Trinity College Dublin – ST3011

MVA - Assignment

Anti-GBM and Weather Data

Laetitia Lachat (17338341) & Marie Kuhn (17320598)

19.12.2017

Content

Content 1

I. INTRODUCTION 2

II. DATA ANALYSIS 2

a. DESCRIPTIVE STATISTICS 2

b. DATA REDUCTION BY PCA 2

c. CLUSTERING 3

Hierarchical clustering 3

K-means clustering 4

Compare clustering solutions 4

d. LOGISTIC REGRESSION 4

Model 1: PCs 5

Model 2: Complete Data 5

Model 3 and 4: Interactions 5

Comparison of Models 6

III. CONCLUSION 6

IV. REFERENCES 6

V. Appendix: 7

I. INTRODUCTION

We were given a data file made up of 4169 observations of 16 variables. The observations took place on consecutive days between 2003 and 2014. The second variable is a binary indicator listing whether a patient in Ireland was diagnosed with Anti-glomerular basement membrane disease (Anti-GBM disease) on the corresponding day. The other 14 variables are weather related variables measured daily at Dublin Airport. Our objective was to establish if there is a connection between environmental/weather data and the diagnosis indicator. A connection would indicate that the Anti-GBM disease is caused by environmental factors.

In the following sections, we first look at what the different variables (diagnosis indicator and weather data) tell us separately. We then perform a dimension reduction on the weather variables. By using different clustering methods, we try to establish a group structure within the data points for days on which a diagnosis occurred. Finally, we carry out a logistic regression to see if our binary indicator can be described by the principal components of the weather data found in the PCA.

II. DATA ANALYSIS

a. DESCRIPTIVE STATISTICS

We started by looking at the different variables separately.

Diagnosis: In general, we are interested in the conditions that led to the infection of the patient, so the conditions before or on the day the patient was infected. We don't know the incubation period or the time between the onset of symptoms and the patient being diagnosed. Even for the better investigated Japanese Kawasaki Disease, there is no precise incubation period known, so we have no base for a sensible assumption. We tried looking at the "waiting time", the number of days between two diagnoses (or rather: two indications of diagnoses). The mean of the waiting times is 50.6 days, the standard deviation is 51 days. **Plot 1** shows the waiting time in days (y-axis) for each of the 79 diagnoses (x-axis). Plotting the histogram of waiting times, we see that they are roughly exp(1/λ) distributed with λ = 50.6 (**Plot 2**). If we assumed the diagnoses independent we could model a Poisson Process for our data, but independence is certainly not what we would want to assume in our analysis.

Weather: A general description of weather variables can be found in **Table 1**. We started by checking the weather variables for correlations since many of the variables measure the same thing (e.g. max and min temperature). The Correlation Plot (**Plot 3**) shows which variables are highly correlated. Having so many highly correlated explanatory variables, we decided to do a PCA to reduce the dimensionality of the data set and the correlation between explanatory variables.

b. DATA REDUCTION BY PCA

For the moment, our data points are given in 14 highly correlated weather variables, so it is difficult to see something pertinent on our plots. By applying a Principal Component Analysis, we extract the internal structure of our data, and are then able to express our data in a reduced number of dimensions. Since we have no reason to prefer one variable over the others and the covariance matrix shows big differences between the variances, we will work with standardized data for the PCA (so we use the correlation matrix). The plot of Eigenvalues (**Plot 4**) suggests considering 3 PCs to describe our data. Those 3 PCs cover almost 73 % of variation, which is enough for our purpose (**R Output 1**).

Analysis of the loadings of the chosen PCs:

- PC1: Mainly influenced by the Temperature variables “Maximum Air Temperature”, “Potential Evapotranspiration”, “Soil Moisture Deficits”, “Moderately drained SMD”, “Poorly drained SMD” and “Well drained SMD”, which we already found to be highly correlated.
- PC2: Data points which have a high value on this PC will have a low value on “Highest Gust”, “Highest Ten Minute Mean Wind Speed” and “Mean Wind Speed”, which means that PC2 is mainly influenced by the Wind Conditions.
- PC3: High positive loadings on “Minimum Air Temperature” and “Precipitation Amount” and a high negative loading on “Sunshine Duration”. (**R Output 2**)

The next step is to express our data in our new variables (our three principal components). From now on, we will only use the data expressed on PC1, PC2 and PC3. **Plot 5** shows our data points projected onto the new variables PC1 and PC2. Days, on which a diagnosis occurred are represented in black, days without a diagnosis in grey. Since there are a lot more days without diagnosis (4090) compared to days with a diagnosis (79), we will now concentrate on the days on which a diagnosis occurred (diagnosis == 1).

c. CLUSTERING

Now, we try to establish if there are different groups to be found within the new data points for days with a diagnosis.

Hierarchical clustering

We chose to first apply a hierarchical clustering on the data points for which the diagnosis == 1. We applied Hierarchical Clustering using a Euclidean dissimilarity measure (since our data is numerical) and Complete Linkage to obtain spherical clusters with a good internal structure.

Usually, Complete Linkage does not display outliers as obviously as Single Linkage would do. However, according to the cluster dendrogram (**Plot 6**), we can see multiple outliers in our data (in this case, the observations 55 and 69). It could be sensible to exclude those points, because they are too odd compared to the other data points. Looking at their values on the weather variables, we see that they differ from the mean values (of all observations) by more than two standard deviations in multiple variables. D55 has a much higher value on “Precipitation Amount. D69 has large deviations in the “Soil Moisture Deficit” variables 11 to 14 and in “Sunshine Duration”, where it has much higher values than the mean (**Plot 7**).

In this case, by observing the dendrogram we can estimate that there are 3 groups. Moreover, if we cut the dendrogram at height + 3*standard deviation (marked by the red line in the dendrogram), we also find 3 groups. However, we can remark that there is only one point in the third group with this method. To confirm our results, we are going to apply k-means clustering with different values of k (especially for k=2 and k=3).

K-means clustering

First, we are going to compute the total within sum of squares values versus k. We are looking for a trade-off between small complexity (which increases with k) and a small WSS (which always decreases with larger k). According the **Plot 8**, we should probably prefer this iterative clustering for k=3.

For k = 3, we had 27 observations in the first group, 21 for the second one and 31 for the third group. The graphical picture of the iterative clustering is shown in **Plot 9**. We can see that our three groups seem consistent. Indeed, our clusters are very compact: the distance between cluster centroids (around 3.2 to 5) correspond to the double of the average distance from cluster centroid (around 1.3 to 2.2) (**R Output 3**).

Looking at the coordinates of the cluster centroids (**R Output 4**), we could conclude that we have the following three groups within our data points for days with diagnoses:

- Group 1: Days with large positive values on PC1, small values on PC2 and PC3
  (= high Temperature, Evapotranspiration and SMD)
- Group 2: Days with large negative values on PC2, negative values on PC1 and positive values on PC3
  (= high Wind Speed and high Precipitation)
- Group 3: Days with large negative values on PC1, large positive values on PC2 and small values on PC3
  (= high Wind Speed and low Precipitation)

Compare clustering solutions

We wish to know if our results are consistent. That is why we are going to compare the hierarchical clustering solution with the iterative k-means clustering solution. In both clustering solutions, we have excluded the observations 69 and 55 (previously considered as outliers). According our **R Output 5**, the first two groups seem to be well fitted. However, for the third group, points are not distributed in the same way.

