BIOS 2049 Final Project

Predicting Earned Run Average in Major League Baseball

Paul C. Brendel

Abstract

Objective: Major League Baseball (MLB) scouts and fantasy baseball participants need an effective system to determine which starting pitchers will have the lowest earned run average (ERA). This study attempted to determine the relationship between ERA, strikeout to walk ratio (K:BB), ground ball to fly ball ratio (GB:FB), first strike percentage (F-Strike%), and fastball velocity (FBv). The study also attempted to build a predictive model for ERA using these predictors.

Methods: A sample of 96 starting pitchers was chosen based on having pitched over 150 innings in the 2013 MLB season. Simple linear regression was performed to assess the individual abilities of K:BB, GB:FB, F-Strike%, and FBv to predict ERA. Multiple linear regression was used to figure out the best model for predicting ERA.

Results: The simple linear regression revealed that each metric, except for GB:FB, was a significant predictor of ERA. With an R^2 of approximately 32%, K:BB was a much better individual predictor than the others. Multiple linear regression revealed that the best model for predicting ERA includes only K:BB and FBv. With an R^2 of only about 35%, this model is still far from ideal for predicting ERA.

Conclusions: K:BB may have been the best predictor since it was the only independent variable directly correlated with opponent on base results. Future studies may want to examine predictors that also have this quality, such as left on base percentage.

Introduction

In a survey conducted in January 2014 by the Harris Poll, Major League Baseball (MLB) was voted as the second most popular sport in America.[1] As any sports fan knows, MLB is big business. As of March 2014, the 30 different MLB franchises were all valued above 485 million dollars, with the prestigious New York Yankees valued at a high of 2.5 billion dollars.[2] The combined desire to satisfy fans and maximize revenue puts great pressure on MLB franchises to put the best possible players on the field in order to win games.

However, it is not only the various MLB scouts and managers that have to make these decisions on evaluating player talent. Millions of fantasy baseball participants also evaluate

MLB players every year. A fantasy sport is a game in which participants act as owners to a team and build a team that competes against other owners based on statistics generated by real individuals in the professional sport. It is estimated by the Fantasy Sports Trade Association that 32 million people aged 12 and above in the U.S. and Canada played fantasy sports in 2010.[3]

In baseball (of either the "real" or "fantasy" variety) the starting pitcher is arguably the most influential position. A game of baseball is won by scoring the most runs, therefore a pitcher that allows the fewest amount of runs is most desirable. There are countless strategies in trying to predict which pitchers will allow the fewest runs or, in other words, have the lowest earned run average (ERA; ERA = 9 * earned runs allowed / innings pitched). Some prefer to "eyeball" talent. Others will look at various pitching metrics to make these conclusions.

One popular metric is fastball velocity. The logic behind this choice is pretty simple; the faster a baseball is pitched, the harder it is to hit. Two other common measures are strikeout-to-walk ratio and ground ball-to-fly ball ratio. The thinking behind these measures is similar – strikeouts and ground balls generally keep opponents off the bases whereas walks and fly balls generally put opponents on the bases. If opponents are on base, they are in a more advantageous position to score runs. Lastly, first pitch strike percentage (the percentage of batters faced that the first pitch was a strike) is commonly examined. There are numerous reasons (beyond the scope of this paper) that this metric is thought to be predictive of ERA beyond the bare fact that a strike is a desirable pitch result.[4] This study plans on establishing the relationship between ERA, fastball velocity, strikeout-to-walk ratio, ground ball-to-fly ball ratio, and first pitch strike percentage. This study will also attempt to build a predictive model for ERA. It is hypothesized that all four measures will significantly contribute to predicting ERA to varying extents.

Methods

Sample Selection:

Starting pitchers were chosen in the sample for this study by selecting those who pitched at least 150 innings during the 2013 season. As a result, 96 starting pitchers are included in the sample. All data was obtained from *www.fangraphs.com*. Table 1 below shows the data for 10 of the 96 pitchers.

Statistical Analysis:

Some descriptive statistics, including mean and standard deviation, were calculated for ERA, fastball velocity in miles per hour (FBv), strikeout-to-walk ratio (K:BB), ground ball-to-fly ball ratio (GB:FB), and first pitch strike percentage (F-Strike%; Table 2). Histograms were plotted for all variables to ascertain normality (Figure 1). Next, scatter plots of ERA vs each of the four dependent variables (FBv, K:BB, GB:FB, F-Strike%) were graphed (Figure 2). These plots were assessed for any non-linear patterns.

