The Moneyball of Formula 1: The Predicative Accuracy of Linear Regression on the Constructors' Championship

Bella McClure

Department of Statistics

University of Connecticut

May 1, 2023

Abstract

Fans of motorsport are accustomed to rooting for their favorite driver hoping they will win the title. In Formula 1 (F1), a parallel Constructors' Championship exists where teams, or constructors, compete for the most points accumulated over the season. Predicting the Constructors' Championship can provide insights into the teams' performances and their likelihood of winning in future seasons. This study explores the use of linear regression to predict the F1 Constructors' Championship based on various factors such as the number of fast laps, poles (first place starts), retirements (crashes during a race), wins, and podiums a given team had. Data were evaluated from the 1999 season until the most recent 2022 racing season. Simple and multiple linear regression were utilized to test the accuracy of such modeling on the results of the Constructors' Championship based on these factors. Results indicate that both multiple and simple linear regression are fairly accurate tools for predicting the results of past Constructors' Championships and the important variables could help team management instruct drivers and set race strategy to maximize the team's chances of winning the overall title.

Keywords: Data analysis, Motor sport, Sports analytics, Statistical modeling

1 Introduction

Formula 1 (F1) has historically been a sport of global attraction. With its first season in 1950, it has grown into a multiple billion dollar grossing sport that fans around the world take pride in watching. This high-class auto racing involves the driving of bespoke race cars around some of the world's most famous racetracks. It is governed by the FIA (Fédération Internationale de l'Automobile), which sets out the technical and sporting regulations for the competition. F1 teams consist of two cars piloted by the designated team drivers taken from a pool of primary and reserve drivers. They compete against each other and drivers from different teams to earn points toward the Drivers' and Constructors' Championships. As in other motorsports, the individual driver with the most points wins the Drivers' Championship.

The Constructors' Championship, something only seen in F1, is awarded to the team that scores the most points over the course of the season. It was introduced in 1958, and since then, it has become an important part of the sport, with teams competing fiercely to win the title and the associated prize money. As with other major sports, F1 teams are primarily owned by wealthy individuals or corporations and each team must design, build and operate their own races cars. Note that this mainly applies to the chassis as the engine can be supplied to customer teams. The championship recognizes the skill and expertise of the teams and is a measure of their success in building a car that can perform consistently across a range of different circuits and conditions. It is also important for understanding the trends and innovations that are shaping the sport, and the impact that these developments are having on the sport's future direction. The combination of the Drivers' and Constructors' Championships acknowledges both the individual skill of the driver and the collective effort of the team, making it a unique and complex motorsport competition worthy of statistical

investigation.

Surprisingly, there are few publicly available peer reviewed articles on the application of statistics in Formula 1. Most of the writings regarding Formula 1 and predictions are done so by simply using intuition. The earliest primary literature was by Bell et al. (2016) by a group of students from the University of Sheffield. They focused mainly on random-coefficient models, for the purposes of finding the rankings of the best driver in history and to quantify certain variables, such as the relative weight of teams and drivers to performance. In a research symposium put on by Valparaiso University, Lasrado (April 29, 2021) presented research on the use of predicative modeling to predict the winner of every race in the 2021 season. Rana et al. (2021) also focused on the prediction of winners, but looked more into the prediction of the overall driver's champion. The work of Rana et al. (2021) is arguably the most thorough paper to date that applies statistical modeling to Formula 1, as it was written for the 2021 International Conference on Industrial Electronics Research and Applications. This paper used models including LASSO Penalized Linear Regression, K Nearest Neighbor Regressor, and Ridge Penalized Linear Regression; however, they also employed an equal number of machine learning algorithms including Neural Networks, Random Forest Regressor, and Extreme Gradient Boost Regressor. Most recently, Sicoie (2022) published a thesis that more generally attempted to give a machine learning framework for predicting race winner and by extension the championship standings. Unfortunately, they found the results to be "not satisfactory" when basing the performance of the models on their specified metric of the correlation coefficient.

