

Daniel Wollschläger

Grundlagen der Datenanalyse mit R

Eine anwendungsorientierte Einführung

3., überarbeitete und erweiterte Auflage

& Springer Spektrum

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	3
1.1.4	Grafische Benutzeroberflächen	4
1.1.5	Weiterführende Informationsquellen und Literatur	5
1.2	Grundlegende Elemente	6
1.2.1	R Starten, beenden und die Konsole verwenden	6
1.2.2	Einstellungen	10
1.2.3	Umgang mit dem workspace	11
1.2.4	Einfache Arithmetik	13
1.2.5	Punktionen mit Argumenten aufrufen	15
1.2.6	Hilfe-Funktionen	16
1.2.7	Zusatzpakete verwenden	16
1.2.8	Empfehlungen und typische Fehlerquellen	18
1.3	Datenstrukturen: Klassen, Objekte, Datentypen	19
1.3.1	Objekte benennen	20
1.3.2	Zuweisungen an Objekte	21
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	31
2.1.4	Elemente benennen	31
2.1.5	Elemente löschen	32
2.2	Logische Operatoren	32
2.2.1	Vektoren mit logischen Operatoren vergleichen	33
2.2.2	Logische Indexvektoren	35
2.3	Mengen	37
2.3.1	Doppelt auftretende Werte finden	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	42
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 bestimmen	59
2.6.6	Faktoren nach Muster erstellen	61
2.6.7	Quantitative in kategoriale Variablen umwandeln	62
2.7	Deskriptive Kennwerte numerischer Daten	63
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	67
2.7.5	Prozentrang, Quartile und Quantile	68
2.7.6	Varianz, Streuung, Schiefe und Wölbung	69
2.7.7	Diversität kategorialer Daten	70
2.7.8	Kovarianz und Korrelation	70
2.7.9	Robuste Streuungsmaße und Kovarianzschätzer	72
2.7.10	Kennwerte getrennt nach Gruppen berechnen	73
2.7.11	Funktionen auf geordnete Paare von Werten anwenden	75
2.8	Matrizen	75
2.8.1	Datentypen in Matrizen	76
2.8.2	Dimensionierung, Zeilen und Spalten	76
2.8.3	Elemente auswählen und verändern	78
2.8.4	Weitere Wege, Elemente auszuwählen und zu verändern	80
2.8.5	Matrizen verbinden	81
2.8.6	Matrizen sortieren	82
2.8.7	Randkennwerte berechnen	83
2.8.8	Beliebige Funktionen auf Matrizen anwenden	83
2.8.9	Matrix Zeilen- oder spaltenweise mit Kennwerten verrechnen	84
2.8.10	Kovarianz- und Korrelationsmatrizen	85
2.9	Arrays	86
2.10	Häufigkeitsauszählungen	88
2.10.1	Einfache Tabellen absoluter und relativer Häufigkeiten	88

