

# Index

## Data sets

affixProductivity, 118  
 alice, 65  
 auxiliaries, 104  
 beginningReaders, 290, 301  
 dative, 4, 33, 148, 279  
 durationsGe, 116  
 durationsOnt, 75, 117  
 dutchSpeakersDist, 136  
 dutchSpeakersDistMeta, 136  
 english, 42, 43, 117, 169, 195, 228  
 etymology, 209, 238  
 faz, 214  
 finalDevoicing, 164, 320  
 havelaar, 52  
 heid, 16, 42  
 imaging, 240  
 latinsquare, 266  
 lexdec, 25, 242  
 lexicalMeasures, 138, 164, 314  
 nesscg, 239  
 nessdemog, 239  
 nessw, 239  
 oldFrench, 129, 160  
 oldFrenchMeta, 129, 160  
 periphrasticDo, 218  
 phylogeny, 143  
 primingHeid, 284  
 ratings, 21, 82, 165  
 regularity, 164, 202  
 selfPacedReadingHeid, 287, 301  
 sizeRatings, 302, 333  
 spanish, 155  
 spanishFunctionWords, 317  
 spanishMeta, 19, 154, 303  
 twente, 231  
 variationLijk, 134  
 ver, 71  
 verbs, 4, 32  
 warlpiri, 42, 301, 304  
 weightRatings, 39  
 writtenVariationLijk, 295, 301

## R

\$, 6, 12, 90  
 &, |, 9  
 ~, 13  
 |, 9, 38  
 ^, 96  
 ^, 168  
 -> (assignment), 3  
 : (sequence operator), 8  
 <- (assignment), 3  
 == (equality), 8  
 = (assignment), 3, 8  
 HPDinterval(), 248  
 I(), 96  
 MASS package, 23, 97  
 N(), 233  
 TukeyHSD(), 106  
 Vm(), 239  
 .RData, 1, 18  
 .Rhistory, 1, 18  
 abline(), 59, 72, 85, 87  
 abs(), 189  
 aggregate(), 17, 263  
 anova(), 104, 166, 200, 203, 253  
 aov(), 107, 264  
 apply(), 272, 283  
 as.character(), 10, 36  
 as.dist(), 136  
 as.factor(), 10  
 as.numeric(), 51, 132  
 attach(), 293  
 attr(), 61  
 barplot(), 21, 32

- 
- biplot(), 124
  - boxplot(), 30
  - bwplot(), 43, 79
  - c(), 7
  - cat(), 190
  - cbind(), 99, 197
  - cex, 36
  - chisq.test(), 74, 113, 163, 172, 198
  - cluster package, 140
  - cmdscale(), 136
  - coda package, 248
  - coef(), 87
  - col, 21
  - collin.fnc(), 182
  - colnames(), 9
  - compare.richness.fnc(), 225, 239
  - confint(), 115
  - consensus() (ape package), 147
  - contr.treatment(), 102
  - cor(), 90, 139, 172
  - cor.test(), 90, 139
  - corres.fnc(), 130
  - corsup.fnc(), 133
  - cut(), 65
  - cutree(), 142
  - data.frame(), 19, 65, 90
  - datadist(), 171
  - dbinom(), 47, 49, 53
  - demo(), 99
  - density(), 25, 99, 172, 188
  - Design, 171
  - detach(), 23
  - dev.off(), 24
  - deviance(), 217
  - dfbetas(), 190
  - dffits(), 189
  - diag(), 162
  - diana(), 138, 140
  - dist(), 129, 140
  - dnorm(), 59
  - dpois(), 54
  - dt(), 63
  - e1071 package, 160
  - equal.count(), 42
  - example(), 5
  - exp(), 5
  - fastbw(), 186, 320
  - fisher.test(), 113
  - fitted(), 110, 172, 247, 297
  - fixef(), 293
  - glm(), 197
  - grid::grid.prompt(), 243
  - growth.fnc(), 223
  - hclust(), 138, 140
  - head(), 4, 223
  - help(), 5
  - install.packages(), 23
  - jitter(), 74, 79, 310
  - jpeg(), 24
  - kappa(), 182
  - kde2d(), 99
  - kruskal.test(), 108
  - ks.test(), 73
  - lattice, 242
  - lattice (package), 123
  - lda(), 155
  - length(), 19, 43, 63
  - levels(), 13
  - library(), 23
  - lines(), 27, 59, 111
  - list(), 16
  - lm(), 86, 107, 165, 174, 262
  - lmer(), 242
  - lmList(), 271
  - lmsreg(), 92
  - lnre(), 232, 326
  - lnre.spc(), 233
  - lnre.vgc(), 235, 327
  - log(), 5
  - lowess(), 34, 93
  - lrm(), 210, 238, 279, 320
  - lty, 59
  - make.reg.fnc(), 270
  - manova(), 158
  - max(), 22
  - mcmcscamp(), 248
  - mean(), 15, 22
  - median(), 22
  - merge(), 17
  - mfrow, 24
  - mfrow(), 175
  - min(), 22, 293
  - mosaicplot(), 33, 42
  - mtext(), 29
  - mvrnorm(), 97
  - mvrnormplot.fnc(), 89
  - names(), 121
  - nchar(), 12
  - nj() (ape package), 143
  - nodelabels() (ape package), 147
  - nrow(), 14
  - objects(), 18
  - ols(), 171
  - options(), 171
  - order(), 11
  - ordered(), 209
  - pairs(), 37, 164
  - pairscor.fnc(), 164, 313
  - panel.abline(), 249
  - panel.xyplot(), 249
  - par(), 24
  - paste(), 30

