

Horse Racing: the Gambler's Fallacy and Hot Hands

Andy Porter, University of Maryland

Bill Evans, Advisor, University of Maryland

Carlos Vegh, Honors Advisor, University of Maryland

1. Introduction

Horse racing in the United States is an economically interesting market. Bettors wager against one another in a pari-mutuel betting system. The track acts as a simplified market where goods are in the form of horses, and prices are in the form of odds (Sauer 1998). The outcome of the race is unknown beforehand and choices are made in uncertainty. These decisions are made throughout the day at racetracks, showcasing choices made over time as well as under uncertainty. This paper investigates the wagering behavior of these bettors as they make decisions over the day. I examine whether gaming decisions are impacted by logical falsehoods known as the gambler's fallacy and hot hands.

Independent events are unaffected by previous events. In the gambling world, it is often thought that gambler's do not fully comprehend the importance of independent events. Slot machine gamblers will save seats on certain machines that are "warming up" over a period of time. The idea is that the machine has to crank out of a winner eventually, thus the slot gamblers are 'due' after a period of losing. This mentality can be

found in many casino games as some gamblers think of past events as information to be used against the casino. A classic roulette example is that if black has been run 10 times in a row at a roulette table, one would incorrectly think that red was due – and place more money on this false insight (Sundali 2006). This is called the gambler’s fallacy. Hot hands is a very similar fallacy but it arrives at the opposite result. For instance, a roulette example is that if seven has won several times in a row, a gambler might think that the number five is more likely than the other numbers. This is called hot hands.

Play on the Maryland lottery demonstrates the gambler’s fallacy. Bettors refrained from buying tickets with numbers that won the day before by 33 percent, and still 10 percent two months later (Clotfelter and Cook, 1993). Similar results are found in greyhound races and thoroughbred races as well. In 13.8 percent of greyhound races the same number dog will win the following race, yet only 13.0 percent of the pool was wagered on them, holding other factors constant (Terrell, 1997). In horse racing, a large study has shown that people bet proportionally more on favorites after long shots have been winning (Metzger 1985). Bettors think the favorite is due after a loss and place their support behind it – despite the fact that it is a new race, with completely different horses.

Important to this study is Metzger’s paper, which used data from 12,000 races to demonstrate the existence of the gambler’s fallacy and hot hands. She found that after a longshot win or series of longshot wins, betting on the favorite was reduced. Similarly after a favorite lost, betting increased on the favorite. These two findings are a demonstration of the gambler’s fallacy and a similar idea, hot hands. Like the gambler’s

fallacy, a bettor believing in hot hands thinks that after an unlikely event occurs previously, it is more likely to happen presently.

The main purpose of Metzger's study was to show the existence of this behavioral phenomenon in an experimental field setting and not to model its magnitude. My paper improves upon Metzger's study by utilizing regression analysis while implementing critical control variables such as field size and race number that she left out of her analysis. Metzger's (1985) speculates that field size could be a reason for the gambler's fallacy and hot hands, in that as the day increased more horses might be run thereby lowering the betting on the favorite. In this study, I have included the number of horses in the race, the race number and in a separate regression the interaction term between the two.

In this paper I have investigated 3016 races from 17 different horse racing tracks from October through November 2006 to test for the existence of the gambler's fallacy and hot hands. The key outcome in my regressions is the percent of money wagered on the favorite in a given race. Bettors have been shown, in aggregate, to be very good at determining the relative probabilities of horses in a given race (Ali 1977, Sauer 1998). Indeed, in this study that the average probability assigned to a favorite by the public is 33.8 percent, while the favorites win about 34.6 percent of all races. These values are very close and in line with other horse racing studies (Coleman 2004).

To study the gambler's fallacy and hot hands, I investigate the effects of the results of previous races on the odds the favorite receives. These odds are converted into

actual probabilities that the aggregate public believes the favorite will win. To study the gambler's fallacy and hot hands, two dummy variables were introduced. One records a success if a longshot (defined here as a horse with under 6 percent probability) has won the previous race. The second measures whether a favorite (defined here as a horse with over 30 percent probability) loses the previous race. These two variables will be used to describe gambler's fallacy and hot hands effects in this study. In some regressions, an interaction term is utilized to detect whether the fallacy variables are biased when they occur at the same instance. Two other variables were introduced to control for fluctuations in the percentage bet on the favorite: the number of horses in a race and the race number (a measure of time during a racing day).

In this paper, the coefficients on the longshot winning and favorite losing variables were both statistically significant at the 5 percent level with the longshot winning variable near the 1 percent significance level. The interaction term between the longshot winning and favorite losing variable is only statistically significant at the 14 percent level.

This paper investigates the effects of previous, unrelated and independent races on the current favorite's probability of winning as defined by the public.

2. Data

The data for this project was taken from two websites, www.equibase.com and www.drf.com, representing Equibase and The Daily Racing Form respectively.

I used information from 3015 races from 17 horse racing tracks in North America during the period from October through November 2006. The tracks used in this paper are: Aqueduct, Bay Meadows, Belmont Park, Churchill Downs, Delaware Park, Finger Lakes, Hawthorne, Hollywood Park, Laurel Park, Los Alamitos, Meadowlands, Penn National, Philadelphia Park, Portland Meadows, Remington Park, Turf Paradise, and Woodbine.

The average amount of races per track in the dataset is 177 per track, the median is 179. The minimum is 101 races, the maximum is 262. The average field size (horses in race) per race is 8.3, the median is 8. Per track, the average minimum field size per race is 6.6, the maximum 9.9. The average minimum number of races is 4.7, the maximum 5.7.

