Introduction

In this study we will be analyzing data related to Major League Baseball. Our goal is to determine what factors have an effect on the win–loss record of each of the 30 major league baseball teams over the past ten years.

The explanatory variables that we will be studying are:

- Team Batting Average
- Team ERA (earned run average)
- Baseball Stadium Dimensions
- Team Payroll
- Average Game Attendance

We will find this information for each team using the following websites:

- www.baseball-reference.com
- www.foxsports.com
- www.espn.com
- www.baseball-almanac.com

After analyzing these data points, using R, we will be able determine which variables have a strong relation to winning percentage and which variables have little or no relation to winning percentage.

We will also be able to take an unknown team from any year between 1998 and 2007 and, given the explanatory variables, determine what their winning percentage should be for that year.

Analysis

Once the data was put in a .csv file, we standardized the data and shifted it 5 to the right to avoid negative values. This allowed us to compare the coefficients that the linear model resulted in. None of the years indicated that any variables had significant variance inflation factors, so all were included in the evaluation. Using the Akaike Information Criterion, we were able to obtain the ideal number of variables to include in the model. Then using stepwise regression, we obtained a final model with coefficients, but before we did this we had to make sure the variance was constant with the funnel function and that outliers were excluded using the lrplot function.

The residuals vs fitted plots and variance vs fitted plots indicated that the data in some years such as 2003 and 2004 was not ideal for multivariate regression. Using the coefficients, we were able to see that in nearly all the years, ERA had a greater effect than batting average, and payroll was a very unstable factor. Payroll even had a negative impact in the 1999 year, and afterwards seems to disappear from the model for 4 years. There seem to be lurking variables that affect the model, such as large contracts signed by players such as Alex Rodriguez that don't result in a gain in win percentage, or use of steroids.

Conclusion

After running the analysis over the span of ten years, we determined that ERA and Batting Average were more important factors and surprising payroll was not as big of a factor. Of the two key variables, the coefficient associated with ERA was usually twice as big as the batting average. This can be the case because ERA stands for Earned Runs Allowed directly relating to the runs that were scored by the opposition, while batting average only shows how often the batters were able to safely reach base and not necessarily score runs for their side. Payroll on the other hand usually had a small coefficient and was only a critical factor in 5 of the 10 years studied. This shows that blindly spending money at players is not always directly translate to a team's wins and needs to be carefully spent. Also, according to the Mitchell Report (attached), a huge lurking variable could be rampant steroid use by players going undetected.

HOW TO USE THE FORMULAS

Once we have the formulas, we can estimate the winning percentage of a team given the year and their explanatory variables.

Example 1:

Calculate the winning percentage of an unknown 1998 team given that their payroll is $45,000,000, their team ERA is 4.14, their team batting average is 0.274, and their average game day attendance is 28,000 fans.

Plugging these numbers into the formula for 1998:

Win % = .187 + (2.96*10^-9)*Payroll + (2.58)*Batting Average – (2.11*10^-6)*Attendance – (9.72*10^-2)*ERA

Winning % = 0.565

Example 2:

Calculate the winning percentage of an unknown 1999 team given that their payroll is $32,000,000, their team ERA is 5.05, and their team batting average is 0.255.

Plugging these numbers into the formula for 1999:

Win % = .168 + (1.25*10^-9)*Payroll + (2.49)*Batting Average – (8.77*10^-2)*ERA

Winning % = 0.399

1998

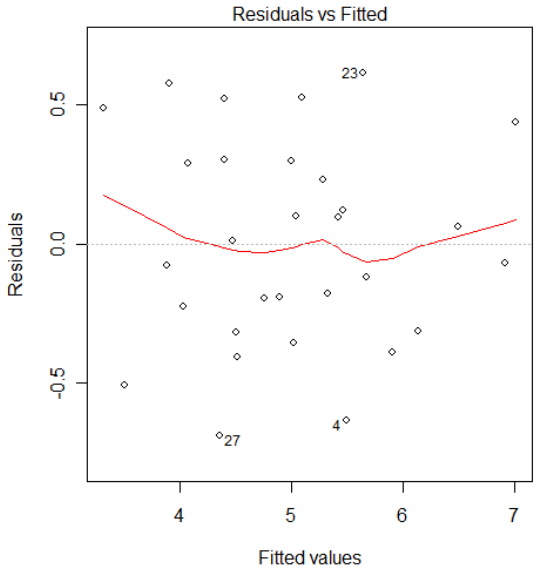
Coefficients:

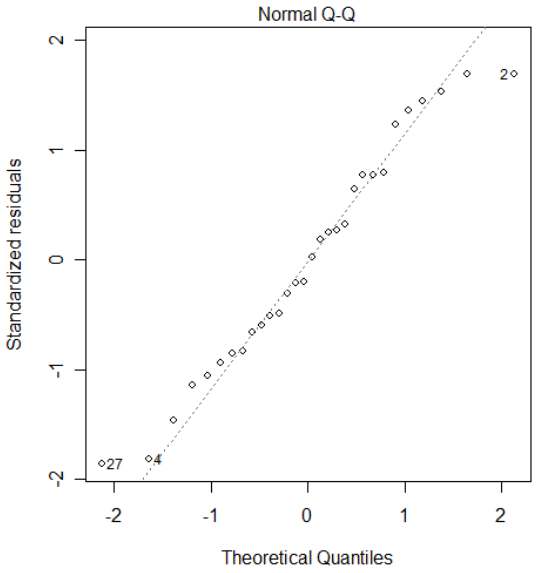
(Intercept):	4.8635
Payroll(x_1):	0.5371
ERA(x_2):	-0.6124
Batting Average (x_3):	0.3705
Attendance (x_4):	-0.2679

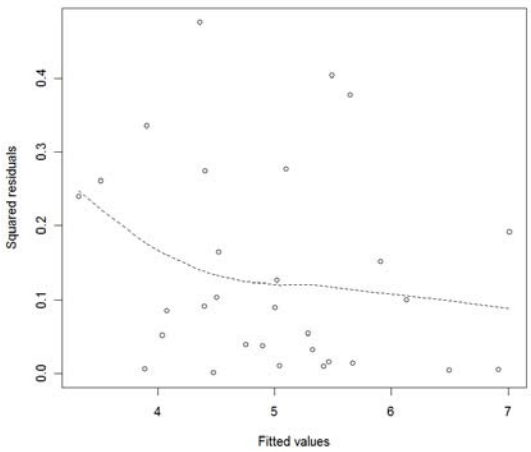
$Y = 4.8635 + .5371\ x_1 - .6124\ x_2 + 0.3705\ x_3 - 0.2679\ x_4$

