Loyola University Maryland
Department of Mathematics and Statistics
ST 765.W01 Project Results – Fall 2022

Student Information

- **Name:** Mike Ogrysko
- **Loyola Email:** mcogrysko@loyola.edu

Project Information

- **Title:** Predictors that Contribute to a Baseball Team's Win Percentage
- **Main Objective or Research Question:**

  Which metrics contribute to the Baltimore Orioles win percentage?
- **Data Source (URL or file)**:

  I compiled the data from this page: https://www.baseball-reference.com/teams/BAL/. I have uploaded the CSV file here: Orioles_Data_1984_2022_MOgrysko.csv.

Modeling Setup

- **Type of Linear Model (*circle one*):**
  - *Multiple Regression*
  - *Logistic Regression*
  - *Analysis of Variance*
  - *Analysis of Covariance*
- **Size of Data Set:**
  - *Number of Observations (or Cases)*: 39
  - *Number of Variables (or Predictors)*: 18
- **Response Variable Description (with units) and Name:**

  W_L_Percentage – This is a percentage calculated by dividing the number of wins by games played.

- **Descriptions, Names, and Types of Predictor Variables:**

Description (with units)	Name	Type
Predictor #1 # of games	G	Integer
Predictor #2 runs scored	R_Scored	Integer
Predictor #3 average batter age	BatAge	Float
Predictor #4 average pitcher age	PAge	Float
Predictor #5 # batters used	NumBat	Integer
Predictor #6 # pitchers used	NumP	Integer
Predictor #7 team hits	Hits_B	Integer
Predictor #8 team home runs	HR_B	Integer
Predictor #9 team stolen bases	SB_B	Integer
Predictor #10 team walks	BB_B	Integer
Predictor #11 team strike outs	SO_B	Integer
Predictor #12 team batting average	BA_B	Float
Predictor #13 team errors	E	Integer
Predictor #14 team earned run average	ERA	Float
Predictor #15 team home runs allowed	HR_P	Integer
Predictor #16 team walks allowed	BB_P	Integer
Predictor #17 team strike outs	SO_P	Integer
Predictor #18 team walks + hits / innings pitched	WHIP	Float

Initial Linear Model:

Naively, start EDA with all records and all predictors. After EDA, remove 2020, 1995, 1994 shortened seasons (based on G). Also, attempt to eliminate multicollinearity by removing Hits_B, WHIP, NumP, and SO_P.

```
#remove 2020, 1995, 1994 shortened seasons
orioles_rev <- orioles[-c(3, 28, 29), ]

#initial lm model
orioles.lm <- lm(W_L_Percentage ~
G+R_Scored+BatAge+PAge+NumBat+HR_B+SB_B+BB_B+SO_B+BA_B+E+ERA+HR_P+BB_P, data=orioles_rev)

summary(orioles.lm)
Call:
lm(formula = W_L_Percentage ~ G + R_Scored + BatAge + PAge +
    NumBat + HR_B + SB_B + BB_B + SO_B + BA_B + E + ERA + HR_P +
    BB_P, data = orioles_rev)

Residuals:
      Min        1Q    Median        3Q       Max
-0.043621 -0.013231  0.000276  0.013121  0.043937

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.505e+00  2.354e+00  -0.640    0.529
G            9.681e-03  1.470e-02   0.659    0.517
R_Scored     2.732e-04  2.471e-04   1.106    0.281
BatAge      -8.619e-03  7.571e-03  -1.138    0.268
PAge         5.960e-03  8.246e-03   0.723    0.478
NumBat       5.481e-04  2.019e-03   0.271    0.789
HR_B         5.417e-04  4.137e-04   1.309    0.205
SB_B         3.126e-04  3.611e-04   0.866    0.396
BB_B         1.370e-04  1.249e-04   1.097    0.285
SO_B         5.781e-05  7.997e-05   0.723    0.478
BA_B         2.157e+00  1.349e+00   1.599    0.125
E           -1.804e-05  3.577e-04  -0.050    0.960
ERA         -1.069e-01  2.135e-02  -5.007  5.9e-05 ***
HR_P        -2.323e-04  2.943e-04  -0.789    0.439
BB_P        -3.228e-05  1.678e-04  -0.192    0.849
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02655 on 21 degrees of freedom
Multiple R-squared:  0.9279,  Adjusted R-squared:  0.8798
F-statistic: 19.31 on 14 and 21 DF,  p-value: 6.766e-09
```

Initial model equation after EDA:

Y = -1.505333e+00 + 9.681101e-03(G) + 2.732327e-04(R_Scored) - 8.618529e-03(BatAge) + 5.959730e-03(PAge) + 5.480562e-04(NumBat) + 5.416946e-04(HR_B) + 3.126486e-04(SB_B) + 1.369688e-04(BB_B) + 5.781457e-05(SO_B) + 2.156984e+00(BA_B) - 1.803625e-05(E) - 1.068803e-01(ERA) - 2.322909e-04(HR_P) - 3.228290e-05(BB_P) + ϵ

Modeling Results

Exploratory Data Analysis (*include* R output):

- *Scatterplot Matrix (splom) of all Variables*

```
#splom all seasons
dev.new(height=800, width=800)
splom(orioles, main="SPLOM Orioles Data 1984-2022", pch=19, cex=.3, xlab=NULL, axis.text.cex = 0.1,
varname.cex = 0.3, axis.line.tck = .3)
```

First, the seasons with fewer games (1994, 1995, and 2020) should be removed from the dataset.

Next, based on sight alone, the following predictors appear to be positively correlated with each other:

- ERA & WHIP
- NumBat & NumP
- R_Scored & Hits_B
- Hits_B & BB_P
- SO_B & SO_P

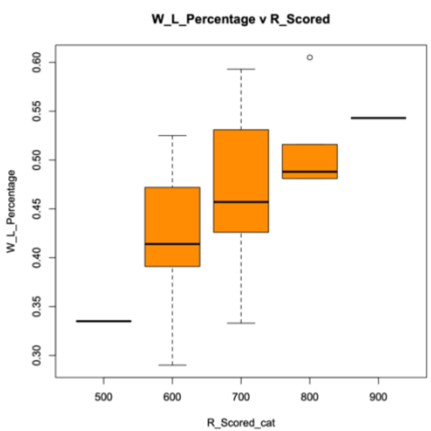
Looking at the correlation, these are confirmed. In addition, G and Hits_B are highly correlated. Some of the correlations make sense when thinking about how baseball is played – Hits and Runs Scored seem like they should be correlated as more hits tend to lead to more runs scored. However, the high correlation between batter strikeouts and pitcher strikeouts is questionable.

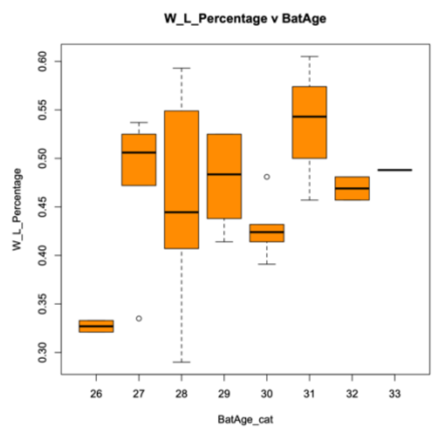
```
#correlation
cor(orioles[,c("G","W_L_Percentage","R_Scored","BatAge","PAge","NumBat","NumP","Hits_B","HR_B","SB_B","BB_B","SO_B","BA_B","E","ERA","HR_P","BB_P","SO_P
","WHIP")])

                          G W_L_Percentage    R_Scored      BatAge        PAge      NumBat         NumP     Hits_B        HR_B
G               1.000000000   -0.006640613  0.72241484  0.20647417 -0.09722911  0.19241375  0.001639119 0.89122729  0.48644814
W_L_Percentage -0.006640613    1.000000000  0.33035947  0.23458121  0.31777226 -0.47002511 -0.509326014 0.12346616  0.24586088
R_Scored        0.722414839    0.330359472  1.00000000  0.59274144  0.31891254 -0.01589833 -0.140354917 0.89366888  0.64265894
BatAge          0.206474170    0.234581214  0.59274144  1.00000000  0.69638742 -0.35543871 -0.454213162 0.50086909  0.22711107
PAge           -0.097229108    0.317772258  0.31891254  0.69638742  1.00000000 -0.28587254 -0.248899751 0.15694136  0.19685731
NumBat          0.192413754   -0.470025106 -0.01589833 -0.35543871 -0.28587254  1.00000000  0.920005246 0.02818062  0.28329672
NumP            0.001639119   -0.509326014 -0.14035492 -0.45421316 -0.24889975  0.92000525  1.000000000 -0.13549207 0.18712698
Hits_B          0.891227294    0.123466160  0.89366888  0.50086909  0.15694136  0.02818062 -0.135492071 1.00000000  0.50775738
HR_B            0.486448145    0.245860877  0.64265894  0.22711107  0.19685731  0.28329672  0.187126978 0.50775738  1.00000000
SB_B            0.320963900   -0.170643118  0.33282496  0.37013347  0.01316094  0.06727939 -0.049514804 0.38407915 -0.28422388
BB_B            0.554808452    0.306063303  0.65644726  0.44311251  0.16648142 -0.33297225 -0.413835963 0.56958163  0.14618918
SO_B            0.485700265   -0.158581681  0.15746675 -0.42872672 -0.34947331  0.72577048  0.649871076 0.24168851  0.56324174
BA_B           -0.065278950    0.331461672  0.50677572  0.74025395  0.59631470 -0.40507233 -0.379586746 0.38703411  0.08701063
E               0.561779810   -0.188108182  0.33978594  0.24569210 -0.10211626 -0.12410887 -0.243479256 0.48601008  0.09840792
ERA             0.068449745   -0.727171193  0.15215682  0.14237183  0.03318879  0.47671495  0.549961152 0.16695409  0.11607741
HR_P            0.487018786   -0.501649514  0.35588488 -0.02573286 -0.11347964  0.62663207  0.590490330 0.41887067  0.52979691
BB_P            0.734232206   -0.175454720  0.75858465  0.50861167  0.09555772  0.17048004  0.046385306 0.83694436  0.35780088
SO_P            0.537340188   -0.013811856  0.45080913  0.01847610  0.04983136  0.68038448  0.556394135 0.49009964  0.70157899
WHIP            0.260195346   -0.579981779  0.37164428  0.37233411  0.04877806  0.19966528  0.221459171 0.42393848  0.05451284
                      SB_B        BB_B        SO_B        BA_B           E         ERA        HR_P        BB_P        SO_P        WHIP
G               0.32096390  0.55480845  0.48570027 -0.06527895  0.56177981  0.06844975  0.48701879  0.73423221  0.53734019  0.26019535
W_L_Percentage -0.17064312  0.30606330 -0.15858168  0.33146167 -0.18810818 -0.72717119 -0.50164951 -0.17545472 -0.01381186 -0.57998178
R_Scored        0.33282496  0.65644726  0.15746675  0.50677572  0.33978594  0.15215682  0.35588488  0.75858465  0.45080913  0.37164428
BatAge          0.37013347  0.44311251 -0.42872672  0.74025395  0.24569210  0.14237183 -0.02573286  0.50861167  0.01847610  0.37233411
PAge            0.01316094  0.16648142 -0.34947331  0.59631470 -0.10211626  0.03318879 -0.11347964  0.09555772  0.04983136  0.04877806
NumBat          0.06727939 -0.33297225  0.72577048 -0.40507233 -0.12410887  0.47671495  0.62663207  0.17048004  0.68038448  0.19966528
NumP           -0.04951480 -0.41383596  0.64987108 -0.37958675 -0.24347926  0.54996115  0.59049033  0.04638531  0.55639414  0.22145917
Hits_B          0.38407915  0.56958163  0.24168851  0.38703411  0.48601008  0.16695409  0.41887067  0.83694436  0.49009964  0.42393848
HR_B           -0.28422388  0.14618918  0.56324174  0.08701063  0.09840792  0.11607741  0.52979691  0.35780088  0.70157899  0.05451284
SB_B            1.00000000  0.37102663 -0.21987315  0.23697990  0.22673108  0.27079504  0.12937815  0.49321570 -0.03501224  0.44751327
BB_B            0.37102663  1.00000000 -0.20123232  0.21446039  0.42561712 -0.13492639 -0.07088543  0.52111929 -0.13233990  0.15184463
SO_B           -0.21987315 -0.20123232  1.00000000 -0.52950315  0.05199378  0.12411919  0.63030411  0.18486078  0.81339678 -0.05553968
BA_B            0.23697990  0.21446039 -0.52950315  1.00000000 -0.06863455  0.18392719 -0.13782819  0.35028715 -0.08184553  0.38477053
E               0.22673108  0.42561712  0.05199378 -0.06863455  1.00000000  0.08679955  0.25424183  0.49270703  0.01985552  0.31062585
ERA             0.27079504 -0.13492639  0.12411919  0.18392719  0.08679955  1.00000000  0.68233466  0.45377099  0.21466746  0.87176044
HR_P            0.12937815 -0.07088543  0.63030411 -0.13782819  0.25424183  0.68233466  1.00000000  0.46353081  0.64307500  0.52113658
BB_P            0.49321570  0.52111929  0.18486078  0.35028715  0.49270703  0.45377099  0.46353081  1.00000000  0.40789443  0.70075084
SO_P           -0.03501224 -0.13233990  0.81339678 -0.08184553  0.01985552  0.21466746  0.64307500  0.40789443  1.00000000  0.07275074
WHIP            0.44751327  0.15184463 -0.05553968  0.38477053  0.31062585  0.87176044  0.52113658  0.70075084  0.07275074  1.00000000
```

