfidenzintervalle der Mittelwertdifferenzen verändern sich ebenso wie das Ergebnis der Methode der Verbindungslinien. Letzeres ist wie folgt zu ersetzen:

	Sorte1	Sorte2	Sorte3	Sorte4	Sorte5	Sorte6
Betrieb	2 4 3 1	3 4 2 1	3 4 2 1	4 3 1 2	4 1 3 2	3 4 2 1
	45 42 41 30	64 58 55 49	43 31 29 28	50 47 40 38	35 33 30 29	34 32 30 29

Korrektur

Zu streichen ist der Satz auf S. 105, denn es gilt nicht:

Da für die Grenzdifferenz des zweifaktoriellen Varianzanalysemodells mit festen Effekten (Modell I) einer vollständigen Kreuzklassifikation mit Wiederholung gilt

$$HSD_{\alpha;ab}^{\overline{AB}\text{ auf gleicher Stufe von B}} = HSD_{\alpha;ab}^{\overline{AB}\text{ auf gleicher Stufe von A}} = HSD_{\alpha;ab}^{\overline{AB}\text{ (allgemein)}} = q_{1-\alpha;\,ab,FG_{Rest}} * s_{Rest} / \sqrt{n} \quad ,$$

ist $HSD_{\alpha;ab}^{\overline{AB}\text{ auf gleicher Stufe von A}} = 11{,}85$.

Richtig ist:

Die Grenzdifferenz $HSD_{\alpha;a}^{\overline{AB}\text{ auf gleicher Stufe von A}}$ ist mit $q_{1-\alpha;\,a,FG_{Rest}} = q_{0,95;4,48} = 3{,}764$

`[probmc("RANGE",.,0.95,48,4)]` :

$$HSD_{\alpha;a}^{\overline{AB}\text{ auf gleicher Stufe von A}} = q_{1-\alpha;\,a,FG_{Rest}} * s_{Rest} / \sqrt{n} = 3{,}764 * 3{,}764 / \sqrt{3} = 8{,}18$$

Die Konfidenzintervalle der Mittelwertdifferenzen verändern sich ebenso wie das Ergebnis der Methode der Verbindungslinien. Letzeres ist wie folgt zu ersetzen:

	Betrieb 1						Betrieb 2						Betrieb 3						Betrieb 4					
Sorte	2	4	5	1	6	3	2	1	4	6	3	5	2	4	3	1	6	5	2	4	1	5	6	3
	49	40	33	30	29	28	55	45	38	30	29	29	64	47	43	41	34	30	58	50	42	35	32	31

Der selbe Fehler setzt sich in der Lösung der Aufgabe 10.2 auf der Seite 154 fort.
Ersetze

$q_{1-\alpha;a*b,FG_{Rest}} = q_{0,95;8,32} = 4.581$ (SAS-Funktion `probmc("RANGE",.,0.95,32,8)` oder Tab. 8.6)
$s_{Rest} = \sqrt{s_{Rest}^2} = \sqrt{0{,}821} = 0{,}906$

ist die Grenzdifferenz der Tukey-Prozedur des zweifaktoriellen Varianzanalysemodells mit festen Effekten (Modell I) einer vollständigen Kreuzklassifikation mit Wiederholung

$$HSD_{\alpha;ab}^{\overline{AB}\text{ auf gleicher Stufe von B}} = HSD_{\alpha;ab}^{\overline{AB}\text{ auf gleicher Stufe von A}} = q_{1-\alpha;\,ab,FG_{Rest}} * s_{Rest} / \sqrt{n} = 4{,}581 * 0{,}906 / \sqrt{3} = 2{,}396 \quad .$$

durch

$q_{1-\alpha;b,FG_{Rest}} = q_{0,95;4,32} = 3{,}832$ (SAS-Funktion `probmc("RANGE",.,0.95,32,4)` oder Tab. 8.6)
$q_{1-\alpha;a,FG_{Rest}} = q_{0,95;4,32} = 3{,}832$ (SAS-Funktion `probmc("RANGE",.,0.95,32,4)` oder Tab. 8.6)
$s_{Rest} = \sqrt{s_{Rest}^2} = \sqrt{0{,}821} = 0{,}906$

sind die Grenzdifferenzen der Tukey-Prozeduren für den Vergleich der mittleren Wirkung der Stufen des Faktors „chemische Behandlung" [auf gleicher Stufe des Faktors „Düngung"] und den Vergleich der mittleren Wirkung der Stufen des Faktors „Düngung" [auf gleicher Stufe des Faktors „chemische Behandlung"] gleich, weil (in diesem Beispiel) a = b = 4:

$$HSD_{\alpha;b}^{\overline{AB}\text{auf gleicher Stufe von A}} = HSD_{\alpha;a}^{\overline{AB}\text{auf gleicher Stufe von B}} = q_{1-\alpha;b,FG_{Rest}} * s_{Rest} / \sqrt{n} = 3{,}832 * 0{,}906 / \sqrt{3} = 2{,}004 \quad .$$

Die Konfidenzintervalle der Mittelwertdifferenzen müssen dementsprechend verändert werden.

Berichte aus der Biologischen Bundesanstalt für Land- und Forstwirtschaft erscheinen seit 1995 in zwangloser Folge.

Heft 35, 1997: Rechtliche Regelungen der Europäischen Union zu Pflanzenschutzmitteln und deren Wirkstoffen (Band A: Richtlinie 91/414/EWG und diesbezüglicher Protokolle) 3. Auflage, Stand: 1. November 1997. Bearbeitet von Dr. Jörg-Rainer Lundehn, 322 S.

