Updates and Errata for *Statistical Data Analytics*

(1st edition, 2015)

Walter W. Piegorsch
University of Arizona

CONTENTS

PREFACE

This listing provides updates and known errata from the first printing (2015) of *Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery* (John Wiley & Sons, Ltd). See the Web Site at

`http://www.math.arizona.edu/~piegorsch/wiley/sda/sda.html`

to find the most up-to-date list. Thanks are due the many readers who have brought these updates to my attention.

Walter W. Piegorsch
Tucson, Arizona

1

Data Analytics and Data Mining

Updates and Errata

- (p. 4, Figure 1) The debate over the DIKW pyramid and how well it represents the Knowledge/Wisdom Hierarchy and the processes of information systems is nontrivial. For example, as implied in the main text Rowley (2007) gives a perspective supporting the pyramid as a representational tool and also provides useful historical context. By contrast, Fricke (2009) questions the stability of the hierarchy and its underlying methodology. Readers are encouraged to peruse these articles, and more recent works that cite them, to gain a broader understanding of what the DIKW pyramid can mean in theory and in practice.

2

Basic Probability and Statistical Distributions

Updates and Errata

- (p. 24, before Example 2.3.2) The last sentence before Example 2.3.2 should read as follows:
  "With large m use the `lfactorial(m)` function, giving $\log(m!)$ so that $m! = \exp\{\log(m!)\}$."

3

Data Manipulation

Updates and Errata

- (p. 52, middle) "...20K is truly twice as 'hot' as 10K, in contrast to Fahrenheit or Celsius".
- (p. 56, top) "Operations such as trimming can improve the stability of an unstable or substandard data set, and will in some instances improve the eventual information they provide. If performed indiscriminately, however, post-sampling data manipulations can just as often – and, perhaps, more often – detrimentally affect the larger analytic enterprise".
- (p. 61, Example 3.4.1) The subjects were participating in a "weight reduction/maintenance program".

4

Data Visualization and Statistical Graphics

Updates and Errata

- (p. 76, initial paragraph) Gelman and Unwin **take** the quote a step further.
- (p. 88, first paragraph) The last sentence of the first paragraph should read as follows: Verzani (2005, Fig. 3.4) gives an instructive graphic.
- (p. 89, Sec. 4.2) The two random samples are:
  $X_i \sim$ i.i.d. $f_X(x)$, $i = 1, \ldots, n$, independent of $Y_j \sim$ i.i.d. $f_Y(y)$, $j = 1, \ldots, m$, where $f_X(x)$ and $f_Y(y)$ are the two samples' p.d.f.s.
- (p. 95, Sec. 4.2.2) Prior to Example 4.2.3:
  "...this graphic device [vertical boxplots] is best reserved for K in, say, the range 2 $\leq K \leq$ 60."

5

Statistical Inference

Updates and Errata

- (p. 129, middle) The sentence beginning with "From (5.24), ..." should read as follows:
  From (5.24), the limits are then $-0.1538 \pm (1.6561)\sqrt{2.1734} = -0.1538 \pm 2.4415$.

- (p. 136, near top) For the Delta method confidence interval's standard error: appeal to (5.7) and find

$$\text{se}[h(\hat{\theta})] = \sqrt{\text{Var}[h(\hat{\theta})]\Big|_{\theta=\hat{\theta}}} \approx \sqrt{\{h'(\hat{\theta})\}^2 \text{ Var}[\hat{\theta}]\Big|_{\theta=\hat{\theta}}} = |h'(\hat{\theta})| \,\text{se}[\hat{\theta}],$$

- (p. 138, middle) Item (3) under the bootstrap generating algorithm should read as follows:
  Repeat steps (1) and (2) a large number, B, of times. Babu and Singh (1983) give theoretical arguments for $B = n(\log n)^2$, although if $n < 60$ a common recommendation is to set at least $B = 2000$ for building confidence intervals.

- (p. 150, Example 5.5.1) The material after Equation (5.51) describing the infarction data's summary statistics should read as follows:
  There, the sample mean was $\overline{X} = 62.8137$ (yrs.) and the sample variance was $S^2 = 69.5908$.

6

Techniques for Supervised Learning: Simple Linear Regression

Updates and Errata

- (p. 183, bottom) Use as the influence criterion $P[F(2, 340) \leq D_i] > \frac{1}{2}$.

7

Techniques for Supervised Learning: Multiple Linear Regression

Updates and Errata

- (p. 201, top) To be clear, the mean square error term in the estimated covariance matrix will involve the weights as well: MSE = $\sum_{i=1}^{n} w_i e_i^2/(n-p-1)$, where the e_i values are the residuals from the WLS fit.
- (p. 202, middle) The expression for the joint, minimal, $100(1-\alpha)\%$ hyper-rectangle should read:

$$\hat{\mu}(\mathbf{x}_h) \pm t_{\alpha/(2H)}(n-p-1)\,\mathrm{se}[\hat{\mu}(\mathbf{x}_h)]$$

- (p. 204, middle) Predictors x_3 and x_4 represent lifestyle factors.
- (p. 212, Example 7.3.1) There are up to $p+1=15$ regression parameters to be estimated from the data.
- (p. 213, top) The **R** function `I()` should be referred to as the 'Inhibit Interpretation' function.
- (p. 218, Item FS.1) The t-statistic should read:

$$T_j = \hat{\beta}_j/\mathrm{se}[\hat{\beta}_j].$$

- (p. 224, top) The modern term for *loess* is named after the "silt-like deposits" formed along river banks.
- (p. 247, middle) Regarding use of Equation (7.43): reject $\mathrm{H}_{ii'j}$ when $P_{ii'j}$ in (7.43) *drops below* $\alpha = 0.10$.

