Assignment-2

Sai Hemanth Nirujogi
G01065588

6.1.

a. *Start R and use these commands to load the data:*
The matrix absorp contains the 100 absorbance values for the 215 samples, while matrix endpoints contain the percent of moisture, fat, and protein in columns 1–3, respectively.

```
data(tecator)
colnames(absorp) <- paste("x", 1:ncol(absorp))
#?tecator
```

b. *In this example, the predictors are the measurements at the individual frequencies. Because the frequencies lie in a systematic order (850–1,050 nm), the predictors have a high degree of correlation. Hence, the data lie in a smaller dimension than the total number of predictors (215). Use PCA to determine the effective dimension of these data. What is the effective dimension?*

```
> dimPCA <- prcomp(absorp, center = TRUE, scale = TRUE)
> summary(dimPCA)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8      PC9
Standard deviation     9.9311 0.9847 0.52851 0.33827 0.08038 0.05123 0.02681 0.01961 0.008564
Proportion of Variance 0.9863 0.0097 0.00279 0.00114 0.00006 0.00003 0.00001 0.00000 0.000000
Cumulative Proportion  0.9863 0.9960 0.99875 0.99990 0.99996 0.99999 0.99999 1.00000 1.000000
```

The effective dimension of the data is 1.

c. *Split the data into a training and a test set, pre-process the data, and build each variety of models described in this chapter. For that model with tuning parameters, what are the optimal values of the tuning parameter(s)?*

```
set.seed(620)
Meatds <- createDataPartition(endpoints[, 3], p = 0.75, list=FALSE)

trainAbsorp <- absorp[Meatds,]
trainProtein <- endpoints[Meatds,3]
testAbsorp <- absorp[-Meatds,]
testProtein <- endpoints[-Meatds,3]
```

25 percent of the sample data is used for the training and a 10-fold cross-validation is used to tune all the models. The results for the 3 models are:

I. Linear Model

```
Linear Regression

163 samples
100 predictors

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 147, 147, 147, 147, 147, 145, ...
Resampling results:

  RMSE      Rsquared  MAE
  1.483321  0.821112  0.9536902
```

II. PLS Model

```
Partial Least Squares

163 samples
100 predictors

Pre-processing: centered (100), scaled (100)
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 147, 147, 147, 147, 147, 145, ...
Resampling results across tuning parameters:

  ncomp  RMSE       Rsquared   MAE
   1     2.9337895  0.1138314  2.5143775
   2     2.3945204  0.4018423  1.9178507
   3     1.9449256  0.6018515  1.4244656
   4     1.7419866  0.6676109  1.3072612
   5     1.1844909  0.8612779  0.9501617
   6     1.1250978  0.8773218  0.9133393
   7     1.0561196  0.8884972  0.8613973
   8     0.9174818  0.9192559  0.7527924
   9     0.8838507  0.9242559  0.7152432
  10     0.8381204  0.9315916  0.6693018
  11     0.7608766  0.9419999  0.5958721
  12     0.6782039  0.9550949  0.5274225
  13     0.6720517  0.9562167  0.5234067
  14     0.6745640  0.9569389  0.5158238
  15     0.6680331  0.9567662  0.5102568
  16     0.7204495  0.9503615  0.5267041
  17     0.7500900  0.9452052  0.5396308
  18     0.7675590  0.9422561  0.5529284
  19     0.7768575  0.9405623  0.5604828
  20     0.8024157  0.9351995  0.5735077

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was ncomp = 15.
```

III. PCR Model

```
Principal Component Analysis

163 samples
100 predictors

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 147, 147, 147, 147, 147, 145, ...
Resampling results across tuning parameters:

  ncomp  RMSE       Rsquared   MAE
   1     2.9369176  0.1123287  2.5179990
   2     2.6782288  0.2426552  2.1708358
   3     2.4270231  0.3916881  1.9659155
   4     1.8019010  0.6469462  1.3439441
   5     1.3424732  0.8145445  1.0440845
   6     1.1437292  0.8741566  0.9264311
   7     1.1509835  0.8721749  0.9317222
   8     1.1368664  0.8728376  0.9271907
   9     1.0298195  0.8954559  0.8455056
  10     0.9371124  0.9151046  0.7651904
  11     0.9616697  0.9107605  0.7811654
  12     0.8509773  0.9285942  0.6762401
  13     0.7684024  0.9407368  0.5971038
  14     0.7077277  0.9516703  0.5468964
  15     0.7042697  0.9522166  0.5494791
  16     0.6973325  0.9537104  0.5478574
  17     0.6633846  0.9577917  0.5200381
  18     0.6709386  0.9568306  0.5210565
  19     0.6819067  0.9552909  0.5254663
  20     0.6730057  0.9563437  0.5138729

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was ncomp = 17.
```

d. *Which model has the best predictive ability? Is any model significantly better or worse than the others?*
e. *Explain which model you would use for predicting the fat content of a sample.*

Based on the results of the 3 models created, the PLS model is the most preferred over the PCR because it handles highly correlated data better.

6.2. *Developing a model to predict permeability (see Sect. 1.4) could save sig- nificant resources for a pharmaceutical company, while at the same time more rapidly identifying molecules that have a sufficient permeability to become a drug:*
a. *Start R and use these commands to load the data:*
> library(AppliedPredictiveModeling)
> data(permeability)
The matrix fingerprints contain the 1,107 binary molecular predictors for the 165 compounds, while permeability contains permeability response.

The dataset consists of 1,107 binary molecular predictors for 165 compounds and the permeability contains the permeability response.

b. *The fingerprint predictors indicate the presence or absence of substructures of a molecule and are often sparse meaning that relatively few of the molecules contain each substructure. Filter out the predictors that have low frequencies using the nearZeroVar function from the caret package. How many predictors are left for modeling?*

388 predictors are left in the sample dataset and the rest 719 are removed after filtering out the predictors that have low frequencies using the nearZerovar.

```
> ncol(nonearZero)
[1] 388
```

c. *Split the data into a training and a test set, pre-process the data, and tune a PLS model. How many latent variables are optimal and what is the corresponding resampled estimate of R2?*

The dataset is divided into the training dataset with 80% and test set with 20%. The PLS model is tuned by dividing the dataset, 8 latent variables are optimal with R^2 is 0.4898897.

d. *Predict the response for the test set. What is the test set estimate of R2?*

```
> rsquarepls
[1] 0.5382089
```

The R^2 for the test set is 0.5382089