STP 429 – Applied Regression

Arizona State University

Tanay Gandhi

Lab #2

Executive Summary

The business surrounding Major League Baseball is a multi-billion dollar industry and is followed religiously by millions of people. Millions of dollars of revenue for each participating team are raised by advertisements, merchandise, sponsorships, game winnings, and much more. Teams are not all worth the same, however. The advertising department of a baseball bat company would rather spend its $5 million budget on the team that won the world series, rather than a team that didn't even make it to their Division Series. This is one example of why it's in a team's best financial interest to identify factors that correlated with their game wins, as well as factors that can be worked on and improved.

This study was developed to identify three significant factors that correlate with number of game wins per year of the St. Louis Cardinals. Potential factors identified are earned run average as pitching team (ERA), fielding percentage as batting team (Fld%_B), walks as pitching team (BB_P), errors committed as batting team (E_B), and stolen bases as batting team (SB).

Using batting and pitching data provided by www.baseball-reference.com from 1882 to 2020

Data

The raw data received from Baseball Reference contained 139 observations, one for each year from 1882 to 2020. However, for the purposes of this study, we only used the observations from years between 1996 and 2019, or n=24 observations. The variables used extensively were team wins, earned run average, fielding percentage, walks (as pitching team), errors committed (as batting team), and stolen bases.

Methodology

Step 1 of this study was analyzing the raw data provided by Baseball Reference site and identifying variables in the data provided that could potentially be used in a predictive model for predicting team wins. This was accomplished by comparing the distributions of 14 initially identified variables with team wins, as well as comparing the individual correlations with team wins, to determine five variables more likely to produce a strong model. Once the five variables were identified, correlation tests were run to determine the strength of the relationship between each variable and team wins. Along with that was an analysis of the scatter plots between each variable and team wins, which helped identify potential interaction terms and potential higher-order terms. Finally, after generating multiple potential models, model effectiveness tests were performed on each to determine the most adequate model.

All statistical calculations were run using SAS software. Exploratory data analysis was done using `proc means`, `proc univariate`, and `proc corr`. The `proc corr` process was used to further analyze each variable. Finally, `proc reg`, was used to generate the regression models.

Results

Two criteria were used in the exploratory data analysis to identify the five variables to further build exploratory regression models from. The was to compare the histogram distribution of each variable against the distribution of teams wins, and identify potential similarities in skewness, modality, variance, etc. The second was running preliminary correlation procedures between each variable and team wins to identify variables with the strongest correlations. The team wins variable (W) is normally distributed and unimodal, with very little discernable skew (see figure 1). Some variable distributions to note are that of earned run average, which had

(figure 2) and fielding percentage (figure 3), both of which had fairly normal distributions that showed similarities with that of team wins. Unsurprisingly, according to table 13, which was the result of the preliminary correlation procedure, earned run average and fielding percentage had the two highest correlations with team wins respectively. The next highest absolute value of correlation was walks as pitching team (BB_P). The absolute value is mentioned as walks is strongly *negatively* correlated with team wins and would be equally useful in a predictive model. It's distribution (figure 4) however varies from team wins, as it is right skewed, possibly due to the outliers seen in the quartile observations in table 8. The last significant correlation to note is between team errors as batting team (E_B) and team wins, which was similarly strongly negative. Finally, the variables with the highest correlation with team wins that were chosen were: earned run average, fielding percentage, walks as pitching team, team errors as batting team, and stolen bases.

After identifying the 5 variables that correlated and related most strongly with team wins, each possible 3-variable regression model was generated, tested, and compared, to identify the best possible predictive model. One model, $W = \beta_0 + \beta_1(FLD\%\_B) + \beta_2(E_B) + \beta_3(SB)$ was removed from testing, as it only contained batting variables. The nine models produced that were compared were:

1. $W = -565.7756 - 10.146(ERA) + 702.75(Fld\%B) + 0.00717(BB\_P)$
2. $W = -1512.907 - 9.4658(ERA) + 1650.2924(Fld\%B) + 0.1596(E\_B)$
3. $W = -411.9387 - 9.6786(ERA) + 550.24711(Fld\%B) - 0.02435(SB)$
4. $W = 138.2848 - 9.9567(ERA) - .0783(E_B) - .0250(SB)$
5. $W = 135.2265 - 10.7443(ERA) - 0.0015(E_B) - 0.0392(SB)$
6. $W = -2279.2537 + 2395.3621(Fld\%B) - 0.0232(BB\_P) + 0.2408(E\_B)$

7. $W = -3914.3385 + 4024.0318(Fld\%B) + 0.4846(E_B) - 0.0365(SB)$
8. $W = 119.5797 - 0.03693(BB_P) - 0.1045(E_B) - 0.0129(SB)$
9. $W = 135.2265 - 10.7443(ERA) - 0.00154(BB_P) - 0.0392(SB)$

I then performed overall tests of statistical adequacy on each of the 9 models to determine their usefulness. The p-values used for these tests can be found in tables 14-24. The generalized null hypothesis for the overall test of utility performed on each model was of the form:

$$H_0: \beta_0 = \beta_1 = \cdots = \beta_n = 0, where\ n = number\ of\ coefficients$$

$$H_\alpha: At\ least\ one\ \beta_i \neq 0$$

With $\alpha = 0.05$, the test on every model but 6, 7, and 8 had the null hypothesis rejected, meaning that we can be 95% confident that models 1, 2, 3, 4, 5, and 9 are statistically adequate, and that we cannot determine with confidence that models 6, 7, and 8 are statistically adequate. Thus, the next step was comparing the adjusted r-squared values of each model to identify the base model that accounts for the most variation. Tables 25-30 contain metrics of each respective model still in consideration. The value considered was Adjusted R-Squared, and model 2's value was 0.3906, which is not very large, but was the largest out of the considered models. It means that model 2 by itself accounts for about 39% of the variability of the data. Another metric to note is the coefficient of variance, which is about 6.93%. In general, models should be below 10%, which this satisfies.

