

Claudia Kegel

**„Analyse eines mehrdimensionalen  
Datensatzes unterstützt durch Xplo-  
Re“**

Hausarbeit zur Vorlesung Einführung in die Softwareumgebung XploRe im  
WS1998/99

## **Inhaltsverzeichnis**

1. EINLEITUNG	1
1.1 Die Daten	1
1.2 Ziel der Analyse und angewandte Methoden	1
2. DATENANALYSE	2
2.1 Univariate Datenanalyse	2
2.1.1 Gegenüberstellung Liberal Art Colleges vs. Universities	6
2.2 Bivariate Datenanalyse	8
2.2.1 Abhängigkeit zwischen den Variablen	8
2.2.2 Regressionsanalyse	11
3. VARIANZANALYSE	14
3.1 Ziel	14
3.2 Annahmen	14
3.3 Die Struktur	16
3.4 Die Hypothesen	16
3.5 Die Teststatistik	17
4. ZUSAMMENFASSUNG	18

# 1. EINLEITUNG

## 1.1 Die Daten

In der vorliegenden Arbeit wird unter Zuhilfenahme der statistischen Software XploRe ein Datensatz analysiert. Er trägt den Namen „College“<sup>1</sup>. Anhand von 8 Merkmalen werden 50 amerikanische Colleges verglichen. Die vorliegenden Daten wurden erhoben an je 25 erstklassigen Liberal Arts Colleges und Universitäten. Dabei wurden folgende 8 Variablen untersucht:

$X_0$ School:	Name jeder Lehranstalt
$X_1$ School-Type:	Art der Schule (Liberal Arts Colleges bzw. Research Universities)
$X_2$ SAT:	Median der SAT-Punkte der Studenten
$X_3$ Acc:	Prozentsatz der akzeptierten Bewerber
$X_4$ \$/Student:	finanzieller Aufwand pro Student in Dollar
$X_5$ Top 10%:	%aler Anteil der Studenten, die sich unter den 10% Besten ihrer High School Graduiertenklassen befanden
$X_6$ %PhD:	%aler Anteil des Lehrkörpers, der über einen PhD-Grad verfügt
$X_7$ Grad%:	Anteil der Studenten an der Einrichtung, die graduieren

Die Variable  $X_0$  dient ausschließlich der Identifikation der einzelnen Beobachtung. Daher wird sie in die weitere Analyse nicht einbezogen, sondern durch die Numerierung der Fälle ersetzt.

## 1.2 Ziel der Analyse und angewandte Methoden

Die Analyse wird unter zwei Aspekten durchgeführt:

- (1) Es wird angenommen, daß die Graduiertenquote eine von den anderen Variablen abhängige Größe ist. Es soll festgestellt, welche Merkmale einen Einfluß auf diese Variable haben. Dieser Einfluß soll durch ein Modell dargestellt werden.

Zur Untersuchung der Abhängigkeit der Variablen  $X_2$  bis  $X_7$  voneinander wird die Korrelation berechnet. Die Modellierung des statistischen Zusammen-

---

<sup>1</sup> entnommen aus der Datenbank <http://lib.stat.cmu.edu/DASL/>

hangs erfolgt über eine Lineare Regressionsanalyse. Das Bestimmtheitsmaß mißt die Güte der (linearen) Regression.

- (2) Hauptziel der Analyse wird es sein, festzustellen, ob sich beide Schulformen in bestimmten Merkmalen signifikant voneinander unterscheiden, d.h., ob die Ausprägung der Variablen School-Type ( $X_1$ ) die anderen Variablen beeinflusst. Dies soll gezeigt anhand der einfaktoriellen Varianzanalyse (ANOVA).

## 2. DATENANALYSE

### 2.1 Univariate Datenanalyse

Im folgenden werden die zu analysierenden Variablen kurz vorgestellt. Es werden die wichtigsten Lage- und Streuungsparameter angegeben. Anhand von Boxplots und Histogramme soll die Verteilung der jeweiligen Variable grafisch dargestellt werden.

#### Nominalskalierte Variable

##### **$X_1$ (School-Type)**

Für die Arbeit in XploRe wurde die Variable numerisch umkodiert. Liberal Art Colleges erhielten die 0, Universities eine 1. Auf die Ermittlung des Modus wurde verzichtet, da er kein sinnvoll interpretierbares Ergebnis liefert. Bei der später durchzuführenden Varianzanalyse wird diese Variable als Faktor verstanden.

#### Metrisch skalierte Variablen

##### **$X_2$ (SAT)**

diskrete Ausprägungen

$x_{\min}$	1109	$x_{\max}$	1400
$\bar{x}$	1264	$x_{0,5}$	1260
$\sigma$	62,33	Schiefe	-0.15455

Der Boxplot Abb. (2.1) zeigt, daß SAT( $X_2$ ) symmetrisch verteilt ist. Dies läßt sich auch am Schiefemaß erkennen, das nahe bei Null liegt. Median und arithmetisches Mittel sind fast identisch und liegen relativ zentral zwischen den beiden Quartilen. Im

Histogramm läßt sich erkennen, daß die Verteilung von  $X_2$  der einer Normalverteilung sehr ähnelt.

