Homework Part Two

Assignment 1: KJ 6.3

Vinicio Haro

DATE:2019-10-25

Dependencies

```
# Package to use knn imputing if
# (!requireNamespace('BiocManager', quietly =
# TRUE)) install.packages('BiocManager')

# BiocManager::install('impute')

options(tinytex.verbose = TRUE)
# Predicitve Modeling
libraries("AppliedPredictiveModeling", "caret", "mice",
    "glmnet", "impute")
# Formatting Libraries
libraries("default", "knitr", "kableExtra", "tidyverse")
# Plotting Libraries
libraries("ggplot2", "grid", "ggfortify", "DataExplorer")
```

(1) Kuhn & Johnson 6.3

> A chemical manufacturing process for a pharmaceutical product was discussed in Sect.1.4. In this problem, the objective is to understand the relationship between biological measurements of the raw materials (predictors), measurements of the manufacturing process (predictors), and the response of product yield. Biological predictors cannot be changed but can be used to assess the quality of the raw material before processing. On the other hand, manufacturing process predictors can be changed in the manufacturing process. Improving product yield by 1% will boost revenue by approximately one hundred thousand dollars per batch: > **(a). Start R and use these commands to load the data:** The data contains 176 observations with 58 variables.BiologicalMaterial07 might be a zero variance predictor and we will investigate further.

```
'data.frame':   176 obs. of  58 variables:
 $ Yield                 : num  38 42.4 42 41.4 42.5 ...
 $ BiologicalMaterial01  : num  6.25 8.01 8.01 8.01 7.47 6.12 7.48 6.94 6.94 6.94 ...
 $ BiologicalMaterial02  : num  49.6 61 61 61 63.3 ...
 $ BiologicalMaterial03  : num  57 67.5 67.5 67.5 72.2 ...
 $ BiologicalMaterial04  : num  12.7 14.6 14.6 14.6 14 ...
 $ BiologicalMaterial05  : num  19.5 19.4 19.4 19.4 17.9 ...
 $ BiologicalMaterial06  : num  43.7 53.1 53.1 53.1 54.7 ...
 $ BiologicalMaterial07  : num  100 100 100 100 100 100 100 100 100 100 ...
 $ BiologicalMaterial08  : num  16.7 19 19 19 18.2 ...
 $ BiologicalMaterial09  : num  11.4 12.6 12.6 12.6 12.8 ...
 $ BiologicalMaterial10  : num  3.46 3.46 3.46 3.46 3.05 3.78 3.04 3.85 3.85 3.85 ...
 $ BiologicalMaterial11  : num  138 154 154 154 148 ...
 $ BiologicalMaterial12  : num  18.8 21.1 21.1 21.1 21.1 ...
 $ ManufacturingProcess01: num  NA 0 0 0 10.7 12 11.5 12 12 12 ...
 $ ManufacturingProcess02: num  NA 0 0 0 0 0 0 0 0 0 ...
```

```
 $ ManufacturingProcess03: num  NA NA NA NA NA NA 1.56 1.55 1.56 1.55 ...
 $ ManufacturingProcess04: num  NA 917 912 911 918 924 933 929 928 938 ...
 $ ManufacturingProcess05: num  NA 1032 1004 1015 1028 ...
 $ ManufacturingProcess06: num  NA 210 207 213 206 ...
 $ ManufacturingProcess07: num  NA 177 178 177 178 178 177 178 177 177 ...
 $ ManufacturingProcess08: num  NA 178 178 177 178 178 178 178 177 177 ...
 $ ManufacturingProcess09: num  43 46.6 45.1 44.9 45 ...
 $ ManufacturingProcess10: num  NA NA NA NA NA NA 11.6 10.2 9.7 10.1 ...
 $ ManufacturingProcess11: num  NA NA NA NA NA NA 11.5 11.3 11.1 10.2 ...
 $ ManufacturingProcess12: num  NA 0 0 0 0 0 0 0 0 0 ...
 $ ManufacturingProcess13: num  35.5 34 34.8 34.8 34.6 34 32.4 33.6 33.9 34.3 ...
 $ ManufacturingProcess14: num  4898 4869 4878 4897 4992 ...
 $ ManufacturingProcess15: num  6108 6095 6087 6102 6233 ...
 $ ManufacturingProcess16: num  4682 4617 4617 4635 4733 ...
 $ ManufacturingProcess17: num  35.5 34 34.8 34.8 33.9 33.4 33.8 33.6 33.9 35.3 ...
 $ ManufacturingProcess18: num  4865 4867 4877 4872 4886 ...
 $ ManufacturingProcess19: num  6049 6097 6078 6073 6102 ...
 $ ManufacturingProcess20: num  4665 4621 4621 4611 4659 ...
 $ ManufacturingProcess21: num  0 0 0 0 -0.7 -0.6 1.4 0 0 1 ...
 $ ManufacturingProcess22: num  NA 3 4 5 8 9 1 2 3 4 ...
 $ ManufacturingProcess23: num  NA 0 1 2 4 1 1 2 3 1 ...
 $ ManufacturingProcess24: num  NA 3 4 5 18 1 1 2 3 4 ...
 $ ManufacturingProcess25: num  4873 4869 4897 4892 4930 ...
 $ ManufacturingProcess26: num  6074 6107 6116 6111 6151 ...
 $ ManufacturingProcess27: num  4685 4630 4637 4630 4684 ...
 $ ManufacturingProcess28: num  10.7 11.2 11.1 11.1 11.3 11.4 11.2 11.1 11.3 11.4 ...
 $ ManufacturingProcess29: num  21 21.4 21.3 21.3 21.6 21.7 21.2 21.2 21.5 21.7 ...
 $ ManufacturingProcess30: num  9.9 9.9 9.4 9.4 9 10.1 11.2 10.9 10.5 9.8 ...
 $ ManufacturingProcess31: num  69.1 68.7 69.3 69.3 69.4 68.2 67.6 67.9 68 68.5 ...
 $ ManufacturingProcess32: num  156 169 173 171 171 173 159 161 160 164 ...
 $ ManufacturingProcess33: num  66 66 66 68 70 70 65 65 65 66 ...
 $ ManufacturingProcess34: num  2.4 2.6 2.6 2.5 2.5 2.5 2.5 2.5 2.5 2.5 ...
 $ ManufacturingProcess35: num  486 508 509 496 468 490 475 478 491 488 ...
 $ ManufacturingProcess36: num  0.019 0.019 0.018 0.018 0.017 0.018 0.019 0.019 0.019 0.019 ...
 $ ManufacturingProcess37: num  0.5 2 0.7 1.2 0.2 0.4 0.8 1 1.2 1.8 ...
 $ ManufacturingProcess38: num  3 2 2 2 2 2 2 2 3 3 ...
 $ ManufacturingProcess39: num  7.2 7.2 7.2 7.2 7.3 7.2 7.3 7.3 7.4 7.1 ...
 $ ManufacturingProcess40: num  NA 0.1 0 0 0 0 0 0 0 0 ...
 $ ManufacturingProcess41: num  NA 0.15 0 0 0 0 0 0 0 0 ...
 $ ManufacturingProcess42: num  11.6 11.1 12 10.6 11 11.5 11.7 11.4 11.4 11.3 ...
 $ ManufacturingProcess43: num  3 0.9 1 1.1 1.1 2.2 0.7 0.8 0.9 0.8 ...
 $ ManufacturingProcess44: num  1.8 1.9 1.8 1.8 1.7 1.8 2 2 1.9 1.9 ...
 $ ManufacturingProcess45: num  2.4 2.2 2.3 2.1 2.1 2 2.2 2.2 2.1 2.4 ...
```

