Project 1

BACON and Hotelling's T-Squared Test

Dana Salem 900191664

Nour Montasser 900191101

Introduction

Statement of Problem

Medical databases generally involve many different types of data such as patient age, blood group, weight, clinical images, patient diagnoses, lab test results and other details from patient treatments. These databases produce a large amount of information. In fact, with the exponential development of information technology in hospitals, the volume of medical data has increased significantly. At the same time, new interest in the analysis of this information has emerged. However, the analysis of this information depends on the quality of the data which is a function of the presence of outliers. By identifying the outliers in medical datasets, for instance the Breast Cancer Wisconsin Dataset, abnormal patterns in health records can be detected, consequently improving the decision making process in terms of diagnosis. Given that a wrong diagnosis by a clinician may lead to the death of a patient, it is crucially important to identify, analyze and treat outliers in the dataset. For that reason, the aim of this project is to undergo an outlier identification process on The Breast Cancer Wisconsin (Diagnostic) Dataset using a statistical methods called the Blocked, Adaptive, Computationally-Efficient Outlier Nominator (BACON) approach. And then Hotelling T^2 test is

applied to examine the difference between the mean of the texture and symmetry components of the benign and malignant patients.

Literature Review

The detection and analysis of outliers is an important step in the field of medicine. To clarify, misdiagnosis can cost the patient extra time and money, allow the patient's health to deteriorate, and put the patient's life at risk (Zylstra et al., 1994). One of the reasons for misdiagnosis could be a result of labeling error which is running tests on mislabeled samples. In fact, researchers have indicated that around 10-15% of the samples are mislabeled in a microarray resulting in potential outlier data points (Zhang et al., 2009). If these outlier data points were to be detected when examining the test results, the patient would not have been influenced by the label noise, correctly diagnosed, and may receive appropriate treatment instead of ineffective treatment, or even harmful treatment consequently increasing their survival rate (Zhang et al., 2009). However, provided that the outlying observation is not a result of a labeling error, they may be unusual clinical cases that may reveal hidden information on the covariate and probably be worth studying further leading to scientific breakthroughs (Segaert P et al. ,2018).

Data Description

Data Source

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

https://archive.ics.uci.edu/ml/citation_policy.html

Summary of Data

The Breast Cancer Wisconsin (Diagnostic) Dataset is a publicly available dataset created by Dr. William H. Wolberg from the University of Wisconsin and was donated to the UCI Machine Learning Repository. It contains measurements of the characteristics of breast cancer tumors with no missing values. The 569 observations were gathered from a digitized image of a fine needle aspirate (FNA) of breast masses of female patients diagnosed with breast cancer. There are 12 variables in this dataset 2 of which are nominal and categorical and the remaining are quantitative. For the Hotelling T2 test this dataset will be divided into two groups malignant patients and benign patients. The description of each variable is given in the below Table 1.0.

Variables Description

Variable Name	**Type**	**Unit of Measurement**	**Description**
id	Nominal	N/A	This is a unique identifier assigned to each patient.
diagnosis	Categorical	N/A	This is a qualitative variable that identifies the type of the tumor whether it is malignant (M) or benign (B).

radius_mean	Quantitative	Millimeters	This feature measures the mean of distances from the center to points on the perimeter of the tumor. It represents the average size of the tumor.
texture_mean	Quantitative	N/A	This feature measures the mean of gray-scale values in the image of the tumor. It represents the variation in the pixel intensities in the image. It is measured on a scale from 0 to 100.
perimeter_mean	Quantitative	Millimeters	This feature measures the mean perimeter of the tumor, which is the length of its boundary.
area_mean	Quantitative	Millimeters Squared	This feature measures the area of the tumor, which is the total number of pixels inside the boundary. It is measured in square mm.

smoothness_mean	Quantitative	N/A	This feature measures the mean local variation in radius lengths of the tumor. It represents how much the radius of the tumor changes at different points along its boundary. It is measured on a scale from 0 to 1.
compactness_mean	Quantitative	N/A	This feature measures the ratio of the perimeter squared to the area of the tumor, minus 1.0. It represents how tightly the tumor is packed together. It is measured on a scale from 0 to 1.
concavity_mean	Quantitative	N/A	This feature measures the mean severity of concave portions of the contour of the tumor. It represents the amount of concavity in the boundary of the tumor. It is measured on a scale from 0 to 1.