`pbinom()`, 49, 52, 53  
`pchisq()`, 65, 68, 74  
`pdf()`, 24  
`pentrace()`, 205  
`persp()`, 99  
`pf()`, 64, 68  
`plclust()`, 140  
`plot()`, 27, 150  
`plot.logistic.fit.fnc()`, 281  
`plot.xmean.ordinaly()`, 213  
`plotcp()`, 151  
`png()`, 24  
`pnorm()`, 59, 60, 73  
`pol()`, 175  
`postscript()`, 24, 25  
`ppois()`, 54  
`prcomp()`, 120, 184  
`predict()`, 96, 153, 156, 200  
`prop.clades()`, 147  
`prop.table()`, 15, 111  
`prop.test()`, 163, 318  
`prune()`, 151  
`pt()`, 63, 68  
`pvals.fnc()`, 248  
`q()`, 18  
`qbinom()`, 49, 52, 53  
`qnorm()`, 59  
`qpois()`, 54, 67  
`qqline()`, 172  
`qqmath()`, 242  
`qqnorm()`, 72, 172, 188  
`qt()`, 63, 229  
`quantile()`, 28, 52, 67, 172, 283  
`quasiF.fnc()`, 262  
`ranef()`, 246  
`range()`, 22, 26  
`rbind()`, 185  
`rbinom()`, 49, 51, 53  
`rcs()`, 177  
`read.table()`, 5  
`resid()`, 172, 213  
`rlnorm()`, 100  
`rm()`, 18  
`rnorm()`, 59, 81, 114  
`round()`, 52, 55, 93  
`rownames()`, 9  
`rpart()`, 150  
`rpois()`, 54, 100  
`rt()`, 63  
`scale()`, 61, 290  
`scan()`, 222  
`sd()`, 61, 172  
`seq()`, 29, 50, 95, 293  
`simulateLatinsquare.fnc()`, 269  
`simulateRegression.fnc()`, 274  
`simulateSplitPlot.fnc()`, 265  
`somers2()`, 153, 281  
`sort()`, 11  
`spc()`, 231, 326  
`splom()`, 123  
`sqrt()`, 4  
`substr()`, 110  
`sum()`, 14  
`summary()`, 89, 121  
`svm()`, 160, 164  
`t()`, 155  
`t.test()`, 75, 79, 82, 103  
`table()`, 43  
`tail()`, 223  
`tapply()`, 15, 81, 107  
`text()`, 36, 150  
`tolower()`, 222, 325  
`toupper()`, 143  
`truehist()`, 23, 43  
`unique()`, 17  
`update()`, 206  
`validate()`, 193  
`var()`, 62  
`var.test()`, 81  
`varclus()`, 182  
`which.influence()`, 190  
`wilcox.test()`, 76, 80, 82, 84  
`with()`, 16  
`write.table()`, 5  
`xaxt`, 29  
`xlab`, 21  
`xlim`, 27  
`xtabs()`, 13, 42, 43, 51, 66  
`xylowess.fnc()`, 40  
`xyplot()`, 40  
`ylim`, 27  
`zipf.fnc()`, 228

