Modeling Performance in Professional Golfers' Association of America Tournaments

Sophia Yin

The Westminster Schools, Atlanta, GA

Abstract

This study models the performance of professional golfers on the Professional Golfers' Association (PGA) Tour using hierarchical models. Continuous outcome variable, such as score relative to par, is analyzed with hierirchical regression model for normally distributed responses, while binary outcome, such as finishing at or under par versus over par, is modeled using a logistic hierarchical model. This approach captures both individual player effects and tournament-level variability, providing insights into factors influencing success on the professional golf circuit.

Keywords: golf, Professional Golfers' Association (PGA), player's performance, hierarchical model

1 Introduction

1.1 Background

Golf has been one of the most widely played and followed sports for several centuries. The objective of the game is for players to use a variety of clubs to strike a ball into a sequence of holes in as few strokes as possible. The sport is played on expansive outdoor courses, typically consisting of 18 holes, each comprising a teeing area, a fairway, and a putting green where the hole is located.

A golfer's score in any given round is influenced not only by individual skill level but also by a range of external factors, including course conditions, weather, and physical attributes of the player. Understanding the effects of these factors on scoring is essential for predicting tournament outcomes and for comparing performance across rounds played under varying conditions.

Unlike many other sports, golf is not contested on a standardized playing surface. Golf courses differ substantially in characteristics such as total yardage, course design, and par, all of which can affect scoring outcomes. In addition, the geographic attributes of a course, including latitude, longitude, and altitude, may influence player performance.

Professional golf tournaments typically consist of four rounds played over four consecutive days. As a result, scoring is affected not only by static course characteristics but also by day-to-day variations, particularly weather conditions such as temperature, precipitation, and wind speed. Performance may also vary by tournament round, as competitive pressure and fatigue can increase as the event progresses.

The Professional Golfers' Association (PGA) Tour is the premier men's professional golf tour worldwide. Tournaments are generally held weekly, beginning on Thursday and concluding on Sunday, with select events designated as PGA Tour signature events. Men's professional golf also features four major championships held annually: The Masters, the PGA Championship, the U.S. Open, and The Open Championship. Entry into major championships is typically determined by performance in other professional events. Although amateur golfers may occasionally qualify to compete in professional tournaments, they are excluded from this study.

1.2 Literature Review

The quantitative analysis of professional golf performance has evolved substantially over the past two decades, driven by improved data availability and advances in statistical modeling. Early research focused primarily on aggregate scoring behavior and broad performance determinants, while more recent studies emphasize shot-level data, hierarchical models, and component-based measures of skill. Collectively, this literature seeks to explain variation in player performance, identify the most important contributors to success, and improve predictive accuracy.

One of the foundational strands of the literature models golf scores using classical statistical distributions. James (2007) provides an early systematic treatment of golf performance, applying statistical methods to tournament scores to understand consistency and scoring variability. Similarly, Grober (2008) models PGA Tour scores as Gaussian random variables, arguing that, despite the bounded and discrete nature of golf scores, normal approximations perform reasonably well at the aggregate level. These studies establish a baseline understanding of score distributions but are limited in their ability to capture hole-level heterogeneity and player-specific effects.

Building on this foundation, subsequent work introduces more nuanced modeling approaches. Hardt and Nettleton (2025) propose a hierarchical framework for modeling hole scores using Hardy distributions, allowing for player-, hole-, and course-level variation. Their approach represents a significant methodological advancement by explicitly accounting for the multilevel structure of golf data and addressing limitations of simpler Gaussian models. This shift toward hierarchical and flexible distributions reflects a broader trend in sports analytics toward models that better reflect the underlying data-generating process.

Another major theme in the literature concerns the determinants of professional golf performance. Peters (2008) examines factors influencing PGA Tour success, highlighting the roles of driving distance, accuracy, and putting, though the analysis is constrained by relatively coarse performance metrics. Later studies leverage richer datasets to refine these insights. Broadie (2012) introduces strokes gained metrics, which decompose performance into distinct components such as driving, approach play, short game, and putting. This framework has become central to modern golf analytics, enabling more precise evaluation of player strengths and weaknesses.

Shot-level data further expand predictive and explanatory power. Leahy (2014) and Atwal, Jones, and Stevenson (2021) utilize proprietary PGA Tour ShotLink data to predict professional golfer performance, demonstrating that granular shot information substantially improves model accuracy relative to aggregate statistics. Korpimies (2020) similarly focuses on prediction, applying machine learning and statistical techniques to forecast PGA Tour success, reinforcing the importance of high-resolution data in performance modeling.