concavepoints_mean	Quantitative	N/A	This feature measures the number of concave portions of the contour of the tumor. It represents the number of inwardly curved sections in the boundary of the tumor. It is measured on a scale from 0 to 1.
symmetry_mean	Quantitative	N/A	This feature measures how symmetric the tumor is. It represents how similar the left and right halves of the tumor are. It is measured on a scale from 0 to 1.
fractaldimension_mean	Quantitative	N/A	This feature measures the mean "coastline approximation" of the tumor. It represents how much the boundary of the tumor is convoluted. It is measured on a scale from 0 to 1.

Table 1.0 – Variable Description

Data Analysis

BACON

A pairwise comparison of the multivariate data is necessary prior to the use of the BACON algorithm for outlier detection. For that reason, the pairwise scatterplots and the Pearson correlation values for the non-qualitative variables were examined. These results allow the formation of expectations on the presence or the absence of outlying observations in the dataset.

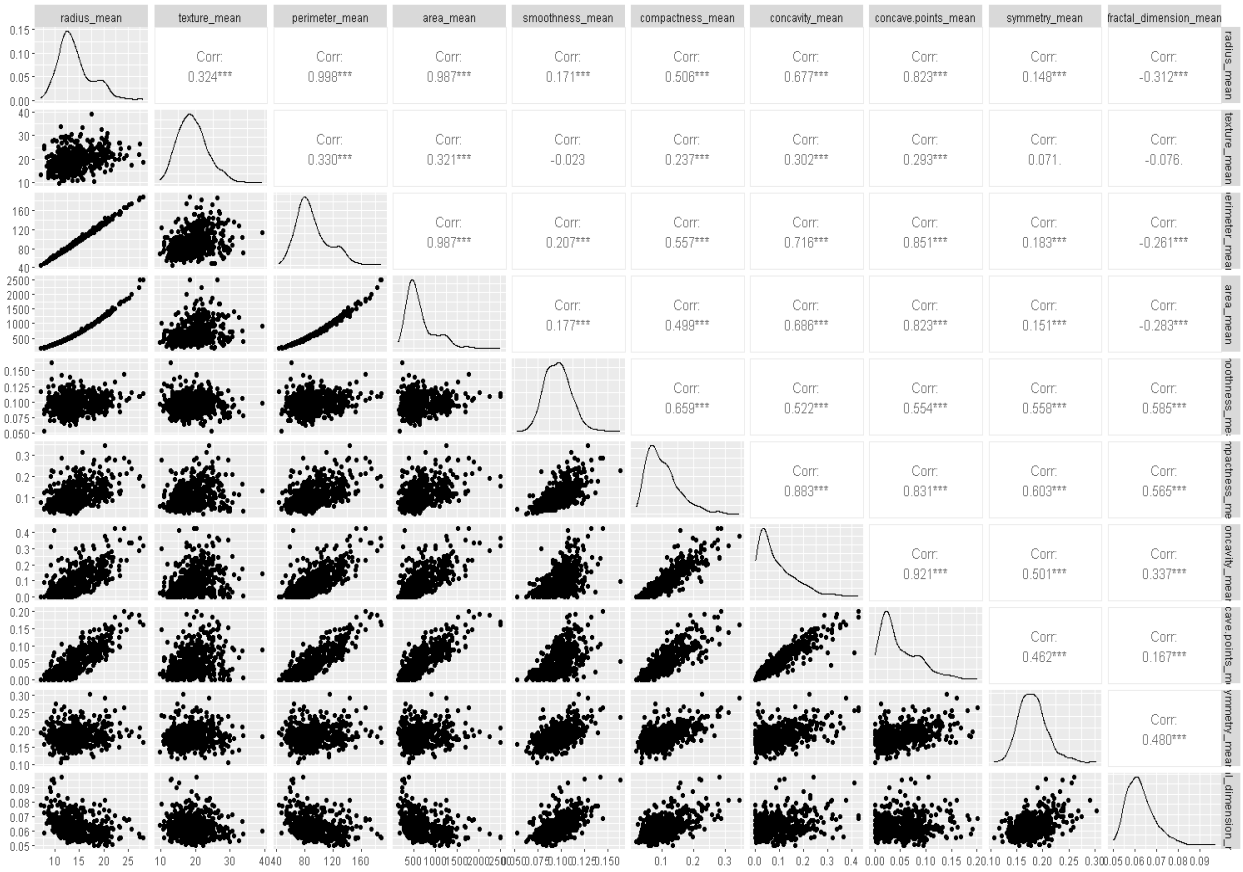

Figure 1.0 – Pairwise Scatterplots and Correlations

Examining the Figure 1.0 indicate that the majority of the variables exhibit a linear relationship. In fact, the positive linear relationships are more common than the negative linear relationships. In addition, it can be inferred from the correlation values that few of the variables have weak

linear relationships with one another. However, the scatterplots also illustrate the presence of some outlying observations since most of the data points are concentrated in a specific area and few points are further away from the bulk of the data implying that the correlation values might be misleading. Moreover, it can be concluded from the diagonal graphs that most of the variables are relatively skewed to the right indicating the presence of extreme values. Although some hypotheses have been formed about the presence of outliers in the dataset from visualizing the scatterplots, the implementation of the BACON algorithm will provide more accurate conclusions about the outlier analysis such as the number of outlying observations and which data points they represent.

