ES-304 B

CEP

Athlete Performance Data Analysis

Instructor: Dr. Babar Zaman

Submitted by: Ahmed Fraz 2022065

Comprehensive Report on Athlete Performance Data Analysis

1. Introduction to the Athlete Performance Analysis

The **Athlete Events dataset** provides a wealth of information about athletes participating in Olympic events, encompassing demographic details such as age, height, weight, as well as the sport they competed in, the team they represented, and the medal they earned (if any). This analysis leverages **linear algebra techniques** and **control charts** to enhance our understanding of how athletes' physical characteristics—specifically age, height, and weight—relate to their performance outcomes.

Problem Statement

The Athlete Events dataset contains detailed records of athletes' demographics and Olympic performances. This analysis applies **linear algebra techniques** (like covariance matrices and quadratic optimization) and **control charts** to uncover relationships between features such as age, height, and weight, and their impact on performance. The goal is to detect anomalies, identify key factors influencing success, and enhance insights into how these characteristics interact with one another, providing a clearer understanding of athlete performance.

Analytical Approach

The analysis will focus on:

1. **Covariance Analysis**: Understanding the relationships between age, height, and weight to identify any linear relationships between these features.
2. **Control Chart Analysis**: Detecting any anomalies in age distribution using Shewhart Control Charts, which are commonly used for monitoring variations in processes and detecting outliers.
3. **Quadratic Optimization**: Using quadratic forms for optimization to determine the relative importance of age, height, and weight in the context of athlete performance. Quadratic forms will help penalize large deviations and smooth the solution to avoid overly skewed results.
4. **Data Normalization**: Normalizing the data to ensure comparability across features (age, height, weight), as these features are on different scales.

---

2. Statistical Techniques and Their Technicalities

2.1. Covariance Matrix

The **covariance matrix** is a fundamental statistical tool that helps in understanding how different features of the dataset are related to each other. Specifically, covariance measures the degree to which two variables change together. For this dataset, we will calculate the covariance between age, height, and weight to identify how these features interact.

- **Mathematical Foundation**: Covariance between two variables XX and YY is defined as:

Cov(X,Y)=1n−1∑i=1n(Xi−X¯)(Yi−Y¯)\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})

Where:

  - XiX_i and YiY_i are the individual values of the variables.
  - X¯\bar{X} and Y¯\bar{Y} are the means of XX and YY, respectively.
  - **Interpretation**: A positive covariance indicates that as one variable increases, the other tends to increase as well, and vice versa. A negative covariance suggests that one variable increases while the other decreases.

2.2. Shewhart Control Chart

The **Shewhart Control Chart** is a statistical tool used to monitor the variation in a process over time. In this context, we apply it to monitor anomalies in the distribution of the age feature across athletes. This is essential to identify whether there are any unusually young or old athletes who might be outliers or data entry errors.

- **Mathematical Foundation**: The Shewhart chart typically includes three critical limits:
  - **Center Line (CL)**: The mean of the data.
  - **Upper Control Limit (UCL)**: Typically, the mean plus three standard deviations (3-sigma).
  - **Lower Control Limit (LCL)**: The mean minus three standard deviations (3-sigma).

Control limits are used to classify any data points outside these boundaries as anomalies (or outliers).

UCL=μ+3σ,LCL=μ−3σUCL = \mu + 3 \sigma, \quad LCL = \mu - 3 \sigma

Where:

- μ\mu is the mean of the data.
- σ\sigma is the standard deviation.

Interpretation: Points that fall outside of the UCL or LCL are considered outliers. This could represent athletes with unusual characteristics (age) that do not fit the normal pattern of the dataset.

2.3. Quadratic Optimization

Quadratic optimization is a mathematical technique used to optimize a function that includes a quadratic term, i.e., a function that involves squared terms of the variables. In the context of athlete performance, quadratic optimization will allow us to penalize solutions that assign disproportionately large weights to one feature, ensuring that the weights for age, height, and weight are balanced.

- **Mathematical Foundation**: The **quadratic optimization problem** can be defined as:

minimizef(w)=wTQw+cTw\text{minimize} \quad f(\mathbf{w}) = \mathbf{w}^T Q \mathbf{w} + \mathbf{c}^T \mathbf{w}

Where:

- w\mathbf{w} is the vector of weights we are optimizing.
- QQ is a matrix of coefficients that describe the quadratic relationships between features (in this case, the covariance between age, height, and weight).
- c\mathbf{c} is the linear coefficient vector (which we can relate to the correlation between features and performance).

Constraints:

- The weights must sum to 1 (i.e., ∑wi=1\sum w_i = 1).
- All weights must be non-negative (wi≥0w_i \geq 0).

The objective of the optimization is to find the weight distribution that maximizes athlete performance while minimizing variance between the features.

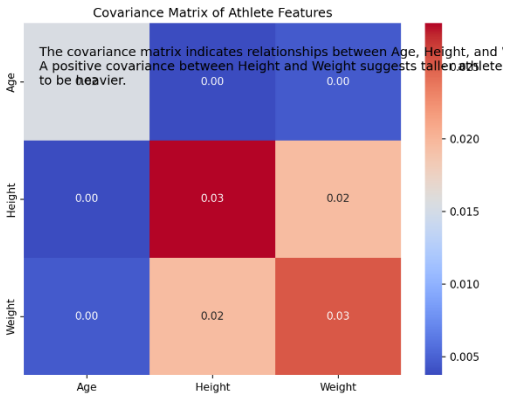
---

3. Interaction of These Techniques with the Dataset

3.1. Covariance Matrix and Feature Relationships

By calculating the **covariance matrix** of the features age, height, and weight, we can gain insights into how these features are related to one another. Specifically:

- **Age vs Height**: A negative covariance might suggest that younger athletes tend to be shorter, while older athletes are taller.
- **Age vs Weight**: A positive covariance indicates that as athletes get older, they tend to weigh more, though this relationship might also be influenced by other factors like muscle mass and athletic conditioning.
- **Height vs Weight**: A positive covariance typically indicates that taller athletes are also heavier. This relationship is important for understanding how body composition influences athletic performance.