Several studies specifically investigate which aspects of the game matter most for winning. Booher et al. (2024) revisit the long-standing debate summarized by the phrase "drive for show, putt for dough," using modern data to reassess the relative importance of driving, approach shots, and putting. Their findings align with Broadie's strokes gained framework, suggesting that ball-striking, particularly approach play, plays a dominant role in determining tournament outcomes. Complementary work by Stöckl, Lamb, and Lames (2012) introduces visualization techniques to assess hole difficulty and its impact on performance rankings, emphasizing the interaction between player skill and course characteristics.

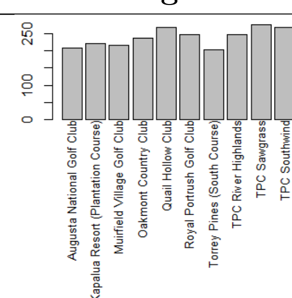
Demographic and career-related factors have also been examined. Tiruneh (2010) analyzes the relationship between age and winning professional golf tournaments, finding evidence of a performance peak followed by a gradual decline. Such results provide context for performance modeling by highlighting systematic patterns beyond pure skill measures.

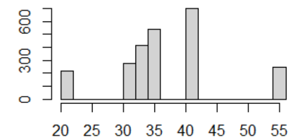
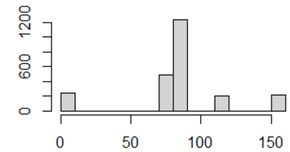
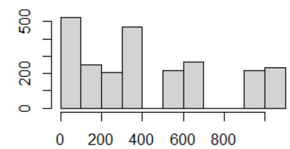
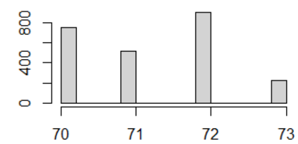
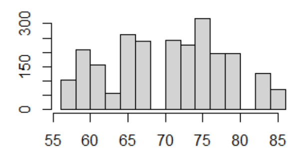
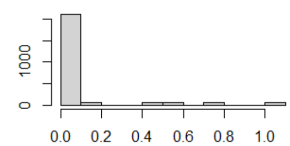
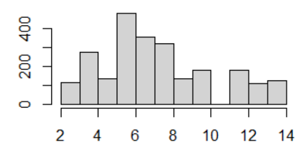
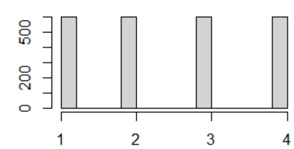
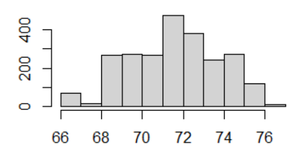
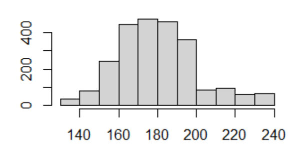
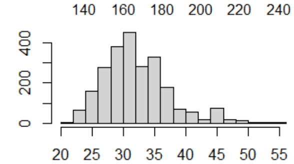
In summary, the literature on golf performance analytics has progressed from simple distributional analyses of scores to sophisticated, data-rich models that decompose skill and account for hierarchical structure. While early studies establish useful baselines, recent work demonstrates the value of shot-level data, strokes gained metrics, and hierarchical modeling frameworks. These advances motivate continued research into flexible statistical models that integrate player, hole, and course effects while maintaining strong predictive performance.

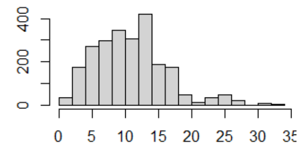
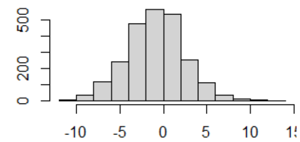
1.3 Data Description

Tournament ranking and score data for all four rounds of each tournament were gathered from the PGA Tour website's individual tournament pages. Player physical characteristics and experience were collected from the PGA Tour website's player bios, and some gaps in data were filled in using the information from the DP World Tour for several players who are also members of the DP Tour. Amateurs and any players with any missing variables (typically height or weight) were omitted. Data from players who were cut from the tournament (eliminated after the second round) and any other players who did not completely finish all four rounds of the tournament were also removed. Two different score outcomes were used: the player's raw score for each round relative to course par (rawdiff) and a binary outcome measuring whether or not the player scored under par for the round (Outcome). Weather data, including average daily temperature, average daily wind speed, and precipitation, were collected from Weather Underground's historical weather records. Below is a description of the variables in the dataset (see Table 1).