Therefore, the mvBACON function in R was used to detect the outliers using the BACON algorithm. In this function, The initial Mahalanobis distances are computed for each of the 569 observations. Then, the initial basic subset is constructed from the computed distances where $c = 4$ resulting in 40 observations. Moreover, the robust Mahalanobis distances are computed for the observations in the basic subset and compared to the adjusted chi-square alpha value = 0.05 to identify the outlying observations. The mvBACON function iterates the previous steps until all the outliers have been detected. The results of executing this function in R are reached after 9 iterations where 47 observations out of the 569 observations are outliers. In percentage terms, the data set consists of 8.26% outliers. The comprehensive output of the mvBACON function is found below in Figure 2.0.

```
> output= mvBACON(x);
rank(x.ord[1:m,] >= p  ==> chosen m =  40
MV-BACON (subset no. 1): 40 of 569 (7.03 %)
MV-BACON (subset no. 2): 450 of 569 (79.09 %)
MV-BACON (subset no. 3): 498 of 569 (87.52 %)
MV-BACON (subset no. 4): 513 of 569 (90.16 %)
MV-BACON (subset no. 5): 517 of 569 (90.86 %)
MV-BACON (subset no. 6): 520 of 569 (91.39 %)
MV-BACON (subset no. 7): 521 of 569 (91.56 %)
MV-BACON (subset no. 8): 522 of 569 (91.74 %)
MV-BACON (subset no. 9): 522 of 569 (91.74 %)
```

Figure 2.0 – Output of BACON Algorithm

Hotelling's T2 Test

To ensure the assumptions of normality and homogeneity of the variance for Hotelling T2 test were met, log transformation was used to normalize the variables as some were skewed to the right as previous shown in the Figure 1.0. The variables texture_mean and symmetry_mean variables were selected to perform the Hotelling T2 on them for medical reasons and that they were the two most approximately normal variables when the log transformation was applied as seen in Figure 3.0 and 4.0. The data was divided into 2 groups, malignant diagnosis data and benign diagnosis data.