Heft 36, 1997: Rechtliche Regelungen der Europäischen Union zu Pflanzenschutzmitteln und deren Wirkstoffen (Band B: Richtlinien, Verordnungen, Entscheidungen und Protokolle zur Wirkstoffprüfung) Stand: 1. November 1997, 3. Auflage. Bearbeitet von Dr. Jörg-Rainer Lundehn, 148 S.

Heft 37, 1997: Zuständigkeiten bei der Prüfung und Zulassung von Pflanzenschutzmitteln und bei der EU-Wirkstoffprüfung. (Stand: Dezember 1997). Bearbeitet von Edelgard Adam, 58 S.

Heft 38, 1997: Inhaltsverzeichnis Amtliche Pflanzenschutzbestimmungen N.F. Band 1, Heft 1 bis Band 63, Heft 5. Bearbeitet von Sigrid von Norsinski, Elke Vogt-Arndt, Richard Voigt, 74 S.

Heft 39, 1998 Wirkstoffdatenblätter zur arbeitsmedizinischen Vorsorgeuntersuchung - Pflanzenschutzmittel -, 1. Folge, Stand: Dezember 1998. Bearbeitet von Dr. Hans-Hermann Schmidt, Dr. Eberhard Hoernicke, Dr. Marion Fathi, Dr. Rudolf Pfeil, 241 S.

Heft 40, 1998: Liste der zugelassenen Pflanzenschutzmittel (Stand: 1. Januar 1998). Bearbeitet von Dr. Achim Holzmann und Andreas Spinti, 69 S.

Heft 41, 1998: 100 Jahre Biologische Bundesanstalt für Land- und Forstwirtschaft – Entwicklung und Organisation des Pflanzenschutzes in Deutschland. Bearbeitet von Dr. Heinrich Brammeier, 296 S.

Heft 42, 1998: 2. BBA-Notifizierer-Konferenz (15./16. Januar 1998). Bearbeitet von Dr. Hartmut Kula und Dr. Jörg-Rainer Lundehn, 193 S.

Heft 43, 1998: Leitlinie: Rückstandsanalysemethoden für die Überwachung, Stand: 21. Juli 1998. Bearbeitet von Dr. Ralf Hänel und Dr. Johannes Siebers.

Heft 44, 1998: Tagungsband zur Antragstellerkonferenz Braunschweig, 10. Juni 1998. Bearbeitet von Edelgard Adam, 176 S.

Heft 45, 1998: Europäische und nationale Regelungen für gentechnisch veränderte Organismen (GVO) (Richtlinien, Entscheidungen, Empfehlungen, Gesetze, Verordnungen und Bekanntmachungen) Stand: 1. Juli 1998. Bearbeitet von Prof. Dr. Günther Deml, Dr. Joachim Schiemann, Dr. Jörg Landsmann, 305 S.

Heft 46, 1998: Einführung in die Biometrie unter Berücksichtigung der Software SAS. Teil 3: Die Varianzanalyse im Feldversuchswesen. Dr. Eckard Moll, 172 S.

Heft 47, 1998: Zuständigkeiten bei der Prüfung und Zulassung von Pflanzenschutzmitteln und bei der EU-Wirkstoffprüfung. (Stand: September 1998). Bearbeitet von Edelgard Adam, 59 S.

Heft 48, 1999: Tropischer und Subtropischer Pflanzenbau. Seine Entwicklung als Teil der Landbauwissenschaften – am Beispiel der Kagera-Region in Tansania/Ostafrika – eine Kurzdarstellung der tansanischen Landwirtschaft. Dr. Heinrich Brammeier, 82 S.

Heft 49, 1999: Art und Menge der in der Bundesrepublik Deutschland abgegebenen und der exportierten Wirkstoffe in Pflanzenschutzmitteln (1987 – 1997). Ergebnisse aus dem Meldeverfahren nach § 19 des Pflanzenschutzgesetzes. Bearbeitet von Dr. Hans-Hermann Schmidt, Dr. Achim Holzmann, Edeltraut Alisch, 77 S.

Heft 50, 1999: Pflanzenschutzmittel im ökologischen Landbau – Probleme und Lösungsansätze. Erstes Fachgespräch am 18. Juni 1998 in Kleinmachnow - Pflanzenstärkungsmittel – Elektronenbehandlung - . Bearbeitet von Dr. Holger Beer und Dr. Marga Jahn, 76 S.

Heft 51, 1999: Wirkstoffdatenblätter zur arbeitsmedizinischen Vorsorgeuntersuchung - Pflanzenschutzmittel - . 2. Folge, Stand. Dezember 1998. Bearbeitet von Dr. Hans-Hermann Schmidt, Dr. Eberhard Hoernicke, Dr. Marion Fathi, Dr. Rudolf Pfeil, 239 S.

Heft 52, 1999: Liste der zugelassenen Pflanzenschutzmittel (Stand: 1. Januar 1999). Bearbeitet von Dr. Achim Holzmann und Andreas Spinti, 63 S.

Heft 53, 1999: Pflanzenschutz im ökologischen Landbau – Probleme und Lösungsansätze. Zweites Fachgespräch am 5. November 1998 in Darmstadt. Die Anwendung kupferhaltiger Pflanzenschutzmittel, ihre Auswirkungen auf den Naturhaushalt und Erörterung der Möglichkeiten, unerwünschte Auswirkungen zu begrenzen. Bearbeitet von Dr. Marga Jahn und Dr. Holger Beer, 85 S.

Heft 54, 1999: Verzeichnis der Wirkstoffe in zugelassenen Pflanzenschutzmitteln (ehemals Merkblatt Nr. 20). Stand: Juli 1999. Bearbeitet von Dr. Walter Dobrat, 265 S.

Heft 55, 2000: Liste der zugelassenen Pflanzenschutzmittel (Stand: 1. Januar 2000). Bearbeitet von Dr. Achim Holzmann, 88 S.