Applied Multivariate Analysis

Project #1 - Outlier Detection

Omar Moustafa

900222400

MACT 4233

Spring 2025

Introduction

Complete Source of the Data:

The data set that will be used and analyzed originates from a global dataset that records life expectancy, body mass index (BMI), and blood pressure (BP) for males and females across different countries and years. It compiles multiple health-related indicators to allow for the analysis of life expectancy trends and factors commonly associated with them. Thus, this data set seeks to explore and visualize the relationships between these key health variables, providing insights into the factors that play into living a longer and healthier life.

What is the Main Objective?

The main objective of this investigation is to compare and contrast two prominent outlier-detection statistical methods which are the Mahalanobis Distance method and the BACON Algorithm. While both are well-renowned and work to detect the presence of outliers in a data set, they differ vastly in their computational efficiency and robustness. By implementing both methods to this dataset, the similarities and differences between them will become clearer, highlighting the circumstances in which each method is more effective. Therefore, this analysis will provide a clearer understanding of the two methods and confirm when Mahalanobis Distance or BACON is preferable for outlier detection as opposed to the other.

Description of the Data

What Are The Observations?

The data set comprises 5,307 observations, each representing a specific country in a given year. It contains multiple necessary health measures for both genders, which allows for an

analysis of variations across different regions and time periods. By examining these observations and detecting potential outliers, valuable insights can be extracted about the existing relationships between BMI, BP, and life expectancy, along with any significant deviations from expected patterns.

Full Definition of the Variables:

Variable Name:	**Variable Type:**	**Unit of Measurement:**
country	Nominal	N/A (Country Name)
year	Interval	Year (1980, 1981, … , 2008)
male_bmi	Quantitative	kg/m²
male_bp	Quantitative	mmHg
male_expectancy	Quantitative	Years
female_bmi	Quantitative	kg/m²
female_bp	Quantitative	mmHg
female_expectancy	Quantitative	Years

Descriptive Statistics:

The data set includes eight variables, six of which are quantitative. Those six being male_bmi, male_bp, male_expectancy, female_bmi, female_bp, and female_expectancy. From the summary statistics table below, it can be seen that the mean male and female BMI values are 23.99 kg/m² and 24.62 kg/m² range between 19–33.9 kg/m² for males and 18.5–35 kg/m² for

females, suggesting variation in body mass indices throughout the observations. The mean BP values for males and females are 130.9 mmHg and 127.4 mmHg, respectively, showing that females have a slightly lower average blood pressure than males. Additionally, the male life expectancy range is 13.10–81.80 years, whereas the female range is 15.70–86.80, indicating that women have quite a bit longer life expectancy. This aligns with the general knowledge and notion that females tend to live longer than males.

```
summary(df_numeric)
```

```
    male_bmi        male_bp      male_expectancy   female_bmi      female_bp     female_expectancy
 Min.   :19.00   Min.   :118.5   Min.   :13.10   Min.   :18.50   Min.   :110.3   Min.   :15.70
 1st Qu.:21.80   1st Qu.:128.0   1st Qu.:56.80   1st Qu.:22.70   1st Qu.:124.4   1st Qu.:61.20
 Median :24.40   Median :131.0   Median :65.10   Median :24.90   Median :127.6   Median :71.60
 Mean   :23.99   Mean   :130.9   Mean   :63.08   Mean   :24.62   Mean   :127.4   Mean   :68.42
 3rd Qu.:25.70   3rd Qu.:133.7   3rd Qu.:70.80   3rd Qu.:26.20   3rd Qu.:130.5   3rd Qu.:76.70
 Max.   :33.90   Max.   :143.1   Max.   :81.80   Max.   :35.00   Max.   :139.5   Max.   :86.80
```

Table 1.1

Displaying and analyzing the correlation matrix is essential as it provides valuable insights into the existing relationships among all quantitative variables. The table below reveals that the trends in both BMI and BP are very similar for both genders, with male_bmi and female_bmi showing a strong positive correlation of 0.883 and male_bp and female_bp presenting a correlation of about 0.84. Furthermore, male_expectancy and female_expectancy also display an almost maximum correlation value of approximately 0.972, suggesting that, similar to BMI and BP, life expectancy is nearly identical between males and females within a country. Conversely, female_bp and female_expectancy show a weaker correlation of roughly -0.033, implying that, for females, blood pressure does not have a significant linear relationship with life expectancy and it being negative implies that the higher the blood pressure, the lower their life expectancy tends to be.