In fact, the published papers found their results to be merely acceptable in terms of the models' capacity to accurately predict the winners of each race or of the championship. Each respective discussion suggested more complex modeling with even more fine tuning, in spite of the fact they already describe the use of some of the most complex modeling available. Despite this, the intent of this work was to employ simple modeling tools. In large part this was driven by the prima facie observation that the sports results are predictable. Formula

1 has long had an issue with competitive balance. The result of these imbalances is that the sport is relatively predictable, arguably the most predictable sport in the world. In terms of predicting the Constructors' Championship, someone who regularly watches could accurately choose the champion with odds around one-third. What other sport allows a one-third chance of selecting the winner each year? The odds of one-third come from the history of the sport, as 98% of the time the winning constructor is one of the top 3 teams from the year prior.

Given the apparent predictable nature of the sport, more straight-forward statistical modeling such as simple and multiple linear regression should be able to, with decent accuracy, predict outcomes. The work described herein will evaluate and identify variables of interest and subsequently apply simple regression techniques to test the accuracy with which regression can predict the past constructor winners. The accuracy metric is defined by the absolute value of the difference between the predicted total percentage of points earned and the true percentage of points earned (prediction error). This paper will first establish the variables of interest, regress on those variables using simple and multiple linear regression, and then conclude with the results of the modeling and a discussion of where this research could go in future analyses.

2 Data Description

The data were taken from both the official Formula 1 website and Wikipedia. The variables of interest were manually transcribed from the source websites and compiled directly into the software. The scope of the study covered the top 5 teams from years 1999-2022. While there are usually ten teams, the study was limited to the top 5 because there is such dominance in the podium places (the top three places) and it is extremely rare to see a team from below score any meaningful number of points, making the mid-fielders and back-markers

much less relevant.[1] The years included in the study were arbitrarily chosen to include only the "modern" era. The cutoff of 1999 was somewhat strategic in that the farther back you go in the history of the sport the more different it becomes in every facet, particularly in the points allocation, making the inconsistencies potentially too radical. That year also marks the beginning of the long stretch of Ferrari dominance and the start of success for one of the most recognizable Formula 1 driver and constructor combinations of Michael Schumacher and Ferrari.

The dependent variable of interest was the percentage of total points, which was standardized as a percentage because the total points available to a given constructor varied almost every season. This variable was calculated with the following formula:

$$PercentageofPoints = \left(\frac{PointsScored}{TotalPointsAvailable}\right) \cdot 100,$$

where the $TotalPointsAvailable$ = points possible over a race weekend $\cdot$ races in a season. The initial independent variables of interest were the number of fast laps, the number of poles, the number of retirements, the number of wins, and the number of podium finishes. These data points were collected and input directly into the software. Note that these variables were recorded for the top 5 teams for every season starting from 1999 until the 2022 season. The descriptive statistics for all variables can be seen in Table 1.

These variables were chosen based on past literature and because on the surface they are the most logical quantitative variables to predict a winner, both driver and constructor. There are variables that may affect the outcome of the Constructors' Championship that are either not easily quantified or unavailable to the public including but not limited to the budget awarded to each team, the reliability of the car, and the top speed versus efficiency of the vehicle. This phenomenon also ties into the reasoning behind choosing to focus on the

[1]Note: The one exception of this would be the 2009 season in which the automaker Brawn won the Constructors' Championship in its first year as a car manufacturer in Formula 1. This can be explained by an FIA rule change regarding the aerodynamic regulation on the car, radically altering the results of the 2009 season.