Table 1. Description of Variables in the Dataset.

Variable	**Description**	**Histogram**
Course	Name of the golf course	

Latitude	Latitude of course location	
Longitude	Longitude of course location	
Altitude	Altitude of course location	
CoursePar	Par of the course	
AvgTemp	Average temperature for the day (°F)	
Precip	Inches of rainfall for the day	
AvgWindspeed	Average windspeed for the day (mph)	
Player	Name of the player	—
Round	Round of the tournament (1, 2, 3, 4)	
Height	Player height (in)	
Weight	Player weight (lbs)	
Age	2025 – birth year	

YearsPro	2025 − year turned pro	
rawdiff	CoursePar − player's score in round	
Outcome	1 if rawdiff < 0, 0 otherwise	

2 Hierarchical Regression Model: Theory

In the dataset, observations from the same player may be correlated across the four rounds, and players participating in the same tournament may also exhibit correlated performances. To capture such dependencies, a three-level hierarchical model is employed, where rounds are nested within players, and players are nested within tournaments.

2.1 Hierarchical Model for a Normal Response

Suppose that the response variable y_{ijm} is continuous and normally distributed. Data are collected longitudinally at times $t_1, \dots, t_a$ for each of n individuals, grouped into c clusters. For each observation, we record predictor variables $(x_{1ijm}, \dots, x_{kijm})$ and one response y_{ijm}, where $i = 1, \dots, n$, $j = 1, \dots, a$, and $m = 1, \dots, c$. Some predictors may correspond to round-level characteristics (level 1), player-level attributes (level 2), or course-level features (level 3).

A general three-level hierarchical linear model for the normal response is defined as

$$y_{ijm} = \beta_0 + \beta_1\, x_{1ijm} + \cdots + \beta_k\, x_{kijm} + \beta_{k+1}\, t_j + u_{1im} + u_{2im}\, t_j + \tau_{1m} + \tau_{2m}\, t_j + \varepsilon_{ijm},$$

where

$$u_{1im}, u_{2im} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2_{u_1} & \sigma_{u_1u_2}\\ \sigma_{u_1u_2} & \sigma^2_{u_2}\end{pmatrix}\right), \quad \tau_{1m}, \tau_{2m} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2_{\tau_1} & \sigma_{\tau_1\tau_2}\\ \sigma_{\tau_1\tau_2} & \sigma^2_{\tau_2}\end{pmatrix}\right),$$

and

$$\varepsilon_{ijm} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Here, u_{1im} and u_{2im} denote the level-2 random intercept and slope, while τ_{1m} and τ_{2m} represent the level-3 random intercept and slope. All u's are assumed independent of the τ's, and both are independent of the residual errors ε_{ijm}.

The fitted hierarchical model specifies the estimated mean response as

$$\widehat{\mathbb{E}}(y) = \widehat{\beta}_0 + \widehat{\beta}_1 \, x_1 + \cdots + \widehat{\beta}_k \, x_k + \widehat{\beta}_{k+1} \, t.$$

The estimated regression coefficients $\widehat{\beta}_1, \ldots, \widehat{\beta}_{k+1}$ are interpreted as follows:

- If a predictor variable x_1 is numeric, then $\widehat{\beta}_1$ represents the expected change in the estimated mean response $\widehat{\mathbb{E}}(y)$ for a one-unit increase in x_1, holding all other predictors constant.
- If x_1 is an indicator variable, then $\widehat{\beta}_1$ measures the difference in the estimated mean response $\widehat{\mathbb{E}}(y)$ between $x_1 = 1$ and $x_1 = 0$, controlling for the remaining predictors.

