

Predicting The Outcome Of NASCAR Races: The Role Of Driver Experience

Mary Allender, University of Portland

ABSTRACT

As national interest in NASCAR grows, the field of sports economics is increasingly addressing various aspects of this sporting contest. The outcome of NASCAR races is of particular interest to fans and thus, models describing and predicting the outcome of NASCAR races are beginning to emerge. This paper builds a model predicting the outcome of NASCAR races using NASCAR data. The outcome is hypothesized to depend on a set of variables and looks in particular, at the importance of driver experience. The findings of this paper conclude that driver years of experience do in fact play a significant role in predicting the outcome of NASCAR races.

INTRODUCTION

The origins of NASCAR reach back to the days of Prohibition when the cars used by moon shiners needed speed while making delivery runs to avoid the authorities in pursuit. More horsepower was needed and so began the quest to modify cars for more horsepower and reliability. Simultaneously, auto racing became a sport. The inaugural auto race at Daytona Beach took place March 8, 1936 (Felden, 2005).

These early races, however, were not officially organized and so races were haphazard and drivers tended to show up randomly. The original tracks often consisted of dirt or sand. Fans were still few in numbers and driving stock cars remained a hobby since it didn't generate enough income to qualify as a job. Over the next ten years, fan interest increased considerably and stock car racing evolved from an occasional, hastily organized race on sand and dirt tracks, to the stadiums and paved tracks we know today. In December, 1947, Bill France Sr., both a driver and race promoter, developed the idea of NASCAR as organized stock car racing subject to specific rules. On February 15, 1948 NASCAR ran its first race at the Daytona Beach road course. The Daytona 500 remains the premier NASCAR race today. This paper proceeds in six parts. Part II analyzes current research on NASCAR as a sport. Part III discusses the data used in the paper. Part IV develops an empirical model that can be used as a tool to predict the outcome of NASCAR races. Part V analyzes the findings of the model and part VI offers conclusions and avenues for further research.

CURRENT RESEARCH

Scholarly research on NASCAR as a sport is relatively new and has taken many different directions. One avenue of research focuses on the reliability of NASCAR vehicles and explores the reasons behind part failure and the extent to which these critical part failures can be reduced. Majety, Dawande, and Rajgopal (1999) show that in general, the typical reliability allocation problem maximizes system reliability subject to a budget constraint. They note that cost is an increasing function of reliability and hence the tradeoff between dollars spent and system reliability. Although the media would have us believe that NASCAR owners are willing to spend virtually unlimited amounts of money to earn a spot in Victory Lane (New York Times, 2/13/06; CBS News, 10/6/05), NASCAR teams themselves acknowledge that in fact, a budget constraint does exist both in the form of willingness to spend money and the rules imposed on the construction of the vehicles themselves although budgets in NASCAR racing are far more substantial than those common to commercially produced vehicles (Wachtel, 2006. Allender (2007) continues with the reliability question and asks whether or not critical part failures in NASCAR vehicles are higher than what we "ought" to expect and explores some reasons as to why in fact they are.

Other lines of research focus on the type of tournament NASCAR represents and the most efficient type of reward structure for rank order tournaments where finish position is all that matters to getting a prize. Becker and Harold (1992), Lynch and Zax (2000) and Maloney and McCormick (2000) use ROT (rank order tournament) theory to investigate the effect of different types of payment structures on the performance of contestants. Along similar lines, Lazear and Rosen (1981), Nabeluff and Stiglitz (1983), and O’Keefe, Viscusi, and Zeckhauser (1984) began to look seriously at a payoff structure that was preferable for the contest organizer. In fact, it was this line of research that began to take the field of sports economics into the realm of serious economic literature Fizez (2006).

Fans of NASCAR are ultimately interested in the outcome of each contest or race. The Nextel Cup Champion for the year, in essence, wins the majority of the points associated with the 38 races NASCAR holds each year at different tracks. Before the season and before each race, popular media focuses much attention on predicting the winner of each race. However, there is little in the sports economics literature that attempts to develop models that help predict the outcome of a NASCAR race. Pfitzner and Rishel (2005) develop a model predicting order of finish in NASCAR races based on variables such as car speed, driver characteristics, and the like. This paper seeks to add to that burgeoning body of literature by developing an empirical model that identifies the most important variables contributing to a driver’s success in a race. Thus, the model can be used as a tool in predicting the outcome of NASCAR races.

THE DATA

The data for this research was taken from the NASCAR website. The 38 races for the 2002 season were used where each race included 43 drivers. The data used included starting position of each driver, finish position of each driver, and individual driver characteristics.

THE MODEL

From a theoretical perspective, this paper hypothesizes a model as follows that may help predict the finish position of individual drivers in NASCAR races.

$$FP = C + SP + TL + PC + DY \tag{1}$$

where

FP represents the finish position of each driver at the end of each race.