Listed below in Figure 2.1 is a summary table of all the included tracks and statistics concerning the number of races used in the study: the average races per racing day per track and the average amount of horses racing in each race per racetrack.

Figure 2.1

	number of Races	avg. Races per day	avg Horses per race
Aqueduct	105	4.96	8.36
Bay Meadows	211	4.73	7.35
Belmont Park	189	5.24	8.32
Churchill Downs	124	5.69	9.90
Delaware Park	106	5.34	6.58

Finger Lakes	153	5.31	7.88
Hawthorne	227	5.05	8.35
Hollywood Park	101	4.73	7.89
Laurel Park	179	5.49	8.93
Los Alamitos	220	5.35	7.93
Meadowlands	202	4.95	7.06
Penn National	153	5.00	8.53
Philadelphia Park	162	5.04	7.96
Portland Meadows	160	4.97	7.59
Remington Park	229	5.31	9.67
Turf Paradise	232	4.97	7.94
Woodbine	262	5.29	9.57
Avg. Total	177.35	5.14	8.28
Avg. of Avgs		5.14	8.23
Median	179.00	5.00	8.00
Min	101	4.73	6.58
Max	262	5.69	9.90
Median	179	5.00	8.00

3. Regression Model

3.1 Dependent Variable

In my regression analysis, the dependent variable investigated is the market assessed probability of a favorite winning. This is written as FavoriteP. This value is taken from the odds given at the end of betting at each race. The odds are then converted into probabilities. The track takes out a percentage of money wagered each race which

manifests itself as the difference of probabilities for all horses in a given race from 100 percent. The percentage over 100 is the takeout rate for the track. Since different tracks use different takeout rates, I divide the probability as posted on the tote board by the total probability (which includes the takeout rate) of all horses in a race to produce the actual probability as defined by the bettors.

3. 2 Independent fallacy variables

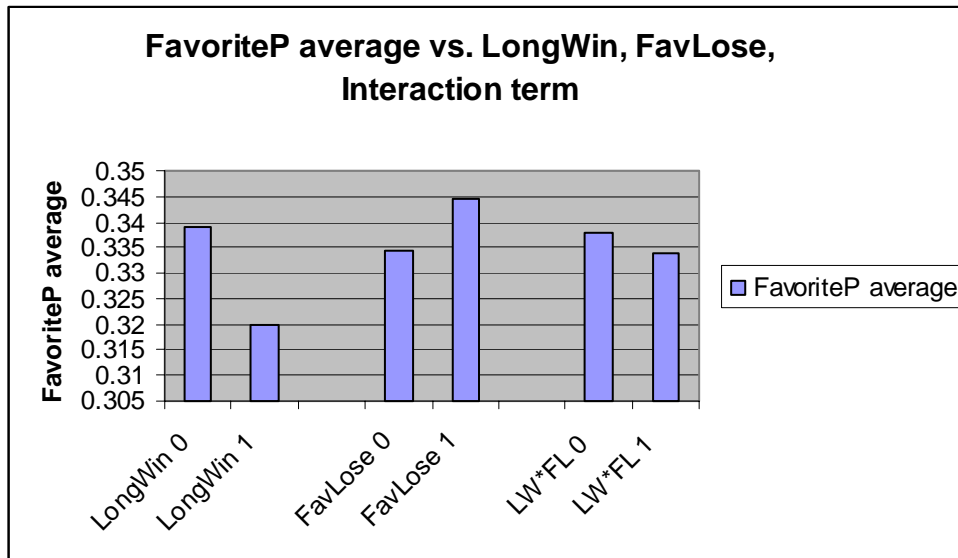
The independent variables measure the gamblers fallacy. I consider whether betting on the current rate is correlated with results from previous races. I construct three variables. The first is a dummy variable that whether the longshot won the previous race, (LongWin). A longshot is measured as a horse with a market assessed probability of winning of less than 6 percent.

Similarly, I construct a variable that measures whether the favorite lost the previous race. A favorite is defined as the top market assessed probability horse in a given race that is also higher than 30 percent.

In addition I include an interaction term dummy for the two above variables (LongWin*FavLose) in some regressions. If a longshot wins a race, the favorite must also lose. Thus I expect multicollinearity between these two variables. This interaction term should tell me whether or not the multicollinearity is strong enough to bias the coefficients of the fallacy variables.

The probabilities that will define longshots and favorites in this study were under 6 percent for longshots, and over 30 percent for favorites. What this means is that the variable LongWin will record a 1 if the public gives a horse a probability of under 6 percent, that horse won the race, and the race is not the first race.

Figure 3.1



The above graph demonstrates how the market assessed probability of the favorite changes in relation to the logical fallacy variables used in this study. FavoriteP, on the y-axis represents the probability assigned to the favorite horse in a race from the posted odds. The odds are converted to probabilities and divided by the total probability to account for the track's individual takeout. The values seen here for FavoriteP are averages of all 3015 races. The x-axis includes the variables LongWin and FavLose which as I describe in the variables section represent events of the previous race. LongWin records a 1 if a horse receiving under 6 percent probability wins the race. FavLose records a 1 if a favorite over 30 percent probability loses a race.

The results of the graph show that bettors had different betting patterns for longshots and favorites. The average bet on the favorite the race after a longshot wins is roughly 2 percent less than it was otherwise.

Conversely, if a 30 percent favorite loses the previous race, betting on the favorite increases by 1 percentage point more on average than otherwise. This graph is a clear example of the gambler's fallacy and hot hands at work. When LongWin records a success, the average amount bet on the favorite decreases. This shows that bettors put more money on non-favorites after a longshot win. This is an example of hot hands. When FavLose records a success, the average amount bet on the favorite increases. This exemplifies the gambler's fallacy in that bettors believe that since the favorite has lost the previous race, it is less likely to lose the current one.