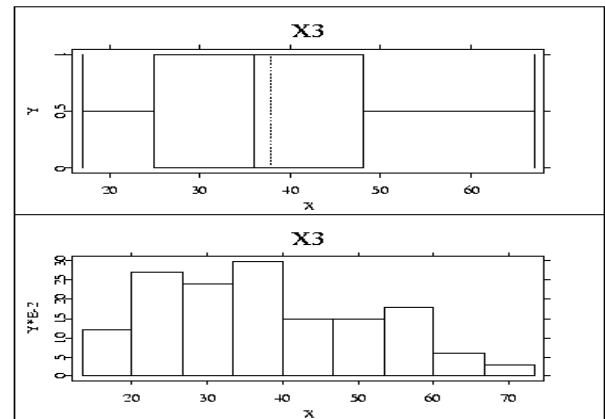
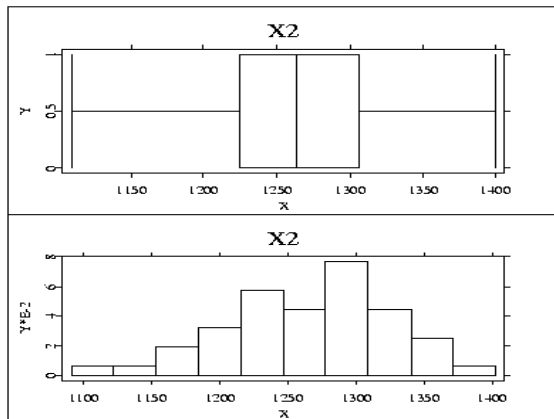


Abb. 2.1 Boxplot und Histogramm für SAT    Abb.2.2 Boxplot und Histogramm für Acc

### $X_3$ (Acceptance)

stetige Ausprägungen in %

$x_{\min}$	17	$x_{\max}$	67
$\bar{x}$	37,84	$x_{0,5}$	36
$\sigma$	13,364	Schiefe	0.3673

Bereits an den Lageparametern läßt sich erkennen, daß die Akzeptanzquote leicht rechtsschief verteilt sein muß. Die Schiefe ist positiv und der Median kleiner als das arithmetische Mittel. Im Boxplot bzw. im Histogramm (Abb. 2.2) ist dies gut zu erkennen. Dennoch weicht die Struktur der Verteilung nicht wesentlich von der einer Normalverteilung ab.

### $X_4$ (\$/Student)

diskrete Ausprägungen

$x_{\min}$	17520	$x_{\max}$	102262
$\bar{x}$	30247	$x_{0,5}$	24718
$\sigma$	15266	Schiefe	2.4159

Die Spannweite ( $x_{\max} - x_{\min}$ ) dieser Variablen weist ein relativ großes Intervall von 84742\$ auf. Dies wird vor allem verursacht durch die extreme Ausprägung von

102262\$, die um mehr als 40000\$ über dem zweithöchsten Wert liegt. Im Boxplot (Abb. 2.3) ist dieser Wert durch einen \* dargestellt und somit als potentieller Ausreißer. Er verzerrt das arithmetische Mittel, das weit über dem Median liegt. \$/Student ( $X_4$ ) ist klar linksschief verteilt (siehe Abb. 2.3). Es läßt sich keine Normalverteilung unterstellen.

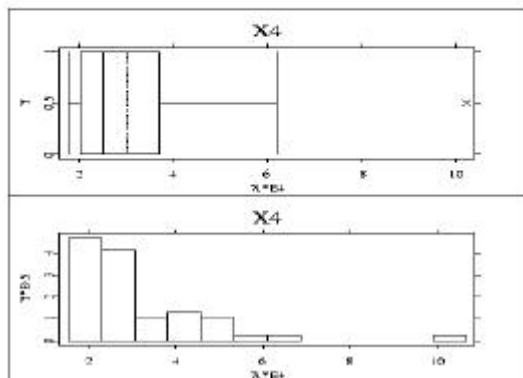


Abb.2.3 Boxplot und Histogramm für \$/Student

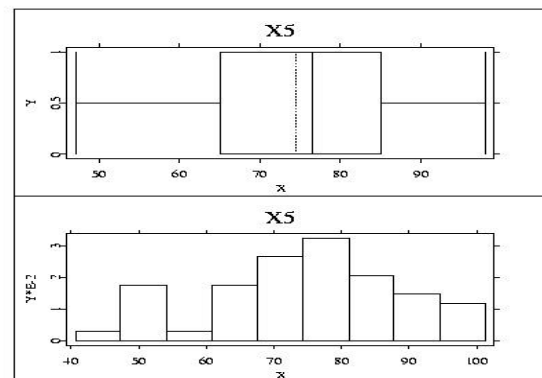


Abb.2.4 Boxplot und Histogramm für Top10%

### $X_5$ (Top10%)

stetige Ausprägungen in %

$X_{\min}$	47	$X_{\max}$	98
$\bar{x}$	74,44	$X_{0,5}$	76
$\sigma$	13,515	Schiefe	-0.23454

Ähnlich wie bei Variable  $X_2$  ist auch diese Variable offensichtlich symmetrisch verteilt. Die Schiefe von  $-0,23454$  deutet eine leicht linksschiefe Verteilung an. Das Histogramm der Abb. 2.4 zeigt, daß die Variable Top10% ( $X_5$ ) annähernd normalverteilt ist.

### $X_6$ (%PhD)

stetige Ausprägungen in %

$X_{\min}$	58	$X_{\max}$	100
$\bar{x}$	90,56	$X_{0,5}$	93
$\sigma$	8,259	Schiefe	-1.4037

Deutlich läßt sich erkennen, daß diese Variable linksschief verteilt ist (siehe Histo-

gramm und Boxplot der Abb. 2.5). Erwartungsgemäß weist dieses Merkmal sehr viele Ausprägungen nahe bei 100 auf. An 75% der Colleges liegt der Anteil der Lehrkräfte mit PhD bei mehr 84%. Bei der Beobachtung Nr. 40, im Boxplot mit einem ° gekennzeichnet, könnte es sich um einen potentiellen Ausreißer handeln. Dieser Eindruck wird verstärkt, zieht man in Betracht, daß die Ausprägungen mit ca. 8% eher gering um den Mittelwert von 90,56% streuen.

