International Journal of Applied Sports Sciences
2025, Vol. 37, No. 2, 237-256.
https://doi.org/10.24985/ijass.2025.37.2.237

ISSN 2233-7946 (Online)
ISSN 1598-2939 (Print)
© Korea Institute of Sport Science

Handling Non-Normal Data in Sport Analytics: The Application of Box-Cox Transformation to MLB and LPGA Data

Seokyong Lee[a], Minseo Kim[a], & Soowoong Hwang[b*]

[a]*Department of Physical Education, Seoul National University, South Korea*
[b]*Division of Sport Studies, College of Sport Studies and Arts, Myongji University, South Korea*

Abstract

Advancements in Information and Communication Technology (ICT) and big data have significantly transformed sports analytics, enabling the collection of complex, multidimensional datasets. However, sports data often exhibit non-normal distributions, skewness, and outliers, which pose challenges for linear models used in association analysis. In this study, the effectiveness of the Box-Cox transformation in addressing these issues was evaluated using datasets from ICT-based sports data, specifically datasets from Major League Baseball (MLB) and the Ladies Professional Golf Association (LPGA). Dependent variable distributions, regression model performance, and residual patterns were compared before and after the transformation. The Box-Cox transformation effectively reduced skewness and improved normality, ensuring that fundamental regression assumptions such as homoscedasticity and linearity were satisfied. Improved model fit was observed across the datasets, as evidenced by higher R^2 values, lower Akaike Information Criterion (AIC) scores, and more evenly distributed residuals. These findings demonstrate that the Box-Cox transformation enhances the reliability and interpretability of regression models in sports analytics, particularly for non-normal data, by addressing both distributional characteristics and residual behaviors.

Key words: sports data transformation, sports ICT, sports big data, sports analytics, regression assumption validity

Introduction

Advances in computing technology, such as improvements in CPU and GPU performance, have enabled the efficient computation of neural networks, while innovations in Information and Communication Technology (ICT), including wearable devices and real-time tracking systems, have transformed data collection and analysis (Ardagna et al., 2016; Huang et al., 2017). These developments have driven sports analytics from simple, frequency-based metrics to detailed analyses incorporating multidimensional data. For instance, in baseball, player performance evaluation now integrates advanced metrics like pitch velocity, spin rate, and movement, collected in real time. Similarly, in soccer, analytics has expanded beyond basic indicators like pass success rates to include metrics such as sprint speeds, heart rate variability, and positional movements. In professional golf now utilizes advanced ball-tracking systems to analyze every shot's trajectory and its impact on performance and financial outcomes. The growing complexity and volume of

Submitted : 26 November 2025
Revised : 23 December 2025
Accepted : 29 December 2025
* Correspondence : sportsict@mju.ac.kr

sports data has facilitated applications in diverse domains, including player performance evaluation, game strategy optimization, and injury risk assessment (Bai & Bai, 2021).

Sports analytics using ICT based big data can be broadly categorized into two primary approaches. The first focuses on developing predictive models, primarily driven by deep learning, which excel in predictive accuracy. These models have been applied to various analyses, such as predicting player performance, assessing injury risks, and forecasting game outcomes (Kumar et al., 2024). By learning complex patterns in data, deep learning enables precise predictions. However, its black-box nature often limits its ability to clearly explain relationships between variables, posing significant challenges to interpretability (Castelvecchi, 2016). While predictive accuracy is important, sports analytics must also provide insight into the underlying processes influencing outcomes (Amendolara et al., 2023). Without understanding how independent variables interact and affect results, the external validity of deep learning models may remain limited.

Given the interpretability limitations of deep learning, researchers frequently employ regression analysis to explore interactions and relationships between variables. Regression analysis assumes a linear relationship between dependent and independent variables, with the slope representing the strength of this relationship. This approach is a robust tool for quantifying the contributions of independent variables to dependent outcomes. However, its reliability depends on satisfying key assumptions such as normality, homoscedasticity, and linearity. Failure to satisfy these assumptions can result in biased coefficient estimates and unreliable interpretations (Osborne & Waters, 2002).

Modern sports data, such as scoring metrics, speed, distance, and playing time, frequently exhibit non-normal distributions. This tendency is further influenced by the increasing complexity of contemporary datasets, which are multidimensional and incorporate player movements, game contexts, and environmental factors. Unlike earlier sports data that often featured relatively simple, frequency-based structures, modern datasets capture the intricate dynamics of real-time interactions and external influences, inherently resulting in more complex distributions. These datasets are increasingly enriched through ICT-based technologies such as GPS tracking, wearable devices, and sensor networks, which enable the real-time collection of highly detailed metrics. However, these detailed metrics often exhibit skewed distributions, extreme values (outliers), or non-linear relationships between variables, posing challenges for conventional regression models to accurately capture underlying patterns.

Conducting regression analysis with non-normally distributed data can lead to biased coefficient estimates and challenges in interpretation. When normality assumptions are violated, the residuals of the model may not meet the required conditions, further compromising the validity of the analysis. Similarly, the lack of homoscedasticity—where residual variance is unequal across levels of an independent variable—makes reliable inference difficult. To address these challenges, researchers have increasingly turned to the Box-Cox transformation, a statistical technique designed to approximate normality in data distributions (Osborne, 2010; Zhang & Yang, 2017; Atkinson et al., 2021). By modifying data using an optimized exponent (λ), the Box-Cox transformation effectively reduces skewness and kurtosis in non-normal datasets. The optimal λ is determined by maximizing the log-likelihood function (Box & Cox, 1964; Zhou & Zou, 2024). Applying this transformation to dependent variables enables regression models to better adapt to the data's underlying distribution, making it easier to satisfy normality assumptions and improving the reliability and interpretability of results.

For example, positively skewed variables such as sprint speeds collected from wearable GPS devices or the frequency of high-intensity runs during a soccer match can be transformed to approximate normality, ensuring that residuals exhibit the required distributional properties for valid coefficient estimation. Additionally, when non-linear relationships exist—such as between a basketball player's cumulative workload (e.g., total accelerations and decelerations during a game) and injury risk metrics derived from wearable sensors—the

transformation can help clarify these relationships by normalizing the dependent variable. Furthermore, by reducing variability in complex, multidimensional datasets, the transformation increases the likelihood of meeting the assumption of homoscedasticity, thereby simplifying the interpretation of results in advanced sports analytics.

While foundational work by Nevill and Atkinson (1997) and Cooper et al. (2007) established critical frameworks for addressing non-linearity and measurement reliability in sports science, a systematic validation of these techniques remains limited in the context of contemporary ICT-based multidimensional big data. This study addresses this gap by providing a systematic comparison between Box-Cox and log transformations, utilizing high-fidelity performance data from MLB and the LPGA. By evaluating these methods against the complex, high-dimensional metrics that characterize the modern era, this research offers a refined analytical framework tailored to the unique data properties of today's technology-driven sports.

Theoretical Background

Characteristics of Sports Data and the Need for the Box-Cox Transformation

Sports data stands apart from other data types due to its multidimensional, dynamic, and context-sensitive nature. Unlike static datasets, sports data is generated in real time during games or player actions, capturing situational complexities that vary across different sports. For example, in baseball, a batter's swing speed and the ball's launch angle are influenced by game-specific factors, while in soccer, sprint speed and player positioning change dynamically based on the match's flow. These characteristics make sports data uniquely challenging for traditional statistical models that rely on simplified assumptions about data distributions (Balague et al., 2013; Morgulev et al., 2018).

A common issue in sports data is its frequent deviation from normality, which underscores the necessity for advanced transformation techniques like the Box-Cox transformation. Variables such as scoring metrics, playing time, and physiological measurements often exhibit positive skewness or extreme values due to the dynamic and context-specific nature of sports (Bai & Bai, 2021). For example, batted ball distance in baseball may feature extreme values resulting from home runs or short fly balls, while sprint distance in soccer often displays skewness due to differences in individual playing styles and situational demands (Rein & Memmert, 2016). In the context of professional golf, performance-related financial data such as total prize money typically exhibits extreme right-skewness due to the 'winner-take-all' nature of tournament prize structures. Such financial outcomes do not follow a normal distribution, as a small percentage of top-tier players earn a disproportionately large share of the total earnings -with the top 10% capturing nearly 55% of the purse- presenting a significant challenge for standard regression-based analysis (Rinehart, 2009). These distributional characteristics can violate the normality assumption of many statistical methods, leading to biased estimates and reduced model reliability.

