ASSIGNMENT\_2

6.1 (a) Start R and use these commands to load the data:

*> library(caret)*

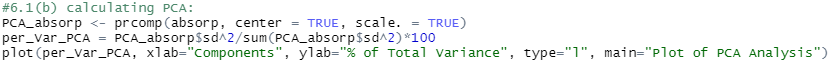
*> data(tecator)*

*> # use ?tecator to see more details*

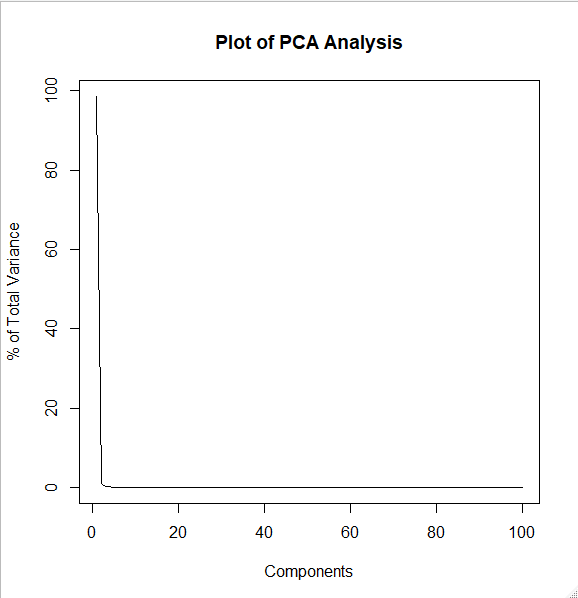
The matrix absorp contains the 100 absorbance values for the 215 samples, while matrix endpoints contains the percent of moisture, fat, and protein in columns 1–3, respectively.



6.1(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 (100). Use PCA to determine the effective dimension of these data. What is the effective dimension?

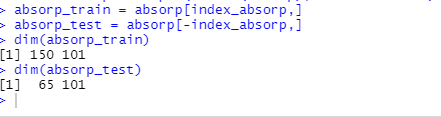


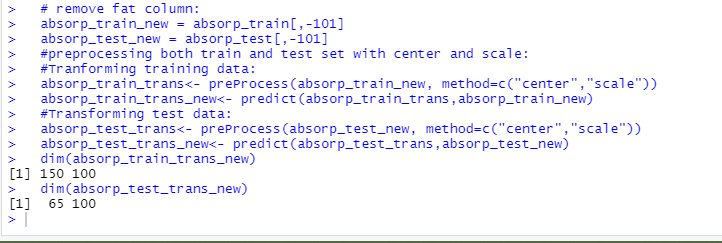




“The **first PCA Component** explains **97.2%** of variance. So, the **effective dimension** of data is **1**”.

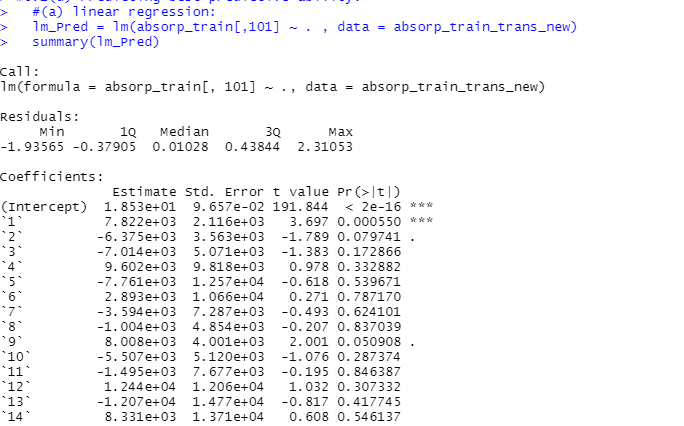
6.1(c) Split the data into a training and a test set, pre-process the data, and build each variety of models described in this chapter. For those models with tuning parameters, what are the optimal values of the tuning parameter(s)?