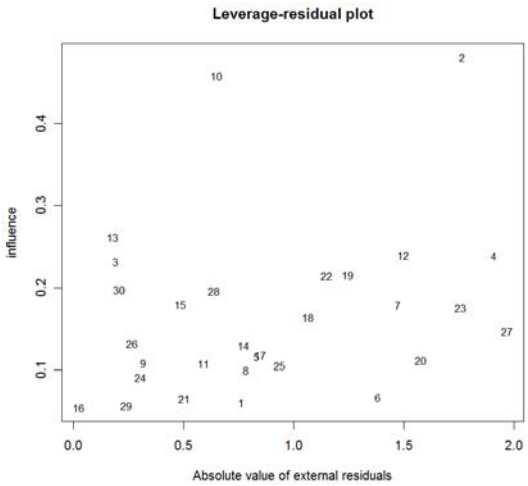
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.69000 -0.29322 -0.02991  0.29688  0.61423

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.86352    0.64835   7.501 7.44e-08 ***
Payroll          0.53707    0.13204   4.067 0.000417 ***
ERA             -0.61237    0.07852  -7.799 3.72e-08 ***
Batting.Average  0.37046    0.09300   3.984 0.000517 ***
Attendance      -0.26787    0.11583  -2.313 0.029262 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4011 on 25 degrees of freedom
Multiple R-Squared: 0.8613,     Adjusted R-squared: 0.8392
F-statistic: 38.82 on 4 and 25 DF,  p-value: 2.213e-10
```

Win % = .187 + (2.96*10^-9)*Payroll + (2.58)*Batting Average – (2.11*10^-6)*Attendance – (9.72*10^-2)*ERA


Residuals vs Fitted
Residuals
Fitted values
23
27
4


Normal Q-Q
Standardized residuals
Theoretical Quantiles
2
27
4


Squared residuals
Fitted values


Leverage-residual plot
influence
Absolute value of external residuals

1999

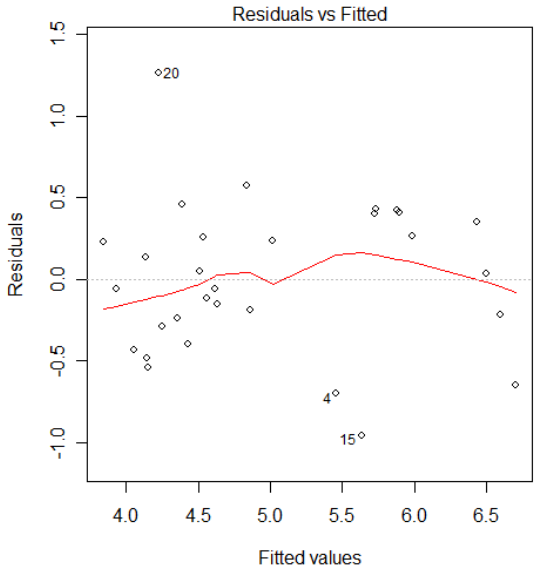
Coefficients:

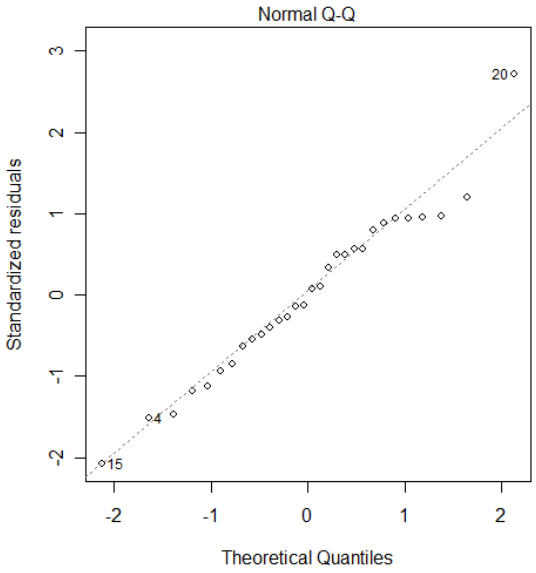
(Intercept):	4.5558
Payroll(x_1):	-0.3513
ERA(x_2):	0.6033
Batting Average (x_3):	0.3408

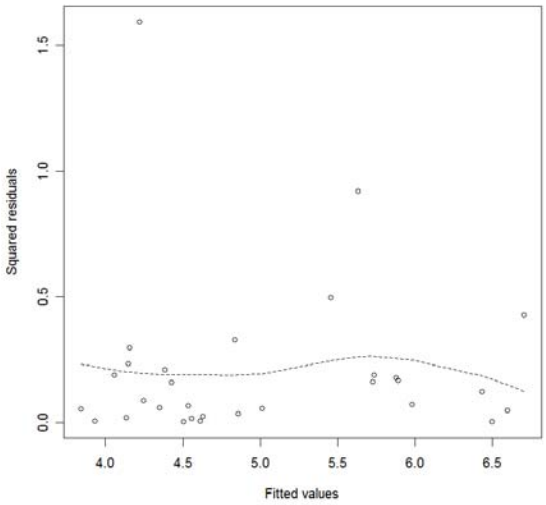
$Y = 4.5558 - .3513\ x_1 + .6033\ x_2 + 0.34083\ x_3$

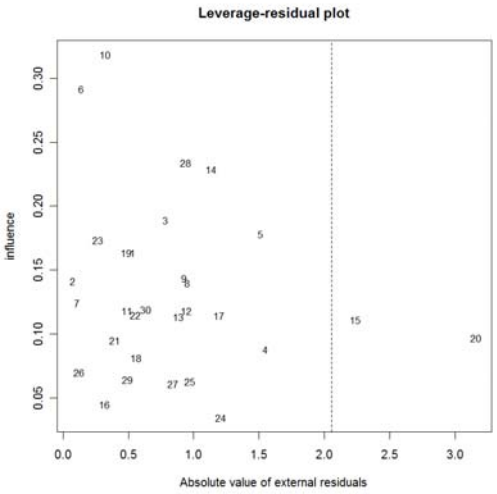
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.95825 -0.28008 -0.01138  0.32726  1.26244

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.5558     0.7945   5.734 4.90e-06 ***
Payroll           0.3513     0.1134   3.098  0.00464 **
ERA              -0.6033     0.1033  -5.840 3.72e-06 ***
Batting.Average   0.3408     0.1081   3.153  0.00405 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4881 on 26 degrees of freedom
Multiple R-Squared: 0.7864,     Adjusted R-squared: 0.7617
F-statistic:  31.9 on 3 and 26 DF,  p-value: 7.239e-09
```