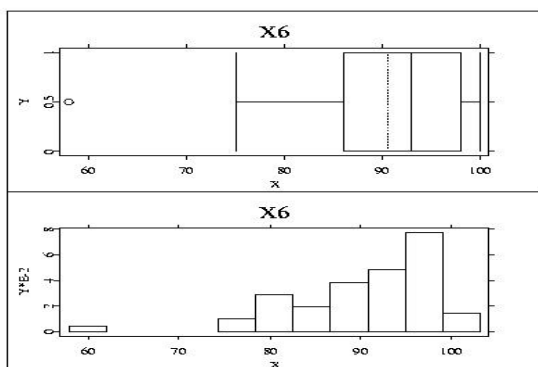


Abb. 2.5 Boxplot und Histogramm für %PhD

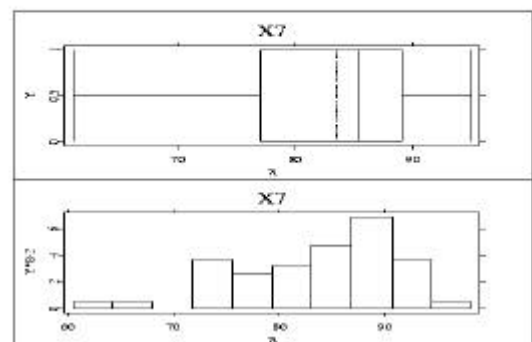


Abb. 2.6 Boxplot und Histogramm für Grad%

### **$X_7$ (Grad%)**

stetige Ausprägungen in %

$X_{\min}$	61	$X_{\max}$	95
$\bar{x}$	83,48	$x_{0,5}$	85
$\sigma$	7,5572	Schiefe	-0.72581

Auch hier liegt keine klare symmetrische Verteilung vor. Das Schiefemaß deutet auf eine linksschiefe Verteilung hin, die sich auch im Histogramm (Abb. 2.6) erkennen läßt. Die Mehrheit der Colleges weist eine hohe Graduiertenquote auf. Unterhalb des Median streuen die Werte weiter als überhalb.

### 2.1.1 Gegenüberstellung Liberal Art Colleges vs. Universities

In der Tab. 2.1 sind die oben für die Variablen untersuchten Lage- und Streuungsparameter, bis auf die Schiefe, getrennt nach Liberal Art Colleges und Universities aufgeführt.

	$X_{\min}$		$X_{\max}$		$\bar{X}$	$X_{0,5}$		$S$		
$X_1$	0	1	0	1	0	1	0	1	0	1
$X_2$	1170	<b>1109</b>	1336	<b>1400</b>	1256,6	<b>1271,3</b>	1255	<b>1280</b>	43,674	<b>76,895</b>
$X_3$	22	<b>17</b>	67	<b>64</b>	40,56	<b>35,12</b>	38	<b>31</b>	12,517	<b>13,875</b>
$X_4$	17520	<b>19365</b>	27879	<b>102262</b>	21756	<b>38739</b>	20377	<b>37137</b>	3455,7	<b>17710</b>
$X_5$	47	<b>52</b>	86	<b>98</b>	67,24	<b>81,64</b>	68	<b>85</b>	10,802	<b>12,175</b>
$X_6$	75	<b>58</b>	98	<b>100</b>	88,24	<b>92,88</b>	90	<b>96</b>	6,6601	<b>9,1484</b>
$X_7$	72	<b>61</b>	93	<b>95</b>	84,12	<b>82,84</b>	85	<b>86</b>	6,0918	<b>8,8679</b>

Tab2.1 Lage- und Streuungsparameter getrennt nach Liberal Art Colleges und Universities

Es fällt auf, daß die Standardabweichungen ( $\sigma$ ) für  $X_1 = 1$  aller Variablen größer sind als die der für  $X_1 = 0$ . Das läßt den Schluß zu, daß sich die Universities untereinander mehr in den einzelnen Merkmalen unterscheiden, als es ihrerseits die Liberal Art Schools tun.

Vergleicht man die Mittelwerte der einzelnen Variablen für die unterschiedlichen Schultypen, ist interessant, daß z.B. das Liberal Art College mit dem höchstem finanziellen Aufwand pro Student ( $X_4$ ), weniger ausgab, als mehr als 50% der Universities. Es läßt sich vermuten, daß in diesem Merkmal deutliche Unterschiede bestehen könnten. Auch die Mittelwerte der Top10%-Quote ( $X_5$ ) differieren offensichtlich. Ebenso wird bei der PhD-Quote der Lehrkräfte ( $X_6$ ) zu prüfen, ob die Differenz von 4% beim arithmetisches Mittel, und sogar 6% beim Median, zufällig oder systematisch ist. Dies soll später anhand der ANOVA analysiert werden.

Mit dem Parallel Coordinate Plot in Abb. 2.7 wird versucht Untergruppen (schwarz = Liberal Art Colleges / rot = Universities) grafisch zu identifizieren. Dabei wurden die Variablen in eine andere Reihenfolge gebracht (1-SAT, 2-Acc, 3-Top10%, 4-%PhD,



5-\$/Student). Anhand der Darstellung lässt sich jedoch keine eindeutige Aussage darüber machen, welche Variable die Schulformen voneinander trennt. Am ehesten kann \$/Student ( $X_4$ ) als eines der trennenden Merkmale erkannt werden: der rote Strahl streut im Plot für 5 sehr weit, während sich der schwarze Strahl immer mehr zusammenzieht.