Simple linear regression was then performed for each independent variable to assess predictive abilities. If the p-value of the overall F test had a value less than .05, then that variable was considered to explain a significant amount of variation in ERA. In addition, the R^2 value was analyzed to determine the total amount of variance in ERA explained by the independent variable. To look for non-normal patterns, jackknife residuals were calculated and plotted against the fitted values (Figure 3). The interaction terms for K:BB and GB:FB, K:BB and F-Strike%, K:BB and FBv, GB:FB and F-Strike%, GB:FB and FBv, and F-Strike% and FBv were created and tested (via the "testparm" command) in order to check for interaction effects. An α value of .05 was used as the cutoff for determining parallelism.

Multiple linear regression was used to determine the best model for predicting ERA. Using all combinations of predictors, 15 models were evaluated (Table 3). To ascertain the best model, the following measures were calculated for each model: R^2, mean squared error (MSE), overall F test p-value, Akaike Information Criteria (AIC), and Bayes Information Criteria (BIC).

Once the best model was selected, stepwise regression using a probability for entry of .30 and a probability for removal of .40 was calculated to confirm the chosen model. Pairwise correlation coefficients were then calculated to examine correlation among independent variables. Residuals for the final model were calculated and plotted as a histogram and vs the fitted values (Figure 4). Since this plot showed a slight fanning pattern, a heteroskedasticity test was performed.

Outliers and influential points were evaluated by calculating the values for leverage and Cook's distance. These values were displayed in box plots then assessed (Figure 5). The final model was refitted if any points were determined to be influential. All analysis was performed using Stata version 13.

Table 1: Sample of 10 observations in data set

Name	ERA	K:BB	GB:FB	F-Strike %	FBv
Justin Masterson	3.45	2.57	2.40	58.5	91.6
Adam Wainwright	2.94	6.26	1.78	64.5	91.1
Jordan Zimmerman	3.25	4.03	1.52	66.9	93.9
Jon Lester	3.75	2.64	1.27	60.9	92.7
Mike Pelfrey	5.19	1.91	1.20	55.3	92.4
Doug Fister	3.67	3.61	2.23	59.0	88.8
Paul Maholm	4.41	2.23	2.07	63.9	87.1
Bartolo Colon	2.65	4.03	1.09	64.9	89.9
Yu Darvish	2.83	3.46	1.08	57.6	92.9
Mike Minor	3.21	3.93	0.82	64.5	90.4

Table 2: Descriptive statistics

Variable	Mean	StDev	Min	Median	Max
ERA	3.72	0.78	1.83	3.58	5.86
K:BB	3.01	1.04	1.01	2.88	6.94

GB:FB	1.38	0.41	0.65	1.34	2.41
F-Strike %	61.3	3.4	50.9	61.2	70.2
FBv	91.1	2.3	81.9	91.5	95.8

Table 3: Analysis of models

Model	Variables	R^2	MSE	F test p-value	AIC	BIC
1	K:BB	.32	.41	<.0001	189.97	195.10
2	GB:FB	.01	.61	.3365	226.19	231.32
3	F-Strike%	.10	.55	.0014	216.71	221.84
4	FBv	.06	.57	.0132	220.83	225.96
5	K:BB, GB:FB	.32	.42	<.0001	191.46	199.16
6	K:BB, F-Strike%	.32	.42	<.0001	191.80	199.49
7	K:BB, FBv	.35	.40	<.0001	187.70	195.39
8	GB:FB, F-Strike%	.11	.55	.0056	218.44	226.13
9	GB:FB, FBv	.08	.57	.0170	220.72	228.41
10	F-Strike%, FBv	.17	.51	.0002	211.43	219.12
11	K:BB, GB:FB, F-Strike%	.33	.42	<.0001	193.19	203.45
12	K:BB, GB:FB, FBv	.36	.40	<.0001	188.45	198.71
13	K:BB, F-Strike%, FBv	.35	.41	<.0001	189.68	199.93
14	GB:FB, F-Strike%, FBv	.18	.51	.0004	212.39	222.65
15	K:BB, GB:FB, F-Strike%, FBv	.36	.40	<.0001	190.37	203.19