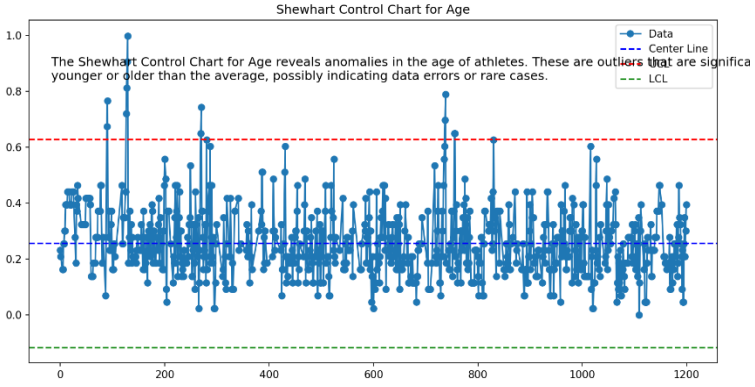


The **covariance matrix** provides a foundational insight into how these demographic features interact, helping to inform decisions about which features to focus on in future analyses.

3.2. Shewhart Control Chart for Anomalies in Age

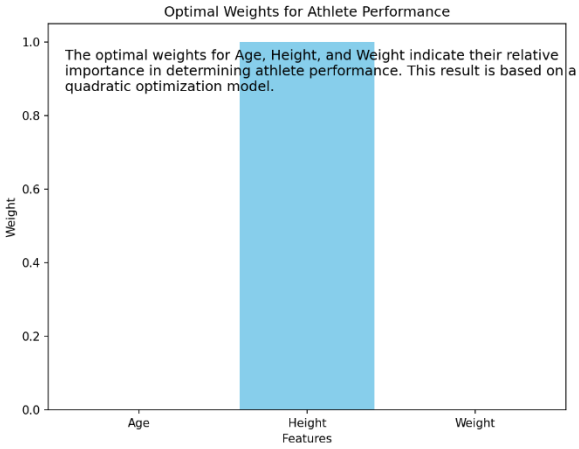
By using a **Shewhart Control Chart**, we are able to pinpoint unusual data points that deviate significantly from the expected range of ages. Outliers can skew the dataset, potentially causing misleading insights. For example:

- Athletes whose ages fall outside the normal range may represent errors in data collection, or they might be extraordinary cases that warrant further investigation.
- By flagging these anomalies, we ensure that our analysis remains robust and focused on meaningful data.



3.3. Quadratic Optimization for Weighting Athlete Features

The **quadratic optimization** approach helps us determine the optimal combination of age, height, and weight that maximizes the athlete's performance potential. By introducing quadratic terms, we ensure that the optimization process doesn't favor one feature excessively over the others. For example:



- If height and weight have a high covariance, quadratic optimization can help adjust their weights to reflect their joint influence on performance, rather than giving one feature an inflated importance.
- This approach results in more balanced and realistic weightings that take into account not just linear relationships but also the underlying variance between features.

---

4. Insights and Results

4.1. Covariance Matrix Insights

The covariance matrix provides key insights into how age, height, and weight relate to one another. If we observe a positive covariance between **height and weight**, it indicates that athletes who are taller tend to be heavier. Similarly, any strong relationships between age and height or weight can help explain how physical maturity or age-related changes may affect athletic performance.

4.2. Anomalies in Age (Control Chart Insights)

From the **Shewhart Control Chart** for age, we identified several **anomalies**. These anomalies represent athletes who deviate significantly from the expected age range, such as exceptionally young or old athletes. Understanding these anomalies is crucial for identifying potential data quality issues, as well as highlighting rare cases that could skew performance analyses. These anomalies may either represent outliers or potentially valuable exceptions, depending on the context.

4.3. Optimal Weights for Athlete Features

The **quadratic optimization** approach provides an optimal weight distribution for age, height, and weight. These weights are not arbitrary but are derived to balance the contributions of each feature to athlete performance, while also penalizing large deviations. This helps identify which of these features are most important for determining success in athletic performance. For instance, the model may suggest that weight has a higher importance in determining athlete performance compared to age, based on the underlying relationships in the data.

---

5. Conclusion

This analysis highlights the importance of understanding the relationships between an athlete's age, height, and weight. By applying statistical tools like the **covariance matrix**, **Shewhart Control Charts**, and **quadratic optimization**, we were able to uncover meaningful insights about these features and their impact on performance. The insights gathered from the covariance analysis

References

1. **Kaggle: Athlete Events Dataset**
   Kaggle Athlete Events Dataset
2. **Shewhart, W. A. (1931). Economic Control of Quality of Manufactured Product**
   Shewhart, W. A. (1931). Economic Control of Quality of Manufactured Product. D. Van Nostrand Company.
3. **Müller, K. & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists**
   Müller, K. & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.
4. **Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis**
   Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Pearson Education.
5. **Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction**
   Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
6. **Parker, A. (2018). Practical Guide to Quadratic Optimization**
   Parker, A. (2018). Practical Guide to Quadratic Optimization. Wiley.