

Grundlagen der Datenanalyse mit R

Eine anwendungsorientierte Einführung

von

Daniel Wollschläger

1. Auflage

Springer 2012

Verlag C.H. Beck im Internet:

www.beck.de

ISBN 978 3 642 25799 5

Zu [Leseprobe](#)

Inhaltsverzeichnis

1	Erste Schritte	1
1.1	Vorstellung	1
1.1.1	Pro und Contra R	1
1.1.2	Typografische Konventionen	3
1.1.3	R installieren	4
1.1.4	Grafische Benutzeroberflächen	5
1.1.5	Weiterführende Informationsquellen und Literatur	5
1.2	Grundlegende Elemente	7
1.2.1	R Starten, beenden und die Konsole verwenden	7
1.2.2	Einstellungen	10
1.2.3	Umgang mit dem workspace	11
1.2.4	Einfache Arithmetik	13
1.2.5	Funktionen mit Argumenten aufrufen	15
1.2.6	Hilfe-Funktionen	16
1.2.7	Zusatzpakete verwenden	16
1.3	Datenstrukturen: Klassen, Objekte, Datentypen	19
1.3.1	Objekte benennen	20
1.3.2	Zuweisungen an Objekte	20
1.3.3	Objekte ausgeben	21
1.3.4	Objekte anzeigen lassen, umbenennen und entfernen	22
1.3.5	Datentypen	23
1.3.6	Logische Werte, Operatoren und Verknüpfungen	24
2	Elementare Dateneingabe und -verarbeitung	27
2.1	Vektoren	27
2.1.1	Vektoren erzeugen	27
2.1.2	Elemente auswählen und verändern	28
2.1.3	Datentypen in Vektoren	30
2.1.4	Elemente benennen	31
2.1.5	Elemente löschen	32
2.2	Logische Operatoren	32

2.2.1	Logische Operatoren zum Vergleich von Vektoren	32
2.2.2	Logische Indexvektoren	35
2.3	Mengen	36
2.3.1	Duplizierte Werte behandeln	37
2.3.2	Mengenoperationen	37
2.3.3	Kombinatorik	39
2.4	Systematische und zufällige Wertefolgen erzeugen	41
2.4.1	Numerische Sequenzen erstellen	41
2.4.2	Wertefolgen wiederholen	43
2.4.3	Zufällig aus einer Urne ziehen	43
2.4.4	Zufallszahlen aus bestimmten Verteilungen erzeugen	44
2.5	Daten transformieren	45
2.5.1	Werte sortieren	45
2.5.2	Werte in zufällige Reihenfolge bringen	46
2.5.3	Teilmengen von Daten auswählen	47
2.5.4	Daten umrechnen	48
2.5.5	Neue aus bestehenden Variablen bilden	51
2.5.6	Werte ersetzen oder recodieren	51
2.5.7	Kontinuierliche Variablen in Kategorien einteilen	53
2.6	Gruppierungsfaktoren	54
2.6.1	Ungeordnete Faktoren	54
2.6.2	Faktoren kombinieren	56
2.6.3	Faktorstufen nachträglich ändern	57
2.6.4	Geordnete Faktoren	59
2.6.5	Reihenfolge von Faktorstufen	59
2.6.6	Faktoren nach Muster erstellen	60
2.6.7	Quantitative in kategoriale Variablen umwandeln	61
2.7	Deskriptive Kennwerte numerischer Daten	62
2.7.1	Summen, Differenzen und Produkte	63
2.7.2	Extremwerte	64
2.7.3	Mittelwert, Median und Modalwert	65
2.7.4	Robuste Maße der zentralen Tendenz	66
2.7.5	Prozentrang, Quartile, Quantile und Interquartilabstand	67
2.7.6	Varianz, Streuung, Schiefe und Wölbung	68
2.7.7	Kovarianz und Korrelation	69
2.7.8	Kennwerte getrennt nach Gruppen berechnen	71
2.7.9	Funktionen auf geordnete Paare von Werten anwenden	73
2.8	Matrizen	73
2.8.1	Datentypen in Matrizen	74
2.8.2	Dimensionierung, Zeilen und Spalten	75
2.8.3	Elemente auswählen und verändern	77
2.8.4	Weitere Wege, Elemente auszuwählen und zu verändern	78
2.8.5	Matrizen verbinden	79
2.8.6	Matrizen sortieren	80
2.8.7	Randkennwerte berechnen	81