2.10.2	Iterationen zählen	90
2.10.3	Absolute, relative und bedingte relative Häufigkeiten in Kreuztabellen	90
2.10.4	Randkennwerte von Kreuztabellen	94
2.10.5	Datensätze aus Häufigkeitstabellen erstellen	94
2.10.6	Kumulierte relative Häufigkeiten und Prozentrang	95
2.11	Fehlende Werte behandeln	96
2.11.1	Fehlende Werte codieren und identifizieren	97
2.11.2	Fehlende Werte ersetzen und umcodieren	98
2.11.3	Behandlung fehlender Werte bei der Berechnung einfacher Kennwerte	99
2.11.4	Behandlung fehlender Werte in Matrizen	100
2.11.5	Behandlung fehlender Werte beim Sortieren von Daten	102
2.11.6	Behandlung fehlender Werte in inferenzstatistischen Tests	102
2.11.7	Multiple Imputation	103
2.12	Zeichenketten verarbeiten	103
2.12.1	Objekte in Zeichenketten umwandeln	103
2.12.2	Zeichenketten erstellen und ausgeben	104
2.12.3	Zeichenketten manipulieren	107
2.12.4	Zeichenfolgen finden	108
2.12.5	Zeichenfolgen extrahieren	109
2.12.6	Zeichenfolgen ersetzen	110
2.12.7	Zeichenketten als Befehl ausführen	111
2.13	Datum und Uhrzeit	112
2.13.1	Datumsangaben erstellen und formatieren	112
2.13.2	Uhrzeit	113
2.13.3	Mit Datum und Uhrzeit rechnen	115
3	Datensätze	117
3.1	Listen	117
3.1.1	Komponenten auswählen und verändern	118
3.1.2	Komponenten hinzufügen und entfernen	120
3.1.3	Listen mit mehreren Ebenen	121
3.2	Datensätze	122
3.2.1	Datentypen in Datensätzen	124
3.2.2	Elemente auswählen und verändern	125
3.2.3	Namen von Variablen und Beobachtungen	127
3.2.4	Datensätze in den Suchpfad einfügen	128
3.3	Datensätze transformieren	129
3.3.1	Variablen hinzufügen und entfernen	130
3.3.2	Datensätze sortieren	131
3.3.3	Teilmengen von Daten mit subsetO auswählen	132
3.3.4	Doppelte und fehlende Werte behandeln	135
3.3.5	Datensätze teilen	136
3.3.6	Datensätze zeilen- oder spaltenweise verbinden	137
3.3.7	Datensätze mit mergeO zusammenführen	138
3.3.8	Organisationsform einfacher Datensätze ändern	141
3.3.9	Organisationsform komplexer Datensätze ändern	143

3.4	Daten aggregieren	147
3.4.1	Punktionen auf Variablen anwenden	147
3.4.2	Punktionen für mehrere Variablen anwenden	150
3.4.3	Punktionen getrennt nach Gruppen anwenden	151
4	Befehle und Daten verwalten	153
4.1	Befehlssequenzen im Editor bearbeiten	153
4.2	Daten importieren und exportieren	154
4.2.1	Daten im Editor eingeben	155
4.2.2	Datentabellen im Textformat	155
4.2.3	R-Objekte	158
4.2.4	Daten mit anderen Programmen austauschen	158
4.2.5	Daten in der Konsole einlesen	165
4.2.6	Unstrukturierte Textdateien	165
4.2.7	Datenqualität sicherstellen	166
4.3	Dateien verwalten	167
4.3.1	Dateien auswählen	167
4.3.2	Dateipfade manipulieren	168
4.3.3	Dateien verändern	169
5	Hilfsmittel für die Inferenzstatistik	171
5.1	Wichtige Begriffe inferenzstatistischer Tests	171
5.2	Lineare Modelle formulieren	172
5.3	Punktionen von Zufallsvariablen	174
5.3.1	Dichtefunktion	175
5.3.2	Verteilungsfunktion	176
5.3.3	Quantilfunktion	177
6	Lineare Regression	179
6.1	Test auf Korrelation	179
6.2	Einfache lineare Regression	180
6.2.1	Deskriptive Modellanpassung	181
6.2.2	Regressionsanalyse	183
6.3	Multiple lineare Regression	186
6.3.1	Deskriptive Modellanpassung und Regressionsanalyse	186
6.3.2	Modell verändern	188
6.3.3	Modelle vergleichen und auswählen	189
6.3.4	Moderierte Regression	191
6.4	Regressionsmodelle auf andere Daten anwenden	194
6.5	Regressionsdiagnostik	195
6.5.1	Extremwerte, Ausreißer und Einfluss	196
6.5.2	Verteilungseigenschaften der Residuen	199
6.5.3	Multikollinearität	201
6.6	Erweiterungen der linearen Regression	203
6.6.1	Robuste Regression	203
6.6.2	Penalisierte Regression	204
6.6.3	Nichtlineare Zusammenhänge	207