The presence of outliers adds another layer of complexity. In sports, outliers often result from exceptional performances or unique game scenarios. For instance, a baseball pitcher achieving an unusually high number of strikeouts or a soccer team recording an extraordinary goal count in a single match are not merely statistical anomalies but carry meaningful insights about performance or strategy. Analytical methods must account for these outliers by preserving their contextual significance while addressing their potential to skew broader patterns.

Traditional transformation methods, such as log or square root transformations, have been commonly employed to address non-normality but often prove insufficient for the complexities of sports data. Log transformation, for instance, is restricted to positive data and cannot handle variables that include zero or negative values, such as point differentials (Lee, 2020). Similarly, square root transformations may reduce skewness but fail to adequately manage extreme values or non-linear relationships (Shi et al., 2013). These

limitations highlight the need for a more flexible and robust approach tailored to the unique distributional and contextual characteristics of sports data.

The Box-Cox transformation addresses these challenges by optimizing the λ parameter to adjust data distributions dynamically. This approach reduces skewness and kurtosis, aligning the data with the assumptions of statistical models while largely preserving key characteristics of the dataset (Marimuthu et al., 2022). Unlike simpler methods, the Box-Cox transformation accommodates the contextual nuances, enabling more accurate modeling of its inherent variability. This makes it particularly suitable for sports analytics, where the interpretability of extreme values and the preservation of data variability are critical. By bridging the gap between statistical rigor and contextual relevance, the Box-Cox transformation offers a methodological framework that enhances the reliability of analytical results across diverse applications, from performance evaluation to strategic decision-making.

Mechanism of the Box-Cox Transformation

The Box-Cox Transformation is a statistical technique proposed by Box and Cox in 1964 to normalize data distributions (Box & Cox, 1964). This method is used to transform non-normal data into forms suitable for statistical models such as linear regression, thereby improving the reliability and predictive accuracy of these models. Specifically, the Box-Cox Transformation reduces skewness and kurtosis, ensuring that the fundamental assumptions of statistical models—such as normality and homoscedasticity—are met (Draper & Smith, 1998).

The transformation operates on a parameter λ, which non-linearly adjusts the data. The transformation is defined as follows:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{(\lambda)}-1}{\lambda}, & if\ \lambda \neq 0 \\ \log(y_i), & if\ \lambda = 0 \end{cases} \quad \text{Equation (1)}$$

In Equation (1), y_i represents the observed value of the original data, and $y_i^{(\lambda)}$ denotes the transformed data. The parameter λ is a transformation coefficient optimized to normalize the data distribution. The goal of the Box-Cox Transformation is to adjust the data distribution to approximate normality, ensuring that the residuals of statistical models exhibit normality.

Unlike the log transformation (λ=0), which applies a fixed and uniform adjustment to the data, the Box-Cox transformation uses Maximum Likelihood Estimation (MLE) to determine the optimal λ value. MLE provides a statistically rigorous framework to estimate the parameter that maximizes the likelihood of the observed data given the model (Marimuthu et al., 2022). This approach minimizes skewness and kurtosis while aligning the transformed data closely with a normal distribution. This is essential for statistical modeling, as it ensures that key assumptions—such as normality and homoscedasticity—are satisfied, thereby enhancing the validity and interpretability of inferential results. The optimization of λ is based on the maximization of the log-likelihood function, expressed as:

$$\ell(\lambda) = -\frac{n}{2}\log\left\{\frac{\sum_{i=1}^{n}(y_i^{(\lambda)}-\bar{y}^{(\lambda)})}{n}\right\}^2 + (\lambda-1)\sum_{i=1}^{n}\log(y_i) \quad \text{Equation (2)}$$

In Equation (2), n is the number of observations, $\bar{y}^{(\lambda)}$ is the mean of the transformed data. The optimal λ value maximizes $\ell(\lambda)$, minimizing skewness and kurtosis while ensuring normality. The result of the transformation is that $y_i^{(\lambda)}$ approximates a normal distribution, making the data suitable for statistical modeling. This adaptive feature of the Box-Cox Transformation distinguishes it from static transformations like the log transformation, which may not adequately address varying degrees of skewness or kurtosis across datasets.

This flexibility is particularly important in sports analytics, where datasets often exhibit non-normality and extreme values. The adaptive nature of the Box-Cox Transformation enables it to preserve critical data characteristics, such as the relative influence of outliers, while ensuring that the data conforms to the assumptions of regression models. This makes the transformation a robust tool for improving the reliability

and predictive accuracy of statistical analyses in contexts where the nuances of data variability are crucial.

Methods

Dataset

This study aimed to explore the effectiveness of the Box-Cox transformation in addressing non-normality in sports data analysis. Representative datasets from two distinct types of sports—baseball and golf—were analyzed. Each sports was chosen to reflect different competition structures: baseball as a representative turn-based sport and golf as a representative individual sports (sports where individual skills and strategies primarily determine the outcome).

Both baseball and golf serve as emblematic examples of modern sports ICT applications, as they generate extensive big data through real-time ball-tracking technologies. These technological advancements have revolutionized sports analytics, enabling precise measurement and detailed analysis of player performance. Such innovations highlight the transformative potential of ICT in shaping data-driven approaches to sports science and analytics. Detailed descriptions of each dataset are provided below.

Major League Baseball (MLB) Dataset

The Baseball Savant platform, a hallmark of sports ICT, provides a wealth of advanced data, including pitch spin rates, batted ball distances, exit velocities, and launch angles, showcasing the cutting-edge capabilities of big data in sports. Leveraging data from Baseball Savant, this study analyzed the cumulative batted ball performance of hitters 133 qualified hitters who met the minimum threshold of 502 plate appearances during the 2024 season using Box-Cox regression. Key variables from the dataset included Fast Swing Rate, Avg Swing Length, Avg Exit Velocity, and Avg Launch Angle, which were selected to capture critical aspects of batting performance.

In this study, 'Fast Swing Rate' was designated as the dependent variable for the primary analysis. Beyond its role as a measure of swing speed and consistency directly relevant to batting performance, its selection serves an exploratory purpose to investigate the mechanical consistency and underlying intent of hitters when executing high-velocity swings. This metric allows for a retrospective examination of how a hitter's deliberate intent to swing fast aligns with technical elements (e.g., swing length, exit velocity, and launch angle) and physical factors (e.g., age and physical fitness), thereby capturing the behavioral consistency of elite performers. Moreover, the use of ICT-based precision data ensures the high reliability and reproducibility of this metric, further supporting its role as a robust dependent variable within this methodological framework. Detailed descriptions of each variable are provided below.

Fast Swing Rate: The percentage of a player's swings that reach a speed of at least 75 MPH

Avg Swing length: The average of total distance (in feet) traveled by the bat's barrel in three-dimensional space (X/Y/Z) from the start of the bat's motion to the point of impact with the ball.

Avg Exit Velocity: The average of exit velocity of the batted ball as tracked by Statcast.

Avg Launch angle: The average of launch angle of the batted ball as tracked by Statcast.

Ladies Professional Golf Association (LPGA) Golfers' Performance Data

The dataset titled "LPGA 2022 Player Performance" was originally sourced from the official LPGA website. Provided under the Public Domain license, the dataset was designed to analyze professional golf player performance and contains 158 observations across 16 variables. For this study, four key variables were selected for Box-Cox regression: totPrize, driveDist, avePutts, and fairPct. Detailed descriptions of these variables are as follows:

totPrize: Total official prize winnings in dollars.

driveDist: Average drive distance on par 4 and 5 holes in yards.
fairPct: Percentage of drives that landed on fairways.
avePutts: Average number of putts per round.

Comparison and Analysis Procedures

In this study, datasets from MLB and LPGA were used to analyze the changes in dependent variables after applying two transformation methods: the Box-Cox transformation and the log transformation. The results of these transformations were compared with the raw data to assess their effects on data normality and regression model performance. The detailed steps are as follows.

As the first step, the descriptive statistics (mean, standard deviation, skewness, and kurtosis) of the raw data for each variable were calculated. These metrics were used to evaluate whether the data tended to follow a normal distribution. Histograms were then used to visualize and evaluate the distributional characteristics of the data. To statistically verify normality, the Shapiro-Wilk test was conducted. The null hypothesis (H0) of the Shapiro-Wilk test is that "the data follow a normal distribution." If the null hypothesis is rejected, it can be concluded that the data do not follow a normal distribution (Shapiro et al., 1968).

In the next step, the Box-Cox transformation and the log transformation were employed for each dependent variable. For the Box-Cox transformation, the optimal λ value was estimated using MLE to minimize skewness and kurtosis while approximating a normal distribution. In parallel, a log transformation (corresponding to λ=0 in the Box-Cox framework) was conducted. After each transformation, the distributional changes in the data were reassessed using histograms and the Shapiro-Wilk test to evaluate normality. This process enabled comparisons among the raw data, Box-Cox-transformed data, and log-transformed data, offering insights into each method's effectiveness in addressing skewness and improving normality.

