Errata

statlearning.com/errata-second-edition

Second Edition

Since the 1st printing (Summer 2021)

In Table 1.1 on page 14, the Credit data set involves information about credit card debt for 400 customers, not for 10,000 customers. *Thanks to Matthew Greenberg*.

In the caption of Figure 2.2 on page 17, “Income” should be in units of $1,000, not units of $10,000. *Thanks to Jacky*.

On page 86, in Table 3.8, the model output for the “South” and “West” regions have been swapped. The 2nd row should be labeled Region[West] and the 3rd row should be labeled Region[South]. *Thanks to Mustafa Anjrini.*

On page 111, the variable "age" is actually "proportion of owner-occupied units built prior to 1940". *Thanks to Umberto Picchini and Denis Kazakov.*

On page 116, the sentence “The syntax lstat:black tells R to include an interaction term between lstat and black. “ should read “The syntax lstat:age tells R to include an interaction term between lstat and age.” *Thanks to Dwight Wynne.*

On page 143, the footnote should mention (4.15), not (4.13). *Thanks to Alejandro Estrada.*

On pages 156-158, some corrections are needed:

- Page 156, Table 4.8: this table results from including Income as a covariate. However, earlier results in this chapter did not include Income as a covariate. When Income is not included as a covariate, the 1st 2 rows of this table are as follows:
    - Predicted default status = No: 9621 244 9865
    - Predicted default status = Yes: 46 89 135

- Page 157, Table 4.9: this table results from including Income as a covariate. However, earlier results in this chapter did not include Income as a covariate. When Income is not included as a covariate, the 1st 2 rows of this table are as follows:
    - Predicted default status = No: 9339 130 9469
    - Predicted default status = Yes: 328 203 531
- Page 158: In the paragraph before Section 4.5, the text should read “this data set has $n=10,000$ and $p=2$”.

Thanks to Marco Bee.

On page 165, the 3rd column of Table 4.10 mentions a z-statistic. This should actually be a t-statistic. *Thanks to Wei Li.*

On page 226, replace the text “the variance is *infinite* so the method cannot be used at all.” with the text “there are infinitely many solutions. Each of these least squares solutions gives zero error on the training data, but typically very poor test set performance due to extremely high variance.\footnote{When $p \gg n$, the least squares solution that has the smallest sum of squared coefficients can sometimes perform quite well. See Section Section 10.8 for a more detailed discussion.} ”

On pages 227-236, the treatment of cross-validation as a way to select the best model among M*0,…,M*p in Step 3 of Algorithms 6.1-6.3 is inadequate. In particular, if cross-validation is used, then it is critical that Step 2 be carried out separately in each training fold of the cross-validation procedure. For instance, if we perform 5-fold cross-validation, then in each fold, we should construct M*1,…,M*p using only the 80% of the data that are in the training set for that fold. Otherwise, if Step 2 is carried out using all n observations, then we may severely underestimate the test error. In light of this issue, the following corrections are needed:

> Page 227, Algorithm 6.1, Step 3. Replace with the following text: “… using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.” The same correction should also be made on page 230, Algorithm 6.2, Step 3, and page 231, Algorithm 6.3, Step 3.
>
> Page 227. In the text, after "… described in Algorithm 6.1." add: "If cross-validation is used to select the best model, then Step 2 is repeated on each training fold, and the validation errors are averaged to select the best value of k. Then the model ${\cal M}_k$ fit on the full training set is delivered for the chosen k."

> Page 235. Just before last paragraph, after "... variance σ^2", the following text should be added: "Note that when cross-validation is used, the sequence of models ${\cal M}_k$ in Algorithms 6.1--6.3 is determined separately for each training fold, and the validation errors are averaged over all folds for each model size k. This means, for example with best-subset regression, that ${\cal M}_k$, the best subset of size k, can differ across the folds. Once the best size k is chosen, we find the best model of that size on the full data set."

Thanks to Marcel Scharth.