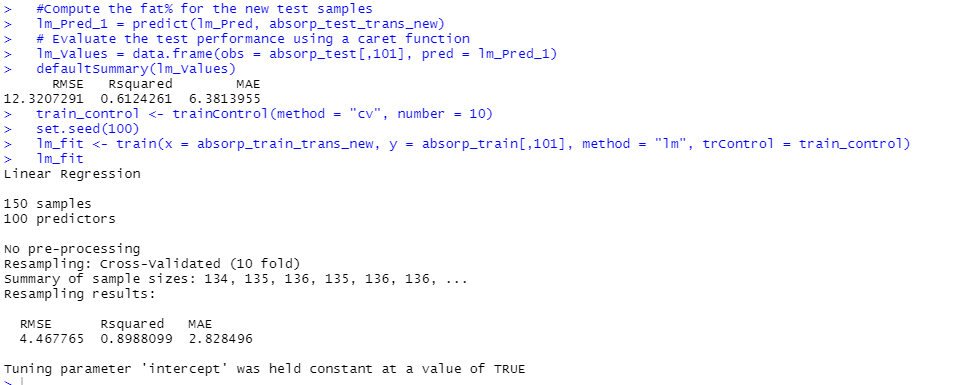




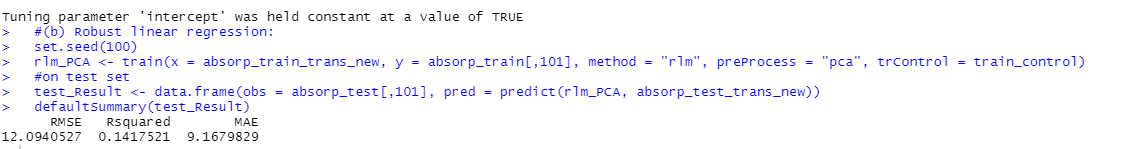
6.1(d) Which model has the best predictive ability? Is any model significantly better or worse than the others?

1. Linear Regression:

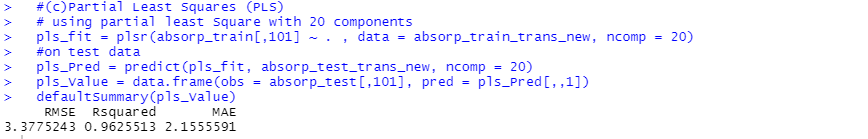




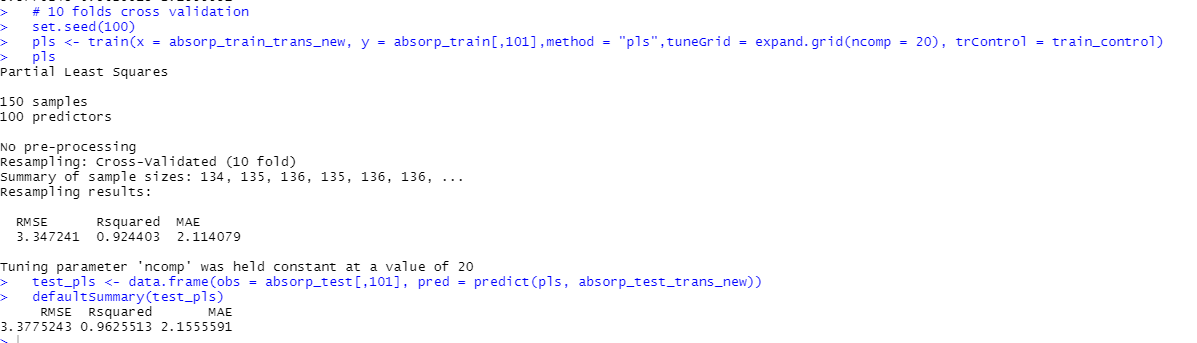
1. Robust Linear Regression:



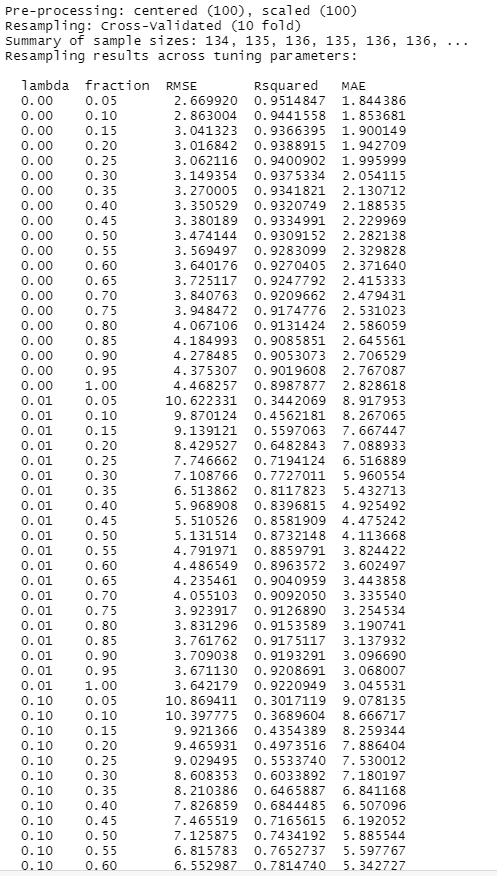
1. Partial Linear Regression:

