
Contents

Preface	ix
1 Learning from data, and tools for the task	1
1.1 <i>The changing world of statistical applications</i>	1
1.1.1 Models, algorithms, and machines that learn	2
1.1.2 Science and statistics	4
1.2 <i>Statistical analysis questions, aims and strategies</i>	5
1.2.1 Terminology — variables, factors, and more!	5
1.2.2 Graphical comparisons	6
1.2.3 Formal model-based comparison	7
1.2.4 How will data be used?	9
1.2.5 The planning of data analysis	9
1.2.6 Subject area knowledge and judgments	10
1.2.7 Notes on sampling from finite populations	10
1.3 <i>Using graphs to make sense of data</i>	11
1.3.1 One-way layouts, perhaps broken down by groups within the data	12
1.3.2 Patterns in univariate time series	14
1.3.3 Bivariate data	15
1.3.4 *Multiple variables and times	17
1.3.5 Graphical displays for categorical data	20
1.3.6 What to look for in plots	21
1.4 <i>Data Summary</i>	22
1.4.1 Counts	23
1.4.2 Summaries of information from data frames	25
1.4.3 Measures of spread — standard deviation and inter-quartile range	27
1.4.4 Correlation	29
1.5 <i>Distributions: models for the random component</i>	30
1.5.1 Discrete distributions	31
1.5.2 Continuous distributions	33
1.5.3 Use of simulation to estimate sampling distributions	34
1.5.4 Graphical checks for normality	36
1.6 <i>Organizing and managing work; reproducible reporting</i>	39
1.6.1 Reproducible reporting — the knitr package	40

1.7	<i>Further Reading and Study</i>	40
1.8	<i>Exercises</i>	41
2	Generalizing from models	46
2.1	<i>Model assumptions</i>	46
2.1.1	Random sampling assumptions – independence	46
2.1.2	The role of non-parametric methods	47
2.1.3	Have all relevant effects been accounted for?	47
2.1.4	The limitations of models	49
2.2	<i>Populations, samples, and sampling distributions</i>	49
2.2.1	Population parameters and sample statistics	49
2.2.2	Using the sample as a window into the population	50
2.2.3	Assessing accuracy – the standard error	51
2.3	<i>Estimation and hypothesis testing</i>	52
2.3.1	The sampling distribution of the t statistic	53
2.3.2	Comparison of means — confidence intervals and tests	54
2.3.3	Comparison of binomial proportions	58
2.3.4	The Pearson or product–moment correlation	59
2.3.5	Contingency tables	59
2.4	<i>A critique of p-value methodology</i>	63
2.4.1	From $p \leq \alpha$ to the probability of an effect?	64
2.4.2	*A more detailed analysis — Positive Predictive Values	65
2.4.3	Interpreting the specific p -value	67
2.4.4	Reproducibility studies	68
2.5	* <i>Resampling methods for standard errors, tests and confidence intervals</i>	69
2.5.1	The one-sample permutation test	69
2.5.2	The two-sample permutation test	69
2.5.3	*Estimating the standard error of the median: bootstrapping	70
2.5.4	Bootstrap estimates of confidence intervals	72
2.6	<i>Regression with a single explanatory variable</i>	73
2.6.1	Straight line regression	73
2.6.2	Fitting models – the model formula	75
2.6.3	The model matrix in regression	78
2.6.4	Iron slag example — check residuals with care!	78
2.6.5	The analysis of variance table	80
2.6.6	Outliers, influence, and robust regression	81
2.6.7	Standard errors and confidence intervals	83
2.6.8	There are two regression lines!	86
2.6.9	*Logarithmic and Power Transformations	86
2.6.10	General forms of non-linear response	88
2.6.11	Size and shape data – allometric growth	88
2.7	<i>Empirical assessment of predictive accuracy</i>	89
2.7.1	The training/test approach, and cross-validation	90
2.7.2	* Bootstrapping	92
2.8	<i>One- and two-way comparisons</i>	95

