Big Data Final Project

Predicting the Probability of a San Antonio Spurs Win

By Jack Stubblefield

Overview

- What combinations of variables are better predictors of a Spurs win? How good of a predictor is difference in team salaries?
- Used box score statistics from each game in the 2016-2017 and 2017-2018 season
- For the logistic regressions, used only data from 2016-2017 season
- Compared results of five logistic regressions with different combinations of presumed significant variables for determining win probability in each game
  - Results include: number of correct predictions, classification rate, MSE, and a residual plot
- Will use one of the logistic regressions from the 2016-2017 season for the 2017-2018 season to see the accuracy of this model on exogenous data

Data Header

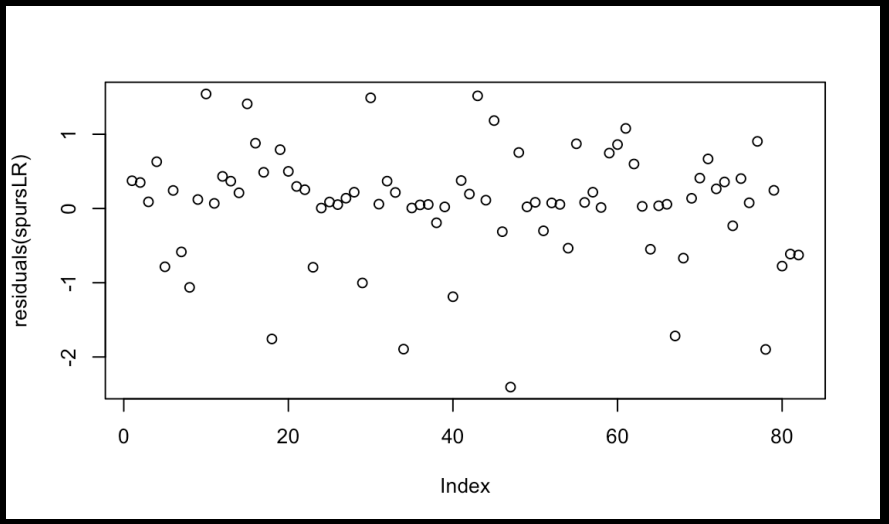
```
> head(spursdata)
  Season Win Result Game.Number Opponent Location FGM FGA FGPerc ThreePtM ThreePtA ThreePtPerc
1   2017   1    Win           1      GSW     Away  47  98  0.480       12       24        0.500
2   2017   1    Win           2      SAC     Away  36  79  0.456        6       18        0.333
3   2017   1    Win           3      NOP     Home  35  83  0.422       10       24        0.417
4   2017   1    Win           4      MIA     Away  37  82  0.451       10       18        0.556
5   2017   0   Loss           5      UTA     Home  33  76  0.434        6       20        0.300
6   2017   1    Win           6      UTA     Away  37  83  0.446        6       20        0.300
  FTM FTA FTPerc REB AST BLK STL TO Fouls OFGM O.FGA OFGPerc O3PM O.3PA O3PPerc OFTM O.FTA
1  23  26  0.885  55  25   3  13 13    19   40    85   0.471    7    33   0.212   13    18
2  24  27  0.889  40  23   8  10  9    26   28    70   0.400    6    20   0.300   32    38
3  18  23  0.783  50  21   7   6  9    15   32    86   0.372    4    22   0.182   11    15
4  22  26  0.846  44  20   5   4 15    26   37    80   0.463    7    20   0.350   18    26
5  19  21  0.905  34  19   5   6  9    15   38    76   0.500   15    31   0.484   15    18
6  20  22  0.910  49  17   8   9 11    25   30    80   0.375   11    26   0.423   15    24
  OFTPerc O.REB O.AST O.BLK O.STL O.TO O.Fouls Salary.Diff
1   0.722    35    24     6    11   16      19    -6.96503
2   0.842    40    22     0     7   15      21   -12.59753
3   0.733    45    14     5     6    9      19    -6.92801
4   0.692    36    19     6     7   12      24    -6.82222
5   0.833    39    22     8     5   10      20   -28.04243
6   0.625    42    15     4     7   12      19   -28.04243
```