Also, the interaction term between LongWin*FavLose shows a .4 percentage point drop between 0 and 1 values indicating that while the two variables LongWin and FavLose do mostly cancel each other out, LongWin is more powerful as was shown above.

The following table shows the averages of the gambler's fallacy variables for each track included in this study. Notation wise, avg. represents average, and avg. of avgs represents the average of values listed in the table as opposed to the total averages, which are averages of all the actual values in the dataset.

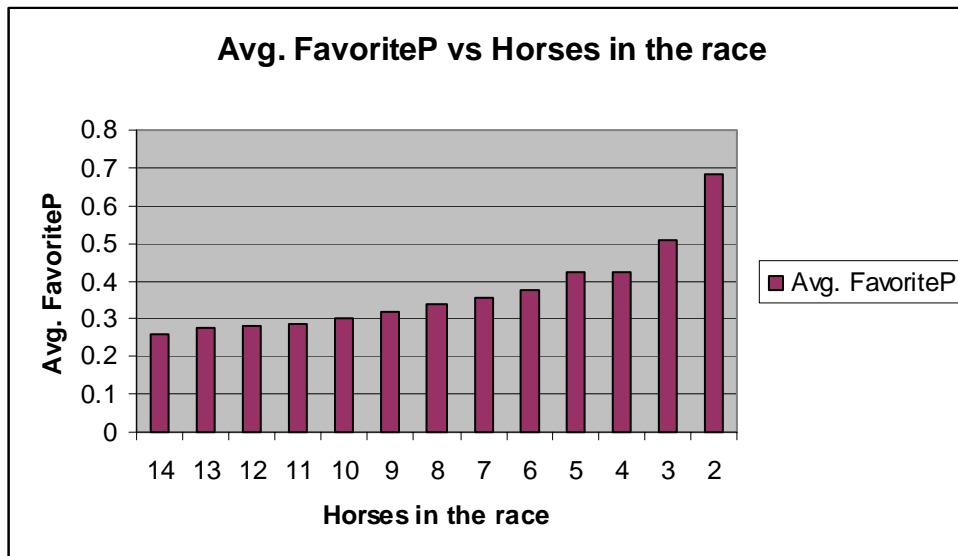
Figure 3.2

	Avg. LongWin	Avg. FavLose	Avg. LongWin* FavLose
Aqueduct	0.152380952	0.323809524	0.095238
Bay Meadows	0.056872038	0.398104265	0.042654
Belmont Park	0.052910053	0.322751323	0.026455
Churchill Downs	0.137096774	0.177419355	0.024194
Delaware Park	0.028301887	0.367924528	0.009434
Finger Lakes	0.052287582	0.281045752	0.039216
Hawthorne	0.061674009	0.264317181	0.017621
Hollywood Park	0.069306931	0.336633663	0.009901
Laurel Park	0.111731844	0.30726257	0.039106
Los Alamitos	0.059090909	0.377272727	0.040909
Meadowlands	0.04950495	0.287128713	0.034653
Penn National	0.071895425	0.352941176	0.03268
Philadelphia Park	0.067901235	0.339506173	0.037037
Portland Meadows	0.0875	0.3875	0.05
Remington Park	0.122270742	0.248908297	0.074236
Turf Paradise	0.103448276	0.387931034	0.077586
Woodbine	0.08778626	0.240458015	0.045802
Avg. Total	0.079933665	0.31641791	0.042454
Avg. Avgs	0.080703522	0.317700841	0.040984
Median (of avgs)	0.069306931	0.323809524	0.039106
Min	0.028301887	0.177419355	0.009434
Max	0.152380952	0.398104265	0.095238

3.3 Independent control variables

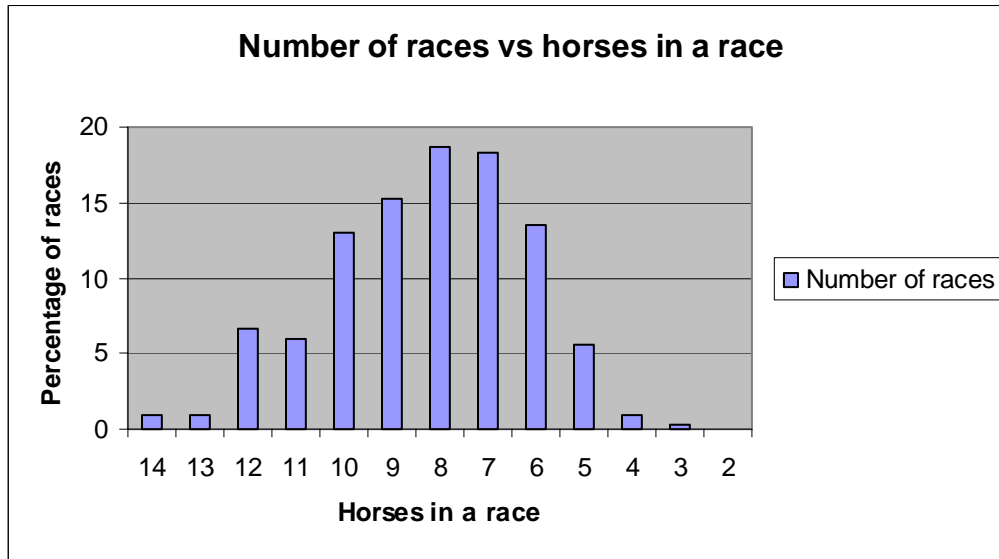
Since each race differs in field size (number of horses in a given race) I include a variable representing field size in my model (Horses) to account for the differences. The field size differs from race to race because of the design from the track's owners or by various maladies which befall the jockeys, trainers and horses. The field size will be affected the FavoriteP dependent variable as more horses enter the race the less people will be able to bet on any individual horse. I include the Horses variable to control for this effect.