Table 1: Descriptive statistics of all variables

Variable	N	Mean	SE Mean	StDev	Variance	Median
Total Point Percentage	120	16.09	1.06	11.64	135.45	16.34
Fastest Lap	120	3.51	0.35	3.80	14.44	2.00
Poles	120	3.60	0.45	4.95	24.50	1.00
Retirements	120	6.33	0.33	3.62	13.10	5.50
Wins	120	3.60	0.43	4.74	22.43	1.00
Podiums	120	10.37	0.81	8.83	77.93	9.00

Constructors' Championship rather than the Drivers' Championship. Much of the other peer review literature focuses on the prediction of the Drivers' Championship, making it slightly more redundant. Additionally, the prediction of a driver winner has more inconsistencies than that of the constructor winner because it is more sensitive to driver input. Though the Constructors' Championship is also based on the drivers, it does not only account for the actions of a single driver but rather the results from both team cars.

As such, this analysis does not have to consider race conditions and driver experience to the same degree someone performing an analysis on the Drivers' Championship would. This also provides stability and a more reliable understanding of the car performance. Lastly, having a dual championship in a sport is rare, making the Constructors' Championship unique and an element of the sport worth acknowledging. Not only is it unique, but it is crucially important to the world of Formula 1. The standings of the Constructors' Championship determine how much each team is awarded at year end, meaning every single point is hotly contested as one point is potentially worth millions of dollars to a team's budget for the coming year. Money is in an input into everything: the expenditure allowed to optimize the car during the season, the number of engineers that can be paid, the salary of the best drivers, and these all require a large budget to fight for the title. Given the overall importance of the constructors to the sport of F1, and a lack of prior analysis, a statistical investigation in this paper was judged worthwhile.

3 Methods

The list of models utilized for analysis were multiple linear regression (MLR) and simple linear regression (SLR). A Stepwise Regression Procedure and a correlation matrix were also performed in conjunction with these models.

The first step to the analysis was running an MLR including all the independent variables. A general multiple linear regression model will take the following form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ...\beta_p X_{ip} + \epsilon_j,$$

where $p =$ the number of independent variables, $\epsilon_j \sim \mathcal{N}(0,\, \sigma^2)$, and the estimated parameters are $\hat{\beta}_0$ and $\hat{\beta}_i$. SLR has the same general form; however instead of multiple predictors, there is only one. The parameter estimates for SLR can be calculated by:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

and

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}.$$

In the analysis, the variables can be listed as

- Y is the percent of the total available points a team earned
- X_1 is the number of fast laps
- X_2 is the number of poles
- X_3 is the number of retirements
- X_4 is the number of wins
- X_5 is the number of podiums

where each variable has a corresponding coefficient of $\hat{\beta}_i$. The SLR analyses used the number of wins and the number of podiums as the sole predictors, whereas the MLR included all the variables.

In SLR, standard error, denoted SE, is the standard deviation of the statistic. The respective standard error calculations are

$$\hat{\sigma}_{\hat{\beta}_0} = \hat{\sigma}\sqrt{\frac{\sum(X_i)^2}{n\sum(X_i - \bar{X}^2)}}$$

and

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}.$$

The parameter estimates and their respective SE calculations can also be calculated for MLR in a similar manner; however, the form with which these equations take is quite complex as they must be adapted for the additional predictors that are entered into the model.

The results of all models will be under the assumptions that error terms (ϵ_j) are independent, have a mean of zero, a common/constant variance of σ^2, and follow a normal distribution. These assumptions will be checked before any discussion of the results to verify the validity of the chosen model; this was accomplished utilizing visual graphics including residual versus fits plots and density plots.

The measure of the goodness-of-fit for both models was determined using the R-squared value, which can be defined as:

$$R - squared = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}.$$

After regressing on the unrestricted model, there was discussion about the Variance Inflation Factor (VIF) of each variable. The VIF is a measure of how the variance of an independent variable is influenced by its interaction with other independent variables present in the model. It can be calculated by

$$VIF = \left(\frac{1}{1 - R_k^2} \right).$$

To get a better sense of the potential issues, a Pearson correlation matrix and table was graphically and computationally run. The visual element is a simple matrix of scatterplots, while the correlation table consists of the correlation coefficients (r) for all possible combinations of variables.[2] Both of these will be found in Section 4 of the paper. The value of such coefficients was calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

though there are multiple ways for this value to be found.