The best way to measure the agreement between two clustering methods is to compute the adjusted rand index (-1 ≤ ARI ≤ 1). Here, the ARI is equal to 0.633. Since it is close to 1, we can conclude that hierarchical and k-means clustering gave us approximately the same clusters.

d. LOGISTIC REGRESSION

Now, we would like to study the relationships between variables. That is why we will try four different models for a logistic regression:

Model 1 (PCs 1): $logit\big(P(Diagnosis = 0)\big) = \alpha + \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3$ (**R Output 6**)

Model 2 (Complete Data 1): $logit\big(P(Diagnosis = 0)\big) = \alpha + \sum_{i=1}^{14} \beta_i W_i$ (**R Output 7**)

Model 3 (PCs 2): Model 1 with Interactions (**R Output 8**)

Model 4 (Complete Data 2): Model 2 with Interactions

We want to express the probability of a having a diagnosis on a certain day P(Diagnosis = 1) as a function of the values of the weather variables on this day: P(Diagnosis = 1) = 1 – P(Diagnosis = 0). In the first model, we want to use the “compressed” variables PC1, PC2 and PC3, which represent the weather variables in a much lower dimensionality. In the second model, we will fit the regression model to the complete weather data consisting of 14 variables. Model 3 and Model 4 are the first two models with interactions.

Model 1: PCs

According the **R Output 6**, we can see that for the parameter α, Pr(>|z|) is lower than 0.05. So we can reject the null hypothesis and conclude that α is a significant parameter to determine if the diagnosis is positive or not. However, for the other parameters (β_1, β_2, β_3), Pr(>[z|) is bigger than 0.05. Thereby we can't reject the null hypothesis, and it would be preferable to consider a simpler model in which those parameters are not included.

Model 2: Complete Data

With the **R Output 7**, 15 parameters are examined. We can see that for the parameter β_{13} and β_{14}, Pr(>|z|) are lower than 0.05 (0.0364 for β13; 0.0107 for β14). We can reject the null hypothesis for those parameters and conclude that β_{13} and β_{14} are meaningful parameters to determine the result of a diagnosis. However, for all the other parameters, Pr(>[z|) is bigger than 0.05. Thereby we can't reject the null hypothesis: it would be better to consider a simpler model in which those parameters are not included.

The parameter 13 represents the poorly drained soil moisture deficits (expressed in mm). The estimate for β13 is negative, so some low poorly drained soil moistures deficits will have a high probability to be assigned to the first category (i.e. a positive diagnosis).

The parameter 14 is the well-drained mean soil temperature (expressed in degrees Celsius). B14 is estimated as negative, thereby a low well drained mean soil temperature has a bigger probability to have a positive diagnosis.

Model 3 and 4: Interactions

By looking the **R Output 8**, we can see that the Pr(>|z|) for α is lower than 0.05. Thereby we don't reject the null hypothesis and we can conclude that α is a significant parameter to determine the result of a diagnosis. Otherwise, all the other parameters have a Pr(>|z|) upper to 0.05. Thereby, the null hypothesis is not rejected and it would be wiser to consider a model where those parameters are not integrated.

For the logistic regression of model 4, we didn't put the results in our report (the summary of the logistic regression is far too big). We can see that Pr(>|z|) is lower than 0.05 for the parameters β_{4x14}, β_{6x12} and β_{8x16}. For those interactions, we reject the null hypothesis and thereby, we can conclude there are significant parameters to determine if a diagnosis is positive or negative. For all the other interactions and parameters, Pr(>|z|) is bigger than 0.05 so it would be preferable to consider a simpler model in which those parameters are not integrated.

The estimate of β_{4x14}, β_{6x12} and β_{8x16} are positive. Thereby the wind direction min mean (expressed in deg) and the well drained mean soil temperature (expressed in degrees Celsius) influence a positive diagnosis. It's the same for the maximum air temperature (in degrees Celsius) and a moderately drained soil moisture deficits (in mm); for the minimum air temperature (in degrees Celsius) and the moderately drained soil moisture deficits (in mm).

Comparison of Models

	AIC	Residual Deviance
Model 1	789.1	783.12 (df = 4165)
Model 2	795.95	765.95 (df = 4154)
Model 3	790.7	776.7 (df = 4162)
Model 4	891.52	679.52 (df = 4063)

According the table, we can see that between the models 1 and 2, the value of AIC and Residual deviance are not very different (around 790 for AIC, around 775 for Residual Deviance).

We can remark that AIC for the model 3 is bigger than for the model 1. Moreover, we didn't find an additional and significant parameter with the model 3, compared to the first model. In conclusion, adding interactions to the first model didn't increase model's predictive ability.

For the model 4, we can conclude that it is not appropriate, compared to the other models. Indeed, its AIC is much bigger than the AIC for the other models (1, 2 and 3): the value is equal to 891.52 (offset of 100 with the AIC of other models). This is confirmed by the results of residual deviance: for the model 4, the value is equal to 679.52, a small value compared to the other models. Thereby, there is a significant difference between the full model and the model 4.

In conclusion, we can say that the model 1 and model 2 are the best proposed models.

III. CONCLUSION

From our research, we can conclude that the variation of our data is mainly influenced by the temperature variables and the wind conditions. Two kinds of days could involve the infection of a person: a day with a high temperature (involving a high soil temperature) and a day with high wind speeds.

As a further analysis, we could carry out a factor analysis and a multidimensional scaling to express our data reduction in different ways. Indeed, it would be useful to find a better model for the logistic regression analysis. Moreover, we could compare clustering methods from different data reduction techniques.