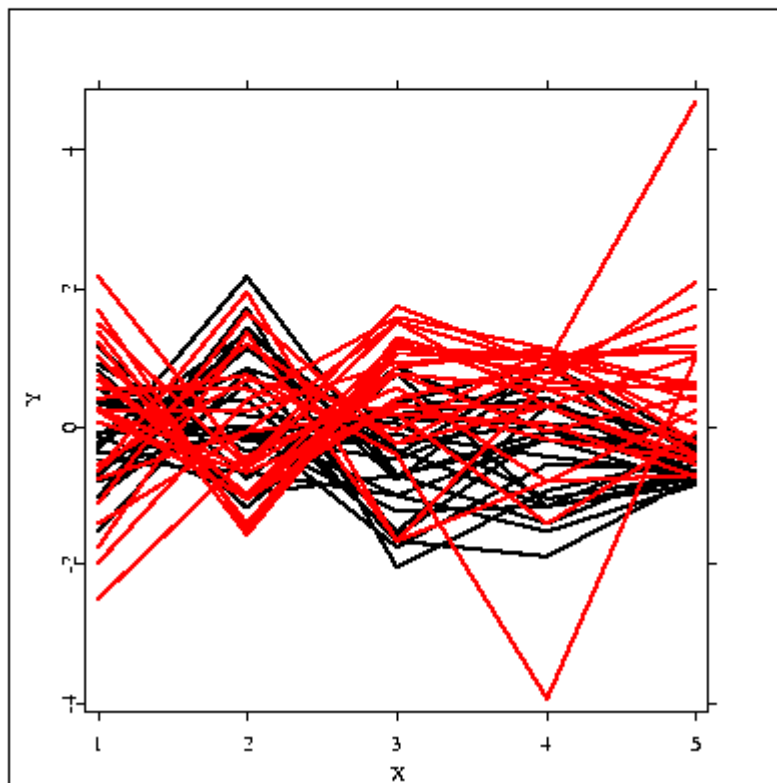


Abb. 2.7 Parallel Coordinate Plot für die Variablen ( $X_2, X_3, X_5, X_6, X_7$ )

## 2.2 Bivariate Datenanalyse

### 2.2.1 Abhängigkeit zwischen den Variablen

In diesem Abschnitt soll die Abhängigkeit der metrischen Variablen untereinander untersucht werden. Definiert man die Variable Grad% als abhängige Variable so kann untersucht werden, welche Variablen die Ausprägung dieser Größe beeinflussen. Es wird erwartet, daß die Graduiertenquote von allen Variablen positiv beeinflusst wird, bis auf die Akzeptanzquote, die negativ wirkt.

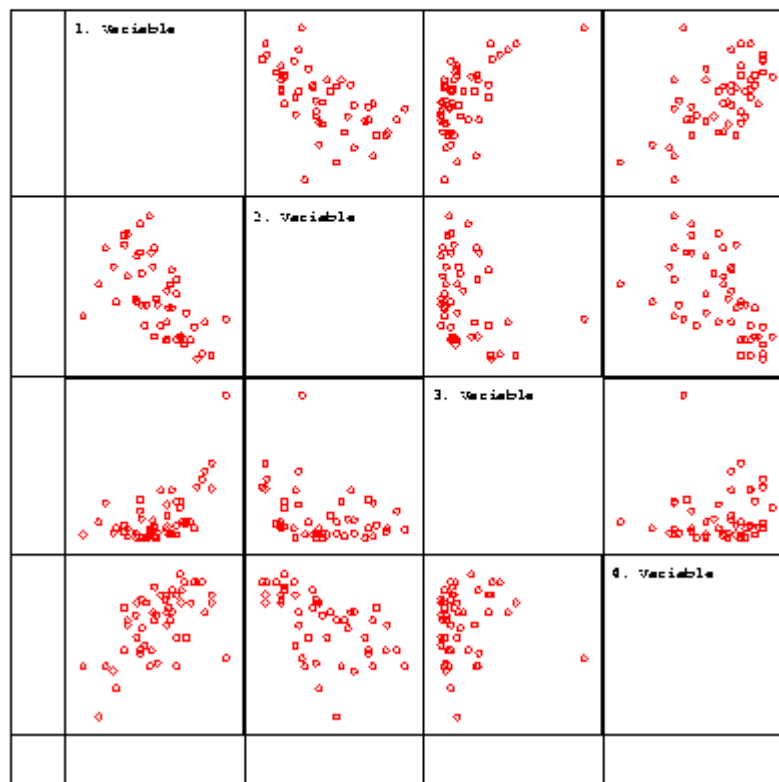
Grafisch läßt sich die Abhängigkeit im mehrdimensionalen Fall durch eine Scatterplotmatrix veranschaulichen. Abb. 2.8 und 2.9 zeigen Punktwolken für verschiedene Variablenkombination des vorliegenden Datensatzes.

Aus der Abb. 2.8 läßt sich für die Variablenpaare SAT ( $X_2$ )/ Grad% ( $X_7$ ) sowie Acc ( $X_3$ )/ Grad% ( $X_7$ ) ein linearer Zusammenhang vermuten. Ebenso sind SAT und Acc stark miteinander korreliert. Für \$/Student ( $X_4$ ) und Grad% ( $X_7$ ) scheint kein eindeutiger Zusammenhang zu bestehen.

