Dancing with the Stars Placement Predictor: A Multiple Regression Model

Madalyn Elwood

MATH330: Regression Models in Business and Social Science

Buena Vista University

Introduction

When brainstorming ideas for this final project, I thought about how to make a model that would be interesting yet sufficient for my paper. While watching the beginning of *Dancing with the Stars*' most recent season, I found myself guessing which contestants would make it to the finale this year. Immediately, I decided that instead of making a prediction based on my own opinions, I would build a model that could make predictions for me. Observing trends that I noticed within the show, I chose my three independent variables as each celebrity's first round score, age, and height.

Methodology

First Round Score

The first variable that I added to my model is each contestant's first round score. I theorized that contestants that score higher in the very first episode would also score higher later in the season, therefore placing higher on the pedestal. Since the number of judges varies by season, I took each contestant's first round score and divided it by the number of judges to find what I call their average judge score (AvgJudgeScore).

Age

Next, I will be discussing the concept of age in my model. I speculated that younger contestants would have more energy than older contestants, thus making it farther in the competition. Especially considering the oldest champion was only 51 at the time of his season, I wanted to test and see if this variable was a good predictor of final placement on the show. I also thought about looking at the logarithm of age; I wanted to test if a jump between younger

contestants, for example, a 20-year-old compared to a 25-year-old, was more impactful than the same jump between older contestants, like 60- and 65-year-olds.

Height

The last variable in my multiple regression model is height. I picked this variable due to the emphasis they put on this characteristic during season 32. On the show, multiple judges and contestants kept mentioning how much harder it was for Harry Jowsey to dance due to him being 6' 5". So, I wanted to test this variable to see if it was a good predictor of contestant placement after all.

Data Collection

To collect my data, I used a few different resources online such as the *Dancing with the Stars* wiki and each participant's Wikipedia page. I also used a shell script to pull information from the internet easily and quickly. I entered all my information into an Excel spreadsheet and transferred it over to SPSS.

Issues

There were various issues within the data set that I had to deal with while making my model. The first was withdrawn contestants. Throughout the history of the show, there has been multiple instances where a celebrity needs to leave the competition in the middle of the season. In those cases, I gave the participant the finishing place of wherever they left the competition. For example, in season 31, Selma Blair withdrew from the competition in week five due to

health-related reasons. Since she had already surpassed four other people, I left her in that fifth-to-last position.

While looking at each season, I realized that there is a variance in the number of both contestants and judges between them. I have already discussed my solution to the judges' inconsistency in the above section 'First Round Scores.' For the contestants, though, I divided each person's place by the total number of participants for that season. This produces a weighted scale that suggests winning against a higher number of people will have a higher weight than being a champion on a season with only a few competitors. In the same sense, scoring eighth place in a season with fifteen other people is more impressive than scoring eighth in a season of eight celebrities, thus it is weighed as such.

The last issue that I had to work around was the special seasons: *All-Stars*, *Juniors*, and *Athletes*. The *All-Stars* season consisted of returning celebrities from past seasons, specifically finalists of past seasons. I felt that this data would skew the results of the overall model. A contestant who finished second in their original season could end up last on this season due to the higher level of competition. Because of this observation, I chose to leave out this set of data and stick to standard, controlled, seasons. In the same sense, I took out all the data from the *Juniors* season as well. This season was for celebrity children, which I believed would have skewed the age variable specifically. However, I kept the *Athletes* season, as occupation was not a variable that I decided to test at this time, and therefore should not influence the overall model.

Results

Equation

After running my model, I found my regression equation to be $Y = 1.071 - .114X_1 + .005X_2 + .00001978X_3$ or otherwise stated as $WeightedPlace = 1.071 - .114(AvgJudgeScore) + .005(Age) + .00001978(Height)$. These equations can be seen in a much cleaner format in 'Appendix A.'

The slope coefficients for this model are -.114 for the average judge score, .005 for age, and .00001978 for height. After filtering away age and height, we predict that a celebrity's weighted placement will decrease by .114 for every added point to their average judge score. At first, this sounded concerning since receiving an increasingly higher score should indicate a higher placement in the season; however, we must recall that first place has a value of one, while last is a greater number. Therefore, it makes sense that an increase by one in first round score would decrease the weighted value by .114, as these values have an inverse relationship.