A Stepwise Regression Procedure was performed to determine if the original model was truly the best fit, given the full model involved five variables. The variable with the largest VIF number was excluded from this procedure as Stepwise is known to not perform as well when there is known multicollinearity in the model, making the total number of candidate terms 4. This specific process permits the re-examination of the variables already incorporated in the model through t-tests of each slope.

The process starts by regressing each predictor individually — i.e., regress Y on X_1, Y on X_2, etc. and then determine which predictor has a t-test p-value that is less than the predetermined α level. The null and alternative hypothesis for such a test is as follows:

$$H_0 : \beta_p = 0, H_a : \beta_p \neq 0 \tag{1}$$

with the appropriate t-statistic being calculated as:

$$t_{\hat{\beta}_p} = \frac{\hat{\beta}_p - \beta_1^{(0)}}{SE_{(\hat{\beta}_p)}}.$$

[2]Note: Only the interactions deemed significant appear in the correlation coefficient table.

Note that $\beta_1^{(0)}$ will equal 0 as it represents the null value, a known constant that may equal the actual unknown parameter, β. A non-zero $\beta_1^{(0)}$ value may be appropriate when testing a different null hypothesis. Thus, the simplified equation can be written as:

$$t_{\hat{\beta}_p} = \frac{\hat{\beta}_p}{SE_{(\hat{\beta}_p)}}.$$

The predictor with the smallest p-value (also equal to the biggest t-statistic in absolute value) is then entered into the model first. You then repeat the process with X_i already in the model by regressing Y on X_i and each other candidate predictor individually. If this step also results in a predictor with a p-value less than the specified α level, that predictor is then entered into the model too. At this point, given there are two predictors in the model, you check if entering the second predictor changed the significance of the first predictor entered by checking the significance of the same test seen in equation 1. If the result of this test is no longer significant, the first predictor entered would be removed from the model. This process continues until there are no more predictors to be entered or removed from the model.

For the purpose of this analysis, the α level to enter and remove a predictor was left at the default of 0.15 used by common statistical software for Stepwise Regression. It is worth noting that Stepwise Procedures are particularly sensitive to these values and results may vary depending on what α level is chosen, which is why it was left at the level that is generally accepted. Despite this, the procedure was favored over other processes such as Backward-Elimination Procedure or Forward-Selection Procedure in that it accounts for the fact that variables entered at an early stage may become unnecessary at a later stage due to a correlation with variables added afterwards. A summary of the findings can be found in Section 4 of this paper.

4 Results

First, a preliminary visual check of the assumption was performed, the plots for which can be seen in Figure 1. Starting with the Normal Probability plot, beside the two noticeable outlying points, the normality assumption seems verified. The larger outlier can be identified as the 2002 Ferrari victory where they captured an incredible 50% of the percentage of total points. On the other end of the spectrum, the other point is of the 2013 season from the Lotus Renault placing fourth and only earning a minuscule 0.16% of total points. It is