2.2 Hierarchical Regression Model for a Binary Response

Using the same notation, a hierarchical logistic regression model for a binary response y_{ijm} is given by

$$\ln\left(\frac{\pi_{ijm}}{1 - \pi_{ijm}}\right) = \beta_0 + \beta_1 \, x_{1ijm} + \cdots + \beta_k \, x_{kijm} + \beta_{k+1} \, t_j + u_{1im} + u_{2im} \, t_j + \tau_{1m} + \tau_{2m} \, t_j,$$

where $\pi_{ijm} = \mathbb{P}(y_{ijm} = 1)$. The fitted model takes the form

$$\ln\left(\frac{\widehat{\pi}}{1 - \widehat{\pi}}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 \, x_1 + \cdots + \widehat{\beta}_k \, x_k + \widehat{\beta}_{k+1} \, t.$$

The estimated coefficients are interpreted as follows. The quantity $\left(\exp\{\widehat{\beta}_1\} - 1\right) \times 100\%$ gives the percent change in the estimated odds $\dfrac{\widehat{\pi}}{1 - \widehat{\pi}}$ in favor of $y = 1$, for a one-unit increase in x_1, with all other variables held constant.

3 Applications and Results

3.1 Hierarchical Model for Binary Outcome

First, we created a binary response variable by dichotomizing players' performance into two categories: "scored under par" and "scored at par or over par." We then modeled the odds of scoring under par, indicating better performance, using a hierarchical binary logistic regression model. The results are presented in Table 2 below. The most complex model that converged included a random intercept at the course level and both a random intercept and a random slope at the player-within-course level. All predictors were standardized prior to analysis (i.e., the mean was subtracted and the result was divided by the standard deviation to obtain a z-score), with the exception of round number.

Table 2. Hierarchical Logistic Regression Results for Binary Performance.

Predictor	**Estimate**	**Std. Error**	z**-value**	p**-value**
Intercept	-0.4365	0.2320	1.881	0.0599
Latitude	-0.2671	0.2768	-0.965	0.3345
Longitude	0.4757	0.2211	2.151	0.0314
Altitude	-0.7346	0.2052	-3.579	0.0003
CoursePar	0.0449	0.3004	0.150	0.8811
AvgTemp	0.0432	0.1825	-0.805	0.4206
Precip	-0.0468	0.0581	0.993	0.3207
AvgWindspeed	-0.4163	0.643	6.473	< 0.0001
Height	0.1152	0.0809	1.424	01544
Weight	-0.0050	0.826	-0.060	09518
Age	-0.2024	0.1718	-1.178	0.2389
YearsPro	0.2163	0.1707	1.267	0.2050
Round	0.0881	0.0480	1.835	0.0665

We write the fitted model as

$$\frac{\widehat{\pi}}{1-\widehat{\pi}} = \exp\Big\{ -0.4365 - 0.2671 \cdot \texttt{Latitude} + 0.4757 \cdot \texttt{Longitude} - 0.7346 \cdot \texttt{Altitude}$$
$$+0.0449 \cdot \texttt{CoursePar} + 0.0432 \cdot \texttt{AvgTemp} - 0.0468 \cdot \texttt{Precip}$$
$$-0.4163 \cdot \texttt{AvgWindspeed} + 0.1152 \cdot \texttt{Height} - 0.0050 \cdot \texttt{Weight} - 0.2024 \cdot \texttt{Age}$$
$$+0.2163 \cdot \texttt{YearsPro} + 0.0881 \cdot \texttt{Round}\Big\}.$$

with the random-effect parameter estimates

$$\hat{\sigma}^2_{u_1} = 2.1972, \quad \hat{\sigma}^2_{u_2} = 0.1317, \quad \hat{\sigma}_{u_1 u_2} = 0.3511, \quad \text{and} \quad \hat{\sigma}^2_{\tau_1} = 0.3366.$$

All these sigmas are statistically significant at the 5% level. Further, in the fitted model, course altitude and average wind speed emerge as highly significant predictors ($p < 0.001$). Longitude is significant at the 5% level, while round reaches significance at the 10% level.

The estimated regression coefficients admit the following interpretations. A one–standard deviation increase in course altitude is associated with a change in the estimated odds of better performance of $\left(\exp -0.7346 - 1\right) \times 100\% = -52.03\%$, corresponding to a 52.03% decrease in the estimated odds of scoring under par. This suggests that higher-altitude courses substantially reduce the likelihood of strong performance, possibly due to environmental factors such as thinner air, altered ball flight dynamics, and reduced physical comfort, all of which may affect shot control and consistency.

Similarly, a one–standard deviation increase in average wind speed leads to a $\left(\exp -0.4163 - 1\right) \times 100\% = -34.05\%$ change in the estimated odds, indicating a 34.05% reduction in the odds of scoring under par. Strong winds increase shot variability and require more technical adjustments, making precise play more difficult and thereby lowering the probability of superior performance.

