Adam Wisowaty
Econ 203

The effect of baseball statistics on a pitcher's earned run average

1. Introduction

In Major League Baseball(MLB), earned run average(ERA) is the most commonly used statistic to evaluate how well a pitcher is playing. ERA is defined by MLB as the representation of how many runs a pitcher gives up per nine innings, this makes it easier to compare pitchers because their statistics are scaled to the same number of innings. ERA is an important statistic that teams look at when evaluating pitchers, as the pitcher with the lowest ERA is considered the best and a pitcher with the highest ERA is considered the worst.

In this study I want to determine what other baseball statistics are most important in determining a pitcher's ERA. To test this I created a regression model that includes the data that I thought would be most important in determining a pitcher's ERA: groundball to flyball ratio, swinging strike percentage, fastball velocity, strikeout percentage, walk percentage and a dummy variable for which league the player plays in.

To evaluate how these variables impact a pitcher's ERA, I created a custom leaderboard at Fangraphs.com by filtering the data to include only pitchers who threw more than 160 innings from 2016 to 2018. I think sorting the data this way reduced the odds of an outlier by requiring sufficient sample size and created relevant information by including only the three most recent years. After creating the leaderboard, I exported the data into Excel and ran a multiple regression for ERA on the independent variables I chose.

After running the regression for the full model, I removed variables that were insignificant and ran another regression for the reduced model. I proved that this model satisfies the assumption of normality of errors by creating a histogram of the standardized residuals. The model also satisfies the assumptions of homoskedasticity, normality of errors, no autocorrelation, no serious outliers and no serious multicollinearity. To prove these assumptions were satisfied I included the histograms, scatter plots of residuals vs. predicted ERA and correlation table for the data. I did not need to transform the dependent variable or introduce any new variables because all of the assumptions were satisfied.

Results of my study confirm that other baseball statistics have a significant impact on a pitcher's ERA. The evidence supports my initial prediction that strikeout percentage, walk percentage and which league a player is in are significant factors in determining a pitcher's ERA at the 10% significance level. However, contrary to my initial prediction, the variables groundball to flyball ratio, swinging strike percentage and fastball velocity did not have any significant correlation to a pitcher's ERA, according to my data.

2. Data

For the study, I use data from Fangraphs.com which is a baseball statistics website that has its data provided by Baseball Info Solutions, a company that has been collecting MLB data for almost two decades. Since the main goal of my study is to interpret the significance that other baseball statistics have on a pitcher's ERA, the dependent variable of the study is pitchers' ERA(*ERA*) and the independent variables are groundball to flyball ratio(*GB/FB*), swinging strike percentage(*SwStr%*), fastball velocity(*vFA*), strikeout percentage(*K%*), walk percentage(*BB%*) and a dummy variable for pitchers' league(*League*) that equals 1 if a pitcher is in the National League(and 0 if the American League).

The descriptive statistics for my model are represented in Table 1 below. I found the mean ERA for 104 pitchers from 2016 to 2018 to be approximately 4, with a maximum of 5.82 and a minimum of 2.26. Figure 1 shows the scatter plots of ERA against the independent variables, based on these charts it appears that the assumption of no serious outliers is satisfied by this data sample. Figures 2 and 4 show the histograms of the standardized residuals for the initial model and reduced model, respectively. These histograms of the residuals appear to be very normally distributed, therefore I concluded that the assumption of normality of errors is satisfied in both the initial and reduced models. Figures 3 and 5 are scatter plots of residuals against the predicted ERA for the initial model and reduced model, respectively. The data on these plots appears to be homoskedastic, so I concluded that the assumption of no heteroskedasticity is satisfied for both the initial and reduced regression models. Table 2 is the correlation table that shows correlation coefficients between each of the individual variables. Using a correlation coefficient cut-off of an absolute value of 0.8 or higher for serious multicollinearity, I found that the variables swinging strike percentage and strikeout percentage had an issue with serious multicollinearity because their correlation coefficient was 0.871. The data for this study was not time series data, so the assumption of no autocorrelation is satisfied.

