Badminton World Tour Analysis

Chang Li

2025-09-13

Table of contents

1 Abstract

This study investigates whether the number of points scored explains the variation in the number of wins among professional badminton players. Using data from the BWF World Tour (2018–2023), I fitted a simple linear regression model with total wins(Wins) as the response and total points scored (pts_for) as the predictor. The model shows a significant positive association between response and predictor variables ($\beta_1 = 0.0169$, $p < 2.2 \times 10^{-16}$, $R^2 = 0.981$) . While the model explains most of the proportion of variability in wins, assumption checks of residual plots indicate violations of homoscedasticity and normality, suggesting that transformation or additional predictors may improve model accuracy in future work.

2 Introduction

The research aims to answer the question: Does the number of points scored explain the variation in the number of wins in professional badminton? This question is significant because identifying the connection between scoring ability and match wins gives us a clear understanding of performance with players. The dataset is from the BWF World Tour (2018–2023). The simple linear regression model focuses on singles matches with wins as the response and total points scored as the predictor. In expectation that players who score more points will, on average, achieve more wins.

3 Data

The dataset contains statistics on 185 professional badminton players who participated in the BWF World Tour between 2018 and 2023. Each row represents one player and includes key variables such as $Wins$ (number of matches won), pts_for (total points scored), $pts_against$ (points conceded), and $matches_played$. This study focuses on 76 observations corresponding to singles players with complete performance records. We restrict our analysis to singles players because doubles events involve team-based factors that cannot accurately reflect an individual player's scoring ability in relation to the number of wins. The dataset also includes efficiency measures such as win percentage and shot percentage, which describe overall player performance (SCORE Network 2023).

4 Methods

The linear regression analysis examines the relationship between the total number of wins the player had ($Wins$) and the key predictor - the total number of points scored by the player (pts_for). With this model, we are able to detect the impact of parameters on the total number of wins and obtain the association between them.

The fitted simple linear regression model is:

$$Wins_i = \beta_0 + \beta_1 \cdot pts\_for_i + \varepsilon_i$$

Where $Wins\_i$ denotes the total number of matches won by i^{th} player, pts_for denotes the total number of points the player scored. ε_i is an error term assumed to follow a normal distribution

with mean zero and constant variance. The intercept β_0 means the expected number of wins when no points are scored, and β_1 represents the average increase in wins for each additional point scored by a player.

Estimation of coefficients is performed through Ordinary Least Squares by R language (R Core Team 2023). We use two-sided t-test for β_1 at significant level α equals to 0.05, along with the 95% confidence interval.We use ANOVA to assess the model's significance by the F-test.

We detect the linear assumptions by using the below plots: residuals vs. fitted plot to check linearity and homoscedasticity, scale-location plot for constant variance, Q-Q plot for normality, and residuals vs. leverage plot and Cook's distance help to find influential observations.