The next step in this study was to identify and test possible interaction terms and quadratic terms with this model to improve it. The scatter plots in figures 7 and 10 plot ERA and E_B against W respectively. Considering these plots and the residuals between these variables and W in regression model 2 (see figures 12 and 14), it was determined that there a possibility of

these variables interacting. Thus, a new regression model, 2a, was generated with this new interaction term between ERA and E_B:

$$W = -1495.66 - 7.4765(ERA) + 1624.9565(Fld\%B) + 0.236(E_B) - 0.0197(ERA)(E\_B)$$

The final item to test was possible quadratic terms that might help the model. It was determined that E_B could possibly be used in a quadratic term for W in a 2nd order model. Therefore, model 2b was generated:

$$W = -1781.8416 - 8.7770(ERA) + 1891.53996(Fld\%B) + 0.6992(E_B) - 0.00242(E_B)^2$$

First, the adjusted r-squared values for these new models were compared with model 2, and it was determined that both were comparatively lower. Then, A nested model f-test was performed on each of these 2 new models to determine if either individual compound model was more comprehensive than model 2. The general form of the hypotheses for these tests were:

$$H_0: Coefficients\ of\ additional\ terms = 0$$

$$H_\alpha: At\ least\ one\ of\ the\ coefficients\ of\ the\ additional\ terms\ does\ not\ equal\ 0$$

Appendix

SAS Code

```
data cardinals;
      set cardinalsFull;
      if Year > 1995 and Year < 2020;
run;

/* exploratory data analysis */

proc univariate data=cardinals;
      var W R_B _2B HR SB BA E_B DP_B Fld__B BatAge ERA BB_P SO_B PAge;
      histogram W R_B _2B HR SB BA E_B DP_B Fld__B BatAge ERA BB_P SO_B PAge;
run;

proc corr data=cardinals;
            var R_B _2B HR SB BA E_B DP_B Fld__B BatAge ERA BB_P SO_P PAge;
            with W;
run;


proc corr data=cardinals plots=scatter;
      var ERA Fld__B BB_P E_B SB;
      with W;
run;

/*Model creation*/

proc reg data=cardinals;
      model W = ERA Fld__B BB_P;

proc reg data=cardinals;
      model W = ERA Fld__B E_B;

proc reg data=cardinals;
      model W = ERA Fld__B SB;

proc reg data=cardinals;
      model W = ERA E_B SB;

proc reg data=cardinals;
      model W = ERA BB_P SB;

proc reg data=cardinals;
      model W = Fld__B BB_P E_B;

proc reg data=cardinals;
      model W = Fld__B E_B SB;

proc reg data=cardinals;
      model W = BB_P E_B SB;

proc reg data=cardinals;
```

```
      model W = ERA BB_P SB;
run;

/*interaction*/

data cardinalsInteraction;
      set cardinals;
      interact = cardinals*E_B;
run;

proc reg data=cardinalsInteraction;
      model W = ERA Fld__B E_B interact;
run;



/*quadratic*/
data cardinalsQuad;
      set cardinals;
      fld_sq = Fld__B*Fld__B;
run;

proc reg data=cardinalsQuad;
      model ERA Fld__B fld_sq;
run;
```

Moments			
N	24	Sum Weights	24
Mean	88.7916667	Sum Observations	2131
Std Deviation	7.84069189	Variance	61.4764493
Skewness	-0.000301	Kurtosis	-0.0404707
Uncorrected SS	190629	Corrected SS	1413.95833
Coeff Variation	8.83043666	Std Error Mean	1.60047453

Table 1 – basic descriptors of team wins (W)

Quantiles (Definition 5)	
Level	Quantile
100% Max	105
99%	105
95%	100
90%	100
75% Q3	94
50% Median	88
25% Q1	84
10%	78
5%	75
1%	73
0% Min	73

Table 2 – Quantile observations of team wins (W)

Moments			
N	24	Sum Weights	24
Mean	3.93375	Sum Observations	94.41
Std Deviation	0.44020561	Variance	0.19378098
Skewness	0.09869153	Kurtosis	-0.0773662
Uncorrected SS	375.8423	Corrected SS	4.4569625
Coeff Variation	11.1904826	Std Error Mean	0.08985659

Table 3 – Descriptors of earned run average as pitching team (ERA)

Quantiles (Definition 5)	
Level	Quantile
100% Max	4.740
99%	4.740
95%	4.650
90%	4.600
75% Q3	4.250
50% Median	3.865
25% Q1	3.680
10%	3.490
5%	3.420
1%	2.940
0% Min	2.940

Table 4 – Quartiles observations of earned run average as pitching team (ERA)

Moments			
N	24	Sum Weights	24
Mean	0.98304167	Sum Observations	23.593
Std Deviation	0.00302855	Variance	9.1721E-6
Skewness	-0.0535665	Kurtosis	-0.4482606
Uncorrected SS	23.193113	Corrected SS	0.00021096
Coeff Variation	0.30807929	Std Error Mean	0.0006182

