Predicting a Driver's Finish in a NASCAR Race

Robert L. Andrews, Virginia Commonwealth University
Department of Management, Richmond, VA. 23284-4000
804-828-7101, rlandrew@vcu.edu

ABSTRACT

This paper examines an interesting problem of building a model using Excel to predict a NASCAR driver's finish in a race based on the data from the previous races for the season. The important modeling decisions are discussed. Several functional capabilities of Excel are illustrated in the model building process. The necessity of having a monotonic function to consistently find a minimum or maximum is also illustrated because the chosen loss function ends up not being monotonic and this means that a truly successful model was not attained with this process. In addition it is obvious that even though the binomial is a discrete distribution giving probability values for x=0 through x=42, it is not a proper distribution for modeling this phenomenon.

INTRODUCTION AND OVERVIEW

NASCAR standings are determined by points earned during the season's races. The number of points a driver earns in a race is determined by his/her finish in the race and his/her ability to be in the lead during the race. Bonus points are earned based on leading the race. Five points are earned if a driver leads for at least a lap and an additional five points are earned by leading for the most number of laps. Hence a driver's bonus points for a race will either be 0, 5 or 10. The points awarded for the place the driver finishes in the race for the purpose of clarity will be referred to in this paper as place points. Place points are not strictly linear based on the place the driver finishes in the race. The first place is awarded 180 place points plus there is the additional consideration of bonus points and these would be five or ten points given the fact that the winner at least was in the lead at the end of the last lap. There is a ten point decrease for second place and a five point decrease in place points awarded between places for the second through sixth place finishes. For the seventh through the eleventh place there is a decrease of four points. Then from the twelfth through the forty-third place there is a decrease of three points per place as is shown in Table 1.

Table 1

Place Points Awarded Based on Finishing Place

Finishing Place	Place Points	Finishing Place	Place Points	Finishing Place	Place Points	Finishing Place	Place Points	Finishing Place	Place Points
1st	180	10th	134	19th	106	28th	79	37th	52
2nd	170	11th	130	20th	103	29th	76	38th	49
3rd	165	12th	127	21st	100	30th	73	39th	46
4th	160	13th	124	22nd	97	31st	70	40th	43
5th	155	14th	121	23rd	94	32nd	67	41st	40
6th	150	15th	118	24th	91	33rd	64	42nd	37
7th	146	16th	115	25th	88	34th	61	43rd	34
8th	142	17th	112	26th	85	35th	58		
9th	138	18th	109	27th	82	36th	55		

The point accumulation situation for a season becomes even more complicated because at the end of the twenty-sixth race of the season the point totals for all drivers qualifying for the "chase" have their point totals adjusted. The top ten drivers and anyone else within 400 points of the points-leader qualify for the "chase." The driver with the most total points will begin the chase with 5,050 points. The driver in second place in the points standing will start with 5,045 and the awarding of points decreases by five points each time through all drivers qualifying for the chase. To put the 5,050 points in perspective, a

driver would only have a total of 4,940 if he or she earned the maximum number of points in all twenty-six races.

Qualifying for the "chase" places a great deal of interest and focus on the twenty-sixth race. Prior to the 2005 twenty-sixth race in Richmond, VA there were six drivers who were virtually sure of being in the chase with the remaining four positions to be decided in the twenty-sixth race. There were ten drivers who could mathematically qualify for the chase. The motivation for this paper came when I was called by a newspaper reporter who asked me if I could come up with a probability of qualifying for the chase for each of these ten drivers based on their performances in the preceding races. At the time I had no clue about how the NASCAR points were awarded for a race. I did know that it was not something that I could do in the two day time-frame for the reporter to be able to write an article prior to the race. The problem did intrigue me and this paper addresses this problem.