5 Results

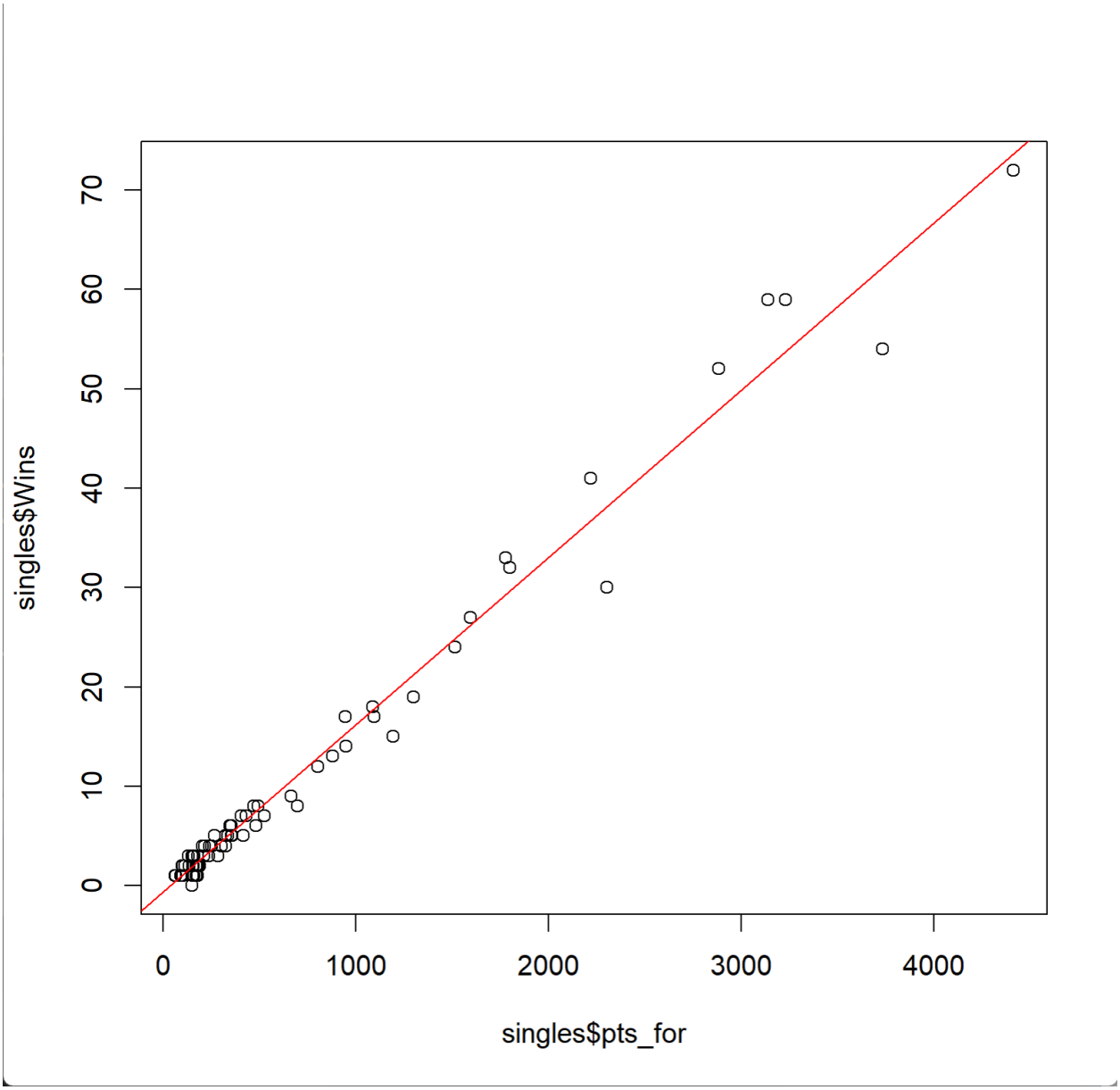

Figure 1: Scatter plot of wins vs points scored

The scatter plot shows a strong positive linear relationship between the number of wins and points scored (pts_for). Most observations fall close to the regression line, suggesting that players who score more points tend to achieve more wins. However, as the values of pts_for increase, some observations deviate from the regression line, indicating that the variability of wins becomes larger at high performance levels. This also reflects that most players achieve relatively moderate scoring and number of wins records; only a few people have very high performance in single matches.

The results of the linear regression analysis indicate that the number of wins is predicted by the point scored(pts_for). This model highlights the central role of scored performance in driving player's

performance: As the number of points scored increases, it also tends to result in an average increase in the total number of wins.

5.0.1 Estimate Coefficient

95% CI for β_0 is [-1.28870213, -0.06415109]

95% CI for β_1 is [0.01631484, 0.01740097]

β_0: in the absence of any points scored, number of wins, on average, fall somewhere between -1.2887 and -0.0642 units.

β_1: for each 1 point scored by players, there will be an average increase in number of wins between 0.0163 and 0.0174 units.

The estimator of intercept are approximate $\hat{\beta}_0 = -0.6764$ with standard error of 0.3073, indicating the true value of β_0 is expected to fall within about ± 0.3073 of the estimate.

The slope coefficient has an estimate of $\hat{\beta}_1 = 0.0169$ with a very small standard error of 0.00027, indicating that the estimate is very precise.

The t-statistic for the slope is 61.85 with a p-value $< 2.2 \times 10^{-16}$, indicating that the slope coefficient is highly significant. We reject the null hypothesis $H_0 : \beta_1 = 0$, and conclude that pts_for has a strongly positive association with wins. The intercept has a t-statistic of -2.20 with a p-value of 0.031, suggesting it is statistically different from zero at the 5% level, although its practical interpretation is limited.

5.0.2 R-Squared

Model	RSE	R Square	Adjusted R Square	DF
pts_for	2.164	0.981	0.9808	74

Predictor: pts_for

RSE (Residual standard error) represents a typical distance between the observed values and the values predicted by the model. RSE = 2.164, which indicates that predictions are typically about 2.16 units away from the true values. This provides an absolute measure of lack of fit. Since the RSE is small, we can conclude that the predicted values $\hat{y}_i$ are approximately close to the true values y_i.