6.6.4	Abhängige Fehler bei Messwiederholung oder Clusterung	208
6.7	Partialkorrelation und Semipartialkorrelation	208
	t-Tests und Varianzanalysen	212
7.1	Tests auf Varianzhomogenität	212
7.1.1	<i>F-Test</i> auf Varianzhomogenität für zwei Stichproben	212
7.1.2	Levene-Test für mehr als zwei Stichproben	213
7.1.3	Fligner-KiEeen-Test für mehr als zwei Stichproben	214
7.2	t-Tests	215
7.2.1	<-Test für eine Stichprobe	215
7.2.2	t-Test für zwei unabhängige Stichproben	217
7.2.3	t-Test für zwei abhängige Stichproben	219
7.3	Einfaktorielle Varianzanalyse (CR- <i>p</i>)	220
7.3.1	Auswertung mit <code>oneway.test</code>	220
7.3.2	Auswertung mit <code>aov()</code>	221
7.3.3	Auswertung mit <code>anova</code>	223
7.3.4	Effektstärke schätzen	223
7.3.5	Voraussetzungen grafisch prüfen	224
7.3.6	Einzelvergleiche (Kontraste)	225
7.4	Einfaktorielle Varianzanalyse mit abhängigen Gruppen (RB- <i>p</i>)	231
7.4.1	Univariat formuliert auswerten und Effektstärke schätzen	232
7.4.2	Zirkularität der Kovarianzmatrix prüfen	235
7.4.3	Multivariat formuliert auswerten mit <code>AnovaO</code>	237
7.4.4	Multivariat formuliert auswerten mit <code>anovaO</code>	238
7.4.5	Einzelvergleiche und alternative Auswertungsmöglichkeiten	239
7.5	Zweifaktorielle Varianzanalyse (CRF- <i>p</i> ?)	239
7.5.1	Auswertung und Schätzung der Effektstärke	240
7.5.2	Quadratsummen vom Typ I, II und III	242
7.5.3	Bedingte Haupteffekte testen	246
7.5.4	Beliebige a-priori Kontraste	249
7.5.5	Beliebige post-hoc Kontraste nach Scheffe	252
7.5.6	Marginale Paarvergleiche nach Tukey	253
7.6	Zweifaktorielle Varianzanalyse mit zwei Intra-Gruppen Faktoren (RBF- <i>pq</i>) . . .	254
7.6.1	Univariat formuliert auswerten und Effektstärke schätzen	254
7.6.2	Zirkularität der Kovarianzmatrizen prüfen	258
7.6.3	Multivariat formuliert auswerten	259
7.6.4	Einzelvergleiche (Kontraste)	260
7.7	Zweifaktorielle Varianzanalyse mit Split-Plot-Design (SPF- <i>p</i> · <i>q</i>)	261
7.7.1	Univariat formuliert auswerten und Effektstärke schätzen	261
7.7.2	Voraussetzungen und Prüfen der Zirkularität	264
7.7.3	Multivariat formuliert auswerten	265
7.7.4	Einzelvergleiche (Kontraste)	266
7.7.5	Erweiterung auf dreifaktorielles SPF- <i>p</i> · <i>qr</i> Design	267
7.7.6	Erweiterung auf dreifaktorielles SPF- <i>pt</i> · <i>r</i> Design	269
7.8	Kovarianzanalyse	270
7.8.1	Test der Effekte von Gruppenzugehörigkeit und Kovariate	270
7.8.2	Beliebige a-priori Kontraste	276

