Contents

1	**Approaching Data Analysis**		**1**
		Introduction	1
	1A	The staircase and the shortcut to inference	2
	1B	Student's true contribution	4
	1C	Distributions and their troubles	7
	1D	A classical example: Wilson and Hilferty's analysis of Peirce's data	10
	1E	Kinds of nonnormality and robustness	12
	1F	The role of vague concepts	17
	1G	Other vague concepts	19
	1H	Indication, determination, or inference	21
		Summary: Data analysis	22
2	**Indication and Indicators**		**25**
	2A	The value of indication	25
	2B	Examples of stopping with indication	27
	2C	Concealed inference	31
	2D	Choice of indicators	32
	2E	An example of choice of indicator	34
	2F	Indications of quality: Cross-validation	36
		Summary: Indication and indicators	40
3	**Displays and Summaries for Batches**		**43**
	3A	Stem-and-leaf	43
	3B	Medians, hinges, etc.	43
	3C	Mids and spreads	48
	3D	Subsampling	49
	3E	Exploratory plotting	49
	3F	Trends and running medians	52
	3G	Smoothing nonlinear regressions	61
	3H	Looking for patterns	64
	3I	Residuals more generally	70
	3J	Plotting and smoothing	75
		Summary: Batch summaries and displays	77

4 Straightening Curves and Plots 79

The idea of straightening out curves 79
4A The ladder of re-expressions 79
4B Re-expressing $y = x^2$ 81
4C The bulging rule 84
4D More complicated curves 87
4E Scatter plots 87
Summary: Straightening curves 87

5 The Practice of Re-Expression 89

5A Kinds of numbers 89
5B Quick logs 92
5C Quick (square) roots and reciprocals 97
5D Quick re-expressions of counted fractions, percentages, etc. 100
5E Matching for powers and logs 102
5F Re-expressions for grades 103
5G Re-expressing ranks 108
5H First aid in re-expression 109
5I What to do with zeros—and infinities 112
Summary: Re-expression 114

6 Need We Re-Express? 117

General hints when re-expressed carrier is $\log x$ 118

7 Hunting out the Real Uncertainty 119

7A How $\sigma/\sqrt{n}$ can mislead 121
7B A further example of the need for direct assessment of variability 122
7C Choosing an error term 123
7D More detailed choices of error terms 125
7E Making direct assessment possible 126
7F Difficulties with direct assessment 129
7G Supplementary uncertainty and its combination with internal uncertainty 129
Summary: Hunting out the real uncertainty 131

8 A Method of Direct Assessment 133

8A The jackknife 133
Appendix to 8A 137
8B Examples with individuals 139
8C Jackknife using groups: Ratio estimation for a sample survey 145
8D A more complex example 148
8E Cross-validation in the example 154
8F Two simultaneous uses of "leave out one" 156

8G Dispersion of the μ's 160
8H Further discussion of the example 161
Summary: The jackknife 162

9 Two- and More-Way Tables 165

9A PLUS analyses 165
9B Looking at two-way PLUS analyses 169
9C Taking advantage of levels 172
9D Polishing additive fits 178
9E Fitting one more constant 190
9F Using re-expression 194
9G Three- and more-way analyses 200
Summary: Two-way tables of responses 201

10 Robust and Resistant Measures 203

10A Resistance 203
10B Robustness 205
10C Robust and resistant estimates of location 206
10D Robust estimates of scale 207
10E Robust and resistant intervals 208
10F Resistant and robust regression 209
10G Multiple-component data 209
10H Closing comment 218
Summary: Resistant and robust techniques 218

11 Standardizing for Comparison 221

11A The simplest case 221
11B Direct standardization 225
11C Precision of directly standardized values 225
11D Difficulties with direct standardization 230
11E Indirect standardizing 233
11F Adjustment for broad categories 240
11G More than two broad categories 249
Summary: Standardizing for comparison 256

12 Regression for Fitting 259

Introduction 259
12A The two meanings of regression 262
12B Purposes of regression 268
12C Graphical fitting by stages 271
12D Collinearity 280
12E Linear dependence, exact and approximate 284
12F Keeping out what is imprecisely measured—Regression as exclusion 287

*12G Which straight line? (Optional) 293
12H Using subsamples 295
Summary: Regression 295

13 Woes of Regression Coefficients 299

13A Meaning of coefficients in multiple regression 299
13B Linear adjustment as a mode of description 303
13C Examples of linear adjustment 305
13D The relative unimportance of the exact carrier 315
13E Proxy phenomena 316
13F Sometimes x's can be "held constant" 318
13G Experiments, closed systems, and physical versus social sciences, with examples 320
*13H Estimated variances are not enough 328
Summary: Woes of regression coefficients 331

14 A Class of Mechanisms for Fitting 333

14A Fitting lines—some through the origin 335
14B Matching as a way of fitting 337
14C Matchers tuned to a single coefficient—and catchers 339
14D Ordinary least squares 341
14E Tuning for ordinary least squares 342
14F Weighted least squares 346
14G Influence curves for location 351
14H Iteratively weighted linear least squares 356
14I Least absolute deviations (Optional) 365
14J Analyzing troubles 369
14K Proof of the statement of Section 13B 374
Summary: Mechanisms for fitting regression 376

15 Guided Regression 381

15A How can we be guided in what to fit? 381
15B Stepwise techniques 387
15C All-subset techniques 391
15D Combined techniques 393
15E Rearranging carriers—judgment components 394
15F Principal components 397
15G How much we are likely to learn? 401
15H Several y's or several studies 402
15I Regression starting where? 402
15J Arbitrary adjustment 404
Summary: Guided regression 404

16 Examining Regression Residuals 407

16A Examining $\hat{y}$ 407
16B Variables—and other carriers 419

16C The next step: Looking with regard to an old variable, t_{old} 423
16D Looking with regard to a new variable, t_{new} 431
16E Looking for additional product terms 435
16F In what order? 446
Summary: Examining regression residuals 448

Appendix: Details About the Need to Re-Express **449**

Problems **467**

Data Exhibits for Problems **552**

Index **579**