Regression Analysis of Overwatch League Data

Andrew Rodgers

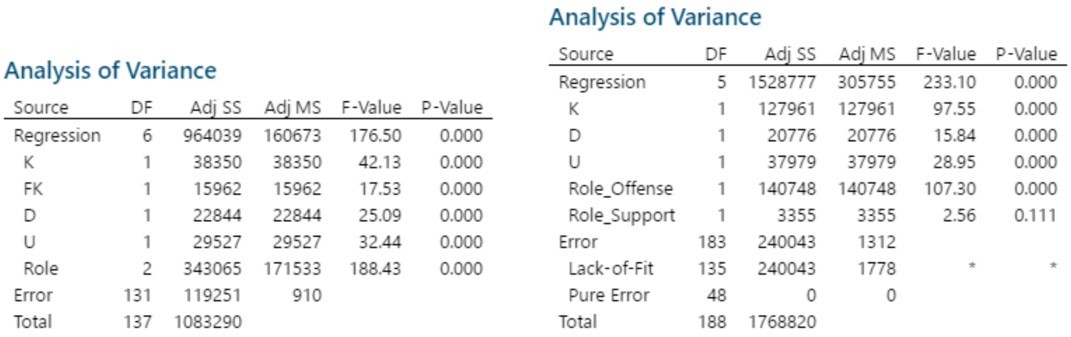
1 Introduction

The goal of this project is to perform a regression on the Overwatch League (OWL) data set. The data is a subset of a stat sheet from recent matches in the 2019 season. Excluded data points are players who did not appear in a match, thus the excluded points are rows of zeros. The quantitative variables measured are Kills(K), Deaths(D),and Ultimates(U). The categorical variable is Role(Tank,Offense,Support). The purpose is to create a model to predict the number of Points earned, given these values for any given player in a future Overwatch League match.

2 Analysis

2.1 Basic MLR Regression

A good place to start the analysis is with the regular, basic regression. From **Figure 1a**, we can see the p-values are around 0, so we can say the model is useful by the F-test. In **Figure 1(b)**, we can see most of the p-value for *Role_Support* is



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	964039	160673	176.50	0.000
K	1	38350	38350	42.13	0.000
FK	1	15962	15962	17.53	0.000
D	1	22844	22844	25.09	0.000
U	1	29527	29527	32.44	0.000
Role	2	343065	171533	188.43	0.000
Error	131	119251	910		
Total	137	1083290			

(a) ANOVA table for the full model

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	1528777	305755	233.10	0.000
K	1	127961	127961	97.55	0.000
D	1	20776	20776	15.84	0.000
U	1	37979	37979	28.95	0.000
Role_Offense	1	140748	140748	107.30	0.000
Role_Support	1	3355	3355	2.56	0.111
Error	183	240043	1312		
Lack-of-Fit	135	240043	1778	*	*
Pure Error	48	0	0		
Total	188	1768820			

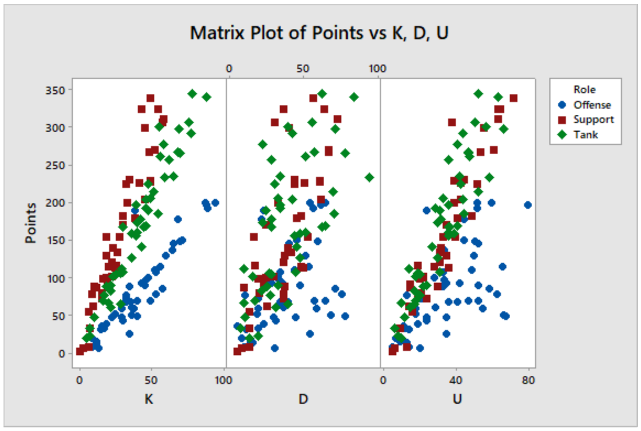
(b) ANOVA table for model with indicator variables

