Regression Analysis

Zach Lorenzen

5/14/2012

Introduction

Gaming and Starcraft II

Competitive gaming is often overlooked as a sport. However, many games are competed in regularly such as Halo and Starcraft II. How much skill could it take to compete competitively though? It doesn't necessarily take a lot of energy to sit in front of a computer or television and play video games, does it? Starcraft II in particular can take a lot of speed, dexterity, and endurance to be able to play at a competitive level.

APM, or actions per minute, is a common way of measuring someone's speed in game, and generally thought as a good measure of one's skill. The faster you play, the better you play. During the game, players must control an army while managing workers and production at their main base. Any time a player clicks his or her mouse, or hits a key on the keyboard, it is an action. Professionals that play and compete in tournaments regularly can often reach an average of three hundred actions per minute, while it isn't unheard of for a player's APM to spike around the six hundred mark at any given time. All of this must be done while they are planning their strategies and tactics, to outplay their opponent.

In this study I collected data from over two thousand Starcraft players on how fast they play, their age, and other factors that relate to their gameplay. In the game, each player chooses one of three races to play as. In this paper when I refer to the race variable, it will not be in reference to a person's race (Black, Hispanic, Caucasian, etc.), but in reference to the player's chosen race in-game. Each race is extremely different and has different buildings, units, strategies, and tactics. I hypothesized that some races, being more aggressive in nature, could require more speed than the others. Players also have the option to play as a random race, randomly being assigned one of the three races every game.

I also asked the players about the league they play in, which I thought would directly correlate with their speed. There are a total of seven different leagues, which the game places you in after you complete a number of 'placement' matches online. The leagues are, from lowest to highest: Bronze, Silver, Gold, Platinum, Diamond, Masters, and GrandMasters. There are only two hundred players in the GrandMasters league for each region, and they are considered the best players from each region.

Collecting Data

There are a couple of potential problems present due to the method of collecting data. In creating the survey, I was presented with the problem as to how to distribute the survey to collect a representative sample of all Starcraft players. At the time of writing this, there are about 3.7 million players worldwide. By the end of the survey, I received approximately 2,500 responses. The only way I could reach these people was by advertising my survey on various places on the internet. I posted my survey on the Facebook walls of a few very well-known people in the Starcraft community. I also posted my survey on the official Starcraft II forum, Team Liquid's forum, and Reddit. Team Liquid is often considered to be where the heart of the Starcraft community is. Team Liquid is one of many professional teams that compete regularly in tournaments. Reddit also receives a lot of traffic, though isn't solely a Starcraft site. Reddit is often considered 'the front page of the internet.' I received the majority of my responses from both Reddit and Team Liquid. The problem with only posting here is that the survey will only reach extremely active players, who are active in the community. This means that I will generally reach the better players, or the players that are striving to improve, rather than the more casual players.

To complicate matters even more, shortly after I posted my survey on these sites, Blizzard (the developer and publisher of Starcraft II) released an update which offered another way to measure APM.

This new method was called EPM, or Effective Actions per Minute, and ignored useless actions, or actions that were superfluous. Several people responded to my survey, asking me if they should answer my survey with the original APM measurement, or EPM. I responded as soon as I could, telling them to use the original measurement since that is what I started with. Unfortunately, many people responded before I could settle the issue, which could cause some misinformation in the survey. This even violates one of the important assumptions we make when using the ordinary least squares method, that there is no measurement error. Therefore we must keep in mind that our results could be biased and inconsistent.

Initial Changes

I will be using APM as the dependent variable, with age, gender, hand size, league, and race as the independent variables. With how the survey was set up, I could not directly use the raw data from the survey itself. I had to make some changes. Originally, since APM could range from anywhere close to zero to 300, I gave ranges for them to answer. I divided the possible answers into ranges of twenty, (one to twenty, twenty one to forty, etc.). So that I could use the data in a regression, I averaged the ranges out (one to twenty I changed to ten, twenty one to forty I changed to thirty, etc.). For age, I took the highest value in the range. So if someone answered they were between the ages of fifteen and eighteen, I listed them as eighteen. The variables for gender, race, and league were each transformed into dummy variables. That is, the value is either a zero or one for each possibility in each variable. For example, race was divided into four separate variables: Zerg, Terran, Protoss, and Random, each representing the possible choices that each player has when choosing a race. If a player answered Terran for race in the survey, after my transformation of the variable it will be recorded as a value of zero for Zerg, Protoss, and Random, and a value of one for Terran.

Regression Data

Basic/Descriptive Statistics

The average reported APM was about 100. This means that the actual average APM is somewhere between 90 and 110. The standard deviation of the APM is about 50. About 70% of the sample is within one standard deviation of the average, and about 95% are within two standard deviations of the average. Therefore about 70% of people have an APM somewhere between 50 and 150, while about 95% of people have an average between 0 and 200.

For the races, the proportions are as follows: 0.22 players play Terran, 0.36 players play Protoss, 0.36 players play Zerg, and the last 0.06 players play as a random race. For league, the proportions are: 0.05 are in Bronze league, 0.09 are in Silver league, 0.15 are in Gold league, 0.24 are in Platinum league, 0.25 are in Diamond league, 0.19 are in Masters League, and 0.01 are in GrandMasters league.

The average age is 22, and the standard deviation is about 4 years. This means that about 70% of players are between 18 and 26. About 95% of players are between 14 and 30. About 0.98 players that answered my survey were male. The average size of a players hand is about 6.75 inches, or 17 cm, measured from the base of their palm to the tip of the middle finger. The standard deviation is about 0.9 inches. About 70% of players have a hand size between about 5.85 and 7.65. About 95% of players have a hand size between 4.95 and 8.55. It should also be noted that there is some missing data with hand size, therefore when I start running regressions; I will delete all samples that are missing this data.

Regressions

I produced some initial regressions and scatter plots which helped me identify some outliers. Those outliers, based on my knowledge of the game and my data, were deleted. That is, there probably aren't any players over the age of 50 that are within the highest caliber of players. Also looking at the basic scatter plot of APM vs. Age, there might be a nonlinear relationship between the two. The scatter plot looked similar to a –log transformation, so I created a new variable that transformed age into the negative natural logarithm of age. Running this regression again, I actually received a lower R square value, and a higher standard error of regression that the original variable yielded. Since the point of transforming variables like this is to induce more linearity, and the entire point of the Ordinary Least Squares method is to minimize the error of regression, I will not use the transformed variable.

Now, instead of looking at one variable at a time, I ran a regression using all five of my variables. The initial regression shows that age and gender have a negative relationship with APM while hand size has a positive relationship with APM. The leagues and races are a little more complicated, as they were separated into dummy variables. Each league, as you go from Bronze to Silver, Silver to Gold, etc. the predicted APM increases. As for race, I predict that Terran and Protoss players would play slower than players who choose random, while Zerg players play the fastest of the four.

There are some problems with this regression, however. There are issues with statistical significance with Gender, Hand Size, the Silver League dummy variable, and all the Race variables. All of these fall short of my 95% confidence interval. This could be because of some multicollinearity within the data. Somewhat high multicollinearity is present among the race terms, and the league terms. However, since there is very little multicollinearity within gender and hand size, we can probably dismiss

those as not significant. It isn't surprising, to say that gender has no influence on how fast someone's APM is. Also, it shouldn't be surprising that hand size isn't significant either. Keyboards are small enough that all but the smallest of hands can reach across them with ease. Therefore I will remove Gender and Hand Size from the regression.