7.8.3	Beliebige post-hoc Kontraste nach Scheffe	277
7.9	Power, Effektstärke und notwendige Stichprobengröße	278
7.9.1	Binomialtest	278
7.9.2	t-Test	280
7.9.3	Einfaktorielle Varianzanalyse	283
8	Regressionsmodelle für kategoriale Daten und Zähldaten	287
8.1	Logistische Regression	288
8.1.1	Modell für dichotome Daten anpassen	288
8.1.2	Modell für binomiale Daten anpassen	290
8.1.3	Anpassungsgüte	291
8.1.4	Vorhersage, Klassifikation und Anwendung auf neue Daten	293
8.1.5	Signifikanztests für Parameter und Modell	295
8.1.6	Andere Link-Funktionen	297
8.1.7	Mögliche Probleme bei der Modellanpassung	297
8.2	Ordinale Regression	298
8.2.1	Modellanpassung	299
8.2.2	Anpassungsgüte	300
8.2.3	Signifikanztests für Parameter und Modell	301
8.2.4	Vorhersage, Klassifikation und Anwendung auf neue Daten	303
8.3	Multinomiale Regression	304
8.3.1	Modellanpassung	305
8.3.2	Anpassungsgüte	306
8.3.3	Signifikanztests für Parameter und Modell	307
8.3.4	Vorhersage, Klassifikation und Anwendung auf neue Daten	308
8.4	Regression für Zähldaten	309
8.4.1	Poisson-Regression	309
8.4.2	Ereignisraten analysieren	311
8.4.3	Adjustierte Poisson-Regression und negative Binomial-Regression	312
8.4.4	Zero-inflated Poisson-Regression	314
8.4.5	Zero-truncated Poisson-Regression	317
8.5	Log-lineare Modelle	317
8.5.1	Modell	317
8.5.2	Modellanpassung	319
9	Survival-Analyse	324
9.1	Verteilung von Ereigniszeiten	324
9.2	Zensierte und gestutzte Ereigniszeiten	325
9.2.1	Zeitlich konstante Prädiktoren	326
9.2.2	Daten in Zählprozess-Darstellung	328
9.3	Kaplan-Meier-Analyse	331
9.3.1	Survival-Funktion und kumulatives hazard schätzen	331
9.3.2	Log-Rank-Test auf gleiche Survival-Funktionen	333
9.4	Cox proportional hazards Modell	334
9.4.1	Anpassungsgüte und Modelltests	337
9.4.2	Survival-Funktion und baseline hazard schätzen	338
9.4.3	Modelldiagnostik	340

9.4.4	Vorhersage und Anwendung auf neue Daten	344
9.4.5	Erweiterungen des Cox PH-Modells	345
9.5	Parametrische proportional hazards Modelle	345
9.5.1	Darstellung über die Hazard-Funktion	345
9.5.2	Darstellung als accelerated failure time Modell	346
9.5.3	Anpassung und Modelltests	347
9.5.4	Survival-Funktion schätzen	348
10	Klassische rionparametrische Methoden	351
10.1	Anpassungstests	351
10.1.1	Binomialtest	352
10.1.2	Test auf Zufälligkeit (Runs-Test)	353
10.1.3	Kolmogorov-Smirnov-Anpassungstest	355
10.1.4	χ^2 -Test auf eine feste Verteilung	358
10.1.5	χ^2 -Test auf eine Verteilungsklasse	359
10.2	Analyse von gemeinsamen Häufigkeiten kategorialer Variablen	361
10.2.1	χ^2 -Test auf Unabhängigkeit	361
10.2.2	χ^2 -Test auf Gleichheit von Verteilungen	362
10.2.3	χ^2 -Test für mehrere Auftretenswahrscheinlichkeiten	363
10.2.4	Fishers exakter Test auf Unabhängigkeit	364
10.2.5	Fishers exakter Test auf Gleichheit von Verteilungen	365
10.2.6	Kennwerte von (2×2) -Konfusionsmatrizen	366
10.2.7	ROC-Kurve und AUC	369
10.3	Maße für Zusammenhang und Übereinstimmung	371
10.3.1	Spearman's ρ und Kendalls τ	371
10.3.2	Zusammenhang kategorialer Variablen	373
10.3.3	Inter-Rater-Übereinstimmung	374
10.4	Tests auf gleiche Variabilität	382
10.4.1	Mood-Test	382
10.4.2	Ansari-Bradley-Test	383
10.5	Tests auf Übereinstimmung von Verteilungen	384
10.5.1	Kolmogorov-Smirnov-Test für zwei Stichproben	384
10.5.2	Vorzeichen-Test	386
10.5.3	Wilcoxon-Vorzeichen-Rang-Test für eine Stichprobe	387
10.5.4	Wilcoxon-Rangsummen-Test / Mann-Whitney- $\{$ -Test	389
10.5.5	Wilcoxon-Test für zwei abhängige Stichproben	390
10.5.6	Kruskal-Wallis- $\{$ -Test für unabhängige Stichproben	390
10.5.7	Friedman-Rangsummen-Test für abhängige Stichproben	392
10.5.8	Cochran-Q-Test für abhängige Stichproben	394
10.5.9	Bowker-Test für zwei abhängige Stichproben	395
10.5.10	McNemar-Test für zwei abhängige Stichproben	396
10.5.11	Stuart-Maxwell-Test für zwei abhängige Stichproben	398
11	Resampling-Verfahren	400
11.1	Bootstrapping	400
11.1.1	Replikationen erstellen	401
11.1.2	Bootstrap-Vertrauensintervalle für β	404