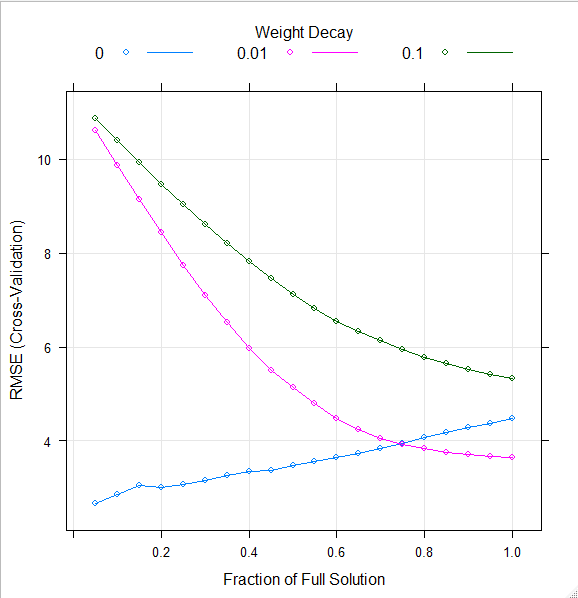


## 10 fold cross validation:



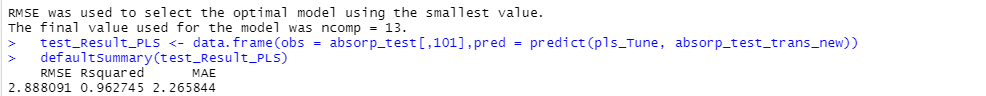
1. Penalised models:



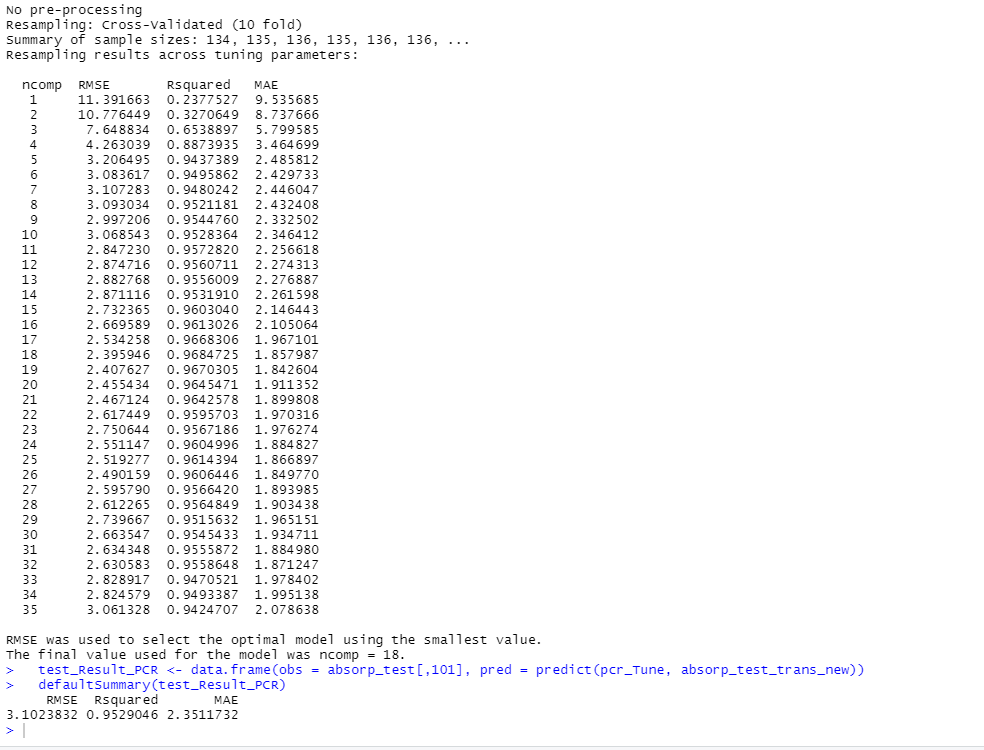


1. Model tuning:

PLS:



PCR:



|  |  |  |
| --- | --- | --- |
| Model | RMSE | R-Squared |
| Linear Regression | 4.46 | 0.898 |
| Robust linear Regression | 12.094 | 0.141 |
| PLR | 3.377 | 0.9625 |
| PLS | 3.347 | 0.924 |
| PLS Model tuning | 2.888 | 0.9627 |
| PCR model tuning | 3.102 | 0.9529 |

6.1(e) Explain which model you would use for predicting the fat content of a

sample.