C represents the intercept term in a regression model.

SP represents the starting position of each driver determined prior to the race during qualifying runs. The coefficient on this explanatory variable is expected to be positive. Starting position based on qualifying times sets the starting position for each of the 43 cars and it appears plausible that those cars starting closer to the front would have a greater probability of finishing toward the front.

TL represents track length. The coefficient on this explanatory variable is expected to be positive. This is based on the hypothesis that the best chance to pass occurs in the turns (Martin, 2005) which are encountered more frequently on shorter rather than longer tracks.

PC represents the percentage of laps under caution. Since caution laps freeze the car positions, we expect the coefficient on this variable to be positive since the greater the percentage of laps under caution, the more difficult it becomes to make up laps when coming from behind.

DY represents the years of experience a driver has in Nextel Cup racing. We hypothesize a negative coefficient on this variable because as driver years of experience increase, we expect that experience to pay off in the form of more wins or finishing further toward the front.

From a theoretical standpoint, it is reasonable to ask whether some of these independent variables posited to explain finish position may be interacting. To assess this possibility, we ran a restricted version of the regression equation 1 above. That is, we ran equation 1 as it is written with no interaction terms. The results are detailed in Table 1. Running the restricted regression tests the null hypothesis that no interaction terms are valid to include. If some of the explanatory variables are interacting, the SSE term on the restricted regression will be relatively high. As Table 1 indicates, the SSE on the restricted regression is high and we decided it was appropriate to use interaction terms.

Table 1

Sample: 1 1440
Included observations: 1440
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	16.81865	1.372632	12.25284	0.0000
SP	0.294190	0.025363	11.59902	0.0000
TL	0.048796	0.471855	0.103413	0.9176
PC	0.695973	5.115689	0.136047	0.8918
DY	-0.137512	0.040933	-3.359486	0.0008
R-squared	0.094788	Mean dependent var		22.00139
Adjusted R-squared	0.092265	S.D. dependent var		12.41622
S.E. of regression	11.82957	Akaike info criterion		7.782549
Sum squared resid	200812.2	Schwarz criterion		7.800856
Log likelihood	-5598.435	F-statistic		37.56604
Durbin-Watson stat	0.408039	Prob(F-statistic)		0.000000

Since driver years of experience and starting position are the only significant variables in the restricted regression, we decided to use those variables as interaction terms. Specifically, we added the interaction terms DY*SP and DY*TL. If the SSE on this regression declines, then we have to reject the null hypothesis that the restricted version is valid. More specifically, the interaction terms tell us:

DY*SP – is the driver years of experience more important to finish position with the starting position effect or without it. The null hypothesis is that the interaction effect makes no difference. If we reject the null, the interaction term is important. Similarly,

DY*TL describes the interaction term between the years of experience with track length. By rejecting the null hypothesis, we must assume that a driver with more years of experience can negotiate different track lengths more effectively than those with fewer years of experience and thus have a greater likelihood of finishing toward the front.

Table 2 tests the model

$$FP = C + SP + DY*DP + TL + PC + DY*TL + DY \tag{2}$$

where the variables remain as described above.

Table 2 shows the results of the regression with the interaction effects. Since the SSE declines, then the results from table 2 depict a more accurate hypothesis about the variables most likely to predict finish position prior to a race. Thus the model described in equation (2) is a better model than equation (1) when predicting the finish order in NASCAR races.

Table 2

Dependent Variable: FP
 Method: Least Squares
 Date: 02/15/07 Time: 19:06
 Sample: 1 1440
 Included observations: 1440

White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12.64312	1.834865	6.890492	0.0000
SP	0.391908	0.040777	9.611083	0.0000
DY*SP	-0.009497	0.003227	-2.942820	0.0033
TL	1.461933	0.812700	1.798859	0.0723
PC	0.759535	5.095338	0.149065	0.8815
DY*TL	-0.131260	0.060681	-2.163126	0.0307
DY	0.257965	0.127305	2.026356	0.0429
R-squared	0.103429	Mean dependent var		22.00139
Adjusted R-squared	0.099675	S.D. dependent var		12.41622
S.E. of regression	11.78119	Akaike info criterion		7.775735
Sum squared resid	198895.4	Schwarz criterion		7.801365
Log likelihood	-5591.529	F-statistic		27.55186
Durbin-Watson stat	0.424131	Prob(F-statistic)		0.000000

