

# *Visualizing Multivariate Data and Models in R*

*A Romance in Many Dimensions*

Here is where the dedication goes ...

---

# Table of contents

---

<b>Preface</b>	<b>v</b>
ONE, TWO, MANY . . . . .	v
Flatland . . . . .	vi
EUREKA! . . . . .	viii
Multivariate scientific discoveries . . . . .	viii
What I assume . . . . .	xi
Conventions used in this book . . . . .	xi
 <b>I Orienting Ideas</b>	 <b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Multivariate vs. multivariable methods . . . . .	3
1.2 Why use a multivariate design . . . . .	3
1.3 Linear models: Univariate to multivariate . . . . .	4
1.4 Visualization is harder . . . . .	4
1.5 Problems in understanding and communicating MLM results . . . . .	5
 <b>2 Getting Started</b>	 <b>7</b>
2.1 Why plot your data? . . . . .	7
2.1.1 Anscombe's Quartet . . . . .	7
2.1.2 One lousy point can ruin your day . . . . .	10
2.1.3 Shaken, not stirred: The 1970 Draft Lottery . . . . .	13
2.2 Plots for data analysis . . . . .	20
2.2.1 Model plots . . . . .	20
2.2.2 Diagnostic plots . . . . .	21
2.2.3 Principles of graphical display . . . . .	21
2.3 What have we learned? . . . . .	21
 <b>II Exploratory Methods</b>	 <b>23</b>
<b>3 Plots of Multivariate Data</b>	<b>25</b>
3.1 Bivariate summaries . . . . .	26
3.1.1 Smoothers . . . . .	27
3.1.2 Stratifiers . . . . .	30
3.1.3 Conditioning . . . . .	33
3.2 Data Ellipses . . . . .	34
3.2.1 Ellipse properties . . . . .	38
3.2.2 R functions for data ellipses . . . . .	41
3.2.3 Example: Penguins data . . . . .	47
3.2.4 Visual thinning . . . . .	50
3.3 Bagplots . . . . .	51
3.4 Non-parametric bivariate density plots . . . . .	53
3.5 Simpson's paradox: marginal and conditional relationships . . . . .	54
3.6 (a) Ignoring species . . . . .	55
3.7 (b) By species . . . . .	56

3.8	(c) Within species . . . . .	56
3.9	Multivariate normality and outliers . . . . .	57
3.9.1	Galton data . . . . .	57
3.9.2	Penguin data . . . . .	60
3.10	Scatterplot matrices . . . . .	63
3.10.1	Visual thinning . . . . .	67
3.11	Corrgrams . . . . .	69
3.12	Generalized pairs plots . . . . .	74
3.13	Parallel coordinate plots . . . . .	79
3.14	Animated tours . . . . .	84
3.14.1	Projections . . . . .	84
3.14.2	Touring methods . . . . .	89
3.15	Network diagrams . . . . .	94
3.15.1	Crime data . . . . .	96
3.15.2	Partial correlations . . . . .	97
3.15.3	Visualizing partial correlations . . . . .	98
3.16	What have we learned? . . . . .	99
3.17	Exercises . . . . .	100
<b>4</b>	<b>Dimension Reduction</b>	<b>103</b>
4.1	<i>Flatland</i> and <i>Spaceland</i> . . . . .	103
4.1.1	Multivariate juicers . . . . .	103
4.2	Principal components analysis (PCA) . . . . .	105
4.2.1	PCA by springs . . . . .	106
4.2.2	Mathematics and geometry of PCA . . . . .	108
4.2.3	Finding principal components . . . . .	113
4.2.4	Visualizing variance proportions: screeplots . . . . .	115
4.2.5	Visualizing PCA scores and variable vectors . . . . .	116
4.3	Biplots . . . . .	121
4.3.1	Constructing a biplot . . . . .	122
4.3.2	Biplots in R . . . . .	123
4.3.3	Example: Crime data . . . . .	123
4.3.4	Biplot contributions and quality . . . . .	126
4.3.5	Supplementary variables . . . . .	128
4.3.6	Example: Diabetes data . . . . .	132
4.4	Nonlinear dimension reduction . . . . .	135
4.4.1	Multidimensional scaling . . . . .	135
4.4.2	t-SNE . . . . .	138
4.5	Application: Variable ordering for data displays . . . . .	142
4.6	Application: Eigenfaces . . . . .	145
4.7	Elliptical insights: Outlier detection . . . . .	151
4.7.1	Example: Penguin data . . . . .	151
4.8	What have we learned? . . . . .	153
<b>III</b>	<b>Univariate Linear Models</b>	<b>159</b>
<b>5</b>	<b>Overview of Linear models</b>	<b>161</b>
5.1	The General Linear Model . . . . .	163
5.1.1	Model formulas . . . . .	164
5.1.2	Model matrices . . . . .	168
5.1.3	Coding factors and contrasts . . . . .	169
5.2	What have we learned? . . . . .	176
<b>6</b>	<b>Plots for univariate response models</b>	<b>177</b>