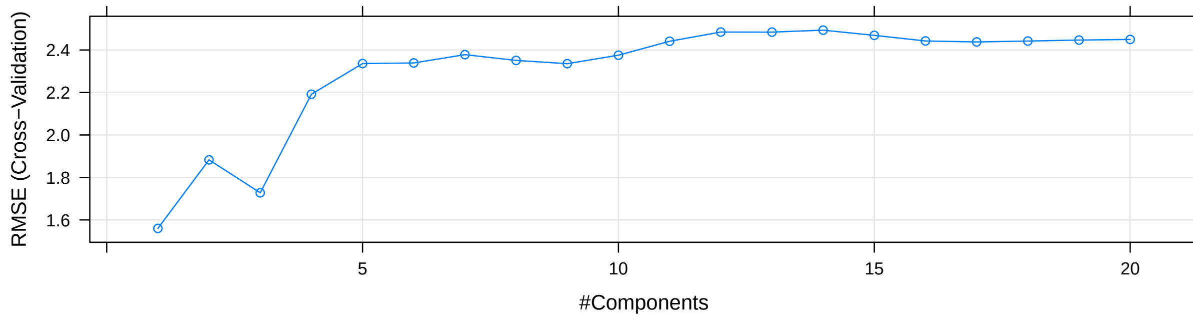
The matrix processPredictors contains the 57 predictors (12 describing the input biological material and 45 describing the process predictors) for the 176 manufacturing runs. yield contains the percent yield for each run.