Finally, regression analyses were performed using the raw (λ = 1), Box-Cox-transformed (λ = determined by MLE), and log-transformed data (λ = 0). A Gaussian error distribution with constant variance was assumed for these models. To ensure a valid and consistent comparison of the Akaike Information Criterion (AIC) across different transformation scales, the Jacobian determinant was incorporated into the log-likelihood function used for both λ estimation and AIC calculation. This adjustment accounts for the change in the geometric scale of the data, thereby enabling a direct comparison between the untransformed and transformed models. Additionally, R^2 and RMSE values are scale-dependent and not directly comparable across different transformations, relative RMSE (normalized by the mean) and qualitative diagnostics of residuals were utilized as supplementary criteria for model evaluation. Model fit was evaluated using metrics such as R^2, Akaike Information Criterion (AIC), and Root Mean Square Error (RMSE), along with an examination of residual plots. This comparative approach underscored the relative strengths and limitations of each transformation method, highlighting the Box-Cox transformation's ability to balance improved regression model assumptions with the preservation of critical characteristics of sports data.

While regression coefficients in Box-Cox and log-transformed models are not directly interpretable in the original units (e.g., percentages or dollars), they provide essential information regarding the direction of influence (Positive/Negative), statistical significance. Furthermore, the magnitude of these coefficients reflects the sensitivity of the transformed dependent variable to a one-unit change in each respective predictor. While direct comparison of magnitudes across variables with different units requires caution, these values effectively quantify the specific contribution of each factor within the modeling framework.

Results

MLB Dataset

Table 1 provides a summary of the MLB dataset, including means, standard deviations, skewness, kurtosis, and Shapiro-Wilk test results for the variables used in the analysis: Fast Swing Rate, Age, Avg Swing Length,

Table 1. Descriptive statistics and normality tests for MLB dataset

Variables	Mean	Std.	Skewness	Kurtosis	Shapiro-Wilk	*p*-value
Fast Swing Rate	25.92	19.23	0.79	2.75	0.92	<.001
Age	27.73	3.65	0.46	3.13	0.98	0.02
Avg Swing Length	7.33	0.41	-0.58	3.86	0.97	0.006
Avg Exit Velocity	89.71	2.25	0.26	3.32	0.99	0.62
Avg Launch Angle	13.60	4.04	0.05	2.73	0.99	0.89

Avg Exit Velocity, and Avg Launch Angle. Given that Fast Swing Rate could potentially be influenced by a player's age, the variable Age was included in the regression analysis to control for its effects. The results then reveal important characteristics of the dataset.

The dependent variable, Fast Swing Rate, exhibits moderate skewness (.79) and moderate kurtosis (2.75), indicating a distribution that is moderately asymmetric with heavier tails. Furthermore, the Shapiro-Wilk test (W = 0.92, p < .001) confirms significant deviation from normality. Among the independent variables, Age shows slight skewness (0.46) and moderate kurtosis (3.13), with its Shapiro-Wilk test (W =.98, p =.02) suggesting a deviation from normality. Similarly, Avg Swing Length displays slight negative skewness (-.58) and higher kurtosis (3.86), with the Shapiro-Wilk test (W = 0.97, p =.006) confirming non-normality. On the other hand, Avg Exit Velocity and Avg Launch Angle exhibit near-symmetric distributions, with minimal skewness (.26 and .05, respectively) and Shapiro-Wilk test results indicating no significant departure from normality (p > .05).

Figure 1 illustrates the distribution of the dependent variable, Fast Swing Rate, under three conditions: Original (untransformed; λ=1), Box-Cox transformation with the MLE-derived optimal λ (0.384), and log transformation (λ=0). While Table 1 provides numerical insights into skewness and kurtosis, these histograms visually depict the effectiveness of each transformation in reducing skewness and bringing the distribution closer to normality.

The log transformation shifts the distribution from its original positive skewness toward a negatively skewed form. This highlights the potential drawback of log transformations in some contexts, as they may over-correct skewness and create distributions that deviate from normality in the opposite direction. This outcome can reduce interpretability, especially for

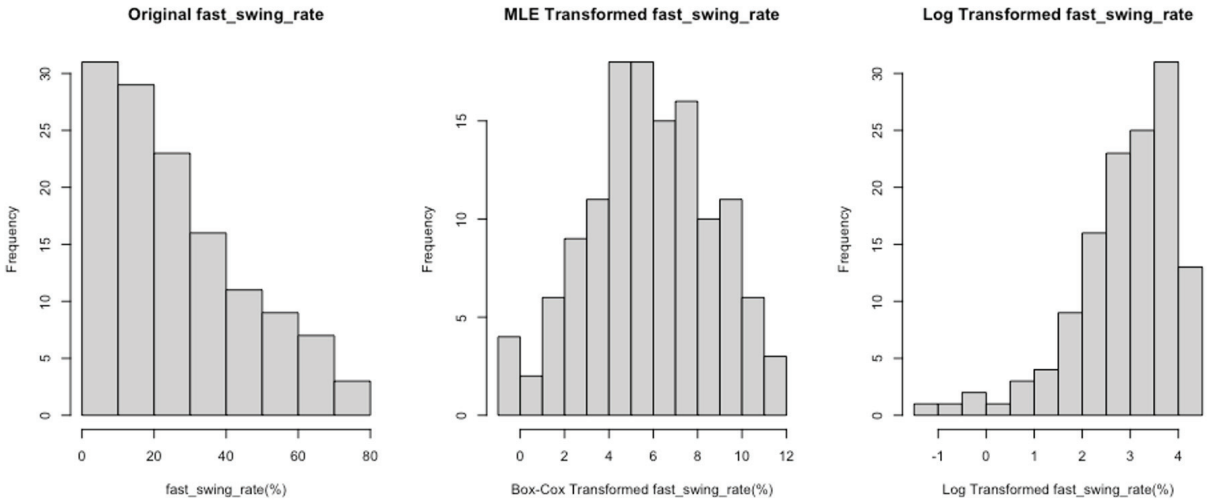

Figure 1. Histogram of fast swing rate for the MLB dataset

Table 2. Comparison of regression models for mlb dataset across different transformations

	λ=1 (Original)		λ=0.384 (Box-Cox transformation)		λ=0 (log transformation)	
	Estimate (p-value)	Std. Error	Estimate (p-value)	Std. Error	Estimate (p-value)	Std. Error
β0	-585.20 (<.001)	41.68	-85.08 (<.001)	5.73	-29.83 (<.001)	2.268
β1 (Age)	0.764 (.012)	0.298	0.104 (.012)	0.04	0.04 (0.017)	0.016
β2 (ASL)	10.18 (<.001)	2.725	2.12 (<.001)	0.375	1.277 (<.001)	0.148
β3 (AEV)	6.22 (<.001)	0.481	0.873 (<.001)	0.066	0.294 (<.001)	0.026
β4 (ALA)	-0.006 (0.982)	0.270	-0.014 (0.712)	0.037	-0.008 (0.561)	0.014
R^2	0.6493		0.6888		0.6598	
AIC	1004.644		492.806		253.585	
Relative RMSE	0.437		0.269		0.215	

Note: dependent variable=Fast Swing Rate; ASL: Avg Swing Length; AEV: Avg Exit Velocity; ALA: Avg Launch Angle. All regression coefficients are reported on the scale of the transformed dependent variable. Relative RMSE: The root mean square error (RMSE) scaled by the mean of each dependent variable to allow comparison across transformations.

datasets where symmetry is critical for analysis. Especially, in sports analytics, extreme values often represent critical outliers, such as extraordinary performance or unique events, which hold significant insights. By overly compressing these extremes, the log transformation may limit the interpretability of such data points.

The Box-Cox transformation, on the other hand, offers a balanced approach. It reduces skewness effectively while preserving the relative scale of extreme values, making it particularly suited for sports data analysis, where such outliers can carry meaningful contextual importance.

Although the visualized distributions demonstrate marked improvements in normality, further evaluation using regression modeling is essential to assess the transformations' impact on satisfying key assumptions, such as homoscedasticity and linearity, which are crucial for generating reliable and interpretable models. Table 2 contains these results, summarizing the key metrics and transformations used in the regression analysis.