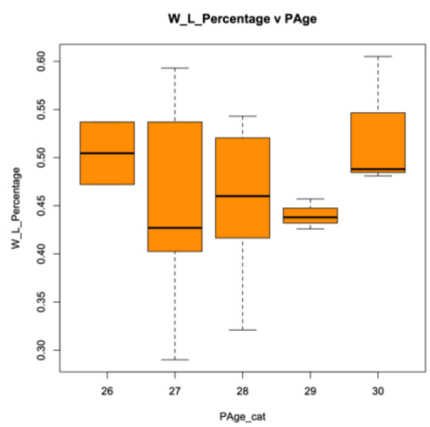
- *Multiple Boxplots of all Variables*

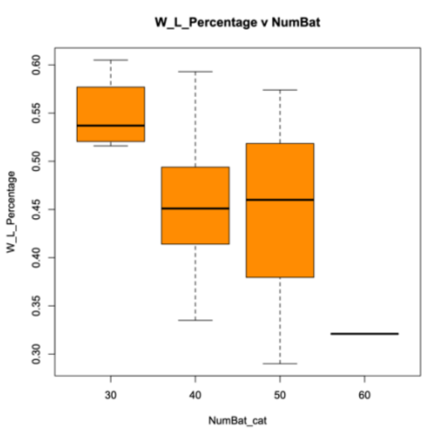
```
#factors for boxplots (seasons removed)
R_Scored_cat = factor(ifelse(orioles_rev$R_Scored < 600, "500", ifelse(orioles_rev$R_Scored < 700,
"600", ifelse(orioles_rev$R_Scored < 800, "700", ifelse(orioles_rev$R_Scored < 900, "800", "900")))))
BatAge_cat = factor(ifelse(orioles_rev$BatAge < 27, "26", ifelse(orioles_rev$BatAge < 28, "27",
ifelse(orioles_rev$BatAge < 29, "28", ifelse(orioles_rev$BatAge < 30, "29", ifelse(orioles_rev$BatAge <
31, "30", ifelse(orioles_rev$BatAge < 32, "31", ifelse(orioles_rev$BatAge < 33, "32", "33"))))))))
PAge_cat = factor(ifelse(orioles_rev$PAge < 27, "26", ifelse(orioles_rev$PAge < 28, "27",
ifelse(orioles_rev$PAge < 29, "28", ifelse(orioles_rev$PAge < 30, "29", "30")))))
NumBat_cat = factor(ifelse(orioles_rev$NumBat < 40, "30", ifelse(orioles_rev$NumBat < 50, "40",
ifelse(orioles_rev$NumBat < 60, "50", "60"))))
NumP_cat = factor(ifelse(orioles_rev$NumP < 20, "10", ifelse(orioles_rev$NumP < 30, "20",
ifelse(orioles_rev$NumP < 40, "30", "40"))))
Hits_B_cat = factor(ifelse(orioles_rev$Hits_B < 1300, "1200", ifelse(orioles_rev$Hits_B < 1400, "1300",
ifelse(orioles_rev$Hits_B < 1500, "1400", ifelse(orioles_rev$Hits_B < 1600, "1500", "1600")))))
HR_B_cat = factor(ifelse(orioles_rev$HR_B < 130, "120", ifelse(orioles_rev$HR_B < 140, "130",
ifelse(orioles_rev$HR_B < 150, "140", ifelse(orioles_rev$HR_B < 160, "150", ifelse(orioles_rev$HR_B <
170, "160", ifelse(orioles_rev$HR_B < 180, "170", ifelse(orioles_rev$HR_B < 190, "180",
ifelse(orioles_rev$HR_B < 200, "190", ifelse(orioles_rev$HR_B < 210, "200", ifelse(orioles_rev$HR_B <
220, "210", ifelse(orioles_rev$HR_B < 230, "220", ifelse(orioles_rev$HR_B < 240, "230",
ifelse(orioles_rev$HR_B < 250, "240", "250"))))))))))))))
SB_B_cat = factor(ifelse(orioles_rev$SB_B < 20, "10", ifelse(orioles_rev$SB_B < 30, "20",
ifelse(orioles_rev$SB_B < 40, "30", ifelse(orioles_rev$SB_B < 50, "40", ifelse(orioles_rev$SB_B < 60,
"50", ifelse(orioles_rev$SB_B < 70, "60", ifelse(orioles_rev$SB_B < 80, "70", ifelse(orioles_rev$SB_B <
90, "80", ifelse(orioles_rev$SB_B < 100, "90", ifelse(orioles_rev$SB_B < 110, "100",
ifelse(orioles_rev$SB_B < 120, "110", ifelse(orioles_rev$SB_B < 130, "120", ifelse(orioles_rev$SB_B <
140, "130", "140")))))))))))))), levels=c("10","20","30","40","50","60","70","80", "90", "100", "110",
"120", "130", "140"))
BB_B_cat = factor(ifelse(orioles_rev$BB_B < 400, "390", ifelse(orioles_rev$BB_B < 410, "400",
ifelse(orioles_rev$BB_B < 420, "410", ifelse(orioles_rev$BB_B < 430, "420", ifelse(orioles_rev$BB_B <
440, "430", ifelse(orioles_rev$BB_B < 450, "440", ifelse(orioles_rev$BB_B < 460, "450",
ifelse(orioles_rev$BB_B < 470, "460", ifelse(orioles_rev$BB_B < 480, "470", ifelse(orioles_rev$BB_B <
490, "480", ifelse(orioles_rev$BB_B < 500, "490", ifelse(orioles_rev$BB_B < 510, "500",
ifelse(orioles_rev$BB_B < 520, "510", ifelse(orioles_rev$BB_B < 530, "520", ifelse(orioles_rev$BB_B <
540, "530", ifelse(orioles_rev$BB_B < 550, "540", ifelse(orioles_rev$BB_B < 560, "550",
ifelse(orioles_rev$BB_B < 570, "560", ifelse(orioles_rev$BB_B < 580, "570", ifelse(orioles_rev$BB_B <
590, "580", ifelse(orioles_rev$BB_B < 600, "590", ifelse(orioles_rev$BB_B < 610, "600",
ifelse(orioles_rev$BB_B < 620, "610", ifelse(orioles_rev$BB_B < 630, "620", ifelse(orioles_rev$BB_B <
640, "630", ifelse(orioles_rev$BB_B < 650, "640", ifelse(orioles_rev$BB_B < 660, "650",
"660"))))))))))))))))))))))))))))
SO_B_cat = factor(ifelse(orioles_rev$SO_B < 900, "800", ifelse(orioles_rev$SO_B < 1000, "900",
ifelse(orioles_rev$SO_B < 1100, "1000", ifelse(orioles_rev$SO_B < 1200, "1100", ifelse(orioles_rev$SO_B
< 1300, "1200", ifelse(orioles_rev$SO_B < 1400, "1300", "1400")))))), levels=c("800","900","1000",
"1100", "1200", "1300", "1400"))
BA_B_cat = factor(ifelse(orioles_rev$BA_B < .240, ".230", ifelse(orioles_rev$BA_B < .250, ".240",
ifelse(orioles_rev$BA_B < .260, ".250", ifelse(orioles_rev$BA_B < .270, ".260", ifelse(orioles_rev$BA_B
< .280, ".270", ".280")))))
E_cat = factor(ifelse(orioles_rev$E < 60, "50", ifelse(orioles_rev$E < 70, "60", ifelse(orioles_rev$E <
80, "70", ifelse(orioles_rev$E < 90, "80", ifelse(orioles_rev$E < 100, "90", ifelse(orioles_rev$E < 110,
"100", ifelse(orioles_rev$E < 120, "110", ifelse(orioles_rev$E < 130, "120", "130")))))))),
levels=c("50","60","70", "80", "90", "100", "110", "120", "130"))
ERA_cat = factor(ifelse(orioles_rev$ERA < 4.0, "<4.00", ifelse(orioles_rev$ERA < 4.5, "4.00 - 4.50",
ifelse(orioles_rev$ERA < 5.0, "4.50 - 5.00", ifelse(orioles_rev$ERA < 5.5, "5.00 - 5.50", "5.50 -
6.00")))))
HR_P_cat = factor(ifelse(orioles_rev$HR_P < 130, "120", ifelse(orioles_rev$HR_P < 140, "130",
ifelse(orioles_rev$HR_P < 150, "140", ifelse(orioles_rev$HR_P < 160, "150", ifelse(orioles_rev$HR_P <
170, "160", ifelse(orioles_rev$HR_P < 180, "170", ifelse(orioles_rev$HR_P < 190, "180",
ifelse(orioles_rev$HR_P < 200, "190", ifelse(orioles_rev$HR_P < 210, "200", ifelse(orioles_rev$HR_P <
220, "210", ifelse(orioles_rev$HR_P < 230, "220", ifelse(orioles_rev$HR_P < 240, "230",
ifelse(orioles_rev$HR_P < 250, "240", ifelse(orioles_rev$HR_P < 260, "250", ifelse(orioles_rev$HR_P <
270, "260", ifelse(orioles_rev$HR_P < 280, "270", ifelse(orioles_rev$HR_P < 290, "280",
ifelse(orioles_rev$HR_P < 300, "290", "300"))))))))))))))))))
BB_P_cat = factor(ifelse(orioles_rev$BB_P < 450, "440", ifelse(orioles_rev$BB_P < 460, "450",
ifelse(orioles_rev$BB_P < 470, "460", ifelse(orioles_rev$BB_P < 480, "470", ifelse(orioles_rev$BB_P <
490, "480", ifelse(orioles_rev$BB_P < 500, "490", ifelse(orioles_rev$BB_P < 510, "500",
ifelse(orioles_rev$BB_P < 520, "510", ifelse(orioles_rev$BB_P < 530, "520", ifelse(orioles_rev$BB_P <
540, "530", ifelse(orioles_rev$BB_P < 550, "540", ifelse(orioles_rev$BB_P < 560, "550",
ifelse(orioles_rev$BB_P < 570, "560", ifelse(orioles_rev$BB_P < 580, "570", ifelse(orioles_rev$BB_P <
590, "580", ifelse(orioles_rev$BB_P < 600, "590", ifelse(orioles_rev$BB_P < 610, "600",
ifelse(orioles_rev$BB_P < 620, "610", ifelse(orioles_rev$BB_P < 630, "620", ifelse(orioles_rev$BB_P <
```