3. Regression Analysis

To determine the effect of other baseball statistics on a pitcher's earned run average, I estimated the initial model to be:

$$ERA = \beta_0 + \beta_1 GB/FB + \beta_2 SwStr\% + \beta_3 vFA + \beta_4 K\% + \beta_5 BB\% + \beta_6 League + \varepsilon$$

As I explained in my introduction, these statistics are what I believe would have the most explanatory power over a pitcher's ERA. These statistics are considered throughout baseball to be some of the most important statistics in determining how well a pitcher is playing and would certainly be taken into consideration by MLB teams when making decisions.

The initial regression results for the study are displayed in Table 3 below. This model appears to be a decent fit, based on the R-squared value being .5180 and the adjusted R-squared value of .4882. With an F-statistic of 17.3747 and a p-value that is very close to zero, the null hypothesis for the overall model validity test can be rejected at the 10% significance level, so the model is valid. However, upon inspecting the p-values for individual t-tests it is clear that the variables for groundball to flyball ratio, swinging strike percentage and fastball velocity are insignificant at the 10% level, so I removed these for the reduced model.

Before constructing the reduced model, I checked to make sure there were no other assumption violations in the model by creating another histogram of standardized residuals and scatter plot of residuals against predicted ERA with the insignificant variables removed. The histogram for the reduced model(Figure 4) appears to be very normally distributed, so the assumption of normality of errors is satisfied. The scatter plot of the residuals for the reduced model(Figure 5) appears to be homoskedastic, so the assumption of no heteroskedasticity is satisfied. As I mentioned earlier, there was an issue with serious multicollinearity between the variables swinging strike percentage and strikeout percentage, but since I am dropping swinging strike percentage in the reduced model this will no longer be a problem.

The regression with a reduced amount of variables also appears to be a decent model for estimating pitcher ERA because R-squared is .5088 and adjusted R-squared increased from .4882 to .4941. The F-statistic for the reduced model is 34.53 and the p-value is very close to zero which means the null hypothesis for the overall model validity test can be rejected, so the model is valid. The individual variables are all significant at the 10% level because all of the p-values for individual variables are less than 10%. To determine whether the initial model or reduced model is more appropriate I conducted a partial F-test. The partial F-test statistic comes out to be .6165 and the critical value for the partial F-test at the 10% level is 2.1411, so the null hypothesis is not rejected and the reduced model appears to be more appropriate.

Therefore my final reduced regression model will be:

$$\boldsymbol{ERA = \beta_0 + \beta_1 K\% + \beta_2 BB\% + \beta_3 League + \varepsilon}$$

4. Empirical Results

I was able to interpret the results of the study by looking at the coefficients of each variable in the reduced model regression results(Table 4). The coefficient for strikeout percentage suggests that a increase of 1% in strikeout percentage leads to a decrease of -.0996 in a pitcher's ERA. The walk percentage coefficient shows that a 1% increase in walk percentage coincides with a .0822 increase in a pitcher's ERA. The league coefficient suggests that a pitcher in the National League would have an ERA -.1910 less than a pitcher in the American League.

These results for the significant variables are exactly what I was expecting to find when I began this study. More strikeouts are considered positive for a pitcher because it produces an out that has no chance to become a hit since the ball is not put into play. Therefore I was expecting strikeout percentage to have a negative linear relationship with a pitcher's ERA. Walks are usually considered a bad thing for pitchers because it gives the offense a free base without needing to hit the ball. Thus, I expected walks to have a positive linear relationship with ERA. National League pitchers have the advantage of no designated hitter—a player that hits instead of the pitcher since pitchers are usually terrible hitters—which makes it slightly easier for them to pitch since they get to face one easier batter. Consequently, I would expect pitchers in the National League to have lower ERAs than pitchers in the American League and the model shows that this is the case with the League variable having a negative coefficient.

5. Summary and Discussion

This study investigated the relationship between baseball statistics and a pitcher's ERA. I found that strikeout percentage, walk percentage and which league a pitcher plays in are all significant factors in determining a pitcher's ERA. However, the study also showed groundball to flyball ratio, swinging strike percentage and fastball velocity are not significant factors to a pitcher's ERA, contrary to my initial thoughts.

One potential shortfall of this study is that the data I used did not include relief pitchers or pitchers who threw less than 160 innings per season from 2016 to 2018. I did not feel as though it would be appropriate to include them in the data because the few innings these pitchers throw is a very small sample size. However, these pitchers often throw harder and are more dominant in the fewer innings that they pitch, so they may have increased the significance of the fastball velocity or swinging strike percentage.