8

Supervised Learning: Generalized Linear Models

Updates and Errata

- (p. 259) The last sentence immediately preceding Sec. 8.1.2 should read as follows:
  As seen there, because the uniform distribution's support space depends upon unknown parameter(s) it cannot be included in the exponential family.

- (p. 260) The log-likelihood equation (8.3) should read as follows:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i(\boldsymbol{\beta}) - b\big(\theta_i(\boldsymbol{\beta})\big)}{a(\varphi)} + c(Y_i, \varphi) \right\}, \tag{8.1}$$

- (p. 260) The estimating equations displayed in the middle of the page should read as follows:

$$\sum_{i=1}^{n} \frac{x_{ij} h'(\eta_i)}{a(\varphi) V(\mu_i)} (Y_i - \mu_i) = 0 \quad \text{for } j = 0, \ldots, p$$

- (pp. 270–271) The last sentence on p. 270 that continues on p. 271 should read as follows:
  Like the Lasso, regularization again produces a form of regression with a built-in variable selector.

- (p. 280, Table 8.6) The caption for Table 8.6 should read as follows:
  A 2×2 contingency table for association between *Alternaria alternata* allergic response and asthma status in six-year-old children.

- (p. 285, Exercise 8.6a) This portion of Exercise 8.6 should read as follows:
  Under the logit link, verify that adding quadratic terms in x_1 and x_2 to the linear predictor will provide no significant improvement in the model fit. Operate at a false positive rate of 5%.

9

Supervised Learning: Classification

Updates and Errata

- (p. 298, Figure 9.2) Notice that *pROC* plots the specificity along the horizontal axis from 1.0 to 0.0.
- (p. 334) Exercise 9.3(d). From the fitted logistic model, estimate the probability that a new patient will be classified with a positive tumor outcome if he presents the following predictor values: $x_1 = 100$, $x_2 = 300$, $x_3 = 50$. Also include a 95% confidence interval for this value.

10

Techniques for Unsupervised Learning: Dimension Reduction

Updates and Errata

- (p. 357, Equation (10.7)) As regards Equation (10.7), n is the size of the sample.

11

Techniques for Unsupervised Learning: Clustering and Association

Updates and Errata

- (p. 375, Table 11.1) The Displacements (in cu. in.) for the final two automobiles should read as follows:

Make & Model	Displacement (cu. in.)
Buick Skylark	350
Ford Torino	302

  All calculations and plots using these data employ the correct values listed above.

- (p. 376, top) The last sentence of the top paragraph should read as follows:
  For example, one might compute the sample correlation between $\mathbf{z}_i$ and $\mathbf{z}_h$, averaging across the p variables; see Hastie et al. (2009, §14.3.2).

- (p. 380, bottom) In the middle of the first paragraph of Example 11.1.3, the text should read as follows:
  Clustering often focuses on the n different genes that represent the observations, where the genes' expression patterns are examined across $p > 1$ conditions such as different disease states, organ/tissue types, ecological species, etc. Alternatively, n individual subjects may be clustered using expression data from p different genes felt to represent presumptive genetic markers. For the latter case, Uhlmann et al. (2012) describe a study where $n = 118$ colorectal cancer patients had their relative gene expression ratios sampled for $p = 4$ different genes that may possibly contribute to cancer progression: osteopontin (Opn), cyclooxigenase-2 (Cox-2), transforming growth factor β (Tgf-β), and matrix metalloproteinase-2 (Mmp-2).

- (p. 389, middle) The penultimate sentence of Example 11.1.4 should read as follows:
  They can be used to service nearest-neighbor queries when, e.g., planning mass-transit circuits, separating water supplies from hazardous waste sites, locating mobile telephone towers, etc.

A

Appendix: Matrix Manipulation

Updates and Errata

- (p. 413, top) The standard definition [of Cayley multiplication] multiplies elementwise across a row and then down a column.
- (p. 418, §A.6.4) The first sentence in §A.6.4 should read as follows:
  A more general decomposition, of which the spectral decomposition in Section A.6.2 can be viewed as a special case, applies to any $n \times p$ matrix $\mathbf{X}$.
- (p. 420, end) In **R**, `cor(X)` calculates a correlation matrix when the input argument `X` is a matrix or data frame.

B

Appendix: Brief Introduction to R

Updates and Errata

- (p. 425) The first sentence of the paragraph preceding the sub-setting threshold illustration in the middle of the page should read as follows:
  R's object-oriented structure allows for creative use of the bracketed indexing feature to create subsetted objects; see `http://cran.r-project.org/doc/manuals/R-intro.html#Index-vectors`.

References

Babu GJ and Singh K 1983 Inference on means using the bootstrap. *Annals of Statistics* **11**(3), 999–1003.

Fricke M 2009 The knowledge pyramid: a critique of the dikw hierarchy. *Journal of Information Science* **35**(2), 131–142.

Gelman A and Unwin A 2013 Infovis and statistical graphics: Different goals, different looks (with discussion). *Journal of Computational and Graphical Statistics* **22**(1), 2–49.

Hastie T, Tibshirani R and Friedman J 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edn. Springer, New York.

Piegorsch WW 2015 *Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery*. John Wiley & Sons, Chichester.

Rowley J 2007 The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science* **33**(2), 163–180.

Uhlmann ME, Georgieva M, Sill M, Linnemann U and Berger MR 2012 Prognostic value of tumor progression-related gene expression in colorectal cancer patients. *Journal of Cancer Research and Clinical Oncology* **138**(10), 1631–1640.

Verzani J 2005 *Using R for Introductory Statistics*. Chapman & Hall/CRC, Boca Raton, FL.