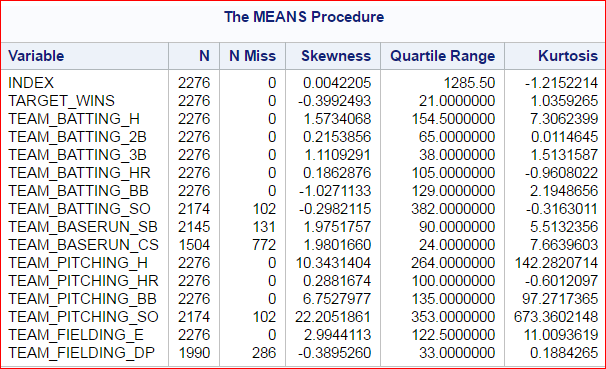
**DATA EXPLORATION**



|  |  |  |  |
| --- | --- | --- | --- |
| Variables As-is to analyse | | Variables transformed if necessary and further analyzed. | |
|  | Normality: looks ok. There is a little skewness to right.  Linearity: Looks ok.  Outliers: There are a few outliers, but aren’t high influences. We can use them as is. |  |  |
|  | Normality: looks good.  Linearity: Looks good.  Outliers: There are a few outliers, but aren’t high influences. We can use them as is. |  |  |
|  | Normality: looks ok. There is a little skewness to right. Skewness > 1.  Linearity: Looks ok.  Outliers: There are a few outliers, but aren’t high influences. We can use them as is.  We can consider log transformation on this variable. |  |  |
|  | Normality: looks ok. Skewness < 1.  Linearity: Looks ok.  Outliers: There are outliers. We need to look at fixing this. |  |  |
|  | Normality: looks ok. Skewness > 1.  Linearity: Looks ok. There is curve or bend.  Outliers: None high impact outliers.  We can consider log transformation on this variable. |  |  |
|  | Normality: looks ok. Skewness > 1.  Linearity: Looks ok. There is curve or bend.  Outliers: None high impact outliers.  There are missing values.  We can consider log transformation on this variable. |  |  |
|  | Normality: No, heavily right skewed. Skewness > 1.  Linearity: Looks ok. There is curve or bend.  Outliers: None high impact outliers. There are missing values.  We can consider log transformation on this variable. |  | Normality: No, heavily right skewed. We should do a log transformation.  There are 2 0 values. We can do value+1 log transformation. |
|  | Normality: No, heavily right skewed. Skewness > 1.  Linearity: Looks ok. There is curve or bend.  Outliers: None high impact outliers. There are missing values.  We can consider log transformation on this variable. | Normality looks good. We will use l\_team\_baserun\_CS |  |
|  | Normality: No, heavily right skewed.  Linearity: Does not look good. |  |  |
|  | Normality: No, a little right skewed.  Linearity: Looks good. | Left skewed data after transformation. Data needs to be looked into for missing, outliers. |  |
|  | Normality: ok.  Linearity: is ok.  Outliers: a few.  Log tranformation, proceed with caution. |  |  |
|  | Normality: ok.  Linearity: Does not look good.  Skewness > 1.  Log tranformation, proceed with caution. |  |  |
|  | Normality: heavy right skew. Skewness > 1.  Linearity: Does not look good.  We can consider log transformation. |  |  |
|  | Normality: Looks good.  Linearity: Looks good. |  |  |

3. Build Models

Approach1:

