EX6

Chathura J Gunasekara

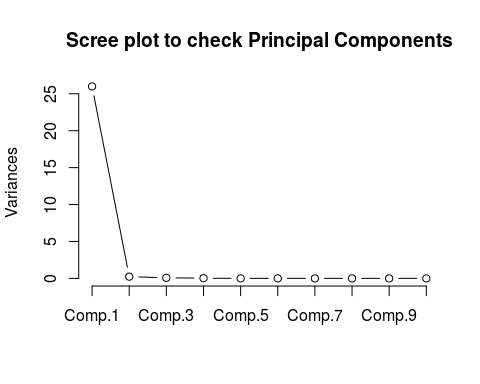
Thursday, October 16, 2014

1. a. Load the data

## Loading required package: lattice  
## Loading required package: ggplot2

1. Use PCA to determine the effective dimension of the data which is number of PCA Componenents. There are two PCA compnenets which is the effective dimention.

## [1] Using Scree Plot to check number of effective compents



## PC1 PC2  
## [1,] -4.3037 -0.4040  
## [2,] 0.9983 0.4865  
## [3,] -7.1414 1.4104  
## [4,] -1.8086 1.1541  
## [5,] 0.9883 -1.1921  
## [6,] 5.5620 0.2265

c.Using the created 2 PCs a traing and testing set is created.

## V1 PC1 PC2  
## 1 22.5 -4.3037 -0.4040  
## 2 40.1 0.9983 0.4865  
## 3 8.4 -7.1414 1.4104  
## 4 5.9 -1.8086 1.1541  
## 5 25.5 0.9883 -1.1921  
## 6 42.7 5.5620 0.2265

1. Ordinary Linear Regression

##   
## Call:  
## lm(formula = V1 ~ ., data = newdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.36 -7.01 -3.23 5.55 30.95   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 18.142 0.763 23.78 < 2e-16 \*\*\*  
## PC1 0.563 0.077 7.31 5.5e-12 \*\*\*  
## PC2 -2.713 0.776 -3.49 0.00058 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.2 on 212 degrees of freedom  
## Multiple R-squared: 0.236, Adjusted R-squared: 0.229   
## F-statistic: 32.8 on 2 and 212 DF, p-value: 3.9e-13

## [1] Testing data

## RMSE Rsquared   
## 10.6231 0.1799

1. Robust Linear Regresssion

##   
## Call: rlm(formula = V1 ~ ., data = newdata)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.82 -5.36 -1.90 6.61 32.47   
##   
## Coefficients:  
## Value Std. Error t value  
## (Intercept) 16.897 0.714 23.649   
## PC1 0.598 0.072 8.286   
## PC2 -3.147 0.727 -4.327   
##   
## Residual standard error: 8.45 on 212 degrees of freedom

## [1] Testing data

## RMSE Rsquared   
## 10.8706 0.1784

1. Partial Least Squares

##   
## Attaching package: 'pls'  
##   
## The following object is masked from 'package:caret':  
##   
## R2  
##   
## The following object is masked from 'package:stats':  
##   
## loadings

## Data: X dimension: 215 2   
## Y dimension: 215 1  
## Fit method: kernelpls  
## Number of components considered: 2  
## TRAINING: % variance explained  
## 1 comps 2 comps  
## X 99.02 100.00  
## V1 19.32 23.63

## [1] Testing data

## , , 1 comps  
##   
## V1  
## 2 18.69  
## 5 18.73  
## 6 21.27  
## 10 26.41  
## 23 10.32  
##   
## , , 2 comps  
##   
## V1  
## 2 17.38  
## 5 21.93  
## 6 20.66  
## 10 27.13  
## 23 12.57

## Partial Least Squares   
##   
## 215 samples  
## 2 predictors  
##   
## Pre-processing: centered, scaled   
## Resampling: Bootstrapped (25 reps)   
##   
## Summary of sample sizes: 215, 215, 215, 215, 215, 215, ...   
##   
## Resampling results  
##   
## RMSE Rsquared RMSE SD Rsquared SD  
## 11 0.2 0.6 0.06   
##   
## Tuning parameter 'ncomp' was held constant at a value of 1  
##