* The model having R-squared value nearer to 1 has the best predictive ability.
* From the above table we can conclude that PLS model has the best predictive ability.
* All the models have similar R-squared values, but Robust Linear Model turns out to be significantly worst with a value of 0.141.

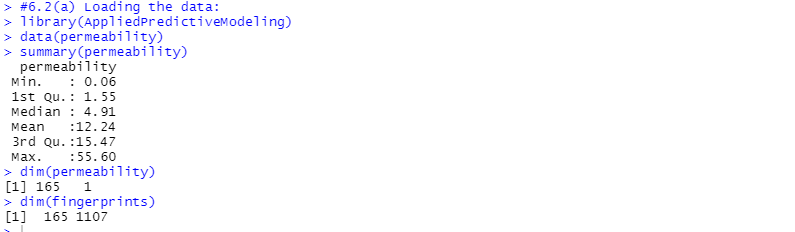
“I would prefer the **PLS Model** for predicting the fat content in the sample”.

6.2(a) Start R and use these commands to load the data:

*> library(AppliedPredictiveModeling)*

*> data(permeability)*

The matrix fingerprints contains the 1,107 binary molecular predictors for the 165 compounds, while permeability contains permeability response.



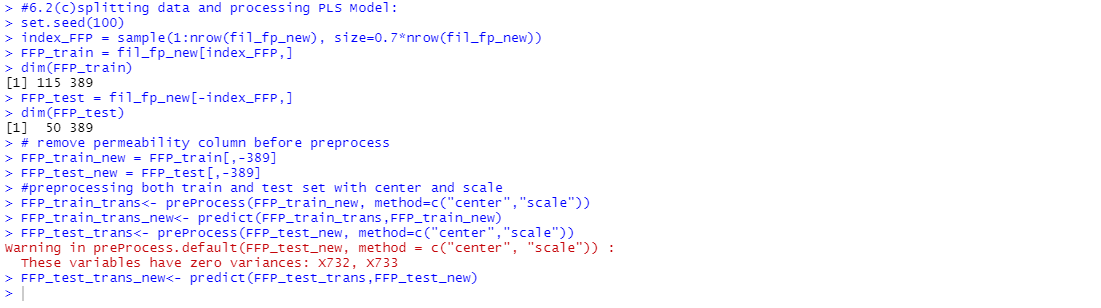
6.2(b) The fingerprint predictors indicate the presence or absence of substructures of a molecule and are often sparse meaning that relatively few of the molecules contain each substructure. Filter out the predictors that have low frequencies using the nearZeroVar function from the caret package. How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

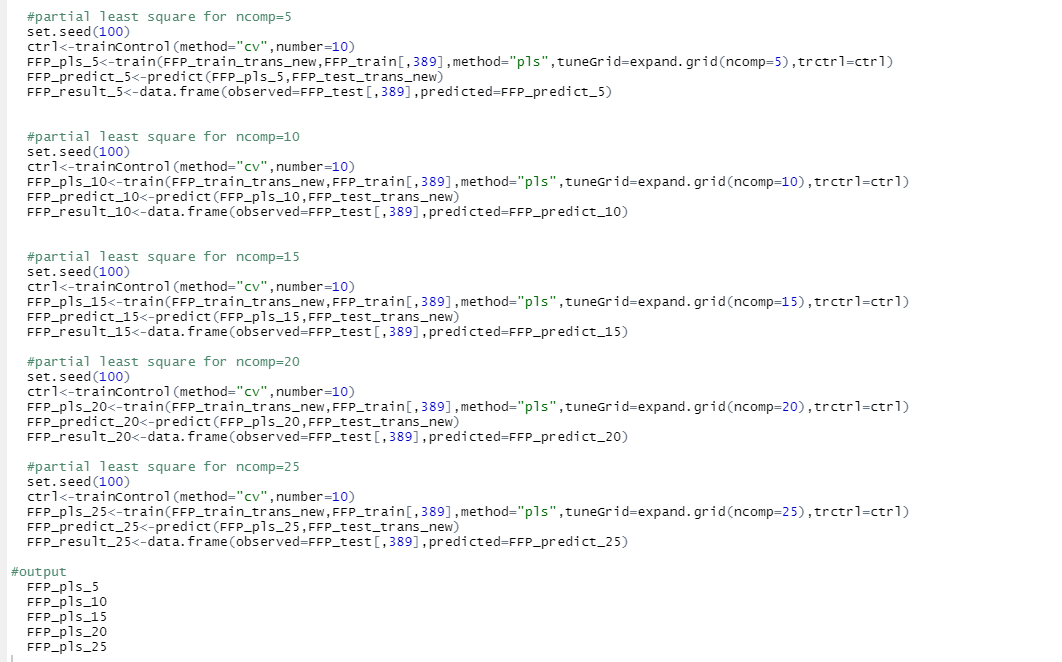


“The ‘nearZeroVar’ has reduced **number of predictors** to **388 from 1107**”.