In der Abb. 2.9 werden  $X_3$  und  $X_7$  noch einmal gegen Top10% ( $X_5$ ) und %PhD ( $X_6$ ) geplottet. Hier zeigt sich, daß die Graduiertenquote sich mit diesen beiden Variablen nicht sehr gut erklären läßt. Der negative Zusammenhang der Akzeptanzquote zeigt sich auch bei diesen beiden Variablen.

Ein Maß für die Abhängigkeit zwischen zwei Variablen ist die Kovarianz. Sind die Variablen unabhängig voneinander, so ist ihre Kovarianz gleich Null. Andererseits bedeutet eine Kovarianz von Null nicht, daß die Variablen unabhängig sind, denn die Kovarianz mißt nur lineare Zusammenhänge. Über den Befehl `cov(x)` gibt XploRe eine Kovarianzmatrix aus, in der jede mögliche Kovarianz zwischen den verschiedenen Variablen angegeben ist. Ist die Kovarianz positiv kann vermutet werden, daß wenn eine Variable eine höhere Ausprägung annimmt, dies auch die andere tut. Für eine Kovarianz kleiner als Null gilt das umgekehrte.

Abb. 2.8 Scatterplotmatrix (Variablen  $X_2, X_3, X_4, X_7$ )



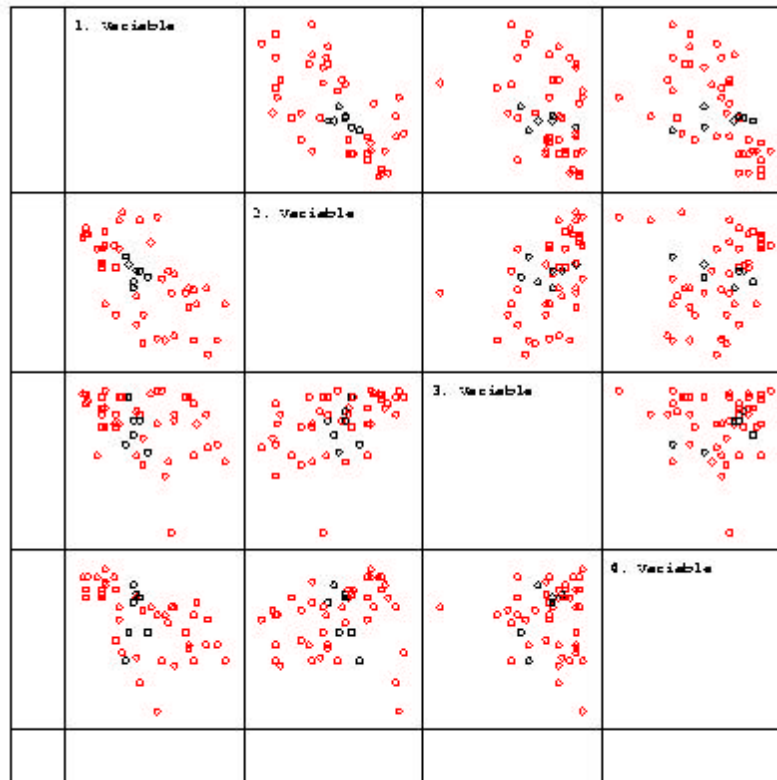


Abb. 2.9 Scatterplotmatrix (Variablen  $X_3$ ,  $X_5$ ,  $X_6$ ,  $X_7$ )

Da die Kovarianz, ebenso wie die Varianz skalenabhängig ist, lassen sich die Kovarianzen untereinander nicht vergleichen. Eine Standardisierung wird erreicht durch den Korrelationskoeffizienten, der zwischen  $-1$  und  $1$  liegt und dasselbe Vorzeichen wie die Kovarianz hat. Die dazugehörige Korrelationsmatrix liefert XploRe mit dem Befehl `corr(x)`.

Für den vorliegenden Datensatz ergibt sich folgende Korrelationsmatrix (Tab. 2.2)

	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	X <sub>5</sub>	X <sub>6</sub>	X <sub>7</sub>
X <sub>2</sub>	1	-0,60675	0,57942	0,50926	0,22094	0,56913
X <sub>3</sub>		1	-0,28882	-0,61633	-0,31166	-0,56221
X <sub>4</sub>			1	0,50995	0,28667	0,051055
X <sub>5</sub>				1	0,44862	0,16134
X <sub>6</sub>					1	-0,055403
X <sub>7</sub>						1

Tab. 2.2 Korrelationsmatrix

Da Acc (X<sub>3</sub>) mit sinkenden Werten offenbar eine höhere Qualität der Bildungseinrichtung bedeutet, ist dieses Merkmal mit allen anderen Variablen, bei denen höhere Werte auch bessere Qualität bedeuten, negativ korreliert.

Auf die Graduiertenquote hat offensichtlich die Anzahl der erreichten SAT-Punkte (X<sub>2</sub>) einen positiven Einfluß, d.h. je höher die durchschnittlich erreichten SAT-Punkte der Studenten sind, desto größer ist die Abschlußquote. Je geringer andererseits die Akzeptanzquote (X<sub>3</sub>) eines Colleges ist, desto mehr Studenten graduieren.

SAT (X<sub>2</sub>) und Acc(X<sub>3</sub>) sind aber auch untereinander stark miteinander korreliert: je höher der Median der SAT-Punkte an einer Uni, desto weniger Bewerber wurden dort akzeptiert. Generell ist für diese beiden Variablen ein deutlicher Einfluß auf alle anderen beobachtbar.