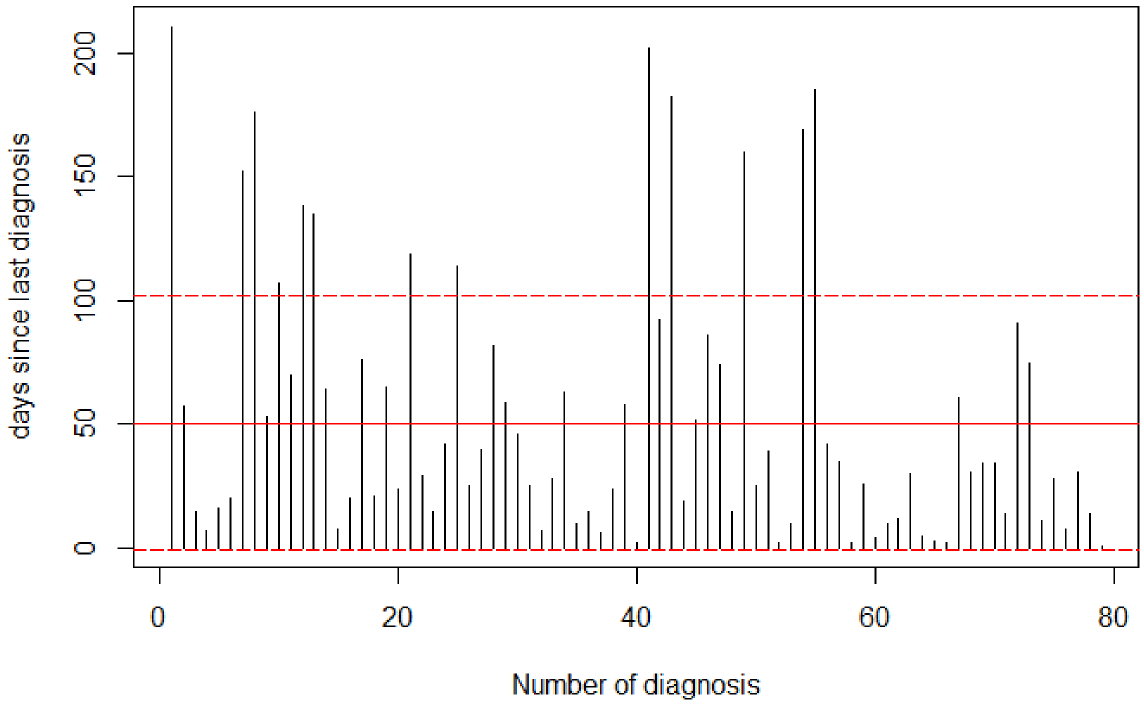
Furthermore, in our research, we assumed that the period incubation is about half-day (as the article suggests it) and that the patient directly saw a doctor (whereas usually a patient waits for a few days to schedule a consultation). It could be useful to try to shift the diagnosis and the other variables to find the incubation period.

IV. REFERENCES

- HOULDING, Brett. *Multivariate Analysis*. Dublin: Trinity College Dublin, School of Computer Science and Statistics. Publications TCD Junior Sophister, Course, 2017.
- WOGAN, Tim. Mysterious illness may be carried by the wind. *Science* [online]. 2014. Available on: http://www.sciencemag.org/news/2014/05/mysterious-illness-may-be-carried-wind (Visited on 17/12/2017).

V. Appendix:

Plot 1 – Days since last diagnosis for the 79 diagnosis indications



Plot 2 – Histogram of waiting times

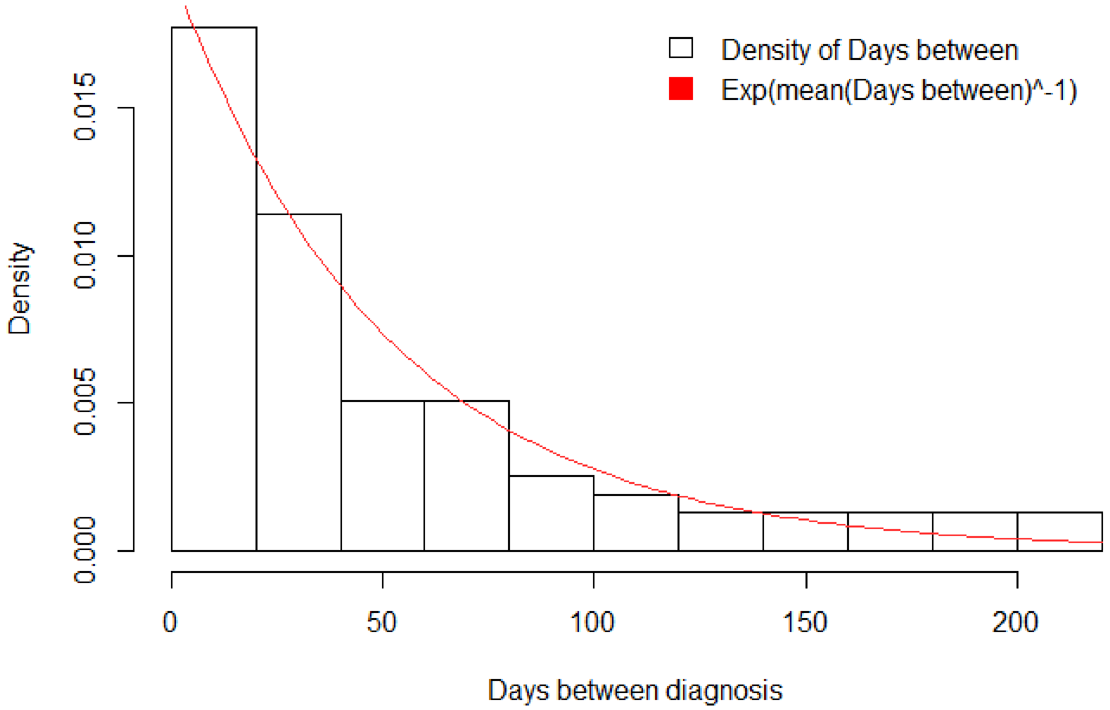


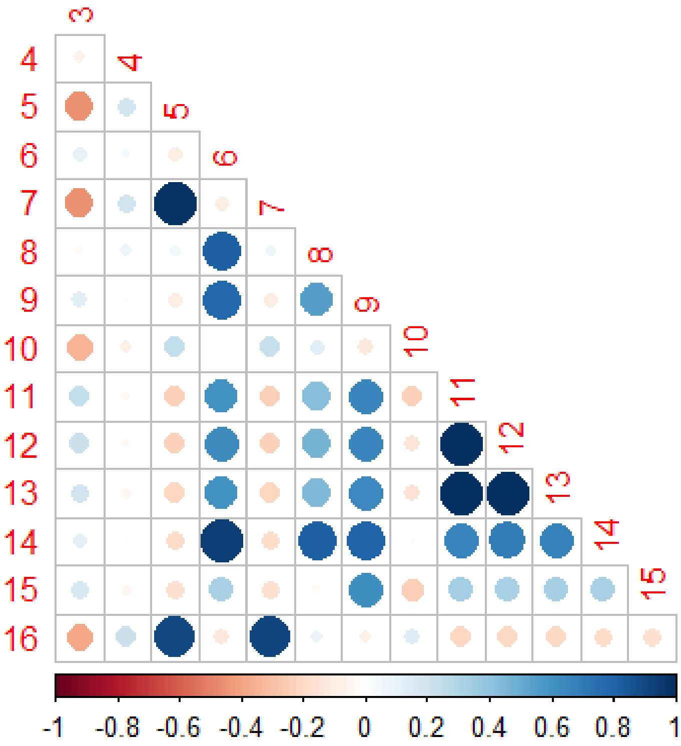
Table 1

	Title	Description	Measured in
1	Date	Date	DD-MM-YY
2	Diagnosis	Indicator whether or not a patient in Ireland was diagnosed with Anti-GBM disease	1 = YES 0 = NO
3	Mean CBL Pressure (hpa)	Mean atmospheric pressure[1]	Hectopascal hpa
4	Wind Direction at max 10 min mean (deg)	Wind direction during the 10 min interval in which the wind was the strongest[2]	North: around 0 deg East: around 90 deg South: around 180 deg West: around 270 deg
5	Highest Gust (kt)	Highest gust	Knots kt
6	Maximum Air Temperature (C)	Maximum air temperature	degrees celsius C
7	Highest ten minute mean wind speed (kt)	Highest ten minute mean wind speed	Knots kt
8	Minimum Air Temperature (C)	Minimum air temperature	Degrees celsius C
9	Potential Evapotranspiration (mm)	Amount of water that would be evaporated and transpired if there were sufficient water available[3]	Millimetres mm
10	Precipitation Amount (mm)	Precipitation amount	Millimetres mm
11	Soil Moisture Deficits (mm)	Amount of rain needed to bring the soil moisture content back to field capacity (the amount of water the soil can hold against gravity)[4]	Millimetres mm
12	moderately drained Soil Moisture Deficits (mm)	SMD for moderately drained soil (saturates on wet winter days, returns to field capacity on first dry day) [5]	Millimetres mm
13	poorly drained Soil Moisture Deficits (mm)	SMD for poorly drained soil (saturates on wet winter days,	Millimetres mm