2.8.8	Beliebige Funktionen auf Matrizen anwenden	81
2.8.9	Matrix zeilen- oder spaltenweise mit Kennwerten verrechnen	82
2.8.10	Kovarianz- und Korrelationsmatrizen	83
2.9	Arrays	84
2.10	Häufigkeitsauszählungen	86
2.10.1	Einfache Tabellen absoluter und relativer Häufigkeiten	86
2.10.2	Iterationen zählen	88
2.10.3	Absolute, relative und bedingte relative Häufigkeiten in Kreuztabellen	89
2.10.4	Randkennwerte von Kreuztabellen	92
2.10.5	Kumulierte relative Häufigkeiten und Prozentrang	92
2.10.6	Diversität kategorialer Daten	94
2.11	Codierung, Identifikation und Behandlung fehlender Werte	94
2.11.1	Fehlende Werte codieren und ihr Vorhandensein prüfen	95
2.11.2	Fehlende Werte ersetzen und umcodieren	96
2.11.3	Behandlung fehlender Werte bei der Berechnung einfacher Kennwerte	97
2.11.4	Behandlung fehlender Werte in Matrizen	98
2.11.5	Behandlung fehlender Werte beim Sortieren von Daten	100
2.12	Zeichenketten verarbeiten	101
2.12.1	Objekte in Zeichenketten umwandeln	101
2.12.2	Zeichenketten erstellen und ausgeben	102
2.12.3	Zeichenketten manipulieren	104
2.12.4	Zeichenfolgen finden	106
2.12.5	Zeichenfolgen ersetzen	108
2.12.6	Zeichenketten als Befehl ausführen	108
2.13	Datum und Uhrzeit	109
2.13.1	Datumsangaben erstellen und formatieren	109
2.13.2	Uhrzeit	110
2.13.3	Berechnungen mit Datum und Uhrzeit	112
3	Datensätze	115
3.1	Listen	115
3.1.1	Komponenten auswählen, verändern und hinzufügen	116
3.1.2	Listen mit mehreren Ebenen	119
3.2	Datensätze	120
3.2.1	Datentypen in Datensätzen	122
3.2.2	Elemente auswählen und verändern	123
3.2.3	Namen von Variablen und Beobachtungen	124
3.2.4	Datensätze in den Suchpfad einfügen	125
3.3	Datensätze transformieren	127
3.3.1	Variablen hinzufügen und entfernen	127
3.3.2	Datensätze sortieren	128
3.3.3	Teilmengen von Daten auswählen	129

3.3.4	Doppelte und fehlende Werte behandeln	132
3.3.5	Datensätze teilen	134
3.3.6	Datensätze zusammenfügen	134
3.3.7	Organisationsform einfacher Datensätze ändern	137
3.3.8	Organisationsform komplexer Datensätze ändern	139
3.4	Daten aggregieren	143
3.4.1	Funktionen auf Variablen anwenden	143
3.4.2	Funktionen für mehrere Variablen anwenden	146
3.4.3	Funktionen getrennt nach Gruppen anwenden	147
4	Befehle und Daten verwalten	151
4.1	Befehlssequenzen im Editor bearbeiten	151
4.2	Daten importieren und exportieren	153
4.2.1	Daten in der Konsole einlesen	153
4.2.2	Daten im Editor eingeben	154
4.2.3	Im Textformat gespeicherte Daten	155
4.2.4	R-Objekte	157
4.2.5	Daten mit anderen Programmen austauschen	157
5	Hilfsmittel für die Inferenzstatistik	163
5.1	Wichtige Begriffe inferenzstatistischer Tests	163
5.2	Lineare Modelle formulieren	164
5.3	Funktionen von Zufallsvariablen	167
5.3.1	Dichtefunktionen	167
5.3.2	Verteilungsfunktionen	168
5.3.3	Quantilfunktionen	169
5.4	Behandlung fehlender Werte in inferenzstatistischen Tests	170
6	Korrelations- und Regressionsanalyse	171
6.1	Test auf Korrelation	171
6.2	Einfache lineare Regression	173
6.2.1	Deskriptive Modellanpassung	173
6.2.2	Regressionsanalyse	176
6.3	Multiple lineare Regression	178
6.3.1	Deskriptive Modellanpassung und Regressionsanalyse	178
6.3.2	Modell verändern	180
6.3.3	Modelle vergleichen und auswählen	181
6.3.4	Moderierte Regression	184
6.4	Regressionsmodelle auf andere Daten anwenden	186
6.5	Kreuzvalidierung von Regressionsmodellen	188
6.6	Regressionsdiagnostik	190
6.6.1	Extremwerte, Ausreißer und Einfluss	191
6.6.2	Verteilungseigenschaften der Residuen	194
6.6.3	Multikollinearität	197
6.6.4	Robuste Regressionsverfahren	199