Win % = .168 + (1.25*10^-9)*Payroll + (2.49)*Batting Average – (8.77*10^-2)*ERA


Residuals vs Fitted
Residuals
Fitted values
20
4
15


Normal Q-Q
Standardized residuals
Theoretical Quantiles
20
4
15


Squared residuals
Fitted values


Leverage-residual plot
influence
Absolute value of external residuals

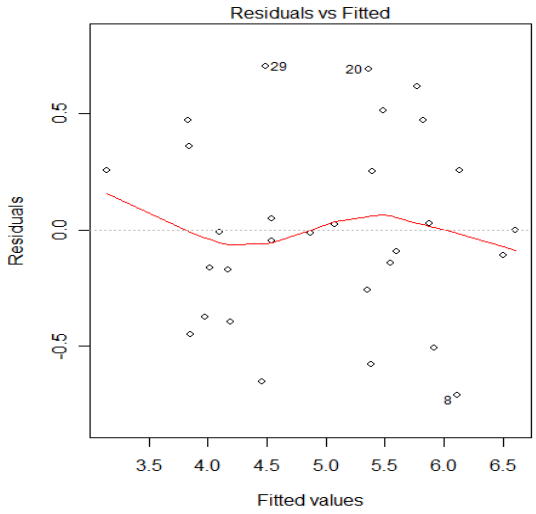
2000

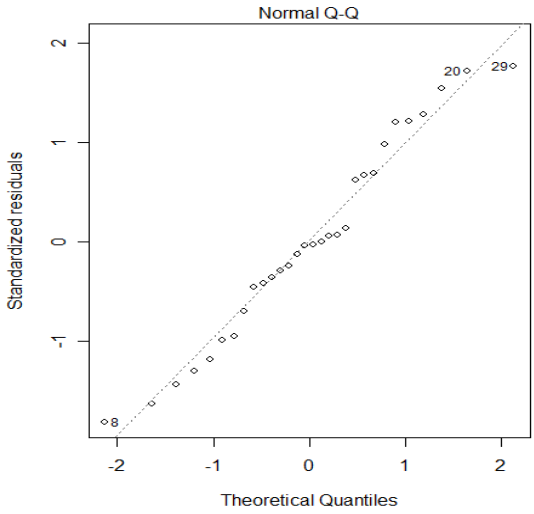
Coefficients:
(Intercept): 6.1267
ERA(x_1): -0.9080
Batting Average(x_2): 0.6827

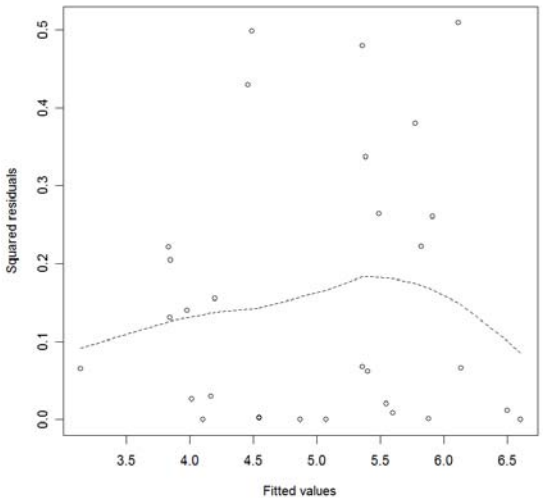
$Y = 5.5293 - .7212\ x_1 + .4251\ x_2$

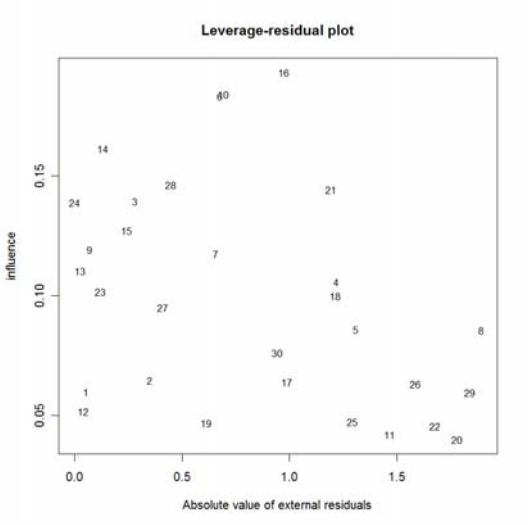
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.71355 -0.23837 -0.01308  0.25691  0.70593

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.12668    0.47028  13.028 3.70e-13 ***
ERA             -0.90799    0.08220 -11.046 1.61e-11 ***
Batting.Average  0.68267    0.08221   8.304 6.51e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4126 on 27 degrees of freedom
Multiple R-Squared: 0.8415,     Adjusted R-squared: 0.8297
F-statistic: 71.66 on 2 and 27 DF,  p-value: 1.589e-11
```

Win % = .131 + (3.63)*Batting Average – (.128)*ERA


Residuals vs Fitted
Residuals
Fitted values
29
20
8


Normal Q-Q
Standardized residuals
Theoretical Quantiles
20
29
8


Squared residuals
Fitted values


Leverage-residual plot
influence
Absolute value of external residuals

2001

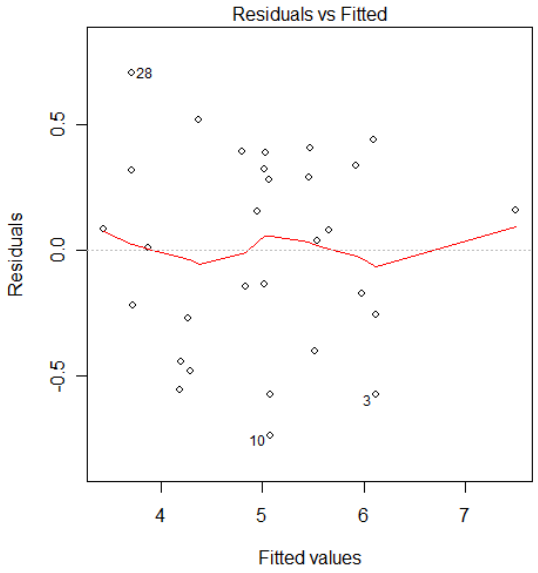
Coefficients:

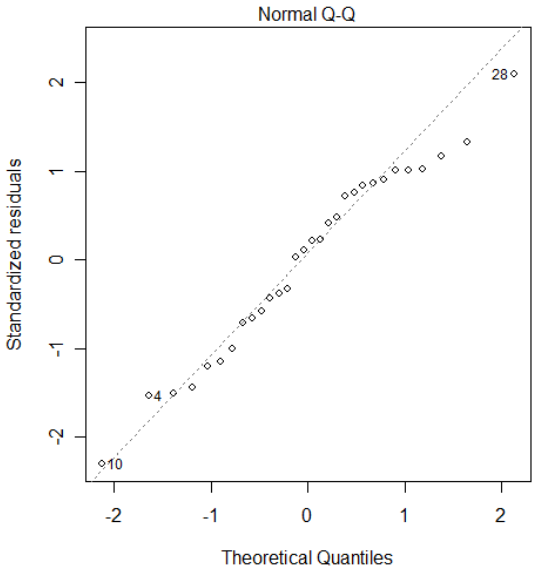
(Intercept):	5.5293
ERA(x_1):	-0.7212
Batting Average(x_2):	0.4251
Attendance (x_3):	0.1903

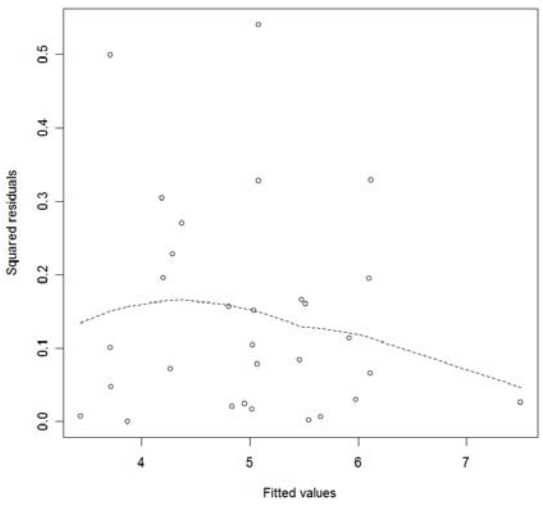
$Y = 5.5293 - .7212\ x_1 + .4251\ x_2 + 0.1903\ x_3$

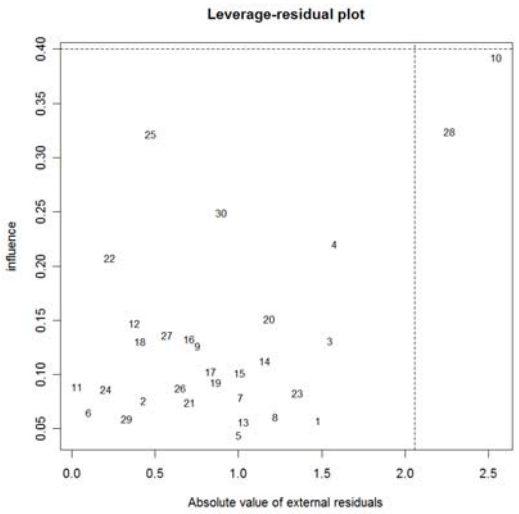
```
Call:
lm(formula = Winning.. ~ ERA + Batting.Average + Attendance,
    data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73558 -0.26648  0.06146  0.32221  0.70654