In contrast, longitude has a positive effect: a one–standard deviation increase corresponds to a $\left(\exp 0.4757 - 1\right) \times 100\% = 60.91\%$ increase in the estimated odds of better performance. This may reflect geographic or regional characteristics correlated with longitude, such as climate patterns, course design styles, or travel-related familiarity, which could systematically influence scoring conditions.

Finally, for each additional round played, the odds increase by $\left(\exp 0.0881 - 1\right) \times 100\% = 9.21\%$. This suggests that players are modestly more likely to perform well as the tournament progresses. Greater familiarity with the course layout, improved strategic adjustments, or heightened competitive focus in later rounds may contribute to this effect.

In summary, the analysis indicates that certain tournament-level characteristics—specifically course altitude, longitude, and round number—as well as weather conditions such as wind speed, are statistically significant predictors of performance. In contrast, none of the player-specific attributes demonstrates statistical significance within the current model.

This outcome may be partially explained by the nature of the response variable, which is binary. At the elite level of competition, players exhibit exceptionally high and relatively homogeneous skill, making it difficult for a binary outcome measure (e.g., under par versus not) to capture subtle differences in individual ability.

To better distinguish among players who already perform at an extraordinary level, a more sensitive performance metric is required. Accordingly, the next phase of the analysis will focus on modeling the raw deviation from course par—that is, the exact margin by which a player finishes above or below par. This continuous outcome provides greater resolution and should allow for a more nuanced assessment of performance differences among top competitors.

3.2 Hierarchical Model for Normal Outcome

Building on the previous analysis, where none of the player-specific attributes were statistically significant in the binary logistic model, we now turn to a more sensitive performance measure: the raw difference between each player's score and the course par. By modeling this continuous outcome, we aim to better capture variation among elite players. The histogram of the raw differences is presented in Figure 1 below.

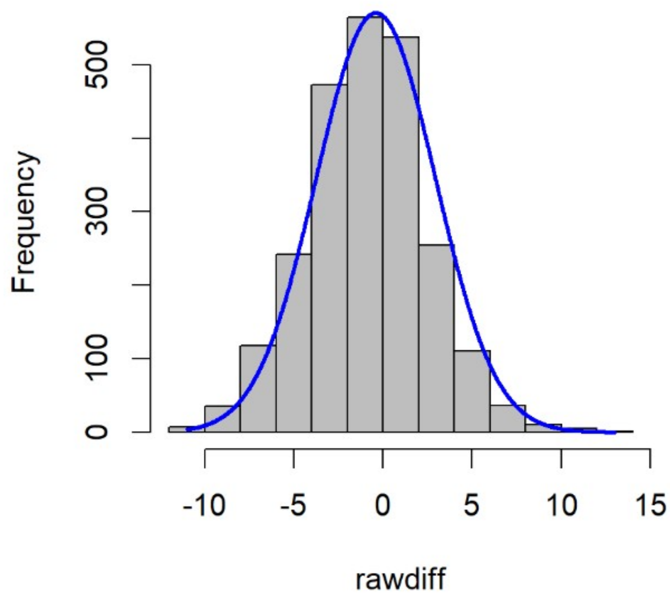


Figure 1. Histogram of raw differences between players' scores and course pars.

As shown in the histogram, the distribution of the raw differences appears approximately normal. This is further assessed by conducting the Wilks test for normality. The resulting p-value is < 0.0001, supporting the null hypothesis that the underlying distribution is normal. Overall, the data fit the normal distribution very well. This pattern may be explained by the Central Limit Theorem. Given the large number of players, and assuming that individual performances are approximately independent and identically distributed, the aggregate distribution of score deviations is expected to approach normality

Therefore, we are well-positioned to fit a hierarchical regression model with a normal response. The results of this model are presented in Table 3 below. Note that the most complex model that converged includes random intercepts only (no random slopes) at the player and course hierarchical levels.

Table 3. Hierarchical Regression for Normally Distributed Raw Differences.