Figure 3.2



As shown in the graph above (Figure 3.2), there is a trend towards increasing FavoriteP with decreasing field size.

Figure 3.3



This graph above (Figure 3.3) shows the distribution of races per field size. Almost 80 percent of races have between 6-10 horses. This is helpful to elucidate why the horses variable is necessary for this model, as most races are between 10-6 horses, but nearly 20 percent of races occur in the extremes where betting on the favorite changes strongly. A control variable is necessary to remove the bias the field size produces.

The final control variable I utilize is the race number (RaceNumber). This is the number given to a race to designate its place in the day. The day starts on race 1. This variable was included control for any persistent variation in betting patterns that occur throughout the day.

$$\text{The model: } FavoriteP = \beta_0 + \beta_1 LongWin + \beta_2 FavLose + \beta_3 Horses + \beta_4 RaceNumber + \mu$$

FavoriteP obviously depends on other variables aside from those in the model. It depends on the actual strength of the favorite horse in each race, something even the most skilled of handicappers cannot perfect. I focus on fluctuations caused by past events, ignoring the skills of the horse itself. The strength of the horse will be captured by the error term in each race.

4. Regression Results

Figure 4.1

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>
Intercept	0.491262	0.006991	70.27442	0	0.477555	0.504969	0.477555
LongWin (< 6%)	-0.01242	0.006041	-2.05609	0.03986	-0.02427	-0.00058	-0.02427
FavLose (> 30%)	0.009867	0.003528	2.796519	0.005198	0.002949	0.016786	0.002949
Horses	-0.01865	0.000816	-22.8735	6.7E-107	-0.02025	-0.01705	-0.02025
RaceNumber	-0.00027	0.000618	-0.43021	0.667072	-0.00148	0.000947	-0.00148
Significance F	0						
R Square	0.162318						
Adjusted R Square	0.161205						
Standard Error	0.088868						
Observations	3015						

The basic regression results are given above (Figure 4.1). Both LongWin and FavLose have p-values of under 5 percent and FavLose with a p-value of 0.5 percent is

under 1 percent significance. LongWin, with a p-value of 3.9 percent is strong as well. The Horses variable is very strong as one would predict, as the more horses enter a race the probability given to any one horse would decrease, the favorite included. Time within a given race day does not seem to be a factor as RaceNumber is clearly not statistically significant with a p-value of 68 percent. In later regressions I have done, I included a set of dummy variables, one for each race number.

The R^2 in this experiment is expectedly low, low but this is to be expected in a regression without detailed information about the characteristics of the horses in the race. That will depend on the given horses' strength in comparison to the field. I leave that for the error term, and focus here on the variables that deal with the gambler's fallacy.

LongWin has a coefficient at -.012, showing that when a longshot under 6 percent wins a race that is not the first race, betting on the favorite for the next race decreases by 1.2 percentage points. This is very interesting because after a longshot has won, bettors bet less on the favorites – choosing a riskier bet. This is in line with Metzger's study of the gambler's fallacy and hot hands. Hot hands occurred in her study when the longshot would win; bettors would bet more on longshots the next race. I have confirmed this result with the added robustness of regression analysis.

Also interesting is FavLose at .009867, this means the when a favorite over 30 percent loses a race, the bettors increase betting on the favorite by roughly 1 percentage points. This is direct agreement with the gambler's fallacy. That is, bettors expect the favorite to win around 1 percentage point more in relation to the field than after it has lost

the previous race. This in addition to the longshot win results, showing that bettors bet less on the favorite after a longshot wins, portrays a complex group of bettors. This group follows longshots after they win by betting less on favorites, and follows favorites after they lose by betting more on them.

Figure 4.2

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.491541751	0.006991849	70.30210997	0	0.477832464	0.505251037
LongWin (< 6%)	-0.021503862	0.008633129	-2.490853628	0.01279718	-0.038431293	-0.00457643
FavLose (>30%)	0.008168519	0.003711698	2.200749749	0.02782945	0.000890797	0.015446241
LongWin*FavLose	0.017782347	0.012076888	1.472427865	0.14100996	-0.005897444	0.041462138
Horses	-0.018665452	0.000815406	-22.89100444	0	-0.020264261	0.017066644
RaceNumber	-0.000203058	0.000619845	-0.327594012	0.74324144	-0.001418421	0.001012305

Significance F	0
R Square	0.162921436
Adjusted R Square	0.161530478
Standard Error	0.088850702
Observations	3015

I include above a second regression with an interaction term between the two fallacy variables (Figure 4.2). I include this because both of these variables could happen at the same time, and have done so in this study. If a race had a favorite with greater than 30 percent of the public's money but lost to a longshot with under 6 percent of the money, both variables would record a 1. Without the interaction term, this value would be -.013 or -1.3 percentage points showing a 1.3 percentage point reduction in betting the

favorite when both of these terms happen at once. When the interaction term is added, this is changed to .004 or a .04 percentage point increase in betting on the favorite. Since the interaction term failed to break even the 10 percent significance level, I cannot use its coefficient in serious discussion. Leaving out the interaction term introduces bias however, as is shown by this regression (Figure 4.2) without it.

With the interaction term, LongWin increases its effect to -2.15 percentage points as it sheds the effect of FavLose and LongWin occurring at the same instance. When the interaction term is in the model as above, LongWin is allowed to just model longshots, not the occurrence of both LongWin and FavLose, decreasing its positive bias. LongWin*FavLose did not pass the 10 percent statistical significance test.