Overall the problem of making predictions about the likelihood of each driver qualifying for the chase is complicated for some of the drivers because their ability to qualify depends on how many points they get relative to the points earned by multiple other drivers. Some drivers can control their own destiny and qualify by earning a specified number of points. For them predicting their ability to qualify can be accomplished by merely developing a model for predicting the points that will be earned in the upcoming race. Another issue is what should be used as inputs into the model. Should I form a model using only the outcomes for the previous races of the season or should I try to include information from previous seasons? Even if I wanted to build a model that includes information from previous seasons it would certainly have to have a major component that uses the current season data. Hence my initial focus was exclusively on building a model that attempts to predict a driver's finish in the last pre-chase race of the season using the results for the preceding races of the season. Such a model is also essential if one attempts to deal with the more complicated situation involving the results of multiple drivers. Hence my attempts at developing this model will be the beginning point the focus of this article.

MODELING POINTS FOR A SINGLE RACE

As indicated in the introduction, the possible points earned by a driver in a race is a discrete variable with integer values but there is no regular interval between the possible values due to difference in the gaps between place points awarded based on where the driver finished and due to the awarding of bonus points. Due to the lack of linearity for the points the initial decision was to first develop a model for the finishing place in a race. Then the finishing place in the race can be easily translated in to a number of place points based on the NASCAR formula. Bonus points will be considered separately, if it is possible to develop a reasonably successful model for finishing place. Bonus points are separate from finishing place because with the exception of the winner, bonus points will not be directly dependent on where a driver finishes in a race. There is of course some statistical dependence here because the chances of earning bonus points for someone finishing 43 will not be the same as for someone finishing second in the race, even though it is possible for the 43rd place finisher to lead one or more laps while the 2nd place may finish without ever leading a lap.

The data that will be used to build the model for each driver were obtained for 2005 from a Fox Sports website http://msn.foxsports.com/nascar/cup/stats. This site has the race results arranged by driver and was chosen over other sites that just have results arranged by race. Excel was the analysis tool that was readily available and was used in the attempt to build a workable model. The results for a driver were first put into a frequency table to register the number finishes for the driver in places one through forty-three.

Having an accident or mechanical failure that would prematurely eliminate the driver from finishing the race clearly impacts the finishing place for the driver. I initially considered trying to separate this out creating an initial estimate of the probability of the driver having an accident or mechanical failure then

developing a model to predict the finishing place in the case of an accident for mechanical failure and another model to predict the finishing place if the driver was still running in the race at the end. With less than 30 races to provide the data for the modeling these would not be adequate to develop models for what amounts to three separate pieces. In fact for the 2005 season Rusty Wallace was running at the end of each race so there would be no information on which to try to determine what place he might finish if he if he had a mechanical failure or race ending accident. Consequently I abandoned this idea and turned my attention toward attempting to build a single model to predict the finishing place for a driver for the race. The possible values for finishing place range from first to forty-third. The goal is to have a model that provides a probability estimate for each of these values. Having a model that merely gives a point estimate of the finishing place as one would get from a regression type model would be of no real value for the original purpose of estimating the probability of a particular driver placing high enough in the 26th race to make it into the chase.

This situation does not fit into a situation for which there is known probability distribution. The binomial distribution with n=42 will provide forty-three probability values that can correspond to the 43 places. The key parameter for the distribution is the probability of success and will be represented by the Greek letter π. Since Excel is the chosen computational tool, it does has a binomial function in its list of statistical functions. For each driver the goal is to use the data from the previous races and try to estimate the value of π. For each value of x, where x is an integer value ranging from 0 to 42, the probability of that x can be calculated using n=42 and the value of π. The value of the finishing place will range from 1 to 43, hence the value of x can be determined by subtracting one from the finishing place and (finishing place) = x +1.

Being able to come up with a specific value of π for a specific driver based on past performance for the season means that a criterion must be established to be able to find a value of π that either maximizes for minimizes the value of the criterion function. Hence a loss function or some method of measuring the amount of disparity between the observed frequencies and the model predicted frequencies was developed. The squared differences between the observed relative frequency and the corresponding modeled probability were used. Hence for each driver there is a frequency table for the driver's finishing places. Since the number of places clearly outnumbers the number of races run, there will be several observed values of zero for Frequency as shown in Table 2. To build a model for a particular driver the goal is select a probability of success that will minimize the sum of squared errors. Table 2 presents a segment of a table for the driver Mark Martin that uses his previous finishes for races in the 2005 season to find an optimal value of π.