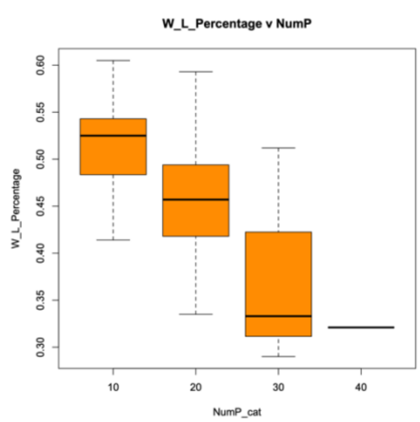
```
640, "630", ifelse(orioles_rev$BB_P < 650, "640",  ifelse(orioles_rev$BB_P < 660, "650",
ifelse(orioles_rev$BB_P < 670, "660",  ifelse(orioles_rev$BB_P < 680, "670", ifelse(orioles_rev$BB_P <
690, "680", "690"))))))))))))))))))))))))))
SO_P_cat = factor(ifelse(orioles_rev$SO_P < 700, "600", ifelse(orioles_rev$SO_P < 800, "700",
ifelse(orioles_rev$SO_P < 900, "800", ifelse(orioles_rev$SO_P < 1000, "900", ifelse(orioles_rev$SO_P <
1100, "1000", ifelse(orioles_rev$SO_P < 1200, "1100", "1200")))))), levels=c("600", "700", "800", "900",
"1000", "1100", "1200"))
WHIP_cat = factor(ifelse(orioles_rev$WHIP < 1.3, "1.20", ifelse(orioles_rev$WHIP < 1.4, "1.30",
ifelse(orioles_rev$WHIP < 1.5, "1.40", "1.50"))))

# boxplots
boxplot(W_L_Percentage ~ R_Scored_cat, data=orioles_rev, main="W_L_Percentage v R_Scored",
col='darkorange')
boxplot(W_L_Percentage ~ BatAge_cat, data=orioles_rev, main="W_L_Percentage v BatAge", col='darkorange')
boxplot(W_L_Percentage ~ PAge_cat, data=orioles_rev, main="W_L_Percentage v PAge", col='darkorange')
boxplot(W_L_Percentage ~ NumBat_cat, data=orioles_rev, main="W_L_Percentage v NumBat", col='darkorange')
boxplot(W_L_Percentage ~ NumP_cat, data=orioles_rev, main="W_L_Percentage v NumP", col='darkorange')
boxplot(W_L_Percentage ~ Hits_B_cat, data=orioles_rev, main="W_L_Percentage v Hits_B", col='darkorange')
boxplot(W_L_Percentage ~ HR_B_cat, data=orioles_rev, main="W_L_Percentage v HR_B", col='darkorange')
boxplot(W_L_Percentage ~ SB_B_cat, data=orioles_rev, main="W_L_Percentage v SB_B", col='darkorange')
boxplot(W_L_Percentage ~ BB_B_cat, data=orioles_rev, main="W_L_Percentage v BB_B", col='darkorange')
boxplot(W_L_Percentage ~ SO_B_cat, data=orioles_rev, main="W_L_Percentage v SO_B", col='darkorange')
boxplot(W_L_Percentage ~ BA_B_cat, data=orioles_rev, main="W_L_Percentage v BA_B", col='darkorange')
boxplot(W_L_Percentage ~ E_cat, data=orioles_rev, main="W_L_Percentage v E", col='darkorange')
boxplot(W_L_Percentage ~ ERA_cat, data=orioles_rev, main="W_L_Percentage v ERA", col='darkorange')
boxplot(W_L_Percentage ~ HR_P_cat, data=orioles_rev, main="W_L_Percentage v HR_P", col='darkorange')
boxplot(W_L_Percentage ~ BB_P_cat, data=orioles_rev, main="W_L_Percentage v BB_P", col='darkorange')
boxplot(W_L_Percentage ~ SO_P_cat, data=orioles_rev, main="W_L_Percentage v SO_P", col='darkorange')
boxplot(W_L_Percentage ~ WHIP_cat, data=orioles_rev, main="W_L_Percentage v WHIP", col='darkorange')
```

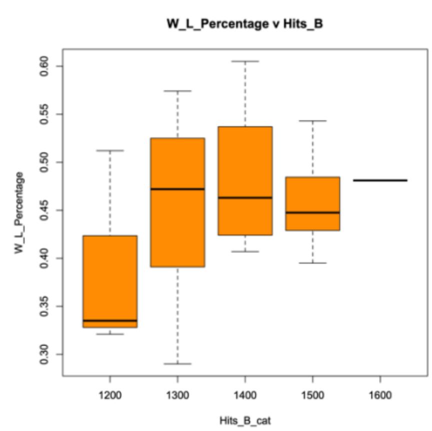


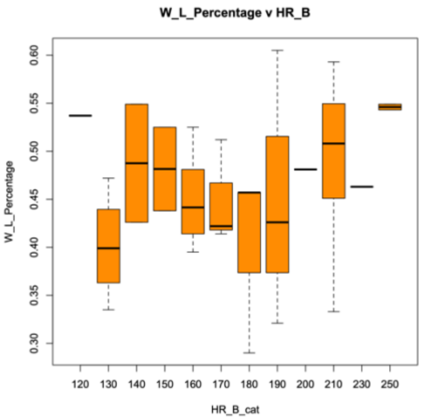


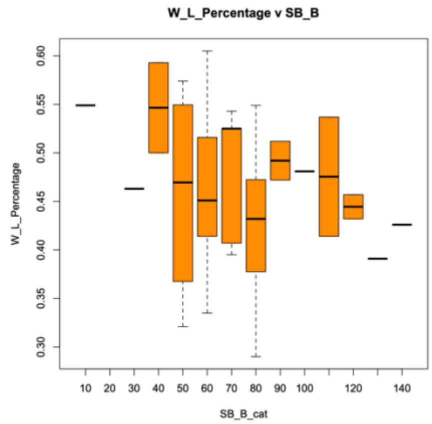


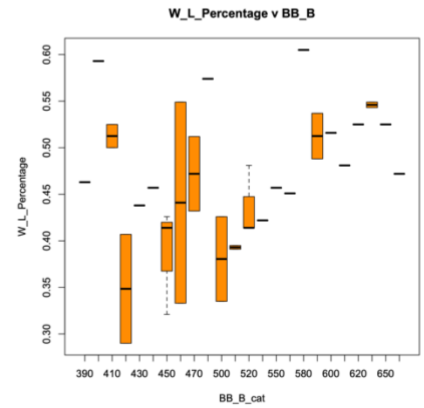


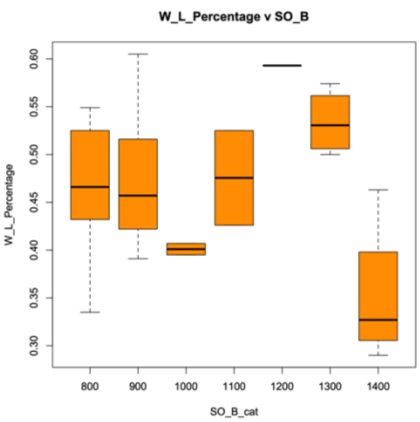


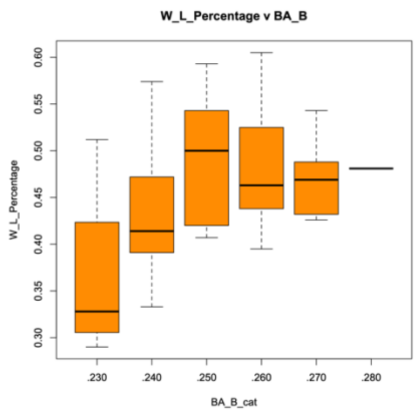


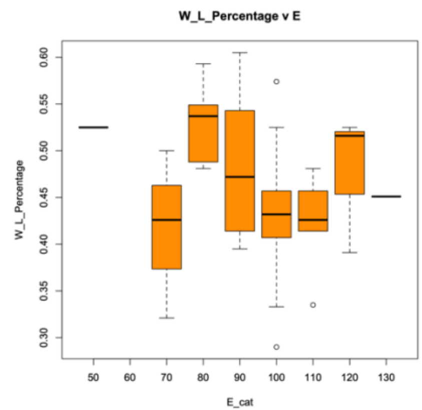

W_L_Percentage v HR_B
W_L_Percentage
HR_B_cat
120 130 140 150 160 170 180 190 200 210 230 250
0.30 0.35 0.40 0.45 0.50 0.55 0.60

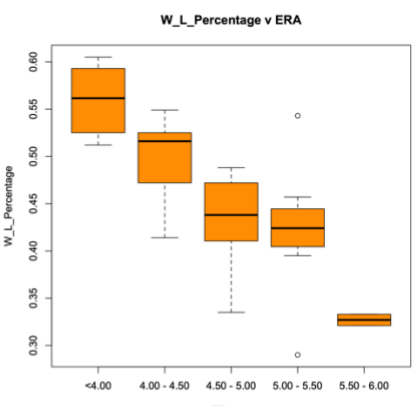

W_L_Percentage v SB_B
W_L_Percentage
SB_B_cat
10 20 30 40 50 60 70 80 90 100 120 140
0.30 0.35 0.40 0.45 0.50 0.55 0.60

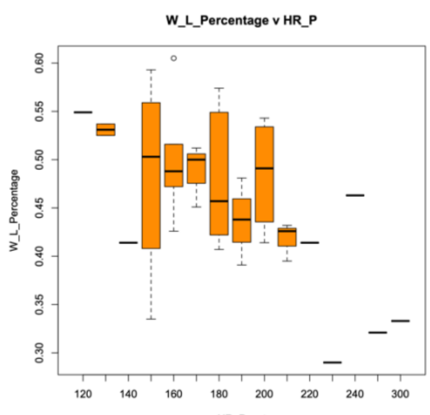

W_L_Percentage v BB_B
W_L_Percentage
BB_B_cat
390 410 430 450 470 500 520 550 580 600 620 650
0.30 0.35 0.40 0.45 0.50 0.55 0.60

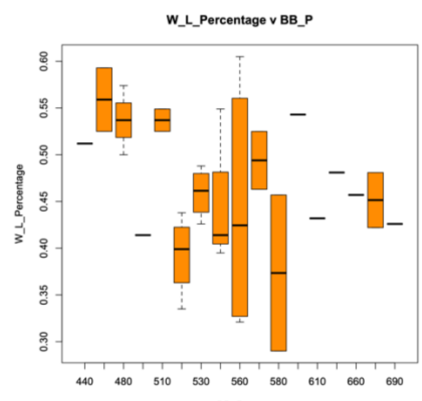

W_L_Percentage v SO_B
W_L_Percentage
SO_B_cat
800 900 1000 1100 1200 1300 1400
0.30 0.35 0.40 0.45 0.50 0.55 0.60

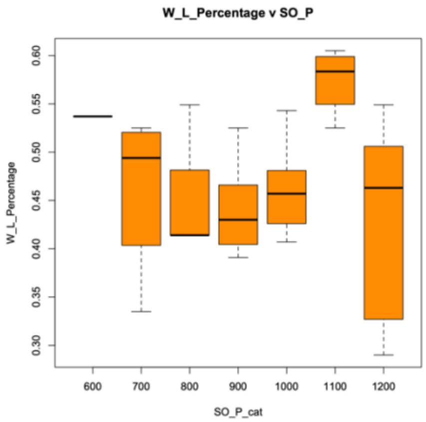

W_L_Percentage v BA_B
W_L_Percentage
BA_B_cat
.230 .240 .250 .260 .270 .280
0.30 0.35 0.40 0.45 0.50 0.55 0.60

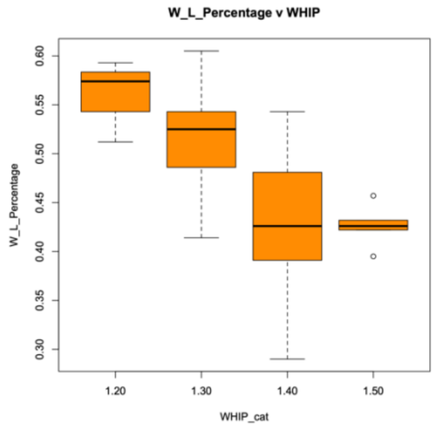

W_L_Percentage v E
W_L_Percentage
E_cat
50 60 70 80 90 100 110 120 130
0.30 0.35 0.40 0.45 0.50 0.55 0.60


W_L_Percentage v ERA
W_L_Percentage
ERA_cat
<4.00 4.00 - 4.50 4.50 - 5.00 5.00 - 5.50 5.50 - 6.00
0.30 0.35 0.40 0.45 0.50 0.55 0.60


W_L_Percentage v HR_P
W_L_Percentage
HR_P_cat
120 140 160 180 200 220 240 300
0.30 0.35 0.40 0.45 0.50 0.55 0.60