This leaves the study with a problem to include biased coefficients or to leave in an interaction term variable that is correlated with the other gambler's fallacy variables and does not have a robust p-value. I argue that leaving the interaction term in the regression equation is the right decision even though it is not below 10 percent because it decreases bias, which is a bigger flaw in the first regression shown above. As a result of the p-value however, I feel that I cannot use its coefficient in any discussion of how its variables effect the equation.

5. Further Regression Results

The gambler's fallacy-type variables both have cutoffs, that is, LongWin is defined as longshots winning the previous race with under 6 percent probability. FavLose

is defined as the favorite losing the previous race with more than 30 percent probability. I include here a regression run (Figure 5.1) with those two variables without the percentage cutoffs. This is defined as the longest of longshots of the race winning the previous race (LW), and the favorite losing the previous race (FL) and an interaction term between the two.

Figure 5.1

	<i>Coefficients</i>	<i>Std. Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.49459	0.00698	70.90415	0.00000	0.48091	0.50826
LongestShotWon	-0.01171	0.01220	-0.95983	0.33722	-0.03563	0.01221
FavoriteLost	0.00016	0.00340	0.04609	0.96325	-0.00650	0.00681
Horses	-0.01885	0.00082	-23.07211	0.00000	-0.02045	-0.01725
RaceNumber	-0.00016	0.00063	-0.24798	0.80416	-0.00139	0.00108

As in shown in bold in the above regression, Longest Longshot, Favorite Losing, and the interaction terms p-values demonstrate that these variables are conventionally statistically insignificant at the 10 percent level. I conclude from this that the cutoffs introduced by the previously discussed variables LongWin and FavLose are much better at describing the gambler's fallacy effects in this study.

Here is a regression run (Figure 5.2) with an additional variable, FavLose2x, which records a 1 if the favorite (over 30 percent) has lost twice in a row.

Figure 5.2

	<i>Standard</i>					
	<i>Coefficients</i>	<i>Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.488110141	0.014526759	33.60076	1.938E-210	0.459626753	0.516593529
LongWin (< 6%)	-0.021528482	0.008631693	-2.49412	0.01268028	-0.0384531	-0.00460386
FavLose (> 30%)	0.01159453	0.004215417	2.750506	0.00598595	0.003329137	0.019859922
LongWin*FavLose	0.01769811	0.012076651	1.465482	0.14289413	-0.00598122	0.041377443
FavLose2x (> 30%)	-0.010641915	0.006148133	-1.73092	0.08356896	-0.02269689	0.001413057
RaceNumber	0.000605863	0.002504317	0.241928	0.80885279	-0.00430448	0.00551621
Horses	-0.018275222	0.00179459	-10.1835	5.7264E-24	-0.02179397	-0.01475647
RNH	-8.94502E-05	0.000294309	-0.30393	0.7612003	-0.00066652	0.000487618
Significance F	0					
R Square	0.16389106					
Adjusted R Square	0.16194468					
Standard Error	0.08882875					
Observations	3015					

Notice in the above regression how the coefficient for FavLose is positive, at around 1.2 percent, while FavLose2x, where the favorite has lost the previous two races, is negative at -1.1 percentage points. Obviously, when the favorite has lost the previous two races, he has lost the previous race as well. These two values would seemingly cancel each other out. This is an interesting twist in relation to the gambler's fallacy and hot hands. Bettors, while wagering more for favorites when they lose, drop back to 'normal' levels when the favorite has lost again. While the p-value for FavLose2x is under the 10 percent mark, at 6 percent it is not exactly robust. In separate regression,

interaction terms of all stripes have been tried – none of which proved to be useful or significant at conventional levels.

In addition to the above regression, other variables such as the favorite winning the last race, the favorite winning the last two races, extreme favorite losing the last race, favorite's losing the last three races, second favorite losing the last race, etc. Of all those tested, only LongWin, FavLose and the additional regression of FavLose2x were shown to be significant at conventional levels of 5 percent or 10 percent.

Also integral to this discussion is the statistics on the favorite winning and losing. In all races in this study, the favorite won 34.6 percent of the races, losing 65.4 percent. The LongWin variable occurred in 6 percent of all races. FavLose occurred in 31.6 percent of races. Both variables occur at the same time in 4.2 percent of all races. The variables I use do not include the first race as it does not have any previous races to utilize the gambler's fallacy or hot hands. This will cause their percentages to be lower than if I did include the first race. Those percentages do illustrate that these occurrences are not rare. FavLose happens with great regularity at 34.6 percent. This shows that favorites that have a probability of over 30 percent and still managing to lose the race are not uncommon, and do present a problem dealt with by gamblers on a regular basis.

As expected Horses is all regressions is highly significant and negative. This shows that an increase of one horse to the field lowers the betting on the favorite by roughly 1.9 percent. This effect does not have any direct implications for the gambler's fallacy variables but by taking it into account in the regression, I have moved it out of the

error term and removed its effects from potentially introducing bias onto the gambler's fallacy variables LongWin and FavLose.

Also it is worth mentioning that horses (field size) could be correlated with race number, meaning that as the race day wore on the number of horses increased or decreased. This could affect the gambler's fallacy variables if it were true because of their time-dependent nature, in that they cannot be on the first race. However a regression (Figure 5.3) including an interaction term between race number and horses (field size) sets this to rest.