To further evaluate the effects of these transformations on regression model performance, Table 2 presents a detailed comparison of regression models for the MLB dataset across three transformation methods: Original (untransformed; λ=1), the Box-Cox transformation using the MLE-based optimal λ=0.384, and the log transformation (λ=0). The results highlight significant differences in model performance and interpretation across the transformations.

The R^2 show moderate improvements in explained variance after transformation. The Box-Cox transformation achieved the highest R^2 value (0.6888), followed by the log transformation (0.6598) and the untransformed model (0.6493). This suggests that transformations help better capture the variability in the dependent variable, Fast Swing Rate, as explained by the independent variables.

The AIC scores indicate the log-transformed model achieves the best model fit, with the lowest AIC value (253.58), followed by the Box-Cox transformation (492.81) and the untransformed model (1004.64). Lower AIC scores emphasize that the log

transformation provides the most parsimonious model among the three approaches.

The regression coefficients (β) demonstrate substantial differences across transformations, particularly for predictors like Age (β_1), Avg Swing Length (β_2), and Avg Exit Velocity (β_3). Across all transformation methods, Avg Swing Length (β_2) and Avg Exit Velocity (β_3) consistently exhibit strong statistical significance ($p < .001$), indicating their robust influence on Fast Swing Rate. Notably, the impact of Age (β_1) becomes less pronounced in the log- transformed model (p=.017) compared to the untransformed and Box-Cox-transformed models (p=.012), suggesting that the choice of transformation method can affect the perceived strength of certain predictors. On the other hand, Avg Launch Angle (β_4) remains statistically insignificant across all transformations ($p > .5$), implying that it has a negligible direct effect on Fast Swing Rate.

The relative RMSE values underscore improved predictive accuracy after transformation. The log-transformed model achieved the lowest RMSE (0.215), followed by the Box-Cox model (0.269), while the untransformed model exhibited the highest RMSE (0.437). These RMSE values align with the visual patterns observed in Figure 2, which illustrates the relationship between actual and predicted values of Fast Swing Rate across three transformation methods: Original (untransformed), the Box-Cox transformation (MLE-derived λ), and the log transformation ($\lambda = 0$).

In the original scale plot (left), the scatterplot exhibits significant dispersion around the diagonal line, particularly for extreme values, indicating greater prediction errors. Furthermore, the uneven spread of points suggests violations of key regression assumptions, such as homoscedasticity and linearity.

The Box-Cox transformation plot (center) substantially reduces the dispersion while still retaining meaningful deviation for extreme observations. For instance, the largest Fast Swing Rate case (78.0) shows a studentized residual close to zero under log transformation (|rstudent| = 0.18), but remains noticeably higher under Box−Cox (|rstudent| = 1.07). A similar pattern holds for other extreme observations, suggesting that Box−Cox mitigates distortion of outlier influence without overcompressing them.

In contrast, the log transformation plot (right) tightly clusters the points along the diagonal, indicating high predictive accuracy. However, this comes at the cost of compressing the data scale, particularly diminishing the relative influence of extreme values. This compression risks oversimplifying the variability inherent in sports data, potentially obscuring meaningful insights about rare or high-variance events. These results align with the findings in Table 2, which highlight the trade-offs between transformations. While the log transformation achieves the best statistical fit, the Box-Cox transformation provides a better balance between predictive accuracy and interpretability,

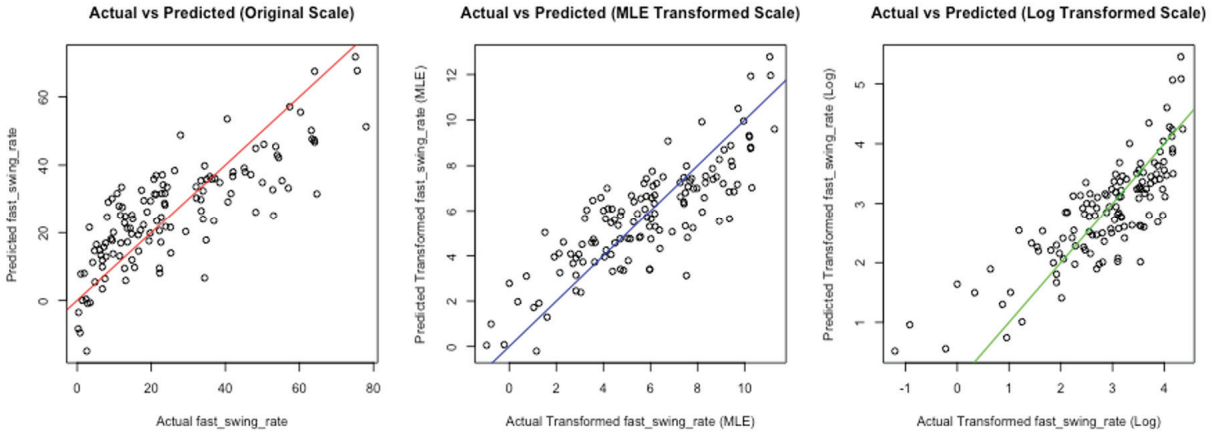

Figure 2. Actual vs. predicted values for *fast swing rate* regression models with original (untransformed), Box-cox transformation (MLE), and log transformation

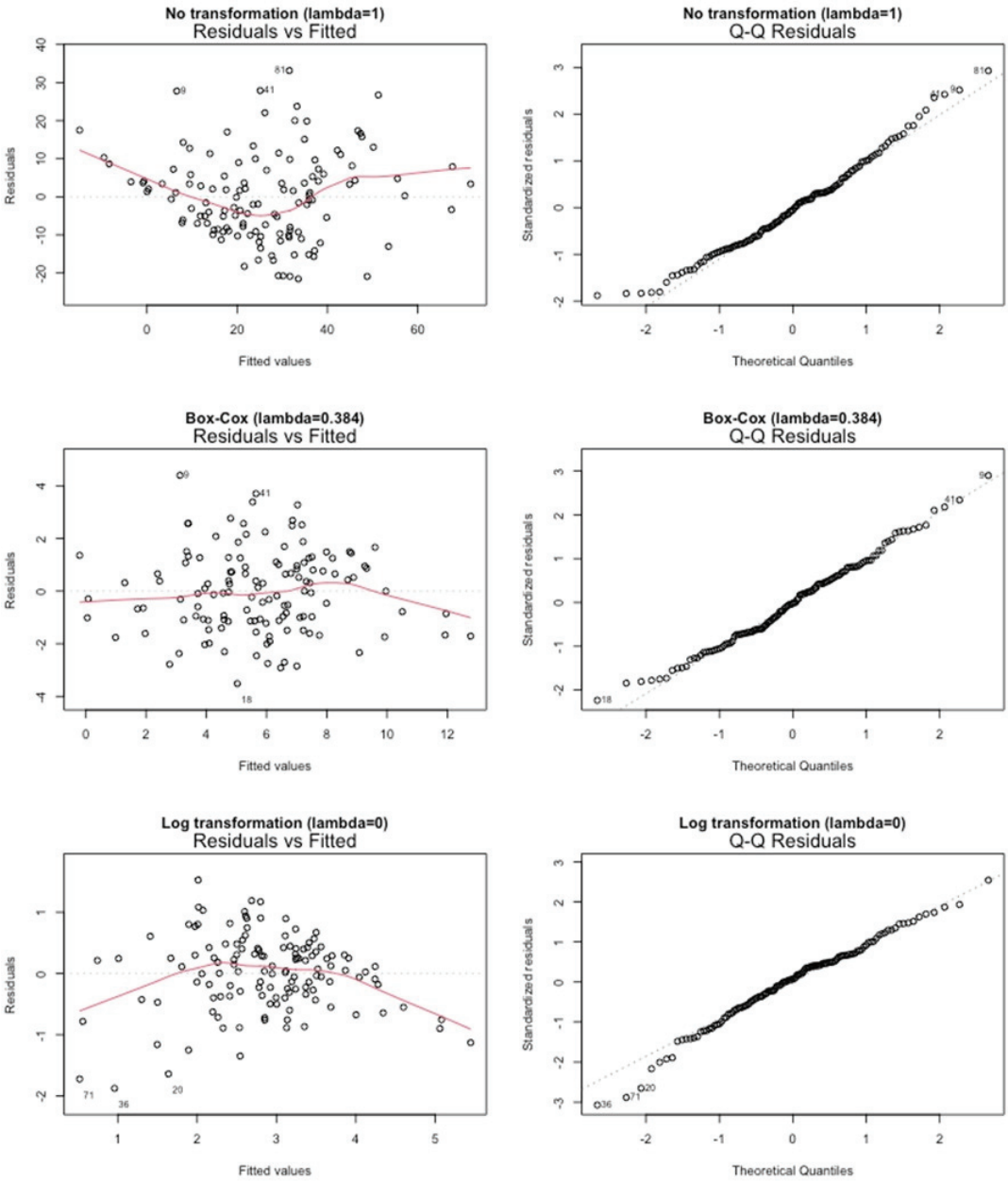

Figure 3. Diagnostic Plots for Residual Analysis of the MLB dataset Across Different Transformations

preserving the nuances of the data that are essential for meaningful analysis in sports contexts. These findings suggest that while transformations enhance prediction accuracy and model reliability, their applicability varies depending on the context of sports analytics.