2.8.1	One-way comparisons	95
2.8.2	Regression versus qualitative anova comparisons – issues of power	98
2.8.3	*Severe multiplicity — the false discovery rate	99
2.8.4	Data with a two-way structure, i.e., two factors	102
2.8.5	Presentation issues	102
2.9	<i>Data with a nested variation structure</i>	103
2.9.1	Degrees of freedom considerations	104
2.9.2	General multi-way analysis of variance designs	105
2.10	* <i>Theories of inference</i>	105
2.10.1	Maximum likelihood estimation	105
2.10.2	Bayesian estimation	106
2.10.3	If there is strong prior information, use it!	107
2.11	* <i>Bayesian regression estimation using the MCMCpack package</i>	107
2.12	<i>Recap</i>	109
2.13	<i>Further reading</i>	110
2.14	<i>Exercises</i>	112
3	Multiple linear regression	118
3.1	<i>Basic ideas: the allbacks book weight data</i>	118
3.1.1	Omission of the intercept term	121
3.1.2	Diagnostic plots	121
3.2	<i>The interpretation of model coefficients</i>	122
3.2.1	Times for Northern Irish hill races	122
3.2.2	Book dimensions, density, and book weight — the oddbooks dataset	126
3.2.3	Mouse brain weight example	128
3.2.4	Issues for obtaining interpretable coefficients	129
3.3	<i>Choosing the model, and checking it out</i>	130
3.3.1	*A more formal approach to the choice of transformation	133
3.3.2	Accuracy estimates, for fitted values and for new observations	135
3.3.3	Choosing the model — deaths from Atlantic hurricanes	137
3.3.4	Strategies for fitting models — suggested steps	139
3.4	<i>Robust regression, outliers, and influence</i>	140
3.4.1	Making outliers obvious — robust regression	141
3.4.2	Leverage, influence, and Cook’s distance	144
3.5	<i>Assessment and comparison of regression models</i>	146
3.5.1	R^2 and adjusted R^2	146
3.5.2	Information measures and model fit	146
3.5.3	Model Comparison using information statistics and/or anova — the nihills data	147
3.5.4	Training/test approaches, and cross-validation	148
3.5.5	Further points and issues	150
3.6	<i>Problems with many explanatory variables</i>	152
3.6.1	Variable selection issues	152
3.7	<i>Multicollinearity</i>	155
3.7.1	The variance inflation factor (VIF)	157

3.8	<i>Errors in x</i>	159
3.9	<i>Multiple regression models – additional points</i>	163
3.9.1	Confusion between explanatory and response variables	163
3.9.2	Missing explanatory variables	164
3.9.3	*Added variable plots	165
3.9.4	* Non-linear methods – an alternative to transformation?	168
3.10	<i>Recap</i>	170
3.11	<i>Further reading</i>	170
3.12	<i>Exercises</i>	172
4	Exploiting the linear model framework	175
4.1	<i>Levels of a factor – using indicator variables</i>	176
4.1.1	Example – sugar weight	176
4.1.2	Different choices for the model matrix when there are factors	178
4.2	<i>Block designs and balanced incomplete block designs</i>	179
4.2.1	Analysis of the rice data, allowing for block effects	179
4.2.2	A balanced incomplete block design	181
4.3	<i>Fitting multiple lines</i>	183
4.4	<i>Methods for fitting smooth curves</i>	186
4.4.1	Polynomial Regression	186
4.4.2	Regression spline bases	189
4.4.3	Other smoothing methods	194
4.4.4	*Quantile regression	194
4.5	* <i>Generalized additive models (GAMs) — Roughness penalty methods</i>	196
4.5.1	Example — the fruitohms data	197
4.5.2	Multiple explanatory variables — dewpoint data	198
4.5.3	*Atlantic hurricanes that made landfall in the US	201
4.6	<i>Further reading</i>	203
4.7	<i>Exercises</i>	203
5	Generalized linear models and survival analysis	207
5.1	<i>Generalized linear models</i>	207
5.1.1	Transformation of the expected value on the left	207
5.1.2	Noise terms need not be normal	208
5.1.3	*Extra-binomial or extra-poisson variation	208
5.1.4	Log odds in contingency tables	209
5.1.5	Logistic regression with a continuous explanatory variable	210
5.2	<i>Logistic multiple regression</i>	212
5.2.1	Selection of model terms, and fitting the model	215
5.2.2	Fitted values	216
5.2.3	A plot of contributions of explanatory variables	217
5.2.4	Cross-validation estimates of predictive accuracy	218
5.3	<i>Logistic models for categorical data – an example</i>	218
5.4	<i>Models for counts — Poisson, quasi-Poisson, and negative binomial</i>	220
5.4.1	Data on aberrant crypt foci	220

