**Chapter 5: Resampling methods**

* Resampling methods involve repeatedly drawing samples from a training set and refitting a model of interested on each sample in order to obtain additional information about the fitted model. For example, in order to estaimte the variability of a linear regression fit, we can repeatedly draw different samples from the trainig data, fit a linear regression to each new sample, and then examine the extent to which the results fits differ.
* The process of evaluating a model’s performance is known as *model assessment*, whereas the process of selecting the proper level of flexibility for a model is known as *model selection*.
* In this chapter, we instead consider a class of methods that estimate the test error rate by *holding out* a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

*Validation set approach*

* Involves randomly dividing the available set of observations into two parts, a training set and a validation or hold out set. The model is fit on the training set, and the fitted model is used to predict the responses for observations in the validation set. The resulting validation set error provides an estaimte of the test error rate
* Drawbacks of the validation approach:
  + The validation estimate of the test error rate can be highly variable, depending on precisely which observations are included in the training set and which observatons are included in the validation set
  + Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to *overestimate* the test error rate for the model fit on the entire dataset

*Leave-one-out cross-validation*

* Closely related to the validation approach, but a single observation is used for the validation set
* MSE1 is a poor estimate since it is highly variable, since it is based upon a single observation
* Repeating this approach *n* times produces *n* squared errors. The LOOCV estaimte for the test MSE is the average of these *n* error estimates
* Advantages of LOOCV over the validation approach:
  + First, it has far less bias, because we repeatedly fit the statistical learning method using training sets that contain *n*-1 observations, almost as many as are in the entire data set. Consequently, the LOOCV approach tends not to overestaimte the test error rate as much as the validation set approach does
  + Second, performing the LOOCV multiple times will always yield the same results
* LOOCV is a very general method, and can be used with any kind of predictive modelling

*K-fold cross validation*

* An alternative to LOOCV is *k-fold CV*. This approach involves randomly dividing the set of observations into *k* groups, or *folds*, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k-1 folds. The mean squared error (MSE), is then computed on the observations in the held-out fold. This procedure is repeated *k* times; each time, a different group of observations is treated as the validation set.
* *K fold* often gives more accurate estimates of the test error rate than does LOOCV. This has to do with a bias-variance trade-off.
* LOOCV will give approximately unbiased estimates of the test error, since each training set contains *n* – 1 observations, whilst performing the *k*-fold CV for say, *k* = 5 or *k* = 10 will lead to an intermediate level of bias.
* Yet LOOCV has a higher variance than k folds: since the mean of many highly correlated quantities has higher variance than does the mean of many quantities that are not as highly correlated, the test error estaimtate resulting from LOOCV tends to have higher variance than does the test error estimate resulting from *k-*fold CV.
* Typically, given these considerations, one performs *k* fold cross validation using *k* = 5 or *k* = 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.

*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
* 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
* The bootstrap approach allows us to use a computer to emulate the process of obtaining new sample sets, so that we can estimate the variability of a without generating additional samples. Rather than repeatedly obtaining independent data sets from the population, we instead obtain datasets by repeatedly sampling observations from the original dataset.