Logistic Regression 1

```
> spursLR=glm(Result~FGA+REB+AST+O3PM+O.AST+Salary.Diff, data=spursdata, subset=1:82,family=binomial)
> summary(spursLR)

Call:
glm(formula = Result ~ FGA + REB + AST + O3PM + O.AST + Salary.Diff,
    family = binomial, data = spursdata, subset = 1:82)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4058  -0.1432   0.1005   0.3960   1.5425

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.04885    4.80031   0.427  0.66951
FGA         -0.22451    0.08133  -2.761  0.00577 **
REB          0.29797    0.10159   2.933  0.00336 **
AST          0.50737    0.13014   3.899 9.67e-05 ***
O3PM        -0.26518    0.13114  -2.022  0.04316 *
O.AST       -0.22610    0.10645  -2.124  0.03367 *
Salary.Diff -0.11330    0.04097  -2.766  0.00568 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93.305  on 81  degrees of freedom
Residual deviance: 46.836  on 75  degrees of freedom
AIC: 60.836

Number of Fisher Scoring iterations: 7
```

```
MyPredictions1617=predict(spursLR, type="response")
#compare actual wins with MyPredictions
cbind(Result[1:82]=="Win",MyPredictions1617[1:82])
#makes an array of "Loss" for all 82 games
PredictedResult1617=array("Loss",dim(spurs1617data)[1])
#replaces Loss by Win for games which MyPredictions >= 0.5
PredictedResult1617[MyPredictions1617 >=0.5]="Win"
PredictedResult1617
```

```
#displays the count of games on which PredictedResult==Result
sum(PredictedResult1617==Result[1:82])
#displays percentage of 82 games that are predicted correctly
mean(PredictedResult1617==Result[1:82])
#displays misclassification rate
1-mean(PredictedResult1617==Result[1:82])
#shows table of number of correctly and incorrectly predicted
#wins and losses; model predicted 62 wins, 56 of which were actually wins
```




```
> sum(PredictedResult1617==Result[1:82])
[1] 71
> #displays percentage of 82 games that are predicted correctly
> mean(PredictedResult1617==Result[1:82])
[1] 0.8658537
> #displays misclassification rate
> 1-mean(PredictedResult1617==Result[1:82])
[1] 0.1341463
>
> table(PredictedResult1617,Result[1:82])

PredictedResult1617 Loss Win
               Loss   15   5
               Win     6  56
```

```
> y=Win[1:82]
> modelpredict=MyPredictions1617[1:82]
> mean((y-modelpredict)^2)
[1] 0.09152417
```