Another potential shortfall of this study is that it does not account for the ballpark a pitcher plays in. Certain stadiums in Major League Baseball are much more conducive to having a lot of offense. For example, the Colorado Rockies stadium always has the most offense because the ball flies farther due to the thinner air at the higher elevation. An ideal study on key statistics that impact a pitcher's ERA would adjust the data for pitchers based on their home stadium.

Figure 1: Scatter Plots of ERA vs independent variables

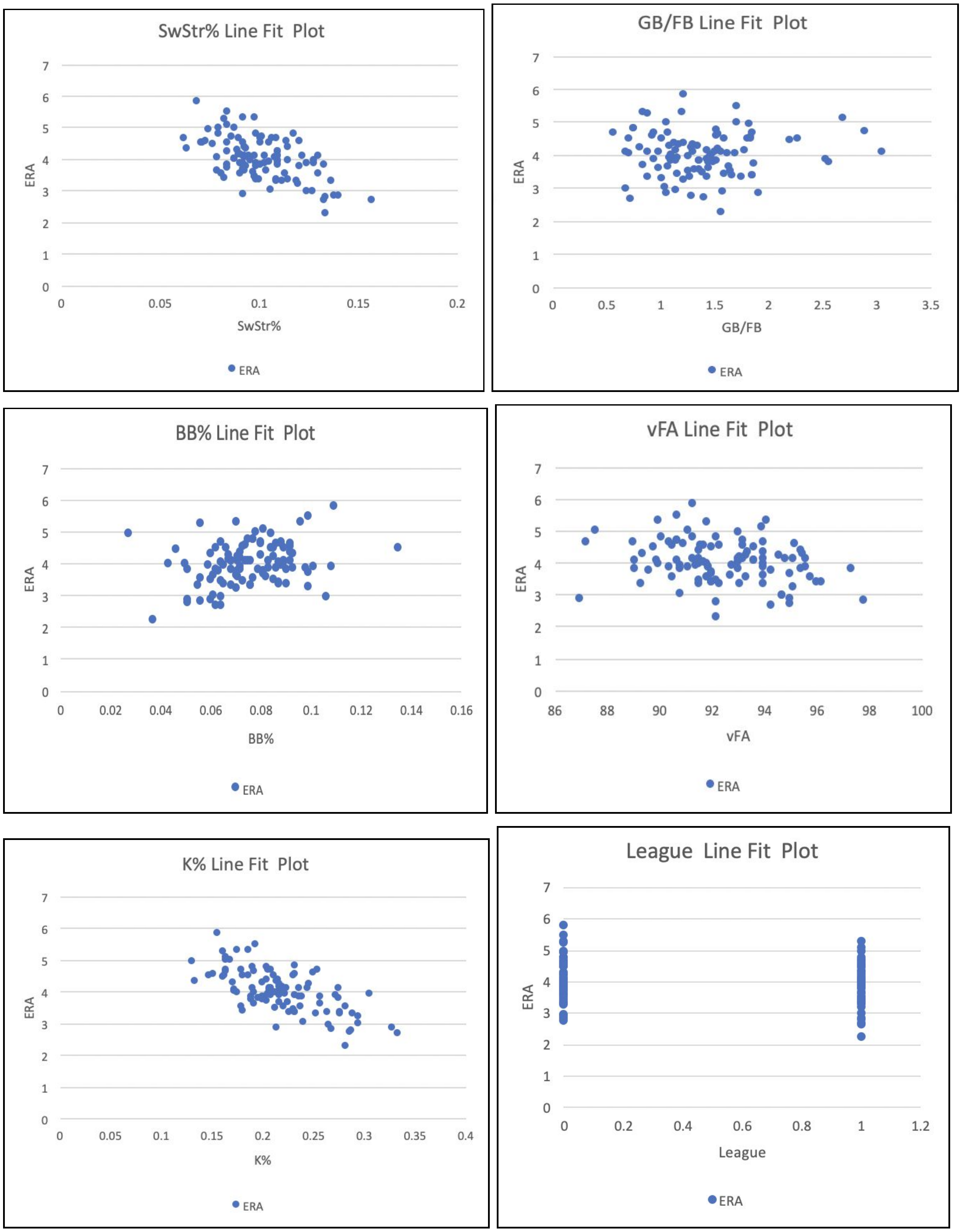


Figure 2: Histogram of standardized residuals for initial model

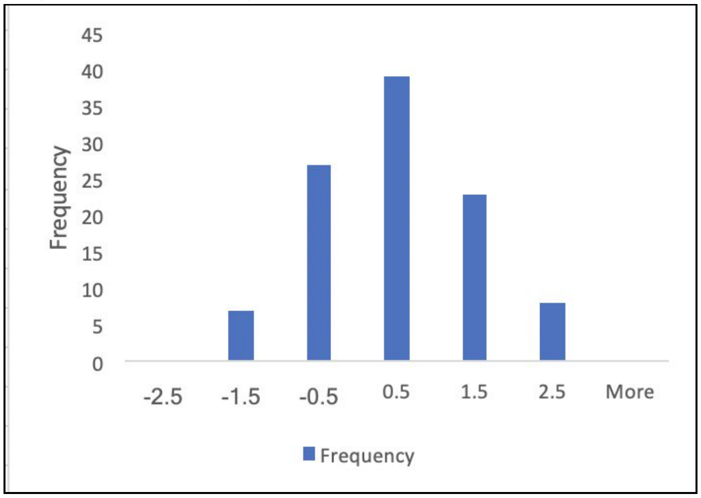


Figure 3: Residuals vs. Predicted ERA for initial model