Holding judge score and height constant, every year increase in age will increase the weighted placement value by .005. This is exciting because if it is significant, it will line up with my theory of younger contestants making it farther in the competition. Lastly, controlling for height, every inch taller a person is increases their weighted placement by .00001978. The small coefficient demonstrates how little of an effect height has on predicting a participant's placement, and we will look more in depth at this when discussing significance later.

The intercept in this model is 1.071, meaning if a contestant scores a zero in the first round, is zero years old, and is zero inches tall, their weighted place value would be 1.071. A weighted value of one or over one means the contestant would end in last place. This observation

of the intercept value is essential but not interesting. We can think of this intercept value as being a correction factor for our variables moving forward.

It is important to remember when interpreting this model, that the predicted value given is not the place of the contestant, but instead it is their weighted place. This means that to use this model in the real world, you must multiply the predicted weighted value by the number of contestants in the season. For example, if a 45-year-old celebrity in a season of 12 contestants recorded a first round score of 20 with four judges (AvgJudgeScore = 5) and is 5' 5" (65 inches), their predicted score is as follows:

$$Y = 1.071 - .114(5) + .005(45) + .00001978(65) = 0.727 = WeightedPlace$$

$$12\ contestants * 0.727 = 8.724 \approx 9$$

This celebrity is predicted to finish the competition around ninth place out of 12 contestants. Now let's look what would happen if there were only eight contestants in the season. Their weighted place would stay the same, so their placement would be as follows:

$$8\ contestants * 0.727 = 5.816 \approx 6$$

We see here that with eight people, our celebrity would finish at sixth place. This sleek aspect of the model makes it interesting, as you can test where someone would stand compared to various numbers of competitors.

Standard Error of Regression and R Squared

This model's standard error of regression (SER) has a value of .23611, as seen under 'Model Summary' in 'Appendix A.' This value represents the average distance observed values are from the regression line. In other words, the standard error of the estimate approximates the accuracy of our model. The low variance within this model is a good estimate of the model's overall precision. When comparing models to one another, a lower SER value will be crucial in determining which model is a better fit for the data. We will see this later under the 'Other Models' section of this paper.

The R squared value in this model is .338. Sometimes, R squared is referred to when comparing two models' suitability for a set of data. Although, we should be cautious when doing so, as R squared only measures the linearity of a model, not its goodness of fit. This is why SER is a better measure to use when comparing models.

Significance

Significance in our model is used to determine whether the relationship between our variables and the predicted value exists or not. We will be working at the 95% confidence level, meaning our p-values will need to be lower than .05 for a variable to be significant. In this model, two of the three variables used are significant.

Looking at the 'Coefficients' table under 'Appendix A,' we see the p-values for each variable in this multiple regression. Average judge score and age's p-values show <.001, which is less than 0.05, making them significant predicators of weighted place. On the other hand, height has a p-value of .994. This is tremendously higher than 0.05, proving its poor prediction

of weighted place. This may play a factor into why the coefficient for height is so small; it has relatively zero effect on the placement of a contestant on *Dancing with the Stars*.

As a whole, the model has a p-value of <.001, as shown under the 'ANOVA' section of 'Appendix A.' Once again using the 95% confidence level, this value is low enough to state the assumption that the model is a significant predictor of a competitor's weighted placement.

Partial Regression Plots

When forming regression models, it is often smart to examine diagnostic plots for any signs of non-linearity. The first partial plot under 'Appendix A' shows the relationship between weighted place and average judge score. This plot clearly shows a strong, negative, linear relationship between the two values, so we are not worried about any problems here. The same goes for age; even though the relationship may not be as defined, we can still notice a possible linear relationship between the variable and weighted place, but this time it is positive. The positive and negative relationships within these plots also tie directly with the signs of the coefficients of these two variables.

Looking at the partial plot for height, we recognize the randomness of the data points in the graph. Since there are no clear signs of linearity, it is safe to assume that height will not be a significant predictor of weighted place. Within these diagnostic plots, we can look for other patterns of the data points that point towards using a transformation. This specific plot, though, shows no signs of a possible change and, therefore, is just random and insignificant.

Multicollinearity

One way to determine multicollinearity is to look at the tolerance and auxiliary R squared values of each variable. The tolerance levels for average judge score, age, and height are .840, .856, and .964 respectively. High tolerance levels produce low auxiliary R squared values. Subtracting each collinearity tolerance from one, we find the auxiliary R squared values to be .160 for average judge score, .144 for age, and .036 for height. Auxiliary R squared for a variable measures the proportion of that variable's variance explained by all the other variables in the model. So, to minimize multicollinearity, we want our auxiliary R squared values to be small. The values that we found for this model are very low, meaning we do not need to worry about multicollinearity in this model. This also means that we can trust that our p-values are not inflated, and therefore are more confident that our two significant X variables belong in the model.