A very high $R^2 = 0.981$ indicates that the predictor accounts for approximately 98.1% of the variance in Wins and captures a large amount of explanatory power. This also suggests that around 98.1% of the variance of Wins is explained by the predictor in the model. Again, a high R^2 only reflects a good fit on the data; it does not imply the model is perfect.

5.0.3 ANOVA

Model	Sum of Squares	DF	Mean Square	F value	Pr(>F)
pts_for	17916.4	1	17916.4	3825.8	< 2.2e-16 ***
Residuals	346.5	74	4.7		

The ANOVA (Analysis of Variance) test results reinforce the model's significance:
The F-statistic of 3825.8 $\gg 1$ shows that variation explained by the predictor is far greater than residual variation, and the corresponding p-value ($< 2.2 \times 10^{-16}$) indicates that the model is highly statistically significant, meaning that the predictor (pts_for) collectively has a meaningful relationship with Wins.

5.0.4 Evaluation of Assumptions

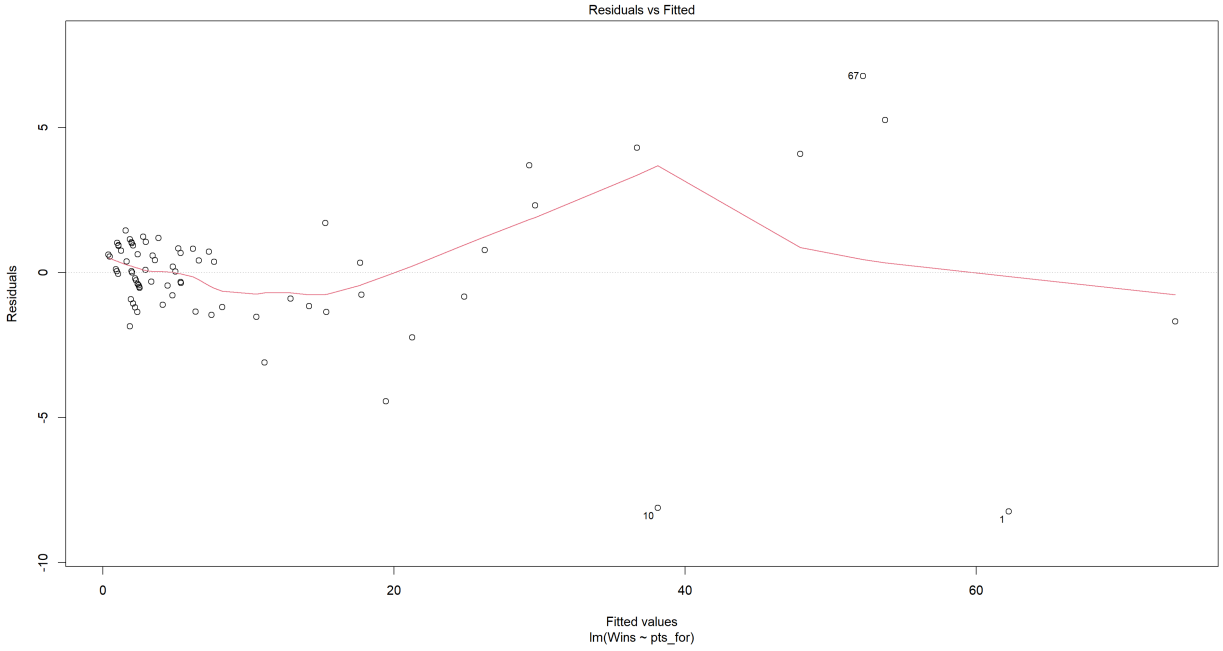

Figure 2: Residuals vs fitted values

- **Residuals vs fitted values**: The residuals are mostly concentrated at low fitted values (0–10). Additionally, as the fitted values increase beyond 20, the residuals show an increasing trend. This suggests that the linear regression model may not fully capture the relationship and there is a violation of the constant variance assumption.