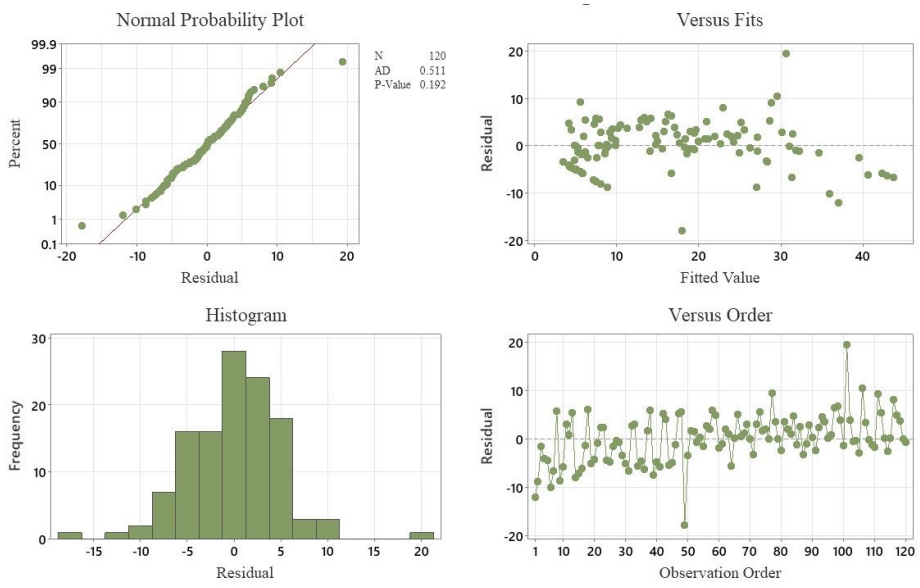

Figure 1: Residual plots to check for normality

not surprising these residuals are outliers as the corresponding data represent the extremes on both ends of the total percentage of points a team could have earned. Moving to the histogram, a bell-shape curve is easily seen, again reaffirming the verification of the normality assumption. The plot of residuals versus fits which attempts to verify the constant and equal variance assumption looks to be random and clustering mainly around the line of 0, verifying this assumption; however, it is worth noting there is a slight linear pattern on the left of the graph just below the line of 0. Lastly, though the independence assumption does not seem blatantly violated (no clear pattern) the data is ordered by year giving a slight linear pattern to the plot. While this limitation is made more acute by the fact that the bulk of

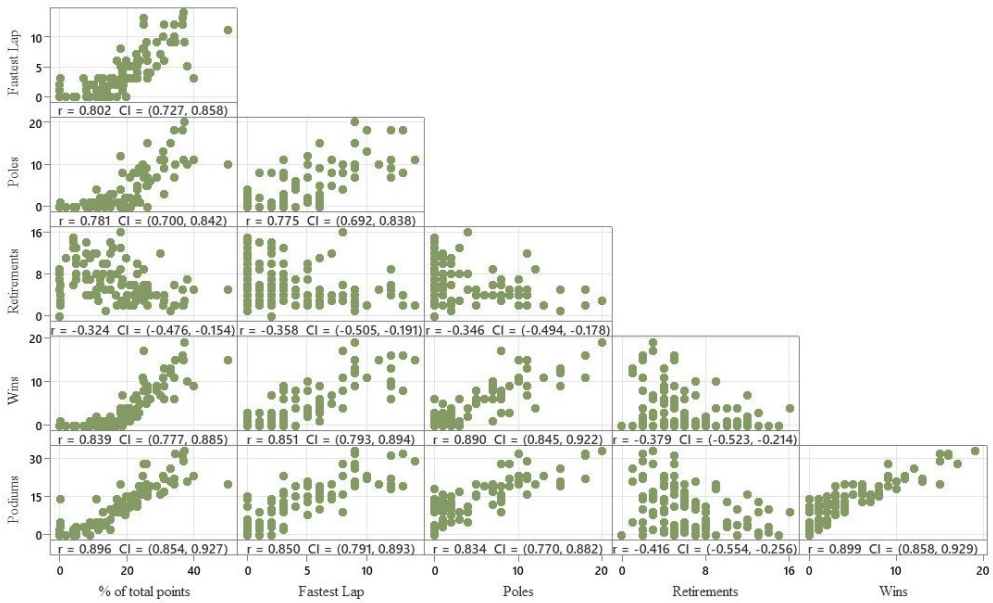

Figure 2: Correlation matrix of all independent variables

Table 2: Pearson correlation coefficient table

	Total Points Percentage	Fastest Lap	Poles	Retirements	Wins
Fastest Lap	0.80				
Poles	0.78	0.78			
Retirements	−0.32	−0.36	−0.35		
Wins	0.84	0.85	0.89	−0.38	
Podiums	0.90	0.85	0.83	−0.42	0.90

the team members remain constant from year to year, key team members do occasionally move to rival teams with outsized impact. The limitation of this violation and a potential remedy are discussed in Section 5.