Other Models

When creating this model, I had to make some decisions on what the best variables would be to test for predicting the weighted placement. While this paper only focuses on three main variables, I had plenty of other ideas that I will share in this section. My hope is to someday revisit this model and look at the impact of a larger number of independent variables on the model's predictions. Nevertheless, I was able to explore the transformation of age rather quickly.

Transformed Age (LOG_Age)

As discussed in the earlier 'Age' section of 'Methodology,' I theorized that using a transformation of age such as a logarithm would influence the accuracy of the model. I believed

that it was possible for an age jump between younger contestants to have more of an impact on the overall placement of a contestant than a same size jump between older contestants. This type of curve that is being described matches the arc of a logarithm.

When running the program with the LOG_Age variable instead of age, I found that there was practically no change at all in the model. This new model had a standard error of regression value of .23619, while the previous model had an SER of .23611. That is a difference of only .00008! In statistics, a value as miniscule as that can be accepted as, essentially, equal. Looking at the R squared values, both models produced the number .338. All these numbers can be found under 'Appendix B.' In theory, when comparing these models, there is no obvious choice. Yet, I kept the model with the untransformed age variable, as it was shown that the transformation was not necessary through these results.

Other Variable Ideas

I had numerous ideas for the independent variables in the model, but I chose the three that I used in this paper based on how significant I anticipated them to be as well as how easy collecting each variable's data seemed. Among these potential variables were sex, fanbase, occupation, number of wins their professional partner has, experience in years their professional has on the show, and first round dance type.

Sex would have been a dummy variable, with male being represented as a value of one and female being zero. It might also be interesting to look at a potential interaction term consisting of age and sex combined. For example, what if being a young female gives a contestant an advantage over an old male.

Regarding their professional partner, this data would simply be a number coming from *Dancing with the Stars*' history. I had debated using either number of previous wins or years of experience within my main model, but I doubted how significant it would be. I understand that significance was not the ultimate goal of this project, although I still wanted to produce the best model that I could come up with. In doing so, I decided to leave out anything to do with the professional and focus solely on the contestant and their characteristics, for now anyways.

Examining occupation and first round dance type would have been interesting, as neither subject is originally given as a numeric value. I would have to create categorical variables to place these factors into the model.

Lastly, my most debated variable for this model was fanbase. Even though *Dancing with the Stars* is a dance show, it is also a reality show, and half of every star's score each week comes from viewer votes. I wanted to include this within the model as I felt it was very important, however, I struggled finding a solid way of measuring a person's popularity. One idea was to use each person's Q-score, but finding this data, for some reason, was extremely difficult as the official Q-Score site does not release its information to the public. Next, I thought about looking at each celebrity's follower count on Twitter and Instagram. I soon realized that this number would not reflect their popularity at the time they were on the show, and finding their past information would be difficult as well. There also were multiple seasons that occurred before either of these platforms were invented.

An idea that I believe may work after more research would be to find the Google search history of each celebrity at the time that they were on the show. This would require coding another shell script or spending numerous hours of research myself on Google Trends. I would

love to come back to this, along with the rest of my ideas, in the future and create one big multivariate regression model.

Conclusion

All in all, I created a multiple regression model that produces the predicted weighted placement of a *Dancing with the Stars* contestant given several factors. These independent variables were an average first round judge score, the person's age, and their height. It was concluded that average judge score and age were significant predictors of this weighted placement. The model was also significant itself overall and had minimal signs of multicollinearity. I also discussed the possibilities for improvement to the model in the future. Now, whenever a new season of the show approaches, I can put my model to the test and see if the results match up with the season finale.