Scheinbar keinen (linearen) Einfluß auf die Graduiertenquote haben die aufgewendeten \$ pro Student (X<sub>4</sub>) und PhD-Quote (X<sub>6</sub>). Sie sind jedoch negativ mit der Akzeptanzquote korreliert, die ihrerseits auf die Graduiertenquote wirkt.

## 2.2.2 Regressionsanalyse

Die Regressionsanalyse ermöglicht es einen linearen Zusammenhang zwischen Variablen zu modellieren. Im bivariaten Fall wird eine abhängige Variable (i.d.R. y) durch eine unabhängige (x) beschrieben. Multivariate Regressionen stellen den linearen Einfluß mehrerer Variablen auf y dar.

Da das Hauptziel der Analyse aber in der Analyse der Unterschiede zwischen Liberal Art Colleges und Universities besteht, soll die Regressionsanalyse hier nur an einem Beispiel gezeigt werden.

Aus dem Scatterplot ist ersichtlich, daß je geringer die Akzeptanzquote eines Colleges ist, desto höher ist die Graduiertenquote. Im linearen Modell kann dieser Zusammenhang folgendermaßen beschrieben werden:

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

$y_i$ : Grad%

$x_i$ : Acc

$\alpha, \beta$ : Parameter

$\varepsilon_i$ : Standardfehler ( $E(\varepsilon) = 0$ )

Mit der Kleinsten-Quadrate-Schätzung lassen sich die Parameter bestimmen. Man erhält:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

wobei unter  $s_{xy}$  bzw.  $s_{xy}$  die empirisch ermittelte Kovarianz bzw. Varianz zu verstehen ist.

Für  $X_3$  und  $X_7$  ergeben sich folgende Parameter<sup>2</sup>:

$$\hat{\alpha} = 95,511$$

$$\hat{\beta} = -0,31793$$

Die beiden Parameter lassen sich wie folgt interpretieren. An einer Universität, an der etwas mehr als 0% der Bewerber akzeptiert werden, ist eine Graduiertenquote von ungefähr 95% zu erwarten. Sinkt die Akzeptanzquote um 1% steigt die Graduiertenquote um 0,32%. In Abb. 2.10 sind die Punktwolke und die Regressionsgerade gra-

---

<sup>2</sup> siehe Makro „Linreg.xpl“ (Diskette)

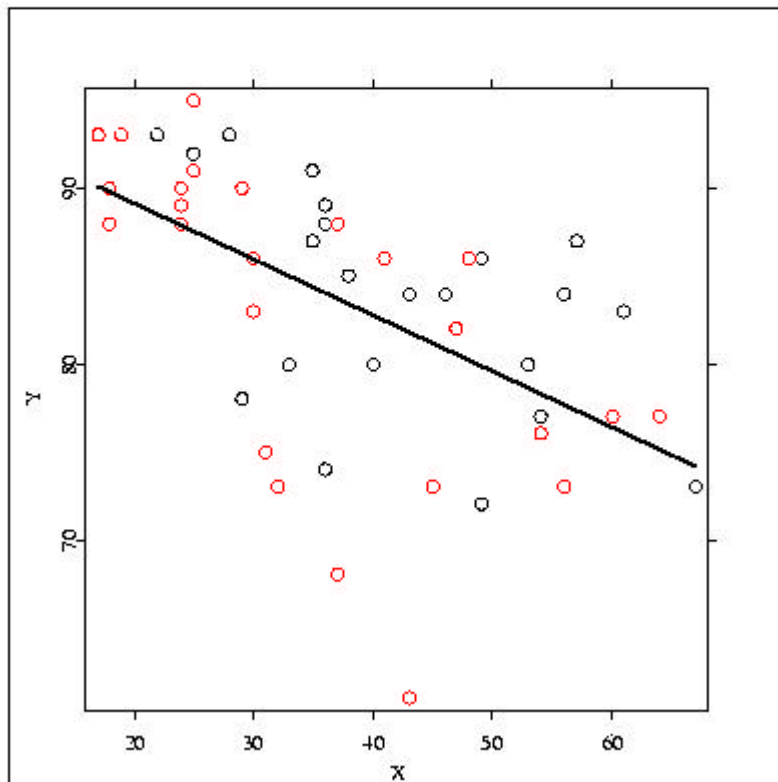


Abb. 2.10 Regressionsgerade und Punktwolke für  $X_3$  und  $X_7$

fisch dargestellt. Die schwarzen Punkte korrespondieren mit den Werten für Liberal Art Colleges, die roten mit den Universities.

Man sieht deutlich, daß vor allem im unteren Bereich des Diagramms die tatsächlichen Beobachtungen sehr stark um die Regressionsgerade streuen. Um festzustellen, wie gut die lineare Regression, den Zusammenhang zwischen  $X_3$  und  $X_7$  anpaßt, kann das Bestimmtheitsmaß  $r^2$  verwendet werden.  $r^2$  gibt an, welcher Teil der Varianz durch die Regression erklärt wird. Im Beispiel erhält man  $r^2 = 0,31608$ , was vermuten läßt, daß ein lineares Modell, den Zusammenhang nicht sehr gut beschreibt.

### 3. VARIANZANALYSE

#### 3.1 Ziel

Hauptziel der Untersuchung ist es festzustellen, ob es wesentliche Unterschiede zwischen Liberal Art Colleges und Universities bestehen. Dazu werden die Variablen  $X_2$  bis  $X_7$  als abhängige Variablen definiert. Einzig unabhängige Variable ist die nominalskalierte Variable  $X_1 = \text{School-Type}$ .