te	11.1.3 Bootstrap-Vertrauensintervalle für	406
■	11.1.4 Lineare Modelle: case resampling	407
j'''	11.1.5 Lineare Modelle: model-based resampling	409
	11.1.6 Lineare Modelle: wild bootstrap	411
p	11.2 Permutationstests	412
	11.2.1 Test auf gleiche Lageparameter in unabhängigen Stichproben	413
	11.2.2 Test auf gleiche Lageparameter in abhängigen Stichproben	415
	11.2.3 Test auf Unabhängigkeit von zwei Variablen	416
t		
12	Multivariate Verfahren	418
	12.1 Lineare Algebra	418
	12.1.1 Matrix-Algebra	418
	12.1.2 Lineare Gleichungssysteme lösen	422
	12.1.3 Norm und Abstand von Vektoren und Matrizen	422
	12.1.4 Mahalanobistransformation und Mahalanobisdistanz	424
	12.1.5 Kennwerte von Matrizen	426
	12.1.6 Zerlegungen von Matrizen	428
	12.1.7 Orthogonale Projektion	430
	12.2 Hauptkomponentenanalyse	433
	12.2.1 Berechnung	434
	12.2.2 Dimensionsreduktion	437
	12.3 Faktorenanalyse	439
	12.4 Multidimensionale Skalierung	446
	12.5 Multivariate multiple Regression	447
	12.6 Hotellings T^2	449
	12.6.1 Test für eine Stichprobe	449
	12.6.2 Test für zwei unabhängige Stichproben	451
	12.6.3 Test für zwei abhängige Stichproben	453
	12.6.4 Univariate Varianzanalyse mit abhängigen Gruppen (RB- p)	454
	12.7 Multivariate Varianzanalyse (MANOVA)	455
	12.7.1 Einfaktorielle MANOVA	455
	12.7.2 Zweifaktorielle MANOVA	456
	12.8 Diskriminanzanalyse	457
	12.9 Das allgemeine lineare Modell	462
	12.9.1 Modell der multiplen linearen Regression	462
	12.9.2 Modell der einfaktoriellen Varianzanalyse	464
	12.9.3 Modell der zweifaktoriellen Varianzanalyse	469
	12.9.4 Parameterschätzungen, Vorhersage und Residuen	473
	12.9.5 Hypothesen über parametrische Funktionen testen	475
	12.9.6 Lineare Hypothesen als Modellvergleiche formulieren	475
	12.9.7 Lineare Hypothesen testen	480
	12.9.8 Beispiel: Multivariate multiple Regression	483
	12.9.9 Beispiel: Einfaktorielle MANOVA	485
	12.9.10 Beispiel: Zweifaktorielle MANOVA	488