Table 5 – Descriptors of fielding percentage as batting team (Fld%_B)

Quantiles (Definition 5)	
Level	Quantile
100% Max	0.9890
99%	0.9890
95%	0.9880
90%	0.9870
75% Q3	0.9850
50% Median	0.9835
25% Q1	0.9805
10%	0.9780
5%	0.9780
1%	0.9780
0% Min	0.9780

Table 6 – Quartile observations of fielding percentage as batting team (Fld&_B)

Moments			
N	24	Sum Weights	24
Mean	508.5	Sum Observations	12204
Std Deviation	58.0816517	Variance	3373.47826
Skewness	0.99770006	Kurtosis	0.96793645
Uncorrected SS	6283324	Corrected SS	77590
Coeff Variation	11.4221537	Std Error Mean	11.8558675

Table 7 – Descriptors of walks as pitching team (BB_P)

Quantiles (Definition 5)	
Level	Quantile
100% Max	667
99%	667
95%	606
90%	593
75% Q3	542
50% Median	500
25% Q1	465
10%	443
5%	440
1%	436
0% Min	436

Table 8 – Quartiles of walks as pitching team (BB_P)

Moments			
N	24	Sum Weights	24
Mean	104.208333	Sum Observations	2501
Std Deviation	19.1015118	Variance	364.867754
Skewness	0.04539112	Kurtosis	-0.3071366
Uncorrected SS	269017	Corrected SS	8391.95833
Coeff Variation	18.3301193	Std Error Mean	3.89907977

Table 9 – Descriptors of errors committed as batting team (E_B)

Quantiles (Definition 5)	
Level	Quantile
100% Max	142.0
99%	142.0
95%	133.0
90%	132.0
75% Q3	118.5
50% Median	101.5
25% Q1	95.0
10%	77.0
5%	75.0
1%	66.0
0% Min	66.0

Table 10 – Quartiles of errors committed as batting team (E_B)

Moments			
N	24	Sum Weights	24
Mean	86.5	Sum Observations	2076
Std Deviation	32.9267963	Variance	1084.17391
Skewness	0.83211417	Kurtosis	0.19779293
Uncorrected SS	204510	Corrected SS	24936
Coeff Variation	38.0656604	Std Error Mean	6.72115415

Table 11 – Descriptors of stolen bases as batting team (SB)

Quantiles (Definition 5)	
Level	Quantile
100% Max	164.0
99%	164.0
95%	149.0
90%	134.0
75% Q3	101.0
50% Median	81.5
25% Q1	61.0
10%	56.0
5%	45.0
1%	35.0
0% Min	35.0

Table 12 – Quartiles of stolen bases as batting team (SB)

Pearson Correlation Coefficients, N = 24 Prob > |r| under H0: Rho=0													
	R_B	_2B	HR	SB	BA	E_B	DP_B	Fld__B	BatAge	ERA	BB_P	SO_P	PAge
W	0.16976 0.4278	0.21036 0.3238	0.02709 0.9000	-0.27543 0.1927	0.17093 0.4245	-0.40931 0.0470	0.21474 0.3136	0.44714 0.0285	0.03420 0.8740	-0.63842 0.0008	-0.42318 0.0394	0.14471 0.4999	-0.00912 0.9662

Table 13 – Correlation coefficients against team wins (W)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	654.03183	218.01061	5.74	0.0053
Error	20	759.92650	37.99633		
Corrected Total	23	1413.95833			

Table 14 – ANOVA for model 1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	657.76564	219.25521	5.80	0.0051
Error	20	756.19270	37.80963		
Corrected Total	23	1413.95833			

Table 17 – ANOVA for model 2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	664.72383	221.57461	5.91	0.0046
Error	20	749.23450	37.46173		
Corrected Total	23	1413.95833			

Table 18 – ANOVA for model 3

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	655.97264	218.65755	5.77	0.0052
Error	20	757.98569	37.89928		
Corrected Total	23	1413.95833			

Table 19 – ANOVA for model 4

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	615.40447	205.13482	5.14	0.0085
Error	20	798.55386	39.92769		
Corrected Total	23	1413.95833			

Table 20 – ANOVA for model 5

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	348.19239	116.06413	2.18	0.1223
Error	20	1065.76594	53.28830		
Corrected Total	23	1413.95833			

Table 21 – ANOVA for model 6

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	359.16656	119.72219	2.27	0.1116
Error	20	1054.79178	52.73959		
Corrected Total	23	1413.95833			

Table 22 – ANOVA for model 7

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	332.18945	110.72982	2.05	0.1396
Error	20	1081.76889	54.08844		
Corrected Total	23	1413.95833			

Table 23 – ANOVA for model 8

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	615.40447	205.13482	5.14	0.0085
Error	20	798.55386	39.92769		
Corrected Total	23	1413.95833			

Table 24 – ANOVA for model 9

Root MSE	6.16412	R-Square	0.4626
Dependent Mean	88.79167	Adj R-Sq	0.3819
Coeff Var	6.94222		

Table 25 – model 1 descriptors

Root MSE	6.14895	R-Square	0.4652
Dependent Mean	88.79167	Adj R-Sq	0.3850
Coeff Var	6.92515		

Table 26 – Model 2 descriptors

Root MSE	6.12060	R-Square	0.4701
Dependent Mean	88.79167	Adj R-Sq	0.3906
Coeff Var	6.89321		