A matrix: 6 × 6 of type dbl

	male_bmi	male_bp	male_expectancy	female_bmi	female_bp	female_expectancy
male_bmi	1.0000000	0.2572285	0.69076615	0.88325774	0.04479440	0.72538926
male_bp	0.2572285	1.0000000	0.13794607	0.19282424	0.83983138	0.19053089
male_expectancy	0.6907661	0.1379461	1.00000000	0.57509345	-0.06177103	0.97235860
female_bmi	0.8832577	0.1928242	0.57509345	1.00000000	0.09770405	0.60931716
female_bp	0.0447944	0.8398314	-0.06177103	0.09770405	1.00000000	-0.03267807
female_expectancy	0.7253893	0.1905309	0.97235860	0.60931716	-0.03267807	1.00000000

Table 1.2

Data Analysis

Outlier Detection Using Mahalanobis Distance:

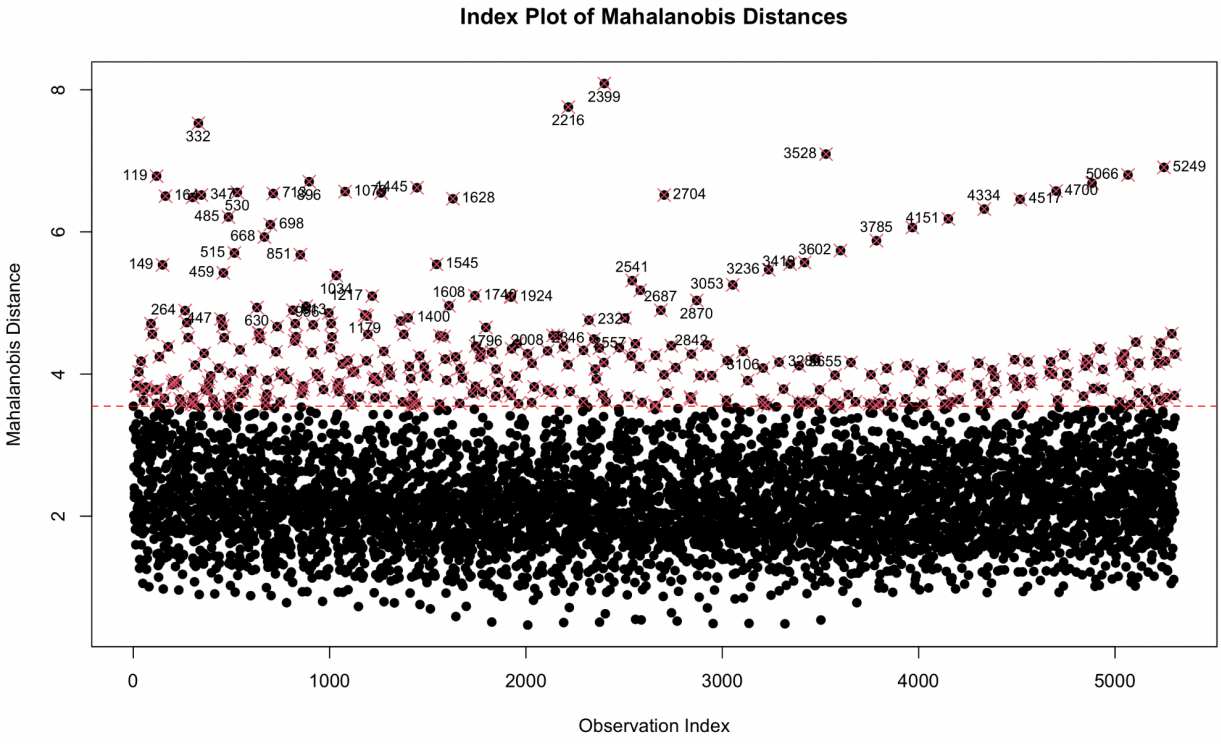

Figure 1.1

The first figure, displayed above, is a scatter plot visualizing the outliers detected using the Mahalanobis Distance method. A total of 323 outliers were identified using this method, indicating that it is highly sensitive to changes within the multivariate distribution. The minimum value among the identified outliers is 16, while the maximum is 5304, demonstrating that these outliers span an extensive range of indices and are highly dispersed throughout the sample. Building on this, the large number of outliers detected suggests that this method is somewhat aggressive when identifying potential outliers, which may include points that are multivariate outliers but may not be particularly extreme from a univariate perspective. The large number of outliers detected demonstrates that values that are even slightly greater than the general trend are susceptible to being classified as outliers under this method.