[1] https://www.met.ie/climate-ireland/pressure.asp
[2] http://snowfence.umn.edu/Components/winddirectionanddegreeswithouttable3.htm
[3] https://en.wikipedia.org/wiki/Evapotranspiration#Potential_evapotranspiration
[4] http://www.met.ie/climate/agri-meteo-data.asp
[5] http://www.met.ie/climate/agri-meteo-data.asp

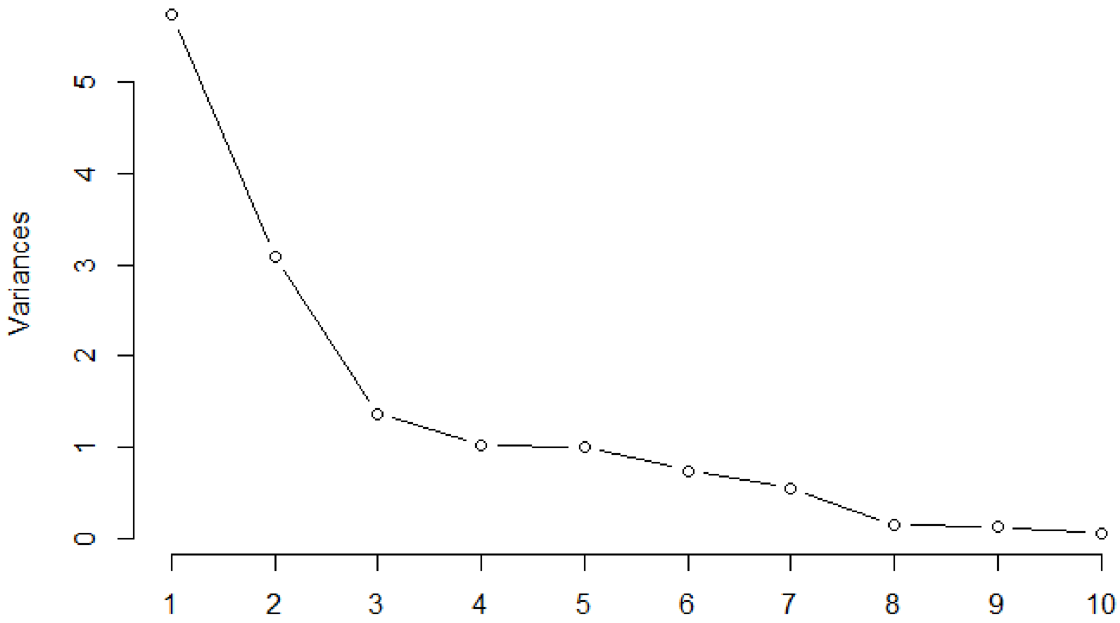
		water surplus is drained at very slow rate) [6]	
14	well drained Mean 10cm soil temperature (C)	Mean 10cm soil temperature for well drained soil (never saturates)	Degrees Celsius C
15	Sunshine duration (hours)	Sunshine duration	Hours h
16	Mean Wind Speed (kt)	Mean wind speed	Knots kt

Plot 3 – Correlation Plot for weather variables

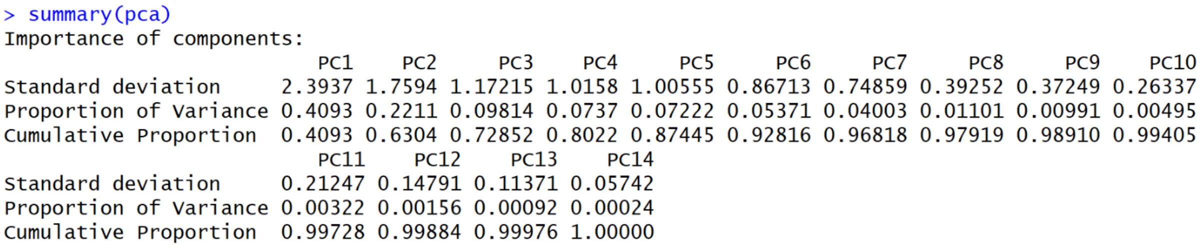


[6] http://www.met.ie/climate/agri-meteo-data.asp

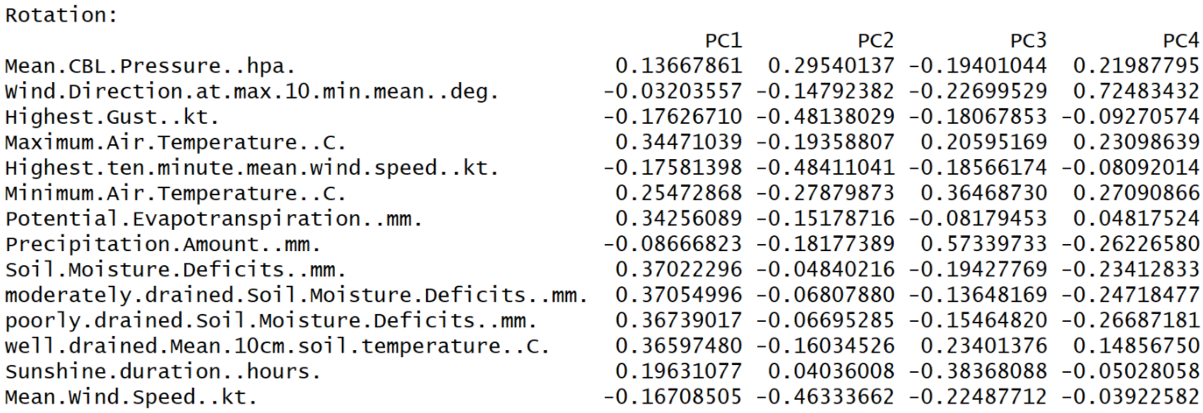
Plot 4: PCA Eigenvalues for weather variables



