Lesson 10 k Fold Cross-Validation

Kevin Cummiskey

February 3, 2020

Review

Last lesson, we discussed three models to predicts the number of wins in a season.

$$Wpct = \beta_0 + \beta_1 RD + \epsilon \tag{1}$$

$$Wpct = \frac{R^2}{R^2 + RA^2} + \epsilon \tag{2}$$

$$Wpct = \frac{R^k}{R^k + RA^k} + \epsilon \tag{3}$$

where $Wpct$ is Win Percentage, R is Runs Scored, RA is Runs Allowed, and ϵ is the random error. Recall that we fit Model 1 to the 1997-2001 seasons.

When prediction is the goal of a statistical model, *overfitting* the model is a big concern. What is *overfitting*?

k-fold cross validation

k-fold cross validation is one method to assess how well a model will predict on new data. In this method, we randomly partition the data into k groups. We set one group (called the *test set*) aside, fit the model on the remaining data (the *training set*), and assess the performance on the test set. We repeat the process k times with each partition as the test set.

Here is a summary of the steps:

1. Randomly partition the data set into k groups.
2. Set one partition (the test set) aside.
3. Fit the model on the remaining data (the training set).
4. Calculate RMSE.
5. Repeat with each partition as the test set.