iv.ridge regresssion

## Loading required package: lars  
## Loaded lars 1.2

## 2 5 6 10 23 24   
## 17.04 22.22 20.64 27.74 11.99 26.23

## Ridge Regression   
##   
## 161 samples  
## 2 predictors  
##   
## Pre-processing: centered, scaled   
## Resampling: Bootstrapped (25 reps)   
##   
## Summary of sample sizes: 161, 161, 161, 161, 161, 161, ...   
##   
## Resampling results across tuning parameters:  
##   
## lambda RMSE Rsquared RMSE SD Rsquared SD  
## 0.000 12 0.3 1 0.09   
## 0.007 12 0.3 1 0.09   
## 0.014 12 0.3 1 0.09   
## 0.021 12 0.3 1 0.09   
## 0.029 12 0.3 1 0.09   
## 0.036 12 0.3 1 0.09   
## 0.043 12 0.3 1 0.09   
## 0.050 12 0.3 1 0.09   
## 0.057 12 0.3 1 0.09   
## 0.064 12 0.3 1 0.09   
## 0.071 12 0.3 1 0.09   
## 0.079 12 0.3 1 0.09   
## 0.086 12 0.3 1 0.09   
## 0.093 12 0.3 1 0.09   
## 0.100 12 0.3 1 0.09   
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was lambda = 0.1.

v.lasso model

## [1] "s" "fraction" "mode" "fit"

## 2 5 6 10 23 24   
## 17.93 17.93 18.35 19.18 16.56 19.00

## PC1 PC2   
## 0.09185 0.00000

(d)For this data set non of the models are siginificantly better or worse than others.A robust linear model with PCA should work because its interpretablity.

1. Robust Linear model will be used because it is easier to implement and give higer R-squred values for the fit and lower RMSE for testing data.