## Topic index

$\alpha$ -level, 68, 105  
 agglomerative clustering, 138  
 Akaike Information Criterion, 206  
 alternative hypothesis, 75  
 analysis of covariance, 108  
 analysis of variance, 101, 107  
 anticonservative, 248  
 arithmetic operators, 2

- assignment (=, <-, ->), 3
- bar plot, 21, 32
- bimodal, 305
- bimodal density, 78
- bimodal distribution, 72
- binomial distribution, 197, 296
- binomial random variable, 46
- biplot, 124, 129
- bivariate standard normal distribution, 87
- BLUP, 247
- Bonferroni correction, 106
- bootstrap, 146, 193, 204
- boxplot, 30
- breakpoint, 215
- British National Corpus, 239
- Brown corpus, 45
- by-item regression, 271
- canceling a command (CONTROL-C, ESC), 2
- CART analysis, 148
- CELEX, 16, 44
- $\chi^2$ -distribution, 63
- chi-squared distance, 129
- chi-squared test, 74, 113
- classification, 148
- classification trees, 148
- clustering, 148
- coefficients, 87
- collinearity, 181
- comments (#), 3
- conditioning plot, 40
- confidence interval, 75, 80
- confidence intervals, 115, 229
- confound, 273
- contingency table, 13, 129
- continuous distribution, 74
- continuous random variable, 44, 57
- correlated, 33
- correlation, 87
- correlation coefficient, 87
- correlation matrix, 125, 139
- correlation test, 90
- correspondence analysis, 129
- cost-complexity pruning, 150
- covariance, 99
- covariance matrix, 125
- CRAN, X
- cross-entropy, 136
- cross-validation, 158, 162
- crossed, 260
- cumulative distribution function, 52, 53, 60, 64
- data frame, 5
- dative alternation, 4
- deciles, 28
- default level, 102
- degrees of freedom, 63, 71
- density, 26
- density estimation, 25
- dependent variable, 13
- deviance residuals, 198
- dfbetas, 190
- dffits, 189
- discrete random variable, 44
- distance matrices, 129
- distances, 136
- divisive clustering, 138
- dummy coding, 102, 239
- Dutch, 16, 42, 52, 71, 75, 104, 164, 202
- eigenvalue rates, 131
- encapsulated PostScript, 25
- English, 4, 169, 239
- equality, 8
- error stratum, 264
- explained variance, 88
- F*-distribution, 63
- factor, 9
- factor analysis, 126
- factor rotation, 127
- fast backwards elimination, 186
- Fisher's exact test of independence, 113
- fitted values, 110
- for loop, 100, 190
- formula, 13, 86, 109
- fractional degrees of freedom, 79
- frequency function, 47
- frequency spectrum, 230
- functions, 4
- generalized linear mixed model, 279
- generalized linear model, 197
- German, 214
- graphical parameters: see `par()`, 24
- grid prompt, 243
- grouping factor, 37
- grouping operator, 38
- growth curve of the vocabulary, 222
- hapax legomena, 223
- Herdan's law, 226
- heteroskedasticity, 33, 188
- high-density line, 48, 51
- highest posterior density intervals, 248
- histogram, 21, 23
- independent random variables, 77
- independent variable, 13
- index of concordance, 281
- indicator variable, 216