Table 2

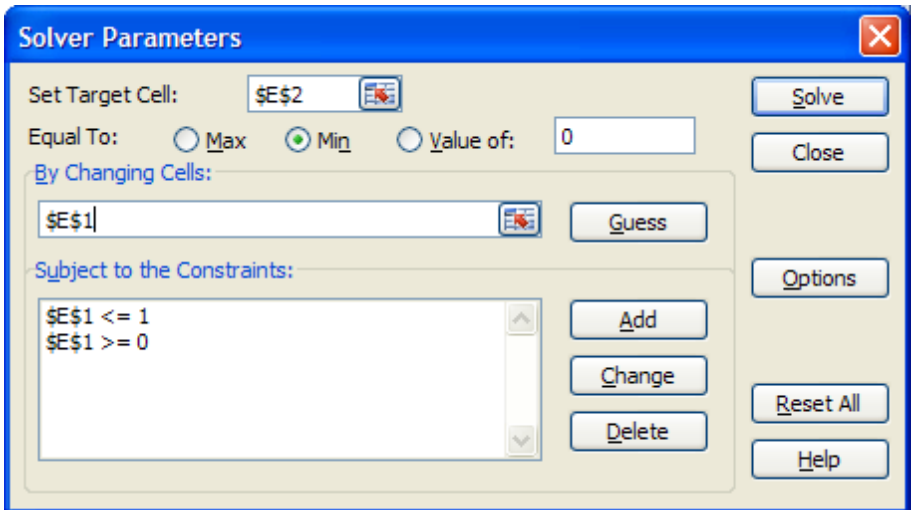
Example Table for Calculating a Value for π

Mark Martin			**p=**	**0.1208**		**0.1868**	**= Maximum B(x) Value**				
	Σ[RF(x) - B(x)]^2 =			**0.1143**		**6**	**= Most Likely Finish**				
Finishing Place	1	2	3	4	5	6	7	8	…	42	43
x for binomial	0	1	2	3	4	5	6	7	…	41	42
Frequency	0	0	4	2	0	1	4	0	…	0	0
RF(x) = Relative Frequency for x	0	0	0.16	0.08	0	0.04	0.16	0	…	0	0
B(x) = binomial probability of x	0.00448	0.02585	0.07283	0.13345	0.17882	0.18678	0.1583	0.11189	…	8.6E-37	2.8E-39
[RF(x) - B(x)]^2	2E-05	0.00067	0.0076	0.00286	0.03198	0.02154	2.9E-06	0.01252	…	7.5E-73	8E-78