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.52926    0.65951   8.384 7.26e-09 ***
ERA             -0.72122    0.07831  -9.209 1.14e-09 ***
Batting.Average  0.42506    0.08032   5.292 1.56e-05 ***
Attendance       0.19031    0.08266   2.302   0.0296 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4082 on 26 degrees of freedom
```

Win % = .102 + (3.23)*Batting Average – (.115)*ERA + (1.73*10^-6)*Attendance


Residuals vs Fitted
Residuals
Fitted values
28
3
10


Normal Q-Q
Standardized residuals
Theoretical Quantiles
28
4
10


Squared residuals
Fitted values


Leverage-residual plot
influence
Absolute value of external residuals

2002

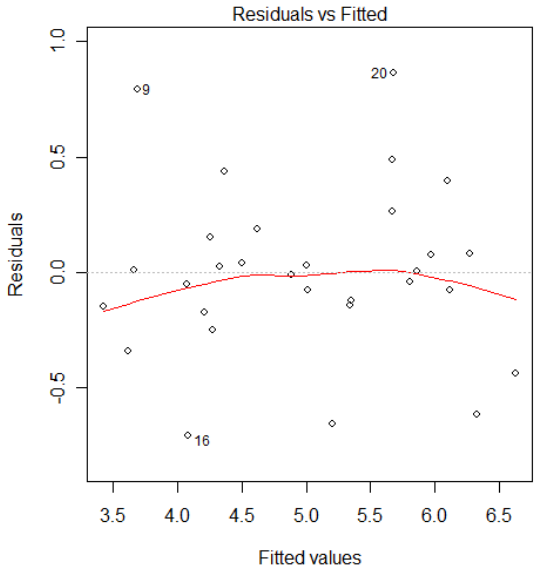
Coefficients:

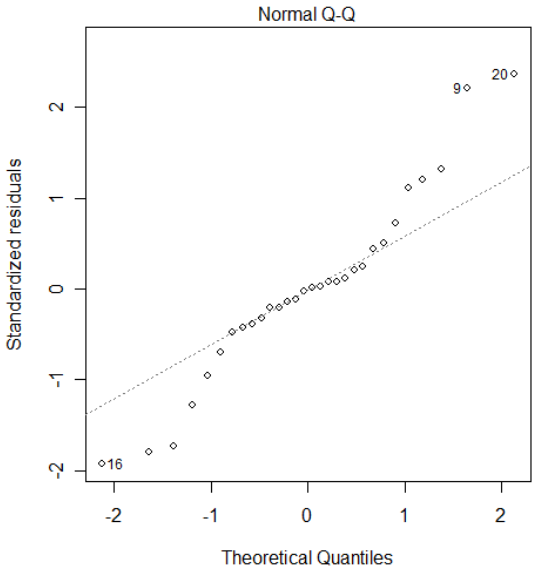
(Intercept):	5.9720
ERA(x_1):	-0.6609
Batting Average(x_2):	0.4665

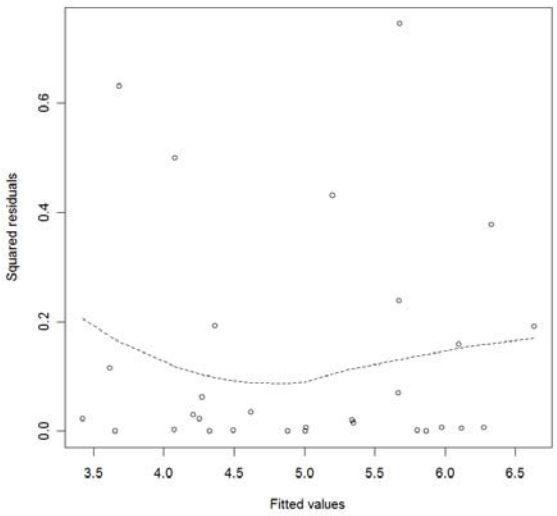
$Y = 5.9702 - .6609\ x_1 + .4665\ x_2$

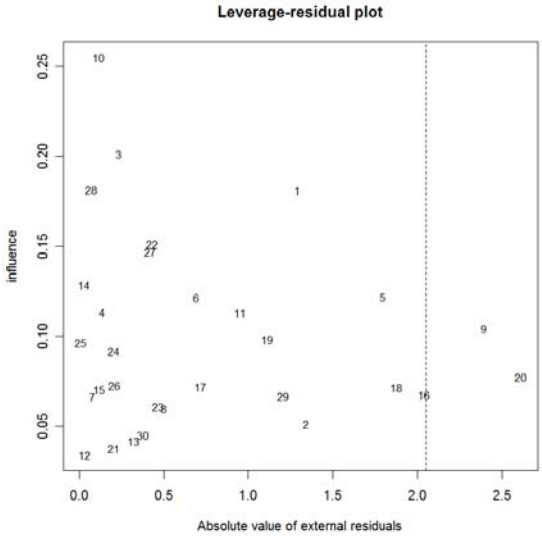
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.706890 -0.147382 -0.003908  0.134561  0.863346

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.97201    0.61882   9.651 3.03e-10 ***
ERA             -0.66091    0.07505  -8.807 2.01e-09 ***
Batting.Average  0.46651    0.07505   6.216 1.20e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3797 on 27 degrees of freedom
Multiple R-Squared: 0.8658,     Adjusted R-squared: 0.8559
F-statistic:  87.1 on 2 and 27 DF,  p-value: 1.677e-12
```