W_L_Percentage v BB_P
W_L_Percentage
BB_P_cat
440 480 510 530 560 580 610 660 690
0.30 0.35 0.40 0.45 0.50 0.55 0.60





Boxplot observations:

Predictor	Notes
G	• Excluded since most seasons should be 162 games
R_Score	• Winning teams score more runs • No teams that scored under 600 runs won more than they lost • R_Scored of 600-700 have a range of outcomes between ~.300 and ~.530 • R_Scored of 700-800 have a range of outcomes between ~.340 and ~.600 • R_Scored of 800-900 have a range of outcomes between ~.480 and ~.520 with an outlier above .600 • R_Scored above 900 has a single outcome of ~.540
BatAge	• Difficult to make a connection between BatAge and W_L_Percentage • Age 26 has lowest range of outcomes – all under ~.340 • Age 27 has range of outcomes between ~.470 and ~.540 – with a min of ~.340 • Age 28 has the largest range of outcomes between ~.300 and ~.600 • Age 29 has range of outcomes between ~.410 and ~.530 • Age 30 has range of outcomes between ~.390 and ~.440 – with max of ~.490 • Age 31 has range of outcomes between ~.450 and ~.600 • Age 32 has range of outcomes between ~.450 and ~.490 • Age 33 has a single outcome of ~.490
PAge	• Difficult to make a connection between PAge and W_L_Percentage • Age 26 has range of outcomes between ~.480 and ~.540 • Age 27 has range of outcomes between ~.290 and ~.600 • Age 28 has range of outcomes between ~.320 and ~.550 • Age 29 has range of outcomes between ~.430 and ~.460 • Age 30 has range of outcomes between ~.490 and ~.610
NumBat	• Winning teams tend to use fewer batters • 30-40 has range of outcomes between ~.520 and ~.610 • 40-50 has range of outcomes between ~.340 and ~.600 • 50-60 has range of outcomes between ~.300 and ~.580 • 60+ has a single outcome of ~.330

NumP	• Winning teams tend to use fewer pitchers • 10-20 has range of outcomes between ~.420 and ~.610 • 20-30 has range of outcomes between ~.340 and ~.600 • 30-40 has range of outcomes between ~.300 and ~.510 • 40+ has a single outcome of ~.330
Hits_B	• 1200-1300 has range of outcomes between ~.330 and ~.510 • 1300-1400 has range of outcomes between ~.300 and ~.580 • 1400-1500 has range of outcomes between ~.410 and ~.600 • 1500-1600 has range of outcomes between ~.400 and ~.550 • 1600+ has a single outcome of ~.490
HR_B	• Difficult to make a connection between HR_B and W_L_Percentage • 120-130 has a single outcome of ~.540 • 250+ has range of outcomes between ~.540 and ~.550 • 190-200 has the largest range of outcomes between ~.330 and ~.610
SB_B	• Difficult to make a connection between SB_B and W_L_Percentage • 10-20 has a single outcome of ~.550 • 50-60, 60-70, and 80-90 have the largest range of outcomes at ~.330 and ~.580, ~.340 and ~.610, and ~.300 and ~.550, respectively
BB_B	• Difficult to make a connection between BB_B and W_L_Percentage • Max is 580-590 and has a single outcome of ~.610 • 460-470 has largest range of outcomes between ~.330 and ~.550
SO_B	• Difficult to make a connection between SO_B and W_L_Percentage • 800-900 has range of outcomes between ~.340 and ~.550 • 900-1000 has range of outcomes between ~.390 and ~.610 • 1000-1100 has range of outcomes between ~.390 and ~.410 • 1100-1200 has range of outcomes between ~.430 and ~.530 • 1200-1300 has a single outcome of ~.600 • 1300-1400 has range of outcomes between ~.510 and ~.580 • 1400+ has range of outcomes between ~.300 and ~.470
BA_B	• Appears that higher batting averages are more prominent in winning teams • .230-.240 has range of outcomes between ~.290 and ~.510 • .230-.240 has range of outcomes between ~.330 and ~.570 • .230-.240 has range of outcomes between ~.410 and ~.600 • .230-.240 has range of outcomes between ~.400 and ~.610 • .230-.240 has range of outcomes between ~.430 and ~.550 • .280+ has a single outcome of ~.480
E	• Difficult to make a connection between E and W_L_Percentage • 50-60 has a single outcome of ~.530 • 130+ has a single outcome of ~.450 • 70-80, 90-100, and 100-110 have the largest range of outcomes of ~.330 and ~.500, ~.400 and ~.610, and ~.340 and ~.530 (min at ~.300 and max at ~.580), respectively

ERA	• Teams with the lowest ERAs tend to have winning records • <4.00 has range of outcomes between ~.510 and ~.610 • 4.00-4.50 has range of outcomes between ~.410 and ~.550 • 4.50-5.00 has range of outcomes between ~.340 and ~.490 • 5.00-5.50 has range of outcomes between ~.400 and ~.460 – with min of ~.300 and max of ~.550 • 5.50-6.00 has range of outcomes between ~.330 and ~.340
HR_P	• Difficult to make a connection between HR_P and W_L_Percentage • 120-130 has a single outcome of ~.550 • 300+ has a single outcome of ~.340 • 150-160 has range of outcomes between ~.340 and ~.600
BB_P	• Difficult to make a connection between BB_P and W_L_Percentage • 440-470 has a single outcome of ~.520 • +690 has a single outcome of ~.440 • 560-570 has the largest range of outcomes between ~.330 and ~.610
SO_P	• Difficult to make a connection between SO_P and W_L_Percentage • 600-700 has a single outcome of ~.540 • +1200 has a range of outcomes between ~.300 and ~.550 (largest range of outcomes) • 1100-1200 has a range of outcomes between ~.530 and ~.610
WHIP	• Teams with the lowest WHIPs tend to have winning records • 1.20-1.30 has range of outcomes between ~.510 and ~.590 • 1.30-1.40 has range of outcomes between ~.410 and ~.610 • 1.40-1.50 has range of outcomes between ~.300 and ~.550 • 1.50+ has range of outcomes between ~.430 and ~.440 – with min of ~.400 and max of ~.460

- *Summary of Data*

```
summary(orioles) #all seasons
       G          W_L_Percentage      R_Scored         BatAge          PAge           NumBat          NumP      
 Min.   : 60.0   Min.   :0.2900   Min.   :274.0   Min.   :26.30   Min.   :26.20   Min.   :32.0   Min.   :15.00  
 1st Qu.:162.0   1st Qu.:0.4155   1st Qu.:683.5   1st Qu.:28.10   1st Qu.:27.70   1st Qu.:42.0   1st Qu.:20.00  
 Median :162.0   Median :0.4630   Median :713.0   Median :29.20   Median :28.10   Median :46.0   Median :22.00  
 Mean   :157.6   Mean   :0.4650   Mean   :715.7   Mean   :29.23   Mean   :28.25   Mean   :46.0   Mean   :23.05  
 3rd Qu.:162.0   3rd Qu.:0.5250   3rd Qu.:762.0   3rd Qu.:30.20   3rd Qu.:28.85   3rd Qu.:48.5   3rd Qu.:26.00  
 Max.   :163.0   Max.   :0.6050   Max.   :949.0   Max.   :33.20   Max.   :30.70   Max.   :62.0   Max.   :42.00  
     Hits_B          HR_B            SB_B             BB_B            SO_B           BA_B       
 Min.   : 523   Min.   : 77.0   Min.   : 19.00   Min.   :164.0   Min.   : 514   Min.   :0.2360  
 1st Qu.:1364   1st Qu.:154.5   1st Qu.: 63.50   1st Qu.:449.0   1st Qu.: 901   1st Qu.:0.2510  
 Median :1434   Median :172.0   Median : 79.00   Median :504.0   Median : 952   Median :0.2590  
 Mean   :1399   Mean   :178.2   Mean   : 78.69   Mean   :505.9   Mean   :1022   Mean   :0.2591  
 3rd Qu.:1495   3rd Qu.:211.0   3rd Qu.: 93.00   3rd Qu.:580.0   3rd Qu.:1122   3rd Qu.:0.2680  
 Max.   :1614   Max.   :257.0   Max.   :144.00   Max.   :660.0   Max.   :1454   Max.   :0.2810  
       E               ERA             HR_P            BB_P            SO_P            WHIP      
 Min.   : 43.00   Min.   :3.430   Min.   : 79.0   Min.   :192.0   Min.   : 487   Min.   :1.241  
 1st Qu.: 87.00   1st Qu.:4.210   1st Qu.:156.0   1st Qu.:515.0   1st Qu.: 885   1st Qu.:1.353  
 Median : 97.00   Median :4.560   Median :180.0   Median :537.0   Median :1007   Median :1.419  
 Mean   : 95.69   Mean   :4.577   Mean   :181.8   Mean   :539.4   Mean   : 990   Mean   :1.413  
 3rd Qu.:107.50   3rd Qu.:4.990   3rd Qu.:205.0   3rd Qu.:579.0   3rd Qu.:1154   3rd Qu.:1.467  
 Max.   :135.00   Max.   :5.840   Max.   :305.0   Max.   :696.0   Max.   :1248   Max.   :1.565  
```

Notes:

- We can see that there were a few seasons that the Orioles did not play 162 games – perhaps the outliers should be removed.
- On average, the Orioles are a losing team posting a mean W_L_Percentage of .465 between 1984 and 2022.
- Minimum W_L_Percentage is .290 and maximum is .605.

Initial Model Run (*include* R output):

```
#remove 2020, 1995, 1994 shortened seasons
orioles_rev <- orioles[-c(3, 28, 29), ]

#initial model run
orioles.lm <- lm(W_L_Percentage ~
G+R_Scored+BatAge+PAge+NumBat+HR_B+SB_B+BB_B+SO_B+BA_B+E+ERA+HR_P+BB_P, data=orioles_rev)
```

- *Coefficient Estimates*

```
orioles.lm$coefficients
  (Intercept)             G      R_Scored        BatAge          PAge        NumBat          HR_B          SB_B          BB_B
-1.505333e+00  9.681101e-03  2.732327e-04 -8.618529e-03  5.959730e-03  5.480562e-04  5.416946e-04  3.126486e-04  1.369688e-04
         SO_B          BA_B             E           ERA          HR_P          BB_P
 5.781457e-05  2.156984e+00 -1.803625e-05 -1.068803e-01 -2.322909e-04 -3.228290e-05
```

- *Statistically Significant Variables*

```
summary(orioles.lm)

Call:
lm(formula = W_L_Percentage ~ G + R_Scored + BatAge + PAge +
    NumBat + HR_B + SB_B + BB_B + SO_B + BA_B + E + ERA + HR_P +
    BB_P, data = orioles_rev)

Residuals:
      Min        1Q    Median        3Q       Max
-0.043621 -0.013231  0.000276  0.013121  0.043937

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.505e+00  2.354e+00  -0.640    0.529
G            9.681e-03  1.470e-02   0.659    0.517
R_Scored     2.732e-04  2.471e-04   1.106    0.281
BatAge      -8.619e-03  7.571e-03  -1.138    0.268
PAge         5.960e-03  8.246e-03   0.723    0.478
NumBat       5.481e-04  2.019e-03   0.271    0.789
HR_B         5.417e-04  4.137e-04   1.309    0.205
SB_B         3.126e-04  3.611e-04   0.866    0.396
BB_B         1.370e-04  1.249e-04   1.097    0.285
SO_B         5.781e-05  7.997e-05   0.723    0.478
BA_B         2.157e+00  1.349e+00   1.599    0.125
E           -1.804e-05  3.577e-04  -0.050    0.960
ERA         -1.069e-01  2.135e-02  -5.007  5.9e-05 ***
HR_P        -2.323e-04  2.943e-04  -0.789    0.439
BB_P        -3.228e-05  1.678e-04  -0.192    0.849
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02655 on 21 degrees of freedom
Multiple R-squared:  0.9279,	Adjusted R-squared:  0.8798
F-statistic: 19.31 on 14 and 21 DF,  p-value: 6.766e-09
```

ERA is the only statistically significant predictor as its p-value is less than 0.001.

