Chapter 5

Resampling Methods is when you repeatedly draw samples from a training set and refit the model of interest on each sample to obtain additional information about the fitted model.

- They can be computationally expensive.

Two most common resampling methods:

1. Cross-validation
   Model Assessment: Evaluating model performance
   Can be used to estimate the test error
   Model Selection: Select appropriate level of flexibility
2. Bootstrap
   Provide a measure of accuracy of a parameter estimate or statistical model.

5.1 Cross-Validation

Holding out: keeping a subset of training observations out of the fitting process and then applying the method to those observations
This is because we need to find the *test* error rate, but often do not have a test set available.
Since the *training* error rate is often very different than the *test* error rate, *holding out* essentially allows us to have our cake and eat it too.

5.1.1 The Validation Set Approach

Step 1: Randomly divide the availble set of observations into two parts:

1. The *training set*
2. The *validation set* or *hold-out set*

Step 2: Fit the model using the training set then use the predicted responses for the observations in the validation set.

- The validation set error rate (usually using MSE for quantitative data) gives an estimate of the *test* error rate.

Drawbacks to the validation set approach\

1. The validation estimate of the test error rate can be highly variable.
   This is because it is decided by which data is included in the training vs. validation set.\
2. Since only a subset of observations are used to train the method, it may *overestimate* the test error rate.

Cross-validation is a refined version of the validation set approach & addresses these two issues.

5.1.2 Leave-One-Out-Cross-Validation (LOOCV)

This approach also splits the data into two parts, but a single observation is used for validation (x_1, x_0) while the rest is used to make up the training set $(x_2, y_2, ..., x_n, y_n)$.\

- LOOCV is a general method that can be used with any kind of predictive modeling.
- The model is then fit on the training set $(n-1)$.
- Prediction $(\hat{y}_1)$ is made for the excluded observation using its value x_1.
- Since (x_1, y_1) wasn't used in the fitting process, $MSE_1 = (y_1 - \hat{y}_2)$ is an approximately unbiased estimate for the test error.
- It is highly variable, though, since it's based on a single observation.
  - To combat the variablity, the process is run multiple times. The result of the LOOCV estimate for the test MSE is the average of these n test error estimates:
    $$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

Advantages to LOOCV

1. It has less bias.
2. Performing LOOCV multiple times will always result in the same results because there is no randomness in the training/validation set splits.

Potential Downsides

1. It might be computationally expensive, since it fits each point.
2. It might take awhile if n is large.

This is easy to vercome when using least squares or polynomial regression:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

- $\hat{y}_i$ is the ith fitted value from the original least squares fit
- h_i is the leverage

The equation holds because the leverage lies between $1/n$ and 1 and reflects the amount an observation influences its own fit.
Thus, the residuals for high-leverage points are inflated the right amount for the equality to hold.

5.1.3 k-Fold Cross-Validation (k-fold CV)

This approach randomly divides a set of observations into k groups (aka *folds*) of roughly equal size.
The first fold is treated as a validation set. The method is then fit on the remaining $k-1$ folds. The MSE_1 is computed on the observations in the validation set.
The process is repeated k times with a different set held out as the validation set.
Since the process is repeated k number of times, the *k-fold CV* estimate is the average of those values:

- $CV_(k) = \frac{1}{k}\sum_{i=1}^{k} MSE_i$

k-fold CV is usually only used when $k=5$ and $k=10$ rather than $k=n$ as is done for $LOOCV$
As a result, *k-fold CV* is less computationally expensive. Since it is a general method it can be applied to most statistical learning methods.

5.1.4 Bias-Variance Trade-Off for k-fold Cross-Validation

k-fold CV often gives a more accurate estimate of the test error rate than *LOOCV*

- *LOOCV* will result in less bias since it is using $n-1$ observations (nearly the same amount as total observations.)
- *k-fold CV* will result in intermediate bias since the training set contains $(k-1)n/k$ observations (more than *LOOCV* but less than the validation set approach)
- But, *LOOCV* has a higher variance than *k-fold CV* does when $k < n$
  - This is because the outputs of *LOOCV* are highly positively correlated with each other since each model is trained on almost identical data.
  - In contrast, *k-fold CV* averages the outputs of the fitted models and are thus less correlated with each other since the overlap between training sets is smaller.

5.1.5 Cross-Validation on Classification Problems

For problems when Y is qualitative, instead of using the MSE to quantify test error we use the number of misclassified observations.
LOOCV for qualitative problems:

- $CV_n = \frac{1}{n}\sum_{i=1}^{n} Err_i$

- where $Err_i = I(y_i \neq \hat{y}_i)$

k-fold CV and *validation set* approaches remain the same.

5.2 The Bootstrap

Bootstrap is a tool that can be used to "quantify the uncertainty associated with a given estimator or statistical learning method."

- It can be used to estimate standard errors of the coefficients in linear regression fits.
  - R does this automatically, but *bootstrap* can be used for other methods which are harder to obtain estimates for.
- It uses the computer to emulate the process of getting new datasets so we can estimate variability without generating more samples.
- It obtains distinct data sets by repeatedly sampling observations from the original dataset.

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

- this equation computes the standard error of the bootstrap estimates.