Table 27 – Model 3 descriptors

Root MSE	6.15624	R-Square	0.4639
Dependent Mean	88.79167	Adj R-Sq	0.3835
Coeff Var	6.93335		

Table 28 – Model 4 descriptors

Root MSE	6.31884	R-Square	0.4352
Dependent Mean	88.79167	Adj R-Sq	0.3505
Coeff Var	7.11647		

Table 29 – Model 5 descriptors

Root MSE	6.31884	R-Square	0.4352
Dependent Mean	88.79167	Adj R-Sq	0.3505
Coeff Var	7.11647		

Table 30 – Model 9 descriptors

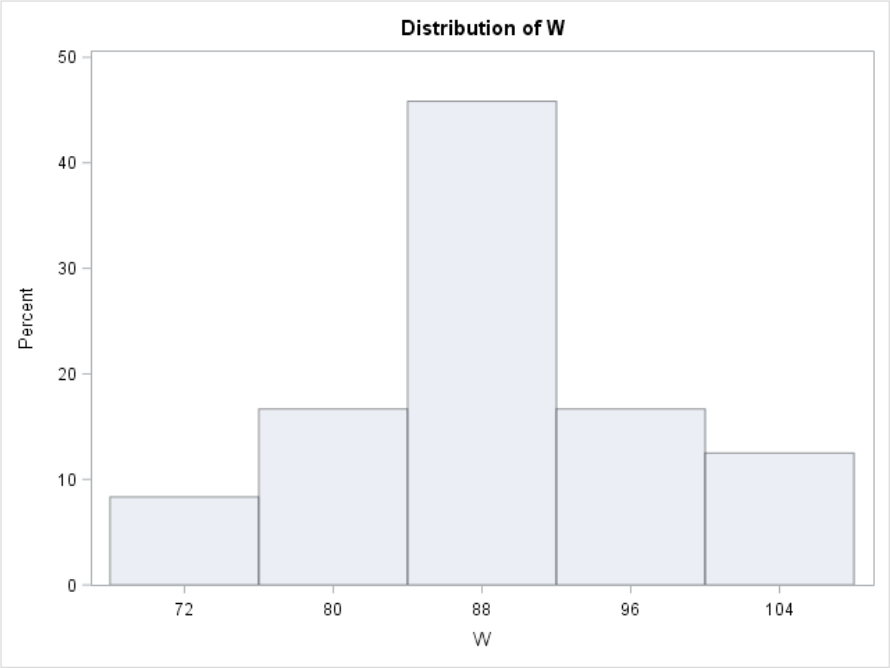

Figure 1 – Distribution of team wins (W)

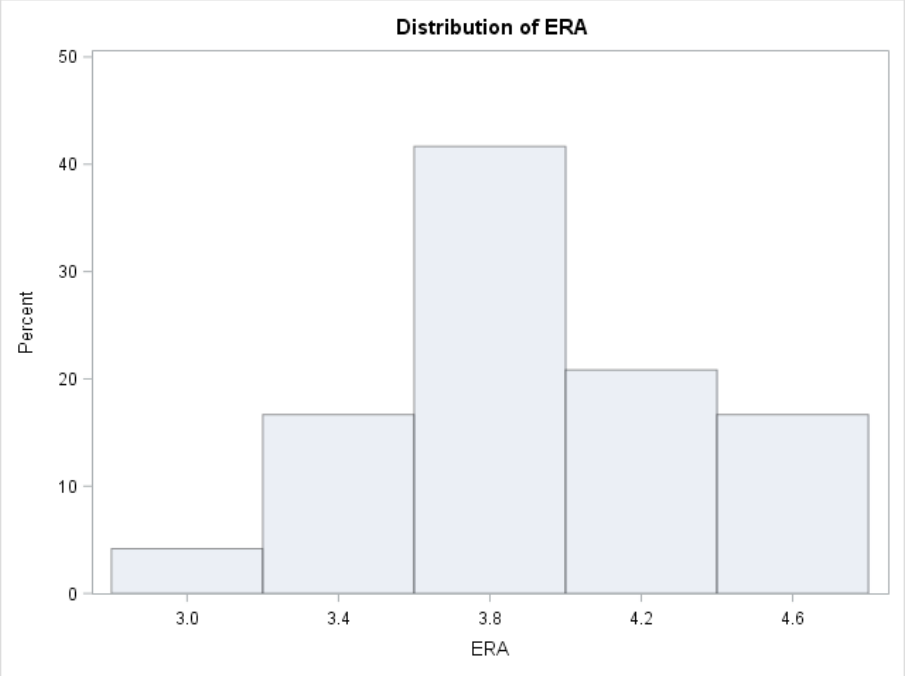

Figure 2 – Distribution of earned run average as pitching team (ERA)

