ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΠΜΣ «Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική»

Multicollinearity: diagnostics and PCA as a method of handling

Πολυσυγγραμικότητα: διαγνωστικοί έλεγχοι και η παλινδόμηση με κύριες συνιστώσες ώς μέθοδος χειρισμού της.

Ιωάννης Κουτσογιώργος

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

Απ. Μπατσίδης: Επίκουρος Καθηγητής Τμ. Μαθηματικών Παν/μιου Ιωαννίνων

Ι. Στεφανίδης: Καθηγητής Παθολογίας/Νεφρολογίας, Ιατρική Σχολή, Πανεπιστήμιο Θεσσαλίας

Χρ.Δοξάνη: Επιστημονικός Συνεργάτης στη Γενετική Φαρμακοεπιδημιολογία, Πανεπιστήμιο Θεσσαλίας

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΛΑΡΙΣΑ 2017

ABSTRACT

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression or logistic regression model are highly linearly related. The purpose of this diploma thesis is to find out two things. The first is to verify through some diagnostic tests that our data set is suffered from multicollinearity and the second one to perform principal component analysis to figure out if this phenomenon is eliminated. The results showed that not only multicollinearity of the data set was eliminated but also the predictive model after PCA was better than the one before PCA.

ΠΕΡΙΛΗΨΗ

Η πολυσυγραμικότητα αναφέρεται σε μία κατάσταση όπου δύο ή περισσότερες επεξηγηματικές μεταβλητές ενός πολλαπλού μοντέλου παλινδρόμησης ή ενός μοντέλου λογιστικής παλινδρόμησης έχουν υψηλή γραμμική συσχέτιση. Ο σκοπός αυτής της διπλωματικής εργασίας είναι να ερευνήσει δύο πράγματα. Το πρώτο είναι να επαληθεύσει μέσω κάποιων διαγνωστικών ελέγχων το γεγονός ότι το δικό μας σύνολο δεδομένων υποφέρει από πολυσυγγραμικότητα και το δεύτερο είναι να εφαρμόσει ανάλυση κύριων συνιστωσών για να ελέγξει άν αυτό το φαινόμενο έχει εξαλειφθεί. Τα αποτελέσματα έδειξαν ότι όχι μόνο εξαλείφθηκε η πολυσυγγραμικότητα στο σύνολο δεδομένων άλλα και το μοντέλο πρόβλεψης μετά την ανάλυση κύριων συνιστωσών ήταν καλύτερο από το μοντέλο πρίν την ανάλυση κύριων συνιστωσών.

1. Introduction

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model or logistic regression model are highly linearly related. From this problem our data set (Wisconsin Diagnostic Breast Cancer) was suffered, as some diagnostic tests verified it. The next step was to perform principal component analysis so to eliminate multicollinearity. The results showed that not only multicollinearity of the data set was eliminated but also the predictive model after PCA was better than the one before PCA.

The rest of this diploma thesis is organized as follows. In the second paragraph the meaning of multicollinearity was described as well as the five methods of multicollinearity detection.

In the third paragraph the ways for dealing with multicollinearity were presented, focusing on the principal component analysis. There were described analytical not only the meaning of PCA but also advantages and disadvantages of this technique.

In the last part of diploma thesis a statistical analysis (through R programming) of the chosen data set was performed presenting all outcome and results from that analysis.

2. Multicollinearity

2.1 Definition of Multicollinearity

According to reference [3] multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model or logistic regression model are highly linearly related. We have perfect multicollinearity if the correlation between two independent variables is equal to 1 or −1. In practice, we rarely face perfect multicollinearity in a data set. More commonly, the issue of multicollinearity arises when there is an approximate linear relationship among two or more independent variables.

2.2 Methods for multicollinearity detection

According to references [1], [2], [5], there are several methods to detect multicollinearity, which are:

- Variance Inflation Factor
- Tolerance
- Condition Number
- Condition Index
- Variance decomposition-proportion

2.2.1 Variance Inflation Factor (VIF)

Variance inflation factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient βk is “inflated” by the existence of correlation among the predictor variables in the model. A VIF of 1 means that there is no correlation among the kth predictor and the remaining predictor variables, and hence the variance of βk is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

2.2.2 Tolerance

Tolerance is a measure of collinearity. The variable’s tolerance is $1\text{-}Rj^2$ (where Rj^2 is the coefficient of determination of a regression of explanator j on all the other explanators) . A small tolerance value indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation and that it should not be added to the regression equation. All variables involved in the linear relationship will have a small tolerance. Some suggest that a

tolerance value less than 0.2 or 0.1 should be investigated further. If a low tolerance value is accompanied by large standard errors and non-significance, multicollinearity may be an issue.

2.2.3 Condition Number

Another measure of the overall multicollinearity of the variables can be obtained by computing the condition number (CN) of the correlation matrix, defined by the ratio of the largest Eigenvalue to smallest Eigenvalue (Chatterjee and Hadi, 2006). The condition number will always be greater than 1. A large condition number (larger than 15) indicates evidence of collinearity.

2.2.4 Condition Index

Most multivariate statistical approaches involve decomposing a correlation matrix into linear combinations of variables. The linear combinations are chosen so that the first combination has the largest possible variance (subject to some restrictions we won't discuss), the second combination has the next largest variance, subject to being uncorrelated with the first, the third has the largest possible variance, subject to being uncorrelated with the first and second, and so forth. The variance of each of these linear combinations is called an eigenvalue. Collinearity is spotted by finding 2 or more variables that have large proportions of variance (.50 or more) that correspond to large condition indices. A rule of thumb is to label as large those condition indices in the range of 30 or larger.

