

- (d) When $n = 5$, what is the probability that the j th observation is in the bootstrap sample?
- (e) When $n = 100$, what is the probability that the j th observation is in the bootstrap sample?
- (f) When $n = 10,000$, what is the probability that the j th observation is in the bootstrap sample?
- (g) Create a plot that displays, for each integer value of n from 1 to 100,000, the probability that the j th observation is in the bootstrap sample. Comment on what you observe.
- (h) We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the j th observation. Here $j = 4$. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

```
> store <- rep(NA, 10000)
> for(i in 1:10000){
  store[i] <- sum(sample(1:100, rep=TRUE) == 4) > 0
}
> mean(store)
```

Comment on the results obtained.

3. We now review k -fold cross-validation.
 - (a) Explain how k -fold cross-validation is implemented.
 - (b) What are the advantages and disadvantages of k -fold cross-validation relative to:
 - i. The validation set approach?
 - ii. LOOCV?
4. Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X . Carefully describe how we might estimate the standard deviation of our prediction.

Applied

5. In Chapter 4, we used logistic regression to predict the probability of `default` using `income` and `balance` on the `Default` data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.
 - (a) Fit a logistic regression model that uses `income` and `balance` to predict `default`.

- (b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:
 - i. Split the sample set into a training set and a validation set.
 - ii. Fit a multiple logistic regression model using only the training observations.
 - iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the `default` category if the posterior probability is greater than 0.5.
 - iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.
 - (c) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.
 - (d) Now consider a logistic regression model that predicts the probability of `default` using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for `student` leads to a reduction in the test error rate.
6. We continue to consider the use of a logistic regression model to predict the probability of `default` using `income` and `balance` on the `Default` data set. In particular, we will now compute estimates for the standard errors of the `income` and `balance` logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the `glm()` function. Do not forget to set a random seed before beginning your analysis.
- (a) Using the `summary()` and `glm()` functions, determine the estimated standard errors for the coefficients associated with `income` and `balance` in a multiple logistic regression model that uses both predictors.
 - (b) Write a function, `boot.fn()`, that takes as input the `Default` data set as well as an index of the observations, and that outputs the coefficient estimates for `income` and `balance` in the multiple logistic regression model.
 - (c) Use the `boot()` function together with your `boot.fn()` function to estimate the standard errors of the logistic regression coefficients for `income` and `balance`.
 - (d) Comment on the estimated standard errors obtained using the `glm()` function and using your bootstrap function.