5.4.2	Moth habitat example	222
5.4.3	*Models with negative binomial errors	226
5.4.4	*Negative binomial versus alternatives — the hurricane deaths data	230
5.5	<i>Additional notes on generalized linear models</i>	232
5.5.1	* Residuals, and estimating the dispersion	232
5.5.2	Standard errors and z - or t -statistics for binomial models	234
5.5.3	Leverage for binomial models	234
5.6	<i>Models with an ordered categorical or categorical response</i>	235
5.6.1	Ordinal Regression Models	235
5.6.2	* Loglinear Models	238
5.7	<i>Survival analysis</i>	238
5.7.1	Analysis of the Aids2 data	239
5.7.2	Right censoring prior to the termination of the study	240
5.7.3	The survival curve for male homosexuals	241
5.7.4	Hazard rates	242
5.7.5	The Cox proportional hazards model	242
5.8	<i>Transformations for proportions and counts</i>	244
5.9	<i>Further reading</i>	245
5.10	<i>Exercises</i>	246
6	Time series models	249
6.1	<i>Time series – some basic ideas</i>	249
6.1.1	Preliminary graphical explorations	249
6.1.2	The autocorrelation and partial autocorrelation function	250
6.1.3	Autoregressive (AR) models	251
6.1.4	* Autoregressive moving average (ARMA) models – theory	253
6.1.5	Automatic model selection?	254
6.1.6	A time series forecast	255
6.2	<i>Regression modeling with ARIMA errors</i>	257
6.3	* <i>Nonlinear time series</i>	263
6.4	<i>Further reading</i>	265
6.5	<i>Exercises</i>	266
7	Multi-level models, and repeated measures	268
7.1	<i>Corn yield data — analysis using <code>aov()</code></i>	270
7.1.1	A More Formal Approach	273
7.2	<i>Analysis using <code>lme4::lmer()</code></i>	275
7.3	<i>Survey data, with clustering</i>	279
7.3.1	Alternative models	279
7.3.2	Instructive, though faulty, analyses	283
7.3.3	Predictive accuracy	284
7.4	<i>A multi-level experimental design</i>	285
7.4.1	The anova table	286
7.4.2	Expected values of mean squares	287
7.4.3	* The analysis of variance sums of squares breakdown	288