Results

The descriptive statistics (Table 2) reveal that ERA has a mean of 3.72±0.78. The range of ERA values went from 1.83 by the 2013 NL Cy Young Award winner Clayton Kershaw to 5.86 by Lucas Harrell with a median of 3.58. The other mean values are as follows: 3.01±1.04 for K:BB; 1.38±0.41 for GB:FB; 61.3±3.4 for F-Strike%; and 91.1±2.3 for FBv. Based on the histograms (Figure 1) it is safe to assume that each of the 5 variables is normally distributed. The scatter plots (Figure 2) show that ERA vs K:BB may have some curvature. Otherwise, the rest of the scatter plots do not point to any non-linear patterns.

Based off the simple linear regression models, each independent variable, except for GB:FB, is a significant predictor of ERA. Despite being significant predictors, the other 3 independent variables did not have very high R^2 values: .32 for K:BB; .10 for F-Strike%; and .06 for FBv. K:BB had the best fit and was able to explain approximately 32% of the variance in ERA, which is much better than the other predictors. The model with K:BB did not, however, have the greatest precision. The greatest precision was seen in the model with F-Strike%, which had the narrowest 95% confidence interval at (-.12, -.03). Upon examining the plots of residuals vs fitted values (Figure 3), it appears that K:BB may exhibit a slight "fanning" pattern. All other plots provide evidence of normality.

Of the six interaction effects tested, one was found to be significant: K:BB and F-Strike% p=.003. The rest of the interaction effects were not significant: K:BB and GB:FB p=.26; K:BB and FBv p=.37; GB:FB and F-Strike% p=.18; GB:FB and FBv p=.77; F-Strike% and FBv p=.12. Consequently, if it is determined that the best model for predicting ERA should include K:BB and F-Strike%, the corresponding interaction term must also be included in the model.

After examining all of the evidence, Model 7, which includes K:BB and FBv, is the best model for predicting ERA. It has an R^2 value that is only 1% less than that of the maximum model. It is tied with Model 12 and the maximum model with the lowest MSE of .40. It is tied with several of the models with having the best overall F test p-value of <.0001. By comparing AIC, Model 7 is ideal because it is within 3 units of the max model and is lower than the AIC values seen in the other models. Using BIC, Model 11 is superior since it is the only model with a difference of less than 2 units compared to the max model.

The stepwise regression yielded a model containing K:BB, GB:FB, and FBv. Since the characteristics of this model differ very little compared to the model with only K:BB and FBv, the model with less predictors (Model 7) is still believed to be the best model. None of the correlation coefficients between independent variables exceeded .80, which indicates that there are not problems with collinearity. Although the plot of residuals vs fitted values of this model appeared to some fanning (Figure 4), the test of heteroskedasticity had a p-value of .53, which indicates that there is constant variance. The histogram of residuals provides no evidence against normality (Figure 4).

A leverage value would be problematic if it is greater than .0625. None of the leverage values exceeded this mark (Figure 5). A Cook's distance value would be problematic if it exceeded .042. Similarly, none of the Cook's distance values surpassed this mark (Figure 5). The final model thus does not need any adjustments for influential points.

The final prediction model is: ERA = 10.40 – 0.41*K:BB – 0.06*FBv. The values for these predictors mean that ERA decreases 0.41 points for every unit increase in K:BB (adjusting for FBv) and that ERA decreases 0.06 points for every unit increase in FBv (adjusting for K:BB). As an example of implementing this model, if a pitcher has a K:BB of 3 and a FBv of 92 miles per hour, then that pitcher's ERA would be predicted to equal 10.40 – 0.41(3) – 0.06(92) or 3.65.

Discussion

This study aimed to create a predictive model for ERA and to examine the relationship between ERA, K:BB, GB:FB, F-Strike%, and FBv. The analysis showed that all 4 predictors, except for GB:FB, were individually significant predictors of ERA. The scatter plot of ERA vs GB:FB shows no real pattern, which confirms this result (Figure 2). It seems that baseball's "conventional wisdom" of groundball pitchers being superior to fly ball pitchers may not have much merit. By a rather wide margin, K:BB was the best individual predictor of ERA. This result may be explained by the fact that it is the only predictor that *directly* correlates with opponents getting on/off base. In other words, a strikeout will always prevent an opponent from getting on base, but an incredibly high velocity (100 MPH) fastball does not guarantee that the opponent will not get on base.