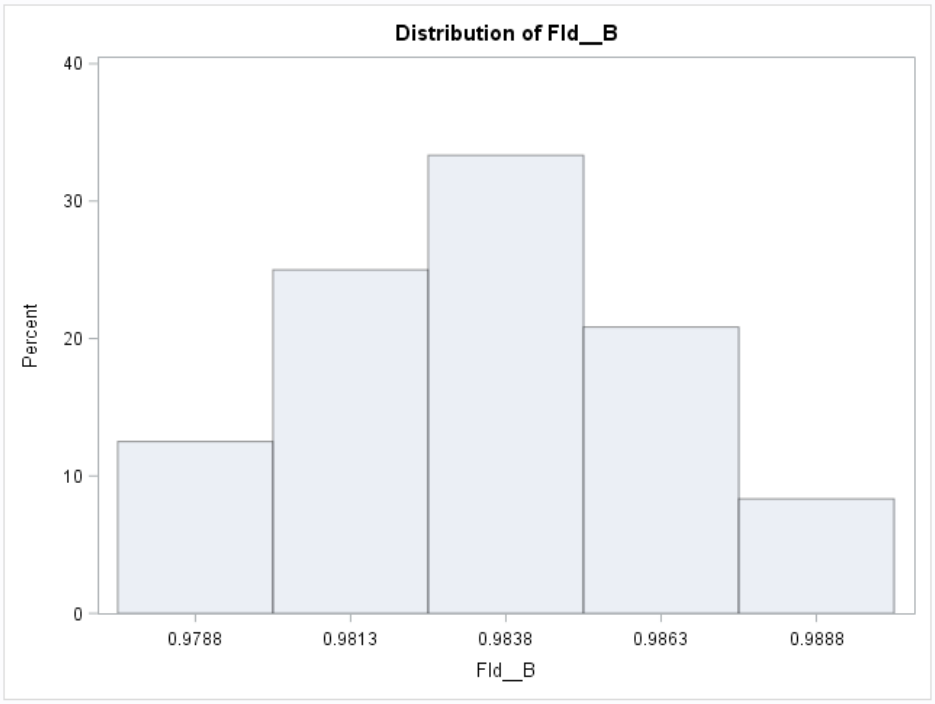

Figure 3 – Distribution of fielding percentage as batting team (Fld%_B)

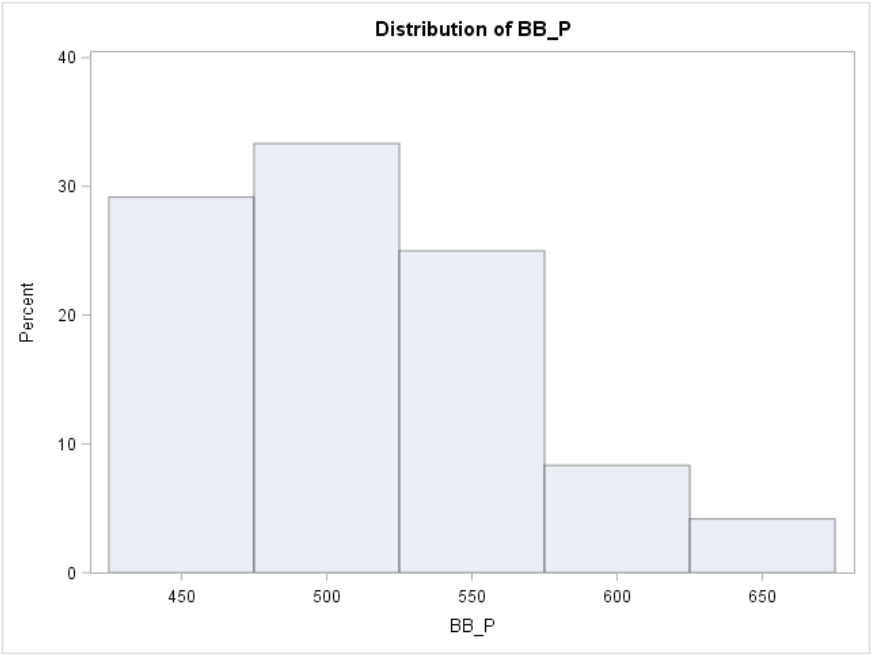

Figure 4 – Distribution of runs as pitching team (BB_P)

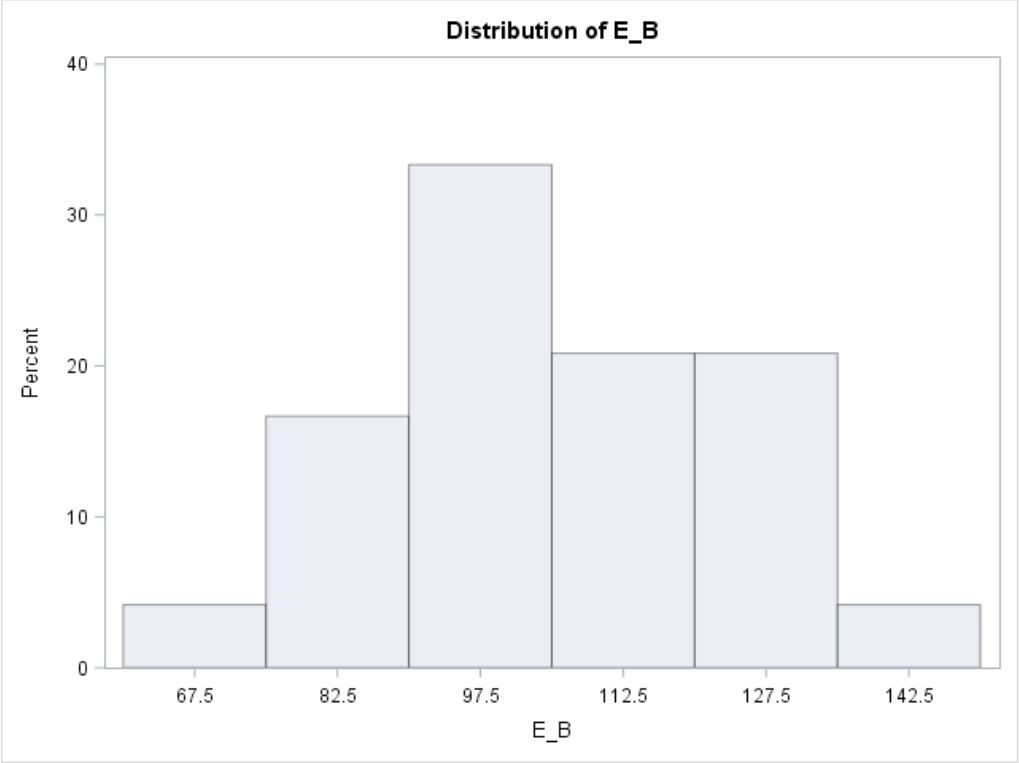

Figure 5 – Distribution of errors committed as batting team (E_B)