Mit der Varianzanalyse ANOVA soll nun getestet werden, ob die Differenzen in den Mittelwerten zwischen beiden Schulformen signifikant von Null verschieden sind. Das bedeutet, daß die Abweichungen zwischen Liberal Art Colleges und Universities nicht zufällig sondern in der Ausprägung von  $X_1$  begründet sind.

Wir definieren dafür  $X_1$  als Faktor mit  $l=0$  und  $l=1$  als Faktorlevels.

Es wird erwartet, daß sich Liberal Art Colleges und Universities in den Variablen  $\$/\text{Student}$  und  $\text{Top10\%}$  unterscheiden, da hier bereits durch deskriptiven Statistiken Differenzen festgestellt wurden.

#### 3.2 Annahmen

Die Anwendung von ANOVA setzt vier Bedingungen voraus:

- (1) Die abhängigen Variablen sind metrisch skaliert; für den Faktor genügt nominales Skalenniveau.
- (2) Es sind mehr als zwei unabhängige Stichproben mit je  $n$  Stichprobenumfängen gegeben.
- (3) Die Stichproben sind normalverteilt mit dem arithmetisches Mittel  $\mu_i$  und der Varianz  $\sigma_i^2$ .
- (4) Es liegt Varianzhomogenität vor, d.h.  $\sigma_i^2 = \sigma_j^2 = \sigma^2$ .

(1) und (2) sind erfüllt. Für die Varianzhomogenität wird vorausgesetzt, daß die unbekannte Varianz der Grundgesamtheit diesem Anspruch genügt.

Aufgrund der Analyse unter 2.1 werden die Grundgesamtheiten der Variablen  $X_2$ ,  $X_3$  und  $X_5$  hier als normalverteilt angenommen. Da der später verwendete F-Test zu den



robusten Tests zählt, führt er dennoch zu guten Ergebnissen, wenn die Grundgesamtheit nicht genau einer Normalverteilung folgt. Aus diesem Grunde wird hier darauf verzichtet, die exakte Verteilung dieser Variablen zu ermitteln.

Für die Variablen  $X_4$ ,  $X_6$  und  $X_7$  ist die Normalverteilung nicht anzunehmen (vgl. Abb. 2.3, 2.5 und 2.6). Um sie dennoch der Normalverteilung anzunähern, müssen sie transformiert werden.

$$X_{4^*} = 1/X_4$$

$$X_{6^*} = (X_6/100)^{1/2} - (1 - X_6/100)^{1/2}$$

$$X_{7^*} = X_7/100^{1/2} - (1 - X_7/100)^{1/2}$$

Student/\$ (Potenztransformation)

gefaltete Wurzeltransformation

gefaltete Wurzeltransformation

Die Verteilungen der transformierten Variablen sind in den Abbildungen 3.1 bis 3.3 dargestellt.

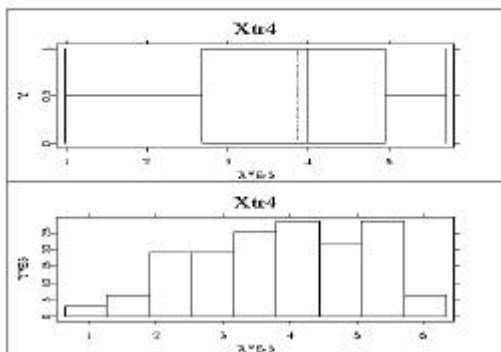


Abb. 3.1 Boxplot und Histogramm der transformierten Variable Student/\$

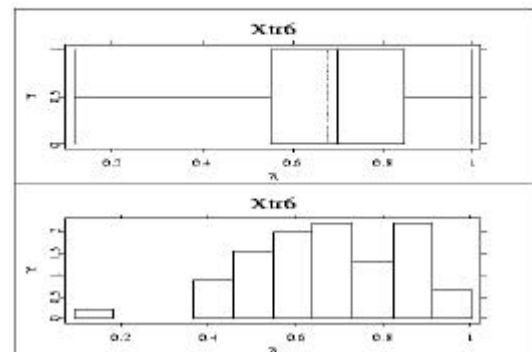


Abb. 3.2 Boxplot und Histogramm der transformierten Variable %PhD\*

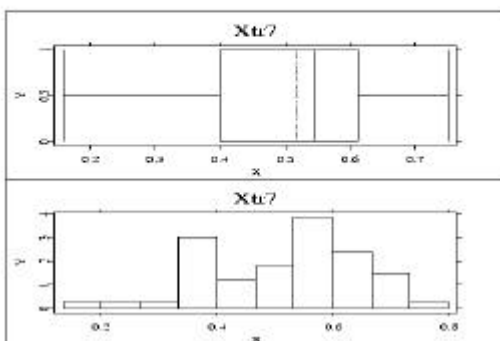


Abb. 3.3 Boxplot und Histogramm der transformierten Variablen Grad%

Vergleicht man die Verteilungen der transformierten Daten mit den Ausgangsdaten in Kapitel 2.1, kann den neuen Variablen eine Normalverteilung unterstellt werden.

### 3.3 Die Struktur

$$x_{ik} = \mu_i + \varepsilon_{ki}$$

$x_{ik}$  : k-te Beobachtung auf dem Level i

$\mu_i$  : Mittelwert der Variable beim Level i

$\varepsilon_{ki}$  : Fehler

Die k-te Beobachtung beim Level i ergibt sich aus dem Mittelwert aller Beobachtungen auf diesem Level und einem Fehler, der im Mittel 0 ist.

