Department of Mathematics

Predicting Baseball Player's Salary Based on Past Performance and Other Factors

Courtney Arand

Advisor: Dr. Maher Qumsiyeh

Abstract

The purpose of this project is to predict a baseball player's salary in the 2016 season based on their performance and other factors in the 2015 season. The factors (regressors) used in this project that could possibly affect the salary (dependent variable) were age, seasons played, games played, wins above replacement (WAR), and batting average. The data was collected from sources such as the Major League Baseball (MLB) website and the USA TODAY website. We used the statistical software package SPSS to analyze the data and obtain a good prediction model.

Model Development

We want to find a model to predict the salary (dependent variable (y)) by using several regressors (independent variables) such as age, seasons played, games played, WAR, and batting average. In order to find out which variables were significant in the model, we ran stepwise procedure in SPSS. This tells us which of the variables will give us the best model to predict a player's salary.

Coefficients[4]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	831.597	152.843		5.441	.000		
	Seasons Played	242.738	20.304	.702	11.955	.000	1.000	1.000
2	(Constant)	324.253	173.403		1.870	.063		
	Seasons Played	254.004	18.924	.735	13.422	.000	.986	1.014
	WAR	167.403	33.088	.277	5.059	.000	.986	1.014
3	(Constant)	2824.274	828.947		3.407	.001		
	Seasons Played	351.185	36.522	1.016	9.616	.000	.250	3.996
	WAR	150.534	32.629	.249	4.613	.000	.958	1.043
	Age	-109.706	35.616	-.329	-3.080	.002	.245	4.081
4	(Constant)	1060.452	1164.610		.911	.364		
	Seasons Played	339.805	36.477	.983	9.316	.000	.245	4.084
	WAR	123.512	34.644	.204	3.565	.000	.830	1.205
	Age	-105.562	35.243	-.316	-2.995	.003	.244	4.094
	Games Played	12.330	5.788	.121	2.130	.035	.838	1.193

a. Dependent Variable: Square Root of Salary

After running the model we realized that transformation is needed on the y variable for model adequacy, especially for normality of the error terms. We found out that the best transformation is using square root of the salary.

Creating the Model

After running linear regression on the data in SPSS, we find out what the coefficients would be for the different factors (regressors).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1060.452	1164.610		.911	.364		
	Seasons Played	339.805	36.477	.983	9.316	.000	.245	4.084
	WAR	123.512	34.644	.204	3.565	.000	.830	1.205
	Age	-105.562	35.243	-.316	-2.995	.003	.244	4.094
	Games Played	12.330	5.788	.121	2.130	.035	.838	1.193

a. Dependent Variable: Square Root of Salary

The following is the model we came up with to predict the square root of a baseball player's salary.

Equation (1): $\hat{y}^{1/2} = 1060.452 - 105.562x_1 + 339.805x_2 + 12.330x_3 + 123.512x_4$

In this model, y is the salary, x_1 is age, x_2 is seasons played, x_3 is games played, and x_4 is the player's WAR.

Measuring Adequacy

Once we found which variables are significant, we can check to see if running linear regression will give an equation that does a good job of estimating a player's salary. To do this, we look at R^2, the normal probability plot, and the residual vs. predicted value plot.

R^2 Value

Model Summary[4]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.779[a]	.607	.597	825.030731469510400

The R^2 value of 0.607 means that the model produced in the linear regression explains 60.7% of the variation in the square root of the salaries.

F-Test

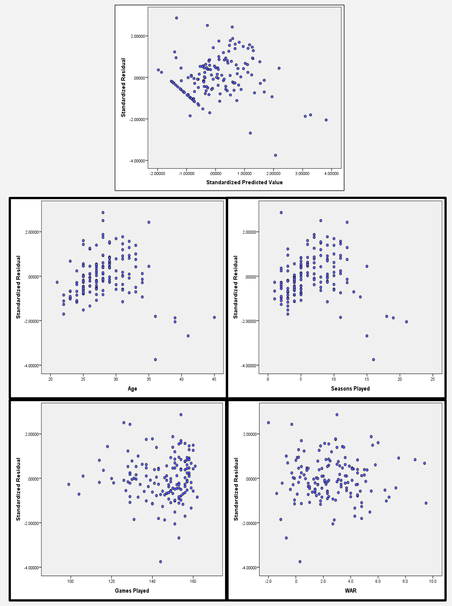
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	145977418.175	4	36494354.544	48.925	.000[b]
	Residual	107413634.332	144	745928.016		
	Total	253391052.507	148			

The chart above gives the F-Statistic and the p-value. This value shows that the model is adequate and that the coefficients of the different regressors are not zero.

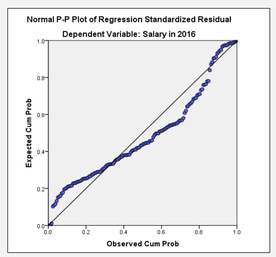
Residual Plots

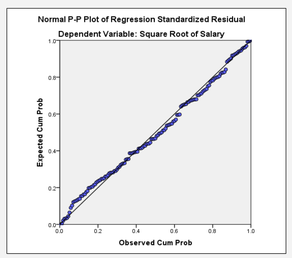
The residual vs. predicted plot shows if there is any relationship between the variance and the mean. This plot after the transformation of the y variable is fairly random which shows that there is not much of a relationship between the mean and variance. The other four graphs show the residuals graphed with the different regressors. They are all fairly random which validates that this is a strong model.



Normal Probability Plot

The normal probability plot on the left shows the original plot before any transformation in y. This shows that the model does not do a good job of predicting salary. After transforming y to the square root of y, the normal probability plot on the right shows that they model is a much better fit.





Conclusion

We were able to develop a model from running procedures in SPSS. After completing different adequacy checks to make sure that the model was a good fit, we can now use the model (1) to predict salaries using the factors age, seasons played, games played, and WAR.