Using the four chosen predictors, it was determined that the best model for predicting ERA included only K:BB and FBv. Despite being the best model, it could only predict approximately 35% of the variance in ERA, which is far from ideal. To improve on the model, it may be a good idea to use metrics like K:BB that are directly correlated with on base activity. One such statistic could be left on base percentage (LOB%), which measures the percentage of base runners that a pitcher strands on base over the course of the season.[5]

The validity of the analysis may be limited by the 150 innings pitched limit that was required of pitchers in order to be included in the sample. This limitation was implemented to ensure that statistics were truly representative of that pitcher over the course of the season. However, there may be bias due to the fact that a pitcher has to be performing relatively successfully in order for the manager to allow the pitcher to pitch that many innings over the season. We may be missing some important data by excluding the pitchers that performed exceptionally poorly in 2013. Furthermore, the findings could be more robust if pitchers from other seasons were included in the sample. Some would argue that the 2013 season was an unusually strong season for pitchers since it came near the end of MLB's "steroid era."[6] Lastly, ERA vs K:BB appeared to exhibit a slightly curved pattern, so a non-linear form of regression may have been the best strategy for choosing the best model.

Appendix

Figure 1: Histograms of distributions of ERA, K:BB, GB:FB, F-Strike%, and FBv

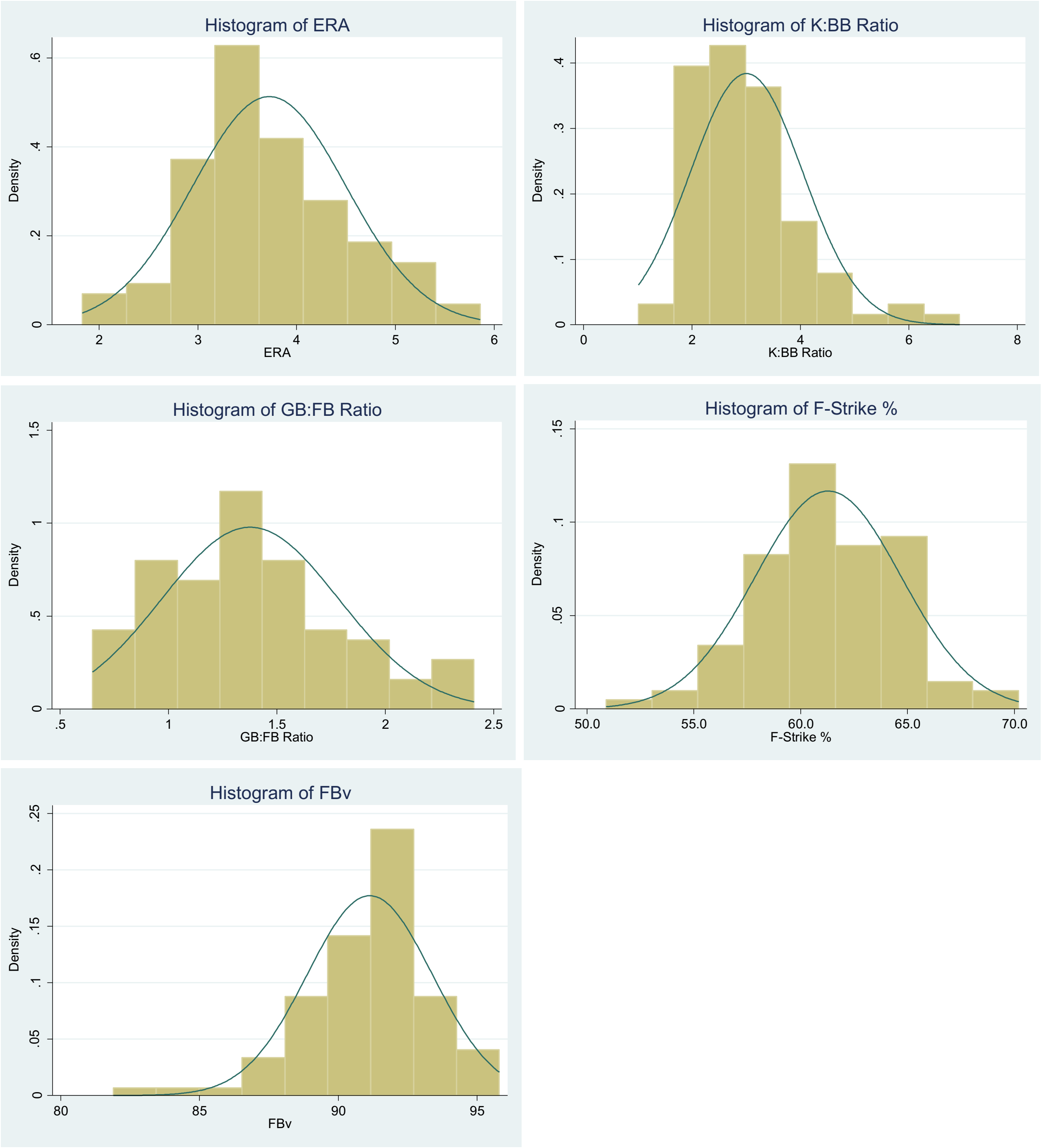


Figure 2: Scatter plots of ERA vs independent variables

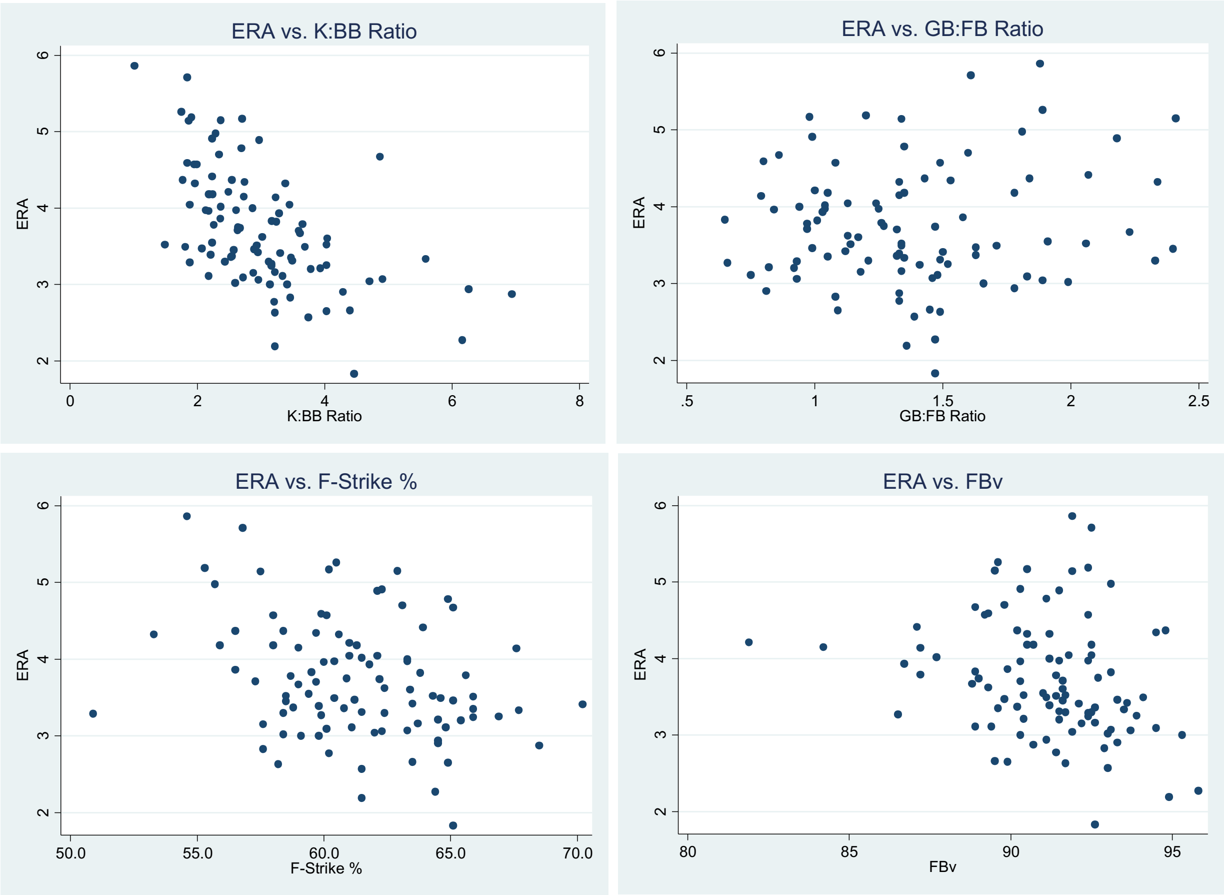
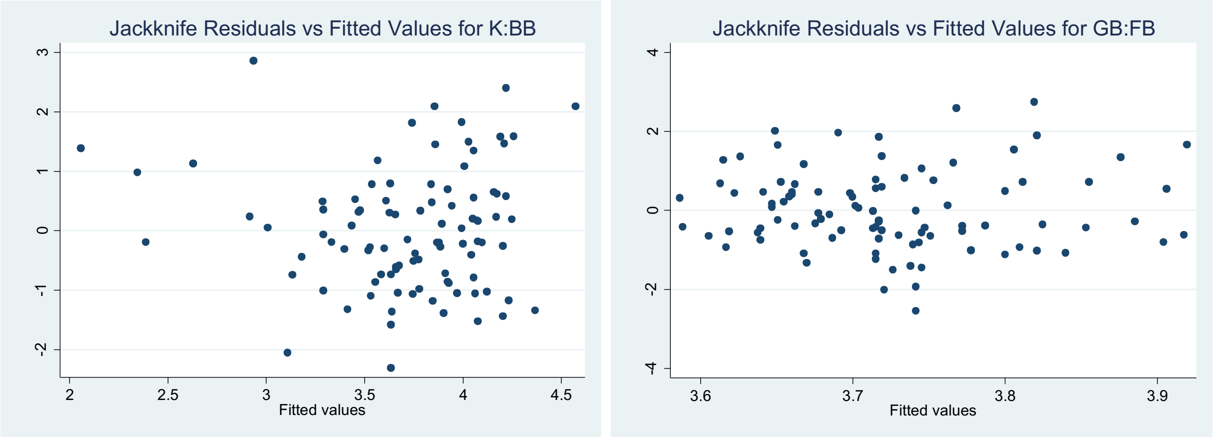