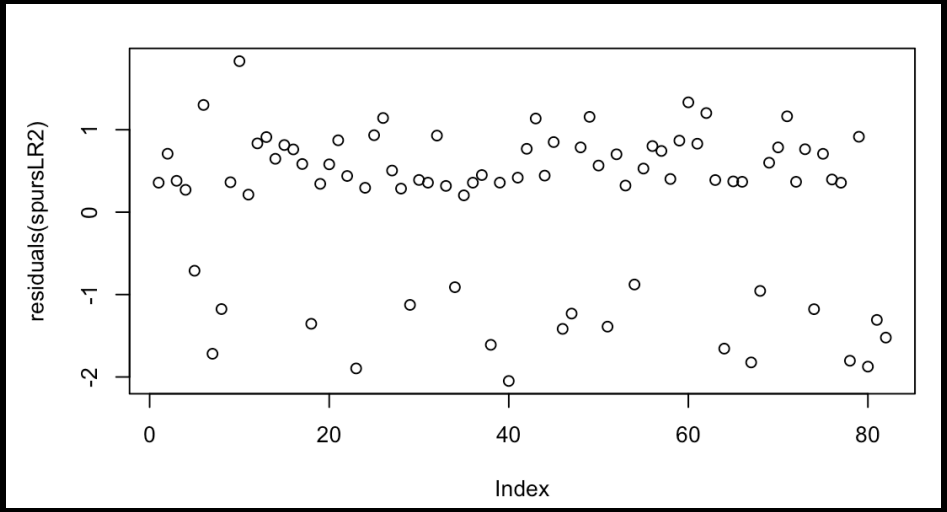
Logistic Regression 2

```
> spursLR2=glm(Result~ThreePtPerc+O3PM,data=spursdata,subset=1:82,family=binomial)
> summary(spursLR2)

Call:
glm(formula = Result ~ ThreePtPerc + O3PM, family = binomial,
    data = spursdata, subset = 1:82)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0485  -0.4824   0.3932   0.7813   1.8315

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.70091    1.43312  -0.489  0.62478
ThreePtPerc 10.20322    3.63089   2.810  0.00495 **
O3PM        -0.24040    0.09539  -2.520  0.01173 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93.305  on 81  degrees of freedom
Residual deviance: 77.121  on 79  degrees of freedom
AIC: 83.121

Number of Fisher Scoring iterations: 5
```

```
#displays the count of games on which PredictedResult==Result
sum(PredictedResult1617==Result[1:82])
#displays percentage of 82 games that are predicted correctly
mean(PredictedResult1617==Result[1:82])
#displays misclassification rate
1-mean(PredictedResult1617==Result[1:82])
#shows table of number of correctly and incorrectly predicted
#wins and losses; model predicted 72 wins,57 of which were actually wins
table(PredictedResult1617,Result[1:82])
```

```
> #displays the count of games on which PredictedResult==Result
> sum(PredictedResult1617==Result[1:82])
[1] 63
> #displays percentage of 82 games that are predicted correctly
> mean(PredictedResult1617==Result[1:82])
[1] 0.7682927
> #displays misclassification rate
> 1-mean(PredictedResult1617==Result[1:82])
[1] 0.2317073
> #shows table of number of correctly and incorrectly predicted
> #wins and losses; model predicted 62 wins,56 of which were actually wins
> table(PredictedResult1617,Result[1:82])

PredictedResult1617 Loss Win
               Loss    6   4
               Win    15  57
```

```
MyPredictions1617=predict(spursLR2, type="response")
#compare actual wins with MyPredictions
cbind(Result[1:82]=="Win",MyPredictions1617[1:82])
#makes an array of "Loss" for all 82 games
PredictedResult1617=array("Loss",dim(spurs1617data)[1])
#replaces Loss by Win for games which MyPredictions >= 0.5
PredictedResult1617[MyPredictions1617 >=0.5]="Win"
PredictedResult1617
```



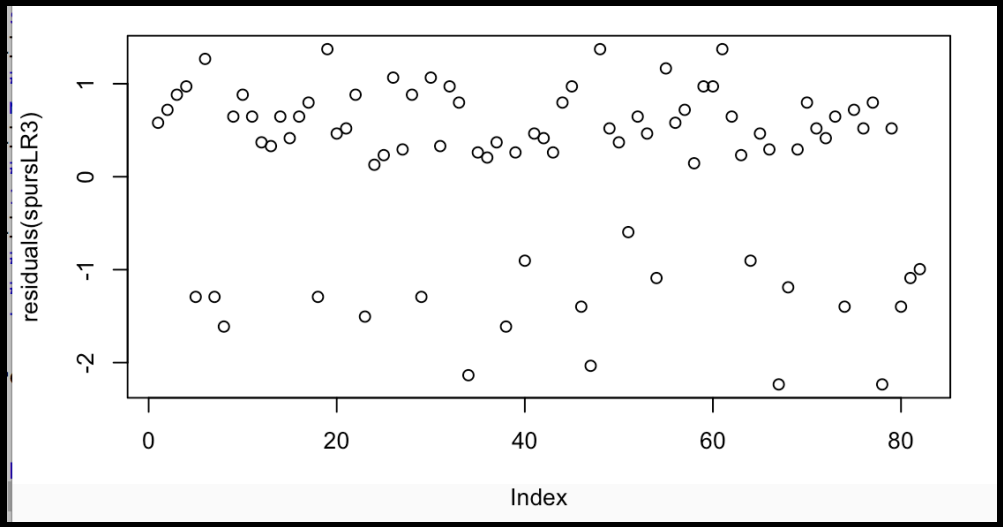

```
> y=Win[1:82]
> modelpredict=MyPredictions1617[1:82]
> mean((y-modelpredict)^2)
[1] 0.1575897
```