Continuing with the analysis, first discussing the full MLR model including fast lap, poles, retirements, wins, and podiums. The fitted regression model is shown below:

$$\hat{Y} = 2.83 + 0.318X_1 + 0.098X_2 + 0.193X_3 + 0.239X_4 + 0.937X_5.$$

At first glance, the model looks to be adequate, with an R-squared value of about 82%. Looking closer however, the VIF for variables X_2, X_4, and X_5 (poles, wins, and podiums) are very large— 4.95, 8.69, and 6.38 respectively.

The general guidelines on what is considered a "good" VIF value are that if it is < 1 there is no multicollinearity present, while a VIF > 10 indicates very strong multicollinearity in the model. Thus, while a VIF of around 3 or 4 may be only slightly concerning, a VIF of 7 would be a stronger suggestion of a problem. To understand this potential multicollinearity problem better, a correlation matrix was run. This allows a visual check of each variable and gives additional information not provided solely by the VIF.

Figure 2 and Table 2 above help to see where the assumption is violated to the highest degree and how severe it is. Notably, the worst offense is between X_4 and X_5 (number of wins vs. number of podiums). This is somewhat expected given that the number of wins is a subset of the number of podiums so intuitively speaking there must be some correlation between the two. With that being said, there are other concerning correlations amongst the independent variables including X_2 versus X_3 (number of fast laps vs. number of wins). The large VIF values bring down the credibility of the model regardless of the proper fit, which prompted the use of a Stepwise Regression Procedure to mitigate the multicollinearity problems. As mentioned, the variable with the largest VIF value (X_4) was excluded from this procedure to ensure proper function of the process.

The summary of the steps taken are seen in Table 3 with the fitted regression equation being

$$\hat{Y} = 3.967 + 0.438X_1 + 1.021X_5.$$

The Stepwise Procedure found that the optimal model included X_1 (the number of fast laps) and X_5 (the number of podiums) with the number of podiums being included first, followed by the number of fast laps. At the threshold of $\alpha = 0.15$, the number of poles earned and the number of retirements were left out of the final model according to the t-tests. Consequently, the VIF for the remaining variables are lower than in the full model, with both having a value of 3.60. With that being said, the R-squared value of around 81% proved to be around the same as the unrestricted model.

Table 3: Stepwise Regression Procedure
Candidate Terms: Fastest Lap, Poles, Retirements, Podiums

	—Step 1—		—Step 2—	
	Coef	P	Coef	P
Constant	3.84		3.97	
Podiums	1.18	0.00	1.02	0.00
Fastest Lap			0.44	0.064
S		5.18		5.13
R-sq		80.35%		80.93%
R-sq(adj)		80.19%		80.59%
Mallow's Cp		5.55		4.03

α to enter = 0.15, α to remove = 0.15

Given the VIF values were still greater than 1, two separate SLR were run using the number of wins and the number of podiums to test the relative importance of each variable individually. Clearly both variables should be highly correlated to the percentage of total points as the winning points are a significant subset of the podium points awarded to the top three finishers. The fitted regression plot and equation for each SLR can be seen in Figure 3 and Figure 4.

Interestingly, the R-squared values are quite different with a difference of around 10% favoring the fit for podiums. It is observable then, that adding wins to the model when the number of podiums is already present does little to increase the accuracy of the model, as the SLR using podiums has an equivalent R-squared value compared to the full MLR model including both. This implies that getting a podium place in itself is more valuable to success in the Constructors' Championship than wins on its own, despite the large drop-off in points for second or third place. This observation potentially puts the priorities of the driver who is seeking victory at odds with team management that may want to guide towards a podium finish over the risk of pushing the car over its failure limit in an attempt for victory. This change in philosophy has parallels to "Moneyball" in baseball where slugging percentage was found to be a greater priority over batting average.