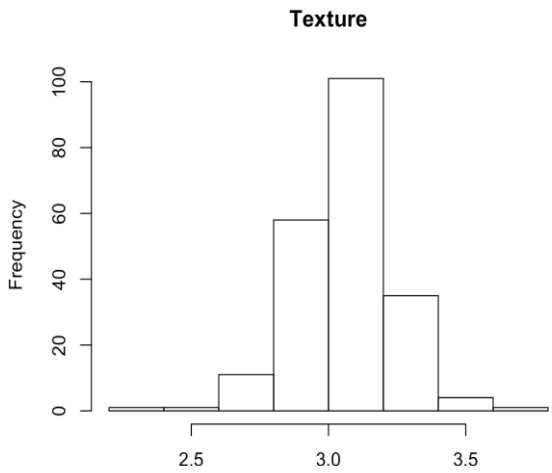

Figure 3.0 – Histogram of Texture Variable

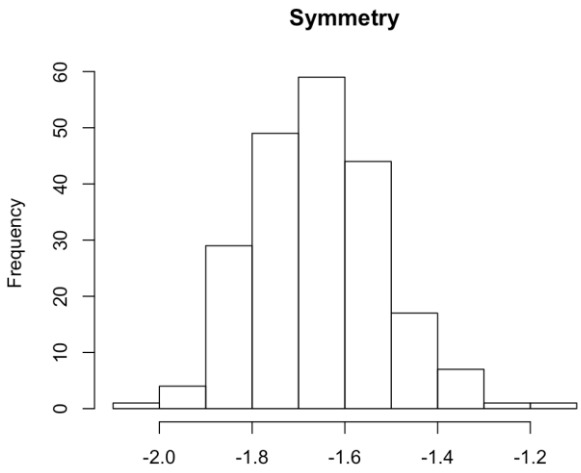

Figure 4.0 – Histogram of Symmetry Variable

Both texture and symmetry are known to be important characteristics of breast tumors. Texture refers to the spatial arrangement of cells within a tumor, while symmetry refers to the degree of similarity between the left and right halves of the breast. Both of these characteristics can be indicative of the malignancy of a tumor, as malignant tumors tend to have more irregular texture and less symmetry compared to benign tumors. Several studies have shown that texture and symmetry can be useful features for distinguishing between malignant and benign breast tumors. To elaborate, a study by Waugh et al. (2016) found that texture features derived from mammograms were able to accurately classify breast tumors as benign or malignant with a high degree of accuracy. Similarly, a study by de Oliveira et al. (2017) found that symmetry features were able to distinguish between benign and malignant breast tumors with high accuracy.

We performed the t2.test on both the full dataset (with outliers) and the cleaned dataset (without outliers) to validate the significance of the outliers. The null hypothesis is that there is no significant difference between the mean vectors of the two groups, similarly the alternative hypothesis is that there is difference between the mean vectors of the two groups.

Results – With outliers

```
        Two-sample Hotelling test

data:  x and y
T2 = 212.55, F = 106.09, df1 = 2, df2 = 566, p-value < 2.2e-16
alternative hypothesis: true difference in mean vectors is not equal to (0,0)
sample estimates:
              texture_mean symmetry_mean
mean x-vector     3.057882     -1.655332
mean y-vector     2.862456     -1.757494
```

Figure 5.0 – Hotelling's T2 Output (With Outliers)

The t2.test function in R was applied to compare the means of two groups, malignant and benign tumors, on the variables texture_mean and symmetry_mean in the breast cancer dataset. The result shows that the two groups differ significantly in terms of these variables, with a p-value of less than 2.2e-16, indicating that the probability of observing such an extreme difference by chance alone is extremely low. The test statistic T2 is 212.55, and the associated F-statistic is 106.09. The degrees of freedom are df1 = 2 and df2 = 566.

The sample estimates show that the mean texture_mean and symmetry_mean values for the malignant group are 3.057882 and -1.655332, respectively, while for the benign group, they are 2.862456 and -1.757494, respectively. This suggests that the malignant tumors tend to have higher texture_mean values and lower symmetry_mean values compared to benign tumors. Overall, these results suggest that texture_mean and symmetry_mean variables can be useful in differentiating between malignant and benign breast tumors.