Data Appendices

Appendix A – Regression Model 1

Equation: $Y = 1.071 - .114X_1 + .005X_2 + .00001978X_3$

Equation: $WeightedPlace = 1.071 - .114(AvgJudgeScore) + .005(Age) + .00001978(Height)$

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
WeightedPlace	367	.06	1.00	.5377	.28870
AvgJudgeScore	367	3.25	9.67	6.4310	1.14268
Height	365	50.00	86.00	68.4767	4.80357
Age	367	15.00	81.00	39.1935	13.66673
Valid N (listwise)	365				

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Age, Height, AvgJudgeScore[b]	.	Enter

a. Dependent Variable: WeightedPlace

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.582[a]	.338	.333	.23611

a. Predictors: (Constant), Age, Height, AvgJudgeScore

b. Dependent Variable: WeightedPlace

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.282	3	3.427	61.480	<.001[b]
	Residual	20.124	361	.056		
	Total	30.406	364			

a. Dependent Variable: WeightedPlace

b. Predictors: (Constant), Age, Height, AvgJudgeScore

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	1.071	.212		5.044	<.001		
	AvgJudgeScore	-.114	.012	-.449	-9.612	<.001	.840	1.191
	Height	1.978E-5	.003	.000	.008	.994	.964	1.038
	Age	.005	.001	.237	5.125	<.001	.856	1.169

a. Dependent Variable: WeightedPlace

Collinearity Diagnostics[a]

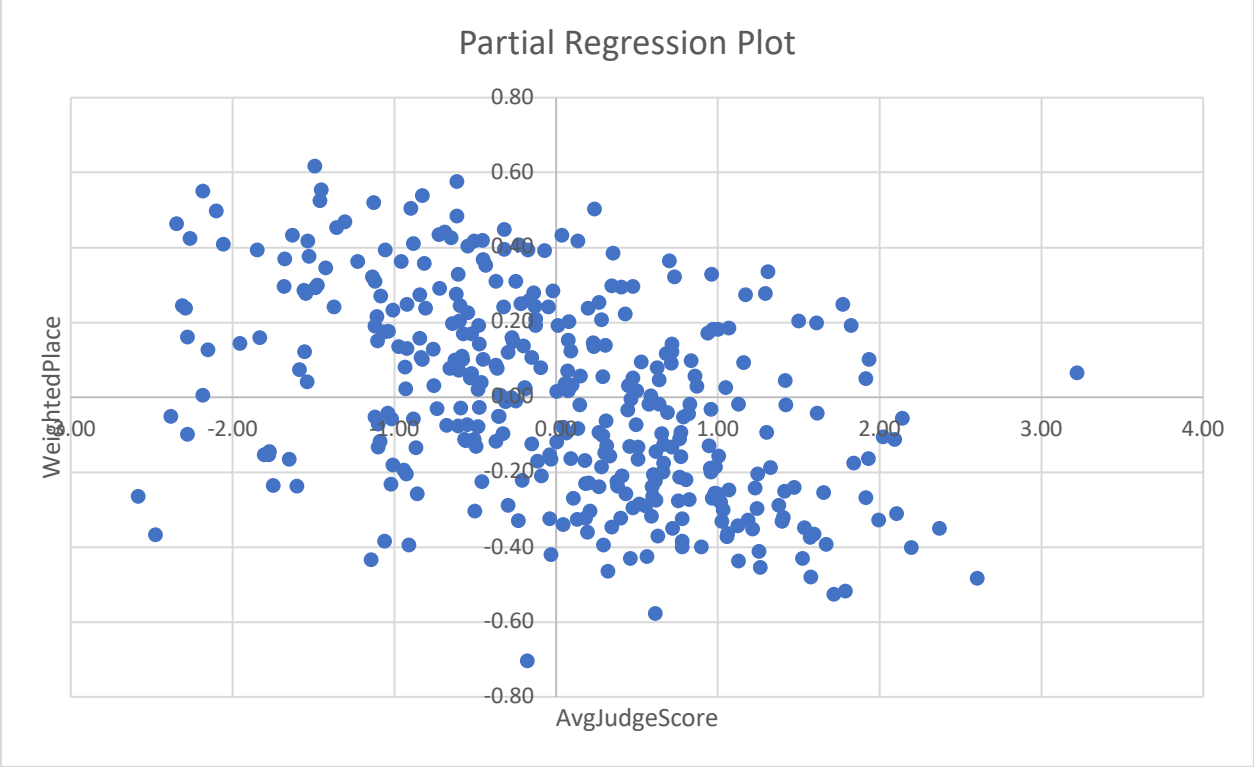
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	AvgJudgeScore	Height	Age
1	1	3.881	1.000	.00	.00	.00	.01
	2	.101	6.209	.00	.06	.00	.64
	3	.017	15.281	.03	.75	.09	.33
	4	.002	43.463	.97	.18	.91	.02