Running the regression without gender and hand size had some interesting effects on the significance of both the race variables, and some league variables. The Silver league variable significance skyrocketed, with a p-value of around .95. The p-values for the race variables generally improved, but not nearly enough. They ended up at .58 for Race-T and .75 for Race-P. Since I left out Race-R from my regression, the coefficient for Race-T and Race-P measure the difference between that variable, and Race-R. And since the p-values do not meet my 95% confidence, they are not statistically different than zero. If the coefficients of these dummy variables are not statistically different than 0, then we can combine the dummy variables that are insignificant with the one left out of the regression. That is, we can combine Race-R, Race-T, and Race-P into one dummy variable. So now we have two dummy variables for the four possible races, Race-RTP and Race-Z. I will leave Race-RTP out of the regression from now on for simplicity's sake.

The same can be done for League-S. Since League-S's coefficient is not statistically different than 0, we can combine the two dummy variables League-S and League-B. I will leave League-BS out of the regression while keeping all the other dummy variables in the regression.

Final Regression

Now, our regression equation has three variables in it, two of which have been broken up into dummy variables, some of which have been combined even. First we have age, which I am treating as an interval level variable, as opposed to breaking it up into dummy variables. This is because there is

such a large range of ages, there would be too many dummy variables to keep track of. So I am leaving age as an interval variable to keep it as simple as possible. In our current regression, age has a coefficient of -1.305. This means that as age increases, our predicted value for APM will decrease. More specifically, for every year older someone is, we would predict someone's APM to decrease by a little more than one.

Our second variable is the League variable. This was initially separated into seven dummy variables. We have combined two of the variables though, so we are left with six. To avoid perfect multicollinearity we leave one of the variables out of the regression equation. I chose to leave out the combined dummy variable, League-BS. That leaves League-G, League-P, League-D, League-M, and League-G. All of the coefficients for these dummy variables are positive, which means that the higher league someone is, we would predict their APM to increase, rather than decrease. Since we left out League-BS, if someone is in Bronze or Silver league, there wouldn't be any change from the constant in our equation. However, if the player is in Gold league, their predicted APM would increase by about 14. If they are in Platinum league, their predicted APM would increase by about 26. Diamond players would have about a 48 APM increase compared to Bronze and Silver players. Master players would have about a 75 APM increase compared to Bronze and Silver players. Finally Grand Master players would have about a 156 increase in APM compared Bronze and Silver players.

Finally our last variable is Race. This was another variable that was changed to dummy variables. However, like league, I was able to combine some of the dummy variables. Not all of them were able to be combined however. Race-R, Race-P, and Race-T were all able to be combined. Race-Z I left as a separate dummy variable though. This means that players that play random, Terran, or Protoss all statistically play at about the same speed, after filtering out the effect of age and league that is. Players that play Zerg, however, seem to play faster than players that play the other races. Zerg players have a predicted APM of about 10 more than the other races.

Our constant coefficient in the regression is about 90. This value is more of an equalizing number, used to balance the equation and make sense. Theoretically, someone who has a zero value for all the other variables would have an APM of 90. However, that doesn't have any meaning in reality, since that person would play either Protoss, Terran, or Random, would be in either Bronze or Silver league, and would be zero years old. Someone who isn't even a year old could never be playing this game, in fact the youngest players are probably about eight or nine years old, and those players probably don't even play competitively. A more reasonable way to look at the variable is that there has to be a starting point somewhere. The coefficients for the other variables (Age, League, and Race) change the predicted value from the starting point of 90.

If we were to write out the regression equation it would look something like:

$$Y = 89.593 + 1.305X + 13.642X_G + 26.280X_P + 47.904X_D + 74.877X_M + 155.789X_{GM} + 10.092X_Z .$$

Where X is age, X_G is Gold league, X_P is Platinum league, X_D is Diamond league, X_M is Masters League, X_{GM} is Grandmaster league, and X_Z is Zerg race. As an example, if we were to look at a twenty year old Zerg player in the Bronze league, I would predict his APM to be about 74. If we were to look at an eighteen year old Terran player in the Grandmaster league, I would predict his APM to be about 222.

Looking at possible problems with this final regression, we can rule out multicollinearity. The smallest tolerance value we have is .491 (for League-D), which is easily high enough to indicate that there isn't a problem concerning multicollinearity in the regression, especially with a sample size of 2309. The R square value is 0.363 (Adjusted R Square is 0.362) and the Standard Error of Regression is 40.023.

Looking at the partial regression plots can be confusing since many of the points are overlapped with other points. There is nothing that I can see that would indicate any major problems, however. The same conclusion is reaches when looking at a graph of the residuals vs. the variables, and the residuals vs. the predicted value.

Conclusion

Finally, looking at the final regression, there are many things of interest to note, some surprising and some not so surprising. First it should be noted that we have solved any problems of multicollinearity by either removing variables with low significance or by collapsing dummy variables. It should also be noted that we checked to see if age had a non-linear relationship with our dependent variable. We concluded that we could not find a nonlinear relationship more appropriate for the model than the general linear relationship.

Gender was removed from the regression because of a low significance, that is, it did not meet our 95% confidence interval. This is not surprising for a couple of reasons. Firstly we did not have a large sample of women. The population of Starcraft players is largely dominated by males, and so we had a very low amount of female responses compared to male responses. It could be that we did not have enough data from women to discover a difference (if there is one). I would hypothesize myself that there is no difference between the speed of women and men in playing Starcraft.

Hand size was also removed from the regression because it did not meet our confidence. Hand size was included in the first place to see if there was a correlation at all. I never thought there really would be, but I thought I needed to check anyway. The keyboard is small enough that most people can reach across it without a problem, so it is not surprising that hand size is not significant in predicting APM.

Age turned out to be a significant variable in predicting APM. This is, too, should not be surprising. Much like the world of more athletic sports, such as football, as players get older, generally they slow down. Likewise, in Starcraft gaming, as a player gets older we would predict their APM to be

slower. Another similarity between the two is the average age is centered on a fairly low age. While an average age for professional football players is around 25 or 26 (according to espn.com), while in Starcraft II, according to my sample, the average age is about 22.