Results – Without Outliers

```
        Two-sample Hotelling test

data:  z and k
T2 = 168.359, F = 84.018, df1 = 2, df2 = 519, p-value < 2.2e-16
alternative hypothesis: true difference in mean vectors is not equal to (0,0)
sample estimates:
              texture_mean symmetry_mean
mean x-vector     3.043627     -1.669541
mean y-vector     2.862240     -1.764289
```

Figure 6.0 – Hotelling's T2 Output (Without Outliers)

The output shows that the calculated T2 statistic is 168.359, with an associated F-statistic of 84.018. The degrees of freedom are df1 = 2 and df2 = 519. The p-value for the test is < 2.2e-16, indicating strong evidence against the null hypothesis. The sample estimates show that the mean

texture_mean and symmetry_mean values for the malignant group are 3.043627 and -1.669541, respectively, while for the benign group, they are 2.862240 and -1.764289, respectively. This suggests that the malignant tumors tend to have higher texture_mean values and lower symmetry_mean values compared to benign tumors. Overall, these results suggest that texture_mean and symmetry_mean variables can be useful in differentiating between malignant and benign breast tumors.

Comparison

The T2 statistic is 168.359 for the dataset without outliers compared to 212.55 for the dataset with outliers, which suggests that the removal of outliers has reduced the distance between the two groups. The F-statistic is also lower for the dataset without outliers (84.018 vs. 106.09), indicating that the difference in mean vectors between the two groups is less significant when outliers are removed. These findings suggest that the outliers were having a significant impact on the analysis and that their removal has resulted in more reliable and accurate results.

It is also important to note that the sample estimates for the mean vectors are very similar between the two datasets, indicating that the presence of outliers does not have a significant impact on the actual mean values for the texture and symmetry variables. Overall, these results suggest that the texture and symmetry variables are significantly different between the malignant and benign groups, regardless of the presence of outliers. However, removing outliers can lead to more reliable results and a better understanding of the underlying differences between the groups.

Conclusion

The findings of the project conclude that the Breast Cancer Wisconsin Dataset has 47 outlying observations accounting for around 8.26% of the dataset that were detected using the BACON

algorithm. Additionally, the insights further indicate the mean vectors of both the texture and symmetry variables, for the with outliers and non-outliers datasets, are not equal for the benign and malignant patients. This is understandable because these are two distinct types of tumors that are differentiated by texture and symmetry. However, it noted that the numerical values of the sample means of the texture and symmetry, before and after deleting the outliers, are inflated indicating that the outlying observations have an influence on the outcome. This reassures the notion that outliers might impact the results thus affecting the decision making process for clinicians. However, in this particular dataset, both the results with and without the outliers reach the same conclusion, but this may not be the case in other datasets where outliers might comprise more than 8.26% of the dataset. For that reason, the analysis and detection of outliers prior to conducting any tests in the medical field is crucially important in order to avoid making any incorrect decisions such as misdiagnosis.

References

Gaspar, J., Catumbela, E., Marques, B., & Freitas, A. (2011). A Systematic Review of Outliers Detection Techniques in Medical Data-Preliminary Study. HEALTHINF, 575-582.

Oliveira, M. A. F., de Oliveira, C. M., & Rodrigues, M. A. (2017). Breast cancer diagnosis using symmetry analysis. In IEEE International Conference on Image Processing (ICIP) (pp. 2604-2608).

Segaert P, Lopes MB, Casimiro S, Vinga S, Rousseeuw PJ. Robust identification of target genes and outliers in triple-negative breast cancer data. Stat Methods Med Res. 2018;962280218794722. [PMC free article] [PubMed]

Waugh, S. A., Purdie, C. A., Jordan, L. B., Vinnicombe, S., & Lerski, R. A. (2016). Texture analysis of mammography and MRI for breast cancer detection: A review. Medical Engineering & Physics, 38(10), 959-968.

Zhang C, Wu C, Blanzieri E, Zhou Y, Wang Y, Du W, Liang Y. Methods for labeling error detection in microarrays based on the effect of data perturbation on the regression model. Bioinformatics. 2009;25(20):2708–2714. [PubMed] [Google Scholar]

Zylstra, S., Bors-Koefoed, R., Mondor, M., Anti, D., Giordano, K., & Resseguie, L. J. (1994). A statistical model for predicting the outcome in breast cancer malpractice lawsuits. Obstetrics and gynecology, 84(3), 392–398.