2.2.5 Variance decomposition-proportion

The variance-decomposition proportions (VD) are the variance proportions of the i-th variable attributable to the j-th eigenvalue. No variable should attribute more than 0.5 to any one eigenvalue (Dormann et al., 2012).

3. Principal Component Analysis (PCA) as a method of handling with multicollinearity

Depending on what the source of multicollinearity is, the solutions will vary. If the multicollinearity has been created by the data collection, collect additional data over a wider X-subspace. If the choice of the linear model has increased the multicollinearity, simplify the model by using variable selection techniques. If an observation or two has induced the multicollinearity, remove those observations. Above all, use care in selecting the variables at the outset. When these steps are not possible, you might try:

- ridge regression and
- principal component analysis (PCA).

In this diploma thesis our interest will be focused on PCA. PCA meaning, advantages and disadvantages are according to reference [4].

3.1 Principal Component Analysis (PCA)

PCA is a commonly used data reduction technique (Abdi and Williams 2010). This method seeks to find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance. The first PC is defined as the linear combination of the predictors that captures the most variability of all possible linear combinations. Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs. Mathematically, the jth PC can be written as:

$$PCj = (aj1 \text{ X Predictor 1}) + (aj2 \text{ X Predictor 2}) + \cdots + (ajP \text{ X Predictor P}).$$

P is the number of predictors. The coefficients aj1, aj2,. . . , ajP are called component weights and help us understand which predictors are most important
to each PC.

3.2 Advantages

The primary advantage of PCA, and the reason that it has retained its popularity as a data reduction method, is that it creates components that are uncorrelated. Some predictive models prefer predictors to be uncorrelated (or at least low correlation) in order to find solutions and to improve the model's numerical stability. PCA preprocessing creates new predictors with desirable characteristics for these kinds of models.

3.3 Disadvantages

PCA seeks predictor-set variation without regard to any further understanding of the predictors (i.e., measurement scales or distributions) or to knowledge of the modeling objectives (i.e., response variable). Hence, PCA can generate components that summarize characteristics of the data that are irrelevant to the underlying structure of the data and also to the ultimate modeling objective.

Because PCA seeks linear combinations of predictors that maximize variability, it will naturally first be drawn to summarizing predictors that have more variation. If the original predictors are on measurement scales that differ in orders of magnitude, then the first few components will focus on summarizing the higher magnitude predictors, while latter components will summarize lower variance predictors. This means that the PC weights will be larger for the higher variability predictors on the first few components. In addition, it means that PCA will be focusing its efforts on identifying the data structure based on measurement scales rather than based on the important relationships within the data for the current problem.

The second caveat of PCA is that it does not consider the modeling objective or response variable when summarizing variability. Because PCA is blind to the response, it is an *unsupervised technique*. If the predictive relationship between the predictors and response is not connected to the predictors' variability, then the derived PCs will not provide a suitable relationship with the response.