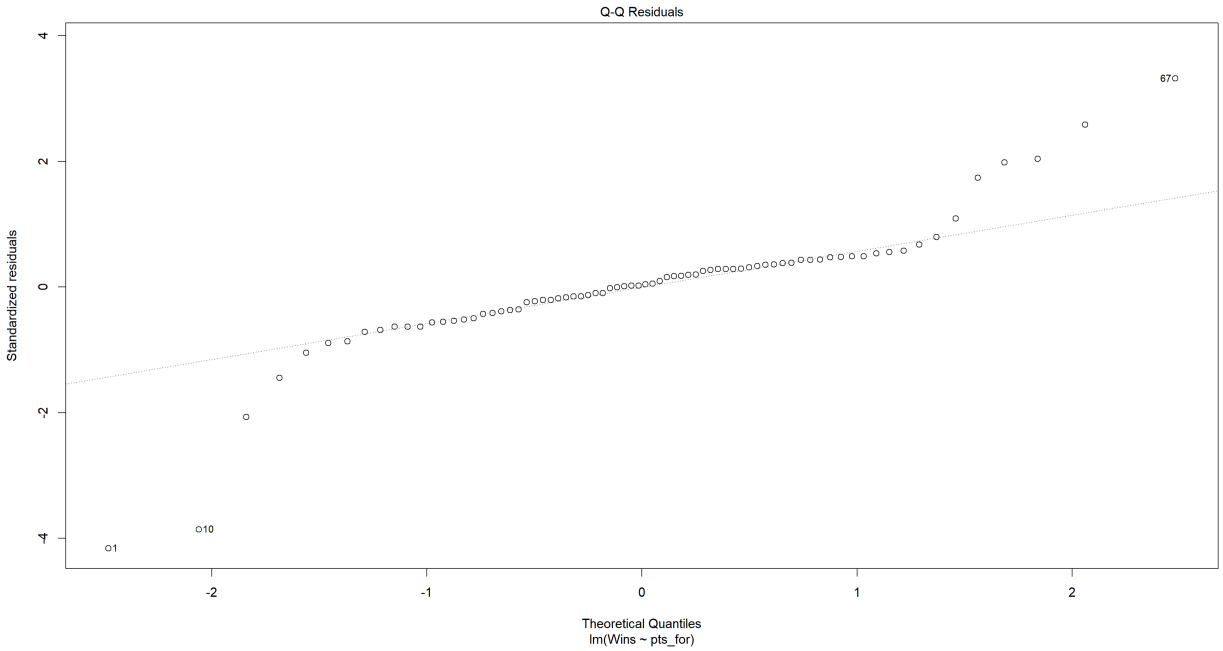

Figure 3: Q-Q residuals

- **Q-Q residuals**: The residuals generally follow the theoretical quantile along the middle range; however, at the lower and upper tails, the points deviate from the theoretical quantile. This indicates that the residuals have heavy tails compared to the normal distribution, thus, normality is not fully satisfied.