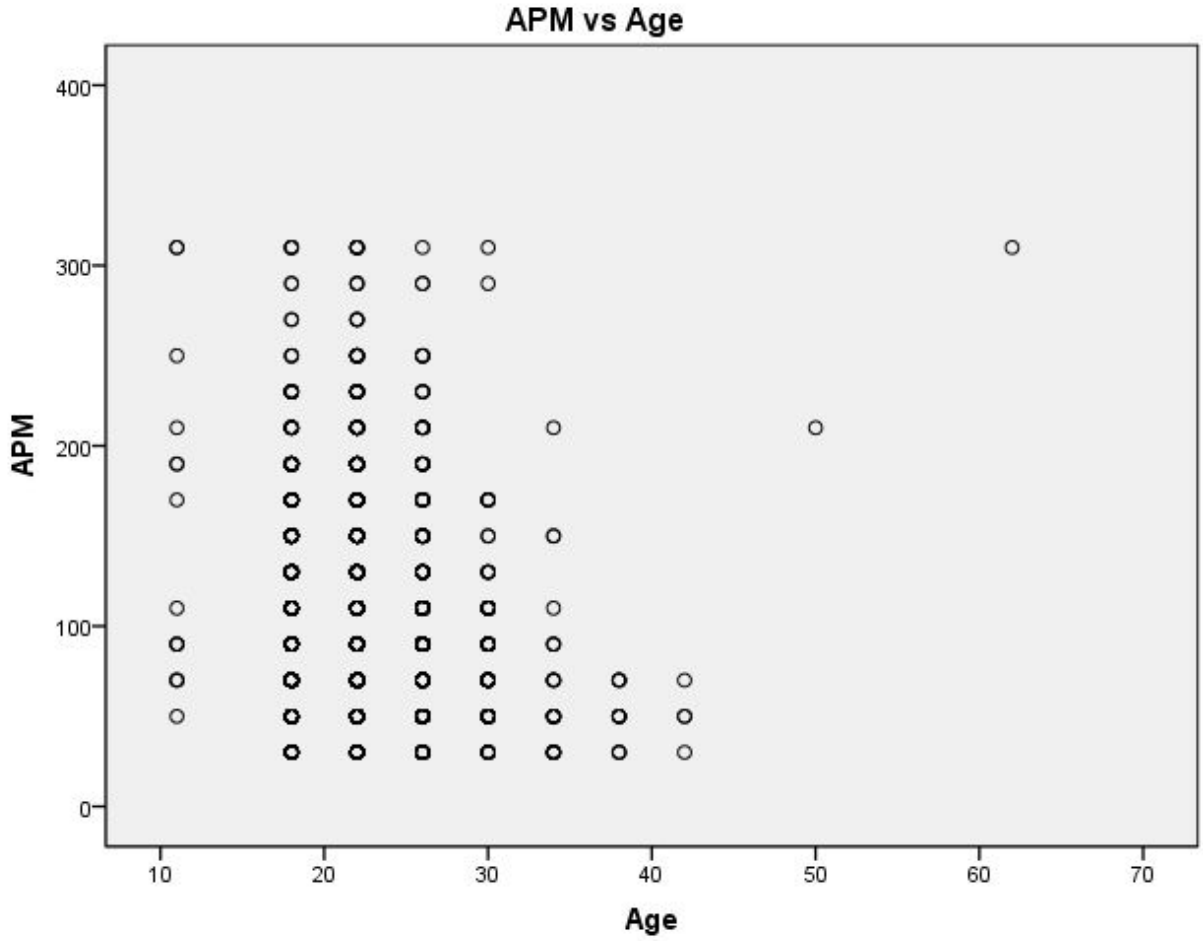
Some race variables turned out to be significant in predicting APM. This partially surprised me, as I guessed that all races were balanced in such a way that they required a similar amount of skill and speed to play. I've seen many people claim that Zerg players play faster than other players, however they claimed this without any evidence to back them up. According to my data, Zerg in fact does play faster than the other race. In general, this race is usually a little more aggressive than the other races, and the mechanics operate quite a bit differently than the other races. The other races were all collapsed into a single dummy variable. This goes along with my guess that race wouldn't be a significant predictor in APM, being that there is no difference between the Protoss, Terran, and Random races. While the Zerg race doesn't have a huge impact on APM, it does have a significant impact.

The most unsurprising result is that League is significant in predicting APM. APM is usually a measure of one's skill at Starcraft. That is, as one's APM gets higher, so does one's skill. In the same way, League is also usually a measure of one's skill at Starcraft. So it isn't surprising in the slightest that league is a significant predictor in APM. It should be noted again that there isn't a significant difference between the Bronze and Silver leagues however. This could be because the difference between the two leagues is not necessarily skill in speed or mechanics, but intuition. Silver players probably have better decision making than Bronze players, who are usually players that have just begun playing the game.

SPSS OUTPUT

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
APM	2311	30	310	101.76	50.304
Race-T	2311	0	1	.22	.416
Race-P	2311	0	1	.36	.480
Race-Z	2311	0	1	.36	.481
Race-R	2311	0	1	.06	.231
League-B	2311	0	1	.05	.227
League-S	2311	0	1	.09	.293
League-G	2311	0	1	.15	.356
League-P	2311	0	1	.24	.427
League-D	2311	0	1	.25	.436
League-M	2311	0	1	.19	.395
League-GM	2311	0	1	.01	.119
Age	2311	11	62	22.15	4.072
Gender	2311	0	1	.98	.155
Hand Size	1810	0	6	3.00	1.187
Valid N (listwise)	1810				



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.168[a]	.028	.028	49.388

a. Predictors: (Constant), logAGE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	96.033	9.613		9.990	.000		
	Age	-1.335	.243	-.105	-5.490	.000	.965	1.036
	Gender	-10.224	6.327	-.031	-1.616	.106	.980	1.020
	Hand Size	1.462	.804	.034	1.819	.069	.984	1.017
	League-S	2.373	4.967	.014	.478	.633	.404	2.472
	League-G	15.651	4.659	.110	3.359	.001	.330	3.032
	League-P	28.975	4.388	.250	6.604	.000	.247	4.043
	League-D	50.591	4.386	.437	11.535	.000	.246	4.064
	League-M	77.615	4.541	.599	17.092	.000	.288	3.473
	League-GM	153.211	8.818	.363	17.374	.000	.808	1.238
	Race-T	-.656	4.397	-.005	-.149	.881	.267	3.742
	Race-P	-1.653	4.214	-.016	-.392	.695	.217	4.605
	Race-Z	7.815	4.219	.075	1.852	.064	.218	4.595

a. Dependent Variable: APM

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	89.632	7.000		12.804	.000		
	League-S	.264	4.495	.002	.059	.953	.402	2.490
	League-G	13.928	4.185	.099	3.328	.001	.313	3.200
	League-P	26.477	3.973	.226	6.665	.000	.241	4.147
	League-D	48.168	3.946	.419	12.205	.000	.234	4.265
	League-M	75.002	4.061	.591	18.467	.000	.270	3.705
	League-GM	155.801	7.841	.369	19.871	.000	.801	1.248
	Race-T	2.213	3.959	.018	.559	.576	.256	3.909
	Race-P	-1.215	3.807	-.012	-.319	.750	.208	4.806
	Race-Z	10.178	3.801	.098	2.677	.007	.208	4.812
	Age	-1.320	.214	-.104	-6.183	.000	.978	1.022

a. Dependent Variable: APM

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	89.593	5.340		16.777	.000		
	Age	-1.305	.213	-.103	-6.124	.000	.981	1.019
	League-G	13.642	3.058	.097	4.461	.000	.585	1.709
	League-P	26.280	2.753	.224	9.545	.000	.502	1.993
	League-D	47.904	2.727	.417	17.569	.000	.491	2.036
	League-M	74.877	2.897	.590	25.849	.000	.530	1.885
	League-GM	155.789	7.303	.369	21.332	.000	.923	1.083
	Race-Z	10.092	1.736	.097	5.812	.000	.996	1.004