Figure 5.3

	<i>Coefficients</i>	<i>Std. Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.48854	0.01453	33.62396	0.00000	0.46005	0.51703
LongWin	-0.02150	0.00863	-2.48962	0.01284	-0.03843	-0.00457
FavLose	0.00814	0.00371	2.19155	0.02849	0.00086	0.01542
LongWin*FavLose	0.01773	0.01208	1.46784	0.14225	-0.00595	0.04142
Horses	-0.01829	0.00180	-10.18752	0.00000	-0.02181	-0.01477
RaceNumber	0.00037	0.00250	0.14729	0.88291	-0.00454	0.00527
RN*Horses	-0.00007	0.00029	-0.23582	0.81358	-0.00065	0.00051

While the p-value for Horses remains very strong, RaceNumber and the interaction term RaceN*Horses are highly insignificant showing conclusively that time is not a factor in this gambler's fallacy model.

In Ali (1977), he found that betting on the favorite decreases during the day and this finding is supported in a weaker extent by Metzger. Below (Figure 5.4) I report estimates from a regression that includes a dummy variable for each race during the day, excluding the first.

Figure 5.4

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.500780944	0.009460032	52.9364932	0	0.482232138	0.5193298
LongWin	-0.02189178	0.008668133	-2.5255474	0.01160292	-0.03888787	-0.0048957
FavLose	0.008598296	0.003806047	2.25911463	0.02394774	0.001135571	0.016061
LongWin*FavLose	0.01834211	0.012113083	1.51423958	0.13007043	-0.00540868	0.0420929
Horses	-0.02023277	0.0013409	-15.088948	1.2138E-49	-0.02286195	-0.0176036
RNH	0.000296511	0.000179715	1.64989036	0.09907005	-5.5867E-05	0.0006489
R2	0.001265806	0.006973719	0.18151092	0.85597883	-0.01240795	0.0149396
R3	-0.00345433	0.007101506	-0.4864222	0.62670337	-0.01737864	0.01047
R4	-0.01350985	0.007419281	-1.8209114	0.06871988	-0.02805725	0.0010375
R5	-0.00021005	0.008022168	-0.0261843	0.97911209	-0.01593956	0.0155195
R6	-0.01560869	0.008914504	-1.7509323	0.08005976	-0.03308785	0.0018705
R7	-0.02493884	0.009902598	-2.5184136	0.0118401	-0.04435541	-0.0055223
R8	-0.0089406	0.010952214	-0.8163278	0.41437748	-0.03041521	0.012534
R9	-0.02876475	0.012641228	-2.2754714	0.02294789	-0.0535511	-0.0039784
R10	-0.0219729	0.016054716	-1.3686257	0.17121877	-0.05345226	0.0095065

Significance F	0
R Square	0.168263535
Adjusted R Square	0.164382098
Standard Error	0.088699483

As shown by this regression, LongWin and FavLose are both still exhibit nearly identical coefficients and both pass the conventional 5 percent significance test. It is worth mentioning that a few race dummies did also pass the 5 percent significance test, the 9th and 7th race. Others came close and a few other race dummies passed below the 10 percent significance test. This could be due to large major stakes races being held on a pre-set and traditional race of the day, where the field is harder to judge and thus the amount bet on the favorite might decrease. From this regression, time does certainly play a factor with relation to the amount bet on the favorite; however it does not change the coefficients of the variables that represent the gambler's fallacy and hot hands, LongWin and FavLose.

The gambler's fallacy and hot hands have been suggested as a possibly reason for longshot bias. Longshot bias has been well documented in horse racing literature. In the UK and North America, researchers since the 1970s have shown that bettors under bet the favorite in relation to its chance of winning, and over bet longshots. An interesting side note is that while favorites are under bet in the UK and North America, in Asia the opposite is true. Favorites are over bet and longshots are under bet.

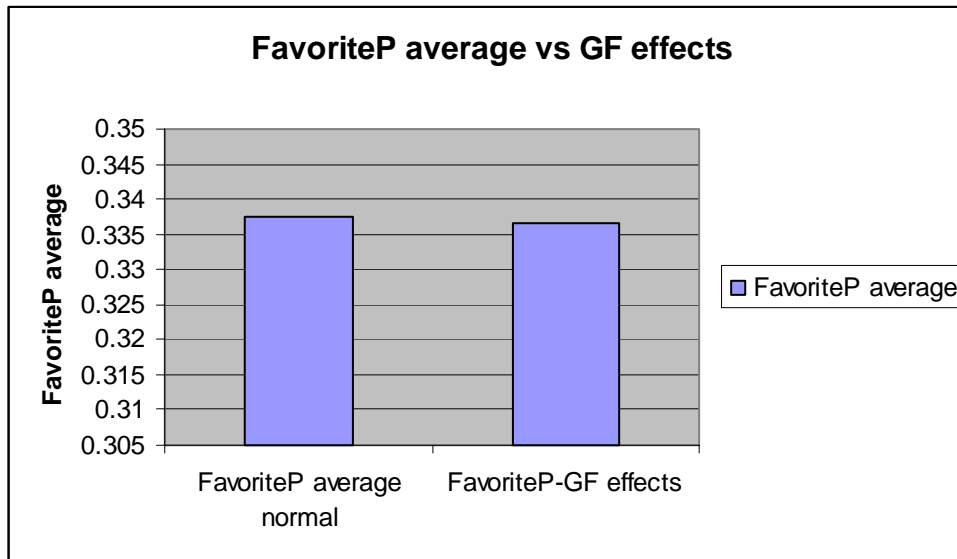
My study has shown the longshot bias as well, with the public giving the favorites 33.8 percent probability while the favorite wins 34.6 percent of the time. These probabilities are called the subjective probability (public) and objective probability

(actual winning percentage). This is in line with all major studies on longshot bias (Coleman 2004).