Figure 1

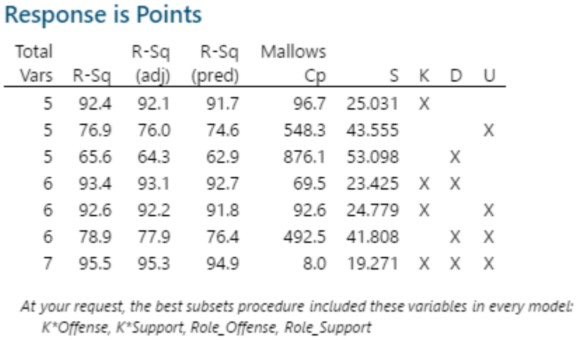
2.2 Variable Selection(Optimal Subsets)

In addition to Role, the data set includes a second categorical variable, Team. We exclude this predictor in our analysis simply because there are 20 responses. If we used a predictor with this many respsonses, we would not have enough data for each group to make an accurate conclusion. An Extra Sum of Squares Test confirms that Team does not contribute to the model. Before we can choose the best set of predictors for the model, we need to create indicator variables for the Role categorical variable, Tank is used as the reference group. In **Figure 2(a)**, we can clearly see in the plot of K versus Points, different roles have significantly differing slope, so we need interaction terms for K. The other plots do not have clearly different slopes, so we will use the additive model for these. With the new interaction terms, we can add the predictors

*K*Role_Support* and *K*Role_Offense* to our model. Looking at **Figure 2(b)**, we can see the best model is the full model, to little surprise.



(a) Matrix plot of Points versus Predictors



Response is Points

Total Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	K	D	U
5	92.4	92.1	91.7	96.7	25.031	X		
5	76.9	76.0	74.6	548.3	43.555			X
5	65.6	64.3	62.9	876.1	53.098		X	
6	93.4	93.1	92.7	69.5	23.425	X	X	
6	92.6	92.2	91.8	92.6	24.779	X		X
6	78.9	77.9	76.4	492.5	41.808		X	X
7	95.5	95.3	94.9	8.0	19.271	X	X	X

*At your request, the best subsets procedure included these variables in every model: K*Offense, K*Support, Role_Offense, Role_Support*

(b) Best Subsets Regressions

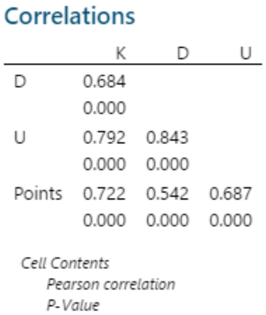
Figure 2

2.3 Stepwise Regression

This section intentionally left blank.

2.4 Multicollinearity

There will always be some correlation between the preditors becuase of the way the game works. For example, Kills and Deaths should be inversely related, and Ultimates and Kills should be related. There is multicollinearity present between the predictors, as demonstrated by **Figure 3(a)**. In **Figure 3(b)**, we can see almost all of the VIF's are above 5, which is another indication of multicollinearity.

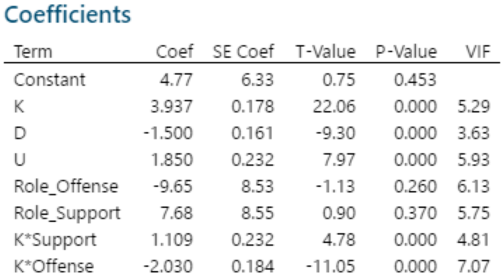


Correlations

	K	D	U
D	0.684		
	0.000		
U	0.792	0.843	
	0.000	0.000	
Points	0.722	0.542	0.687
	0.000	0.000	0.000

Cell Contents
Pearson correlation
P-Value

(a) Correlation Matrix



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.77	6.33	0.75	0.453	
K	3.937	0.178	22.06	0.000	5.29
D	-1.500	0.161	-9.30	0.000	3.63
U	1.850	0.232	7.97	0.000	5.93
Role_Offense	-9.65	8.53	-1.13	0.260	6.13
Role_Support	7.68	8.55	0.90	0.370	5.75
K*Support	1.109	0.232	4.78	0.000	4.81
K*Offense	-2.030	0.184	-11.05	0.000	7.07

(b) VIF values for predictors

Figure 3

2.5 Outliers

Looking at the 4-in-1 plots, there is only one real outlier in our regression. Calculating Cook's Distance for the data, we see this point (row 72) has a significantly higher Cook's Distance, 0.138, and a much larger DFITS, 1.227, than is expected.