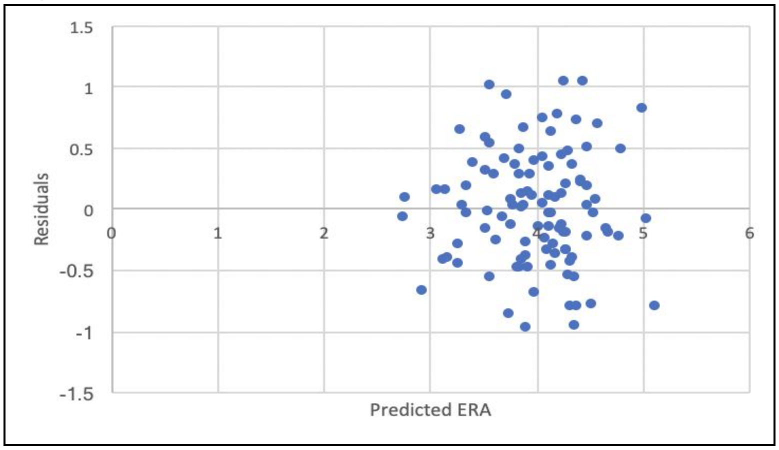


Figure 4: Histogram of standardized residuals for reduced model

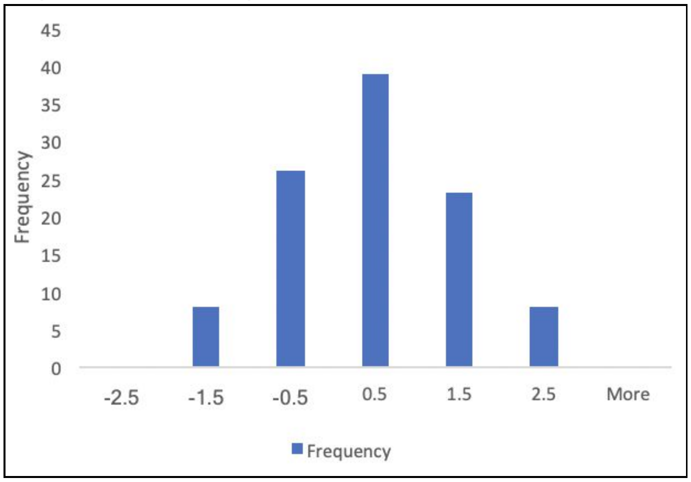


Figure 5: Residuals vs. Predicted ERA for reduced model

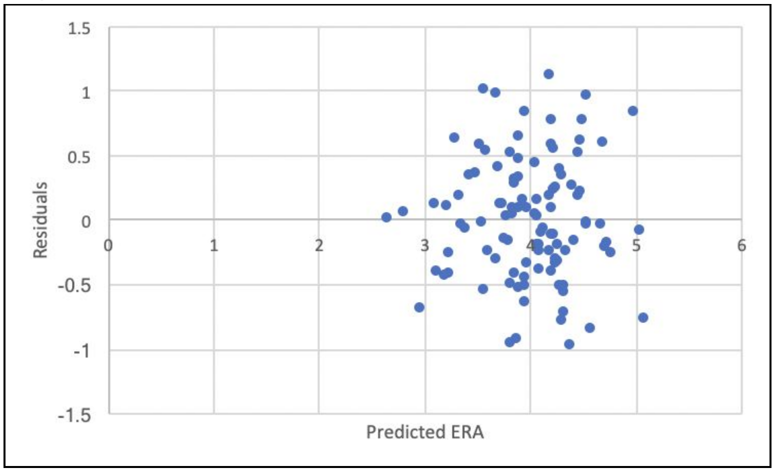
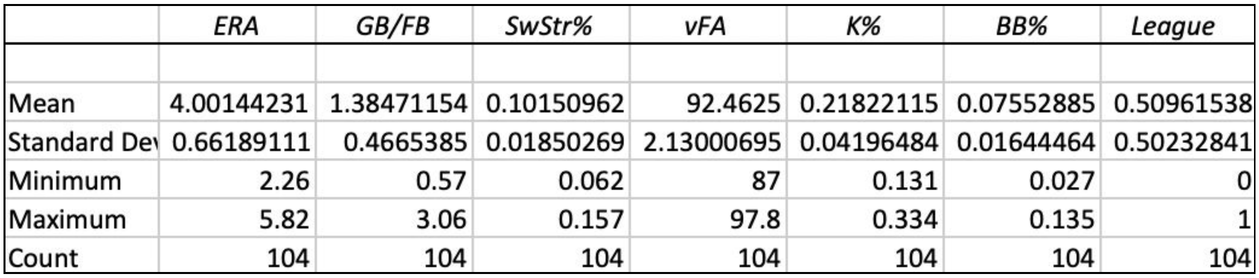
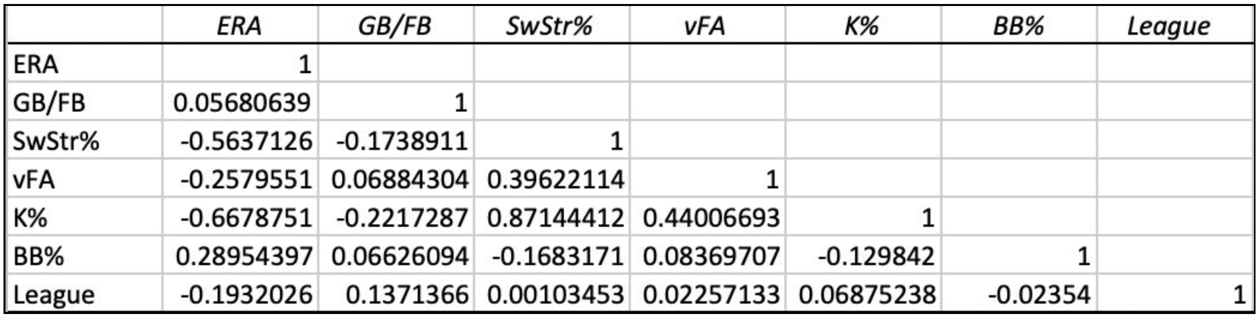