Predictor	**Estimate**	**Std. Error**	***t*-value**	***p*-value**
ntercept	-0.1524	0.4574	-0.3333	0.7389
Latitude	0.1287	0.5851	0.2200	0.8259
Longitude	-0.8253	0.4872	-1.6941	0.0903
Altitude	0.9112	0.4513	2.0190	0.0435
CoursePar	-0.5917	0.5742	-1.0305	0.3028
AvgTemp	-0.1809	0.2002	-0.9034	0.3663
Precip	-0.1021	0.0681	-1.4993	0.1338
AvgWindspeed	0.8437	0.0710	11.8912	0.0000
Height	-0.1274	0.0727	-1.7518	0.0798
Weight	-0.0038	0.0745	-0.0504	0.9598
Age	0.2226	0.1563	1.4244	0.1543
YearsPro	-0.1814	0.1550	-1.1700	0.2420
Round	-0.1091	0.0516	-2.1122	0.0347

The fitted model has the estimated mean raw score difference written as

$$\widehat{\mathbb{E}}(\text{rawdiff}) = -0.1524 + 0.1287 \cdot \texttt{Latitude} - 0.8253 \cdot \texttt{Longitude} + 0.9112 \cdot \texttt{Altitude}$$

$$-0.5917 \cdot \texttt{CoursePar} - 0.1809 \cdot \texttt{AvgTemp} - 0.1021 \cdot \texttt{Precip} + 0.8437 \cdot \texttt{AvgWindspeed}$$

$$-0.1274 \cdot \texttt{Height} - 0.0038 \cdot \texttt{Weight} + 0.2226 \cdot \texttt{Age} - 0.1814 \cdot \texttt{YearsPro} - 0.1091 \cdot \texttt{Round},$$

with the random-effect parameter estimates $\hat{\sigma}^2_{u_1} = 0.1362$ (p-value$= 0.1475$), $\hat{\sigma}^2_{\tau_1} = 1.8801$ (p-value< 0.0001), and $\hat{\sigma}^2 = 7.9684$. Note that the random intercept variance at the course level is not statistically significant, suggesting that, after accounting for the observed course-level covariates (such as altitude, par, and weather conditions), there is little remaining unexplained variability between courses. In contrast, the random intercept variance at the player level is highly significant, indicating substantial residual heterogeneity among players. This means that even after controlling for observed player characteristics (e.g., age, height, years as a professional), players still differ meaningfully in their average performance relative to par.

At a 5% significance level, two predictors are statistically significant: `AvgWindspeed` and `Round`. First, holding all other variables constant, a one–standard-deviation increase in `AvgWindspeed` is associated with an increase of approximately 0.8446 strokes in the estimated mean raw score difference. Since higher raw differences correspond to worse performance relative to par, this indicates that stronger wind conditions significantly worsen predicted performance. This result is consistent with the binary outcome model and aligns with intuition: higher wind speeds require players to adjust aim and club selection, increasing shot difficulty and the likelihood of error.

Second, `Round` was not standardized. Therefore, its coefficient represents the expected change in the mean raw score difference for a one-unit increase in the tournament round. Holding all other variables constant, progressing by one round is associated with a decrease of approximately 0.1091 strokes in the estimated mean raw score difference, indicating a slight improvement in performance as the tournament advances. This may reflect increased familiarity with the course or greater competitive focus in later rounds.

At a 10% significance level, two additional predictors are significant: `Altitude` and `Height`. A one–standard-deviation increase in `Altitude`, holding all other variables constant, is associated with an increase of approximately 0.9112 strokes in the estimated mean raw score difference, implying worse performance at higher-altitude courses. This finding is consistent with the binary outcome model and may reflect environmental challenges associated with altitude.

Finally, a one–standard-deviation increase in `Height` is associated with a decrease of approximately 0.1274 strokes in the predicted raw score difference, indicating a modest improvement in performance. A possible explanation is that taller players may be able to generate greater swing distance, potentially reducing the number of strokes required to complete a hole.

4 Conclusion and Discussion

This study investigated the determinants of professional golf performance using hierarchical regression models for both a binary outcome (scoring under par) and a continuous outcome (raw deviation from course par). By explicitly accounting for the multilevel structure of the data—rounds nested within players and players nested within courses—we were able to separate player-level and course-level sources of variation while controlling for weather and tournament characteristics.

Across both modeling frameworks, environmental and tournament-level variables consistently emerged as important predictors of performance. In particular, course altitude and average wind speed were strong and robust determinants. Higher altitude and stronger wind were associated with worse performance in both the binary and normal models, reinforcing the substantial role of environmental difficulty in shaping scoring outcomes. Longitude also exhibited a statistically significant effect in the logistic model, potentially reflecting broader geographic or regional differences in course setup, climate, or travel demands.