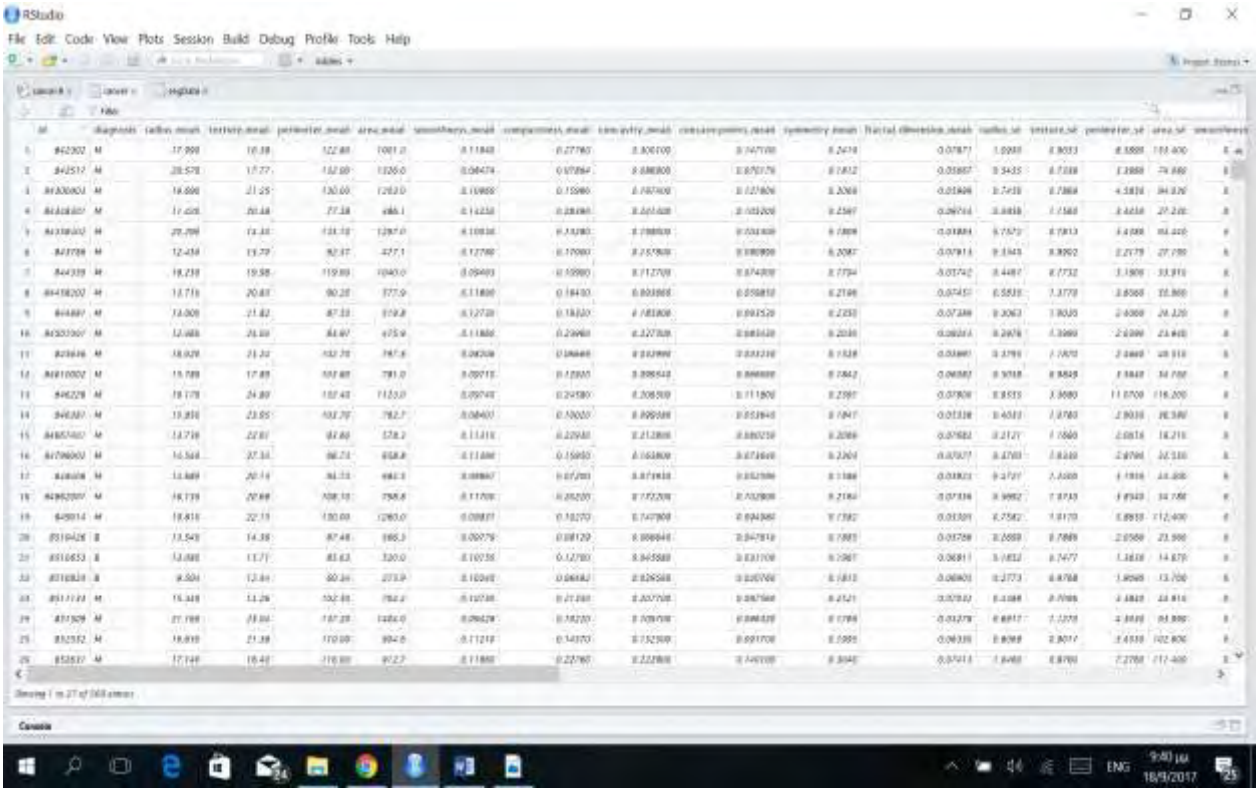
4. Practical application and results

The data set I chose is called Wisconsin Diagnostic Breast Cancer (WDBC) and it was downloaded from reference [6]. Except from attribute information which are ID number and Diagnosis (dependent variable) (M = malignant, B = benign) there are also ten real-valued features (independent variables) which are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

These features were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.
Also, the mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

This data set consists of 569 observations of 32 variables as shown below:



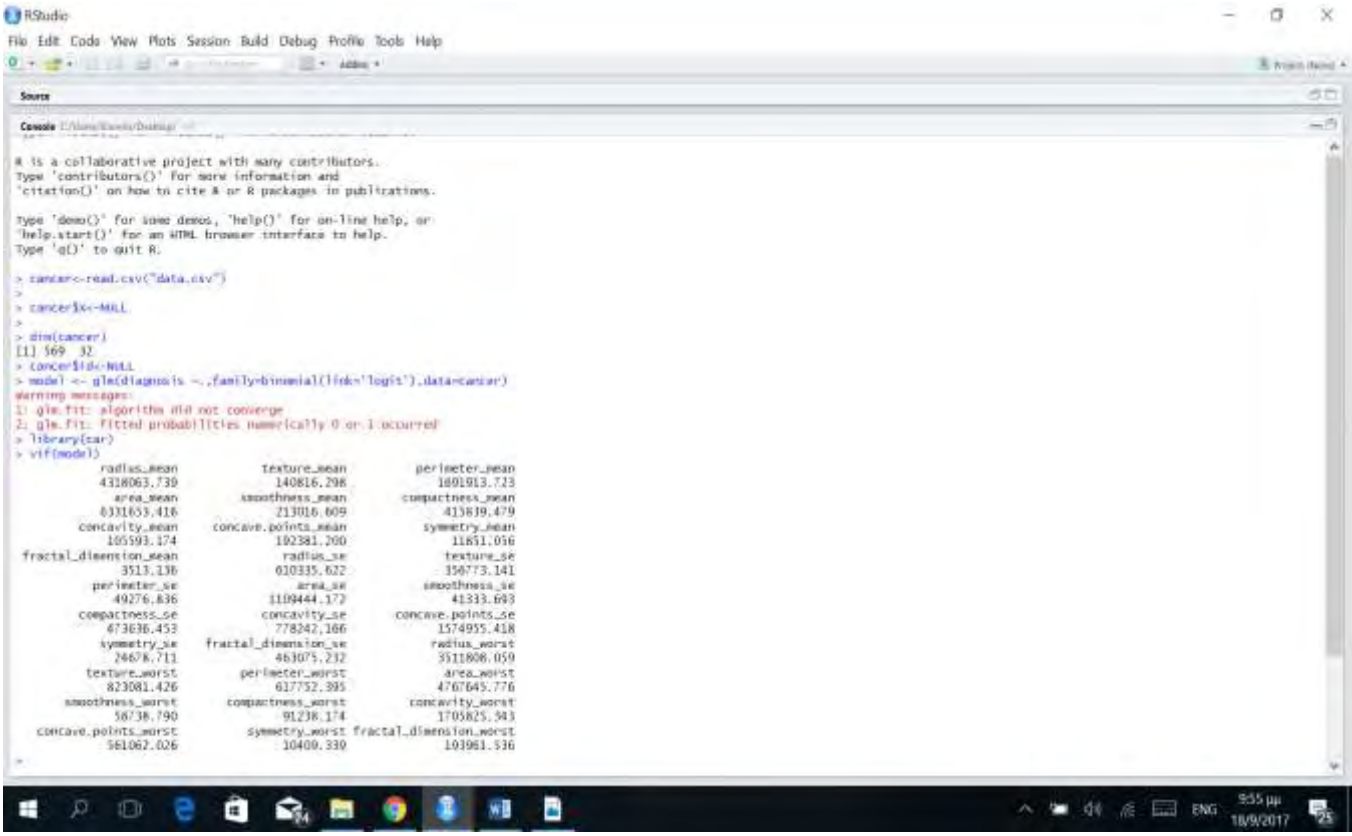
This study of the data set consists of these stages:

- Test for multicollinearity using diagnostic tests
- If multicollinearity appears, use PCA to eliminate phenomenon
- Test if the model, after PCA handling, has better predictive power than the model before it