Win % = -.153 + (4.24)*Batting Average – (-.108)*ERA


Residuals vs Fitted
Residuals
Fitted values
9
20
16


Normal Q-Q
Standardized residuals
Theoretical Quantiles
9
20
16


Squared residuals
Fitted values


Leverage-residual plot
influence
Absolute value of external residuals

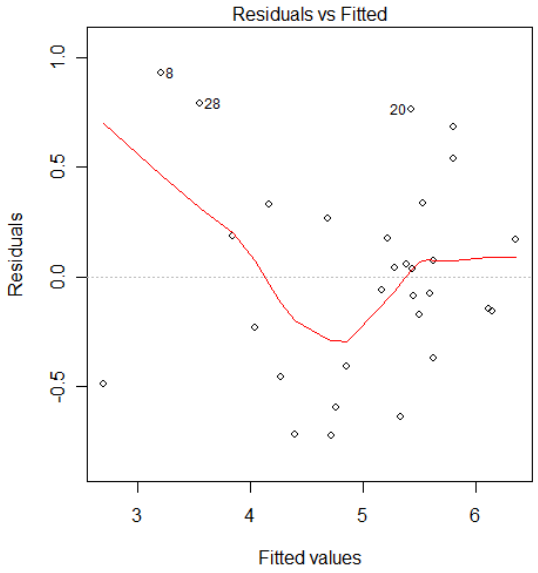
2003

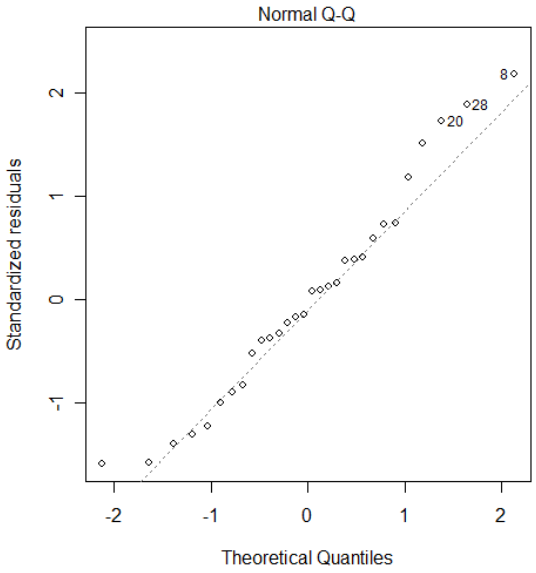
Coefficients:
(Intercept): 5.636
ERA(x_1): -0.7077
Batting Average(x_2): 0.580

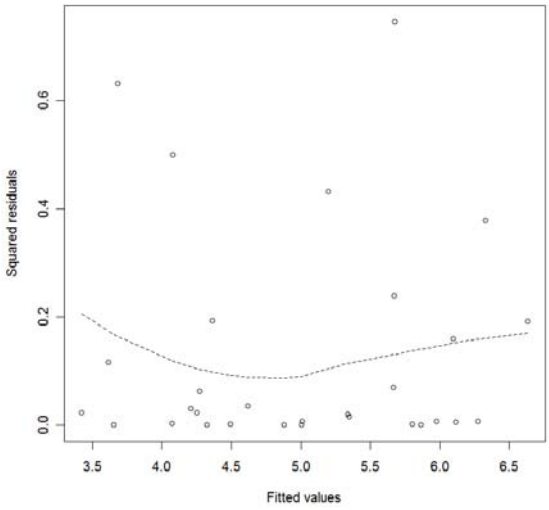
$Y = 5.636 - .7077\ x_1 + .580\ x_2$

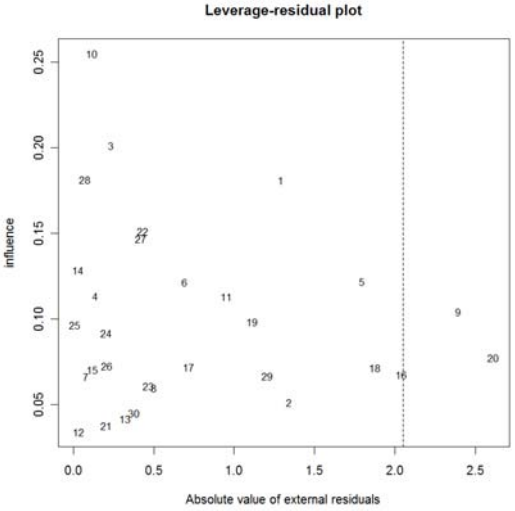
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.706890 -0.147382 -0.003908  0.134561  0.863346