The diagnostic plots in Figure 3 illustrate the residuals for regression models under three different transformations of the dependent variable: no transformation (λ=1), Box-Cox transformation (λ =0.384), and log transformation (λ=0). Each transformation is evaluated using two metrics: Residuals vs Fitted Values and Q-Q Plots of Residuals.

In the original model (λ=1), the residuals vs fitted

plot shows a distinct curvature and uneven spread, indicating violations of linearity and homoscedasticity. Furthermore, the Q-Q plot demonstrates significant deviations from the diagonal line, particularly at the tails, suggesting that the residuals are not normally distributed. These patterns highlight substantial issues with regression assumptions in the untransformed model.

The Box-Cox transformation model (λ=0.384) exhibits marked improvements in residual behavior. The residuals vs fitted plot shows a more random distribution of residuals around the horizontal axis, with reduced curvature and improved homoscedasticity. Additionally, the Q-Q plot indicates that the residuals align more closely with the diagonal line, demonstrating an enhanced approximation of normality. These improvements suggest that the Box-Cox transformation effectively addresses the skewness in the original data while preserving the relative scale of the values.

In the log-transformed model (λ=0), the residuals vs fitted plot shows a more consistent pattern compared to the untransformed model, but minor curvature persists, indicating slight departures from linearity. The Q-Q plot shows good alignment with the diagonal line in the center but reveals minor deviations at the tails. While the log transformation improves normality, its effect is less balanced compared to the Box-Cox transformation.

Overall, these diagnostic plots suggest that the Box-Cox transformation provides the best balance between improving regression assumptions and maintaining the interpretability of the data. The log transformation, while effective in normalizing the data, compresses the scale of extreme values, which can diminish the interpretability of important outliers—an essential consideration in sports analytics, where extreme values often signify meaningful insights. The untransformed model, by contrast, demonstrates the poorest fit, underscoring the necessity of transformations for addressing non-normality and improving model assumptions. These findings highlight the Box-Cox transformation as a practical and statistically robust approach for handling non-normal data in sports contexts.

LPGA Dataset

Table 3 presents a summary of the LPGA dataset, including means, standard deviations, skewness, kurtosis, and Shapiro-Wilk test results for the dependent and independent variables. The dependent variable, totPrize, exhibits substantial and extreme skewness (2.73) and kurtosis (13.04), indicating a highly right-skewed distribution with heavy tails. This is further supported by the Shapiro-Wilk test ($W = 0.71$, $p < .001$), which confirms a significant deviation from normality. In contrast, the independent variables—driveDist, fairPct, and avePutts—demonstrate relatively low skewness and kurtosis, with Shapiro-Wilk test results indicating no significant departures from normality ($p > .05$).

Figure 4 illustrates the distribution of the dependent variable (totPrize) under the three conditions: Original (untransformed) (λ=1), Box-Cox transformation with MLE-derived λ=0.10, and log transformation (λ=0). The log transformation appears to achieve a distribution closer to normality, as indicated by reduced skewness and kurtosis values.

Conversely, the Box-Cox transformation achieves balance by reducing skewness while preserving the relative scale of extreme values. While the visualized distributions suggest improvements in normality, the

Table 3. Descriptive Statistics and Normality Tests for LPGA Dataset

Variables	Mean	Std.	Skewness	Kurtosis	*Shapiro-Wilk*	*p*-value
totPrize	522580.73	666704.20	2.73	13.04	0.71	<.001
driveDist	257.14	9.32	-0.01	2.57	0.993	0.745
fairPct	73.67	5.97	-0.08	2.86	0.995	0.885
avePutts	30.08	0.6	0.14	3.17	0.993	0.712

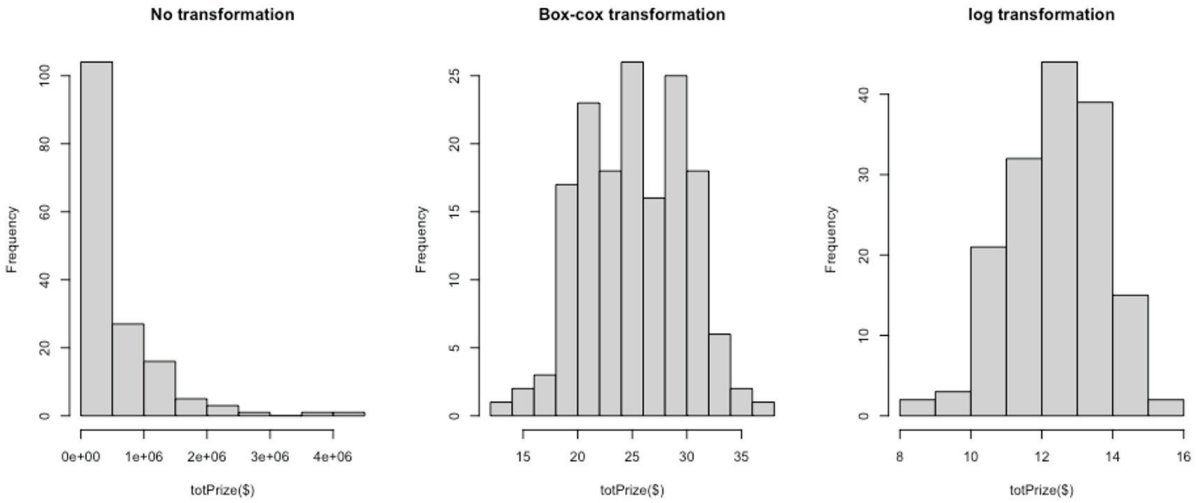

Figure 4. Histogram of totPrize for the LPGA Dataset

necessity for further regression modeling with Box-Cox transformation derives from its potential to better satisfy regression assumptions such as homoscedasticity and linearity, which are critical for reliable and interpretable models.

To further evaluate the effects of these transformations on regression model for the 158 observations, Table 4 presents a detailed comparison of regression models for the LPGA dataset across three transformation methods: Original (untransformed) (λ=1), the Box-Cox transformation using the MLE-based optimal λ=0.10, and the log transformation (λ=0). The R^2 indicate an improvement in explained variance after transformation, with both the Box-Cox transformation and the log transformation achieving R^2>0.4, compared to 0.2501 for the untransformed model. However, the AIC scores suggest that the log-transformed model (λ=0) achieves the best model fit, with the lowest AIC value of 463.11,

Table 4. Comparison of regression models for LPGA dataset across different transformations

	λ=1 (Original)		λ=0.10 (Box-Cox transformation)		λ=0 (log transformation)	
	Estimate (*p*-value)	Std. Error	Estimate (*p*-value)	Std. Error	Estimate (*p*-value)	Std. Error
β0	1456669 (0.633)	3044153	26.17 (0.164)	18.69	12.73 (0.02)	5.37
β1 (driveDist)	30936 (<.001)	6281	0.272 (<.001)	0.038	0.078 (<.001)	0.01
β2 (fairPct)	44505 (<.001)	9867	0.426 (<.001)	0.059	0.123 (<.001)	0.02
β3 (avePutts)	-404481 (<.001)	78132	-3.41 (<.001)	0.479	-0.978 (<.001)	0.14
R^2	0.2501		0.4072		0.4085	
AIC	4649.492		857.284		463.106	
Relative RMSE	1.101		0.140		0.082	

Note: dependent variable=totPrize. All regression coefficients are reported on the scale of the transformed dependent variable. The RMSE scaled by the mean of the dependent variable to allow comparison across transformations.

compared to 857.28 for the Box-Cox model and 4649.49 for the untransformed model. The relative RMSE values also underscore improved predictive accuracy after transformation. The log transformation achieves the lowest RMSE (0.082), followed by the Box-Cox transformation (0.140), while the untransformed model exhibits the highest RMSE (1.101).