6.7	Partialkorrelation und Semipartialkorrelation	200
6.8	Logistische Regression	202
6.8.1	Modellanpassung für dichotome Daten	202
6.8.2	Modellanpassung für binomiale Daten	204
6.8.3	Vorhersage und Klassifikation	205
6.8.4	Signifikanztest	207
7	Parametrische Tests für Dispersions- und Lageparameter von Verteilungen	209
7.1	Tests auf Varianzhomogenität	209
7.1.1	<i>F</i> -Test auf Varianzhomogenität für zwei Stichproben	209
7.1.2	Levene-Test für mehr als zwei Stichproben	211
7.1.3	Fligner-Killeen-Test für mehr als zwei Stichproben	212
7.2	<i>t</i> -Tests	212
7.2.1	<i>t</i> -Test für eine Stichprobe	212
7.2.2	<i>t</i> -Test für zwei unabhängige Stichproben	214
7.2.3	<i>t</i> -Test für zwei abhängige Stichproben	216
7.3	Einfaktorielle Varianzanalyse (CR- <i>p</i>)	217
7.3.1	Auswertung mit <code>oneway.test()</code>	217
7.3.2	Auswertung mit <code>aov()</code>	218
7.3.3	Auswertung mit <code>anova()</code>	220
7.3.4	Effektstärke schätzen	220
7.3.5	Grafische Prüfung der Voraussetzungen	221
7.3.6	Einzelvergleiche (Kontraste)	223
7.4	Einfaktorielle Varianzanalyse mit abhängigen Gruppen (RB- <i>p</i>)	229
7.4.1	Univariat formulierte Auswertung und Schätzung der Effektstärke	229
7.4.2	Zirkularität der Kovarianzmatrix prüfen	232
7.4.3	Multivariat formulierte Auswertung mit <code>Anova()</code>	234
7.4.4	Multivariat formulierte Auswertung mit <code>anova()</code>	235
7.4.5	Einzelvergleiche (Kontraste)	236
7.5	Zweifaktorielle Varianzanalyse (CRF- <i>pq</i>)	236
7.5.1	Auswertung und Schätzung der Effektstärke	236
7.5.2	Quadratsummen vom Typ I, II und III	239
7.5.3	Bedingte Haupteffekte testen	243
7.5.4	Beliebige a-priori Kontraste	245
7.5.5	Beliebige post-hoc Kontraste nach Scheffé	248
7.6	Zweifaktorielle Varianzanalyse mit zwei Intra-Gruppen Faktoren (RBF- <i>pq</i>)	249
7.6.1	Univariat formulierte Auswertung und Schätzung der Effektstärke	249
7.6.2	Zirkularität der Kovarianzmatrizen prüfen	253
7.6.3	Multivariat formulierte Auswertung	254
7.6.4	Einzelvergleiche (Kontraste)	255
7.7	Zweifaktorielle Varianzanalyse mit Split-Plot-Design (SPF- <i>p · q</i>)	255