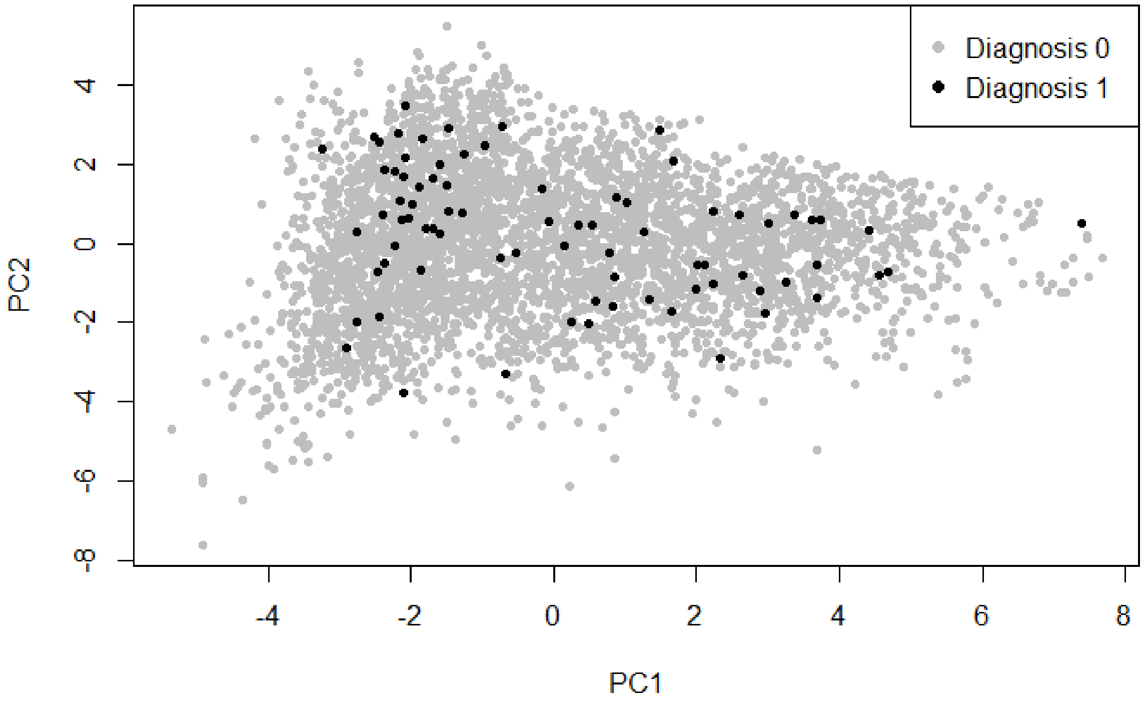
R Output 1: PCA Summary Weather Variables


```
> summary(pca)
Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.3937 1.7594 1.17215 1.0158 1.00555 0.86713 0.74859 0.39252 0.37249 0.26337
Proportion of Variance 0.4093 0.2211 0.09814 0.0737 0.07222 0.05371 0.04003 0.01101 0.00991 0.00495
Cumulative Proportion  0.4093 0.6304 0.72852 0.8022 0.87445 0.92816 0.96818 0.97919 0.98910 0.99405
                          PC11    PC12    PC13    PC14
Standard deviation     0.21247 0.14791 0.11371 0.05742
Proportion of Variance 0.00322 0.00156 0.00092 0.00024
Cumulative Proportion  0.99728 0.99884 0.99976 1.00000
```

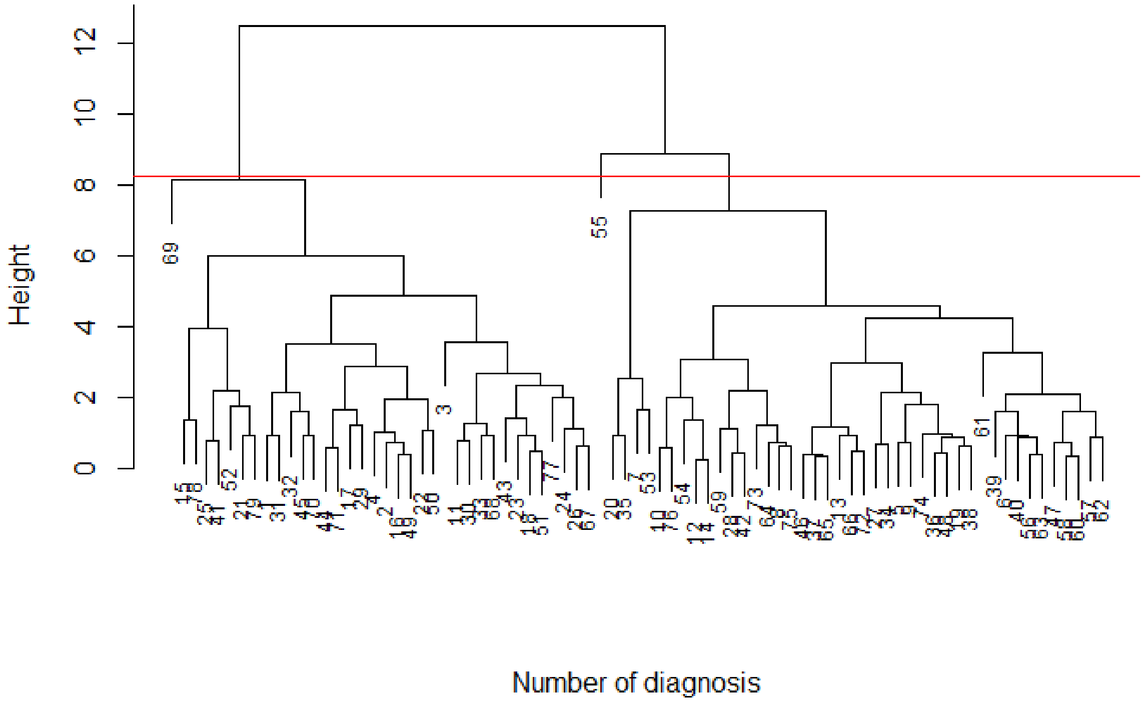
R Output 2: Coordinates of PCs 1 to 4


```
Rotation:
                                                      PC1         PC2         PC3         PC4
Mean.CBL.Pressure..hpa.                        0.13667861  0.29540137 -0.19401044  0.21987795
Wind.Direction.at.max.10.min.mean..deg.       -0.03203557 -0.14792382 -0.22699529  0.72483432
Highest.Gust..kt.                             -0.17626710 -0.48138029 -0.18067853 -0.09270574
Maximum.Air.Temperature..C.                    0.34471039 -0.19358807  0.20595169  0.23098639
Highest.ten.minute.mean.wind.speed..kt.       -0.17581398 -0.48411041 -0.18566174 -0.08092014
Minimum.Air.Temperature..C.                    0.25472868 -0.27879873  0.36468730  0.27090866
Potential.Evapotranspiration..mm.              0.34256089 -0.15178716 -0.08179453  0.04817524
Precipitation.Amount..mm.                     -0.08666823 -0.18177389  0.57339733 -0.26226580
Soil.Moisture.Deficits..mm.                    0.37022296 -0.04840216 -0.19427769 -0.23412833
moderately.drained.Soil.Moisture.Deficits..mm. 0.37054996 -0.06807880 -0.13648169 -0.24718477
poorly.drained.Soil.Moisture.Deficits..mm.     0.36739017 -0.06695285 -0.15464820 -0.26687181
well.drained.Mean.10cm.soil.temperature..C.    0.36597480 -0.16034526  0.23401376  0.14856750
Sunshine.duration..hours.                      0.19631077  0.04036008 -0.38368088 -0.05028058
Mean.Wind.Speed..kt.                          -0.16708505 -0.46333662 -0.22487712 -0.03922582
```