- inflation in surprise, 106  
interaction, 42, 109, 154, 166  
intercept, 59, 85, 103  
inverse, 52, 60  
inverse transformation, 32
- jitter, 79  
jpeg, 24
- knots (of spline), 177  
Kolmogorov-Smirnov test, 73, 79  
Kruskal-Wallis rank sum test, 108
- latent semantic analysis, 127  
latent variable, 128  
Latin Square design, 267  
law of large numbers, 229  
least squares, 171  
least squares regression, 86  
levels (of a factor), 9  
leverage, 189  
lexical richness, 222  
linear combination, 96  
linear discriminant analysis, 155  
linear discriminants, 154  
linear model, 86, 96  
linearity assumption, 95  
link function, 196  
list, 16  
LNRE distributions, 229  
loadings, 124  
log link function, 296, 297  
logarithmic transformation, 31, 71, 92  
logit, 196  
lognormal distribution, 223  
lognormal random variable, 100  
lognormal-Poisson distribution, 100  
long data format, 5, 202
- Markov chain Monte Carlo sampling, 248  
maximum likelihood, 195  
mean, 58  
mean squared error, 194  
median, 21, 28  
missing data, 133  
mixed-effects regression, 270  
mode, 21, 305  
model criticism, 71  
model likelihood, 204  
mosaic plot, 33, 112, 305  
multicollinearity, 37  
multidimensional scaling, 136  
multimodal, 140  
multiple comparisons, 105  
multivariate analysis of variance, 158  
multivariate data, 118
- negative subscripting, 37  
neighbor joining, 143  
nested, 261  
noise, 74  
non-parametric test, 77  
non-sequential ANOVA table, 175  
normal distribution, 58  
null deviance, 198, 204  
null-hypothesis, 68, 75, 115
- one-tailed test, 70, 71, 75, 290  
one-way analysis of variance, 101  
optimism, 194  
ordered factor, 209  
ordinal logistic regression, 209  
orthogonal predictors, 181  
outliers, 27, 91, 92, 188, 190, 311  
overdispersion, 199
- paired observations, 82  
paired random variables, 77  
paired *t*-test, 83  
parabola, 95  
parameters of the binomial distribution, 46  
parametric test, 77  
partial effects, 175  
partitioning, 138  
pdf, 24  
penalized maximum likelihood estimation, 205  
penalty, 205  
perspective plot, 99  
phylogenetic classification, 142  
phylogeny estimation, 143  
png, 24  
Poisson, 54  
Poisson distribution, 54, 100, 296  
polynomial, 175  
population, 46  
population probabilities, 48  
posterior distribution, 248  
PostScript, 24  
power, 69, 77, 265  
predictor, 13  
principal components analysis, 119, 163  
probability, 44  
probability density function, 47, 53, 65  
probability distribution, 20, 44  
probability of failure, 46  
probability of success, 46  
productivity, 118  
prompting, 243  
proportional odds model, 212  
proportions test, 163
- quadratic term, 95  
quantile function, 52, 53, 60