Logistic Regression 3

```
> spursLR3=glm(Result~AST,data=spursdata,subset=1:82,family=binomial)
> summary(spursLR3)

Call:
glm(formula = Result ~ AST, family = binomial, data = spursdata,
    subset = 1:82)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2344  -0.4149   0.4645   0.7782   1.3722

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.25706    1.50932  -2.821 0.004795 **
AST          0.23812    0.06976   3.413 0.000642 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93.305  on 81  degrees of freedom
Residual deviance: 76.529  on 80  degrees of freedom
AIC: 80.529

Number of Fisher Scoring iterations: 5
```

```
MyPredictions1617=predict(spursLR3, type="response")
#compare actual wins with MyPredictions
cbind(Result[1:82]=="Win",MyPredictions1617[1:82])
#makes an array of "Loss" for all 82 games
PredictedResult1617=array("Loss",dim(spurs1617data)[1])
#replaces Loss by Win for games which MyPredictions >= 0.5
PredictedResult1617[MyPredictions1617 >=0.5]="Win"
PredictedResult1617
```

```
#displays the count of games on which PredictedResult==Result
sum(PredictedResult1617==Result[1:82])
#displays percentage of 82 games that are predicted correctly
mean(PredictedResult1617==Result[1:82])
#displays misclassification rate
1-mean(PredictedResult1617==Result[1:82])
#shows table of number of correctly and incorrectly predicted
#wins and losses; model predicted 72 wins,57 of which were actually wins
table(PredictedResult1617,Result[1:82])
```




```
> #displays the count of games on which PredictedResult==Result
> sum(PredictedResult1617==Result[1:82])
[1] 63
> #displays percentage of 82 games that are predicted correctly
> mean(PredictedResult1617==Result[1:82])
[1] 0.7682927
> #displays misclassification rate
> 1-mean(PredictedResult1617==Result[1:82])
[1] 0.2317073
> #shows table of number of correctly and incorrectly predicted
> #wins and losses; model predicted 72 wins,57 of which were actually wins
> table(PredictedResult1617,Result[1:82])

PredictedResult1617 Loss Win
               Loss    6   4
               Win    15  57
```

```
> y=Win[1:82]
> modelpredict=MyPredictions1617[1:82]
> mean((y-modelpredict)^2)
[1] 0.1517373
```

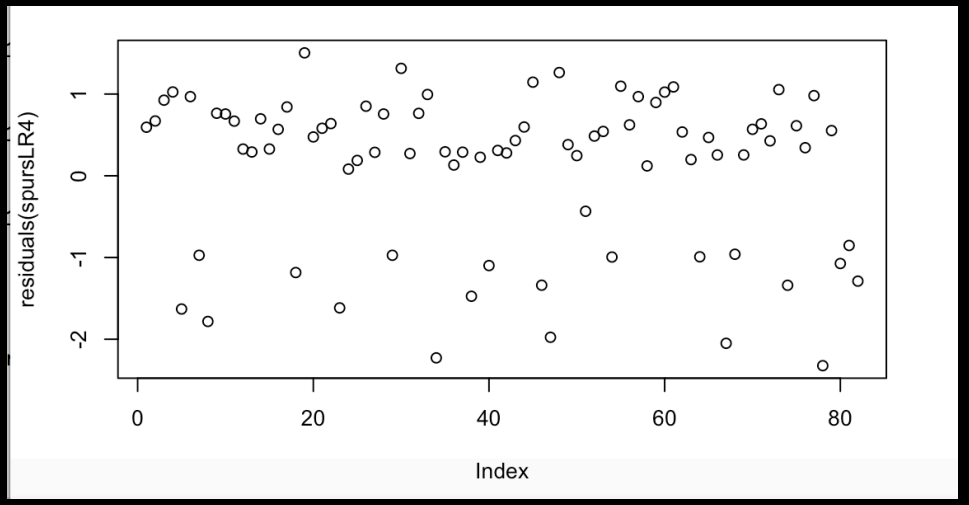
Logistic Regression 4

```
> spursLR4=glm(Result~AST+Salary.Diff, data=spursdata,subset=1:82,family=binomial)
> summary(spursLR4)

Call:
glm(formula = Result ~ AST + Salary.Diff, family = binomial,
    data = spursdata, subset = 1:82)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3237  -0.3053   0.4042   0.7411   1.5041

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.00645    1.66363  -3.009 0.002618 **
AST          0.25413    0.07441   3.415 0.000637 ***
Salary.Diff -0.04279    0.02525  -1.695 0.090163 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93.305  on 81  degrees of freedom
Residual deviance: 73.546  on 79  degrees of freedom
AIC: 79.546
```

```
#displays the count of games on which PredictedResult==Result
sum(PredictedResult1617==Result[1:82])
#displays percentage of 82 games that are predicted correctly
mean(PredictedResult1617==Result[1:82])
#displays misclassification rate
1-mean(PredictedResult1617==Result[1:82])
#shows table of number of correctly and incorrectly predicted
#wins and losses; model predicted 70 wins,58 of which were actually wins
table(PredictedResult1617,Result[1:82])
```

```
> #displays the count of games on which PredictedResult==Result
> sum(PredictedResult1617==Result[1:82])
[1] 67
> #displays percentage of 82 games that are predicted correctly
> mean(PredictedResult1617==Result[1:82])
[1] 0.8170732
> #displays misclassification rate
> 1-mean(PredictedResult1617==Result[1:82])
[1] 0.1829268
> #shows table of number of correctly and incorrectly predicted
> #wins and losses; model predicted 64 wins,54 of which were actually wins
> table(PredictedResult1617,Result[1:82])

PredictedResult1617 Loss Win
               Loss    9   3
               Win    12  58
```

```
MyPredictions1617=predict(spursLR4, type="response")
#compare actual wins with MyPredictions
cbind(Result[1:82]=="Win",MyPredictions1617[1:82])
#makes an array of "Loss" for all 82 games
PredictedResult1617=array("Loss",dim(spurs1617data)[1])
#replaces Loss by Win for games which MyPredictions >= 0.5
PredictedResult1617[MyPredictions1617 >=0.5]="Win"
PredictedResult1617
```




```
> y=Win[1:82]
> modelpredict=MyPredictions1617[1:82]
> mean((y-modelpredict)^2)
[1] 0.1446324
```