2.6 Transformations

If we exlcude the single outlier, the data is very nice, with an excellent symmetry. Looking at histograms of the predictors, we can see all of the distributions are roughly normal or symmetrical. Scatter plots show no clear nonlinear patterns. Thus, we have no reason to suspect our data needs any transformations.

2.7 Assumptions

All of the usual assumptions for regressions are met in this case. The 4-in-1 plot shows no correlated errors, normally distributed residuals, and constant variance. We assume the data is linearly related, so we have the correct model function.

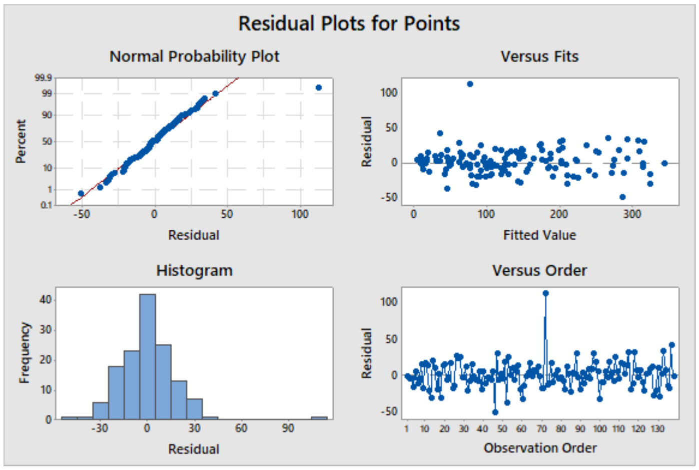


Figure 4 : 4-in-1 plot

3 Conclusions

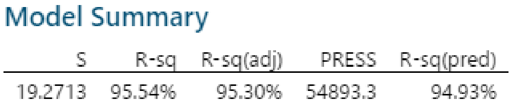
To little surprise, using the full model with the interaction terms gives us the best model.

$$Points = 4.77 + 3.937 * K - 1.5 * D + 1.850 * U - 9.65 * Role\_Offense \\ + 7.68 * Role\_Support + 1.109 * K * Role\_Support - 2.03 * K * Role\_Offense \quad (1)$$

We found no evidence any transformations were needed to the data, or that any subset of the predictors would give a better model. In this case there is multicollinearity, which is due to the game rules and scoring system.

The full model works best because of the game rules and scoring system. In fact, the actual Points calculation is a linear combination of Kills, Deaths, and Ultimates, along with 2 other statistics not provided to us. The coefficients in this calculation differ between Roles. The game is designed so that more Ultimates leads to more Kills, which leads to less Deaths; explaining most of the multicollinearity.

Looking at the predicted R-squared of our model, we can say it is fairly good at predicting values based on new data. Applying the model to a small set of test data, our model overestimated Points, but not by a significant margin, likely due to random chance.



Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
19.2713	95.54%	95.30%	54893.3	94.93%

While performing the analyses, I was intrigued by the lack of effect that Team had on the model. Maybe it was because there are 20 teams and there would be many fewer observations per team. The more likely answer is that the consistency of skills at the professional gaming level is so high that the most likely variables to affect the model would be hard to quantify; such as teamwork, play style, or something else during the match. Based on this data alone, it does not appear that the best and worst teams would have very different models, but perhaps a larger set of observations would indicate this.