- *ANOVA Table*

```
Analysis of Variance Table

Response: W_L_Percentage
          Df   Sum Sq  Mean Sq  F value    Pr(>F)    
G          1 0.004435 0.004435   6.2942 0.0203854 *  
R_Scored   1 0.039449 0.039449  55.9851 2.366e-07 ***
BatAge     1 0.003406 0.003406   4.8331 0.0392670 *  
PAge       1 0.004781 0.004781   6.7854 0.0165382 *  
NumBat     1 0.037539 0.037539  53.2744 3.468e-07 ***
HR_B       1 0.005485 0.005485   7.7835 0.0109756 *  
SB_B       1 0.000024 0.000024   0.0346 0.8541800    
BB_B       1 0.000826 0.000826   1.1721 0.2912427    
SO_B       1 0.009945 0.009945  14.1131 0.0011605 ** 
BA_B       1 0.000028 0.000028   0.0403 0.8429144    
E          1 0.012131 0.012131  17.2164 0.0004547 ***
ERA        1 0.071976 0.071976 102.1453 1.608e-09 ***
HR_P       1 0.000413 0.000413   0.5861 0.4524565    
BB_P       1 0.000026 0.000026   0.0370 0.8492574    
Residuals 21 0.014797 0.000705                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- *Model Fit (Multiple R^2)*

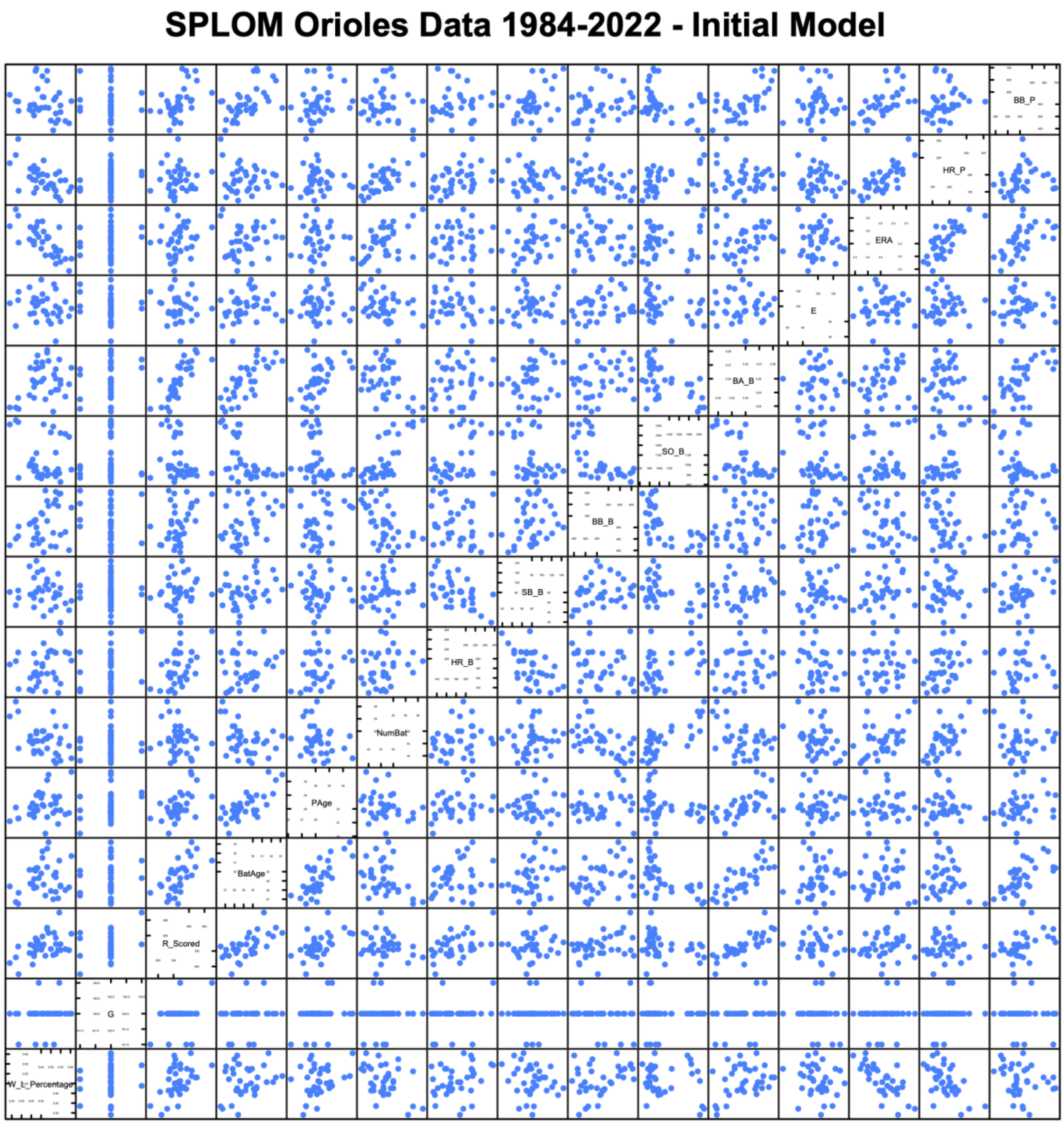
```
From summary(orioles.lm):
Adjusted R-squared:  0.8798
```

Returns the following initial model equation:
Y = -1.505333e+00 + 9.681101e-03(G) + 2.732327e-04(R_Scored) - 8.618529e-03(BatAge) + 5.959730e-03(PAge) + 5.480562e-04(NumBat) + 5.416946e-04(HR_B) + 3.126486e-04(SB_B) + 1.369688e-04(BB_B) + 5.781457e-05(SO_B) + 2.156984e+00(BA_B) - 1.803625e-05(E) - 1.068803e-01(ERA) - 2.322909e-04(HR_P) - 3.228290e-05(BB_P) + ϵ

Initial Model Diagnostics (*include* R output):

- *Scatterplot Matrix*

```
#splom initial model
dev.new(height=800, width=800)
splom(~
orioles_rev[,c("W_L_Percentage","G","R_Scored","BatAge","PAge","NumBat","HR_B","SB_B","BB_B","SO_B","BA_
B","E","ERA","HR_P","BB_P")],main="SPLOM  Orioles  Data  1984-2022  -  Initial  Model",  pch=19,  cex=.3,
xlab=NULL, axis.text.cex = 0.1, varname.cex = 0.3, axis.line.tck = .3)
```



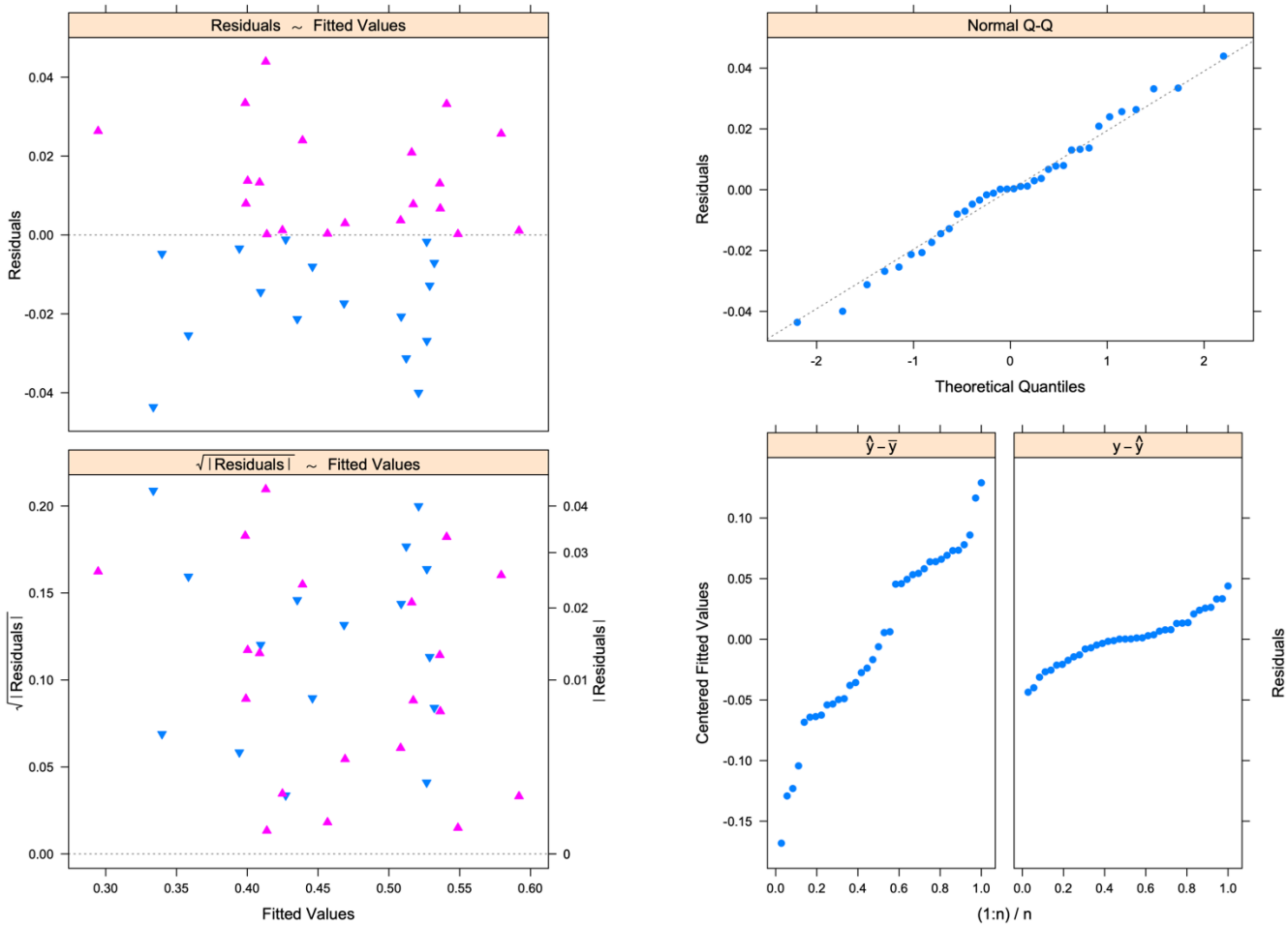
Glancing at the splom, multicollinearity among the predictors is not obvious. Looking at the correlation, this seems to be true as well. However, looking at the vif, there are several predictors with values over 5.

```
vif(orioles.lm)
        G  R_Scored    BatAge      PAge    NumBat      HR_B      SB_B      BB_B      SO_B      BA_B         E
 1.806011 16.736121  8.009806  3.001281  6.983753  9.812272  5.067443  4.951531 12.693933 13.553403  1.781247
      ERA      HR_P      BB_P
 7.369204  5.965589  5.460601
```