Der Datensatz wird nun nicht mehr als eine Stichprobe verstanden, sondern er besteht nun aus zwei verschiedenen Stichproben für Liberal Art Colleges und Universities.

### 3.4 Die Hypothesen

$$\mathbf{H}_0 : \mu_0 = \mu_1 = \mu$$

⇒ Sowohl Liberal Art Colleges als auch Universities weisen dieselben Mittelwerte für die jeweilige Variable auf bzw. sie entstammen derselben Grundgesamtheit.

$$\mathbf{H}_1 : \mu_0 \neq \mu_1$$

⇒ Liberal Art Colleges und Universities entstammen verschiedenen Grundgesamtheiten und haben für das Merkmal nicht denselben Mittelwert.

### 3.5 Die Teststatistik<sup>3</sup>

Der F-Test vergleicht die Differenzen der Summe der Abweichungsquadrate (SS) im vollen und im reduzierten Fall. Als Prüfgröße erhält man:

$$F = \frac{\{SS(\text{reduced}) - SS(\text{full})\}}{\{df(r) - df(f)\}} \cdot \frac{df(f)}{SS(\text{full})}$$

df(f) bzw. df(r) : Anzahl der Freiheitsgrade

df(f) = n – p            n = Gesamtzahl der Beobachtungen

p = Anzahl der Faktoren

df(r) = n – 1

Als Signifikanzniveau legen wir  $\alpha = 0,05$  fest.

df(f) = 48

df(r) = 49

Der kritische F-Wert (1,48) beim Signifikanzniveau liegt im Intervall von  $4,085 > f^* > 4,034$ . Die Nullhypothese wird abgelehnt, falls  $F > f^*$

*Für den Datensatz erhält man folgende Ergebnisse:*

X<sub>2</sub>:            F = 0,68518

X<sub>3</sub>:            F = 2,1187

X<sub>4\*</sub>:           F = 40,788

X<sub>5</sub>:            F = 19,567

X<sub>6\*</sub>:           F = 7,093

X<sub>7\*</sub>:           F = 0,21363

Für die Variablen Student/\$ (X<sub>4\*</sub>), Top 10% (X<sub>5</sub>) und %PhD (X<sub>6\*</sub>) wird die Nullhypothese auf einem Signifikanzniveau von  $\alpha = 0,05$  verworfen. Es kann somit davon

---

<sup>3</sup> siehe Makro „ANOVA.xpl“ (Diskette)

ausgegangen werden, daß sich Liberal Art Colleges und Universities in diesen Eigenschaften unterscheiden. Vergleicht man die Daten aus Tab. 2.1 so läßt sich schlußfolgern:

- An Universities ist die Quote derjenigen Studenten, die bereits in der High School zu den besten zählten im Durchschnitt höher als in Liberal Art Colleges.
- Der Anteil derjenigen Lehrkräfte, die über einen PhD verfügen ist an Liberal Art Colleges im Durchschnitt etwas geringer.
- Liberal Art Colleges haben mehr Studenten pro ausgegebenem Dollar, d.h. daß Universities umgekehrt mehr Dollar pro Student ausgeben.

#### **4. ZUSAMMENFASSUNG**

Im Rahmen der vorliegenden Datenanalyse wurden 6 metrische und 1 nominalskaliertes Merkmal von 50 amerikanischen Colleges untersucht. Ausgehend von einer univariaten Datenanalyse, in der Lage- und Streuungsparameter sowie Verteilung der einzelnen Variablen untersucht wurde, schloß sich eine Untersuchung der Abhängigkeiten zwischen den Variablen an. Es wurde eine Regressionsanalyse durchgeführt, um den Einfluß der Quote der Bewerber, die akzeptiert werden auf die Graduiertenquote zu modellieren.

Im 3. Kapitel wurde mittels einer ANOVA getestet, in welchen Merkmalen sich Liberal Art Colleges und Universities signifikant voneinander unterscheiden. Dabei wurde festgestellt, daß Universities höhere finanzielle Ausgaben pro Student haben, der Anteil der Studenten, die zu den Top 10% in der High School gehörten höher ist und der Anteil der Lehrkräfte mit PhD-Grad größer ist als an Liberal Art Colleges.

Damit wurde festgestellt, daß es deutliche Unterschiede zwischen Liberal Art Colleges und Universities in bestimmten Merkmalen gibt.



Claudia Kegel  
Anna-Seghers-Str. 59  
12489 Berlin

Matrikel-Nr. 116477

**„Analyse eines mehrdimensionalen  
Datensatzes unterstützt durch Xplo-  
Re“**

Hausarbeit zur Vorlesung Einführung in die Softwareumgebung XploRe im  
WS1998/99

## **Inhaltsverzeichnis**

3. EINLEITUNG	1
1.1 Die Daten	1
1.2 Ziel der Analyse und angewandte Methoden	1
4. DATENANALYSE	2
2.1 Univariate Datenanalyse	2
2.1.1 Gegenüberstellung Liberal Art Colleges vs. Universities	6
2.2 Bivariate Datenanalyse	8
2.2.1 Abhängigkeit zwischen den Variablen	8
2.2.2 Regressionsanalyse	11
3. VARIANZANALYSE	14
3.1 Ziel	14
3.2 Annahmen	14
3.3 Die Struktur	16
3.4 Die Hypothesen	16
3.5 Die Teststatistik	17
4. ZUSAMMENFASSUNG	18