4.1 Test for multicollinearity using diagnostic tests

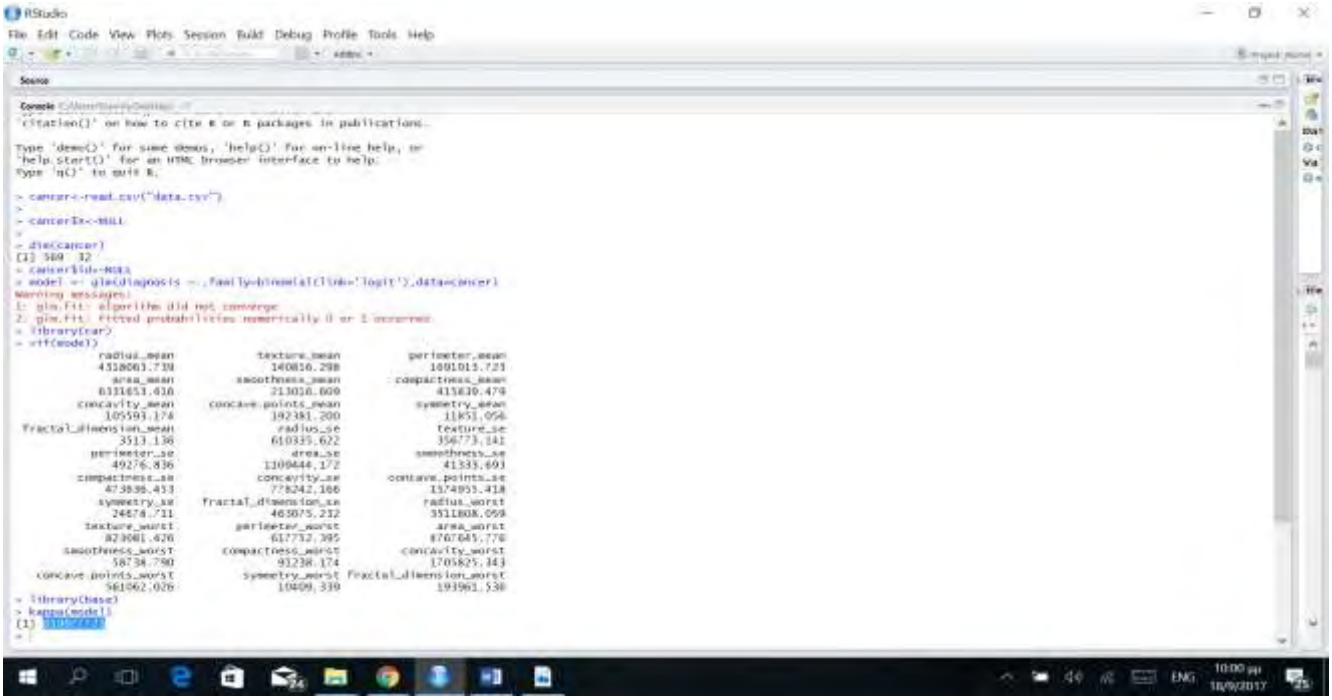
In this paragraph we will use the diagnostics in order to check if multicollinearity is present.

1. Variance Inflation Factor (VIF)



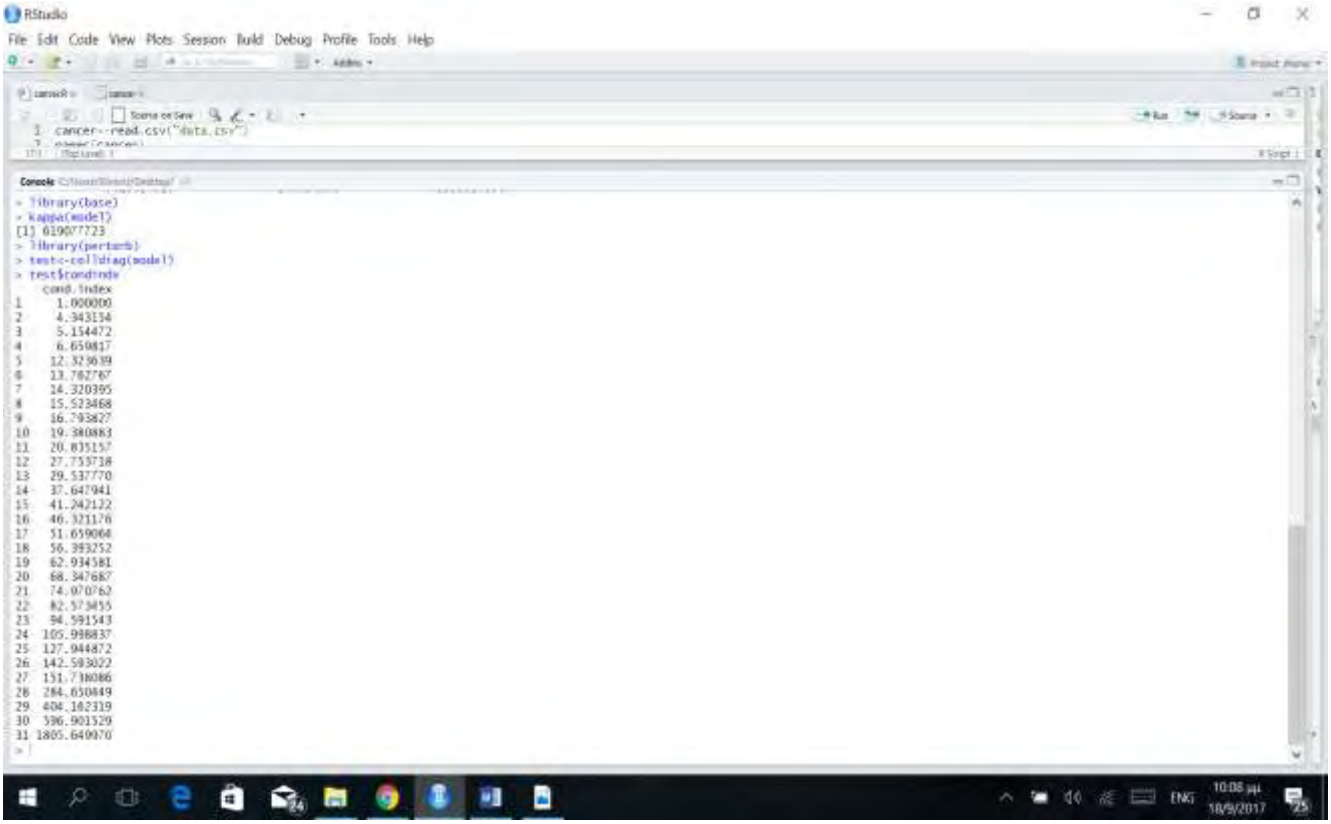
Variance inflation factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the predictors. As we see VIFs are exceeding the number 10 so it's a sign of serious multicollinearity.