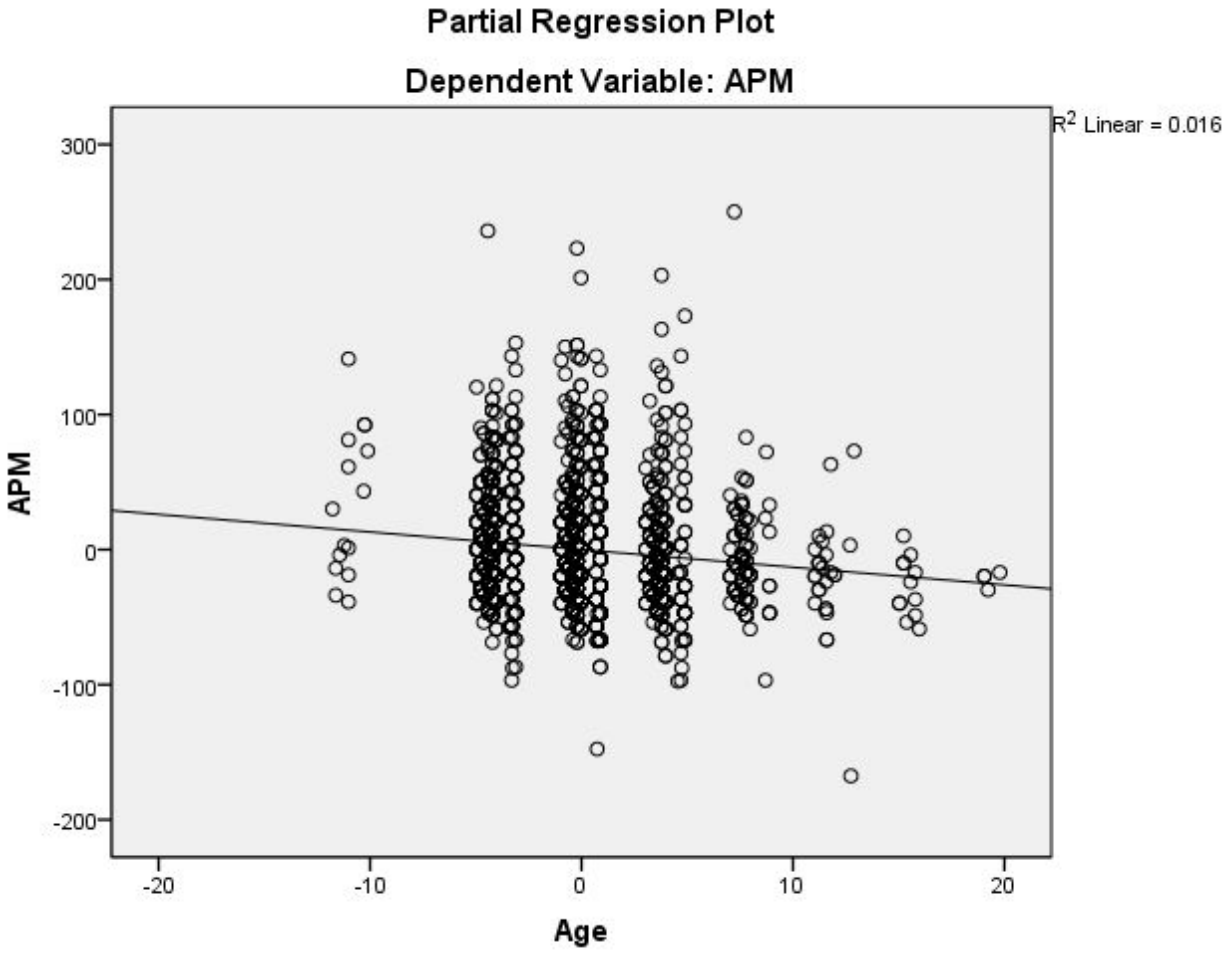
a. Dependent Variable: APM

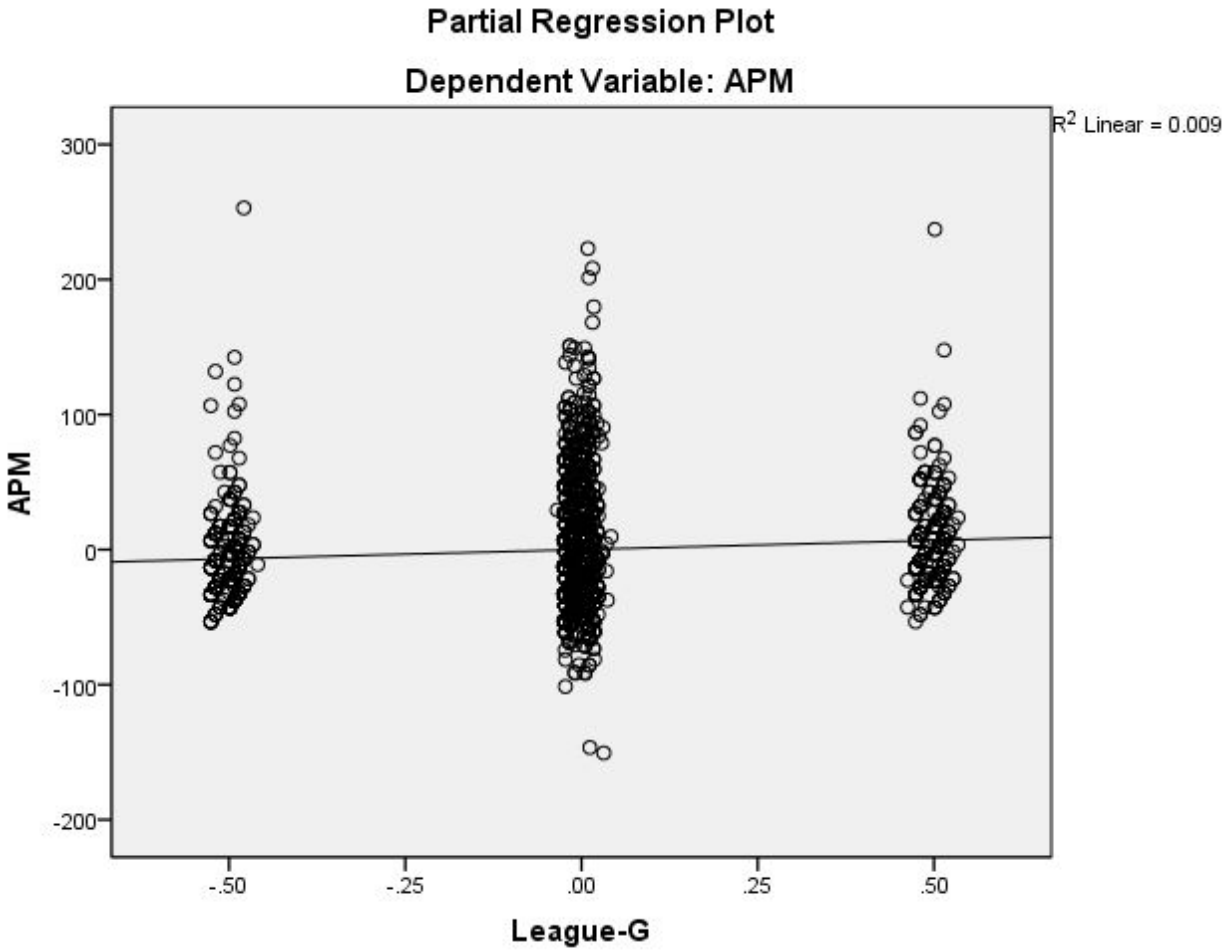
Model Summary[b]