Comparing the fit of the model produced by the Stepwise Regression and the SLR using

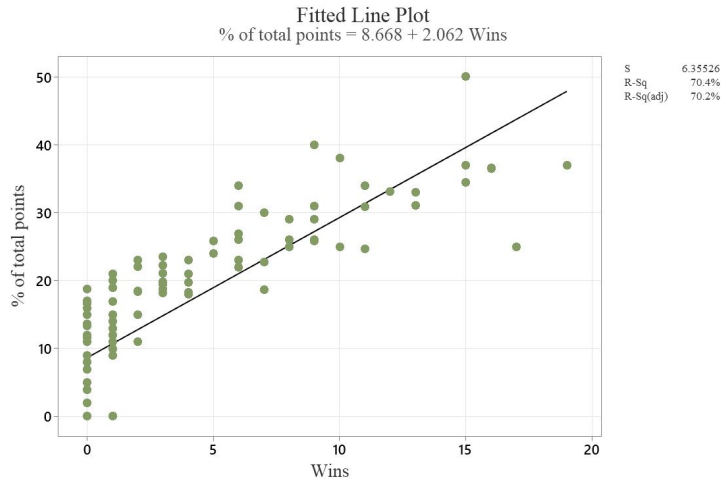

Figure 3: Single linear regression using wins as the sole predictor

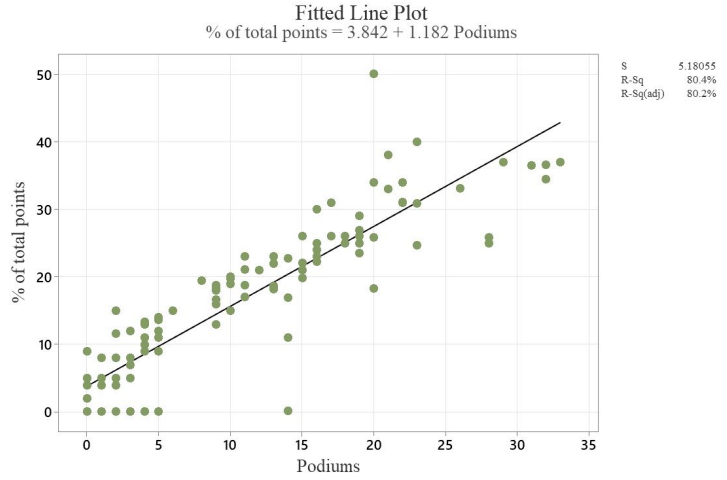

Figure 4: Single linear regression using podiums as the sole predictor

the number of podiums, the Stepwise model does technically have a higher R-squared but the difference is less than 1%, which is negligible. Using the SLR also completely eliminates potential for multicollinearity issues.

Due to these factors, the data were plugged back into only the SLR using podiums as the sole predictor when trying to test the accuracy. Additionally, given the model was meant to test the accuracy of predicting the winning constructor, only the victors of every season were included in Table 4. The average error was only about 5 percentage of total points earned, which is not terribly high. By simply looking at the table, there are some noticeable

Table 4: Summary of Results

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Constructor's Champion	Fer	Fer	Fer	Fer	Fer	Fer	R	R	Fer	Fer	Br	RB	RB	RB	RB	MB	MB	MB	MB	MB	MB	MB	MB	RB
Constructor's % of Total Points	31	38	40	50	25	37	26	29	31	25	26	26	34	22	31	37	37	37	33	31	35	33	26	25
Predicted* Constructor's % of Total Points	24	29	31	27	23	38	24	25	30	25	22	26	30	20	30	40	42	43	35	31	42	29	37	37

*SLR including variable = 'podiums'; Fer = Ferrari; R = Renault; Br = Brawn; RB = Redbull; MB = Mercedes Benz

patterns. The SLR model pretty consistently underestimated the percentage (with some exceptions most notably in 2022), and was fairly consistent in terms of its error across the different seasons. When comparing the model predictions for first place versus other podium places, it became clear the underestimation got worse as you went down in podium places—i.e., the model was better at predicting the percentage of total points of the team who placed first compared to the team who placed third or fourth.