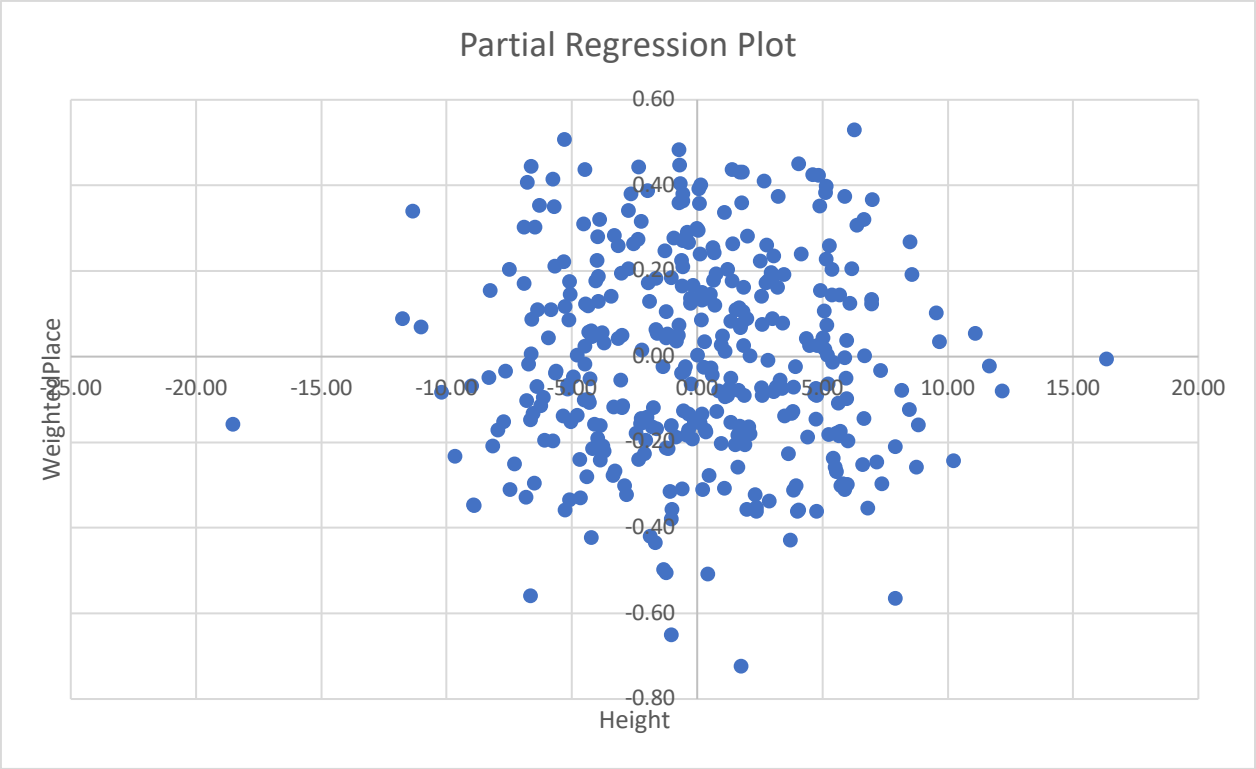
a. Dependent Variable: WeightedPlace

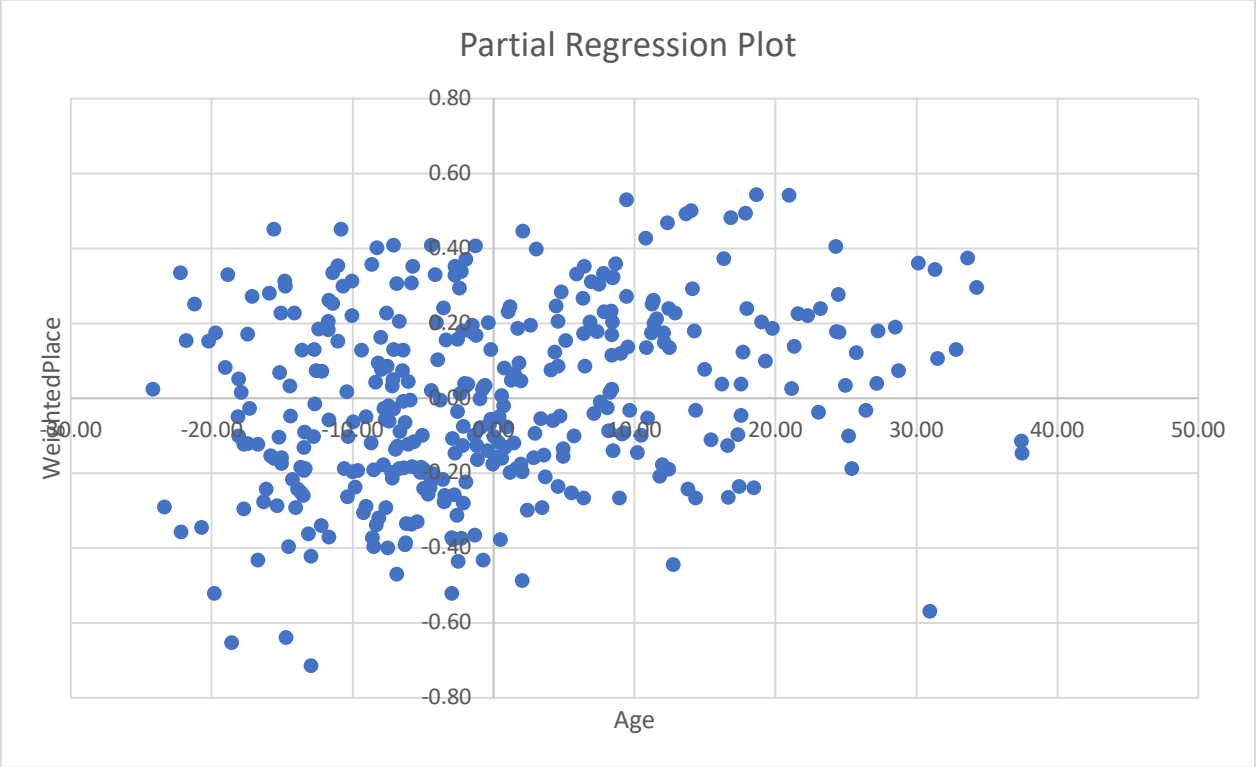
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.1500	1.0790	.5389	.16807	365
Residual	-.72358	.52936	.00000	.23513	365
Std. Predicted Value	-2.314	3.214	.000	1.000	365
Std. Residual	-3.065	2.242	.000	.996	365

a. Dependent Variable: WeightedPlace




Partial Regression Plot
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
-0.80
-25.00
-20.00
-15.00
-10.00
-5.00
0.00
5.00
10.00
15.00
20.00
WeightedPlace
Height


Partial Regression Plot
0.80
0.60
0.40
0.20
0.00
-0.20
-0.40
-0.60
-0.80
-30.00
-20.00
-10.00
0.00
10.00
20.00
30.00
40.00
50.00
WeightedPlace
Age

Appendix B – Regression Model 2 (Transformed LOG_Age)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
WeightedPlace	367	.06	1.00	.5377	.28870
AvgJudgeScore	367	3.25	9.67	6.4310	1.14268
Height	365	50.00	86.00	68.4767	4.80357
LOG_Age	367	2.71	4.39	3.6080	.35184
Valid N (listwise)	365				

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	LOG_Age, Height, AvgJudgeScore[b]	.	Enter

a. Dependent Variable: WeightedPlace
