

## 5.2 The Bootstrap

The **bootstrap** is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method. As a simple example, the bootstrap can be used to estimate the standard errors of the coefficients from a linear regression fit. In the specific case of linear regression, this is not particularly useful, since we saw in Chapter 3 that standard statistical software such as **R** outputs such standard errors automatically. However, the power of the bootstrap lies in the fact that it can be easily applied to a wide range of statistical learning methods, including some for which a measure of variability is otherwise difficult to obtain and is not automatically output by statistical software. bootstrap

In this section we illustrate the bootstrap on a toy example in which we wish to determine the best investment allocation under a simple model. In Section 5.3 we explore the use of the bootstrap to assess the variability associated with the regression coefficients in a linear model fit.

Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of  $X$  and  $Y$ , respectively, where  $X$  and  $Y$  are random quantities. We will invest a fraction  $\alpha$  of our money in  $X$ , and will invest the remaining  $1 - \alpha$  in  $Y$ . Since there is variability associated with the returns on these two assets, we wish to choose  $\alpha$  to minimize the total risk, or variance, of our investment. In other words, we want to minimize  $\text{Var}(\alpha X + (1 - \alpha)Y)$ . One can show that the value that minimizes the risk is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}, \quad (5.6)$$

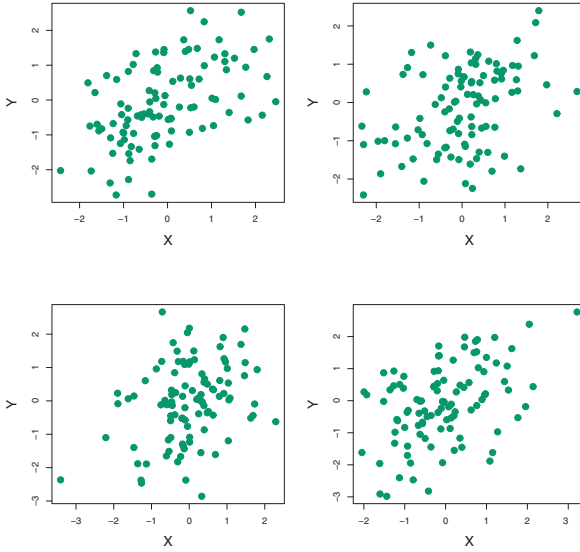
where  $\sigma_X^2 = \text{Var}(X)$ ,  $\sigma_Y^2 = \text{Var}(Y)$ , and  $\sigma_{XY} = \text{Cov}(X, Y)$ .

In reality, the quantities  $\sigma_X^2$ ,  $\sigma_Y^2$ , and  $\sigma_{XY}$  are unknown. We can compute estimates for these quantities,  $\hat{\sigma}_X^2$ ,  $\hat{\sigma}_Y^2$ , and  $\hat{\sigma}_{XY}$ , using a data set that contains past measurements for  $X$  and  $Y$ . We can then estimate the value of  $\alpha$  that minimizes the variance of our investment using

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}. \quad (5.7)$$

Figure 5.9 illustrates this approach for estimating  $\alpha$  on a simulated data set. In each panel, we simulated 100 pairs of returns for the investments  $X$  and  $Y$ . We used these returns to estimate  $\sigma_X^2$ ,  $\sigma_Y^2$ , and  $\sigma_{XY}$ , which we then substituted into (5.7) in order to obtain estimates for  $\alpha$ . The value of  $\hat{\alpha}$  resulting from each simulated data set ranges from 0.532 to 0.657.

It is natural to wish to quantify the accuracy of our estimate of  $\alpha$ . To estimate the standard deviation of  $\hat{\alpha}$ , we repeated the process of simulating 100 paired observations of  $X$  and  $Y$ , and estimating  $\alpha$  using (5.7),



**FIGURE 5.9.** Each panel displays 100 simulated returns for investments  $X$  and  $Y$ . From left to right and top to bottom, the resulting estimates for  $\alpha$  are 0.576, 0.532, 0.657, and 0.651.

1,000 times. We thereby obtained 1,000 estimates for  $\alpha$ , which we can call  $\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{1,000}$ . The left-hand panel of Figure 5.10 displays a histogram of the resulting estimates. For these simulations the parameters were set to  $\sigma_X^2 = 1, \sigma_Y^2 = 1.25$ , and  $\sigma_{XY} = 0.5$ , and so we know that the true value of  $\alpha$  is 0.6. We indicated this value using a solid vertical line on the histogram. The mean over all 1,000 estimates for  $\alpha$  is

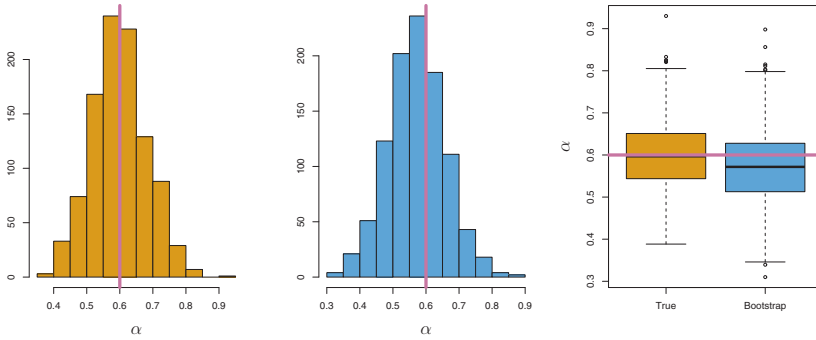
$$\bar{\alpha} = \frac{1}{1,000} \sum_{r=1}^{1,000} \hat{\alpha}_r = 0.5996,$$

very close to  $\alpha = 0.6$ , and the standard deviation of the estimates is

$$\sqrt{\frac{1}{1,000 - 1} \sum_{r=1}^{1,000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083.$$

This gives us a very good idea of the accuracy of  $\hat{\alpha}$ :  $\text{SE}(\hat{\alpha}) \approx 0.083$ . So roughly speaking, for a random sample from the population, we would expect  $\hat{\alpha}$  to differ from  $\alpha$  by approximately 0.08, on average.

