

Einführung in die Statistik mit R

(R I, II & III)

Wintersemester 2006/2007, Sommersemester 2007 & Wintersemester 2007/2008

und

Lineare Modelle mit R:

Regression und Varianzanalyse

(R IV)

Sommersemester 2008

Dr. Gerrit Eichner
Mathematisches Institut der
Justus-Liebig-Universität Gießen
Arndtstr. 2, D-35392 Gießen, Tel.: 0641/99-32104
E-Mail: gerrit.eichner@math.uni-giessen.de
URL: www.math.uni-giessen.de/gerrit.eichner

Inhaltsverzeichnis

1	Einführung	1
1.1	Was ist R und woher kommt es?	1
1.2	Unsere Unix-Umgebung	1
1.3	Grundsätzliches zur Arbeit mit R	4
1.3.1	Aufruf von R	4
1.3.2	Befehlseingabe am R-Prompt	5
1.3.3	Benutzerdefinierte Objekte: Zulässige Namen, speichern und löschen	6
1.3.4	Die “command history”	7
1.3.5	R-Demonstrationen	7
1.3.6	Die R Online-Hilfe	7
1.3.7	Beenden von R	8
2	Datenobjekte: Strukturen, Attribute und elementare Operationen	9
2.1	Konzeptionelle Grundlagen	9
2.1.1	Atomare Strukturen/Vektoren	9
2.1.2	Rekursive Strukturen/Vektoren	9
2.1.3	Weitere Objekttypen und Attribute	10
2.1.4	Das Attribut „Klasse“ (“class”)	10
2.2	numeric-Vektoren: Erzeugung und elementare Operationen	10
2.2.1	Beispiele regelmäßiger Zahlenfolgen: <code>seq()</code> und <code>rep()</code>	11
2.2.2	Elementare Vektoroperationen	13
2.3	Arithmetik und Funktionen für numeric-Vektoren	14
2.3.1	Elementweise Vektoroperationen: Rechnen, Runden, Formatieren	14
2.3.2	Zusammenfassende und sequenzielle Vektoroperationen	15
2.3.3	“Summary statistics”: <code>summary()</code>	16
2.4	logical-Vektoren und logische Operatoren	18
2.4.1	Elementweise logische Operationen	18
2.4.2	Zusammenfassende logische Operationen	19
2.5	character-Vektoren und elementare Operationen	20
2.5.1	Zusammensetzen von Zeichenketten: <code>paste()</code>	20
2.5.2	Benennung von Vektorelementen: <code>names()</code>	21
2.5.3	Operationen für character-Vektoren	21
2.6	Indizierung und Modifikation von Vektorelementen: []	22
2.6.1	Indexvektoren	22
2.6.2	Zwei spezielle Indizierungsfunktionen: <code>head()</code> und <code>tail()</code>	23
2.6.3	Indizierte Zuweisungen	23
2.7	Faktoren und geordnete Faktoren: Definition und Verwendung	24
2.7.1	Erzeugung von Faktoren: <code>factor()</code>	25
2.7.2	Änderung der Levelsortierung und Zusammenfassung von Levels	26
2.7.3	Erzeugung von geordneten Faktoren: <code>ordered()</code>	26
2.7.4	Änderung der Levelordnung und Zusammenfassung von Levels bei geordneten Faktoren	26
2.7.5	Klassierung und Erzeugung von geordneten Faktoren: <code>cut()</code>	27
2.7.6	Tabellierung von Faktoren und Faktorkombinationen: <code>table()</code>	27
2.7.7	Faktor(en)gruppierete Funktionsanwendung: <code>tapply()</code>	27
2.8	Matrizen: Erzeugung, Indizierung, Modifikation und Operationen	28
2.8.1	Grundlegendes zu Arrays	28
2.8.2	Erzeugung von Matrizen: <code>matrix()</code>	28
2.8.3	Benennung von Spalten und Zeilen: <code>dimnames()</code> , <code>colnames()</code> , <code>rownames()</code>	29
2.8.4	Erweiterung mit Spalten und Zeilen: <code>cbind()</code> und <code>rbind()</code>	30