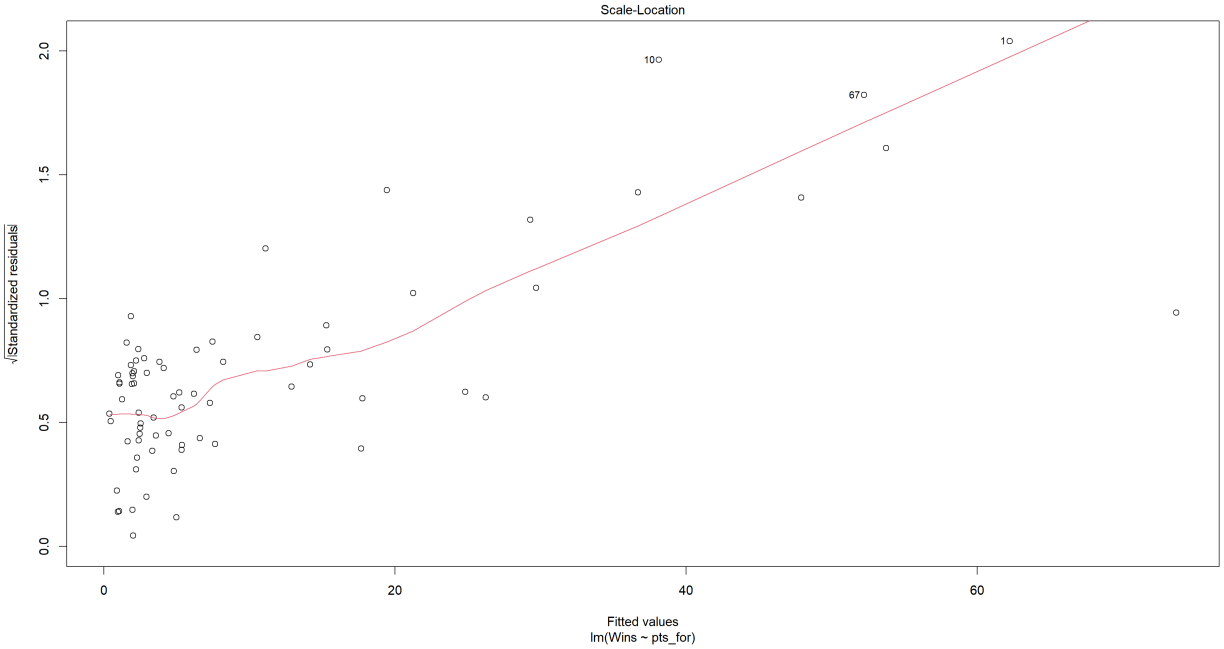

Figure 4: Scale-location plot

- **Scale-location plot**: The scale-location plot evaluates homoscedasticity. In this plot, we see a noticeable trend showing that variance of residuals increases as fitted values increase. Thus, this suggests a violation of the homoscedasticity assumption.

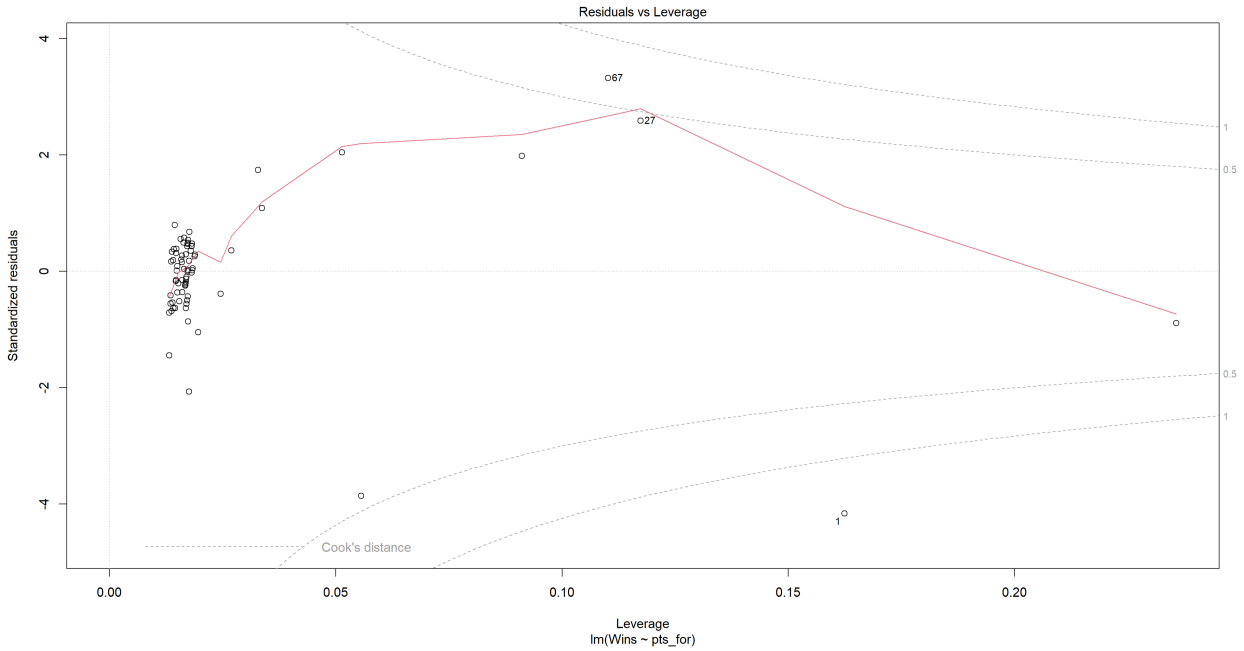

Figure 5: Residuals vs leverage

- **Residuals vs leverage**: This plot detects outliers and influential observations. In this plot, three observations (such as 1, 27, and 67) lie close to or beyond the Cook's distance contours. This shows that these observations have crucial impact on the regression model.

6 Discussion

This analysis answers the main question by showing that scoring ability has a strong and statistically significant association with wins in professional badminton singles matches. The main strength of the analysis is its high interpretability, as seen in the very considerable R^2 and precise slope estimate. However, the obvious weaknesses include violation of the homoscedasticity, heavy-tailed residuals, and influential data points that may crucially affect the conclusion.

The transformations could be considered as one of the future works, additionally, including more predictors (e.g., a dummy variable for category differences between Singles and Doubles). Further research could also explore longitudinal performance trends across different years and tournaments.

7 Reproducibility

All analysis code, raw data, and documentation are available in GitHub repository:
https://github.com/Chang-Lii/badminton_worldtour_analysis.git

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

SCORE Network. 2023. "Badminton World Tour 2018–2023 Dataset." https://data.scorenetwork.org/badminton/badminton_worldtour_2018-23.html.