```
cor(orioles[,c("G","R_Scored","BatAge","PAge","NumBat","HR_B","SB_B","BB_B","SO_B","BA_B","E","ERA","HR_
P","BB_P")])
                   G    R_Scored      BatAge        PAge      NumBat        HR_B        SB_B        BB_B
G         1.00000000  0.72241484  0.20647417 -0.09722911  0.19241375  0.48644814  0.32096390  0.55480845
R_Scored  0.72241484  1.00000000  0.59274144  0.31891254 -0.01589833  0.64265894  0.33282496  0.65644726
BatAge    0.20647417  0.59274144  1.00000000  0.69638742 -0.35543871  0.22711107  0.37013347  0.44311251
PAge     -0.09722911  0.31891254  0.69638742  1.00000000 -0.28587254  0.19685731  0.01316094  0.16648142
NumBat    0.19241375 -0.01589833 -0.35543871 -0.28587254  1.00000000  0.28329672  0.06727939 -0.33297225
HR_B      0.48644814  0.64265894  0.22711107  0.19685731  0.28329672  1.00000000 -0.28422388  0.14618918
SB_B      0.32096390  0.33282496  0.37013347  0.01316094  0.06727939 -0.28422388  1.00000000  0.37102663
BB_B      0.55480845  0.65644726  0.44311251  0.16648142 -0.33297225  0.14618918  0.37102663  1.00000000
SO_B      0.48570027  0.15746675 -0.42872672 -0.34947331  0.72577048  0.56324174 -0.21987315 -0.20123232
BA_B     -0.06527895  0.50677572  0.74025395  0.59631470 -0.40507233  0.08701063  0.23697990  0.21446039
E         0.56177981  0.33978594  0.24569210 -0.10211626 -0.12410887  0.09840792  0.22673108  0.42561712
ERA       0.06844975  0.15215682  0.14237183  0.03318879  0.47671495  0.11607741  0.27079504 -0.13492639
HR_P      0.48701879  0.35588488 -0.02573286 -0.11347964  0.62663207  0.52979691  0.12937815 -0.07088543
BB_P      0.73423221  0.75858465  0.50861167  0.09555772  0.17048004  0.35780088  0.49321570  0.52111929
                SO_B        BA_B           E         ERA        HR_P       BB_P
G         0.48570027 -0.06527895  0.56177981  0.06844975  0.48701879 0.73423221
R_Scored  0.15746675  0.50677572  0.33978594  0.15215682  0.35588488 0.75858465
BatAge   -0.42872672  0.74025395  0.24569210  0.14237183 -0.02573286 0.50861167
PAge     -0.34947331  0.59631470 -0.10211626  0.03318879 -0.11347964 0.09555772
NumBat    0.72577048 -0.40507233 -0.12410887  0.47671495  0.62663207 0.17048004
HR_B      0.56324174  0.08701063  0.09840792  0.11607741  0.52979691 0.35780088
SB_B     -0.21987315  0.23697990  0.22673108  0.27079504  0.12937815 0.49321570
BB_B     -0.20123232  0.21446039  0.42561712 -0.13492639 -0.07088543 0.52111929
SO_B      1.00000000 -0.52950315  0.05199378  0.12411919  0.63030411 0.18486078
BA_B     -0.52950315  1.00000000 -0.06863455  0.18392719 -0.13782819 0.35028715
E         0.05199378 -0.06863455  1.00000000  0.08679955  0.25424183 0.49270703
ERA       0.12411919  0.18392719  0.08679955  1.00000000  0.68233466 0.45377099
HR_P      0.63030411 -0.13782819  0.25424183  0.68233466  1.00000000 0.46353081
BB_P      0.18486078  0.35028715  0.49270703  0.45377099  0.46353081 1.00000000
```

- *Residual and Normal Plots*

```
#residual and normal plots
dev.new(height=800, width=800)
lmplot(orioles.lm)
```



On the residual plot of the fitted values, we can see that the residuals are randomly spaced around the horizontal axis.

The Normal QQ plot shows normality except for a few places where it diverges from the line of normality.

Summary of Initial Model Diagnostics:
Before running the initial model, we eliminated 3 seasons based on games played, reducing the number of records to 36. In addition, we eliminated 4 predictors, Hits_B, WHIP, NumP, and SO_P, due to concerns over multicollinearity.

Initial model:
```
orioles.lm <- lm(W_L_Percentage ~
G+R_Scored+BatAge+PAge+NumBat+HR_B+SB_B+BB_B+SO_B+BA_B+E+ERA+HR_P+BB_P, data=orioles_rev)
```

After running the model, we found that there was only one significant predictor, ERA; however, the Adjusted R-Squared value for the model was 0.8798. Looking at the SPLOM and a correlation matrix, we did not see any signs of multicollinearity among the predictors; however, there were several predictors that had a variance inflation factor over 5.

Keeping all predictors and using the following coefficients:

```
orioles.lm$coefficients
  (Intercept)             G      R_Scored        BatAge          PAge        NumBat          HR_B          SB_B          BB_B
-1.505333e+00  9.681101e-03  2.732327e-04 -8.618529e-03  5.959730e-03  5.480562e-04  5.416946e-04  3.126486e-04  1.369688e-04
         SO_B          BA_B             E           ERA          HR_P          BB_P
 5.781457e-05  2.156984e+00 -1.803625e-05 -1.068803e-01 -2.322909e-04 -3.228290e-05
```

Gave us the following equation for the initial model:

Y = -1.505333e+00 + 9.681101e-03(G) + 2.732327e-04(R_Scored) - 8.618529e-03(BatAge) + 5.959730e-03(PAge) + 5.480562e-04(NumBat) + 5.416946e-04(HR_B) + 3.126486e-04(SB_B) + 1.369688e-04(BB_B) + 5.781457e-05(SO_B) + 2.156984e+00(BA_B) - 1.803625e-05(E) - 1.068803e-01(ERA) - 2.322909e-04(HR_P) - 3.228290e-05(BB_P) + ϵ

Based on the number of statistically significant predictors and the number of predictors with VIF over 5, a stepwise procedure could be used to find a parsimonious model.

Improvements to Linear Model (*circle one*):

- *Stepwise Procedures*

```
#subset
models <- leaps::regsubsets(W_L_Percentage ~
G+R_Scored+BatAge+PAge+NumBat+HR_B+SB_B+BB_B+SO_B+BA_B+E+ERA+HR_P+BB_P, data = orioles_rev, nbest=2)
models.summary <- summaryHH(models)
tmp <- (models.summary$cp < 10)
models.summary[tmp,]

                        model p   rsq    rss adjr2     cp   bic stderr
3                        R-ER 3 0.897 0.0212 0.890  0.119 -70.9 0.0254
5                     R-SO-ER 4 0.909 0.0187 0.900 -1.405 -71.8 0.0242
6                      R-E-ER 4 0.908 0.0189 0.900 -1.246 -71.6 0.0243
7                   R-SO-E-ER 5 0.913 0.0178 0.902 -0.782 -70.2 0.0239
8                 R-SO-BA_-ER 5 0.913 0.0179 0.902 -0.634 -70.0 0.0240
9               R-SO-BA_-E-ER 6 0.916 0.0172 0.902  0.443 -67.7 0.0240
10           R-SO-BA_-ER-BB_P 6 0.915 0.0174 0.901  0.689 -67.3 0.0241
11      R-SO-BA_-ER-HR_P-BB_P 7 0.918 0.0169 0.901  1.978 -64.8 0.0241
12      R-BB_B-SO-BA_-ER-BB_P 7 0.918 0.0169 0.901  2.011 -64.8 0.0242
13      R-BtA-P-N-HR_B-BA_-ER 8 0.920 0.0165 0.899  3.421 -62.1 0.0243
14    R-BB_B-SO-BA_-E-ER-BB_P 8 0.920 0.0165 0.899  3.429 -62.1 0.0243
15 G-R-BtA-N-HR_B-BB_B-BA_-ER 9 0.921 0.0161 0.898  4.874 -59.3 0.0244
16 R-BtA-P-N-HR_B-BB_B-BA_-ER 9 0.921 0.0161 0.898  4.913 -59.3 0.0245

Model variables with abbreviations
                                                                      model
ER                                                                      ERA
HR_P                                                                   HR_P
R-ER                                                           R_Scored-ERA
BA_-ER                                                             BA_B-ERA
R-SO-ER                                                   R_Scored-SO_B-ERA
R-E-ER                                                       R_Scored-E-ERA
R-SO-E-ER                                               R_Scored-SO_B-E-ERA
R-SO-BA_-ER                                          R_Scored-SO_B-BA_B-ERA
R-SO-BA_-E-ER                                      R_Scored-SO_B-BA_B-E-ERA
R-SO-BA_-ER-BB_P                                R_Scored-SO_B-BA_B-ERA-BB_P
R-SO-BA_-ER-HR_P-BB_P                      R_Scored-SO_B-BA_B-ERA-HR_P-BB_P
R-BB_B-SO-BA_-ER-BB_P                      R_Scored-BB_B-SO_B-BA_B-ERA-BB_P
R-BtA-P-N-HR_B-BA_-ER             R_Scored-BatAge-PAge-NumBat-HR_B-BA_B-ERA
R-BB_B-SO-BA_-E-ER-BB_P                  R_Scored-BB_B-SO_B-BA_B-E-ERA-BB_P
G-R-BtA-N-HR_B-BB_B-BA_-ER     G-R_Scored-BatAge-NumBat-HR_B-BB_B-BA_B-ERA
R-BtA-P-N-HR_B-BB_B-BA_-ER R_Scored-BatAge-PAge-NumBat-HR_B-BB_B-BA_B-ERA

model with largest adjr2
7

Number of observations
36

#steps with full linear model
models.step <- step(orioles.lm)

Start:  AIC=-250.69
W_L_Percentage ~ G + R_Scored + BatAge + PAge + NumBat + HR_B +
    SB_B + BB_B + SO_B + BA_B + E + ERA + HR_P + BB_P

           Df Sum of Sq      RSS     AIC
- E         1 0.0000018 0.014799 -252.68
- BB_P      1 0.0000261 0.014824 -252.62
- NumBat    1 0.0000519 0.014849 -252.56
```

```
- G          1 0.0003056 0.015103 -251.95
- PAge       1 0.0003680 0.015165 -251.80
- SO_B       1 0.0003683 0.015166 -251.80
- HR_P       1 0.0004390 0.015236 -251.63
- SB_B       1 0.0005282 0.015326 -251.42
<none>                   0.014797 -250.69
- BB_B       1 0.0008478 0.015645 -250.68
- R_Scored   1 0.0008615 0.015659 -250.65
- BatAge     1 0.0009131 0.015711 -250.53
- HR_B       1 0.0012080 0.016005 -249.86
- BA_B       1 0.0018021 0.016599 -248.55
- ERA        1 0.0176629 0.032460 -224.41

Step:  AIC=-252.68
W_L_Percentage ~ G + R_Scored + BatAge + PAge + NumBat + HR_B +
    SB_B + BB_B + SO_B + BA_B + ERA + HR_P + BB_P

           Df Sum of Sq      RSS     AIC
- BB_P       1 0.0000276 0.014827 -254.61
- NumBat     1 0.0000611 0.014860 -254.53
- G          1 0.0003104 0.015110 -253.93
- SO_B       1 0.0003724 0.015172 -253.79
- PAge       1 0.0003909 0.015190 -253.74
- HR_P       1 0.0004911 0.015290 -253.51
- SB_B       1 0.0005610 0.015360 -253.34
<none>                   0.014799 -252.68
- BB_B       1 0.0008633 0.015662 -252.64
- R_Scored   1 0.0008712 0.015670 -252.62
- BatAge     1 0.0010589 0.015858 -252.19
- HR_B       1 0.0013109 0.016110 -251.63
- BA_B       1 0.0020496 0.016849 -250.01
- ERA        1 0.0176617 0.032461 -226.41

Step:  AIC=-254.61
W_L_Percentage ~ G + R_Scored + BatAge + PAge + NumBat + HR_B +
    SB_B + BB_B + SO_B + BA_B + ERA + HR_P

           Df Sum of Sq      RSS     AIC
- NumBat     1 0.0000856 0.014912 -256.41
- SO_B       1 0.0003565 0.015183 -255.76
- HR_P       1 0.0004638 0.015291 -255.50
- PAge       1 0.0004955 0.015322 -255.43
- G          1 0.0005094 0.015336 -255.40
- SB_B       1 0.0005441 0.015371 -255.32
- BB_B       1 0.0008443 0.015671 -254.62
- R_Scored   1 0.0008469 0.015674 -254.62
<none>                   0.014827 -254.61
- HR_B       1 0.0014398 0.016267 -253.28
- BatAge     1 0.0015242 0.016351 -253.09
- BA_B       1 0.0020534 0.016880 -251.94
- ERA        1 0.0290151 0.043842 -217.59

Step:  AIC=-256.41
W_L_Percentage ~ G + R_Scored + BatAge + PAge + HR_B + SB_B +
    BB_B + SO_B + BA_B + ERA + HR_P

           Df Sum of Sq      RSS     AIC
- G          1  0.000494 0.015406 -257.23
- HR_P       1  0.000512 0.015425 -257.19
- PAge       1  0.000519 0.015431 -257.18
- R_Scored   1  0.000800 0.015712 -256.53
<none>                   0.014912 -256.41
```

```
- BB_B      1  0.000880 0.015793 -256.34
- SB_B      1  0.001002 0.015914 -256.07
- SO_B      1  0.001183 0.016096 -255.66
- BatAge    1  0.001535 0.016447 -254.88
- HR_B      1  0.001580 0.016492 -254.78
- BA_B      1  0.002064 0.016976 -253.74
- ERA       1  0.033949 0.048861 -215.68

Step:  AIC=-257.23
W_L_Percentage ~ R_Scored + BatAge + PAge + HR_B + SB_B + BB_B +
    SO_B + BA_B + ERA + HR_P