7.7.1	Univariat formulierte Auswertung und Schätzung der Effektstärke	256
7.7.2	Voraussetzungen und Prüfen der Zirkularität	258
7.7.3	Multivariat formulierte Auswertung	259
7.7.4	Einzelvergleiche (Kontraste)	260
7.7.5	Erweiterung auf dreifaktorielles SPF- $p \cdot qr$ Design	261
7.7.6	Erweiterung auf dreifaktorielles SPF- $pq \cdot r$ Design	263
7.8	Kovarianzanalyse	264
7.8.1	Test der Effekte von Gruppenzugehörigkeit und Kovariate	264
7.8.2	Beliebige a-priori Kontraste	269
7.8.3	Beliebige post-hoc Kontraste nach Scheffé	271
7.9	Power, Effektstärke und notwendige Stichprobengröße	271
7.9.1	Binomialtest	271
7.9.2	t -Test	273
7.9.3	Einfaktorielle Varianzanalyse	276
8	Klassische nonparametrische Methoden	281
8.1	Anpassungstests	281
8.1.1	Binomialtest	282
8.1.2	Test auf Zufälligkeit (Runs-Test)	284
8.1.3	Kolmogorov-Smirnov-Anpassungstest	286
8.1.4	χ^2 -Test auf eine feste Verteilung	288
8.1.5	χ^2 -Test auf eine Verteilungsklasse	289
8.2	Analyse von gemeinsamen Häufigkeiten kategorialer Variablen	291
8.2.1	χ^2 -Test auf Unabhängigkeit	291
8.2.2	χ^2 -Test auf Gleichheit von Verteilungen	292
8.2.3	χ^2 -Test für mehrere Auftretenswahrscheinlichkeiten	293
8.2.4	Fishers exakter Test auf Unabhängigkeit	294
8.2.5	Fishers exakter Test auf Gleichheit von Verteilungen	296
8.2.6	Kennwerte von (2×2) -Konfusionsmatrizen	297
8.3	Maße für Zusammenhang und Übereinstimmung	300
8.3.1	Spearmans ρ und Kendalls τ	300
8.3.2	Zusammenhang kategorialer Variablen	302
8.3.3	Inter-Rater-Übereinstimmung	303
8.4	Tests auf gleiche Variabilität	310
8.4.1	Mood-Test	311
8.4.2	Ansari-Bradley-Test	312
8.5	Tests auf Übereinstimmung von Verteilungen	313
8.5.1	Kolmogorov-Smirnov-Test für zwei Stichproben	313
8.5.2	Vorzeichen-Test	315
8.5.3	Wilcoxon-Vorzeichen-Rang-Test für eine Stichprobe	316
8.5.4	Wilcoxon-Rangsummen-Test / Mann-Whitney- U -Test	317
8.5.5	Wilcoxon-Test für zwei abhängige Stichproben	319
8.5.6	Kruskal-Wallis- H -Test für unabhängige Stichproben	319
8.5.7	Friedman-Rangsummen-Test für abhängige Stichproben	320

8.5.8	Cochran- Q -Test für abhängige Stichproben	322
8.5.9	Bowker-Test für zwei abhängige Stichproben	323
8.5.10	McNemar-Test für zwei abhängige Stichproben	324
8.5.11	Stuart-Maxwell-Test für zwei abhängige Stichproben	325
9	Resampling-Verfahren	329
9.1	Bootstrapping	329
9.1.1	Bootstrap-Vertrauensintervalle für μ	330
9.1.2	Bootstrap-Vertrauensintervalle für Regressionsparameter ..	333
9.1.3	Bootstrap-Tests in varianzanalytischen Designs	336
9.2	Permutationstests	337
9.2.1	Test auf gleiche Lageparameter in unabhängigen Stichproben	338
9.2.2	Test auf gleiche Lageparameter in abhängigen Stichproben ..	340
9.2.3	Test auf Unabhängigkeit von zwei Variablen	342
10	Multivariate Verfahren	343
10.1	Lineare Algebra	343
10.1.1	Matrix-Algebra	344
10.1.2	Lineare Gleichungssysteme lösen	346
10.1.3	Norm und Abstand von Vektoren und Matrizen	347
10.1.4	Mahalanobistransformation und Mahalanobisdistanz	348
10.1.5	Kennwerte von Matrizen	351
10.1.6	Zerlegungen von Matrizen	353
10.1.7	Orthogonale Projektion	355
10.2	Hauptkomponentenanalyse	358
10.3	Faktorenanalyse	364
10.4	Multidimensionale Skalierung	370
10.5	Multivariate multiple Regression	372
10.6	Hotellings T^2	374
10.6.1	Test für eine Stichprobe	374
10.6.2	Test für zwei unabhängige Stichproben	376
10.6.3	Test für zwei abhängige Stichproben	378
10.6.4	Univariate Varianzanalyse mit abhängigen Gruppen (RB- p)	379
10.7	Multivariate Varianzanalyse (MANOVA)	380
10.7.1	Einfaktorielle MANOVA	380
10.7.2	Zweifaktorielle MANOVA	381
10.8	Diskriminanzanalyse	382
10.9	Das allgemeine lineare Modell	386
10.9.1	Modell der multiplen linearen Regression	387
10.9.2	Modell der einfaktoriellen Varianzanalyse	389
10.9.3	Modell der zweifaktoriellen Varianzanalyse	394
10.9.4	Parameterschätzungen, Vorhersage und Residuen	398
10.9.5	Hypothesen über parametrische Funktionen testen	400
10.9.6	Lineare Hypothesen als Modellvergleiche formulieren ..	401