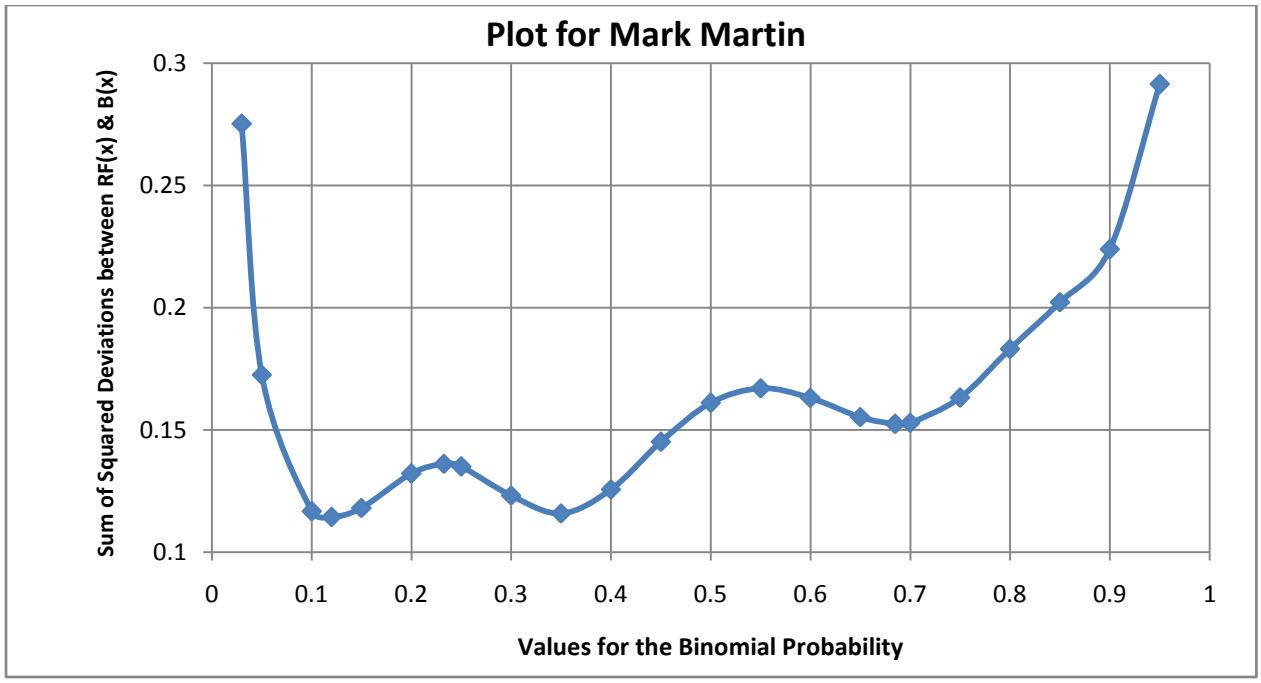
In Table 2 the values for Finishing Place range from 1 to 43 since there are 43 drivers in each NASCAR race. To get the value of x for the binomial the value of one is subtracted from each of the Finishing

Place values. The Frequency row contains values for the count of the number of times the driver finished in this place. The RF(x) row measures the relative frequency or proportion of the 2005 races for which the driver finished in this place. The B(x) row is the binomial probability calculated by the binomial function in Excel using the value for x from the table, number trials equal to 42 and the value for the probability of success equal to the value of π at the top of the worksheet. The $[\ RF(x) - B(x)\ ]^2$ row contains the values for the squared difference between the relative frequency row value and the binomial probability row value.

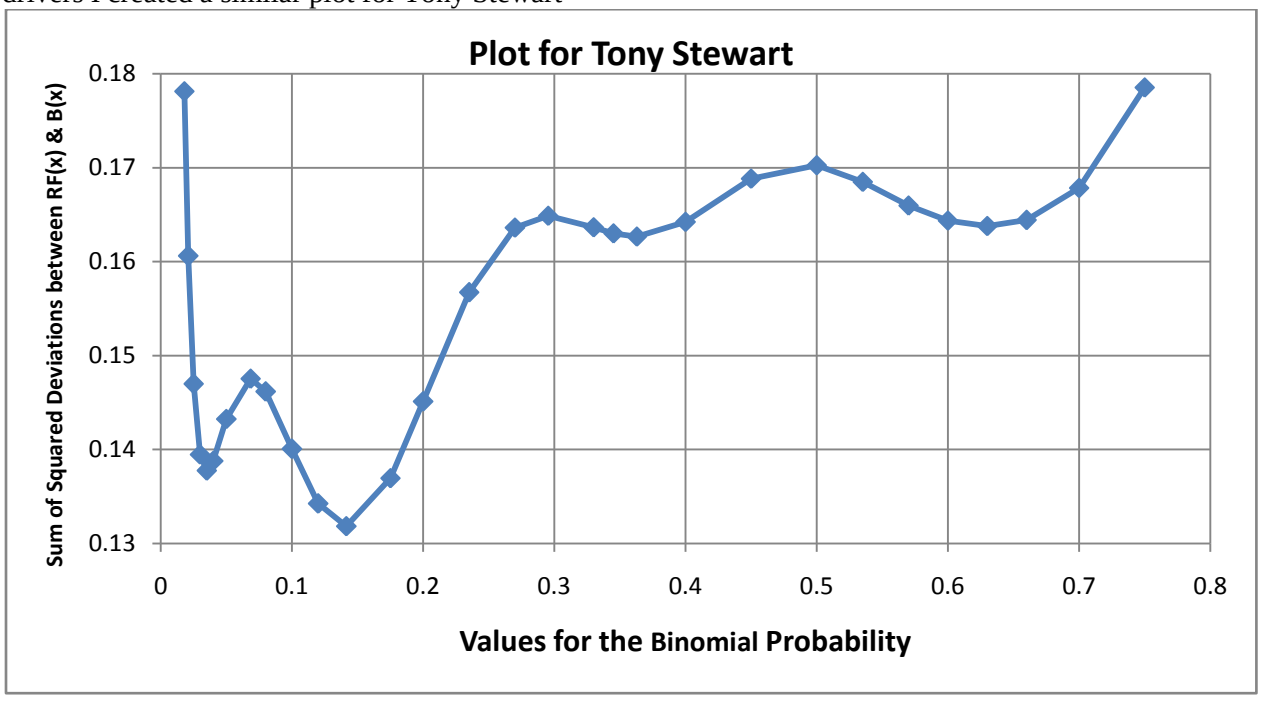
The sum of the squared differences was placed at the top of the sheet appearing below the value for π. The goal was to find a value of π that would minimize this sum of squared differences. Solver was used to accomplish this. To use Solver a formula must be located in cell, E2 in this example. Excel will alter the value in a designated input cell to either maximize, minimize or solve for a specified value of the function. The formula in the target cell must either directly or indirectly depend on the value in the cell Excel has been told to change. The value of π was determined by solver to minimize the sum of the squared differences for the 2005 season results for Mark Martin. The solver parameters below were used for the complete table like the one in Table 2 above. To make sure that the value for π was not outside the valid range for probabilities, between 0 and 1, I sent up constraints with these values as the limits for the cell containing the probability value. I discovered though that I obtained three different minimum values when I used different initial values for π, which tends to indicate that this sum of squared error function is not monotonic with a single local minimum over the range from 0 to 1.