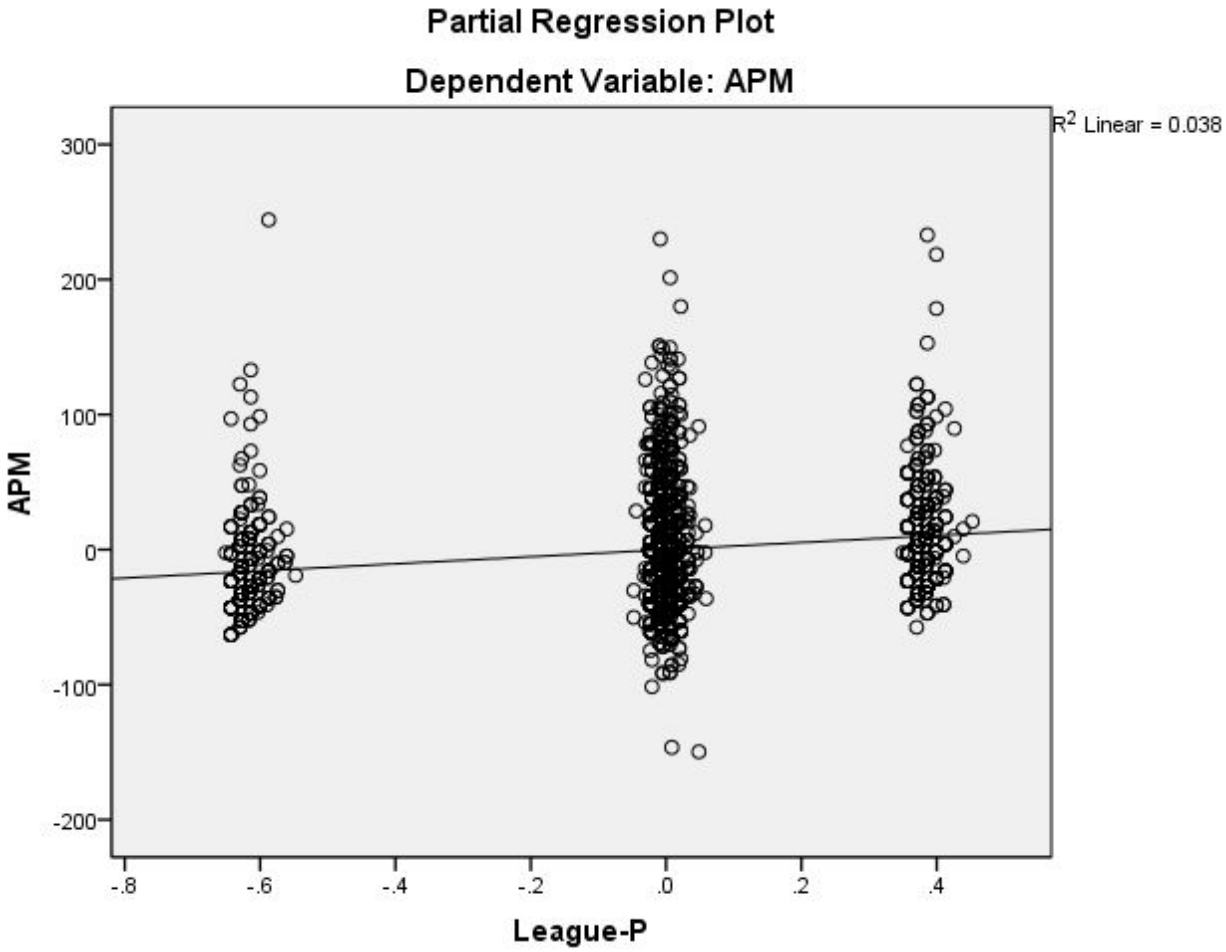
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.603[a]	.363	.362	40.023

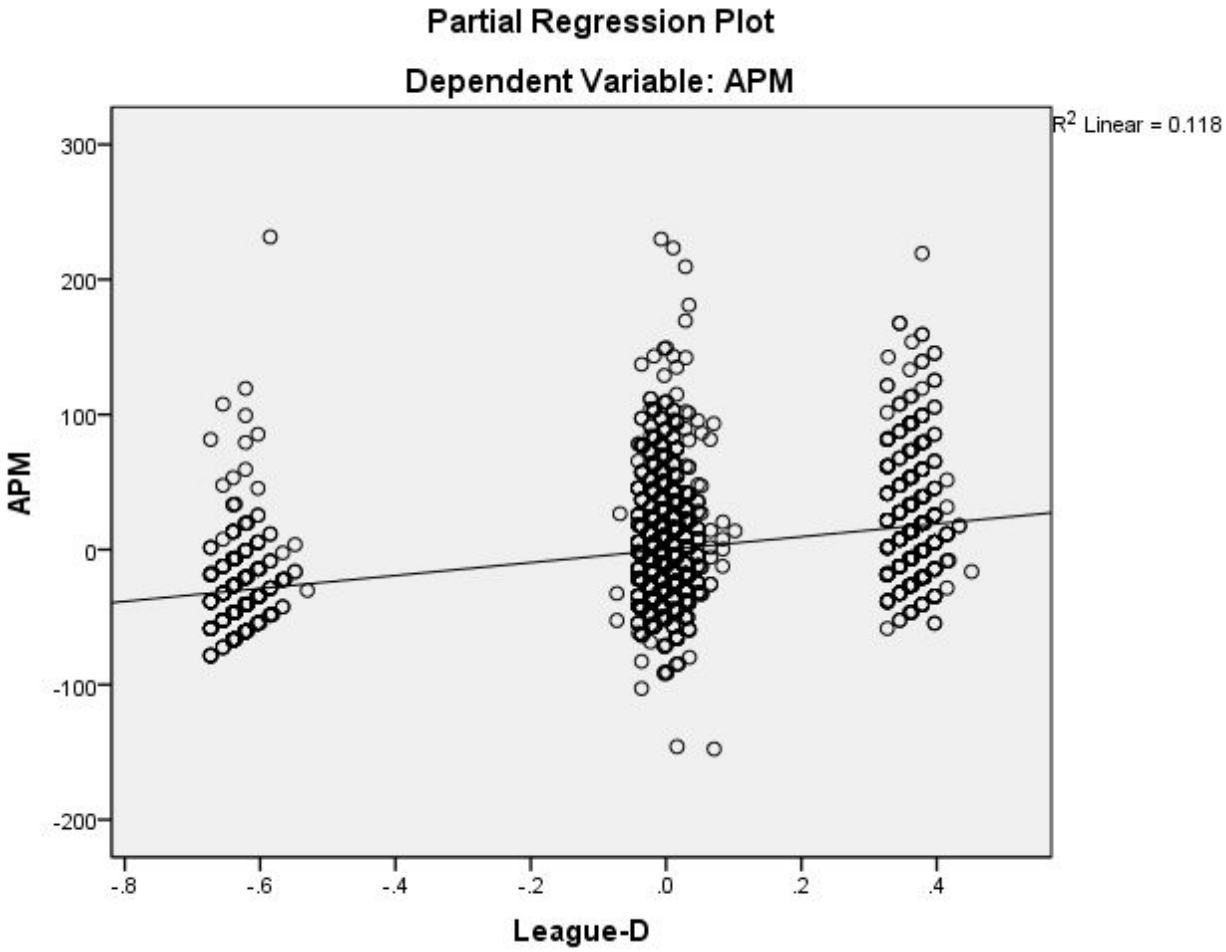
a. Predictors: (Constant), Race-Z, League-GM, Age, League-G, League-M, League-P, League-D