Table 1: Descriptive Statistics



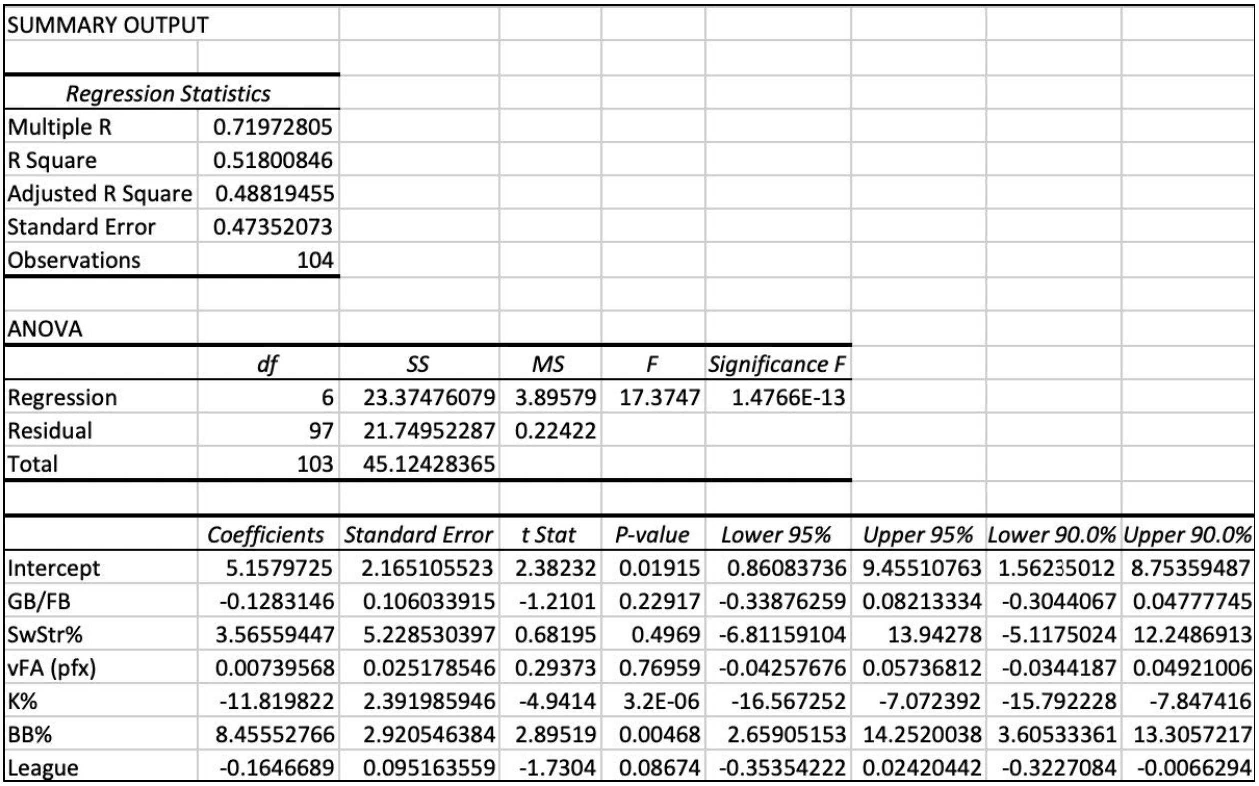
	ERA	*GB/FB*	*SwStr%*	*vFA*	*K%*	*BB%*	*League*
Mean	4.00144231	1.38471154	0.10150962	92.4625	0.21822115	0.07552885	0.50961538
Standard Dev	0.66189111	0.4665385	0.01850269	2.13000695	0.04196484	0.01644464	0.50232841
Minimum	2.26	0.57	0.062	87	0.131	0.027	0
Maximum	5.82	3.06	0.157	97.8	0.334	0.135	1
Count	104	104	104	104	104	104	104