2.8.5	Matrixdimensionen und Indizierung von Elementen: <code>dim()</code> , <code>[]</code> , <code>head()</code> et al.	30
2.8.6	Einige spezielle Matrizen: <code>diag()</code> , <code>col()</code> , <code>lower.tri()</code> & Co.	32
2.8.7	Wichtige Operationen der Matrixalgebra	32
2.8.8	Zeilen- und spaltenweise Anwendung von Operationen: <code>apply()</code>	33
2.8.9	Statistikspezifische Matrixfunktionen und Zeilen- bzw. Spaltenzusammenfassungen: <code>cov()</code> , <code>colSums()</code> , <code>colMeans()</code> und Verwandte	34
2.8.10	Erzeugung spezieller Matrizen mit Hilfe von <code>outer()</code>	34
2.9	Listen: Konstruktion, Indizierung und Verwendung	35
2.9.1	Erzeugung und Indizierung: <code>list()</code> und <code>[] []</code> , <code>head()</code> sowie <code>tail()</code>	35
2.9.2	Benennung von Listenelementen und ihre Indizierung: <code>names()</code> und <code>\$</code>	36
2.9.3	Komponentenweise Anwendung von Operationen: <code>lapply()</code> und <code>sapply()</code>	37
2.10	Data Frames: Eine Klasse „zwischen“ Matrizen und Listen	38
2.10.1	Indizierung von Data Frames: <code>[]</code> und <code>\$</code> sowie <code>head()</code> und <code>tail()</code>	38
2.10.2	Erzeugung von Data Frames: <code>data.frame()</code>	39
2.10.3	„Summary statistics“ und Struktur eines Data Frames: <code>summary()</code> und <code>str()</code>	40
2.10.4	Komponentenweise Anwendung von Operationen: <code>lapply()</code> und <code>sapply()</code>	40
2.10.5	Anwendung von Operationen auf nach Faktoren gruppierte Zeilen: <code>by()</code>	41
2.10.6	„Öffnen“ und „Schließen“ von Data Frames und der Suchpfad: <code>attach()</code> , <code>detach()</code> und <code>search()</code>	42
2.11	Abfrage und Konversion der Klasse von Datenobjekten sowie Abfrage von NA, NaN, Inf und NULL	43
3	Import und Export von Daten	45
3.1	Datenimport aus einer Datei: <code>scan()</code> und <code>read.table()</code>	45
3.1.1	Die Funktion <code>scan()</code>	45
3.1.2	Die Beispieldaten „SMSA“	48
3.1.3	Die Funktion <code>read.table()</code>	49
3.2	Datenexport in eine Datei	52
3.2.1	Die Funktionen <code>write()</code> , <code>sink()</code> und <code>write.table()</code>	52
4	Elementare explorative Grafiken	54
4.1	Grafikausgabe am Bildschirm und in Dateien	54
4.2	Explorative Grafiken für univariate Daten	54
4.2.1	Die Häufigkeitsverteilung diskreter Daten: Balken-, Flächen- und Kreisdiagramme sowie Dot Charts	55
4.2.2	Die Verteilung metrischer Daten: Histogramme, „stem-and-leaf“-Diagramme, Boxplots und Q-Q-Plots	57
4.2.3	Zur Theorie und Interpretation von Boxplots und Q-Q-Plots	60
4.3	Explorative Grafiken für multivariate Daten	62
4.3.1	Die Häufigkeitsverteilung bivariat diskreter Daten: Mosaikplots	62
4.3.2	Die Verteilung multivariat metrischer Daten: Streudiagramme	63
4.3.3	Die Verteilung trivariat metrischer Daten: Bedingte Streudiagramme („coplots“)	66
4.3.4	Eine weitere Möglichkeit multivariater Darstellung: „Stars“	68
5	Wahrscheinlichkeitsverteilungen und Pseudo-Zufallszahlen	69
5.1	Die eingebauten Verteilungen	69
5.2	Bemerkungen zu Pseudo-Zufallszahlen in R	71