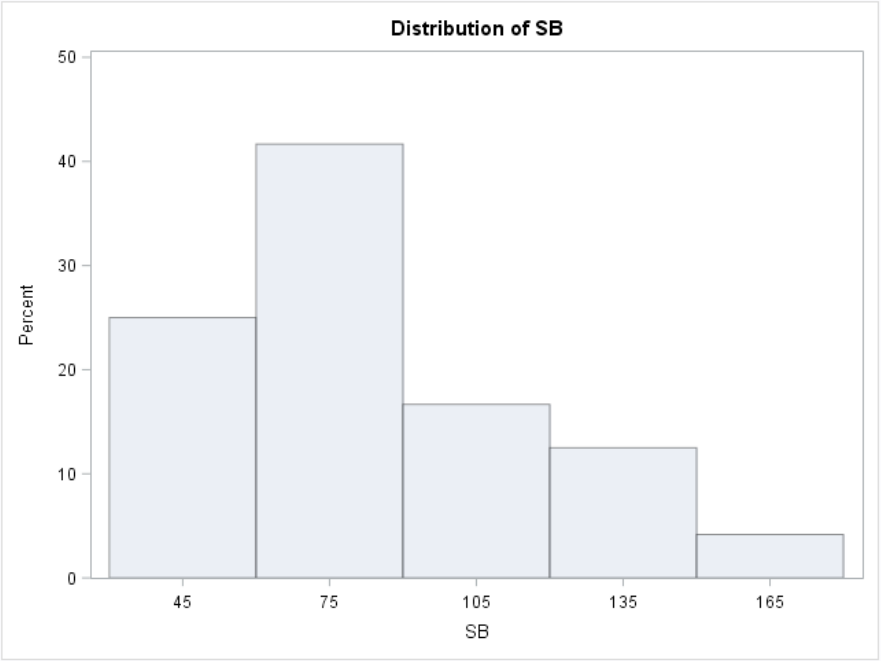

Figure 6 – Distribution of stolen bases as batting team (SB)

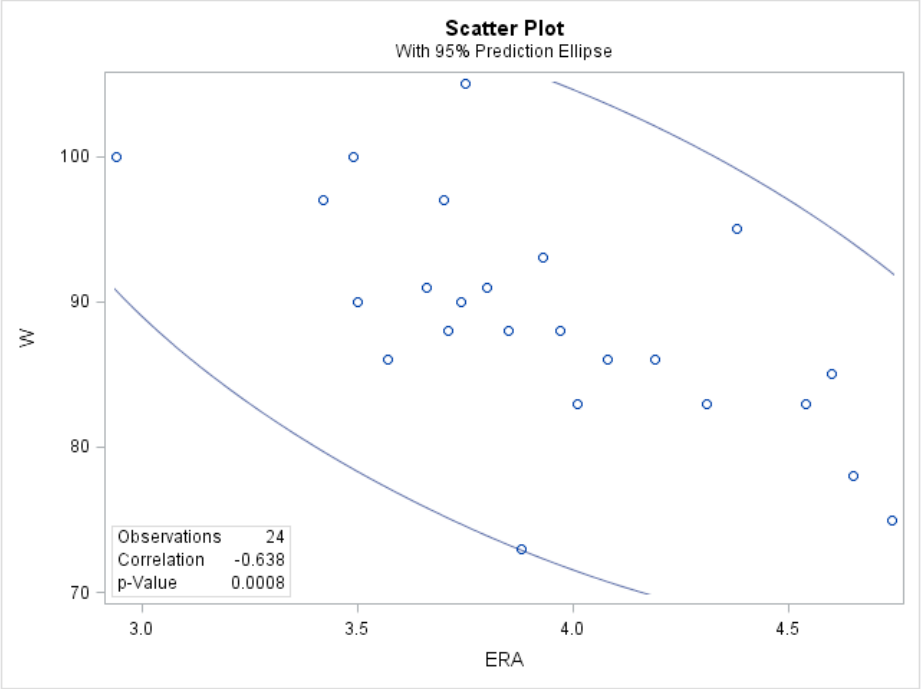

Figure 7 – Scatter plot between earned run average (ERA) and wins (W)