           Df Sum of Sq      RSS     AIC
- HR_P      1  0.000352 0.015758 -258.42
- BB_B      1  0.000617 0.016023 -257.82
- PAge      1  0.000799 0.016206 -257.41
<none>                  0.015406 -257.23
- SB_B      1  0.000974 0.016381 -257.03
- SO_B      1  0.001071 0.016477 -256.81
- R_Scored  1  0.001164 0.016571 -256.61
- HR_B      1  0.001388 0.016794 -256.13
- BatAge    1  0.001654 0.017061 -255.56
- BA_B      1  0.001922 0.017329 -255.00
- ERA       1  0.036228 0.051635 -215.69

Step:  AIC=-258.42
W_L_Percentage ~ R_Scored + BatAge + PAge + HR_B + SB_B + BB_B +
    SO_B + BA_B + ERA

           Df Sum of Sq      RSS     AIC
- PAge      1  0.000711 0.016469 -258.83
- BB_B      1  0.000817 0.016575 -258.60
- SB_B      1  0.000829 0.016587 -258.58
<none>                  0.015758 -258.42
- SO_B      1  0.000969 0.016727 -258.27
- HR_B      1  0.001133 0.016891 -257.92
- R_Scored  1  0.001134 0.016892 -257.92
- BatAge    1  0.001500 0.017258 -257.15
- BA_B      1  0.002197 0.017955 -255.72
- ERA       1  0.114751 0.130509 -184.31

Step:  AIC=-258.83
W_L_Percentage ~ R_Scored + BatAge + HR_B + SB_B + BB_B + SO_B +
    BA_B + ERA

           Df Sum of Sq      RSS     AIC
- SB_B      1  0.000534 0.017003 -259.68
- BB_B      1  0.000730 0.017199 -259.27
- BatAge    1  0.000823 0.017292 -259.08
- HR_B      1  0.000939 0.017408 -258.84
<none>                  0.016469 -258.83
- SO_B      1  0.001167 0.017636 -258.37
- R_Scored  1  0.001386 0.017855 -257.92
- BA_B      1  0.002109 0.018579 -256.49
- ERA       1  0.115734 0.132204 -185.85

Step:  AIC=-259.68
W_L_Percentage ~ R_Scored + BatAge + HR_B + BB_B + SO_B + BA_B +
    ERA

           Df Sum of Sq      RSS     AIC
- BB_B      1  0.000362 0.017365 -260.93
```

```
- HR_B      1  0.000405 0.017408 -260.84
- BatAge    1  0.000460 0.017463 -260.72
<none>                  0.017003 -259.68
- SO_B      1  0.001010 0.018013 -259.61
- BA_B      1  0.001576 0.018579 -258.49
- R_Scored  1  0.004179 0.021182 -253.77
- ERA       1  0.126568 0.143571 -184.88

Step:  AIC=-260.93
W_L_Percentage ~ R_Scored + BatAge + HR_B + SO_B + BA_B + ERA

           Df Sum of Sq      RSS     AIC
- HR_B      1  0.000224 0.017589 -262.46
- BatAge    1  0.000439 0.017804 -262.03
- SO_B      1  0.000649 0.018014 -261.61
<none>                  0.017365 -260.93
- BA_B      1  0.001275 0.018640 -260.38
- R_Scored  1  0.011938 0.029303 -244.09
- ERA       1  0.132300 0.149665 -185.38

Step:  AIC=-262.46
W_L_Percentage ~ R_Scored + BatAge + SO_B + BA_B + ERA

           Df Sum of Sq      RSS     AIC
- BatAge    1  0.000284 0.017874 -263.89
<none>                  0.017589 -262.46
- BA_B      1  0.001101 0.018690 -262.28
- SO_B      1  0.002168 0.019758 -260.28
- R_Scored  1  0.020557 0.038146 -236.59
- ERA       1  0.136147 0.153736 -186.42

Step:  AIC=-263.89
W_L_Percentage ~ R_Scored + SO_B + BA_B + ERA

           Df Sum of Sq      RSS     AIC
- BA_B      1  0.000867 0.018740 -264.18
<none>                  0.017874 -263.89
- SO_B      1  0.003311 0.021184 -259.77
- R_Scored  1  0.020435 0.038309 -238.44
- ERA       1  0.139186 0.157059 -187.65

Step:  AIC=-264.18
W_L_Percentage ~ R_Scored + SO_B + ERA

           Df Sum of Sq      RSS     AIC
<none>                  0.018740 -264.18
- SO_B      1  0.002483 0.021223 -261.70
- R_Scored  1  0.074477 0.093217 -208.43
- ERA       1  0.142580 0.161320 -188.68

#step1
orioles.step.lm <- lm(W_L_Percentage ~ R_Scored + SO_B + ERA, data=orioles_rev)
summary(orioles.step.lm)

Call:
lm(formula = W_L_Percentage ~ R_Scored + SO_B + ERA, data = orioles_rev)

Residuals:
      Min        1Q    Median        3Q       Max
-0.048541 -0.016116 -0.000057  0.018764  0.048845
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.575e-01  5.755e-02   7.950 4.49e-09 ***
R_Scored     6.644e-04  5.891e-05  11.277 1.11e-12 ***
SO_B         4.459e-05  2.165e-05   2.059   0.0477 *
ERA         -1.149e-01  7.364e-03 -15.603  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0242 on 32 degrees of freedom
Multiple R-squared:  0.9087,  Adjusted R-squared:  0.9001
F-statistic: 106.2 on 3 and 32 DF,  p-value: < 2.2e-16

vif(orioles.step.lm)

R_Scored     SO_B      ERA
1.144576 1.119939 1.055084

cor(orioles_rev[,c("R_Scored","SO_B","ERA")])

           R_Scored        SO_B        ERA
R_Scored  1.0000000 -0.29331557 0.17278176
SO_B     -0.2933156  1.00000000 0.09225928
ERA       0.1727818  0.09225928 1.00000000

anova(orioles.step.lm)
Analysis of Variance Table

Response: W_L_Percentage
          Df   Sum Sq  Mean Sq  F value    Pr(>F)
R_Scored   1 0.043884 0.043884  74.9350 6.829e-10 ***
SO_B       1 0.000057 0.000057   0.0981    0.7562
ERA        1 0.142580 0.142580 243.4642 < 2.2e-16 ***
Residuals 32 0.018740 0.000586
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#step2 - dropping SO_B because of anova table
orioles.step.lm2 <- lm(W_L_Percentage ~ R_Scored + ERA, data=orioles_rev)
summary(orioles.step.lm2)

Call:
lm(formula = W_L_Percentage ~ R_Scored + ERA, data = orioles_rev)

Residuals:
      Min        1Q    Median        3Q       Max
-0.038853 -0.016924 -0.004613  0.015334  0.045463

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.219e-01  5.064e-02   10.30 7.59e-12 ***
R_Scored     6.261e-04  5.859e-05   10.69 2.99e-12 ***
ERA         -1.126e-01  7.628e-03  -14.76 4.28e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02536 on 33 degrees of freedom
Multiple R-squared:  0.8966,  Adjusted R-squared:  0.8903
F-statistic: 143.1 on 2 and 33 DF,  p-value: < 2.2e-16
```

```
vif(orioles.step.lm2)

R_Scored      ERA
1.030772 1.030772

cor(orioles_rev[,c("R_Scored","ERA")])

          R_Scored       ERA
R_Scored 1.0000000 0.1727818
ERA      0.1727818 1.0000000

anova(orioles.step.lm2)

Analysis of Variance Table

Response: W_L_Percentage
          Df   Sum Sq  Mean Sq F value    Pr(>F)
R_Scored   1 0.043884 0.043884  68.236 1.537e-09 ***
ERA        1 0.140154 0.140154 217.929 4.276e-16 ***
Residuals 33 0.021223 0.000643
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- *Ridge Regression*
- *Lasso Technique*
- *Principal Components Regression*

Final Model Run (*include* R output):

From the stepwise procedure, use a model consisting of only the R_Scored and ERA predictors.

```
orioles.step.lm2 <- lm(W_L_Percentage ~ R_Scored + ERA, data=orioles_rev)
summary(orioles.step.lm2)

Call:
lm(formula = W_L_Percentage ~ R_Scored + ERA, data = orioles_rev)

Residuals:
      Min        1Q    Median        3Q       Max
-0.038853 -0.016924 -0.004613  0.015334  0.045463

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.219e-01  5.064e-02   10.30 7.59e-12 ***
R_Scored     6.261e-04  5.859e-05   10.69 2.99e-12 ***
ERA         -1.126e-01  7.628e-03  -14.76 4.28e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02536 on 33 degrees of freedom
Multiple R-squared:  0.8966,  Adjusted R-squared:  0.8903
F-statistic: 143.1 on 2 and 33 DF,  p-value: < 2.2e-16

vif(orioles.step.lm2)

R_Scored      ERA
1.030772 1.030772

cor(orioles_rev[,c("R_Scored","ERA")])

          R_Scored       ERA
R_Scored 1.0000000 0.1727818
ERA      0.1727818 1.0000000

anova(orioles.step.lm2)

Analysis of Variance Table

Response: W_L_Percentage
          Df   Sum Sq  Mean Sq F value    Pr(>F)
R_Scored   1 0.043884 0.043884  68.236 1.537e-09 ***
ERA        1 0.140154 0.140154 217.929 4.276e-16 ***
Residuals 33 0.021223 0.000643
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- *Coefficient Estimates*

```
orioles.step.lm2$coefficients
  (Intercept)      R_Scored           ERA
 0.5218788166  0.0006261264 -0.1126009062
```

- *Statistically Significant Variables*

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.219e-01  5.064e-02   10.30 7.59e-12 ***
R_Scored     6.261e-04  5.859e-05   10.69 2.99e-12 ***
ERA         -1.126e-01  7.628e-03  -14.76 4.28e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R_Scored and ERA are statistically significant as the p-values for both are under 0.001.

- *ANOVA Table*

```
Analysis of Variance Table

Response: W_L_Percentage
          Df   Sum Sq  Mean Sq F value    Pr(>F)
R_Scored   1 0.043884 0.043884  68.236 1.537e-09 ***
ERA        1 0.140154 0.140154 217.929 4.276e-16 ***
Residuals 33 0.021223 0.000643
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R_Scored and ERA are statistically significant at 0.001.

- *Model Fit (Multiple R^2)*

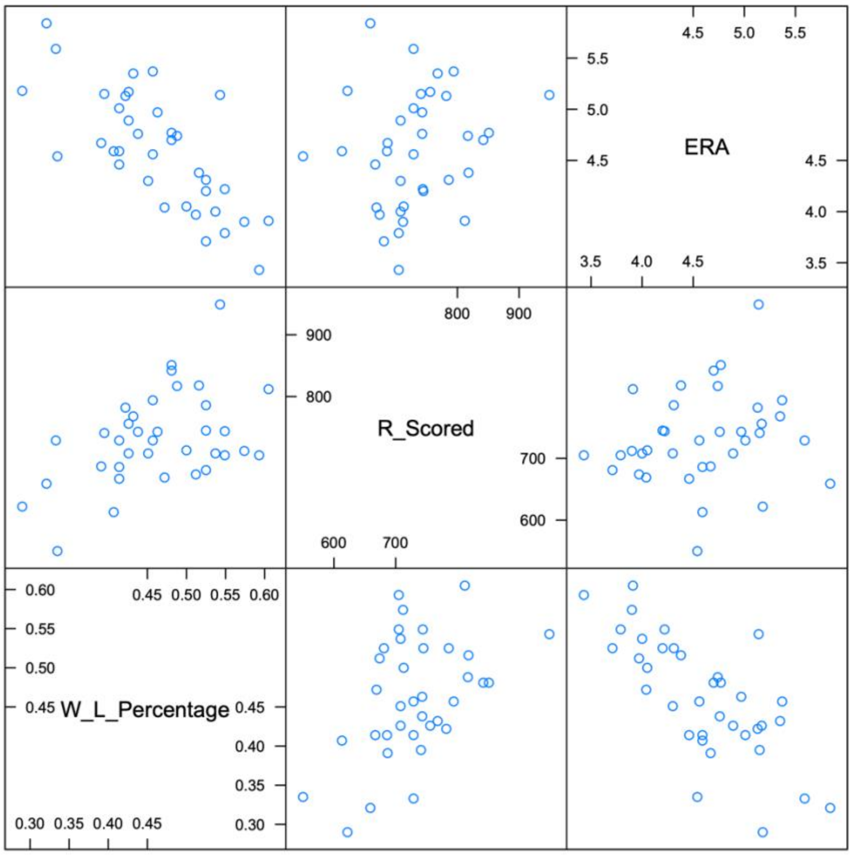
```
From summary(orioles.step.lm2):
Adjusted R-squared:  0.8903
```

