Department of Mathematics

Best Predictors of Player's Goals in Soccer

Christopher Anderson

Advisor: Dr. Maher Qumsiyeh

Abstract

The purpose of this project is to see what offensive statistics are best for predicting a soccer player's future goals using the 2017-18, 2018-19, and 2019-20 (as of 3/9) season's data. The regressors (variables) tried in this project that could affect goal output (dependent variable) were shots, shots on goal, expected goals, minutes (total playing time), games (number of appearances), position, team, player, and league. The data was collected from Football (soccer) reference and analyzed using the Statistical Analysis System (SAS). A model predicting the response variable (goals) in terms of some significant regressors was obtained.

Model Development

We want to find a model that can predict the goals (dependent variable, y) scored by a soccer player in the coming season by using several regressors (independent variables). To do this, we first took all of the regressors and broke them up into two groups: quantitative (a numerical measurement) and qualitative (a non-numerical measurement) variables. The quantitative variables were shots, shots on goal, expected goals, minutes and games. The qualitative variables were team, league, position, and player. We then performed a Stepwise procedure on the quantitative variables in SAS to find which of these factors we need to keep in the model. We took the model from the Stepwise procedure and subsequently ran a General Liner Model procedure to see which qualitative variables would be significant in the model. We then performed a Variance Inflation Factors (VIF) Procedure to test multicollinearity and a Cp test to see why an insignificant variable was left in the model by the Stepwise procedure. VIF was high between shots, shots on goal, and expected goals, so we took the square root of shots on goal. Shots seems to be insignificant but after running a Cp test we saw that the lowest Cp was the model that included shots.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	xG		xG	1	0.8333	0.8333	31.3515	804.79	<.0001
2	sog1			2	0.0179	0.8512	12.8894	19.27	<.0001
3	Games		Games	3	0.0046	0.8559	9.5885	5.12	0.0250
4	Minutes		Minutes	4	0.0036	0.8595	7.4699	4.06	0.0457
5	Shots		Shots	5	0.0030	0.8625	6.0000	3.47	0.0644

Creating the Model

After running the linear regression, the following coefficients were obtained for the significant variables.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-3.09586	1.01634	-3.05	0.0027
sog1		1	1.93422	0.42379	4.56	<.0001
xG	xG	1	0.88223	0.07322	12.05	<.0001
Minutes	Minutes	1	0.00217	0.00098122	2.21	0.0288
Shots	Shots	1	-0.03044	0.01634	-1.86	0.0644
Games	Games	1	-0.23871	0.07615	-3.13	0.0021

The following is the best model we came up with to predict the goals scored by a soccer player in the coming season.

$$\hat{y} = -3.096 + 1.934x_1 + 0.882x_2 + 0.00217x_3 - 0.0304x_4 - 0.239x_5$$

Where y is the goals, x_1 is squared root of shots on goal, x_2 is expected goals, x_3 is minutes, x_4 is shots, and x_5 is games.

Model Adequacy Checks

Once we found which variables were significant we ran a linear regression to determine if the equation is a good predictor for goals. Thus, we looked at the R^2,the normal probability plot and the residual vs. predicted value plot.

R^2 Value

Root MSE	2.86243	R-Square	0.8625
Dependent Mean	12.42331	Adj R-Sq	0.8581
Coeff Var	23.04079		

The R^2 value of 0.863 means that the model produced in the linear regression explains 86.3% of the variation in the goals.

F-Test

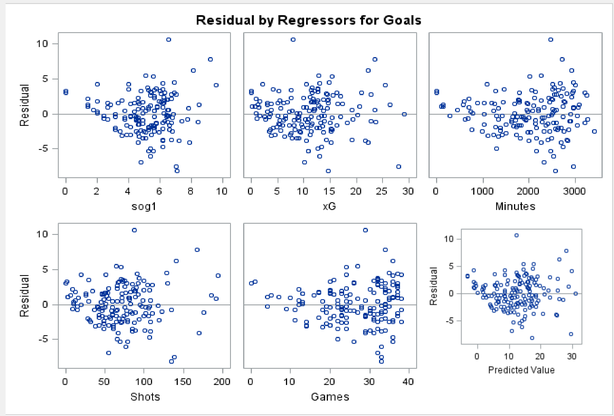
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8069.41174	1613.88235	196.97	<.0001
Error	157	1286.37967	8.19350		
Corrected Total	162	9355.79141			

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.

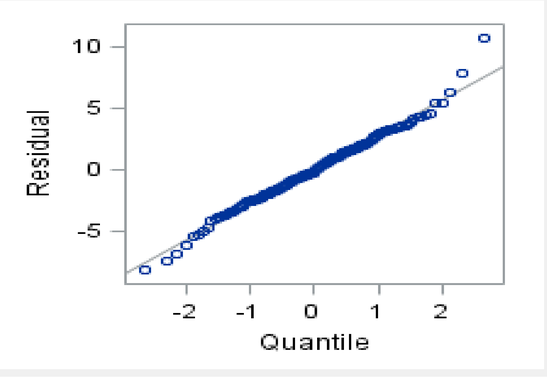
Residual Plots

The residual vs. predicted plot shows if there is any relationship between the variance and the mean. The five graphs show that there is no relationship as they are all random, validating that this is a strong model.



Normal Probability Plot

The normality plot below shows that the model is a good fit for predicting future goals.



Conclusion

We were able to develop a model by using different techniques in SAS. After performing different adequacy checks to make sure that the model was a good fit, we can now use the model to predict goals scored by a player using the variables shots, the square root of shots on goal, expected goals, minutes, and games.