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.97201    0.61882   9.651 3.03e-10 ***
ERA             -0.66091    0.07505  -8.807 2.01e-09 ***
Batting.Average  0.46651    0.07505   6.216 1.20e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3797 on 27 degrees of freedom
Multiple R-Squared: 0.8658,     Adjusted R-squared: 0.8559
F-statistic:  87.1 on 2 and 27 DF,  p-value: 1.677e-12
```

Win % = -.15 + (4.08)*Batting Average – (.099)*ERA


Residuals vs Fitted
Residuals
Fitted values
8
28
20


Normal Q-Q
Standardized residuals
Theoretical Quantiles
8
28
20


Squared residuals
Fitted values


Leverage-residual plot
influence
Absolute value of external residuals

2004

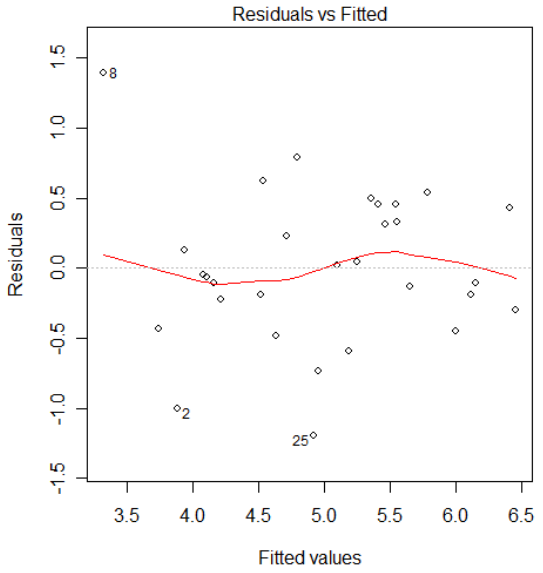
Coefficients:

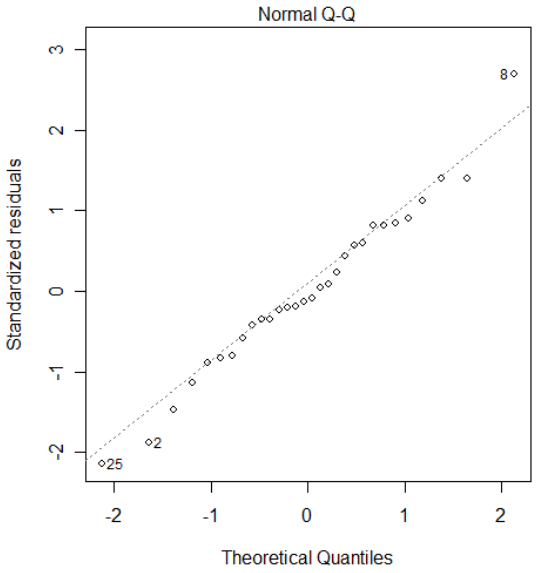
(Intercept):	4.2734
ERA(x_1):	-0.5301
Batting Average(x_2):	0.3979
Payroll (x_3):	0.2775

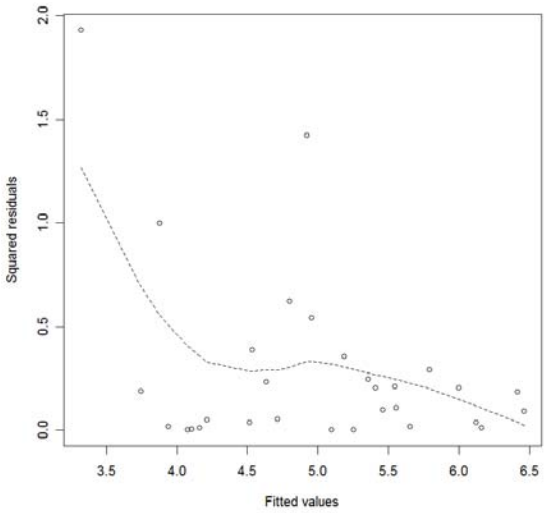
$Y = 4.2734 - .5301\ x_1 + .3979\ x_2 + 0.2775\ x_3$

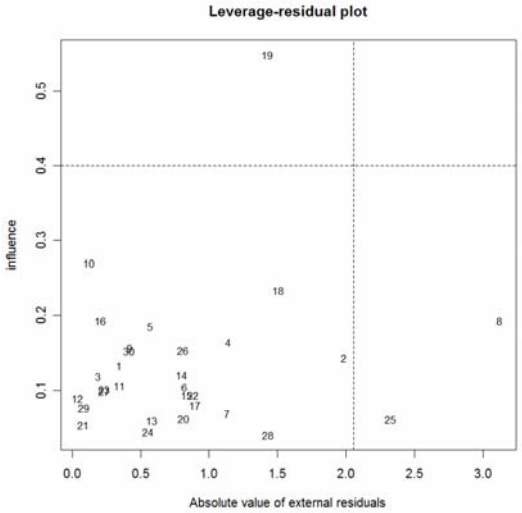
```
Residuals:
     Min       1Q   Median       3Q      Max
-1.19239 -0.28030 -0.05498  0.40319  1.38931

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.2734     0.9550   4.475 0.000134 ***
ERA              -0.5301     0.1093  -4.848 5.02e-05 ***
Batting.Average   0.3979     0.1128   3.528 0.001580 **
Payroll           0.2775     0.1149   2.416 0.023021 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.573 on 26 degrees of freedom
Multiple R-Squared: 0.7056,     Adjusted R-squared: 0.6716
F-statistic: 20.77 on 3 and 26 DF,  p-value: 4.45e-07
```

Win % = -.042 + (3.41)*Batting Average – (.095)*ERA + (7.07*10^-10)*Payroll


Residuals vs Fitted
Residuals
Fitted values
8
2
25


Normal Q-Q
Standardized residuals
Theoretical Quantiles
8
2
25


Squared residuals
Fitted values


Leverage-residual plot
influence
Absolute value of external residuals

2005

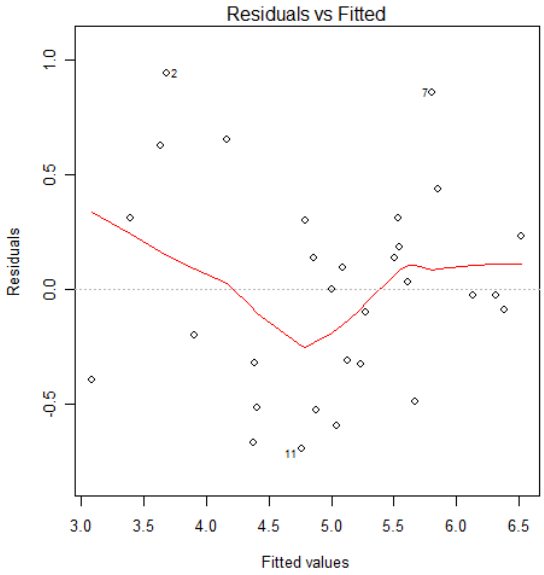
Coefficients:

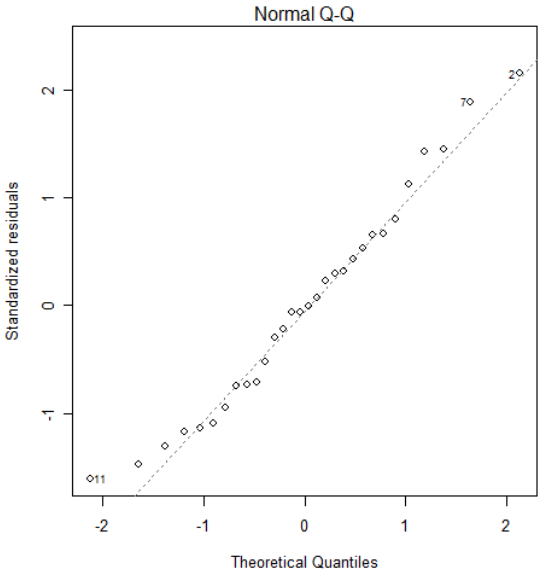
(Intercept):	6.71699
ERA(x_1):	-0.77258
Batting Average(x_2):	0.37448
Payroll (x_3):	0.22505
Dimensions(x_4):	0.17034

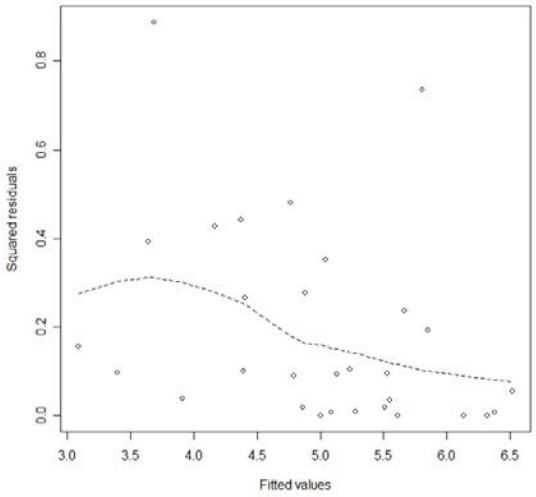
$Y = 6.71699 - 0.77258x_1 + 0.37448x_2 + 0.22505x_3 + 0.17034x_4$

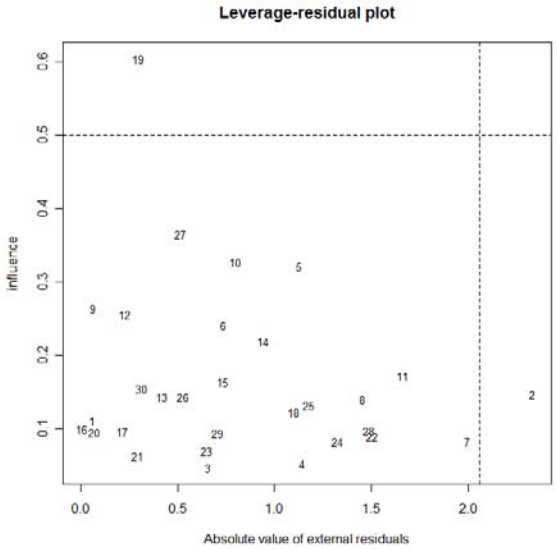
```
> summary(y)