Subsequently I plotted the sum of squared differences for Mark Martin as shown in the graph below. From it one can clearly see that there are in fact three local minima and two local maxima over the plotted range. This explains why I obtained three different values for the minimum (.1208, .3474 & .6866) when I varied the starting value for π.

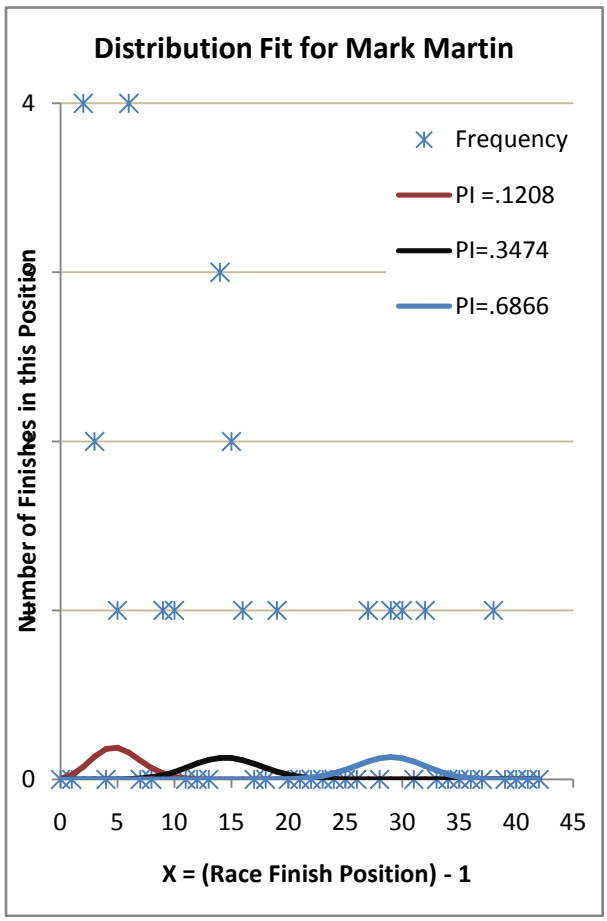


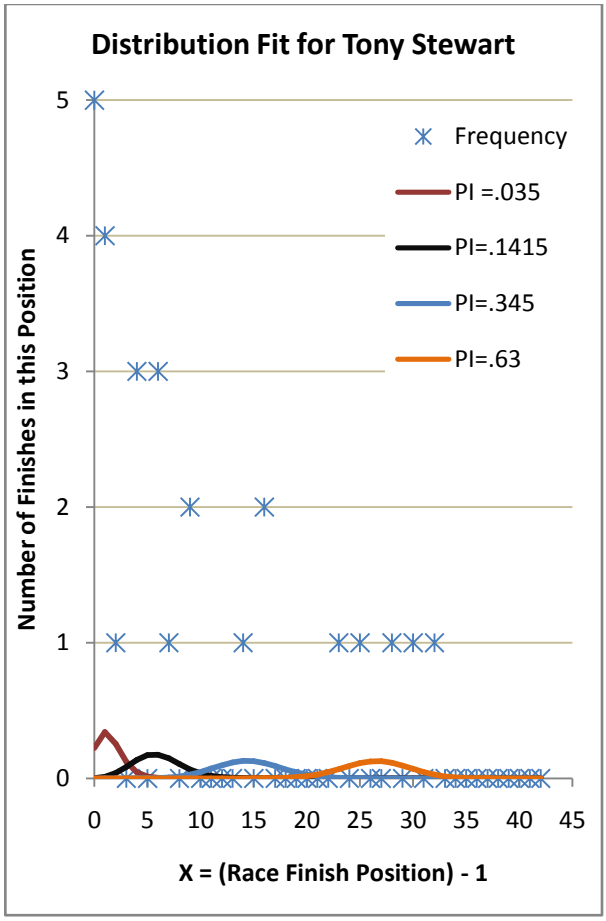
To see if this problem was unique to the data for Mark Martin or if similar problems existed for other drivers I created a similar plot for Tony Stewart



The plot for Tony Stewart shows that there are four local minima (.035, .1415, .345 & .630) and three local maxima for this function. As can be seen from the respective plots, for Mark Martin the global minimum is at $\pi = .1208$, while the global minimum for Tony Stewart is at $\pi = .1415$. Based on the binomial model this means that the expected finish place for Mark Martin would be 6.07 while the expected finish place for Tony Stewart would be 6.94. In the races used to build the model, the average finish for Tony Stewart was 10.07, while the average finish for Mark Martin was 13.7. We are seeing a disparity between the model and what was observed.

To better understand what is happening for these two drivers a separate graph was created for each showing the distribution of their respective finishes and the fitted distributions for the local minima values of π.





From the distribution graph for Mark Martin one can see that the three different fitted binomial distributions attempt to have the hump in a location that is allowing the curve to rise up toward a cluster of finishes above it. The four different fitted binomial distributions do the same thing for Tony Stewart. It is also clear that none of distributions is really doing a respectable job of approximating the distribution of finishes. The spread for the binomial is too narrow compared to the distribution of finishes. The binomial has its maximum variability with π=.5 and for π=.5 the standard deviation is about 2.6. Using the data for the previous finishes for both drivers the standard deviation for Mark Martin was 10.4 and the value for Tony Stewart was 10.2. For the 16 drivers who were in the chase or had the mathematical possibility of qualifying for the chase in the last race for the 2005 season, the standard deviations of their finishes ranged from 7.9 for Jeremy Mayfield to 14.1 for Kurt Busch. This further enforces the fact that the spread of the binomial is not adequate for modeling driver finishes. Hence to be successful in creating a method for estimating the probability of the finish of a driver the probability distribution will have to be much wider than the binomial.

SUMMARY

It is clear that I have not accomplished the stated goal of developing a model that would provide a reasonable estimate of a NASCAR driver's finish in a late season race based on the results for the season's previous races. I have clearly established that the binomial distribution, which will give discrete probabilities for each of the possible finish values and can be easily calculated with Excel, is not a proper distribution to use due to its limited variability. Also I will not be able be quoted in the newspaper with my model's probabilities for the likelihood of the different drivers qualifying for the chase when NASCAR comes to Richmond this year. Maybe with more thought and work I will have a model ready for next year.