2. Condition Number



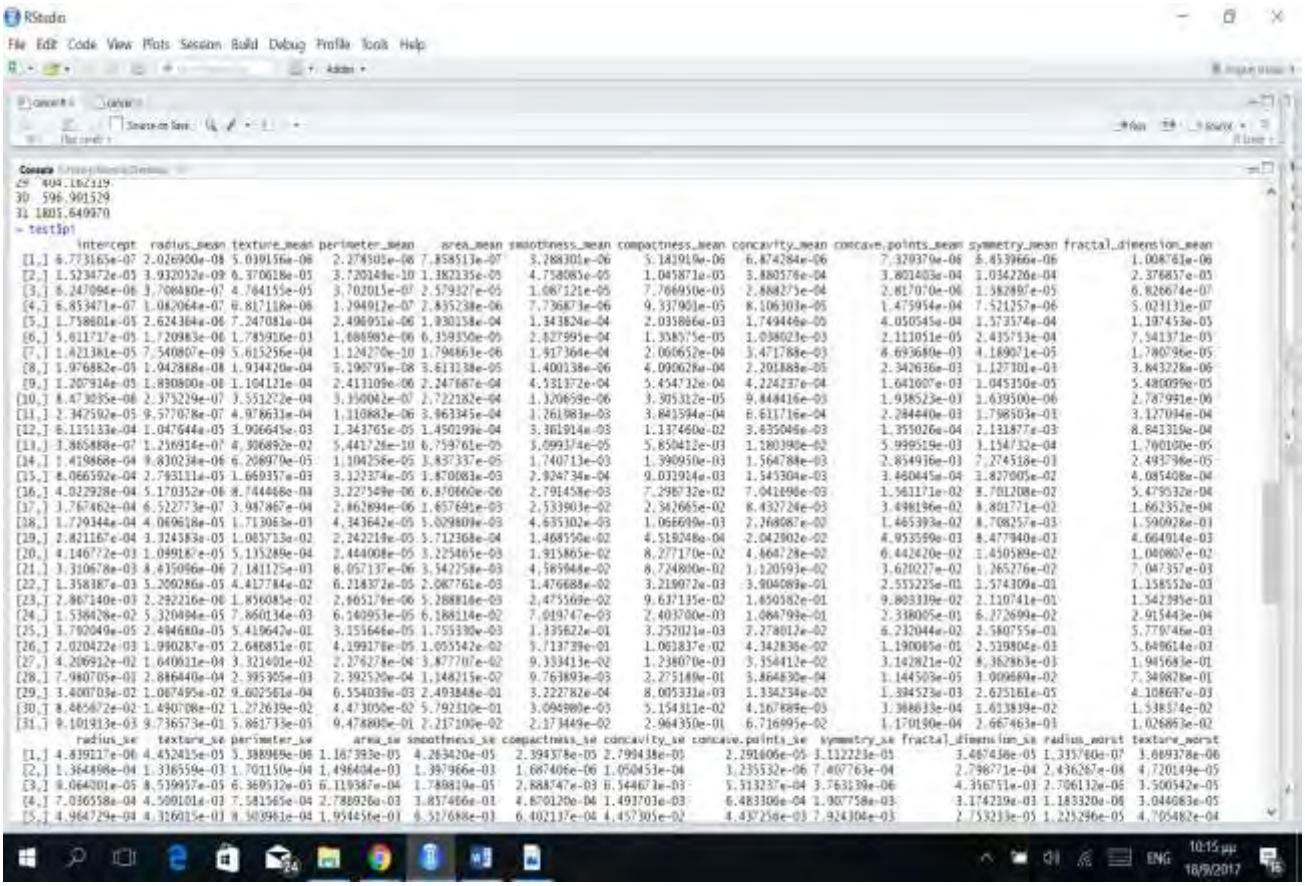
Condition's number value is 619077723 which is above 15. This difference in values indicates multicollinearity problem.

3. Condition Index



As we see there are values between 30-100 (14 through 23) which indicate moderate multicollinearity problem and there are values above 100 (24 through 31) which indicate serious multicollinearity problem.