Call:
lm(formula = Winning.. ~ ERA + Batting.Average, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93239 -0.29906 -0.03101  0.28227  1.27151

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.7233     0.6937   8.251 7.39e-09 ***
ERA              -0.6941     0.1010  -6.875 2.19e-07 ***
Batting.Average   0.5494     0.1010   5.442 9.31e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5421 on 27 degrees of freedom
Multiple R-Squared: 0.7264,     Adjusted R-squared: 0.7061
F-statistic: 35.83 on 2 and 27 DF,  p-value: 2.523e-08
```

Win % = .575 + (4.39*10^-10)*Payroll + (3.43)*Batting Average – (9.1*10^-2)*ERA – (1.76*10^-3)*Dimensions


Residuals vs Fitted
Residuals
Fitted values
2
7
11


Normal Q-Q
Standardized residuals
Theoretical Quantiles
2
7
11


Squared residuals
Fitted values


Leverage-residual plot
Influence
Absolute value of external residuals

2006

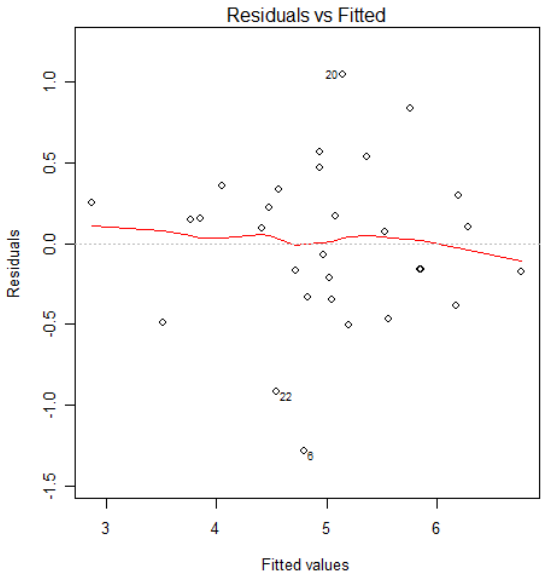
Coefficients:

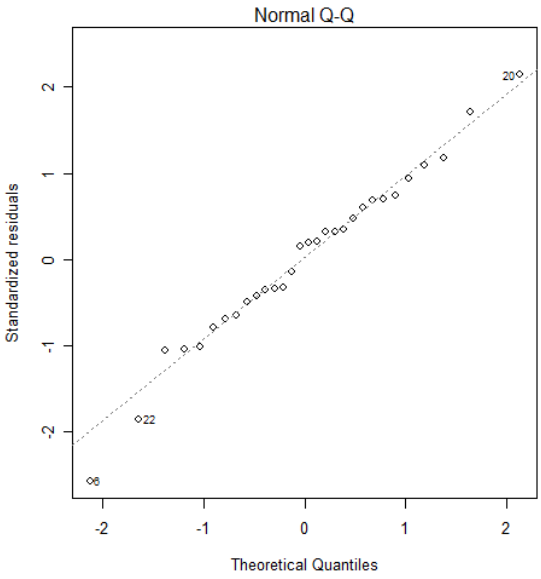
(Intercept):	5.61919
ERA(x_1):	-0.66522
Payroll (x_2):	0.33827
Batting Average(x_3):	0.20311

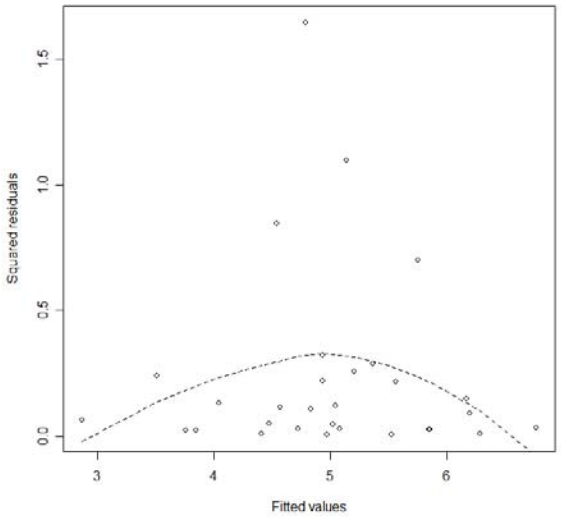
$Y = 5.61919 - 0.66522x_1 + 0.33827x_2 + 0.20311x_3$

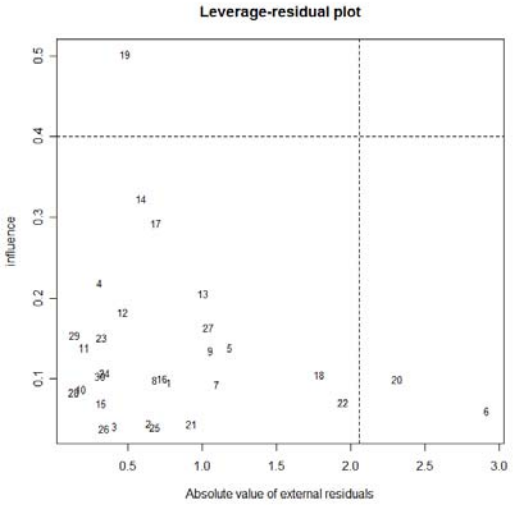
```
Call:
lm(formula = Winning.. ~ ERA + Payroll + Batting.Average, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.28357 -0.29908  0.08165  0.28859  1.04764