13 Vorhersagegüte prädiktiver Modelle	492
13.1 Kreuzvalidierung linearer Regressionsmodelle	492
13.1.1 f_c -fache Kreuzvalidierung	493
13.1.2 Leave-One-Out Kreuzvalidierung	494
13.2 Kreuzvalidierung verallgemeinerter linearer Modelle	495
13.3 Bootstrap-Vorhersagefehler	496
14 Diagramme erstellen	499
14.1 Grafik-Devices	499
14.1.1 Aufbau und Verwaltung von Grafik-Devices	499
14.1.2 Grafiken speichern	501
14.2 Streu- und Liniendiagramme	502
14.2.1 Streudiagramme mit <code>plotO</code>	502
14.2.2 Datenpunkte eines Streudiagramms identifizieren	504
14.2.3 Streudiagramme mit <code>matplot()</code>	505
14.3 Diagramme formatieren	505
14.3.1 Grafikelemente formatieren	505
14.3.2 Farben spezifizieren	508
14.3.3 Achsen formatieren	511
14.4 Säulen- und Punktdiagramme	511
14.4.1 Einfache Säulendiagramme	512
14.4.2 Gruppierte und gestapelte Säulendiagramme	512
14.4.3 Dotchart	515
14.5 Elemente einem bestehenden Diagramm hinzufügen	516
14.5.1 Koordinaten in einem Diagramm identifizieren	517
14.5.2 In beliebige Diagrammbereiche zeichnen	518
14.5.3 Punkte	519
14.5.4 Linien	520
14.5.5 Polygone	522
14.5.6 Punktionsgraphen	525
14.5.7 Text und mathematische Formeln	526
14.5.8 Achsen	528
14.5.9 Fehlerbalken	529
14.5.10 Rastergrafiken	532
14.6 Verteilungsdiagramme	534
14.6.1 Histogramm und Schätzung der Dichtefunktion	534
14.6.2 Stamm-Blatt-Diagramm	536
14.6.3 Boxplot	537
14.6.4 Stripchart	539
14.6.5 Quantil-Quantil-Diagramm	540
14.6.6 Empirische kumulierte Häufigkeitsverteilung	542
14.6.7 Kreisdiagramm	542
14.6.8 Gemeinsame Verteilung zweier Variablen	543
14.7 Daten interpolieren und fiten	547
14.7.1 Lineare Interpolation und LOESS-Glätter	547
14.7.2 Splines	548

14.8	Multivariate Daten visualisieren	549
14.8.1	Höhenlinien und variable Datenpunktsymbole	550
14.8.2	Dreidimensionale Gitter und Streudiagramme	552
14.8.3	Bedingte Diagramme für mehrere Gruppen mit <code>ggplot2</code>	553
14.8.4	Bedingte Diagramme für mehrere Gruppen mit <code>lattice</code>	560
14.8.5	Matrix aus Streudiagrammen	561
14.8.6	Heatmap	563
14.9	Mehrere Diagramme in einem Grafik-Device darstellen	565
14.9.1	<code>layoutO</code>	565
14.9.2	<code>parCmfrow</code> , <code>mfc</code> , <code>fig</code>)	567
14.9.3	<code>split.screenO</code>	569
15	R als Programmiersprache	571
15.1	Kontrollstrukturen	571
15.1.1	Fallunterscheidungen	571
15.1.2	Schleifen	574
15.2	Eigene Funktionen erstellen	577
15.2.1	Funktionskopf	577
15.2.2	Funktionsrumpf	578
15.2.3	Fehler behandeln	579
15.2.4	Rückgabewert und Funktionsende	581
15.2.5	Eigene Funktionen verwenden	581
15.2.6	Generische Funktionen	582
15.3	Funktionen analysieren und verbessern	584
15.3.1	Quelltext fremder Funktionen begutachten	584
15.3.2	Funktionen zur Laufzeit untersuchen	585
15.3.3	Effizienz von Auswertungen steigern	587
	Literaturverzeichnis	590