2. a.Loading data

1. Remove the near zero variance predictors.

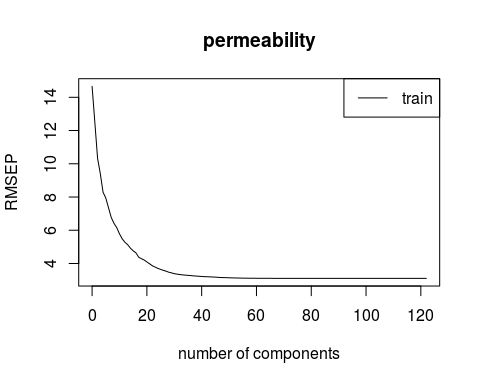
## [1] 388

## [1] New Dimension of the fingerprint is:

## [1] 165 388

c.Split the data and train a PLS model

## Data: X dimension: 123 388   
## Y dimension: 123 1  
## Fit method: kernelpls  
## Number of components considered: 122  
## TRAINING: % variance explained  
## 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps  
## X 27.22 42.56 49.89 53.22 62.49 66.94  
## permeability 27.29 50.33 58.79 68.02 70.66 74.84  
## 7 comps 8 comps 9 comps 10 comps 11 comps 12 comps  
## X 69.13 71.79 75.02 76.94 78.87 80.99  
## permeability 78.70 80.85 82.34 84.43 85.99 87.04  
## 13 comps 14 comps 15 comps 16 comps 17 comps 18 comps  
## X 83.36 84.77 86.18 87.58 88.13 89.31  
## permeability 87.75 88.72 89.43 89.95 91.06 91.43  
## 19 comps 20 comps 21 comps 22 comps 23 comps 24 comps  
## X 90.45 91.13 91.75 92.18 92.68 93.14  
## permeability 91.73 92.20 92.61 93.04 93.32 93.60  
## 25 comps 26 comps 27 comps 28 comps 29 comps 30 comps  
## X 93.61 93.96 94.29 94.51 94.81 95.00  
## permeability 93.80 94.00 94.16 94.37 94.50 94.65  
## 31 comps 32 comps 33 comps 34 comps 35 comps 36 comps  
## X 95.25 95.49 95.77 96.00 96.27 96.51  
## permeability 94.74 94.81 94.87 94.92 94.96 95.01  
## 37 comps 38 comps 39 comps 40 comps 41 comps 42 comps  
## X 96.77 96.97 97.14 97.32 97.51 97.74  
## permeability 95.05 95.09 95.13 95.16 95.20 95.22  
## 43 comps 44 comps 45 comps 46 comps 47 comps 48 comps  
## X 97.91 98.05 98.16 98.25 98.32 98.44  
## permeability 95.24 95.26 95.29 95.32 95.36 95.37  
## 49 comps 50 comps 51 comps 52 comps 53 comps 54 comps  
## X 98.54 98.6 98.69 98.79 98.86 98.93  
## permeability 95.39 95.4 95.41 95.43 95.44 95.45  
## 55 comps 56 comps 57 comps 58 comps 59 comps 60 comps  
## X 99.02 99.08 99.14 99.20 99.28 99.33  
## permeability 95.46 95.47 95.48 95.48 95.49 95.49  
## 61 comps 62 comps 63 comps 64 comps 65 comps 66 comps  
## X 99.39 99.44 99.48 99.52 99.57 99.6  
## permeability 95.49 95.50 95.50 95.50 95.50 95.5  
## 67 comps 68 comps 69 comps 70 comps 71 comps 72 comps  
## X 99.63 99.66 99.68 99.70 99.72 99.75  
## permeability 95.50 95.51 95.51 95.51 95.51 95.51  
## 73 comps 74 comps 75 comps 76 comps 77 comps 78 comps  
## X 99.76 99.79 99.81 99.83 99.85 99.87  
## permeability 95.51 95.51 95.51 95.51 95.51 95.51  
## 79 comps 80 comps 81 comps 82 comps 83 comps 84 comps  
## X 99.88 99.90 99.91 99.92 99.93 99.94  
## permeability 95.51 95.51 95.51 95.51 95.51 95.51  
## 85 comps 86 comps 87 comps 88 comps 89 comps 90 comps  
## X 99.95 99.96 99.97 99.98 99.98 99.98  
## permeability 95.51 95.51 95.51 95.51 95.51 95.51  
## 91 comps 92 comps 93 comps 94 comps 95 comps 96 comps  
## X 99.99 99.99 100.00 100.00 100.37 100.74  
## permeability 95.51 95.51 95.51 95.51 95.51 95.51  
## 97 comps 98 comps 99 comps 100 comps 101 comps  
## X 101.11 101.48 101.85 102.22 102.60  
## permeability 95.51 95.51 95.51 95.51 95.51  
## 102 comps 103 comps 104 comps 105 comps 106 comps  
## X 102.97 103.34 103.71 104.08 104.45  
## permeability 95.51 95.51 95.51 95.51 95.51  
## 107 comps 108 comps 109 comps 110 comps 111 comps  
## X 104.82 105.19 105.56 105.93 106.30  
## permeability 95.51 95.51 95.51 95.51 95.51  
## 112 comps 113 comps 114 comps 115 comps 116 comps  
## X 106.67 107.04 107.42 107.79 108.16  
## permeability 95.51 95.51 95.51 95.51 95.51  
## 117 comps 118 comps 119 comps 120 comps 121 comps  
## X 108.53 108.90 109.27 109.64 110.01  
## permeability 95.51 95.51 95.51 95.51 95.51  
## 122 comps  
## X 110.38  
## permeability 95.51

 40 variables are optimal according the above graph as it will explain 91% of the variation in the data.

## Partial Least Squares   
##   
## 123 samples  
## 388 predictors  
##   
## Pre-processing: centered, scaled   
## Resampling: Bootstrapped (25 reps)   
##   
## Summary of sample sizes: 123, 123, 123, 123, 123, 123, ...   
##   
## Resampling results across tuning parameters:  
##   
## ncomp RMSE Rsquared RMSE SD Rsquared SD  
## 1 13 0.3 2 0.1   
## 2 12 0.4 2 0.1   
## 3 12 0.4 2 0.1   
## 4 12 0.4 2 0.1   
## 5 12 0.4 2 0.2   
## 6 12 0.4 2 0.1   
## 7 12 0.4 2 0.1   
## 8 12 0.4 2 0.1   
## 9 12 0.4 2 0.1   
## 10 12 0.4 2 0.2   
## 11 12 0.4 2 0.2   
## 12 12 0.4 2 0.2   
## 13 12 0.4 2 0.2   
## 14 12 0.4 2 0.2   
## 15 13 0.4 2 0.2   
## 16 13 0.4 2 0.2   
## 17 13 0.4 2 0.2   
## 18 13 0.4 2 0.2   
## 19 14 0.4 2 0.2   
## 20 14 0.4 2 0.2   
## 21 14 0.4 2 0.2   
## 22 14 0.3 3 0.2   
## 23 14 0.3 3 0.2   
## 24 15 0.3 3 0.2   
## 25 15 0.3 3 0.2   
## 26 15 0.3 3 0.2   
## 27 15 0.3 3 0.2   
## 28 15 0.3 3 0.2   
## 29 16 0.3 3 0.2   
## 30 16 0.3 3 0.1   
## 31 16 0.3 3 0.1   
## 32 16 0.3 3 0.1   
## 33 16 0.3 3 0.1   
## 34 17 0.3 3 0.1   
## 35 17 0.3 3 0.1   
## 36 17 0.2 3 0.1   
## 37 17 0.2 3 0.1   
## 38 17 0.2 3 0.1   
## 39 18 0.2 4 0.1   
## 40 18 0.2 4 0.1   
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was ncomp = 6.