(b). A small percentage of cells in the predictor set contain missing values. Use an imputation function to fill in these missing values (e.g., see Sect. 3.8). Our missing data plot shows that the target variable is complete. Manufactuing process 03 is missing 8.52 percent of entires. There are more predictors missing less than 3 percent of their entries. This is an ideal situation to impute variables. The impute package is not available in CRAN. We need to install it directly from BiocManager. We utilize knn method to impute missing values across all variables with missing data. We essentially use k nearest neighbors toimpute the missing values. For each variable with missing data, we use Euclidean distance to identify the k nearest neighbors. If we are missing a coordinate to compute the distance, the package uses the average distance from the closest non missing coordinates. This package assumes that not all variables are missing data. Some other methods of imputation include using the mean or median of each variable to fill in the NA's however the impute package allows KNN to be done in a single line.

(c). Split the data into a training and a test set, pre-process the data, and tune a model of your choice from this chapter. What is the optimal value of the performance metric? We can see several predictors that ar quite correlated with each other. We can use a function to apply a correlation threshold and remove pairwise correlations. We removed any pairwise correlation greater than .7 (arbitrary choice). We are essentially being proactive when it comes to avoiding multicolinearity. We will be fitting a partial least squares model using the train function. We specify method to pls and request the 20 best fits based on RMSE. We build the model on the features that were selected from dropping variables that had pairwise correlation. We also use 10 fold cross validation. On a high level, this means that we will parition the training data into k equally sized sets and retain one of those ki sets to validate our model. The plot parameter revealed the the optimal value of components. In terms of r squared , ncomp 13 is the ideal parameter.

```
Partial Least Squares

144 samples
 35 predictor

Pre-processing: centered (35), scaled (35)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 130, 131, 128, 130, 130, 129, ...
Resampling results across tuning parameters:

  ncomp  RMSE      Rsquared   MAE
   1     1.560007  0.3891078  1.232125
   2     1.882913  0.4172155  1.292349
   3     1.727661  0.4592812  1.251877
   4     2.191698  0.4292503  1.400555
   5     2.336186  0.4337308  1.432827
   6     2.339050  0.4369019  1.428144
   7     2.378060  0.4476128  1.421328
   8     2.351002  0.4544916  1.406405
   9     2.335555  0.4576866  1.391024
  10     2.375508  0.4549352  1.400848
  11     2.441100  0.4528699  1.419688
  12     2.484580  0.4490261  1.432653
  13     2.484220  0.4533374  1.433991
  14     2.493490  0.4543693  1.437662
  15     2.468935  0.4584692  1.428240
  16     2.442578  0.4614583  1.420099
  17     2.437957  0.4630998  1.418461
  18     2.442169  0.4634011  1.419379
  19     2.446810  0.4635929  1.420389
  20     2.449727  0.4643199  1.421220

RMSE was used to select the optimal model using the smallest value.
```

```
The final value used for the model was ncomp = 1.
```



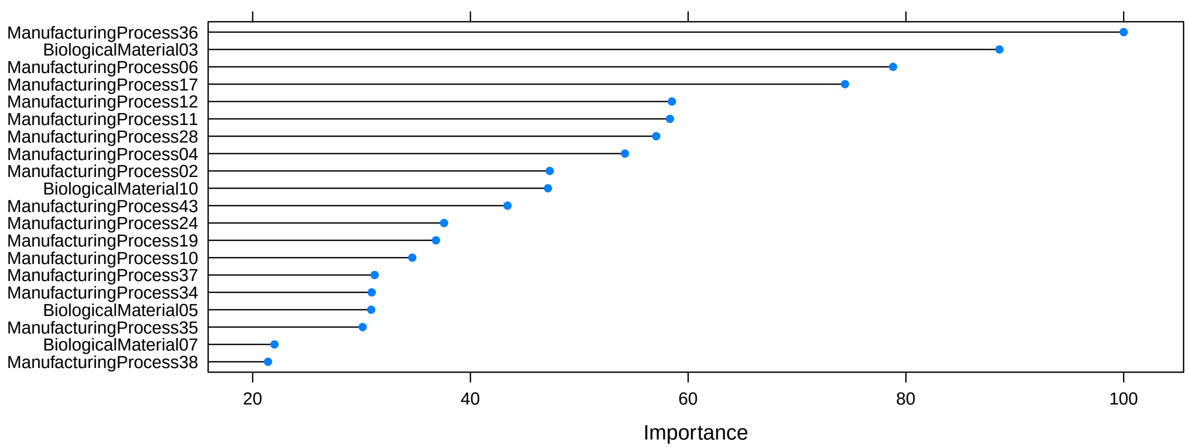
(d). Predict the response for the test set. What is the value of the performance metric and how does this compare with the resampled performance metric on the training set? The test data produces a RMSE of 1.5444438, r squared of 0.4223426 and MAE 1.2908355.Recall the metrics from nComp13. With training data we got RMSE of 2.392602, R squared of 0.4211762 and MAE of 1.325125. There is a decrease in rMSE, however we still get roughly 40 percent of the data variability explained when using the training data vs test data. The problem did NOT specifiy to pick the best model but rather a model of our choice, however we can speculate on how to potentially imporve our results. I think given the type of data, we would benefit from applying method of smoothing splines. Splines balance the overall goodness of fit by applying the derivative of functions generated on noisy data. I would also recommed additive regression methods.

```
     RMSE  Rsquared       MAE
1.4657792 0.3327395 1.1385856
```