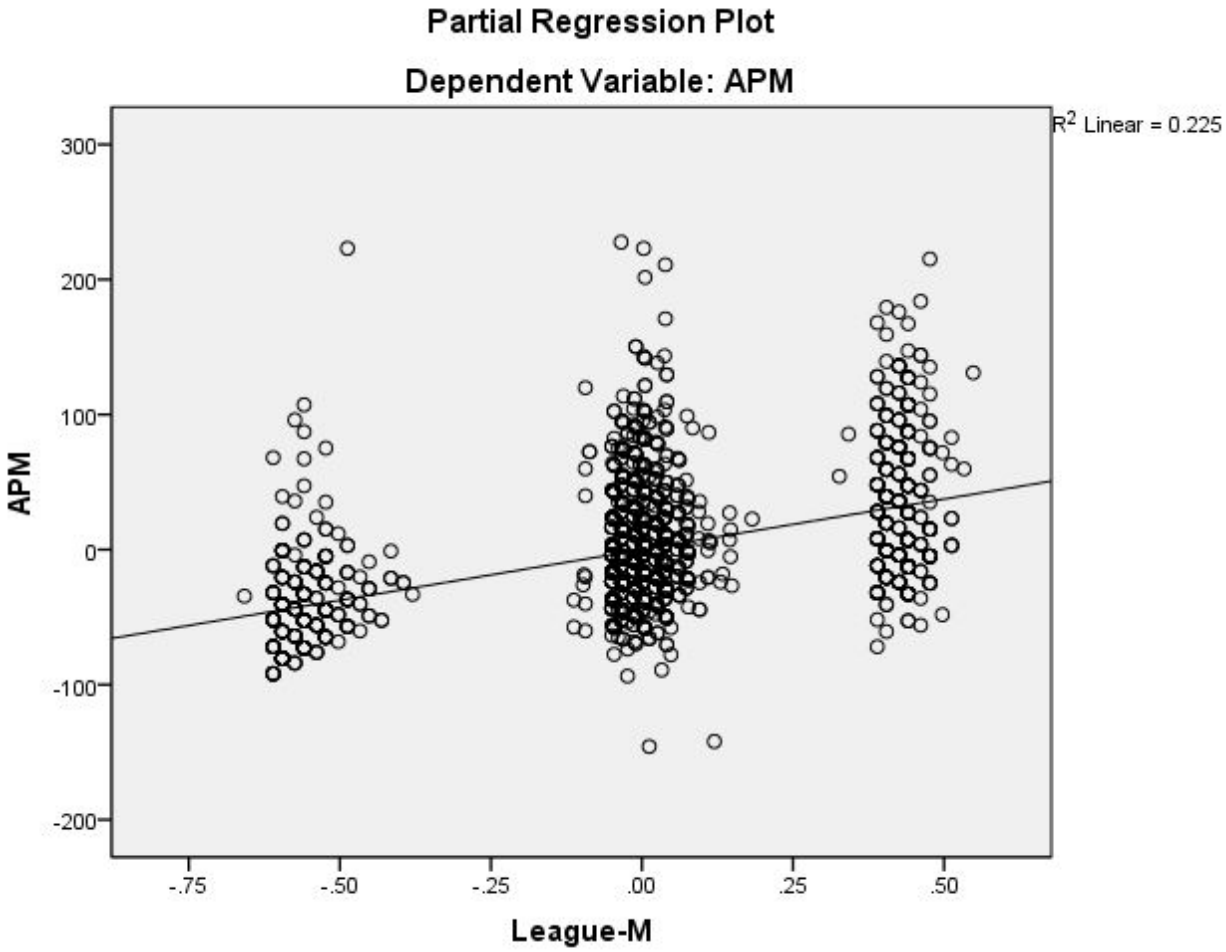
b. Dependent Variable: APM

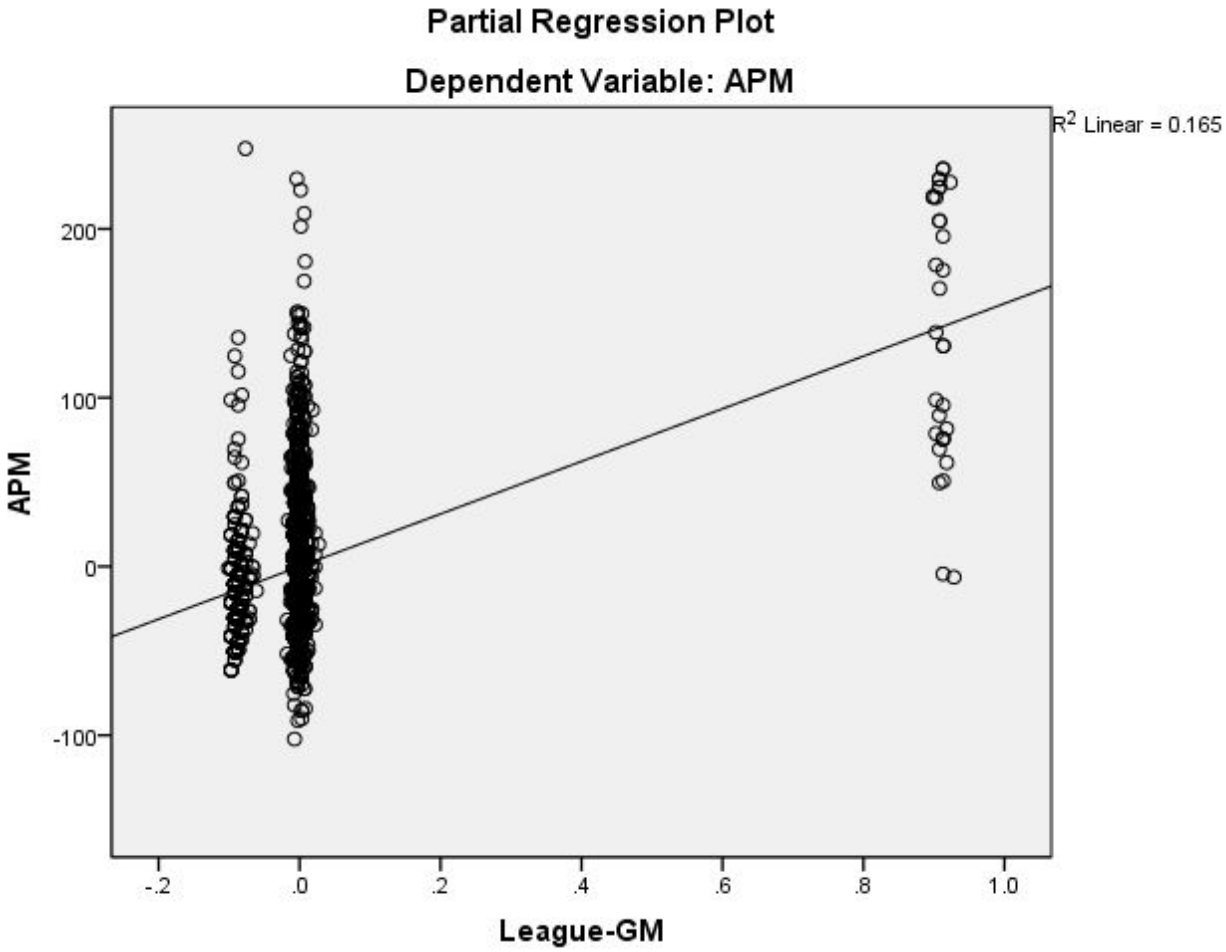

Partial Regression Plot
Dependent Variable: APM
R^2 Linear = 0.016
APM
Age
300
200
100
0
-100
-200
-20
-10
0
10
20