d.

e. i.Apply robust liner model

## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps  
## Warning: 'rlm' failed to converge in 20 steps

## Robust Linear Model   
##   
## 123 samples  
## 388 predictors  
##   
## Pre-processing: principal component signal extraction, scaled, centered   
## Resampling: Bootstrapped (25 reps)   
##   
## Summary of sample sizes: 123, 123, 123, 123, 123, 123, ...   
##   
## Resampling results  
##   
## RMSE Rsquared RMSE SD Rsquared SD  
## 13 0.3 2 0.1   
##   
##

f.

3.

a.Load Data

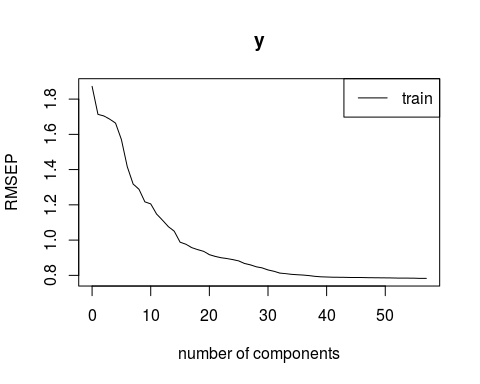
b.Impute missing values

## Bioconductor version 3.0 (BiocInstaller 1.16.0), ?biocLite for help  
## BioC\_mirror: http://bioconductor.org  
## Using Bioconductor version 3.0 (BiocInstaller 1.16.0), R version 3.1.1.  
## Installing package(s) 'impute'

##   
## The downloaded source packages are in  
## '/tmp/RtmpLUhPil/downloaded\_packages'

c.Split the data in to training and testing set. Train a PLS model

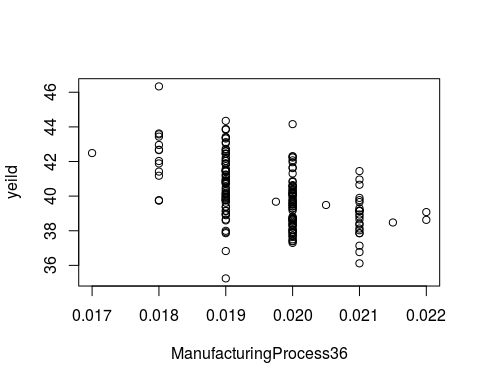
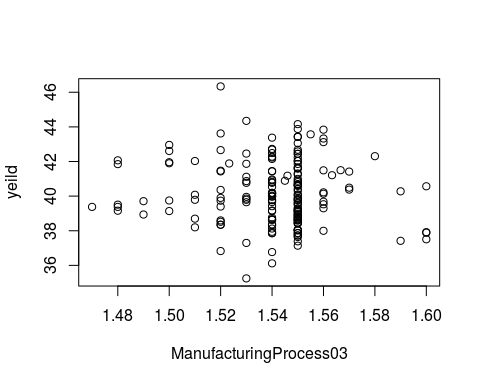
## Data: X dimension: 132 57   
## Y dimension: 132 1  
## Fit method: kernelpls  
## Number of components considered: 57  
## TRAINING: % variance explained  
## 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps  
## X 73.80 81.46 86.44 97.86 99.86 99.90 99.93 99.95  
## y 16.23 17.10 18.87 21.01 29.66 42.77 50.45 52.64  
## 9 comps 10 comps 11 comps 12 comps 13 comps 14 comps 15 comps  
## X 99.96 99.99 99.99 99.99 100.00 100.00 100.00  
## y 57.73 58.54 62.41 64.64 66.94 68.49 72.12  
## 16 comps 17 comps 18 comps 19 comps 20 comps 21 comps 22 comps  
## X 100.00 100.00 100.00 100.00 100.00 100.0 100.00  
## y 72.79 73.85 74.46 74.97 75.98 76.5 76.88  
## 23 comps 24 comps 25 comps 26 comps 27 comps 28 comps 29 comps  
## X 100.00 100.00 100.0 100.00 100.00 100.00 100.00  
## y 77.13 77.42 77.8 78.53 78.92 79.45 79.75  
## 30 comps 31 comps 32 comps 33 comps 34 comps 35 comps 36 comps  
## X 100.0 100.00 100.00 100.00 100.00 100.00 100.00  
## y 80.3 80.65 81.14 81.29 81.46 81.57 81.67  
## 37 comps 38 comps 39 comps 40 comps 41 comps 42 comps 43 comps  
## X 100.00 100.00 100.00 100.00 100.00 100.00 100.00  
## y 81.79 81.98 82.11 82.15 82.21 82.23 82.25  
## 44 comps 45 comps 46 comps 47 comps 48 comps 49 comps 50 comps  
## X 100.00 100.00 100.0 100.00 100.00 100.00 100.00  
## y 82.27 82.29 82.3 82.33 82.34 82.36 82.38  
## 51 comps 52 comps 53 comps 54 comps 55 comps 56 comps 57 comps  
## X 100.0 100.00 100.00 100.00 100.00 100.00 100.15  
## y 82.4 82.42 82.44 82.45 82.46 82.51 82.51