Coleman (2004) specifically discusses the gambler's fallacy and hot hands as being a potential reason for the longshot bias, in that bettors by utilizing the gambler's fallacy and hot hands do so with a result of under betting the favorite. This theory has not been investigated previously. In Metzger's (1985) study of the gambler's fallacy and hot hands, she does not use this as a springboard to investigate the longshot bias.

In my study I took the coefficients from the first regression and added their effects to a new estimated FavoriteP. This means that when a LongWin or FavLose records a 1, I have subtracted the coefficient from the FavoriteP given in the race. This will show how bettors would bet on the favorite had they not had the gambler's fallacy or hot hands on their minds. In my study favorites win 34.6 percent of the time, and bettors bet on them 33.8 percent of the time. With the coefficients added to each instance of LongWin and FavLose, it is estimated that bettors bet 33.7 percent on favorites, a .1 percent decrease, shown in Figure 5.4 below.

Figure 5.4

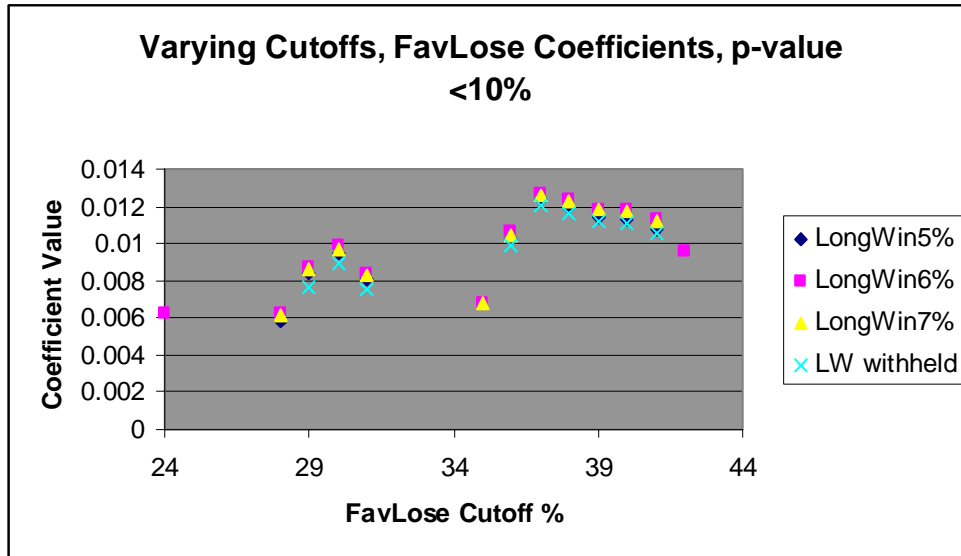


This shows that longshot bias is not explained the gambler’s fallacy and hot hands. It is still possible that other, perhaps more robust variables dealing with the gambler’s fallacy and hot hands could explain the longshot bias but in this study it showed that it actually made it worse. I would recommend more studies to investigate this interesting theory.

The cutoff value of 6 percent assigned to a longshot is arbitrary in the sense that there is no commonly accepted value for a “longshot”. Similarly, there is no commonly accepted value designating a remarkable favorite. In this study, I chose 30 percent to include the median and average values of favorites. I now include a plot representing regressions run using varying cutoff values of FavLose, as well as varying cutoff values of LongWin. The plot that follows are included to demonstrate that the values chosen throughout the paper may be arbitrary, but those coefficients are not flukes. The chart (figure 5.5) below plots the percent cutoff that defines a favorite on the x-axis, the

coefficient value on the y-axis, for varying cutoff values of LongWin. In this chart, I plot the regression coefficients from FavLose cutoff values from 24-44 percent and LongWin cutoffs from 5-7 percent that were statistically significant at the 10 percent level.

Figure 5.5



From the 80 regressions run, 40 percent had p-values statistically significant at the 10 percent level. This shows that for a wide range of cutoffs values for FavLose and LongWin are relevant, and the cutoff values of 6 percent for LongWin and 30 percent for FavLose are not flukes whatsoever. Looking at the graph above, there is some interesting structure to the FavLose coefficients. In particular, the values from 37-41 percent are steadily decreasing. This can be interpreted as the impact (the coefficient) decreasing as the previous losing favorite becomes stronger. This is puzzling from the gambler's fallacy perspective, as one would expect the coefficients to slope upwards with increasing cutoffs. This would mean that as the favorite who lost the previous race increases in strength, the impact of the gambler's fallacy (the coefficient) would increase. This

interesting and conflicting result merits a study that delves deeper into how these cutoff values.

5. Conclusion

The dummy variables LongWin, defined as a longshot receiving under 6 percent of the public's money wins the previous race, and FavLose, defined as a favorite receiving over 30 percent of the public's money loses the previous race, were both shown to be significant at conventional levels (5 percent). Another variable, FavLose2x which recorded a 1 if the favorite lost two races in a row, was significant at the 10 percent level.

LongWin has a negative coefficient, meaning that if a longshot as defined in this study wins, the next race the public will bet less money on the favorite in the race. FavLose has a positive coefficient, showcasing the opposite. LongWin has a relatively higher magnitude but doesn't occur in nearly as many races as FavLose. This is an example of hot hands, a variant of the gambler's fallacy discussed throughout this paper, whereby the bettors embrace a longshot after it has won the previous race. This, like the gambler's fallacy is a statistical misconception. It would be interesting to see of other examples of hot hands in other markets.