(e). Which predictors are most important in the model you have trained? Do either the biological or process predictors dominate the list? Manufacturing Process 36 is the most important predictor followed by BiologicalMaterial03.Overall, the process is doinated by manufacturing process predictors.

```
, , 1 comps

                           .outcome
BiologicalMaterial03    0.247356455
BiologicalMaterial05    0.088394378
BiologicalMaterial07   -0.063928707
BiologicalMaterial09    0.040444529
BiologicalMaterial10    0.133123649
ManufacturingProcess01 -0.059898507
ManufacturingProcess02 -0.133579881
ManufacturingProcess03 -0.045630016
ManufacturingProcess04 -0.152606990
ManufacturingProcess05  0.050681758
ManufacturingProcess06  0.220434418
ManufacturingProcess07 -0.027636631
ManufacturingProcess08  0.008692913
ManufacturingProcess10  0.098791122
ManufacturingProcess11  0.163991396
ManufacturingProcess12  0.164484979
ManufacturingProcess16 -0.016371809
```

```
ManufacturingProcess17 -0.208290130
ManufacturingProcess19  0.104785161
ManufacturingProcess20 -0.030279376
ManufacturingProcess21 -0.006562046
ManufacturingProcess22  0.004510278
ManufacturingProcess23 -0.042809642
ManufacturingProcess24 -0.106829029
ManufacturingProcess25  0.003305285
ManufacturingProcess28  0.160523715
ManufacturingProcess34  0.088526155
ManufacturingProcess35 -0.086230402
ManufacturingProcess36 -0.278795312
ManufacturingProcess37 -0.089267175
ManufacturingProcess38 -0.062296854
ManufacturingProcess39  0.005792439
ManufacturingProcess41 -0.019524610
ManufacturingProcess43  0.122890032
ManufacturingProcess45 -0.005554439
```



> **(f). Explore the relationships between each of the top predictors and the response. How could this information be helpful in improving yield in future runs of the manufacturing process?** We are unable to change biological process but make alterations to the raw input materials that go into the biological process. Based on the importance of bio process 3, we could perhaps explore making changes into the raw materials. Manufacturing process 36 is the most important. I suggest using experimental design to compare that particular process with the other manufacturing processes. We want to see why a process such as 19 is not as important as 36.

If we examine our correlations, we see that ManufacturingProcess36 has strong negative correlation with Yield. That variable would be one that merits furthur analysis into why it has such a negative correlation with yield.

```
                            Yield ManufacturingProcess36
Yield                   1.0000000            -0.52500284
ManufacturingProcess36 -0.5250028             1.00000000
BiologicalMaterial03    0.4450860            -0.46578804
ManufacturingProcess17 -0.4258069            -0.03947942
ManufacturingProcess11  0.3302385             0.10435956
ManufacturingProcess06  0.3878354            -0.25131410
```

	BiologicalMaterial03	ManufacturingProcess17
Yield	0.44508598	-0.42580687
ManufacturingProcess36	-0.46578804	-0.03947942
BiologicalMaterial03	1.00000000	-0.09760502
ManufacturingProcess17	-0.09760502	1.00000000
ManufacturingProcess11	-0.09185407	-0.54602913
ManufacturingProcess06	0.18373279	-0.25603100

	ManufacturingProcess11	ManufacturingProcess06
Yield	0.33023849	0.3878354
ManufacturingProcess36	0.10435956	-0.2513141
BiologicalMaterial03	-0.09185407	0.1837328
ManufacturingProcess17	-0.54602913	-0.2560310
ManufacturingProcess11	1.00000000	0.2889326
ManufacturingProcess06	0.28893257	1.0000000