- quantile-quantile plot, 53, 72  
 quartiles, 28  
 quasi-*F*, 262
- $\rho$ , 87  
 $r$ , 87  
 $r_s$ , 91  
*R*-squared, 88  
 random intercepts, 247  
 random noise, 114  
 random number generator, 53  
 random numbers, 51, 81  
 random regression, 271  
 random slopes, 248  
 random variable, 20, 44  
 rank-frequency step function, 228  
 recursive partitioning, 149  
 reference level, 102  
 register variation, 118  
 regression line, 87  
 regression towards the mean, 277  
 rejection regions, 75  
 relative frequency, 44  
 remainder operator, 65  
 repeatable factors, 241  
 residual deviance, 199, 204  
 residual standard error, 90, 172  
 residuals, 172  
 restricted cubic spline, 176  
 rotation matrix, 124
- S-PLUS, xii  
 sample, 46, 51  
 scatterplot, 33, 84  
 scatterplot matrix, 37  
 scatterplot smoother, 34, 39  
 semantic transparency, 78  
 sequence, 8  
 sequential ANOVA table, 167  
 Shapiro-Wilk test for normality, 73, 76  
 shingle, 42  
 shrinkage, 205, 275, 277  
 significance level, 68  
 simple main effect, 165  
 skew, 71  
 skewed distributions, 76  
 skewness, 92
- slope, 59, 85  
 Somers'  $D_{xy}$ , 281  
 sorting, 10  
 Spearman correlation, 91, 140  
 standard deviation, 58, 121  
 standard error, 89  
 standard normal distribution, 59, 62  
 standardization, 61  
 statistical significance, 68, 114  
 string, 8  
 subscripting, 6, 7, 130  
 supervised methods, 118  
 supplementary data, 134  
 supplementary rows, columns, 133  
 support vector machines, 160
- t*-distribution, 63  
*t*-test, 75, 79, 103  
*t*-value, 89  
 test statistic, 68, 73–75  
 tick marks, 28  
 ties, 74, 79  
 tokens, 222  
 treatment, 275  
 treatment coding, 102  
 Tukey's HSD test, 106  
 two-tailed test, 70, 71  
 type I error rate, 265  
 type-token ratio, 223, 224  
 types, 222
- uniform random variable, 57  
 unsupervised methods, 118
- validation, 193, 204  
 variable, 3  
 variance, 62  
 vector, 6–8  
 vocabulary growth rate, 223  
 vocabulary richness, 222
- Warlpiri, 42, 301, 305  
 Welch *t*-test, 79  
 Wilcoxon test, 76, 80, 84  
 WordNet, 21
- Zipf's law, 226

### Author index

- Balota, 40  
 Bates, x, 242  
 Becker, ix  
 Belsley, 181, 183
- Biber, 126  
 Bresnan, 4, 13, 32, 81, 116, 148, 279  
 Burrows, 163

- 
- Carroll, J.B., 224  
 Chambers, ix  
 Chitashvili, 165  
 Clark, 263  
 Cochran, 262  
 Cortese, 40  
 Crawley, ix, x, 200  
 Cueni, 148, 279
- Dalgaard, ix  
 Dalton-Puffer, 239  
 Dickinson, 263  
 Dumas, 127  
 Dunn, 142
- Ellegård, 218  
 Ernestus, 75, 116, 117, 129, 134, 164, 295
- Faraway, 242  
 Feldman, 40  
 Foley, 142  
 Forster, 263  
 Francis, 20, 45  
 Frauenfelder, 116
- Gale, 230  
 Good, 223, 230  
 Gremmen, 260  
 Guiraud, 224
- Haerdle, 25  
 Harrell, 195, 221  
 Hellwig, 116  
 Herdan, 224, 226  
 Hoover, 224
- Keune, 134, 295  
 Khmaladze, 229  
 Kroch, 219  
 Kučera, 20, 45  
 Kuh, 181
- Landauer, 127  
 Levinson, 142  
 Lieber, 71  
 Lorch, 271
- Miller, 21  
 Moscoso del Prado, 208
- Mulken, 129  
 Murtagh, 129  
 Myers, 271
- Nikitina, 148, 279
- O'Shannessy, 42, 301  
 Orlov, 226
- Paradis, 143, 146  
 Perdijk, 289  
 Plag, 239  
 Pluymaekers, 75, 116, 117  
 Pollman, 214
- Quené, 272
- Raaijmakers, 260, 266, 268  
 Reesink, 142  
 Ripley, xi, 23, 25, 35
- Sampson, 230  
 Satterthwaite, 262  
 Schreuder, 40, 116, 202, 208, 209  
 Schrijnemakers, 260  
 Sergent-Marschall, 40  
 Sichel, 224  
 Spassova, 154  
 Spieler, 40  
 Sproat, 12
- Tabak, 202, 208, 209  
 Terrill, 142  
 Tweedie, 224
- Van den Bergh, 272  
 Van Hout, 134, 295  
 Venables, xi, 23, 25, 35  
 Verzani, xi  
 Vulcanović, 219, 221
- Welsch, 181  
 Wilks, xi
- Yap, 40  
 Yule, 224
- Zipf, 42, 224, 226, 306