Interestingly, FavLose2x was negative. This could mean that when the favorite loses twice the bettors switch between the gambler's fallacy and hot hands. FavLose's positive coefficient invokes the gambler's fallacy, as the public thinks the favorite is more likely to win after a previous loss. However, once the favorite has lost twice in a

row the coefficient switches negatively, showing an embrace of hot hands. That is after seeing the favorite lose twice in a row bettors abandon the favorite thinking that they are more prone to losing. That last bit is speculation however, it could also be the case that bettors after two favorite loses abandon logical fallacies all together, as the coefficients for FavLose and FavLose2x as shown in the regression roughly cancel each other out. This happens every time FavLose2x occurs because FavLose must have occurred as well.

I infer causality between the gambler's fallacy and hot hands variables and my dependent variable, FavoriteP due to the time difference between the variables LongWin and FavLose and FavoriteP. I do not simply believe that this study implies correlation from these results because of the time the variables were recorded. Since LongWin and FavLose are both variables that by definition occur before the betting of the next race of the day, for those variables to be correlated with FavoriteP implies directly that FavoriteP is following LongWin and FavLose.

In this study I have used 17 different racetracks with on average 177 races per track. Because of this, I have not used interaction terms between the variables. I could potentially be overlooking specific tracks in my dataset that have an extremely high proportion of bettors believing in logical fallacies, therefore contaminating the rest of the dataset. While I cannot say for with certainty that this is not the case, it does not seem very plausible given the high number of tracks and the lack of a dominating track presence in the data pool and the relative agreement among the variables between tracks as measured by the summary statistics.

The underlying reasons bettors might use for adopting the gambler's fallacy and hot hands are many. One could be the allure of a big payoff. Indeed, longshots under 6 percent have comparatively very large payoffs due to their unlikeliness of winning. Bettors could see the payoff on the tote board and salivate on the possibility of winning it themselves, thereby putting less money than they would have on the favorite. In addition however the relatively low coefficient of -2.15 percentage points does indicate that it is not an overriding factor, or at least the bettors who would follow a big payday are lower in number in relation to the hardliner handicappers who would discourage adopting the gambler's fallacy and hot hands as statistical truths. In a following study, it would be interesting to record the payoffs as an independent variable and investigate if previous high payoffs are indeed correlated with lower betting on the favorite.

The gambler's fallacy and hot hands are followed by gamblers despite being statistically incorrect at a basic level. Some believers have been adamant of their logistical cult-like devotion to these fallacies. A quick look to the wikipedia discussion page for the gambler's fallacy entry reveals much about some faithful individuals. Suffice to say that while there are undoubtedly some, perhaps most individuals who utilize the gambler's fallacy do so quickly and maybe subconsciously, there are some who go out of their way to write long, faithful and statistically incorrect arguments believing in the gambler's fallacy. I speculate that these diehards are low in number when compared to the overall horse track betting populace.

In future studies, I would very much like to investigate whether hot hands and gambler's fallacy are wholly relegated to gambling, and not other facets of life. If so, then

these studies concerning horse racing could be dismissed as simply niches of strange phenomenon, rather than illuminating examples of economic markets. But, if not then this data might be globally illuminating.

For instance, if a relatively unknown small startup suddenly did very well in the stock market, would that increase investment in other relatively unknown startups? It can be said that the populace attending and gambling might have different statistical reasoning abilities than the general populace, and they might specifically be subject to the gambler's fallacy more than non-bettors. However, I think it is more convincing that the gamblers are just people from the populace, who have an average set of statistical tools like everyone else, and use these in pursuits they deem worthwhile. In this case those pursuits involve gambling. In this view I would estimate that should an event occur, that people in general would color their statistical reasoning surrounding that event as more or less likely to happen, even though both events are unrelated and independent.

Also, in a future study I would like to see models of FavoriteP against the actual payoffs received from the winner of the previous race. From this, I could test whether or not big payoffs decreased betting on the favorite as I have hypothesized earlier in the paper. This can be extrapolated into the examples used above as well. Payoffs could represent payoffs in any sense, for example the lottery. Do huge jackpot winners that are televised increase purchase of lottery tickets? These and other similar questions could provide interesting answers.

Gamblers are different from people who do not gamble, in that they intentionally risk their money for the potential for of a greater reward. In everyday life for most people, their money is not risked. Year long salaries are sought, tenure at universities is yearned for, and generally risk is avoided. Risk, the gambler's fallacy and hot hands then do not seem to enter into everyday life.

In this study I investigated the logical fallacies manifesting themselves as the gambler's fallacy and hot hands. These fallacies were shown in this study to be significant at conventional levels for 3015 horse races at 17 tracks in North America. I conclude that these fallacies are indeed present in horse racing and present a compelling picture of the misunderstanding of statistics by gamblers, and possibly the general public as a whole. Studies involving non-gambling related decision making are needed to expand this discussion and contribute to this study of these irrational economic agents, people.

Bibliography

Ali, M. M., Probability and utility estimates for racetrack bettors, *Journal of Political Economy*, 1977, 85, 803–15.

Clotfelter, C. T. and Cook, P. J., The gamblers fallacy in lottery play, *Management Science*, 1983, 12, 1521–5.

Coleman, Les, New light on the longshot bias, *Applied Economics*, 2004, 36, 315-326

Terrell, D., Biases in assessments of probabilities: new evidence from greyhound racing, Louisiana State University Economics, 1997, *Working Paper*, No 97–28.

Metzger, M. A. Biases in betting: an application of laboratory findings, *Psychological Reports*, 1985, 56, 883–8.

Sauer, Raymond D. The Economics of Wagering Markets, *Journal of Economic Literature*, 1998, 36, 2021-2064

Sundali, James, Biases in casino betting, the hot hand and the gambler's fallacy, *Judgement and Decision Making*, 1, 2006, 1-12