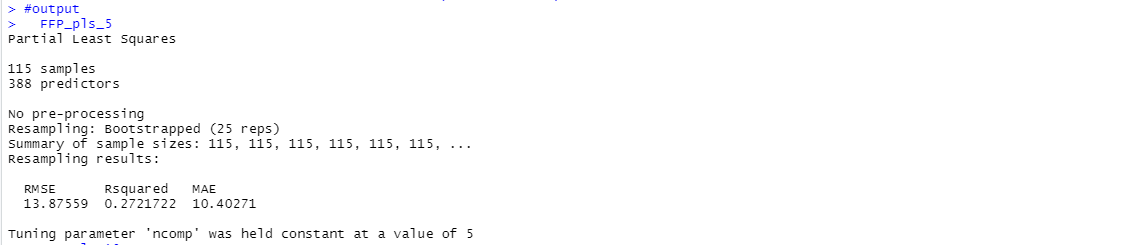
6.2(c) Split the data into a training and a test set, pre-process the data, and tune a PLS model. How many latent variables are optimal and what is the corresponding resampled estimate of *R*2?

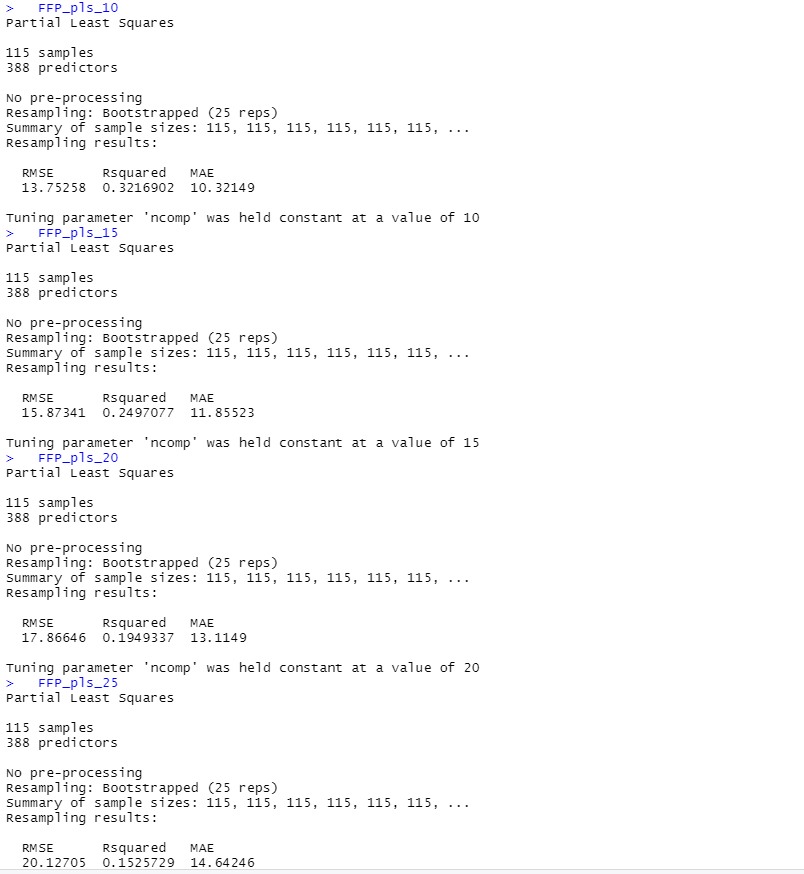
6.2(d)Predict the response for the test set. What is the test set estimate of *R*2?





OUTPUT:





|  |  |
| --- | --- |
| ncomp | R square |
| 5 | 0.2721 |
| 10 | 0.3216 |
| 15 | 0.2490 |
| 20 | 0.1949 |
| 25 | 0.1525 |

* I have used different ncomp values like 5,10,15,20,25 to determine the response on the test data.
* By comparing the all values of R square for different ncomp values, it is clear that ncomp=10 gives best result.

“**ncomp = 10** has the best result”