10.9.7 Lineare Hypothesen testen	405
10.9.8 Beispiel: Multivariate multiple Regression	408
10.9.9 Beispiel: Einfaktorielle MANOVA	410
10.9.10 Beispiel: Zweifaktorielle MANOVA	414
11 Diagramme erstellen	417
11.1 Grafik-Devices	417
11.1.1 Aufbau und Verwaltung von Grafik-Devices	417
11.1.2 Grafiken speichern	419
11.2 Streu- und Liniendiagramme	421
11.2.1 Streudiagramme mit <code>plot()</code>	421
11.2.2 Datenpunkte eines Streudiagramms identifizieren	424
11.2.3 Streudiagramme mit <code>matplot()</code>	425
11.3 Diagramme formatieren	426
11.3.1 Grafikelemente formatieren	426
11.3.2 Farben spezifizieren	428
11.3.3 Achsen formatieren	430
11.4 Säulen- und Punktdiagramme	431
11.4.1 Einfache Säulendiagramme	431
11.4.2 Gruppierte und gestapelte Säulendiagramme	432
11.4.3 Dotchart	436
11.5 Elemente einem bestehenden Diagramm hinzufügen	437
11.5.1 Koordinaten in einem Diagramm identifizieren	438
11.5.2 In beliebige Diagrammbereiche zeichnen	438
11.5.3 Punkte	440
11.5.4 Linien	441
11.5.5 Polygone	443
11.5.6 Funktionsgraphen	447
11.5.7 Text und mathematische Formeln	448
11.5.8 Achsen	450
11.5.9 Fehlerbalken	451
11.5.10 Rastergrafiken	454
11.6 Verteilungsdiagramme	456
11.6.1 Histogramm und Schätzung der Dichtefunktion	456
11.6.2 Stamm-Blatt-Diagramm	459
11.6.3 Boxplot	460
11.6.4 Stripchart	461
11.6.5 Quantil-Quantil-Diagramm	463
11.6.6 Empirische kumulierte Häufigkeitsverteilung	464
11.6.7 Kreisdiagramm	465
11.6.8 Gemeinsame Verteilung zweier Variablen	466
11.7 Daten interpolieren und fitten	469
11.7.1 Lineare Interpolation und polynomiale Glätter	469
11.7.2 Splines	471
11.8 Multivariate Daten visualisieren	472

11.8.1 Höhenlinien und variable Datenpunktsymbole	473
11.8.2 Dreidimensionale Gitter und Streudiagramme	476
11.8.3 Bedingte Diagramme für mehrere Gruppen	477
11.8.4 Matrix aus Streudiagrammen	480
11.9 Mehrere Diagramme in einem Grafik-Device darstellen	483
11.9.1 <code>layout()</code>	483
11.9.2 <code>par(mfrow, mfcol, fig)</code>	485
11.9.3 <code>split.screen()</code>	487
12 R als Programmiersprache	491
12.1 Kontrollstrukturen	491
12.1.1 Fallunterscheidungen	491
12.1.2 Schleifen	494
12.2 Eigene Funktionen erstellen	497
12.2.1 Funktionskopf	497
12.2.2 Funktionsrumpf	498
12.2.3 Rückgabewert	500
12.2.4 Eigene Funktionen verwenden	501
12.2.5 Generische Funktionen	501
12.3 Funktionen analysieren und verbessern	503
12.3.1 Quelltext fremder Funktionen begutachten	503
12.3.2 Funktionen zur Laufzeit untersuchen	504
12.3.3 Effizienz von Auswertungen steigern	506
Literaturverzeichnis	509
Index	517
R-Funktionen, Klassen und Schlüsselwörter	527
Zusatzpakete	535