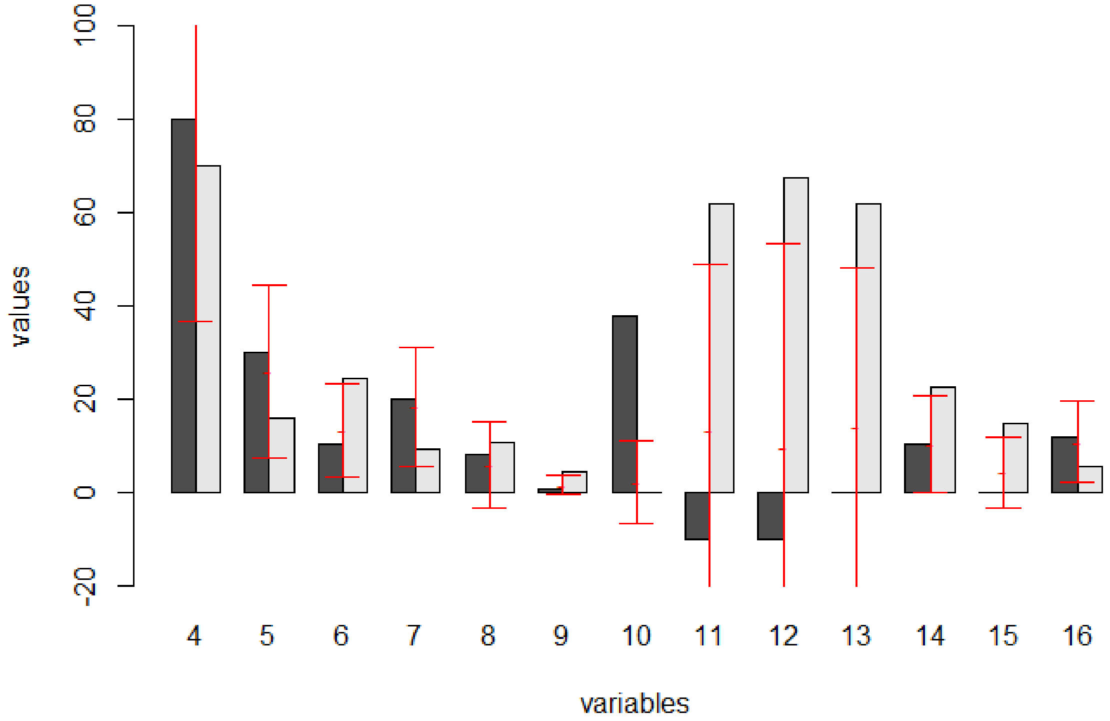
Plot 5: Scatter plot all data points projected on PC1 and PC2



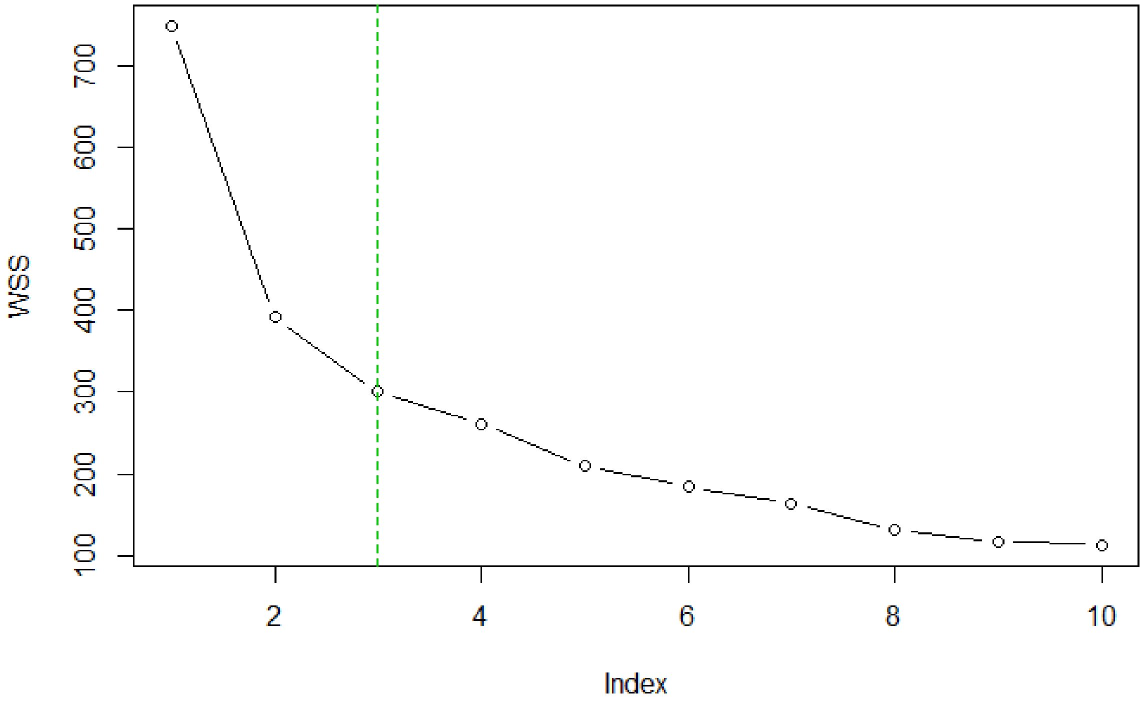
Plot 6: Dendrogram for Hierarchical Clustering (Complete Linkage, Euclidean distance) on new 3 dimensions



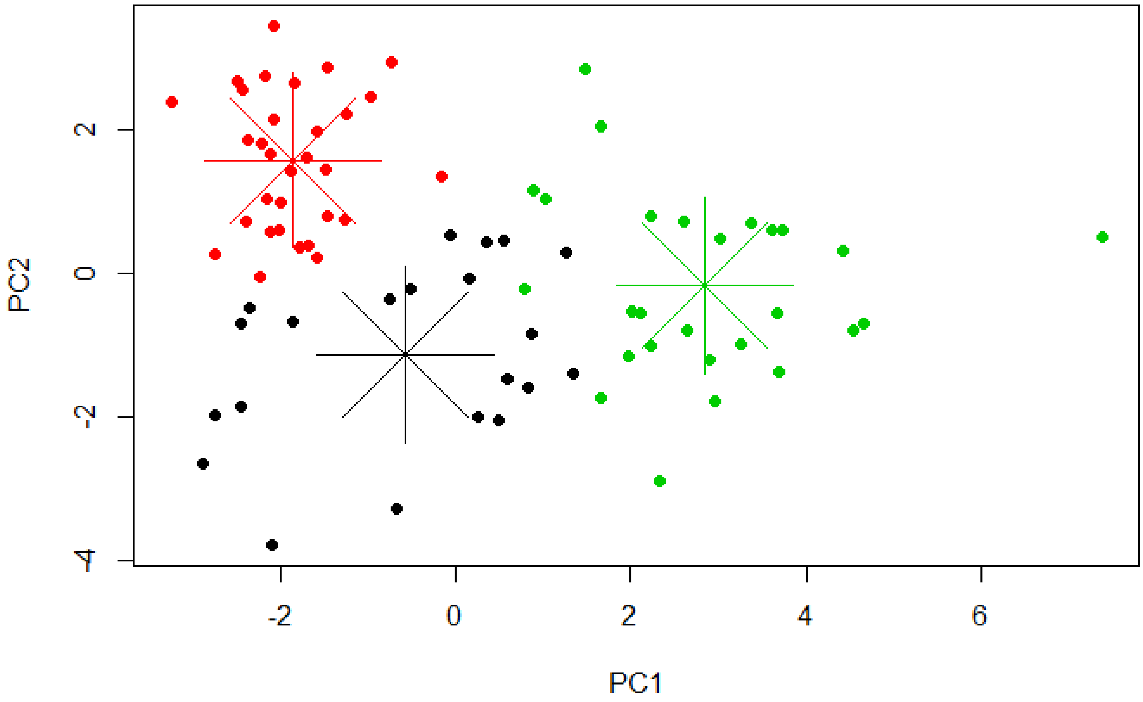
Plot 7: Observations D55 (dark) and D69 (light) compared to $mean \mp 2 * sd$ (arrows in red)