4. Variance decomposition-proportion

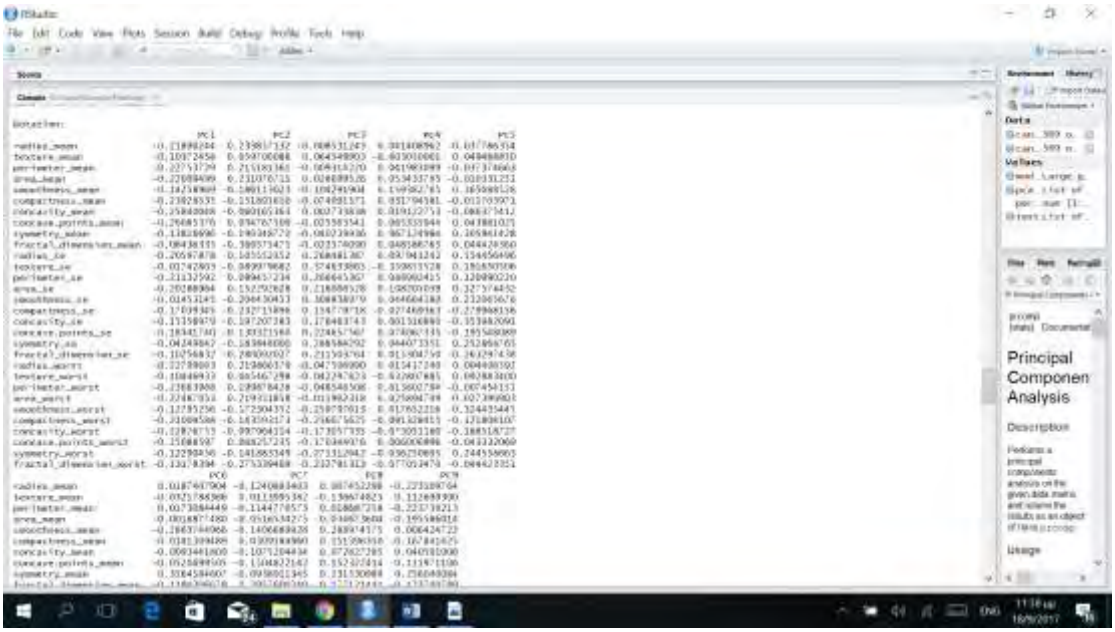


There are at least two regression coefficients with variance-decomposition proportion bigger than 50%. This fact leads us to multicollinearity problem.

5. Tolerance

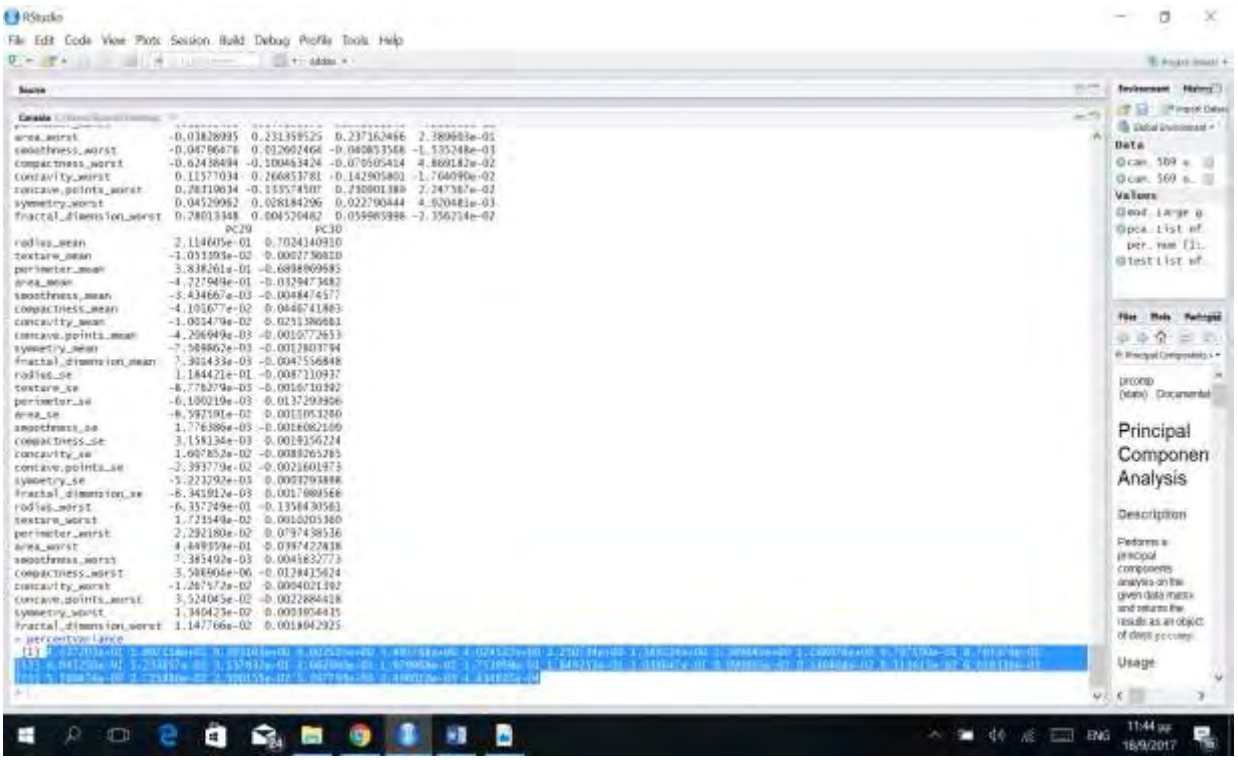
This diagnostic test is not applicable to specific data set as the dependent variable (diagnosis) is binary.