Table 2: Correlations



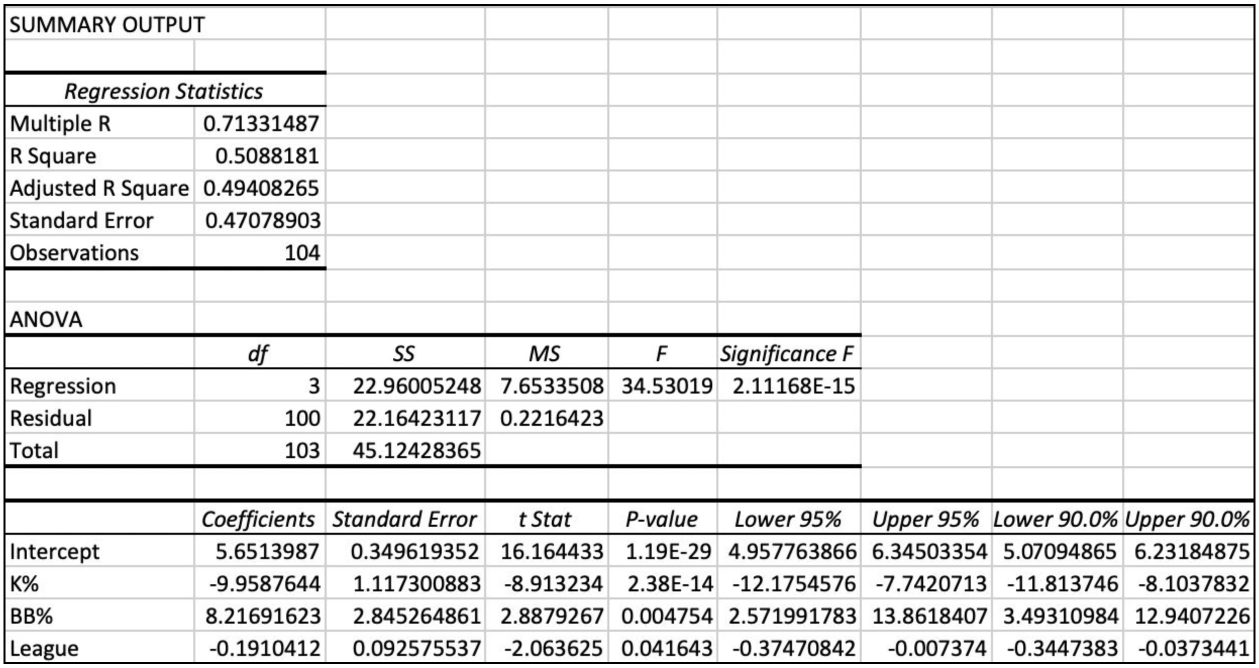
	ERA	*GB/FB*	*SwStr%*	*vFA*	*K%*	*BB%*	*League*
ERA	1						
GB/FB	0.05680639	1					
SwStr%	-0.5637126	-0.1738911	1				
vFA	-0.2579551	0.06884304	0.39622114	1			
K%	-0.6678751	-0.2217287	0.87144412	0.44006693	1		
BB%	0.28954397	0.06626094	-0.1683171	0.08369707	-0.129842	1	
League	-0.1932026	0.1371366	0.00103453	0.02257133	0.06875238	-0.02354	1

Table 3: Initial Model Regression Results



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.71972805							
R Square	0.51800846							
Adjusted R Square	0.48819455							
Standard Error	0.47352073							
Observations	104							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	6	23.37476079	3.89579	17.3747	1.4766E-13			
Residual	97	21.74952287	0.22422					
Total	103	45.12428365						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	5.1579725	2.165105523	2.38232	0.01915	0.86083736	9.45510763	1.56235012	8.75359487
GB/FB	-0.1283146	0.106033915	-1.2101	0.22917	-0.33876259	0.08213334	-0.3044067	0.04777745
SwStr%	3.56559447	5.228530397	0.68195	0.4969	-6.81159104	13.94278	-5.1175024	12.2486913
vFA (pfx)	0.00739568	0.025178546	0.29373	0.76959	-0.04257676	0.05736812	-0.0344187	0.04921006
K%	-11.819822	2.391985946	-4.9414	3.2E-06	-16.567252	-7.072392	-15.792228	-7.847416
BB%	8.45552766	2.920546384	2.89519	0.00468	2.65905153	14.2520038	3.60533361	13.3057217
League	-0.1646689	0.095163559	-1.7304	0.08674	-0.35354222	0.02420442	-0.3227084	-0.0066294

Table 4: Reduced Model Regression Results



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.71331487							
R Square	0.5088181							
Adjusted R Square	0.49408265							
Standard Error	0.47078903							
Observations	104							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	22.96005248	7.6533508	34.53019	2.11168E-15			
Residual	100	22.16423117	0.2216423					
Total	103	45.12428365						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	5.6513987	0.349619352	16.164433	1.19E-29	4.957763866	6.34503354	5.07094865	6.23184875
K%	-9.9587644	1.117300883	-8.913234	2.38E-14	-12.1754576	-7.7420713	-11.813746	-8.1037832
BB%	8.21691623	2.845264861	2.8879267	0.004754	2.571991783	13.8618407	3.49310984	12.9407226
League	-0.1910412	0.092575537	-2.063625	0.041643	-0.37470842	-0.007374	-0.3447383	-0.0373441