In practice, however, the procedure for estimating  $\text{SE}(\hat{\alpha})$  outlined above cannot be applied, because for real data we cannot generate new samples from the original population. However, the bootstrap approach allows us to use a computer to emulate the process of obtaining new sample sets,



**FIGURE 5.10.** Left: A histogram of the estimates of  $\alpha$  obtained by generating 1,000 simulated data sets from the true population. Center: A histogram of the estimates of  $\alpha$  obtained from 1,000 bootstrap samples from a single data set. Right: The estimates of  $\alpha$  displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of  $\alpha$ .

so that we can estimate the variability of  $\hat{\alpha}$  without generating additional samples. Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set.

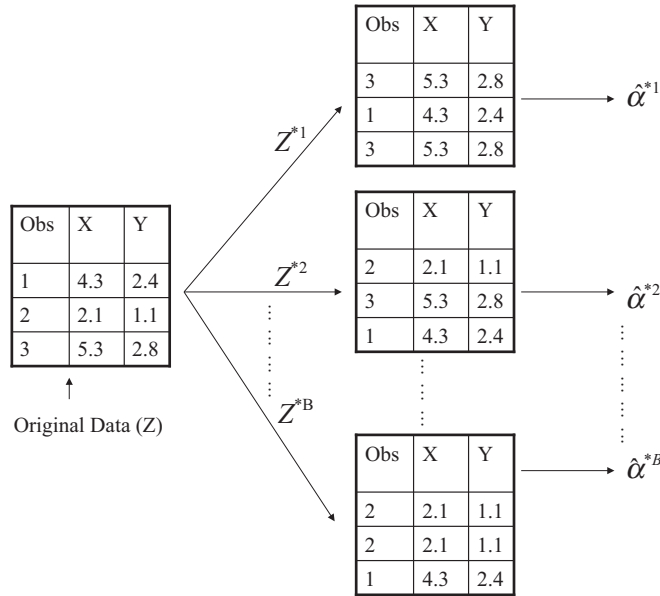
This approach is illustrated in Figure 5.11 on a simple data set, which we call  $Z$ , that contains only  $n = 3$  observations. We randomly select  $n$  observations from the data set in order to produce a bootstrap data set,  $Z^{*1}$ . The sampling is performed with replacement, which means that the same observation can occur more than once in the bootstrap data set. In this example,  $Z^{*1}$  contains the third observation twice, the first observation once, and no instances of the second observation. Note that if an observation is contained in  $Z^{*1}$ , then both its  $X$  and  $Y$  values are included. We can use  $Z^{*1}$  to produce a new bootstrap estimate for  $\alpha$ , which we call  $\hat{\alpha}^{*1}$ . This procedure is repeated  $B$  times for some large value of  $B$ , in order to produce  $B$  different bootstrap data sets,  $Z^{*1}, Z^{*2}, \dots, Z^{*B}$ , and  $B$  corresponding  $\alpha$  estimates,  $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \dots, \hat{\alpha}^{*B}$ . We can compute the standard error of these bootstrap estimates using the formula

replacement

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1} \sum_{r=1}^B \left( \hat{\alpha}^{*r} - \frac{1}{B} \sum_{r'=1}^B \hat{\alpha}^{*r'} \right)^2}. \quad (5.8)$$

This serves as an estimate of the standard error of  $\hat{\alpha}$  estimated from the original data set.

The bootstrap approach is illustrated in the center panel of Figure 5.10, which displays a histogram of 1,000 bootstrap estimates of  $\alpha$ , each computed using a distinct bootstrap data set. This panel was constructed on the basis of a single data set, and hence could be created using real data.



**FIGURE 5.11.** A graphical illustration of the bootstrap approach on a small sample containing  $n = 3$  observations. Each bootstrap data set contains  $n$  observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of  $\alpha$ .

Note that the histogram looks very similar to the left-hand panel which displays the idealized histogram of the estimates of  $\alpha$  obtained by generating 1,000 simulated data sets from the true population. In particular the bootstrap estimate  $SE(\hat{\alpha})$  from (5.8) is 0.087, very close to the estimate of 0.083 obtained using 1,000 simulated data sets. The right-hand panel displays the information in the center and left panels in a different way, via boxplots of the estimates for  $\alpha$  obtained by generating 1,000 simulated data sets from the true population and using the bootstrap approach. Again, the boxplots are quite similar to each other, indicating that the bootstrap approach can be used to effectively estimate the variability associated with  $\hat{\alpha}$ .

### 5.3 Lab: Cross-Validation and the Bootstrap

In this lab, we explore the resampling techniques covered in this chapter. Some of the commands in this lab may take a while to run on your computer.