All independent variables maintained their statistical significance ($p < .001$) and directional effects across all transformation methods, indicating that the relationships between predictors and the dependent variable totPrize remained stable despite adjustments in scale. The coefficients decreased in magnitude across transformations, with the largest values observed in the untransformed model and progressively smaller values in the Box-Cox and log-transformed models. However, this reduction in coefficient size reflects changes in the scale of measurement introduced by the transformations rather than a diminished influence of the predictors. The consistency of these relationships across models underscores the robustness of the findings. While the transformations adjusted the scale of the coefficients, their effects and significance remained stable.

The relationship between actual and predicted values across the three transformation methods is visualized in Figure 5. In the first plot (original scale), the predicted values deviate significantly from the diagonal line, especially for extreme values, indicating a lack of model fit. In the second plot (MLE-transformed scale), the predicted values align more closely with the diagonal, reflecting an improved fit after addressing skewness through the Box-Cox transformation. The third plot (log-transformed scale) demonstrates the most linear alignment with the diagonal line, showing that the log transformation effectively regularized the scale of the dependent variable. However, the log transformation's tendency to overly diminish the relative impact of extreme values, particularly evident in sports data, may limit its interpretability in contexts where outliers carry critical importance. These visualizations reinforce the nuanced trade-offs between statistical performance and contextual relevance across transformation methods.

The summary statistics presented in the previous table 4 provide a quantitative overview of the regression models across different transformations. However, these numerical summaries alone cannot fully capture the extent to which key regression assumptions, such as homoscedasticity and normality, are satisfied. To address this, diagnostic plots were employed to visually evaluate the residual patterns and normality of residuals for the three transformations: Original (untransformed; λ=1), MLE-based Box-Cox transformation (λ=0.10), and log transformation (λ=0) (Figure 6).

The Residuals vs Fitted plot for the untransformed model (λ=1) revealed a systematic pattern and uneven spread of residuals, indicating a violation of the homoscedasticity assumption. The residuals showed greater variance at the extremes of the fitted values,

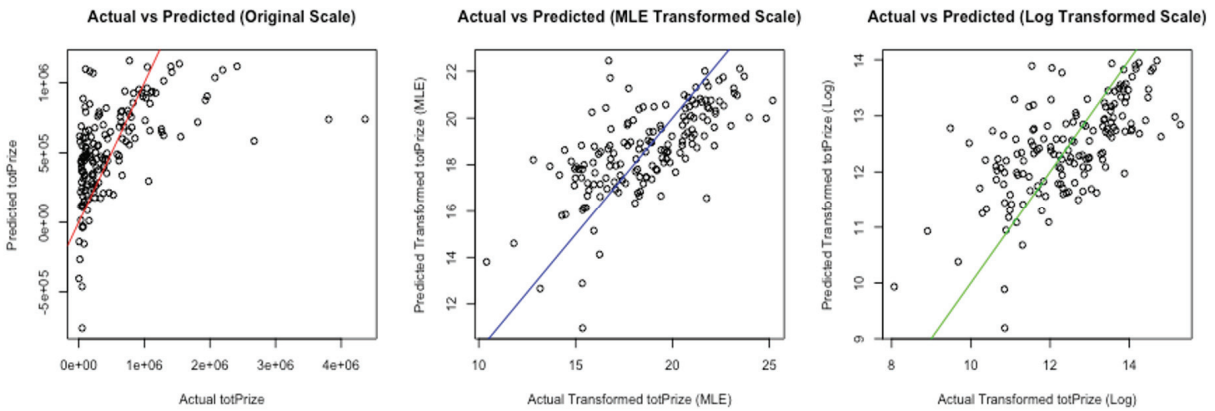

Figure 5. Actual vs. predicted values for LPGA totPrize models

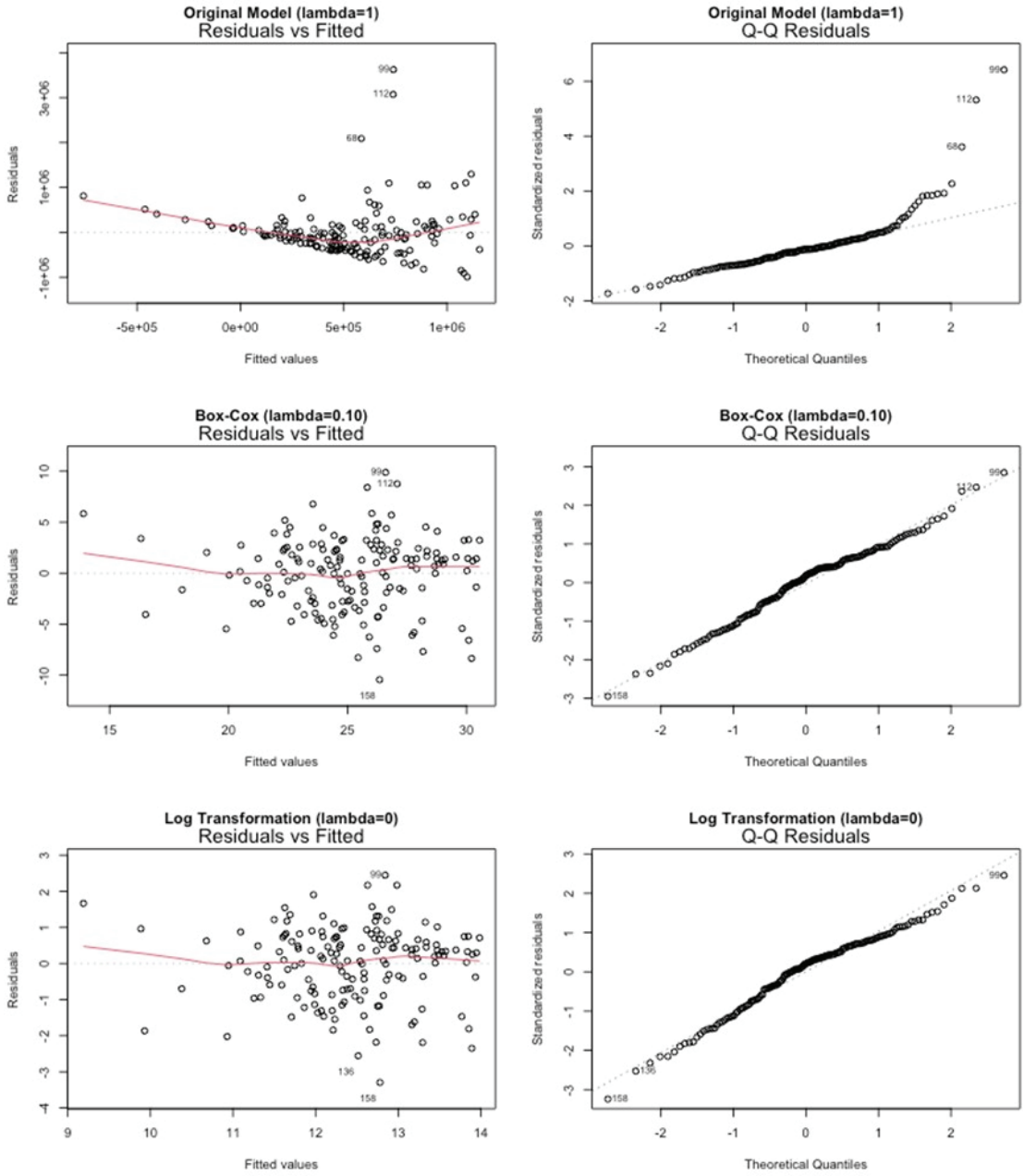

Figure 6. Diagnostic plots for residual analysis of the LPGA dataset across different transformation

suggesting that the model was not adequately capturing the variance structure in the data. In contrast, the MLE-based transformation (λ=0.10) reduced the systematic pattern and produced a more consistent spread of residuals across the fitted values, reflecting improved homoscedasticity. The log transformation (λ =0) also exhibited uniform residual variance, though the effect of extreme values was notably diminished.

The Q-Q Residuals plot for the untransformed model showed substantial deviations from the theoretical normal distribution line, particularly in the tails, suggesting that the residuals were heavily influenced by outliers. The Box-Cox transformation (λ=0.10) substantially reduced this deviation, bringing the

residuals closer to normality. The log transformation (λ=0) achieved the highest alignment with the theoretical line, effectively normalizing the residuals. However, the strong suppression of extreme values by the log transformation may hinder interpretability, particularly in sports analytics, where outliers often carry significant contextual meaning.

Overall, the comparison of diagnostic plots demonstrates that the MLE-based transformation strikes a balance between improving regression assumptions and preserving the interpretability of extreme values. While the log transformation provides the highest degree of normality, its tendency to overly diminish the influence of extreme values limits its applicability in contexts where such values are analytically meaningful.