5 Discussion

To recapitulate, this research attempted to utilize less complex modeling to predict the Constructors' Championship in F1. A significant relationship between the variables evaluated was observed. This was not entirely surprising given the obvious overlap between certain variables, particularly wins and podiums. Both SLR and MLR produced good results with the best fit obtained when using the full model including all variables; however, the credibility of this model is questionable given the obvious multicollinearity problems. This was a known potential problem going into the analysis, and given the severity of the issue, further steps were taken to reduce the effects. The Stepwise Regression Procedure did somewhat mitigate the issue utilizing only two of the five variables, but given near equivalent results were obtained with the SLR considering just podiums, this variable is seen to have the largest importance overall. As such, the SLR with podiums is the simplest method for predicting F1's constructors' champion. The accuracy was acceptable, with the most accurate results (having the smallest error) being obtained for the first place constructor, which is an ideal result.

The idea that the number of podiums might be more important than number of wins has important implications for F1 team management. One can expect resistance from F1 drivers if team bosses were to instruct drivers to settle for podium places over pushing for victory. F1 drivers did not get to the pinnacle of motorsport by settling for anything less than victory. During a race one rarely hears a team instruct their driver over the radio to settle for third over pushing for higher steps on the podium; more often one is witness to drivers crashing out in their relentless drive to achieve victory.

At a minimum these results argue that teams should question their current practice of letting their drivers fight for victory from lower podium places, especially towards the closing stages of a race. Further statistical analysis of the Drivers' Championship is needed to understand if the two parallel F1 championships are influenced by the same (or different) variables to the same degree as this championship. The lessons from "Moneyball" of identifying the appropriate variables for both driver and constructor will clarify the extent to which their priorities are aligned.

Though the model was fairly accurate and provided insights, it is worth noting there are limitations to this study. The data collection process was complicated by the fact that other potentially good predictors have no public data, such as true budget of a team for a given season. Additionally, given the slight violation of the independence assumption lowering the credibility of the results, different modeling besides regression may be warranted, such as time series modeling. Lastly, there are some variables that would be great to include that simply do not follow a linear trend, such as the tire degradation of each car, aerodynamic drag, and downforce of the car. In this way, the study was limited to those variables that would likely have a linear property, which restricted the scope of predictors that could be considered. If future researchers can gain access to more data that are currently carefully guarded secrets of F1 teams, it could help improve the predictive accuracy which would allow for more insights to be made into the sport. Furthermore, utilizing a statistical tool that allows for variables that follow a curve rather than a line, such as non-linear regression,

would optimize the predictive abilities of the model and is an additional investigation that researchers could focus on moving forward.

References

Bell, A., Smith, J., Sabel, C. E., Jones, K., and Bell, A. (2016), "Formula for success Multilevel modelling of Formula One Driver and Constructor performance,1950-2014," *Journal of Quantitative Analysis in Sports*, pp. 99–112.

Lasrado, S. (April 29, 2021), "Predicting winners in the Formula 1 car racing season," Unpublished technical report.

Rana, R., Pandey, D., Mishra, S., Nehra, N., Deshwal, D., and Sangwan, P. (2021), "Predicting Standings in F1 Sports Driver's Championship using Lasso Penalised Regression," in *2021 International Conference on Industrial Electronics Research and Applications (ICIERA)*, IEEE, pp. 1–5.

Sicoie, H. (2022), "Machine Learning Framework for Formula 1 Race Winner and Championship Standings Predictor," Ph.D. thesis, Tilburg University.