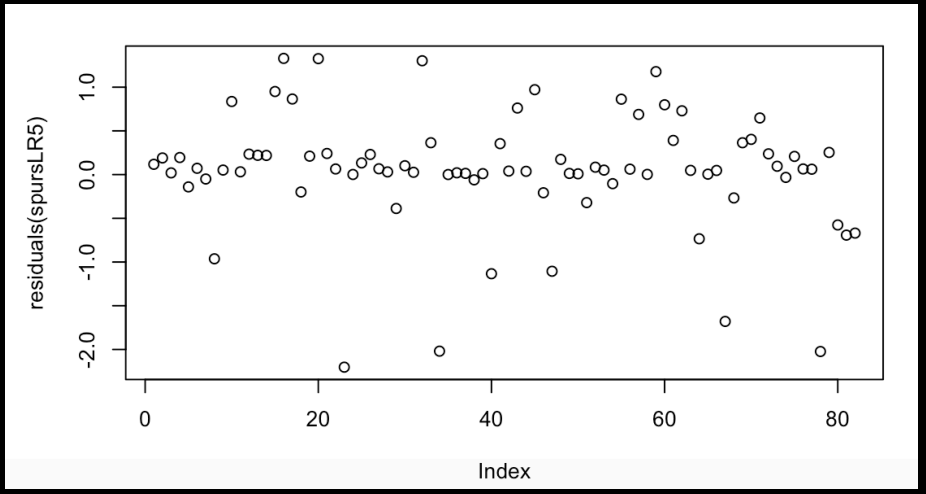
Logistic Regression 5

```
> spursLR5=glm(Result~REB+AST+ThreePtPerc+STL+TO+OFGPerc+O3PM+OFTPerc+O.AST+
spursdata,subset=1:82,family=binomial)
> summary(spursLR5)

Call:
glm(formula = Result ~ REB + AST + ThreePtPerc + STL + TO + OFGPerc +
    O3PM + OFTPerc + O.AST + Salary.Diff, family = binomial,
    data = spursdata, subset = 1:82)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.20248  -0.02399   0.06172   0.24040   1.32882

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  18.67321   10.82811   1.725  0.08462 .
REB           0.03316    0.07945   0.417  0.67636
AST           0.40854    0.14682   2.783  0.00539 **
ThreePtPerc  17.09797    8.45393   2.022  0.04313 *
STL           0.14758    0.20736   0.712  0.47664
TO           -0.02789    0.16887  -0.165  0.86881
OFGPerc     -48.02905   19.24481  -2.496  0.01257 *
O3PM         -0.24550    0.18512  -1.326  0.18478
OFTPerc      -8.93942    5.36891  -1.665  0.09591 .
O.AST        -0.23514    0.14288  -1.646  0.09981 .
Salary.Diff  -0.08504    0.05562  -1.529  0.12625
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93.305  on 81  degrees of freedom
Residual deviance: 36.187  on 71  degrees of freedom
AIC: 58.187
```

```
#displays the count of games on which PredictedResult==Result
sum(PredictedResult1617==Result[1:82])
#displays percentage of 82 games that are predicted correctly
mean(PredictedResult1617==Result[1:82])
#displays misclassification rate
1-mean(PredictedResult1617==Result[1:82])
#shows table of number of correctly and incorrectly predicted
#wins and losses; model predicted 61 wins,57 of which were actually wins
table(PredictedResult1617,Result[1:82])
```

```
> #displays the count of games on which PredictedResult==Result
> sum(PredictedResult1617==Result[1:82])
[1] 74
> #displays percentage of 82 games that are predicted correctly
> mean(PredictedResult1617==Result[1:82])
[1] 0.902439
> #displays misclassification rate
> 1-mean(PredictedResult1617==Result[1:82])
[1] 0.09756098
> #shows table of number of correctly and incorrectly predicted
> #wins and losses; model predicted 63 wins,58 of which were actually wins
> table(PredictedResult1617,Result[1:82])

PredictedResult1617 Loss Win
               Loss   17   4
               Win     4  57
```

```
MyPredictions1617=predict(spursLR5, type="response")
#compare actual wins with MyPredictions
cbind(Result[1:82]=="Win",MyPredictions1617[1:82])
#makes an array of "Loss" for all 82 games
PredictedResult1617=array("Loss",dim(spurs1617data)[1])
#replaces Loss by Win for games which MyPredictions >= 0.5
PredictedResult1617[MyPredictions1617 >=0.5]="Win"
PredictedResult1617
```




```
> y=Win[1:82]
> modelpredict=MyPredictions1617[1:82]
> mean((y-modelpredict)^2)
[1] 0.07053186
```

