# FIVE-FOLD CROSS VALIDATION

### Five-Fold Cross-Validation

We have, so far, looked upon various models in order to find a better performing one. However, it can be difficult to determine if these improvements in scores result from the captures of better relationships within our model or if we are just overfitting the model. In order to clarify this aspect we use validation techniques such as *k*-Fold Cross Validation.

In the cross-validation, the training set is divided into sub-samples, and each single sub-sample will be saved as the data for the verification of model while the other *k-1* groups of sample will be used for training. Cross-validation is repeated *k* times, of which each sub-sample is verified once. The average number of results or other combinations are used, and a single estimate is finally obtained. The advantage of this method is that it repeatedly uses randomly generated sub-samples for training and validation. Actually, 10-Fold Cross Validation is the most commonly used.

In our experiment, *k* has a specific value, 5, the reference to the model with be 5-Fold Cross-Validation.

### Mean Square Error (MSE)

Mean Square Error (MSE) is used to evaluate the quality of an [estimator](https://en.wikipedia.org/wiki/Estimator) (parameter) or a predictor (some [random variable](https://en.wikipedia.org/wiki/Random_variable)), in other words,is the average of the square of the errors. MSE satisfies the equation as below:

MSE (*T*) = var (*T*) + (bias (*T*))2

where bias(*T*) = E(*T*) - **

Ususally, if the MSE of one model is larger, the error of this model will be larger.

## Predictions

Cross-validation for linear regression can be easily achieved in R Studio. The function ‘cv.lm’ can predict the accuracy for multiple linear regression. By using the function ‘cv.lm’, the output will be displayed, including the analysis of variance table and the observations in five test sets. The whole output can be found in Appendix 2. The output gives the plots of the cross-validation predicted value of two models as below in Figure 15 and 16. It is quite hard to say whether **dataModel** or **finalModel** is the better because the five regression lines all seems parallel in both plots.

Thus, the focus of the output should be the comparison of overall ms (mean square) of both finalModel and dataModel. From the output, the overall ms of finalModel is 268 whilst which of dataModel is 255. It can be predicted that dataModel is a bit more suitable for this case than finalModel.

|  |
| --- |
| Preditedcrossvalidation-final.png  Figure 15: Cross-Validation predicted values for FinalModel |

|  |
| --- |
| Predictedobserveddata.png  Figure 16: Cross Validation Predicted Values for DataModel |

Also, by using get\_mse function, the output turns out that the MSE value of finalModel is 258 while the MSE value of dataModel is 248. Although the discrepancy between these two model is not so large, it is clear that dataModel is better than finalModel in this case.

Appendix 2 Output