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.61919    0.83514   6.728 3.86e-07 ***
ERA             -0.66522    0.09759  -6.816 3.10e-07 ***
Payroll          0.33827    0.10413   3.249  0.00319 **
Batting.Average  0.20311    0.10282   1.975  0.05893 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5156 on 26 degrees of freedom
Multiple R-Squared: 0.7616,     Adjusted R-squared: 0.7341
F-statistic: 27.69 on 3 and 26 DF,  p-value: 2.961e-08
```

Win % = .532 + (1.41)*Batting Average – (.0102)*ERA + (6.52*10^-10)*Payroll


Residuals vs Fitted
Residuals
Fitted values
20
22
6


Normal Q-Q
Standardized residuals
Theoretical Quantiles
20
22
6


Squared residuals
Fitted values


Leverage-residual plot
Influence
Absolute value of external residuals

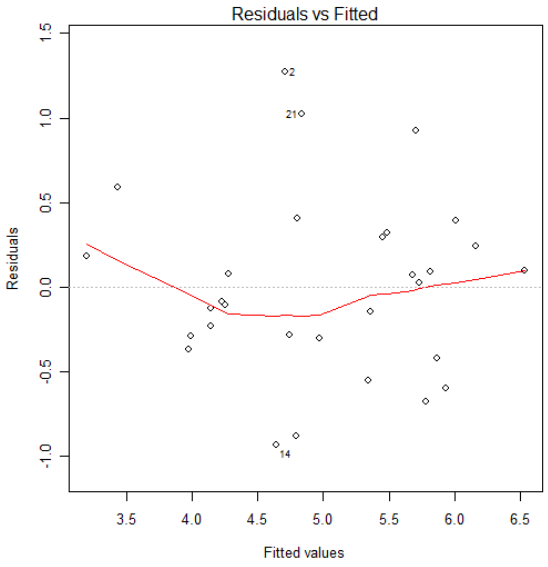
2007

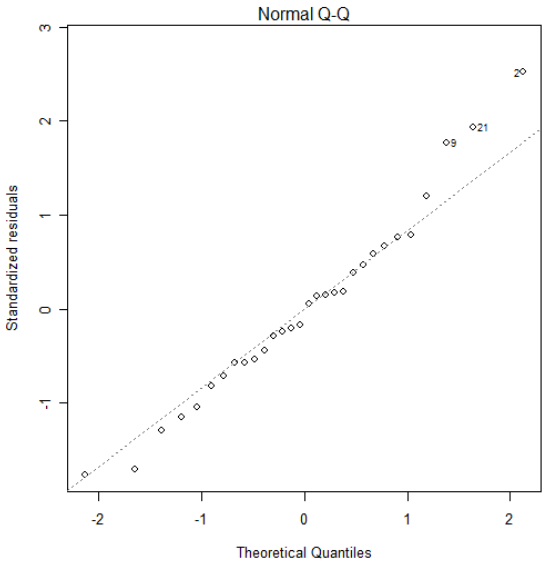
Coefficients:
(Intercept): 5.7233
ERA(x_1): -0.6941
Batting Average(x_2): 0.5494

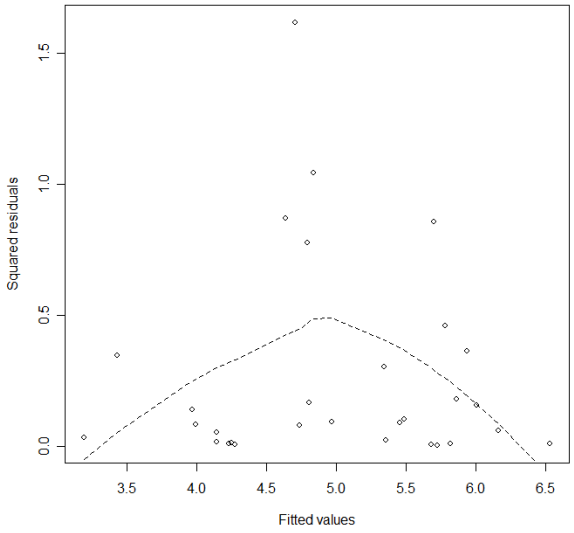
$Y = 5.7233 - 0.6941x_1 + 0.5494x_2$

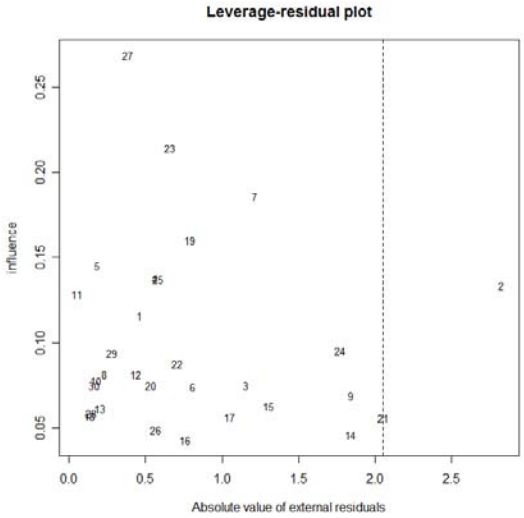
```
> summary(y)

Call:
lm(formula = Winning.. ~ ERA + Batting.Average, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93239 -0.29906 -0.03101  0.28227  1.27151

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.7233     0.6937   8.251 7.39e-09 ***
ERA              -0.6941     0.1010  -6.875 2.19e-07 ***
Batting.Average   0.5494     0.1010   5.442 9.31e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5421 on 27 degrees of freedom
Multiple R-Squared: 0.7264,     Adjusted R-squared: 0.7061
F-statistic: 35.83 on 2 and 27 DF,  p-value: 2.523e-08
```

Win % = .202 + (2.73)*Batting Average – (.097)*ERA


Residuals vs Fitted
Residuals
Fitted values
2
21
14
1.5
1.0
0.5
0.0
-0.5
-1.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5


Normal Q-Q
Standardized residuals
Theoretical Quantiles
20
21
9
3
2
1
0
-1
-2


Squared residuals
Fitted values
1.5
1.0
0.5
0.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5


Leverage-residual plot
Influence
Absolute value of external residuals
27
23
7
19
5
25
11
1
29
22
24
12
20
6
3
9
15
17
26
16
14
21
2
0.25
0.20
0.15
0.10
0.05
0.0
0.5
1.0
1.5
2.0
2.5