Outlier Detection Using The BACON Algorithm:

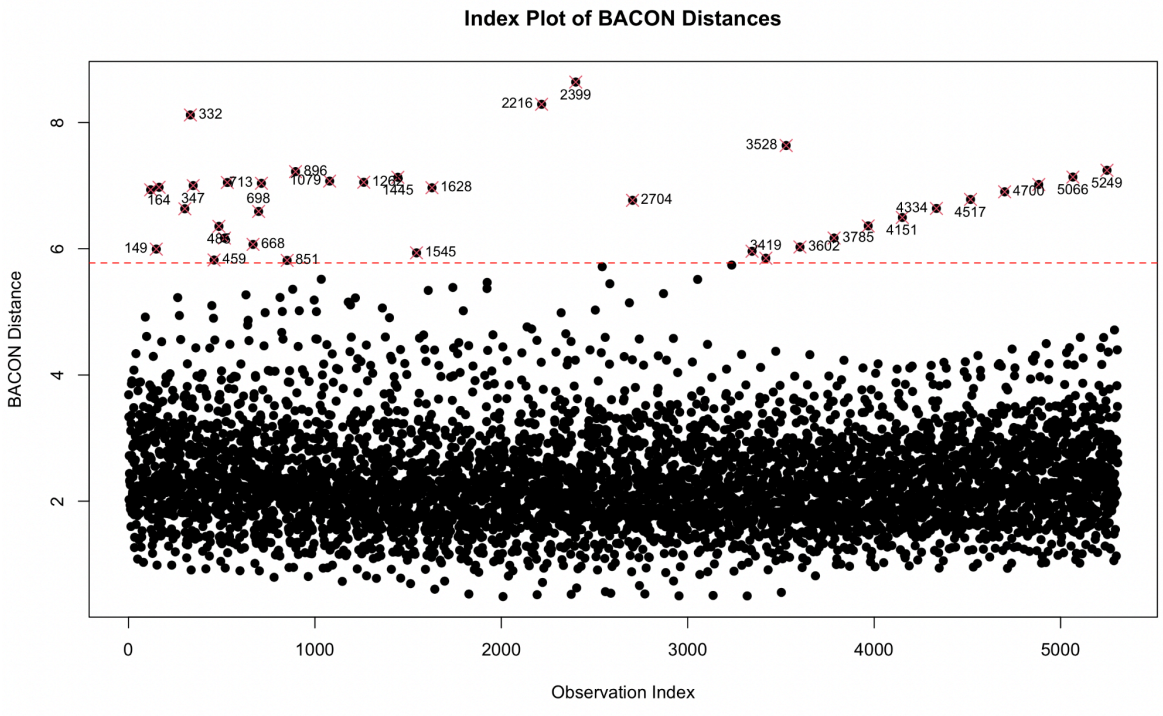

Figure 1.2

Moving onto the second figure, the above scatter plot visualizes the outliers detected using the BACON Algorithm. This second method identified only a fraction of the number of outliers that were identified using Mahalanobis Distance, with only 36 points being identified as outliers. The BACON Algorithm is widely known for being computationally efficient and more robust than Mahalanobis Distance as it focuses on the more extreme values while impervious to misclassifying a regular point as an outlier, namely swamping. The relatively minimal number of outliers detected demonstrates that BACON is much more conservative in its approach and methods, and it will likely end up excluding a large handful of borderline cases that were flagged as outliers by Mahalanobis Distance.

Q-Q Plot Analysis of Squared MD and BD:

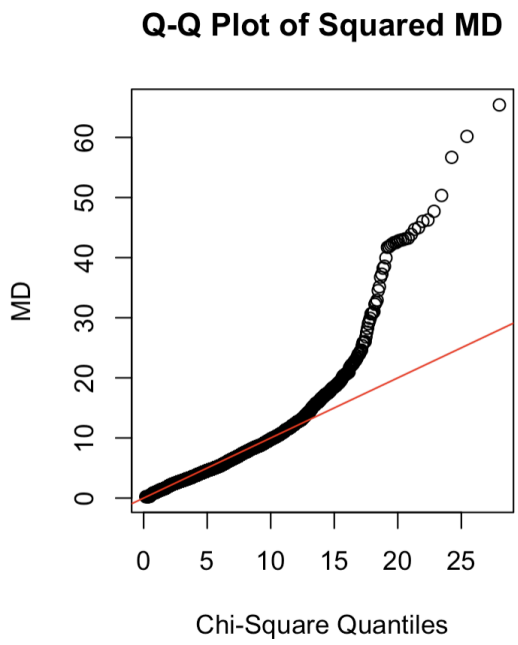

Figure 1.3a

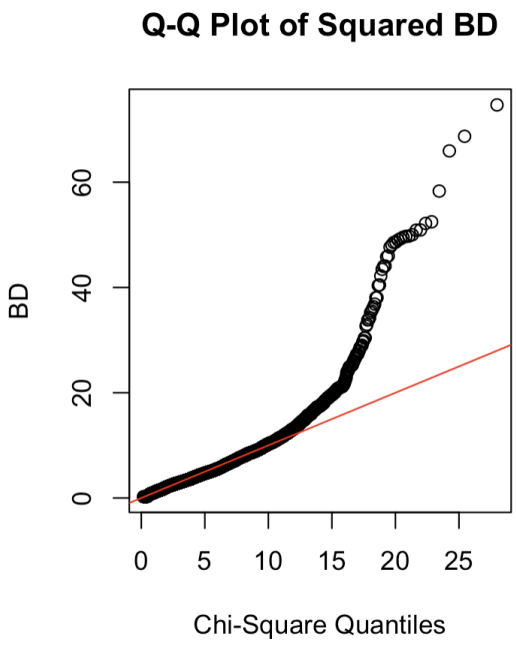

Figure 1.4b

The visualizations and corresponding analyses of the two scatter plots are followed by the two side-by-side Q-Q plots displayed above. This further evaluates the distribution of outliers by comparing the Squared Mahalanobis Distance (MD) and the Squared BACON Distance (BD) to theoretical Chi-Square quantiles. While both generally align with the expected distribution, there are notable differences in their treatment of outliers. **Figure 1.3a**, which represents Squared MD, shows greater deviation at the upper quantiles, meaning it works to capture not just extreme outliers but even moderate ones as well. Conversely, the upper quantiles of **Figure 1.3b**, which represents Squared BD, remain more aligned with the theoretical line, suggesting that the most extreme outliers will be flagged and labeled as such.

Comparing The Two Methods:

When comparing the two methods, there are significant differences regarding the sensitivity and selectivity of the outliers. It turns out that BACON detects a much more limited subset of points, whereas Mahalanobis Distance identifies a considerably larger number of points as outliers. Despite this standout contradiction, both methods agree on a set of common outliers where these overlapping outliers strongly indicate that these specific points are likely to be extreme outliers. This is because these points were detected as outliers by the two methods, regardless of the different steps and approaches taken by each method.

Looking at the two index plots, the common outliers between them appear to be located at a Mahalanobis Distance of slightly less than 6 and above, such as the indices 3419, 3602, 3785, and several more. In **Figure 1.1**, representing the Mahalanobis Distance method, it can be seen that all of the points with a Mahalanobis Distance of approximately 3.5 and above are labeled as outliers. However, in **Figure 1.2**, representing the BACON Algorithm, most points falling

roughly between 3.5 and 6 appear still within the general range of the main cluster, only slightly elevated above the majority. This shows that these points, while they are somewhat distant, are not extreme outliers or deviations. In contrast, only the points with a Mahalanobis Distance of approximately 6 or greater were flagged as outliers as they stood out more prominently and deviated more significantly from the large cluster, making them strong candidates for extreme outliers classification.

This perspective is further supported by the side-by-side comparison of the Q-Q plot for each method where greater deviations of the MD plot from the theoretical plot compared to that of the BD plot also reveal the Mahalanobis Distance method captures a large amount of points and labels them outliers whereas the BACON Algorithm is less sensitive to minor changes, making it the more cautious approach of the two for outlier detection.

Conclusion

In conclusion, Mahalanobis Distance and the BACON Algorithm methods offer different perspectives on detecting and identifying outliers, with Mahalanobis Distance classifying significantly more points as outliers than the BACON Algorithm. BACON looks to be more cautious and selective, concentrating only on those that are extreme outliers, whereas Mahalanobis Distance, on the other hand, is more sensitive to even the slightest of outliers and deviations from the main cluster of points.

With that said, the two methods still overlap in certain areas, which indicates that they both reliably identify the most extreme outliers. This supports the notion that the BACON Algorithm helps identify the more critical and extreme outliers. On the contrary, Mahalanobis Distance does so in addition to the more subtle and slight outliers. Therefore, one method may be more suitable and ideal depending on the application, where BACON helps recognize the most

extreme outliers while Mahalanobis Distance helps detect a much wider range of outliers, both slight and extreme.