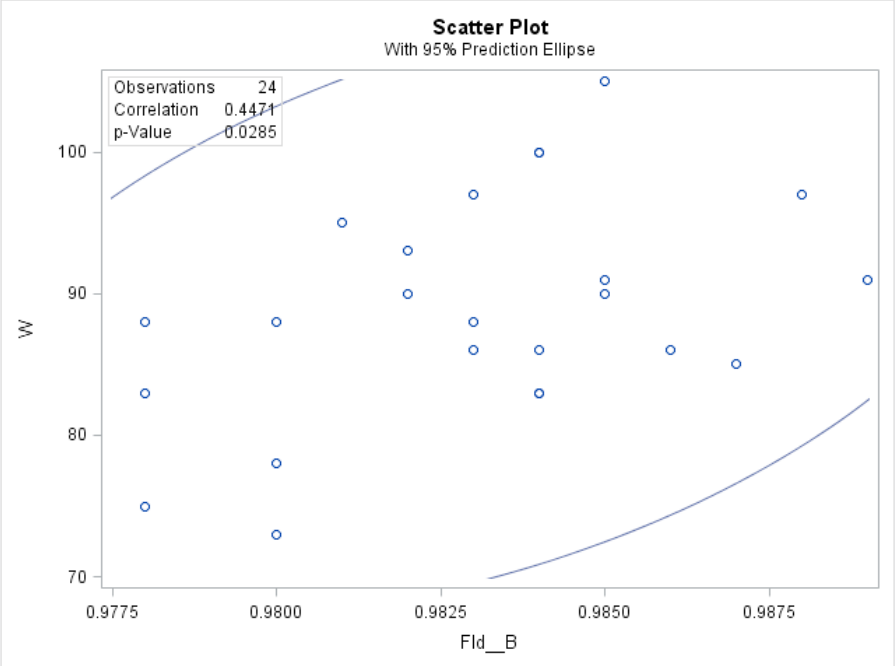

Figure 8 – Scatter plot between fielding percent (Fld%_B) and wins (W)

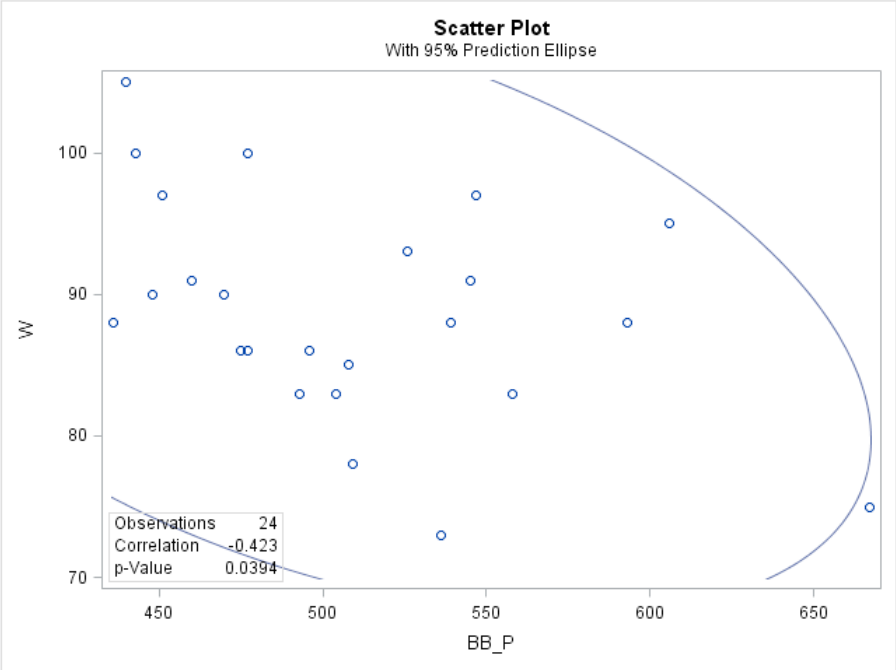

Figure 9 – Scatter plot between runs as pitching team (BB_P) and wins (W)

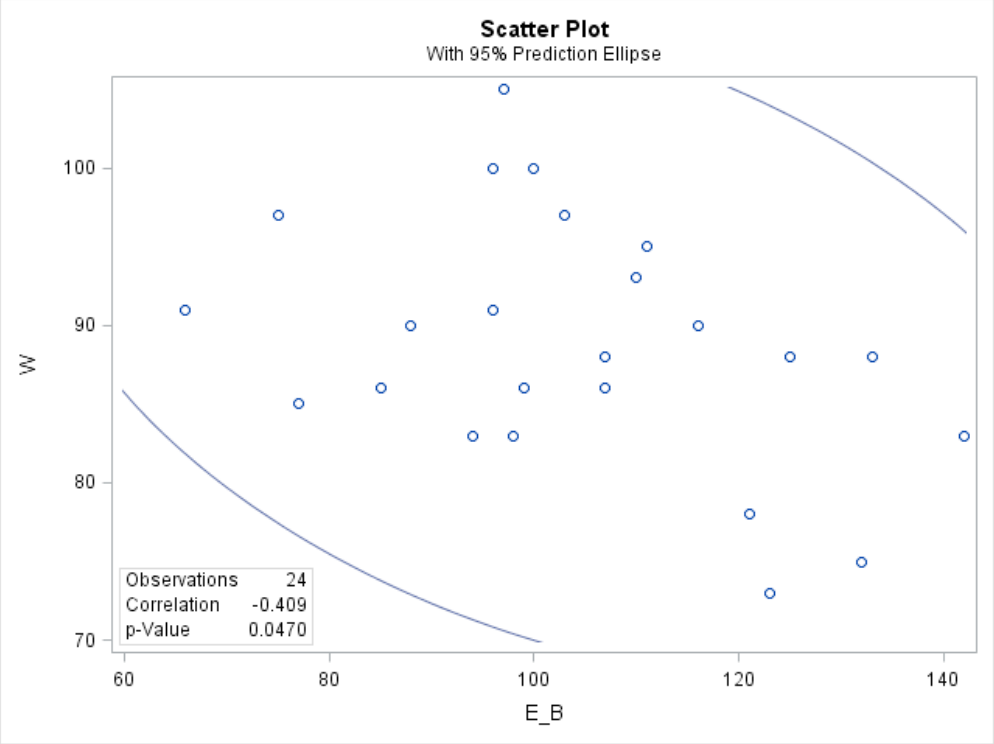

Figure 10 – Scatter plot between errors as batting team (E_B) and wins (W)