Round number showed evidence of improving performance as tournaments progressed. In the logistic model, later rounds were associated with higher odds of scoring under par, and in the normal model, later rounds were associated with slightly lower raw score deviations. This pattern may reflect course familiarity, strategic adaptation, or increased competitive focus as the tournament advances.

In contrast, player-specific physical and demographic characteristics (height, weight, age, and years as a professional) were generally not statistically significant at conventional levels. Even when marginal effects were detected (e.g., height at the 10% level in the normal model), their magnitude was relatively small. This finding suggests that among elite PGA Tour players—who already represent a highly selected and homogeneous population—observable physical traits explain little additional variation in round-level performance. Instead, performance differences appear to arise more from situational and environmental conditions than from measurable demographic characteristics.

The comparison between the binary and continuous outcomes provides an important methodological insight. The binary model identified only a limited number of significant predictors, likely because dichotomizing performance into "under par" versus "not under par" discards substantial information. The normal hierarchical model, using the full raw score deviation, offered greater sensitivity and revealed additional meaningful effects, such as the impact of altitude and height. This demonstrates the value of modeling performance on a continuous scale when the goal is to detect subtle differences among highly skilled competitors.

The random-effects structure further highlights the multilevel nature of golf performance. In the normal model, significant player-level random intercept variance indicates persistent heterogeneity in baseline performance across players, even after controlling for observable characteristics. Conversely, limited residual course-level variation suggests that once measurable course and weather characteristics are included, much of the between-course variability is explained.

Several limitations should be noted. First, the analysis relies on round-level aggregate data rather than shot-level metrics such as strokes gained, which are known to capture finer components of skill. Second, weather data were aggregated at the daily level and may not fully reflect within-round variability. Third, potential interactions, such as whether certain players are more resilient to wind or altitude, were not explored. Future research could extend this framework by incorporating shot-level data, interaction terms, or dynamic performance measures to better isolate individual skill components.

In conclusion, the findings emphasize that tournament context and environmental conditions play a central role in shaping professional golf performance. While elite players differ in baseline ability, external factors such as wind and altitude substantially influence scoring outcomes. Modeling raw score deviations within a hierarchical framework provides a nuanced and statistically robust approach to understanding these effects, offering both explanatory insight and a foundation for improved predictive modeling in professional golf analytics.

Supplemental Materials

The data and R code are publicly available and can be found in the following GitHub repository:

`https://github.com/syinzr/Modeling-Performance-in-PGA-Tournaments`

Acknowledgments

I would like to thank Dr. Olga Korosteleva, Professor of Statistics at California State University, Long Beach, for her guidance. This article would not have been possible without her support throughout every stage of the project. She has been a constant source of encouragement and has helped me grow my confidence in statistics and research. I would also like to express gratitude towards my parents for inspiring and motivating me throughout this process and encouraging me to pursue my interest in mathematics.

References

[1] James, N. (2007). The Statistical Analysis of Golf Performance, *International Journal of Sports Science & Coaching*, 2(1_suppl), 231-249.
[2] Peters, A. C. (2008). Determinants of Performance on the PGA Tour, *Issues in Political Economy*, 17.
[3] Hardt, M. and D. Nettleton. (2025). A hierarchical approach to modeling golf hole scores with Hardy distributions, *Journal of Quantitative Analysis in Sports*, 21(2), 97-115.
[4] Korpimies, S. (2020). Predicting Players' Success on the PGA-Tour, Master's Thesis,

Aalto University, School of Business, Information and Service Management Degree Program.

[5] Atwal, J., Jones, B., and Stevenson, O. (2021). Improving Analytics in Golf (dissertation).

[6] Leahy, B. (2014). Predicting Professional Golfer Performance Using Proprietary PGA Tour "Shotlink" Data.

[7] Grober, R.D. (2008). PGA Tour Scores as a Gaussian Random Variable. arXiv: Applications.

[8] Broadie, M. (2012). Assessing Golfer Performance on the PGA TOUR. Interfaces 42(2):146-165.

[9] Booher, A., Rao, E., Valentovish, B., and Wang, J. (2024). Drive for Show, Putt for Dough? Unveiling Golf's Winning Formula.

[10] Gizachew Tiruneh. (2010). Age and Winning Professional Golf Tournaments.

[11] Stöckl, M., Lamb, P., and Lames, M. (2012). A Model for Visualizing Difficulty in Golf and Subsequent Performance Rankings on the PGA Tour. International Journal of Golf Science, 1, 10-24.