6	Definition neuer Funktionen	72
6.1	Ein Beispiel	72
6.2	Syntax der Funktionsdefinition	73
6.3	Spezifizierung von Funktionsargumenten	74
6.3.1	Argumente mit default-Werten	74
6.3.2	Variable Argumentezahl	75
6.3.3	Zuordnung von Aktual- zu Formalparametern beim Funktionsaufruf	76
6.3.4	Rückgabewert einer Funktion	76
6.4	Kontrollstrukturen: Bedingte Anweisungen, Schleifen, Wiederholungen	77
7	Weiteres zur elementaren Grafik	79
7.1	Grafikausgabe	79
7.2	Elementare Plotfunktionen	79
7.3	Layoutfunktionen und Grafikparameter	80
7.4	Einige (auch mathematisch) nützliche Plotfunktionen	85
7.4.1	Stetige Funktionen: <code>curve()</code>	85
7.4.2	Geschlossener Polygonzug: <code>polygon()</code>	85
7.4.3	Beliebige Treppenfunktionen: <code>plot()</code> in Verbindung mit <code>stepfun()</code>	85
7.4.4	Die empirische Verteilungsfunktion: <code>plot()</code> in Verbindung mit <code>ecdf()</code>	85
7.4.5	„Fehlerbalken“: <code>errbar()</code> im Package <code>Hmisc</code>	86
7.4.6	Mehrere Polygonzüge „auf einmal“: <code>matplot()</code>	87
7.5	Interaktion mit Plots	87
8	Para- und nicht-parametrische Inferenzstatistik in „klassischen“ Ein- und Zweistichprobenproblemen für metrische Daten	89
8.1	Auffrischung des Konzepts statistischer Tests	89
8.1.1	Motivation anhand eines Beispiels	89
8.1.2	Null- & Alternativhypothese, Fehler 1. & 2. Art	89
8.1.3	Konstruktion eines Hypothesentests im Normalverteilungsmodell	91
8.1.4	Der p -Wert	93
8.2	Konfidenzintervalle für den Erwartungswert der Normalverteilung	95
8.3	Eine Hilfsfunktion für die explorative Datenanalyse	96
8.4	Ein Einstichproben-Lokationsproblem	98
8.4.1	Der Einstichproben- t -Test	98
8.4.2	Wilcoxons Vorzeichen-Rangsummentest	99
8.5	Zweistichproben-Lokations- und Skalenprobleme	102
8.5.1	Der Zweistichproben- F -Test auf Gleichheit der Varianzen	102
8.5.2	Der Zweistichproben- t -Test bei unbekanntem, aber gleichen Varianzen	103
8.5.3	Die Welch-Modifikation des Zweistichproben- t -Tests	104
8.5.4	Wilcoxons Rangsummentest (Mann-Whitney U-Test)	105
8.6	Das Zweistichproben-Lokationsproblem für verbundene Stichproben	108
8.6.1	Die Zweistichproben- t -Tests bei verbundenen Stichproben	109
8.6.2	Wilcoxons Vorzeichen-Rangsummentest für verbundene Stichproben	111
8.7	Tests auf Unkorreliertheit	112
8.7.1	Der Pearsonsche Test auf Unkorreliertheit	113
8.7.2	Der Spearmansche Rangtest auf Unkorreliertheit	114
8.8	Die Formelversionen der Funktionen für die Zweistichproben-tests	117
8.9	Testgüte und Bestimmung des Stichprobenumfangs für Lokationsprobleme im Normalverteilungsmodell	119
8.9.1	Der zweiseitige Einstichproben-Gaußtest	119
8.9.1.1	Herleitung der Gütefunktion	119
8.9.1.2	Interpretation und Veranschaulichung der Gütefunktion	120