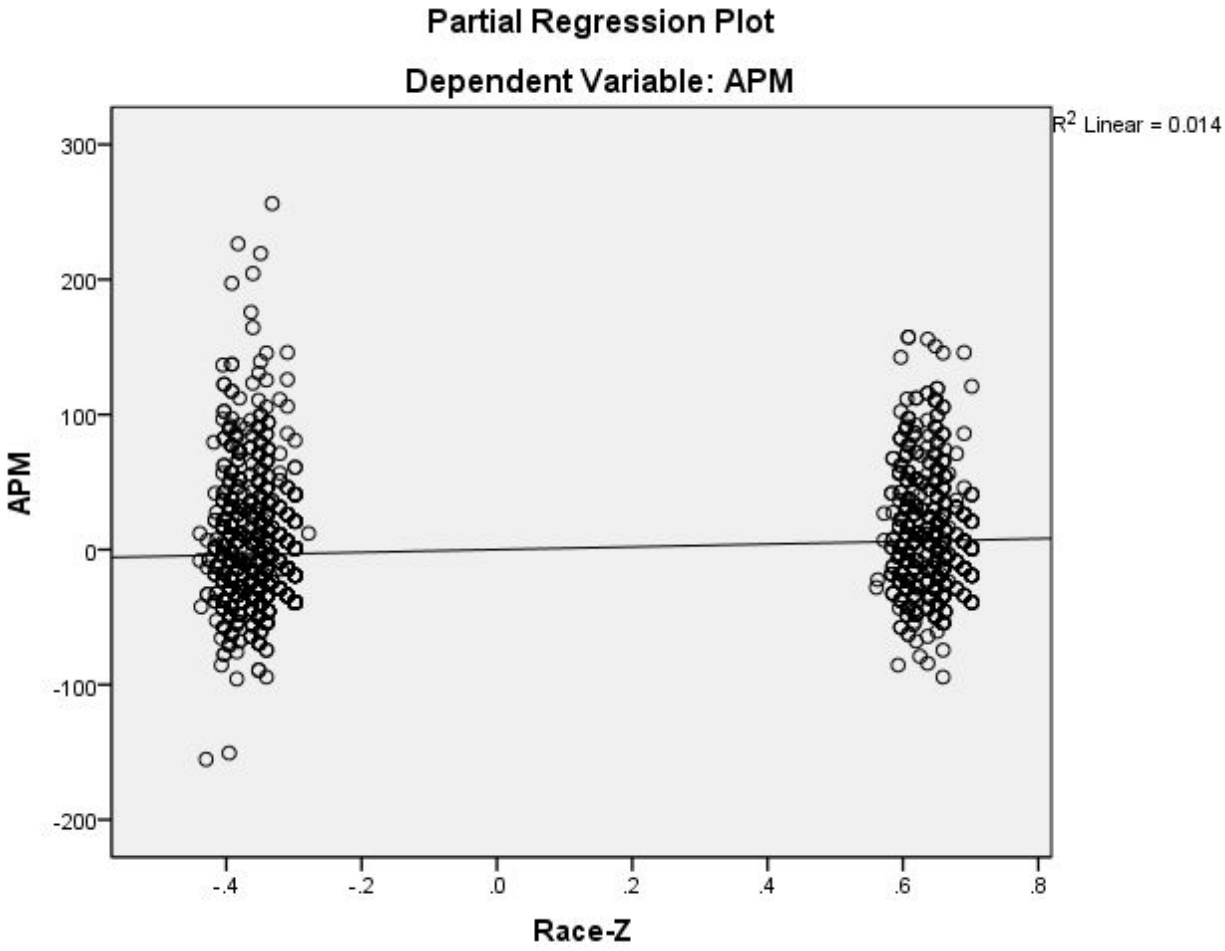

Partial Regression Plot
Dependent Variable: APM
R^2 Linear = 0.009
APM
300
200
100
0
-100
-200
-.50
-.25
.00
.25
.50
League-G

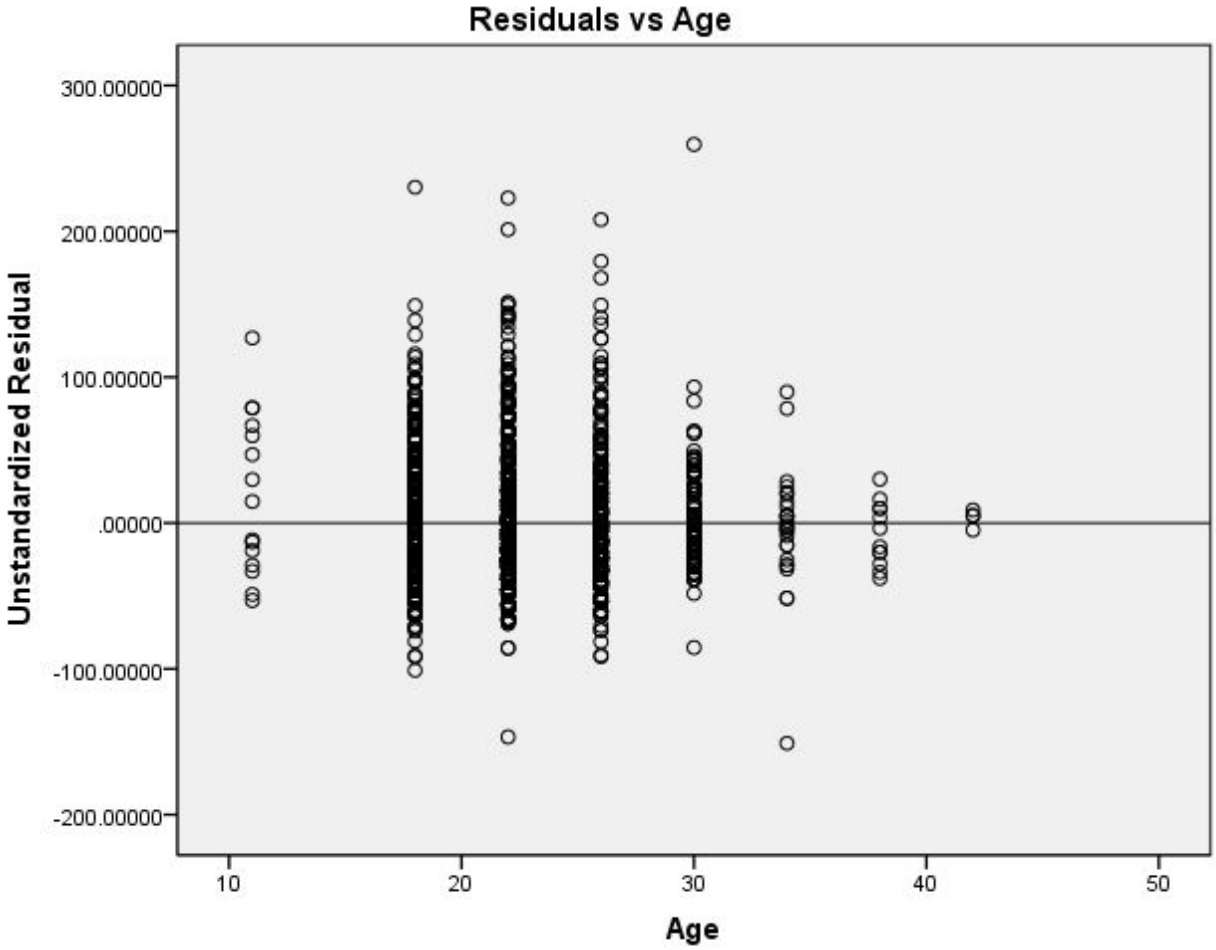

Partial Regression Plot
Dependent Variable: APM
R2 Linear = 0.038
APM
300
200
100
0
-100
-200
-.8
-.6
-.4
-.2
.0
.2
.4
League-P