On page 320, the last line of R code should say “plot(gam.lo, se = TRUE, col = "green")” rather than “plot.Gam(gam.lo, se = TRUE, col = "green")”. *The authors.*

On page 401, “how many training errors are misclassified” should read “how many training observations are misclassified”. *Thanks to Denis Kazakov.*

On the third bullet on page 432, the text “This model has 1,409 parameters” should be replaced with “This model has 1,345 parameters”. *Thanks to Declan Clarke.*

In Table 10.2 on page 433, “# Parameters” for “Neural Network” should read 1345 instead of 1409. *Thanks to Oliver Angelil.*

On page 433, “its tempting to always go for” should read “it’s tempting to always go for”. *Thanks to Michael Baznik.*

On page 434, “we give a briefoverview here” should read “we give a brief overview here”. *Thanks to Michael Baznik.*

On page 437, after (10.31), “is also popular as an additional form or regularization” should be replaced with “is also popular as an additional form of regularization”. *Thanks to Michael Baznik.*

On page 439, second bullet: “These includes” should read “these include”. *Thanks to Michael Baznik.*

On page 443: immediately before Section 10.1, we plan to add the following text in the next printing: “Since the first printing of this book, the torch package has become available as an alternative to the keras package for deep learning. While torch does not require a python installation, the current implementation appears to be a fair bit slower than keras. A version of this lab that makes use of torch is available on the book website. \footnote{Many thanks to Daniel Falbel and Sigrid Keydana for preparing the torch version of this lab.}"

On pages 448 and 451: several code blocks require some small changes, because in keras 2.6.0, the function predict_classes() has been deprecated. See the Chapter 10 Jupyter notebook, Rmarkdown file, or .R file on the Resources page for the corrected lab. *Thanks to*

Zhu Wang.

On page 479, the denominator of the x-axis label for Figure 11.7 is missing a vertical bar "|". *Thanks to Zhe Huang*.

On page 508, immediately under Figure 12.4, the variable name "UrbanPop" is misspelled as "UrpanPop". *Thanks to Christian Schroder*.

On page 509, the sentence "Each principal component vector is unique, up to a sign flip." should be replaced with "While in theory the principal components need not be unique, in almost all practical settings they are (up to sign flips)." *Thanks to Michael Causey*.

On page 515, the sentence "Table 12.3 illustrates the setup" should say "Table 12.2 illustrates the setup". *Thanks to V. Anand*.

On page 536, the sentence "We now omit 20 entries in the 50 × 2 data matrix at random." should read "We now omit 20 entries in the 50 × 4 data matrix at random." *Thanks to Tammo Ricklefs*.

On the top of page 550, the text "can be obtaining by" should read "can be obtained by". *Thanks to Jesse Onland*.

On page 553, line 14: the text "mean blood pressure of mice in the control group" should say "the expected blood pressure of mice in the control group", to avoid confusion between sample mean and population mean. Similarly, "mean blood pressure of mice in the treatment group" should say "the expected blood pressure of mice in the treatment group". *The authors*.

On page 553, line -3: the text "the mean value of the jth biomarker among mice in the control group equals the mean value of the jth biomarker among mice in the treatment group" should say "the expected value of the jth biomarker among mice in the control group equals the expected value of the jth biomarker among mice in the treatment group". *The authors*.

On page 554, beginning of Section 13.1: the text "is the coefficient" should be changed to "is the true coefficient", to avoid confusion between the coefficient estimate and the population value of the coefficient. *The authors*.

On page 555, item 1 in the middle of the page: the text "the coefficient" should be changed to "the true coefficient", again to avoid confusion between the coefficient estimate and the population quantity. *The authors*.

On page 555, item 2 in the middle of the page: the text "mean blood pressure" should be changed to "expected blood pressure". *The authors*.

On page 555, second-to-last paragraph: the text “mean blood pressure” should be changed to “expected blood pressure”. *The authors*.

On pages 564-565, there is a small notational issue. On the left-hand side of (13.6), FWER(alpha) should be replaced with FWER. And on the left-hand side of the first equation on page 565, FWER(alpha) should be replaced with FWER(alpha/m). *The authors*.

On page 593, in both questions 6(c) and 6(d), “false discovery rate” should be replaced with “false discovery proportion”. *Thanks to Jesse Onland*.

Since the release of v1.2 of the ISLR2 package

The help page for the BrainCancer dataset should state that the variable time measures “Survival time, in months”, rather than age, in years. This will be corrected in the next version of the ISLR2 package. *The authors*.

The help page for the Credit dataset should state that the variable “Income” is in units of $1,000, not units of $10,000. This has been corrected in v1.3. *Thanks to Erica Chauvet*.

There was a problem with the Auto dataset in v1.1 of the ISLR2 package. The issue has been resolved in v1.2. *Thanks to Christian Schroder*.

The file install.R which is used to install the ‘keras’ and related packages in chapter 10 has been changed. Update to ISLR2 version 1.3 for the updated version. *Thanks to Shahrdad Shadab*.