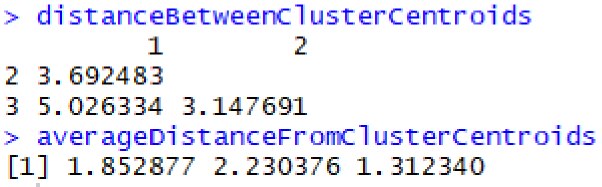
Plot 8: Total of the within sum of squares values versus k



Plot 9: K-means clustering with k = 3, projected data on PC1 and PC2

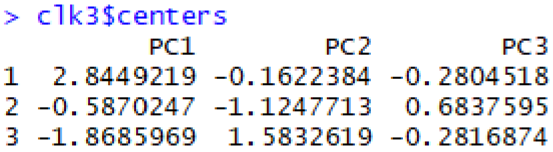


R Output 3: Distance between Cluster Centroids, Average Distance from Cluster Centroids for k-Means (k=3)



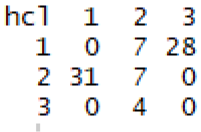
```
> distanceBetweenClusterCentroids
         1        2
2 3.692483
3 5.026334 3.147691
> averageDistanceFromClusterCentroids
[1] 1.852877 2.230376 1.312340
```

R Output 4: Cluster centroids for k=3



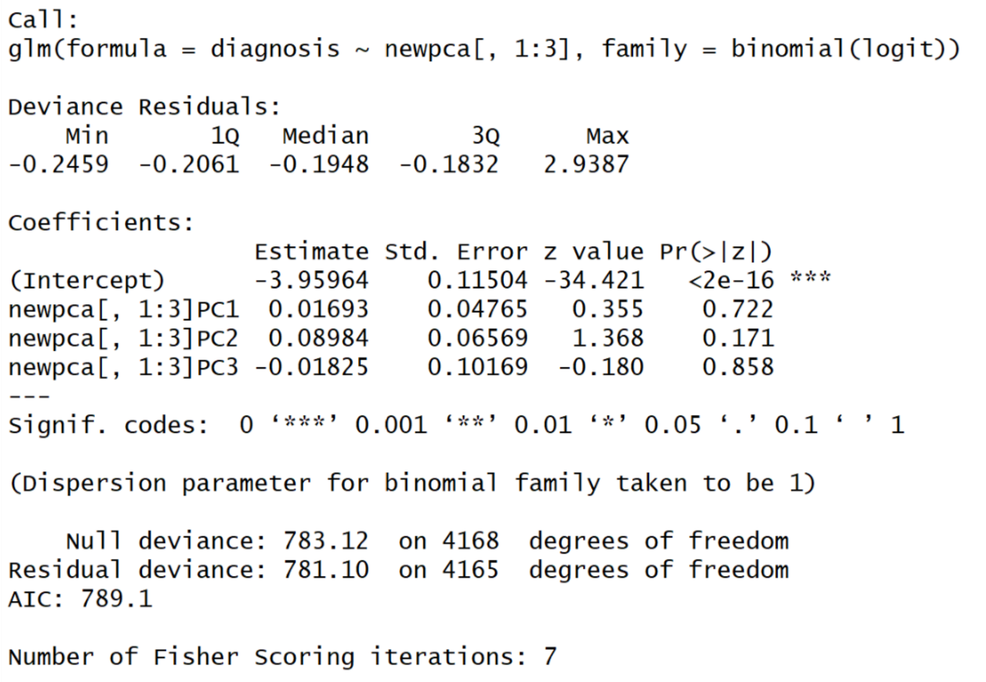
```
> clk3$centers
         PC1        PC2        PC3
1  2.8449219 -0.1622384 -0.2804518
2 -0.5870247 -1.1247713  0.6837595
3 -1.8685969  1.5832619 -0.2816874
```

R Output 5: Different classifications between hierarchical clustering and iterative one



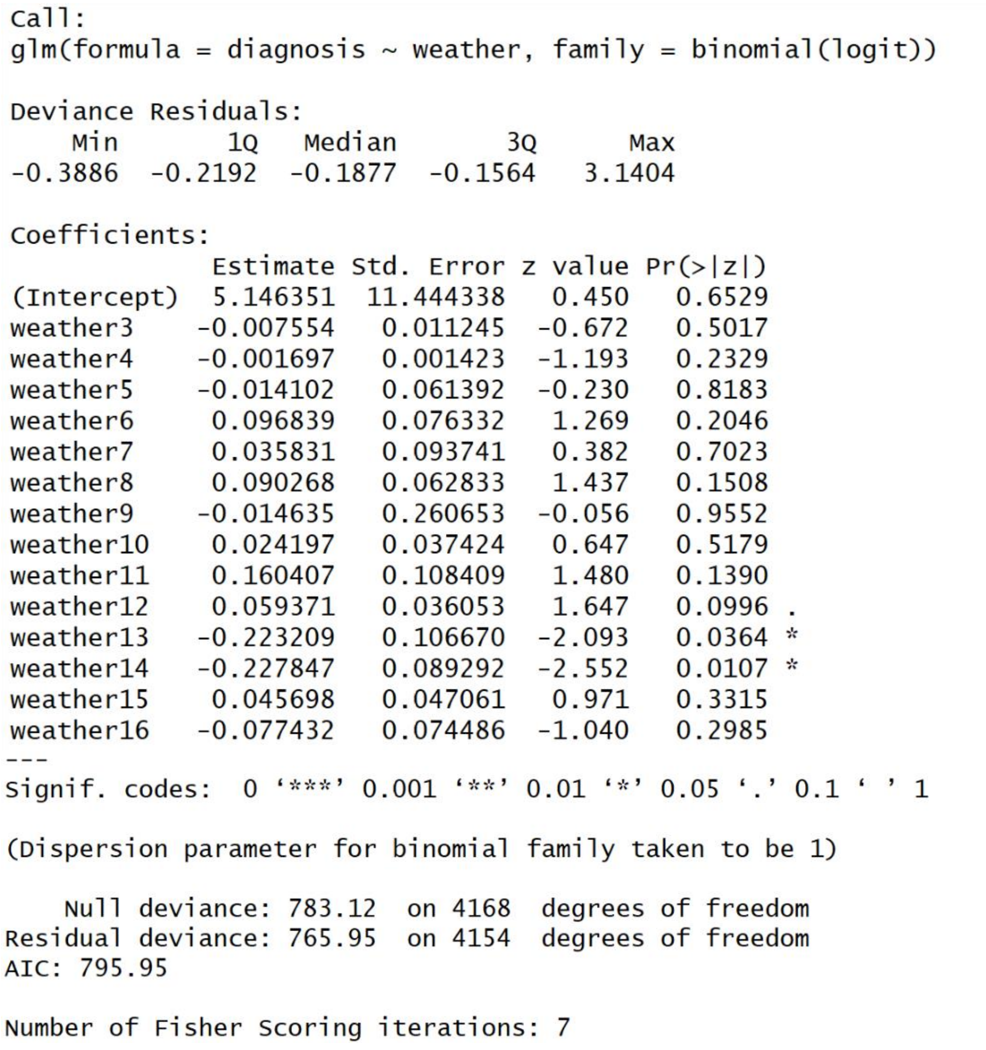
```
hcl  1  2  3
  1  0  7 28
  2 31  7  0
  3  0  4  0
```