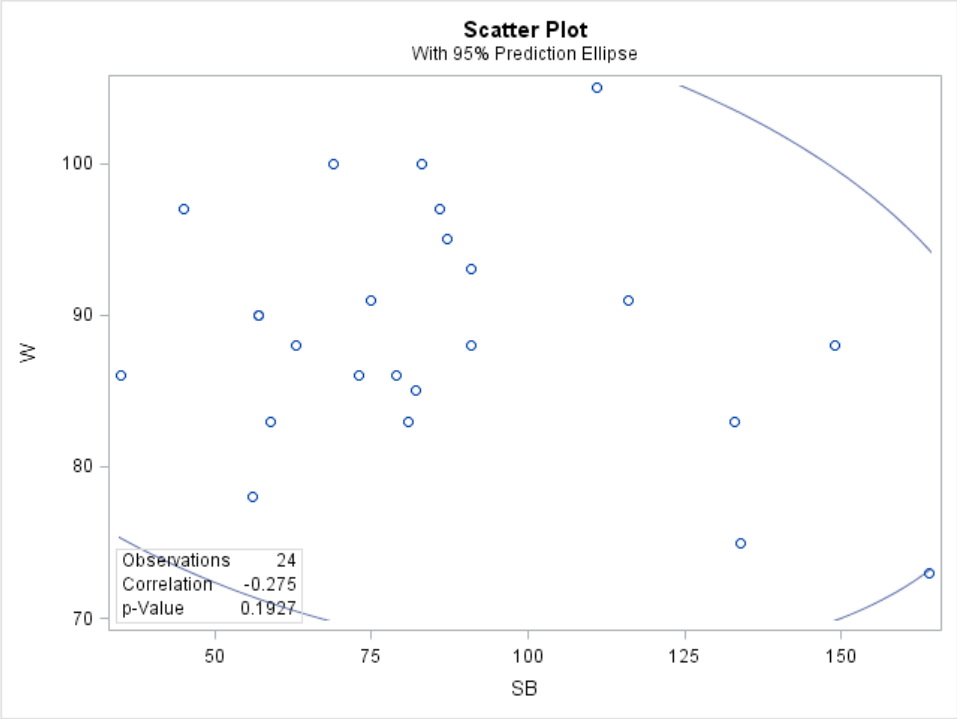

Figure 11 – Scatter plot between stolen bases (SB) and wins (W)

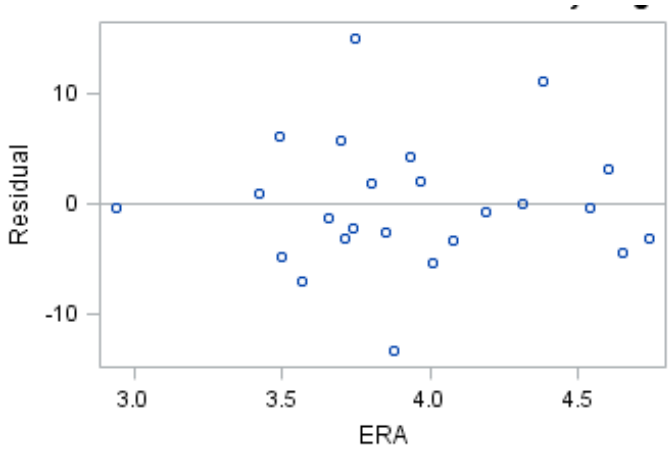

Figure 12 – Residual regressors between ERA and W in model 2

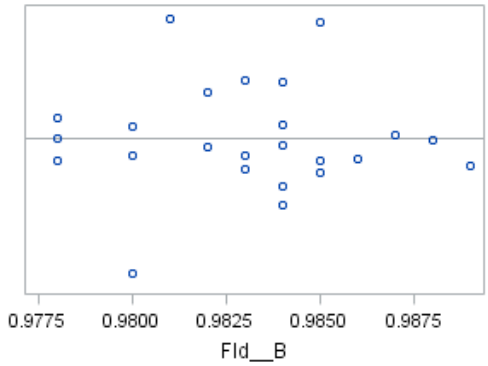

Figure 13 – Residual regressors between Fld%B and W in model 2

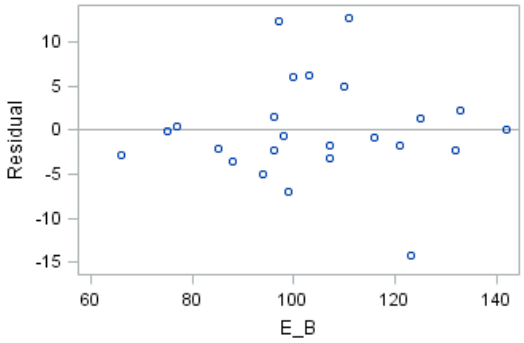

Figure 14 – Residual regressors between E_B and W in model 2