6.1	The “regression quartet”	178
6.1.1	Example: Duncan’s occupational prestige	178
6.1.2	Diagnostic plots	181
6.1.3	Example: Canadian occupational prestige	182
6.2	Other Model plots	184
6.3	Coefficient displays	185
6.3.1	Displaying coefficients	186
6.3.2	Visualizing coefficients	187
6.3.3	More useful coefficient plots	189
6.4	Added-variable and related plots	192
6.4.1	Properties of AV plots	195
6.4.2	Marginal - conditional plots	196
6.4.3	Prestige data	198
6.4.4	Component + Residual plots	198
6.5	Effect displays	202
6.5.1	Prestige data	204
6.6	Outliers, leverage and influence	208
6.6.1	The leverage-influence quartet	209
6.6.2	Influence plots	216
6.6.3	Duncan data	216
6.6.4	Influence in added-variable plots	217
6.7	What have we learned?	220
<b>7</b>	<b>Topics in Linear Models</b>	<b>221</b>
7.1	Ellipsoids in data space and $\beta$ space	221
7.1.1	Coffee, stress and heart disease	223
7.2	Measurement error	227
7.2.1	OLS is BLUE	227
7.2.2	Errors in predictors	227
7.2.3	Coffee data: $\beta$ space	231
7.3	What have we learned?	232
<b>8</b>	<b>Collinearity &amp; Ridge Regression</b>	<b>235</b>
8.1	What is collinearity?	235
8.1.1	Visualizing collinearity	237
8.1.2	Data space and $\beta$ space	238
8.2	Measuring collinearity	241
8.2.1	Variance inflation factors	241
8.2.2	Collinearity diagnostics	243
8.3	Tableplots	245
8.4	Collinearity biplots	245
8.5	Remedies for collinearity: What can I do?	248
8.6	Ridge regression	252
8.6.1	Properties of ridge regression	252
8.6.2	The <code>genridge</code> package	255
8.7	Univariate ridge trace plots	255
8.8	Bivariate ridge trace plots	258
8.8.1	Visualizing the bias-variance tradeoff	259
8.9	Low-rank views	262
8.9.1	Biplot view	266
8.10	What have we learned?	267
<b>IV</b>	<b>Multivariate Linear Models</b>	<b>269</b>

<b>9 Hotelling's <math>T^2</math></b>	<b>271</b>
9.1 $T^2$ as a generalized $t$ -test	271
9.2 $T^2$ properties	272
Example: Mathscore data	273
9.3 HE plot and discriminant axis	277
9.3.1 <code>heplot()</code>	277
9.4 Discriminant analysis	280
9.5 More variables	281
9.5.1 Biplots	284
9.5.2 Testing mean differences	284
9.6 Variance accounted for: Eta square ( $\eta^2$ )	286
9.7 What we've learned	287
9.8 Exercises	287
<b>10 Multivariate Linear Models</b>	<b>289</b>
10.1 Structure of the MLM	290
10.1.1 Assumptions	291
10.2 Fitting the model	292
10.2.1 Example: Dog food data	292
10.2.2 Sums of squares	294
10.2.3 How big is $SS_H$ compared to $SS_E$ ?	296
10.3 Multivariate test statistics	298
10.3.1 Testing contrasts and linear hypotheses	298
10.4 ANOVA $\rightarrow$ MANOVA	301
10.4.1 Example: Father parenting data	303
10.4.2 Ordered factors	309
10.4.3 Example: Adolescent mental health	309
10.4.4 Factorial MANOVA	314
10.5 MRA $\rightarrow$ MMRA	314
10.5.1 Example: NLSY data	315
10.5.2 Example: School data	321
10.6 Model diagnostics for MLMs	325
10.6.1 Multivariate normality of residuals	325
10.6.2 Distance plot	327
10.6.3 Multivariate influence	328
10.7 ANCOVA $\rightarrow$ MANCOVA	330
10.7.1 Example: Paired-associate tasks and academic performance	331
10.8 What we've learned	337
<b>11 Visualizing Multivariate Models</b>	<b>339</b>
11.1 HE plot framework	340
11.2 HE plot construction	341
11.2.1 MANOVA model	345
11.3 HE plots	347
11.4 Significance scaling	348
11.5 Visualizing contrasts and linear hypotheses	349
11.6 HE plot matrices	351
11.7 Low-D views: Canonical analysis	351
11.7.1 Coefficients	353
11.7.2 Canonical scores plot	354
11.7.3 Canonical HE plot	354
11.8 Factorial MANOVA	356
11.9 Quantitative predictors: MMRA	361
11.10 Canonical correlation analysis	364