Discussion

Advancements in ICT technology have transformed sports data from simple frequency-based records to vast repositories of big data. This shift has introduced diverse data distributions that often deviate from normality, posing challenges to regression analysis—a widely used method for examining linear relationships between variables. As Cooper et al. (2007) and Vagenas et al. (2018) highlighted, addressing non-normality is critical for ensuring the reliability and interpretability of sport analytical outcomes. Their work demonstrates the importance of methodological adaptations when traditional assumptions about data distribution are violated.

Despite the necessity of addressing non-normality, sports analytics has relied heavily on log transformations to address non-normality (Nevill & Atkinson, 1997; Reid et al., 2010; Atkinson & Batterham, 2012). While log transformations are effective in meeting statistical assumptions, they often compress extreme values excessively, potentially reducing their interpretive value. This limitation is critical in sports contexts where outliers often represent significant performance metrics or strategic insights. For instance, Lionel Messi's record-setting 50 goals in the 2011–12 La Liga season reflect not just statistical anomalies but also tactical superiority and unique team dynamics. Over-transforming such data risks obscuring its analytical and contextual significance. Similarly, in golf, extraordinary prize earnings in a single tournament can indicate unique strategies or exceptional performances, and in baseball, extreme values such as high strikeout rates or exceptional hit distances often reflect key player characteristics.

Preserving the intrinsic meaning of outliers is essential in sports data analysis to ensure that variability is not overly suppressed, allowing meaningful interpretation within the context of performance evaluation. This balance between statistical rigor and contextual relevance is particularly critical in sports analytics. In line with this, Empacher et al. (2023) emphasized that outliers in sports data are not merely statistical noise but carry critical information about game contexts or meaningful outcomes. Suppressing or removing these outliers can result in the loss of valuable insights during data interpretation.

The findings of the current study revealed that both transformations enhanced model fit compared to untransformed models. Specifically, the Box-Cox transformation recorded higher R^2 values than the log transformation in baseball data and produced comparable R^2 values in golf data, demonstrating its effectiveness in improving explanatory power.

The coefficient of determination (R^2) is calculated as $R^2 = 1 - (SSR/SST)$, where SSR represents the residual sum of squares, and SST represents the total sum of squares. The finding that the Box-Cox transformation recorded higher or similar R^2 values suggests its ability to flexibly adjust data distributions while effectively capturing data variability within the model, minimizing distortion. This highlights its advantage over the log transformation, which may compress data excessively, potentially failing to reflect some variability in the model.

In terms of model simplicity, the log transformation achieved the lowest AIC among all models. AIC, calculated as $AIC = -2 * \ln(L) + 2k$, where L represents the maximum likelihood and k the number of parameters, assesses the balance between model complexity and explanatory power. The lower AIC of

the log-transformed model suggests that it effectively normalized the data while maintaining a simple structure. However, this simplicity may come at the expense of reducing the contribution of outliers, as observed in prior research (Hoaglin & Velleman, 1995; Baesens et al., 2009; Khakifirooz et al., 2021), which cautioned that log transformations could diminish the interpretability of critical extreme values in data. This aligns with the present study's findings, emphasizing the need for careful consideration when applying log transformations.

The analysis of predictive accuracy, as measured by RMSE, provided further insights into the trade-offs between transformation methods. While empirical results indicate that the log transformation achieved the lowest RMSE, suggesting reduced prediction errors, this finding requires a nuanced interpretation within the context of sports data. The lower RMSE in log-transformed models often stems from the aggressive compression of informative outliers—such as superstar athletes or extreme performance cases—which can create a statistical illusion of higher predictive accuracy by artificially minimizing residual variance.

As observed in our results, the extreme case of a 78.0 Fast Swing Rate yielded a studentized residual close to zero under log transformation (|rstudent| = 0.18), whereas it remained meaningfully higher under Box−Cox (|rstudent| = 1.07). This disparity, also illustrated in the "Actual vs Predicted" plots (Figures 2 and 5), indicates that the log transformation squashes unique performance signals to achieve tighter clustering. In contrast, the Box-Cox transformation mitigates the distortion of outlier influence without overcompressing them, prioritizing contextual integrity over the mere minimization of numerical error. Therefore, despite the slightly higher RMSE, the Box-Cox transformation serves as a more valid alternative for sports analytics by preserving the variability that defines elite performance.

Residual and Q-Q plots (Figures 3 and 6) reinforced these findings. The original models showed significant deviations from normality and heteroscedasticity, while the Box-Cox transformation improved these assumptions without diminishing the relative importance of extreme values. Although the log transformation further enhanced the normality of residuals, its compression of data raises concerns about potential information loss.

Beyond statistical validation, the Box-Cox transformed models offer actionable insights for practitioners in baseball and golf. For the MLB Fast Swing Rate model, the results indicate that while Average Exit Velocity ($\beta_3 \approx 0.87$) is a significant predictor, Average Swing Length ($\beta_2 \approx 2.12$) exerts an even more dominant positive influence. This suggests that achieving a high swing rate is not merely a function of raw power; rather, it requires a swing mechanics that ensures a sufficient acceleration zone (swing length) before impact. Furthermore, the relatively small coefficient for Age ($\beta_1 \approx 0.10$ implies that elite hitters may mitigate age-related declines in swing speed through disciplined physical conditioning and technical adjustments.

In the LPGA total prize money (totPrize) model, the findings underscore the paramount importance of short-game proficiency. Average Putts (avePutts, $\beta_3 \approx -3.41$) emerged as the most powerful negative predictor, meaning that reducing even a single putt per round has a far greater relative impact on earnings than increasing driving distance. Additionally, the positive influence of Fairway Percentage (fairPct, $\beta_2 \approx 0.43$) compared to Driving Distance (driveDist, $\beta_1 \approx 0.27$) suggests that a strategy prioritizing stability and accuracy may be more economically efficient for professional golfers than an aggressive focus on distance alone. By utilizing the Box-Cox transformation, these models preserve the impact of top-tier performers while providing a reliable framework for identifying these key performance drivers.

These findings underscore the critical balance required in selecting transformation methods for sports data analysis. While the log transformation effectively normalizes data distributions and optimizes regression assumptions, its tendency to compress extreme values can limit its interpretive value in contexts where such values are essential for understanding performance and strategy. In contrast, the Box-Cox transformation provides a robust alternative by addressing skewness and improving normality without excessively

diminishing the impact of outliers. This approach not only enhances the explanatory power of regression models but also preserves the variability and contextual significance of extreme values, which are often pivotal in sports analytics.

In sports contexts, outliers frequently embody extraordinary performances, unique tactical decisions, or game-defining moments, making their accurate representation critical for meaningful analysis. By maintaining the integrity of these extreme values, the Box-Cox transformation aligns better with the dual demands of statistical rigor and practical applicability. As demonstrated in this study, this balance allows for a more nuanced interpretation of sports data, ensuring that analytical outcomes remain relevant and actionable within the competitive and strategic landscapes of sports.

Limitations & Future Directions

While this study demonstrates the utility of Box-Cox and log transformations in addressing the non-normality of sports data, several limitations and areas for improvement should be acknowledged. First, both transformations alter the original scale of the data, necessitating back-transformation for interpretation and estimation. This process can introduce re-transformation bias, leading to discrepancies between predictions made on the transformed scale and their counterparts on the original scale (Manning, 1998). Specifically, the mean values calculated in transformed and back-transformed data may differ, particularly for datasets with highly skewed distributions or non-normal residuals (Asuero & Bueno, 2011). For instance, in sports contexts, the bias may obscure subtle yet meaningful variations in player performance metrics when interpreting back-transformed results. While this issue is less problematic for symmetric data with minimal outliers and homoscedastic residuals, such conditions are rare in sports data, where extreme values often hold critical contextual significance. Researchers must carefully evaluate these trade-offs, balancing statistical rigor with the interpretability of results, particularly in datasets where outliers represent meaningful performance metrics.

Second, the Box-Cox transformation is limited to positive data values. Although the datasets in this study contained only positive dependent variables, datasets with negative values require an Offset Addition approach. This method involves adding a constant to all data points, shifting the minimum value above zero to enable the transformation (Huang et al., 2023). However, the choice of offset is critical. An excessively large or small constant can distort the original data distribution or obscure meaningful patterns (Riani et al., 2023). Researchers must carefully balance the need for transformation with preserving the interpretive integrity of the data.