Final model: Y = 0.5218788166 + 0.0006261264 (R_Scored) - 0.1126009062 (ERA) + ϵ

Final Model Diagnostics (*include* R output):

- *Scatterplot Matrix*

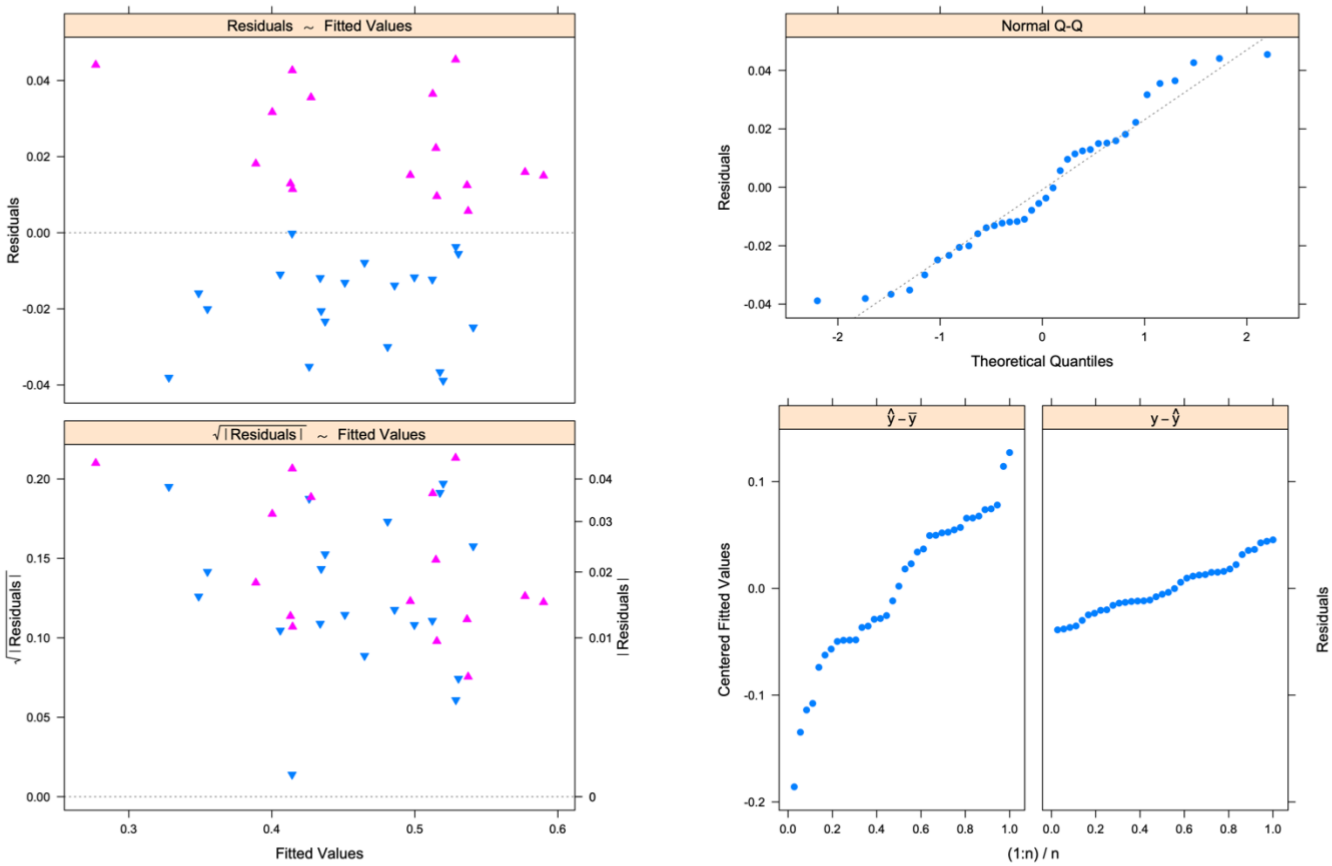
```
#splom final
dev.new(height=400, width=400)
splom(~ orioles_rev[,c("W_L_Percentage","R_Scored","ERA")],
      axis.text.cex=.7, xlab=NULL,
      auto.key=list(space="right", border=TRUE))
```



From the splom, we can observe that there does not appear to be much correlation between R_Scored and ERA; however, there is a semblance of a correlation between the response, W_L_Percentage, and the predictors, R_Scored and ERA. It appears that R_Scored has a positive correlation with W_L_Percentage and ERA has a negative correlation with W_L_Percentage. This would make sense – teams win more games when they score a lot of runs (positive correlation) and their pitchers do not give up many runs (negative correlation).

- ## *Residual and Normal Plots*

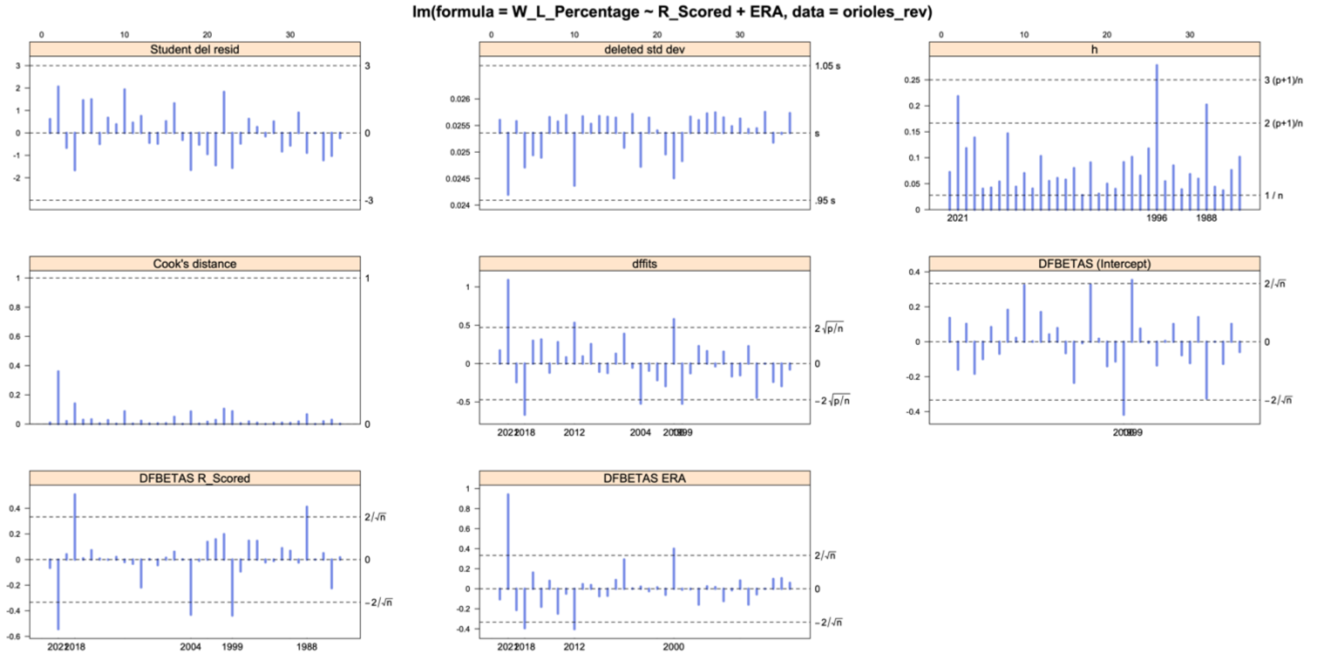
```
#final residual
dev.new(height=400, width=400)
lmplot(orioles.step.lm3)
```



On the residual plot of the fitted values, we can see that the residuals are randomly spaced around the horizontal axis. At first, the Normal QQ plot seems like it shows normality except for the tails; however, zooming in we can see several places where it appears to diverge from normality.

- ## *Case Statistics*

```
#case statistics
orioles.case.step <- case(orioles.step.lm2)
orioles.case.step.trellis <-
  plot(orioles.case.step, orioles.step.lm2, par.strip.text=list(cex=1.2),
       layout=c(3,3), main.cex=1.6, col=likertColor(2)[2], lwd=4)
dev.new(height=800, width=800)
orioles.case.step.trellis

                    Noteworthy Observations
Student del resid
deleted std dev
h                   2 26 32
Cook's distance
dffits              2 4 10 18 22 23
DFBETAS (Intercept) 22 23
DFBETAS R_Scored    2 4 18 23 32
DFBETAS ERA         2 4 10 22
```



Leverage and Difference in Fits identify record 2 as a noteworthy observation. Here are the notable variables for this record (2021):

- 2nd lowest W_L_Percentage
- 4th lowest R_Scored
- Highest ERA

This all makes sense – In the season where the Orioles had their second fewest wins, the team posted its highest ERA and 4th lowest R_Scored.

Difference in Fits identifies record 4 as a noteworthy observation as well. Here are the notable variables for this record (2018):

- Lowest W_L_Percentage
- 3rd lowest R_Scored
- 5th Highest ERA

To determine if the data is significantly different from a normal distribution, we can run a Shapiro-Wilk test:

```
shapiro.test(resid(orioles.step.lm2))

        Shapiro-Wilk normality test

data:  resid(orioles.step.lm2)
W = 0.95646, p-value = 0.1669
```

The p-value is greater than 0.05, so we assume that the data is not significantly different from normal distribution – and we will keep the remaining records and current model.

Final Linear Model and Summary of Model Results:

```
#step 2 - final
orioles.step.lm2 <- lm(W_L_Percentage ~ R_Scored + ERA, data=orioles_rev)
summary(orioles.step.lm2)

Call:
lm(formula = W_L_Percentage ~ R_Scored + ERA, data = orioles_rev)

Residuals:
      Min        1Q    Median        3Q       Max
-0.038853 -0.016924 -0.004613  0.015334  0.045463

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.219e-01  5.064e-02   10.30 7.59e-12 ***
R_Scored     6.261e-04  5.859e-05   10.69 2.99e-12 ***
ERA         -1.126e-01  7.628e-03  -14.76 4.28e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02536 on 33 degrees of freedom
Multiple R-squared:  0.8966,  Adjusted R-squared:  0.8903
F-statistic: 143.1 on 2 and 33 DF,  p-value: < 2.2e-16

vif(orioles.step.lm2)

R_Scored      ERA
1.030772 1.030772

cor(orioles_rev[,c("R_Scored","ERA")])

          R_Scored       ERA
R_Scored 1.0000000 0.1727818
ERA      0.1727818 1.0000000

anova(orioles.step.lm2)

Analysis of Variance Table

Response: W_L_Percentage
          Df   Sum Sq  Mean Sq F value    Pr(>F)
R_Scored   1 0.043884 0.043884  68.236 1.537e-09 ***
ERA        1 0.140154 0.140154 217.929 4.276e-16 ***
Residuals 33 0.021223 0.000643
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After running stepwise procedures, we reduced the number of predictors down from 14 to 2: R_Scored (runs scored) and ERA. This model has an Adjusted R-Squared value of 0.8903, which is better than the initial model at 0.8798. Both ERA and R_Scored are statistically significant as their p-values are less than 0.001. Looking at the VIF and correlation, there is no evidence of multicollinearity between the two predictors. These predictors contribute most to the Baltimore Orioles' Win/Loss Percentages (W_L_Percentage).

This model produced the following coefficients:

```
  (Intercept)      R_Scored           ERA
 0.5218788166  0.0006261264 -0.1126009062
```

With these coefficients, we produced the following model equation:

$Y = 0.5218788166 + 0.0006261264\ (R\_Scored) - 0.1126009062\ (ERA) + \boldsymbol{\epsilon}$

According to our model, we can observe the following:
- W_L_Percentage increases by 0.0006261264 when R_Scored increases by 1
- W_L_Percentage decreases by 0.1126009062 when ERA increases by 1

In other words, the Orioles' Win/Loss Percentage (W_L_Percentage) is higher when the runs scored by the team is higher (R_Scored) and the team's ERA is lower.