 From the above figure 40 components have optimal RMSE. d.

e.

## Overall  
## BiologicalMaterial01 0.0037636  
## BiologicalMaterial02 0.0061627  
## BiologicalMaterial03 0.0074524  
## BiologicalMaterial04 0.0056757  
## BiologicalMaterial05 0.0043282  
## BiologicalMaterial06 0.0064829  
## BiologicalMaterial07 0.0096931  
## BiologicalMaterial08 0.0061005  
## BiologicalMaterial09 0.0032605  
## BiologicalMaterial10 0.0054177  
## BiologicalMaterial11 0.0037374  
## BiologicalMaterial12 0.0053204  
## ManufacturingProcess01 0.0043896  
## ManufacturingProcess02 0.0028928  
## ManufacturingProcess03 0.0016489  
## ManufacturingProcess04 0.0032305  
## ManufacturingProcess05 0.0018557  
## ManufacturingProcess06 0.0055780  
## ManufacturingProcess07 0.0048562  
## ManufacturingProcess08 0.0042153  
## ManufacturingProcess09 0.0099172  
## ManufacturingProcess10 0.0021014  
## ManufacturingProcess11 0.0041500  
## ManufacturingProcess12 0.0001294  
## ManufacturingProcess13 0.0072740  
## ManufacturingProcess14 0.0047149  
## ManufacturingProcess15 0.0043144  
## ManufacturingProcess16 0.0001571  
## ManufacturingProcess17 0.0081822  
## ManufacturingProcess18 0.0016796  
## ManufacturingProcess19 0.0013171  
## ManufacturingProcess20 0.0018221  
## ManufacturingProcess21 0.0029222  
## ManufacturingProcess22 0.0028299  
## ManufacturingProcess23 0.0046533  
## ManufacturingProcess24 0.0064237  
## ManufacturingProcess25 0.0045760  
## ManufacturingProcess26 0.0047087  
## ManufacturingProcess27 0.0034934  
## ManufacturingProcess28 0.0029033  
## ManufacturingProcess29 0.0092933  
## ManufacturingProcess30 0.0052514  
## ManufacturingProcess31 0.0029302  
## ManufacturingProcess32 0.0095203  
## ManufacturingProcess33 0.0094482  
## ManufacturingProcess34 0.0026805  
## ManufacturingProcess35 0.0046266  
## ManufacturingProcess36 0.1180427  
## ManufacturingProcess37 0.0107505  
## ManufacturingProcess38 0.0040080  
## ManufacturingProcess39 0.0088175  
## ManufacturingProcess40 0.0038981  
## ManufacturingProcess41 0.0039788  
## ManufacturingProcess42 0.0079372  
## ManufacturingProcess43 0.0044322  
## ManufacturingProcess44 0.0065352  
## ManufacturingProcess45 0.0140504

ManufacturingProcess variables dominate the list with higher weighted sums.

f.   ManufacturingProcess36 has a high overall weight and when we plot yeild and ManufacturingProcess36 has a high correlation can be seen. But for ManufacturingProcess03 , it is low. So information about ManufacturingProcess36 is helpfull in improving the yeild.