Team 2 - Homework Two

Assignment 2: KJ 7.2; KJ 7.5

Juliann McEachern

10/23/19

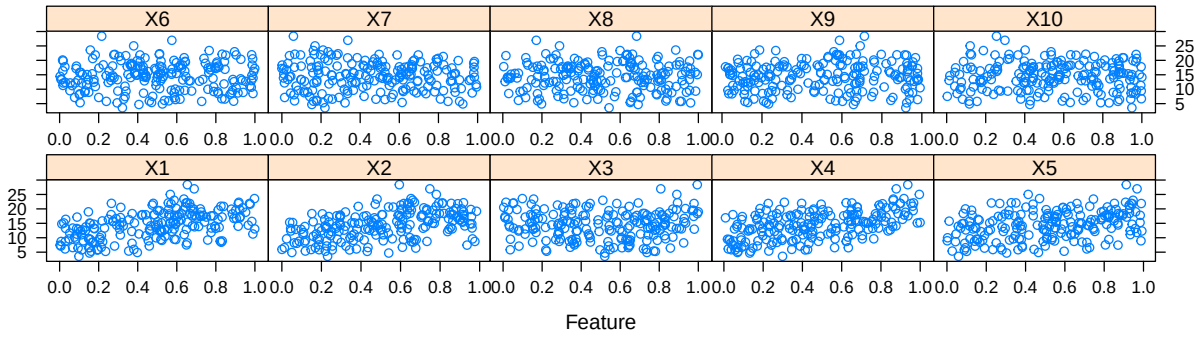
Dependencies

```
# predictive modeling
libraries("mlbench", "caret", "mice", "AppliedPredictiveModeling")

# Formatting Libraries
libraries("default", "knitr", "kableExtra")

# Plotting Libraries
libraries("ggplot2", "grid", "ggfortify")
```

(1) Kuhn & Johnson 7.2

> Friedman (1991) introduced several benchmark data sets create by simulation. One of these simulations used the following nonlinear equation to create data: $y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + N(0, \sigma^2)$; where the x values are random variables uniformly distributed between $[0, 1]$ (there are also 5 other non-informative variables also created in the simulation).

The package `mlbench` contains a function called `mlbench.friedman1` that simulates these data:



(a) Tune several models on these data. For example:

```
knnModel <- train(x = trainingData$x, y = trainingData$y,
    method = "knn", preProc = c("center", "scale"),
    tuneLength = 10)
knnModel

knnPred <- predict(knnModel, newdata = testData$x)
```

```r
## The function 'postResample' can be used to get
## the test set performance values
postResample(pred = knnPred, obs = testData$y)
```

Model 1:

Train set model & performance:

```
Linear Regression

200 samples
 10 predictor

Pre-processing: principal component signal extraction (10), centered
 (10), scaled (10)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 200, 200, 200, 200, 200, 200, ...
Resampling results:

  RMSE      Rsquared   MAE
  2.497918  0.7557065  1.992082

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Test set performance values:

```
     RMSE  Rsquared       MAE
2.6970680 0.7084666 2.0600540
```

Model 2:

Train set model & performance:

```
Partial Least Squares

200 samples
 10 predictor

Pre-processing: principal component signal extraction (10), centered
 (10), scaled (10)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 200, 200, 200, 200, 200, 200, ...
Resampling results across tuning parameters:

  ncomp  RMSE      Rsquared   MAE
  1      2.694689  0.7067492  2.128219
  2      2.509652  0.7453022  1.974252
  3      2.518209  0.7440264  1.984267
  4      2.520068  0.7437803  1.984709
  5      2.519658  0.7438896  1.984688
```

```
RMSE was used to select the optimal model using the smallest value.
The final value used for the model was ncomp = 2.
```

Test set performance values:

```
    RMSE Rsquared      MAE
2.685591 0.710292 2.052676
```

Model 3:

Train set model & performance:

```
Multivariate Adaptive Regression Spline

200 samples
 10 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 200, 200, 200, 200, 200, 200, ...
Resampling results:

  RMSE      Rsquared   MAE
  1.840203  0.8617361  1.433966

Tuning parameter 'nprune' was held constant at a value of 10

Tuning parameter 'degree' was held constant at a value of 1
```

Test set performance values:

```
    RMSE Rsquared      MAE
1.776575 0.872700 1.358367
```

(b) Which models appear to give the best performance? Does MARS select the informative predictors (those named X1-X5)?

Table 1: Train Set Performance

	RMSE	**RSquared**	**MAE**
lmModel	2.497918	0.7557065	1.992082
plsModel	2.509652	0.7453022	1.974252
marsModel	1.840203	0.8617361	1.433966

Table 2: Test Set Performance

	RMSE	**Rsquared**	**MAE**
lmPerf	2.697068	0.7084666	2.060054
plsPerf	2.685591	0.7102920	2.052676
marsPerf	1.776575	0.8727000	1.358367

(2) Kuhn & Johnson 7.5

Exercise 6.3 describes data for a chemical manufacturing process. Use the same data imputation, data splitting, and pre-processing steps as before and train several nonlinear regression models.

(a) Which nonlinear regression model gives the optimal resampling and test set performance?

(b) Which predictors are most important in the optimal nonlinear regression model? Do either the biological or process variables dominate the list? How do the top ten important predictors compare to the top ten predictors from the optimal linear model?

(c) Explore the relationships between the top predictors and the response for the predictors that are unique to the optimal nonlinear regression model. Do these plots reveal intuition about the biological or process predictors and their relationship with yield?