7.4.4	The variance components	290
7.4.5	The mixed model analysis	291
7.4.6	Predictive accuracy	293
7.5	<i>Within and between subject effects</i>	294
7.5.1	Model selection	295
7.5.2	Estimates of model parameters	297
7.6	<i>A mixed model with a beta-binomial error</i>	298
7.6.1	The beta-binomial	298
7.6.2	Diagnostic checks	301
7.6.3	LT99s and confidence intervals – model comparisons	302
7.6.4	Lethal time estimates and confidence intervals	303
7.7	<i>Observation level random effects — the moths dataset</i>	305
7.8	<i>Repeated measures in time</i>	307
7.8.1	Example – random variation between profiles	309
7.8.2	Orthodontic measurements on children	312
7.9	<i>Further notes on multi-level and other models with correlated errors</i>	316
7.9.1	Different sources of variance – complication or focus of interest?	316
7.9.2	Predictions from models with a complex error structure	317
7.9.3	An historical perspective on multi-level models	317
7.9.4	Meta-analysis	318
7.9.5	Functional data analysis	319
7.9.6	Error structure in explanatory variables	319
7.10	<i>Recap</i>	319
7.11	<i>Further reading</i>	319
7.12	<i>Exercises</i>	321
8	Tree-based Classification and Regression	322
8.1	<i>The uses of tree-based methods</i>	323
8.1.1	Problems for which tree-based regression may be used	323
8.2	<i>Detecting email spam – an example</i>	324
8.2.1	Choosing the number of splits	326
8.3	<i>Terminology and methodology</i>	327
8.3.1	Choosing the split – regression trees	327
8.3.2	Within and between sums of squares	328
8.3.3	Choosing the split – classification trees	329
8.3.4	Tree-based regression versus loess regression smoothing	329
8.4	<i>Predictive accuracy, and the cost-complexity tradeoff</i>	331
8.4.1	Cross-validation	331
8.4.2	The cost-complexity parameter	332
8.4.3	Prediction error versus tree size	332
8.5	<i>Data for female heart attack patients</i>	333
8.5.1	The one-standard-deviation rule	335
8.5.2	Printed Information on Each Split	335
8.6	<i>Detecting email spam – the optimal tree</i>	336
8.7	<i>The randomForest package</i>	338

8.7.1	Prior probabilities	339
8.7.2	A low-dimensional representation of observations	340
8.7.3	Models with a complex error structure	341
8.8	<i>Additional notes on tree-based methods</i>	341
8.8.1	Differences between <code>rpart()</code> and <code>randomForest()</code>	341
8.8.2	Tree-based methods, versus other approaches	343
8.8.3	Further notes	343
8.9	<i>Further reading and extensions</i>	344
8.10	<i>Exercises</i>	344
9	Multivariate data exploration and discrimination	347
9.1	<i>Multivariate exploratory data analysis</i>	348
9.1.1	Scatterplot matrices	348
9.1.2	Principal components analysis	348
9.1.3	Multi-dimensional scaling	353
9.2	<i>Discriminant analysis</i>	354
9.2.1	Example – plant architecture	355
9.2.2	Logistic regression	356
9.2.3	Linear discriminant analysis	357
9.2.4	An example with more than two groups	358
9.3	<i>*High-dimensional data — RNA-Seq gene expression</i>	360
9.3.1	Data and design matrix setup	361
9.3.2	From <i>p</i> -values to false discovery rate (FDR)	363
9.4	<i>High dimensional data from expression arrays</i>	364
9.4.1	Classifications and associated graphs	365
9.4.2	The mean-variance relationship	368
9.4.3	Graphs derived from the cross-validation process	372
9.4.4	Estimating contrasts, and calculating False Discovery Rates	373
9.5	<i>Further reading</i>	374
9.6	<i>Exercises</i>	375
10	Regression on Discriminant or Principal Component Scores	378
10.1*	<i>Propensity scores in regression comparisons – labor training data</i>	379
10.1.1	Regression comparisons	381
10.1.2	A strategy that uses propensity scores	384
10.2	<i>Principal component scores in regression</i>	390
10.3	<i>Further reading</i>	393
10.4	<i>Exercises</i>	393
	<i>Epilogue</i>	395
A	The R system – additional notes	397
A.1	<i>Data input from the RStudio menu, and data structures</i>	397
A.2	<i>Missing values NaNs, and Infinities</i>	399
A.3	<i>Factors</i>	400

<i>A.4 Dynamic graphics – the rgl and rggobi packages</i>	403
<i>A.5 Plotting characters, symbols, line types and colors</i>	404