RESULTS

The model depicted in Table 2 is superior to that depicted in Table 1 since Table 2 shows a decline in the SSE and an increase in R^2 . This tells us that the interaction terms are important in explaining finish position. We reject the null hypothesis that the restricted model as detailed in Table 1 is valid. Although both track length (TL) and percentage of laps run under caution (PC) are insignificant in the restricted equation, we continue to include them in the model that includes the interaction terms because theoretically, they may add to the overall explanatory power of the model. The coefficient on SP is significant at the 1% level and is positive as hypothesized. The interaction term DY*SP is also significant at the 1% level and has a negative coefficient as hypothesized. In other words, as the number of driver years of experience increase and work together with the given starting position, the driver moves further toward the front. The coefficient on track length is insignificant but positive as hypothesized. The low t-value is most likely explained by the fact that there is not much variation in track length around the mean, which generates a high standard error of the coefficient and would thus contribute to a lower t value. The coefficient on PC is positive as hypothesized but insignificant. The coefficient on the interaction term DY*TL is significant at the 5% level and we hypothesize a negative sign as generated on this term. This means that as the driver years of experience increase and work together with the ability to negotiate different track lengths, the driver can be expected to move further toward the front. Finally, the coefficient on DY is positive as hypothesized and significant at the 5% level.

The F statistic is high and shows that the overall explanatory power of the model is statistically significant. The R^2 is 10% meaning that our model explains 10% of the variation in finish position among drivers in NASCAR races. Low values for R^2 tend to be expected for cross sectional data since multicollinearity may be operative and we may have omitted some important explanatory variables.

CONCLUSION

This paper set out to develop an empirical model based on theoretical hypotheses to explain the finish position of drivers in NASCAR races. The model clearly identifies starting position and driver years of experience as the most important variables that explain the finish position of each driver. Furthermore, the interaction between these two variables is also significant in explaining finish position. The practical implications of this model then, are

that drivers and their crews should focus heavily on perfecting the likely performance of the car they use for qualifying purposes.

This paper offers suggestions for further research. In order to improve R^2 , it may be advisable to explore the option of including additional explanatory variables. Another avenue worth exploring is how best to frame and utilize the variable associated with caution laps. Theoretically, the number of laps under caution is totally unpredictable prior to each race. Or is it? Are there some races that involve more crashes and hence caution laps than others? If that is not the case, then the randomness of caution laps would be picked up in the error term and contribute to a lower R^2 . On the other hand, again theoretically, the number of caution laps that occur during a race should have a significant effect on the outcome because caution laps allow for pit stops that give the crew time to make adjustments, add gasoline, and change tires, all of which should affect finish position.

Another avenue of further research will certainly be to enhance this model using several seasons and hence develop this model as a pooled time series cross sectional logit model.

REFERENCES

1. Allender, Mary, 2007. Are There a Higher than Expected Number of Early Life Critical Part Failures in NASCAR Vehicles? A Reliability Study, *The Sport Journal*, vol. 25, #1.
2. Becker, B.E. and M.A. Harold.1992. The Incentive Effects of Tournament Compensation Schemes. *Administrative Science Quarterly*, 37: 336-50.
3. Depken, C.A., and D.P. Wilson. 2004a. The Efficiency of the NASCAR Racing System: Initial Empirical Evidence. *Journal of Sports Economics*, 5(4): 371-86.
4. Fizel, John, ed., 2006. *Handbook of Sports Economics Research*. M.E. Sharpe.
5. Lazear, E.P. and S. Rosen. 1981. Rank Order Tournaments as Optimal Labor Contracts. *Journal of Political Economy*, 89(5): 841-64.
6. Lynch, J.G. and J.S. Zax.2000. The Rewards to Running: Prize Structure and Performance in Professional Road Racing. *Journal of Sports Economics*, 1(4):323-40.
7. Majety, Subba Rao, Miland Dawande, Jayant Rajgopal, 1999. Optimal Reliability Allocation with Discrete Cost-Reliability Data for Component. *Operations Research*, vol. 47, No 6.
8. Maloney, M.T. and McCormick. 2000. The Response of Workers to Wages in Tournaments. *Journal of Sports Economics*, 1(2): 99-123.
9. Martin, Mark. 2005. *NASCAR for Dummies*, 2d ed., Wiley.
10. Nalebuff, B.J. and J.E. Stiglitz. 1983. Prizes and Incentives: Towards a General Theory of Compensation and Competition. *Bell Journal of Economics*, 14 (Spring):21-34.
11. O’Keefe, Mary, Kip Viscusi, and Richard Zeckhauser. 1984. Economic Contests: Comparative Reward Schemes. *Journal of Labor Economics* 2(1): 27-56.
12. Pfitzner, Barry and Tracy Rishel. 2005. Do Reliable Predictors Exist for the Outcomes of NASCAR Races? *The Sport Journal*, vol. 8, no. 2.
13. Von Allmen, P. 2001. Is the Reward System in NASCAR Efficient? *Journal of Sports Economics*, 2(1): 62-79.
14. Wachtel, Gene. Engineer, Hendricks Motorsports, Personal Interview, 2006.

NOTES