b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.581[a]	.338	.332	.23619

a. Predictors: (Constant), LOG_Age, Height, AvgJudgeScore
b. Dependent Variable: WeightedPlace

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.267	3	3.422	61.350	<.001[b]
	Residual	20.139	361	.056		
	Total	30.406	364			

a. Dependent Variable: WeightedPlace

b. Predictors: (Constant), LOG_Age, Height, AvgJudgeScore

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.603	.253		2.388	.017		
	AvgJudgeScore	-.114	.012	-.451	-9.672	<.001	.844	1.185
	Height	.000	.003	-.008	-.183	.855	.958	1.044
	LOG_Age	.194	.038	.237	5.098	<.001	.852	1.174

a. Dependent Variable: WeightedPlace

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	AvgJudgeScore	Height	LOG_Age
1	1	3.963	1.000	.00	.00	.00	.00
	2	.030	11.472	.00	.61	.01	.06
	3	.006	26.488	.01	.11	.39	.68
	4	.002	48.117	.99	.28	.60	.26

a. Dependent Variable: WeightedPlace

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.1271	1.0367	.5389	.16795	365
Residual	-.67597	.54674	.00000	.23522	365
Std. Predicted Value	-2.452	2.964	.000	1.000	365
Std. Residual	-2.862	2.315	.000	.996	365

a. Dependent Variable: WeightedPlace