4.2 PCA as a method of handling multicollinearity



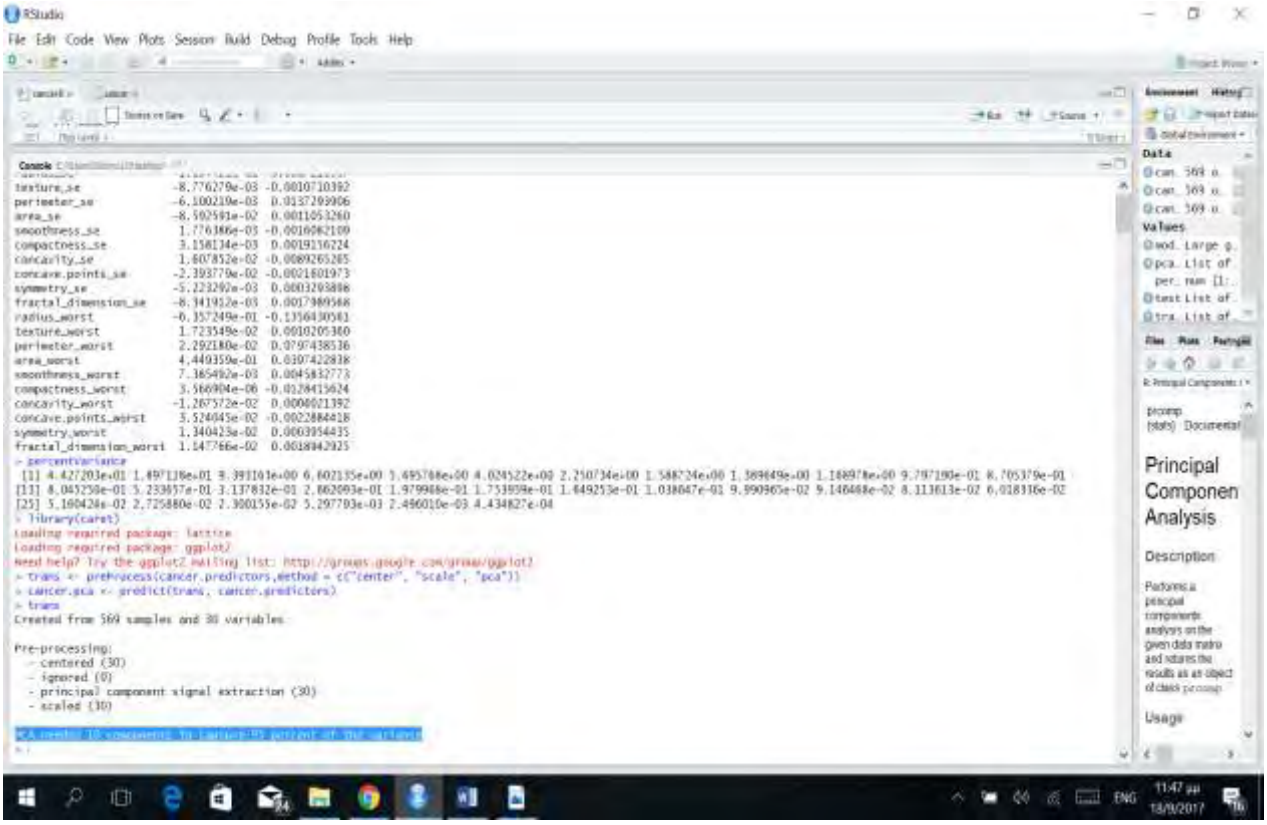
There are 30 PCs that are shown above. PC1 is defined as the linear combination of the predictors that captures the most variability of all possible linear combinations. Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs.

The percentage of variability that is captured by each PC is:

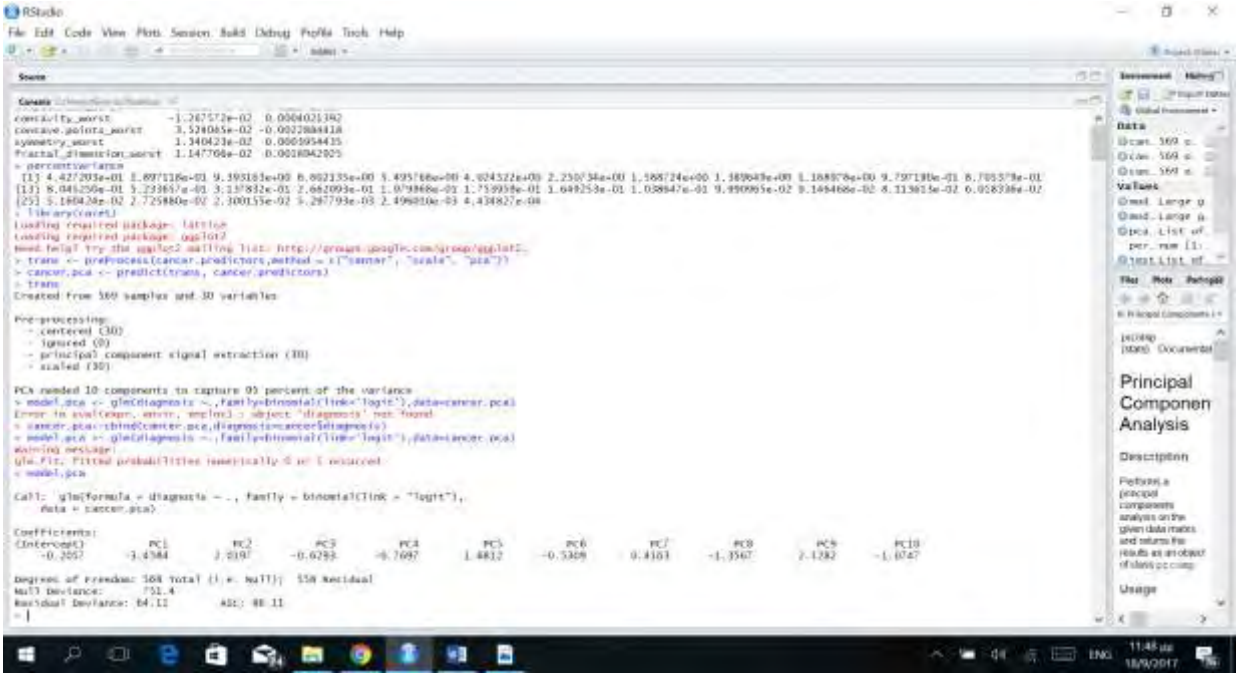


4.3 Test if the model, after PCA handling, has better predictive power than the model before it.

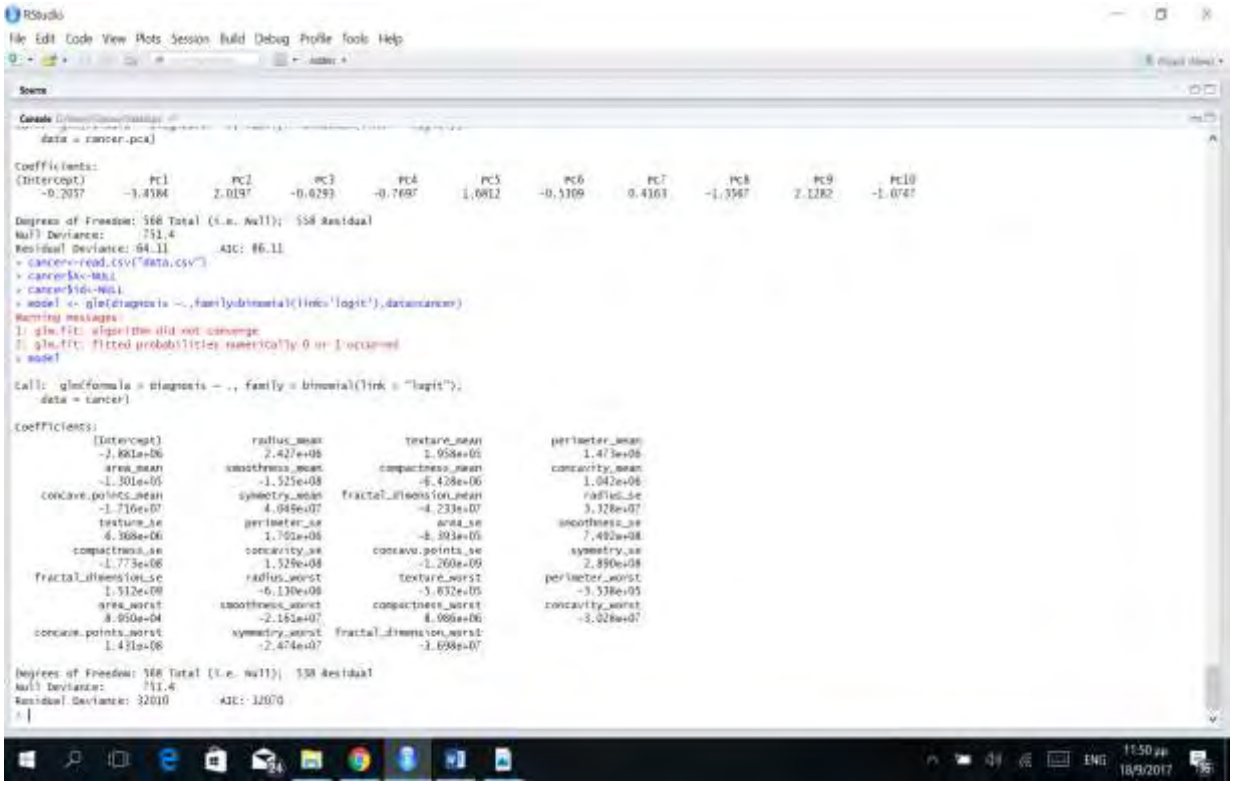
PCA needed 10 components to capture 95%of the variance.



Model after PCA



Model before PCA



In conclusion, we ended up with 2 results:

I. We eliminate multicollinearity of the data set, through principal component analysis (PCA)
II. The predictive model after PCA is better than the one before PCA. AIC metric indicates that. In the model after PCA AIC is 86.11 and in the other is 32070. And we know that smaller AIC number leads to better model.

5. R code

```
cancer<-read.csv("data.csv")

names(cancer)

cancer$X<-NULL

cancer$id<-NULL

model <- glm(diagnosis ~.,family=binomial(link='logit'),data=cancer)

#multicollinearity check with vif

library(car)

vif(model)

#multicollinearity check with Condition indexes and variance decomposition
proportions

library(perturb)

test<-colldiag(model)

#condition indexes

test$condindx

#variance decomposition proportions

test$pi

#multicollinearity check with condition number

library(base)

kappa(model)

cancer.predictors<-cancer[,-1]

pca.cancer <- prcomp(cancer.predictors,center = TRUE, scale. = TRUE)

percentVariance <- pca.cancer$sd^2/sum(pca.cancer$sd^2)*100

percentVariance

library(caret)

trans <- preProcess(cancer.predictors,method = c("center", "scale", "pca"))

cancer.pca <- predict(trans, cancer.predictors)

cancer.pca<-cbind(cancer.pca,diagnosis=cancer$diagnosis)

model.pca <- glm(diagnosis ~.,family=binomial(link='logit'),data=cancer.pca)
```

6. References

1. *Chatterjee and Hadi, 2006*
2. *Dormann et al. (2012)*
3. *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, "An Introduction to Statistical Learning" 2013.*
4. *Max Kuhn, Kjell Johnson, "Applied Predictive Modeling" 2013.*
5. *O'Brien(2007), Belsley (1991), Farrar, Glauber (1967), Wichers (1975), Kuman (1975) and O'Hagan,McCabe (1975), "Articles of Wikipedia"*
6. *Website "www.kaggle.com"*