R Output 6: Logistic Regression Diagnosis ~ PCs



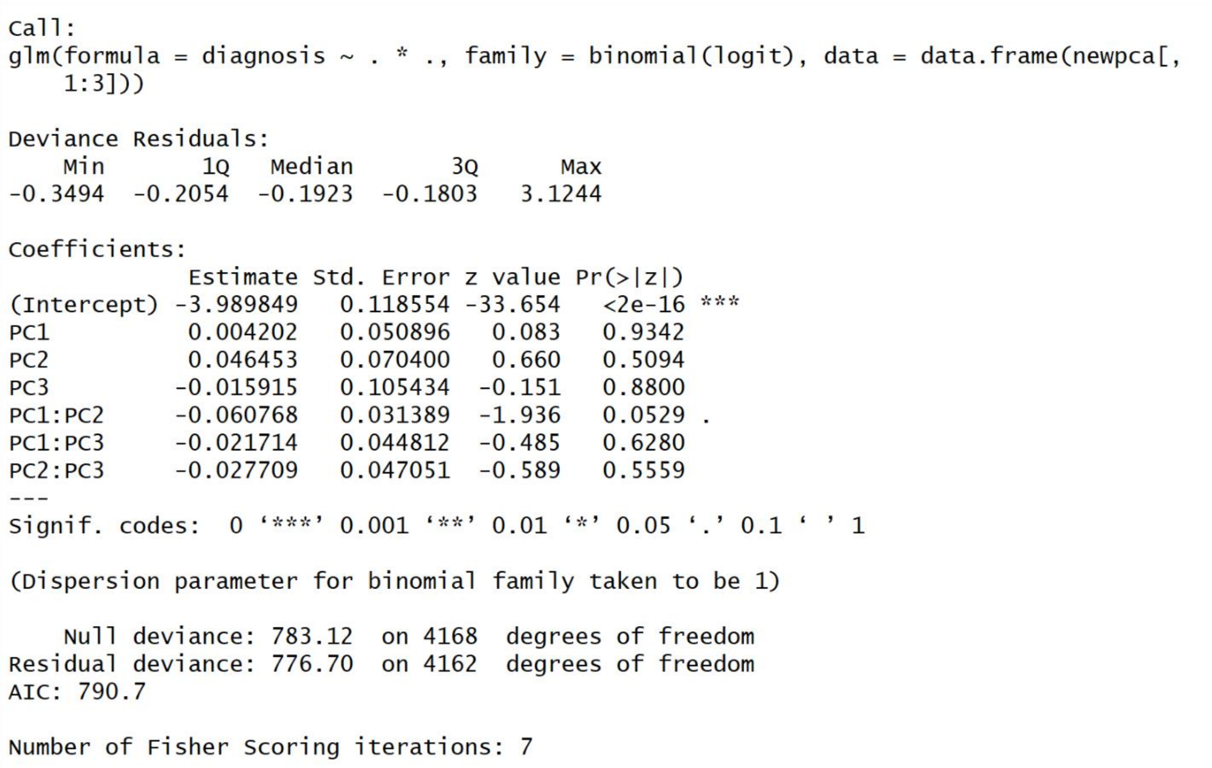
```
Call:
glm(formula = diagnosis ~ newpca[, 1:3], family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2459  -0.2061  -0.1948  -0.1832   2.9387

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.95964    0.11504 -34.421   <2e-16 ***
newpca[, 1:3]PC1  0.01693    0.04765   0.355    0.722
newpca[, 1:3]PC2  0.08984    0.06569   1.368    0.171
newpca[, 1:3]PC3 -0.01825    0.10169  -0.180    0.858
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 783.12  on 4168  degrees of freedom
Residual deviance: 781.10  on 4165  degrees of freedom
AIC: 789.1

Number of Fisher Scoring iterations: 7
```

R Output 7: Logistic Regression Diagnosis ~ Weather



```
Call:
glm(formula = diagnosis ~ weather, family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3886  -0.2192  -0.1877  -0.1564   3.1404

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.146351  11.444338   0.450   0.6529
weather3    -0.007554   0.011245  -0.672   0.5017
weather4    -0.001697   0.001423  -1.193   0.2329
weather5    -0.014102   0.061392  -0.230   0.8183
weather6     0.096839   0.076332   1.269   0.2046
weather7     0.035831   0.093741   0.382   0.7023
weather8     0.090268   0.062833   1.437   0.1508
weather9    -0.014635   0.260653  -0.056   0.9552
weather10    0.024197   0.037424   0.647   0.5179
weather11    0.160407   0.108409   1.480   0.1390
weather12    0.059371   0.036053   1.647   0.0996 .
weather13   -0.223209   0.106670  -2.093   0.0364 *
weather14   -0.227847   0.089292  -2.552   0.0107 *
weather15    0.045698   0.047061   0.971   0.3315
weather16   -0.077432   0.074486  -1.040   0.2985
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 783.12  on 4168  degrees of freedom
Residual deviance: 765.95  on 4154  degrees of freedom
AIC: 795.95

Number of Fisher Scoring iterations: 7
```

R Output 8: Logistic Regression: Diagnosis ~ PCs including Interactions



```
Call:
glm(formula = diagnosis ~ . * ., family = binomial(logit), data = data.frame(newpca[,
    1:3]))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3494  -0.2054  -0.1923  -0.1803   3.1244

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.989849   0.118554 -33.654   <2e-16 ***
PC1          0.004202   0.050896   0.083   0.9342
PC2          0.046453   0.070400   0.660   0.5094
PC3         -0.015915   0.105434  -0.151   0.8800
PC1:PC2     -0.060768   0.031389  -1.936   0.0529 .
PC1:PC3     -0.021714   0.044812  -0.485   0.6280
PC2:PC3     -0.027709   0.047051  -0.589   0.5559
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 783.12  on 4168  degrees of freedom
Residual deviance: 776.70  on 4162  degrees of freedom
AIC: 790.7

Number of Fisher Scoring iterations: 7
```