Third, this study focused on datasets from baseball and golf, sports where ICT technologies are extensively used for performance tracking. These datasets provided clear examples of non-normal distributions. However, sports with high real-time variability, such as soccer or basketball, may pose unique challenges due to the dynamic nature of play. For instance, Rein and Memmert (2016) highlight that temporal dependencies in soccer datasets complicate traditional analytical approaches, emphasizing the need for adaptable transformation methods capable of addressing the dynamic strategies, situational contexts in such sports. Future studies should explore the applicability of transformation methods in these contexts, taking into account factors such as temporal variability and game-specific situational dynamics.

Fourth, it is important to note that the findings of this study are most applicable to positive continuous variables, such as swing rates, driving distances, and prize money. While Box-Cox and log transformations work well for these types of data, other common sports metrics—such as counts (e.g., goals or fouls) or binary outcomes (e.g., win/loss)—might require different analytical methods. For these variables, Generalized Linear Models (GLMs) could be better suited than a standard linear regression with transformations. Future research should compare these different models to identify the most robust approach for various types of sports data.

Finally, while this study highlights the strengths of

the Box-Cox and log transformations, alternative methods warrant exploration. For example, Yeo-Johnson transformations handle both positive and negative values without requiring offset addition, offering flexibility for datasets with mixed data ranges. Moreover, machine learning-based normalization techniques, which adapt dynamically to complex data characteristics, hold promise for addressing non-normality in modern sports analytics. By integrating these methods and expanding the scope of analysis to include diverse sports and contexts, future research can enhance the precision, interpretability, and practical utility of transformation approaches in sports analytics.

Acknowledgments

We thank the anonymous reviewers for their careful review and constructive suggestions.

Author Contributions

Conceptualization: Soowoong Hwang
Data curation: Minseo Kim
Formal analysis: Seokyong Lee
Investigation: Seokyong Lee, Minseo Kim
Project administration: Soowoong Hwang
Writing-original draft prepartion: Seokyong Lee
Writing -review and editing: Soowoong Hwang

Conflict of Interest

The authors declare no conflict of interest.

References

Amendolara, A., Pfister, D., Settelmayer, M., Shah, M., Wu, V., Donnelly, S., Johnston, B., Peterson, R., Sant, D., Kriak, J., & Bills, K. (2023). An overview of machine learning applications in sports injury prediction. *Cureus,* **15(9).** https://doi.org/10.7759/cureus.46170

Ardagna, C. A., Ceravolo, P., & Damiani, E. (2016). Big data analytics as-a-service: Issues and challenges. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 3638-3644). IEEE. https://doi.org/10.1109/BigData.2016.7841029

Asuero, A. G., & Bueno, J. M. (2011). Fitting straight lines with replicated observations by linear regression. IV. Transforming Data. *Critical Reviews in Analytical Chemistry*, **41(1)**, 36-69. https://doi.org/10.1080/10408347.2010.523589

Atkinson, A. C., Riani, M., & Corbellini, A. (2021). The box-cox transformation: Review and extensions. *Statistical Science*, **36(2)**, 239-255 https://doi.org/10.1214/20-STS778

Atkinson, G., & Batterham, A. M. (2012). The use of ratios and percentage changes in sports medicine: time for a rethink?. *International Journal of Sports Medicine*, **33(07)**, 505-506. https://doi.org/10.1055/s-0032-1316355

Baesens, B., Mues, C., Martens, D., & Vanthienen, J. (2009). 50 years of data mining and OR: upcoming trends and challenges. *Journal of the Operational Research Society*, **60**, S16-S23. https://doi.org/10.1057/jors.2008.171

Bai, Z., & Bai, X. (2021). Sports big data: management, analysis, applications, and challenges. *Complexity*, **2021(1)**, 6676297. https://doi.org/10.1155/2021/6676297

Balague, N., Torrents, C., Hristovski, R., Davids, K., & Araújo, D. (2013). Overview of complex systems in sport. *Journal of Systems Science and Complexity*, **26(1)**, 4-13. https://doi.org/10.1007/s11424-013-2285-0

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological),* **26(2)**, 211-243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature News*, **538(7623)**, 20. https://doi.org/10.1038/538020a

Cooper, S. M., Hughes, M., O'Donoghue, P., & Nevill, M. A. (2007). A simple statistical method for assessing the reliability of data entered into sport performance analysis systems. *International Journal of Performance Analysis in Sport*, **7(1)**, 87-109. https://doi.org/10.1080/24748668.2007.11868390

Draper, N.R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). John Wiley & Sons.

Empacher, C., Kamps, U., & Volovskiy, G. (2023). Statistical prediction of future sports records based on record values. *Stats*, **6(1)**, 131-147. https://doi.org/10.3390/stats6010008

Hoaglin, D. C., & Velleman, P. F. (1995). A critical look at some analyses of major league baseball salaries. *The American Statistician*, **49(3)**, 277-285. https://doi.org/10.1080/00031305.1995.10476165

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708). https://doi.org/10.48550/arXiv.1608.06993

Huang, Z., Zhao, T., Lai, R., Tian, Y., & Yang, F. (2023). A comprehensive implementation of the log, Box-Cox and log-sinh transformations for skewed and censored precipitation data. *Journal of Hydrology*, **620**, 129347. https://doi.org/10.1016/j.jhydrol.2023.129347

Khakifirooz, M., Tercero-Gómez, V. G., & Woodall, W. H. (2021). The role of the normal distribution in statistical process monitoring. *Quality Engineering*, **33(3)**, 497-510. https://doi.org/10.1080/08982112.2021.1909731

Kumar, G. S., Kumar, M. D., Reddy, S. V. R., Kumari, B. S., & Reddy, C. R. (2024). Injury prediction in sports using artificial intelligence applications: A brief review. *Journal of Robotics and Control (JRC)*, **5(1)**, 16-26. https://doi.org/10.18196/jrc.v5i1.20814

Lee, D. K. (2020). Data transformation: A focus on the interpretation. *Korean Journal of Anesthesiology*, **73(6)**, 503-508. https://doi.org/10.4097/kja.20137

Manning, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, **17(3)**, 283-295. https://doi.org/10.1016/S0167-6296(98)00025-3

Marimuthu, S., Mani, T., Sudarsanam, T. D., George, S., & Jeyaseelan, L. (2022). Preferring Box-Cox transformation, instead of log transformation to convert skewed distribution of outcomes to normal in medical research. *Clinical Epidemiology and Global Health*, **15**, 101043. https://doi.org/10.1016/j.cegh.2022.101043

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. International *Journal of Data Science and Analytics*, **5**, 213-222. https://doi.org/10.1007/s41060-017-0093-7

Nevill, A. M., & Atkinson, G. (1997). Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *British Journal of Sports Medicine*, **31(4)**, 314-318. https://doi.org/10.1136/bjsm.31.4.314

Osborne, J. W., & Waters, E. (2002). Four Assumptions of Multiple Regression That Researchers Should Always Test. Practical Assessment, *Research & Evaluation,* **8(2)**, 2. https://doi.org/10.7275/r222-hv23

Osborne, J. (2010). Improving your data transformations:

Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, **15(1).** https://doi.org/10.7275/qbpc-gk17

Reid, M., McMurtrie, D., & Crespo, M. (2010). The relationship between match statistics and top 100 ranking in professional men's tennis. *International Journal of Performance Analysis in Sport*, **10(2)**, 131-138. https://doi.org/10.1080/24748668.2010.11868509

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus*, 5, 1-13. https://doi.org/10.1186/s40064-016-3108-2

Riani, M., Atkinson, A. C., & Corbellini, A. (2023). Automatic robust Box-Cox and extended Yeo-Johnson transformations in regression. *Statistical Methods & Applications*, **32(1)**, 75-102. https://doi.org/10.1007/s10260-022-00640-7

Rinehart, K. L. (2009). The economics of golf: An investigation of the returns to skill of PGA Tour golfers. *Major Themes in Economics*, **11(1)**, 57-70. https://scholarworks.uni.edu/mtie/vol11/iss1/6

Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American statistical association*, **63(324)**, 1343-1372. https://doi.org/10.1080/01621459.1968.10480932

Shi, P. J., Hu, H. S., & Xiao, H. J. (2013). Logistic regression is a better method of analysis than linear regression of arcsine square root transformed proportional diapause data of Pieris melete (Lepidoptera: Pieridae). *Florida Entomologist*, **96(3)**, 1183-1185. https://doi.org/10.1653/024.096.0361

Zhang, T., & Yang, B. (2017). Box-cox transformation in big data. *Technometrics,* **59(2)**, 189-201. https://doi.org/10.1080/00401706.2016.1156025

Zhou, H., & Zou, H. (2024). The nonparametric Box–Cox model for high-dimensional regression analysis. *Journal of Econometrics*, **239(2)**, 105419. https://doi.org/10.1016/j.jeconom.2023.01.025