Figure 3: Residuals vs fitted values for independent variables



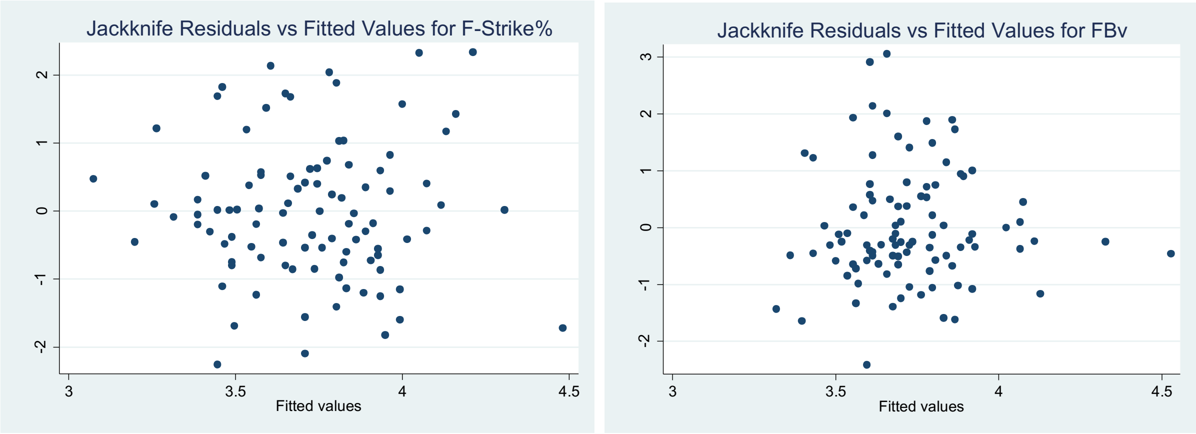


Figure 4: Residuals vs fitted values for best model

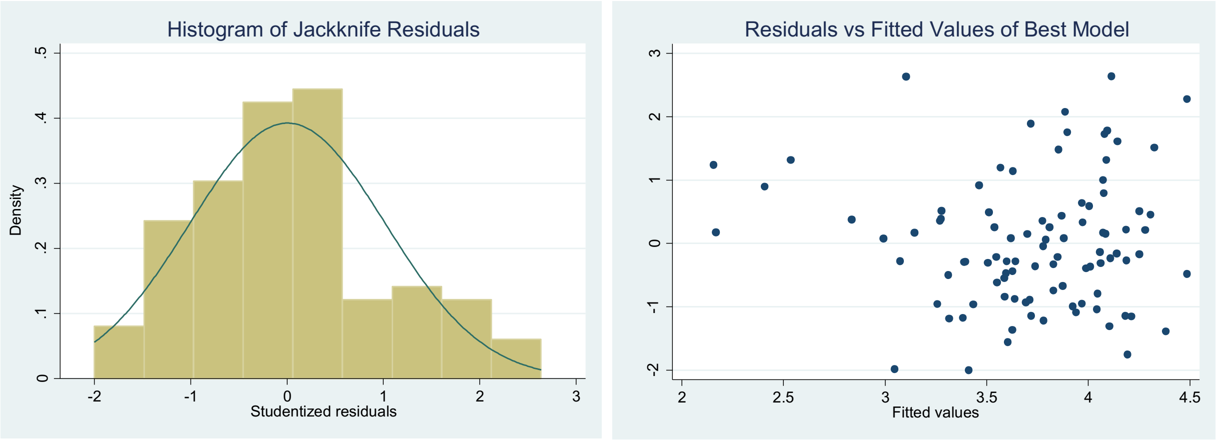
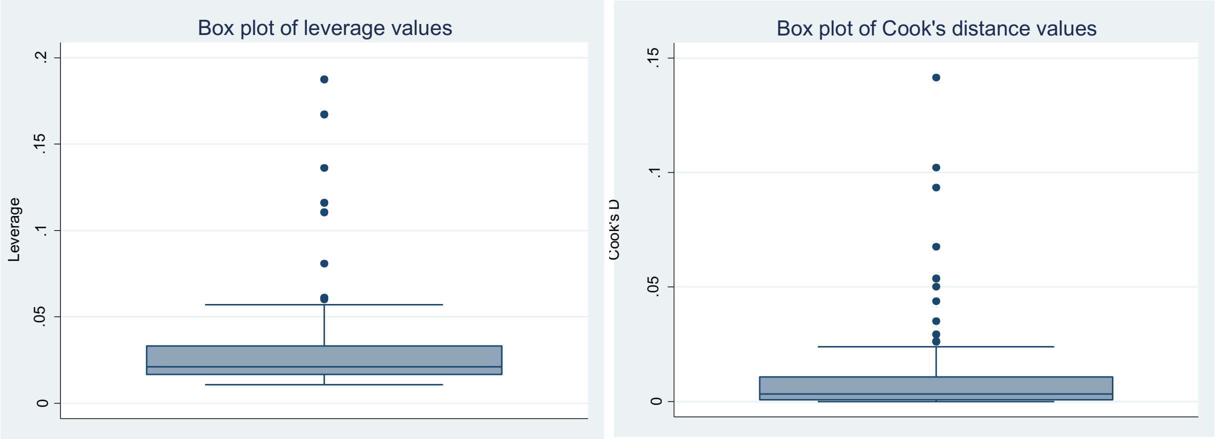


Figure 5: Box plots for evaluating outliers and influential points



References

1. Rovell, Darren. "NFL Most Popular for 30th Year In row." *ESPN*. ESPN Internet Ventures, 26 Jan. 2014. Web. 15 Apr. 2014.
2. "The Business of Baseball." *Forbes*. Forbes Magazine, n.d. Web. 17 Apr. 2014.
3. Clapham, Kyle. "Fantasy Sports Becoming Big Business as Popularity Continues to Rise." *Medill Reports - Chicago*. N.p., 14 May 2012. Web. 18 Apr. 2014.
4. "Plate Discipline (O-Swing%, Z-Swing%, Etc.)." *FanGraphs Sabermetrics*. N.p., n.d. Web. 19 Apr. 2014.
5. "LOB%." *FanGraphs Sabermetrics*. N.p., n.d. Web. 19 Apr. 2014.
6. "The Steroids Era." *ESPN*. ESPN Internet Ventures, 5 Dec. 2012. Web. 17 Apr. 2014.