|  |  |
| --- | --- |
| Use all variables as is, eliminate high p value variables manually. |  |
| Model 2: glmselect selection=forward cvmethod=random  Data cleaning:   1. Update missing values with median.   The following columns were updated.  TEAM\_BATTING\_H = 1454  TEAM\_BATTING\_2B = 238  TEAM\_BATTING\_3B = 47  TEAM\_BATTING\_HR = 102  TEAM\_BATTING\_BB = 512  TEAM\_BATTING\_SO = 750  TEAM\_BASERUN\_SB = 101  TEAM\_BASERUN\_CS = 49  TEAM\_FIELDING\_E = 159  TEAM\_FIELDING\_DP = 149  TEAM\_PITCHING\_BB = 536.5  TEAM\_PITCHING\_H = 1518  TEAM\_PITCHING\_HR = 107  TEAM\_PITCHING\_SO = 813.5   1. Delete outliers.   Delete target\_wins in(0,12)  TEAM\_PITCHING\_H= 30132  Note: Removed column TEAM\_BATTING\_HBP as I had trouble formatting the column to numeric. |  |
| Model 3:    Data cleaning:   1. Update missing values with median.   The following columns were updated.  TEAM\_BATTING\_H = 1454  TEAM\_BATTING\_2B = 238  TEAM\_BATTING\_3B = 47  TEAM\_BATTING\_HR = 102  TEAM\_BATTING\_BB = 512  TEAM\_BATTING\_SO = 750  TEAM\_BASERUN\_SB = 101  TEAM\_BASERUN\_CS = 49  TEAM\_FIELDING\_E = 159  TEAM\_FIELDING\_DP = 149  TEAM\_PITCHING\_BB = 536.5  TEAM\_PITCHING\_H = 1518  TEAM\_PITCHING\_HR = 107  TEAM\_PITCHING\_SO = 813.5   1. Delete outliers.   Delete target\_wins in(0,12)  TEAM\_PITCHING\_H= 30132   1. Log transformed team\_wins.   Note: Removed column TEAM\_BATTING\_HBP as I had trouble formatting the column to numeric. |  |
| Model 4: selection=stepwise cvmethod=random  Data cleaning:   1. Update missing values with median.   The following columns were updated.  TEAM\_BATTING\_H = 1454  TEAM\_BATTING\_2B = 238  TEAM\_BATTING\_3B = 47  TEAM\_BATTING\_HR = 102  TEAM\_BATTING\_BB = 512  TEAM\_BATTING\_SO = 750  TEAM\_BASERUN\_SB = 101  TEAM\_BASERUN\_CS = 49  TEAM\_FIELDING\_E = 159  TEAM\_FIELDING\_DP = 149  TEAM\_PITCHING\_BB = 536.5  TEAM\_PITCHING\_H = 1518  TEAM\_PITCHING\_HR = 107  TEAM\_PITCHING\_SO = 813.5   1. Delete outliers.   Delete target\_wins in(0,12)  TEAM\_PITCHING\_H= 30132  Note: Removed column TEAM\_BATTING\_HBP as I had trouble formatting the column to numeric. |  |

Interpretations

Model considered 4, Adj R2 = 0.5358, Root MSE = 0.106

Log(Y) = β0 + β1 \* x1 + β2 \* x2 + β3 \* x3 + β4 \* x4+ β5 \* x5 + β6 \* x6 + β7 \* x7

Or

Log(TARGET\_WINS) = β0 + β 1\*TEAM\_BATTING\_H+ β2 \* (TEAM\_BATTING\_HR)2 + β3 \* TEAM\_BATTING\_SO +

β4\* TEAM\_PITCHING\_HR + β5\* log(TEAM\_PITCHING\_BB) + β6 \* log(TEAM\_FIELDING\_E) + β7 \* TEAM\_FIELDING\_DP

Where:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | β0= | 2.44 |
| x1= | Team\_BATTING\_H | β1= | 0.0003 |
| x2= | Team\_BATTING\_HR\_2 | β2= | -0.00002 |
| x3= | TEAM\_BATTING\_SO | β3= | -0.0004 |
| x4= | TEAM\_PITCHING\_HR | β4= | 0.0076 |
| x5= | L\_TEAMPITCHING\_BB | β5= | 0.3759 |
| x6= | L\_TEAM\_FIELDING\_E | β6= | -0.2357 |
| x7= | TEAM\_FIELDING\_DP | β7= | -0.0015 |

= 2.44 + 0.0003 \*TEAM\_BATTING\_H – 0.000017\* (TEAM\_BATTING\_HR)2 – 0.00037 \* TEAM\_BATTING\_SO +0.0076\* TEAM\_PITCHING\_HR +0.3759\* log(TEAM\_PITCHING\_BB) – 0.2356 \* log(TEAM\_FIELDING\_E) – 0.001\* TEAM\_FIELDING\_DP

Target\_wins = e 2.44 \* e 0.0003 \*TEAM\_BATTING\_H \* e – 0.000017\* (TEAM\_BATTING\_HR)2 \* e – 0.00037 \* TEAM\_BATTING\_SO \* e 0.0076\* TEAM\_PITCHING\_HR \* e 0.3759\* log(TEAM\_PITCHING\_BB) \* e – 0.2356 \* log(TEAM\_FIELDING\_E) \* e – 0.001\* TEAM\_FIELDING\_DP

**Target\_wins = 11.47 \* e (0.0003 \*TEAM\_BATTING\_H) \* e (– 0.000017\* (TEAM\_BATTING\_HR)2) \* e (- 0.00037 \* TEAM\_BATTING\_SO) \* e 0.0076\* TEAM\_PITCHING\_HR \* (TEAM\_PITCHING\_BB) 0.3759 \* (TEAM\_FIELDING\_E) -0.2356 \* e – 0.001\* TEAM\_FIELDING\_DP**

The predicted median Target wins when all variables are held constant in 11.47.