8.9.1.3	Verwendungen für die Gütefunktion	122
8.9.1.4	Das Problem der unbekanntem Varianz	123
8.9.2	Der zweiseitige Einstichproben- <i>t</i> -Test	124
8.9.2.1	Herleitung der Gütefunktion	124
8.9.2.2	Verwendung der Gütefunktion	125
8.9.3	Der einseitige Einstichproben- <i>t</i> -Test	127
8.9.3.1	Gütefunktion: Herleitung, Eigenschaften und Veranschaulichung	127
8.9.3.2	Verwendung der Gütefunktion	128
8.9.4	Die Zweistichproben- <i>t</i> -Tests	129
8.9.4.1	Zwei verbundene Stichproben	130
8.9.4.2	Zwei unverbundene Stichproben	131
9	Zur Inferenzstatistik und Parameterschätzung für Nominaldaten	133
9.1	Bernoulli-Experimente mit <code>sample()</code>	133
9.2	Einstichprobenprobleme im Binomialmodell	134
9.2.1	Der exakte Test für die Auftrittswahrscheinlichkeit <i>p</i> : <code>binom.test()</code> . . .	134
9.2.2	Der approximative Test für <i>p</i> : <code>prop.test()</code>	135
9.2.3	Konfidenzintervalle für <i>p</i>	136
9.3	Mehrstichprobentests im Binomialmodell	138
9.3.1	Zur Theorie der approximativen <i>k</i> -Stichproben-Binomialtests (Pearsons X^2 -Tests)	139
9.3.2	Zur Implementation der <i>k</i> -Stichproben-Binomialtests: <code>prop.test()</code> . . .	141
9.3.2.1	Der Fall $k = 2$ Stichproben	141
9.3.2.2	Der Fall $k \geq 3$ Stichproben	142
9.4	Testgüte und Bestimmung von Stichprobenumfängen im Binomialmodell	143
9.4.1	Einseitiger und zweiseitiger Einstichprobentest	143
9.4.2	Einseitiger und zweiseitiger Zweistichprobentest: <code>power.prop.test()</code> . .	144
9.5	Tests im Multinomialmodell	146
9.5.1	Multinomial-Experimente mit <code>sample()</code>	146
9.5.2	Der approximative χ^2 -Test im Multinomialmodell: <code>chisq.test()</code>	147
9.6	Kontingenztafeln	148
9.6.1	χ^2 -Test auf Unabhängigkeit zweier Faktoren und auf Homogenität	148
9.6.1.1	Zum Fall der Unabhängigkeit	149
9.6.1.2	Zum Fall der Homogenität	150
9.6.1.3	Der approximative χ^2 -Test auf Unabhängigkeit und der approximative χ^2 -Test auf Homogenität: <code>chisq.test()</code>	151
9.6.2	Fishers Exakter Test auf Unabhängigkeit zweier Faktoren	153
9.6.2.1	Die Implementation durch <code>fisher.test()</code>	153
9.6.2.2	Der Spezialfall der (2×2) -Tafel: Die Odds Ratio	154
9.6.3	Kontingenztafeln für $k \geq 2$ Faktoren und ihre Darstellung: <code>xtabs()</code> & <code>fTable()</code>	156
9.6.3.1	Der Fall bereits registrierter absoluter Häufigkeiten	157
9.6.3.2	Der Fall explizit aufgeführter Levelkombinationen	159
10	Einführung in die lineare Regression	162
10.1	Die einfache lineare Regression	164
10.2	Die multiple lineare Regression	166
10.3	Zur Syntax von Modellformeln	169
10.4	Zur Interaktion stetiger Covariablen	171
10.5	Modelldiagnose I: Residualanalyse und Transformationen des Modells	174
10.5.1	Grafische Residualanalyse	174
10.5.2	Varianz stabilisierende Transformationen	176

10.5.3	Linearisierende Transformationen	177
10.5.4	Symmetrisierung des Designs und spezielle linearisierbare Regressions- funktionen	178
10.6	Modifizierung eines linearen Regressionsmodells	179
10.6.1	Die Funktion <code>update()</code>	180
10.6.2	Das Entfernen eines Terms: <code>drop1()</code>	180
10.6.3	Das Hinzufügen eines Terms: <code>add1()</code>	183
10.6.4	Akaikes Informationskriterium AIC	184
10.6.4.1	Die Diskrepanz	185
10.6.4.2	Die Kullback-Leibler-Diskrepanz und AIC	187
10.6.4.3	AIC im Modell der linearen Regression	188
10.7	Modelldiagnose II: Ausreißer, Extrempunkte, einflussreiche Punkte und Residual- analyse	190
10.7.1	Grafische Identifizierung	190
10.7.2	Inferenzstatistische Residualanalyse	191
10.7.3	Quantitative Identifizierung einflussreicher Punkte und Quantifizierung ihres Einflusses	193
10.7.3.1	Einfluss eines Punktes auf $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_n)'$	193
10.7.3.2	Einfluss eines Punktes auf $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_{p-1})'$	194
10.7.3.3	Einfluss eines Punktes auf $\hat{\sigma}^2$	195
10.7.4	Zusammenfassung und Umsetzung	195
10.7.4.1	Die Funktion <code>influence.measures()</code> und <code>Co.</code>	195
10.7.4.2	Die Funktion <code>summary.lm()</code>	197
10.7.5	Zur Unabhängigkeitsannahme der Fehler	198
10.8	Schätz- und Prognosewerte sowie Konfidenz- und Toleranzintervalle	199
10.8.1	Schätzwerte für die Regressionsfunktion und grafische Darstellung	199
10.8.2	Punktweise Konfidenzintervalle für die Regressionsfunktion und für ihre Parameter	203
10.8.3	Simultane Konfidenzintervalle für die Regressionsfunktion und simultane Konfidenzbereiche für ihren Parametervektor	204
10.8.4	Ein Konfidenzband für die Regressionsfunktion	207
10.8.5	Punktweise und simultane Toleranzintervalle für zukünftige Response-Werte	209
10.8.6	Die Formeln im Spezialfall der einfachen linearen Regression	211
10.8.6.1	Konfidenzintervalle für die Parameter der Regressionsgeraden	211
10.8.6.2	Konfidenzintervalle für die Regressionsgerade	212
10.8.6.3	Toleranzintervalle zukünftiger Response-Werte	213
	Literaturverzeichnis	215