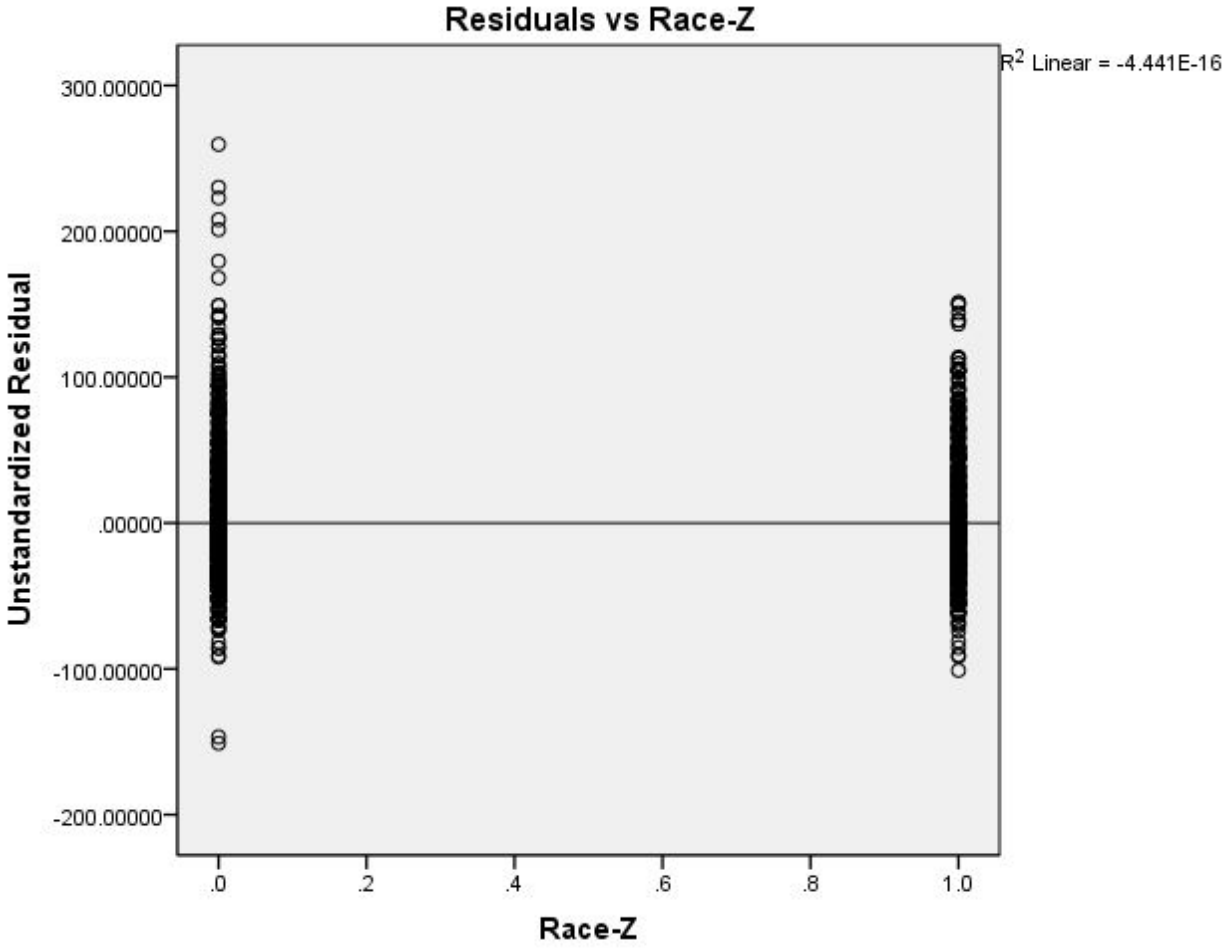

Partial Regression Plot
Dependent Variable: APM
R^2 Linear = 0.118
APM
300
200
100
0
-100
-200
-.8
-.6
-.4
-.2
.0
.2
.4
League-D

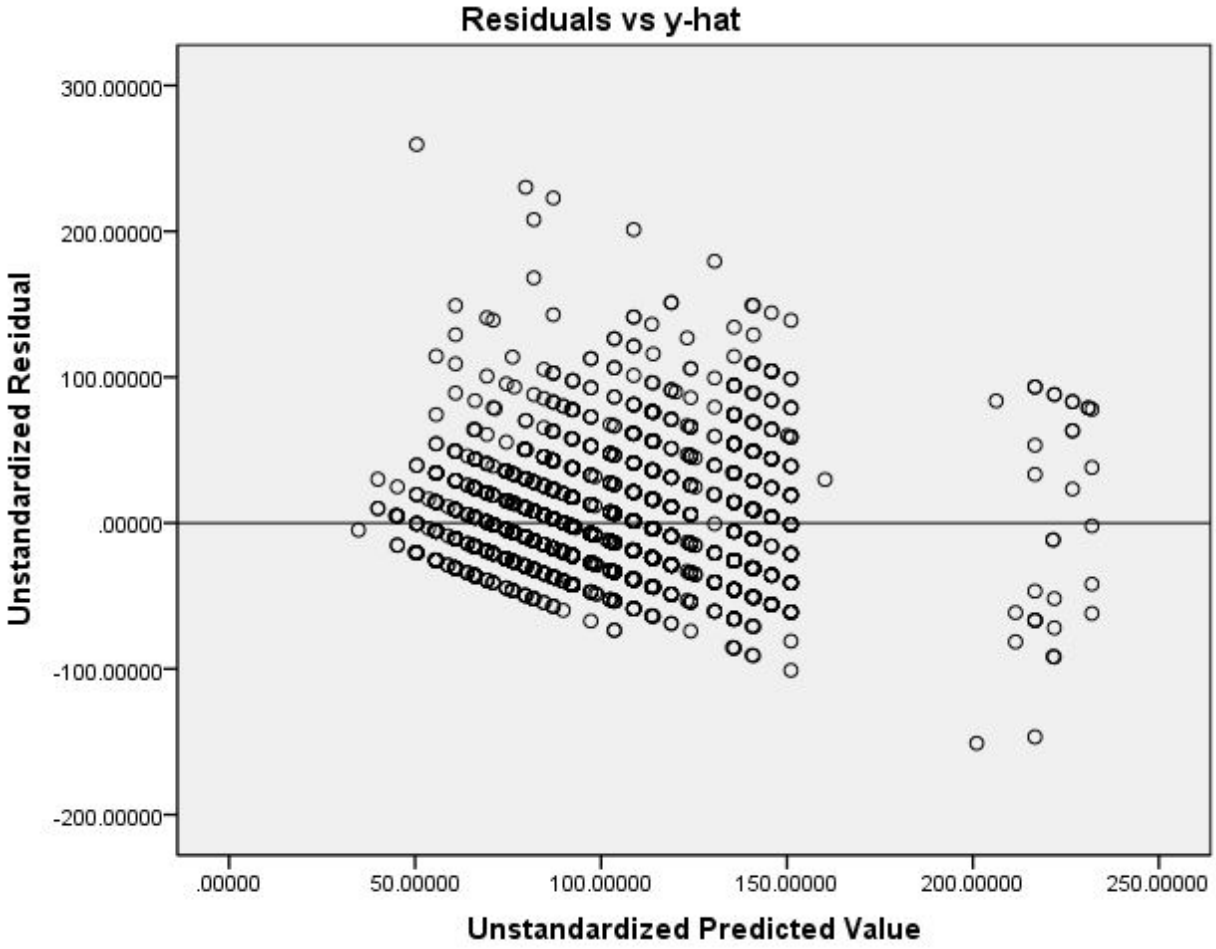

Partial Regression Plot
Dependent Variable: APM
R^2 Linear = 0.225
APM
300
200
100
0
-100
-200
-.75
-.50
-.25
.00
.25
.50
League-M


Partial Regression Plot
Dependent Variable: APM
R2 Linear = 0.165
APM
200
100
0
-100
-.2
.0
.2
.4
.6
.8
1.0
League-GM


Partial Regression Plot
Dependent Variable: APM
R^2 Linear = 0.014
APM
300
200
100
0
-100
-200
-.4
-.2
.0
.2
.4
.6
.8
Race-Z


Residuals vs Age
Unstandardized Residual
300.00000
200.00000
100.00000
.00000
-100.00000
-200.00000
10
20
30
40
50
Age


Residuals vs Race-Z
R2 Linear = -4.441E-16
300.00000
200.00000
100.00000
.00000
-100.00000
-200.00000
Unstandardized Residual
.0
.2
.4
.6
.8
1.0
Race-Z


Residuals vs y-hat
Unstandardized Residual
300.00000
200.00000
100.00000
.00000
-100.00000
-200.00000
.00000
50.00000
100.00000
150.00000
200.00000
250.00000
Unstandardized Predicted Value