Finding Accuracy of Model for the 2017-2018 Season

- Chose logistic regression 1

```
MyPredictions1718=predict(spursLR, type="response",spursdata)
#compare actual wins with MyPredictions
cbind(Result[83:164]=="Win",MyPredictions1718[83:164])
#makes an array of "Loss" for all of the games from the 2017-2018 season
PredictedResult1718=array("Loss",dim(spursdata)[1])
#replaces Loss by Win for games which MyPredictions >= 0.5
PredictedResult1718[MyPredictions1718 >=0.5]="Win"
PredictedResult1718
```

```
> y2=Win[83:164]
> modelpredict2=MyPredictions1718[83:164]
> mean((y2-modelpredict2)^2)
[1] 0.2174843
```

```
#displays the count of games which our prediction
#equals the actual outcome for the game for 2018 using 2017 regression
sum(PredictedResult1718[83:164]==Result[83:164])
#displays percentage of 82 games from 2017-2018 season that are predicted correctly
mean(PredictedResult1718[83:164]==Result[83:164])
#displays misclassification rate
1-mean(PredictedResult1718[83:164]==Result[83:164])
#shows table of number of correctly and incorrectly predicted
#wins and losses; model predicted 44 wins,33 of which were actually wins
table(PredictedResult1718[83:164],Result[83:164])
```

```
> #displays the count of games which our prediction
> #equals the actual outcome for the game for 2018 using 2017 regression
> sum(PredictedResult1718[83:164]==Result[83:164])
[1] 57
> #displays percentage of 82 games from 2017-2018 season that are predicted correctly
> mean(PredictedResult1718[83:164]==Result[83:164])
[1] 0.695122
> #displays misclassification rate
> 1-mean(PredictedResult1718[83:164]==Result[83:164])
[1] 0.304878
> #shows table of number of correctly and incorrectly predicted
> #wins and losses; model predicted 44 wins,33 of which were actually wins
> table(PredictedResult1718[83:164],Result[83:164])

      Loss Win
  Loss   24  14
  Win    11  33
```

```
> #displays total count of games for both seasons
> #using regression from 2017 season data
> sum(PredictedResult1718==Result[1:164])
[1] 128
> #displays percentage of 164 games from both seasons that are predicted correctly
> mean(PredictedResult1718==Result[1:164])
[1] 0.7804878
> #displays misclassification rate for both seasons
> 1-mean(PredictedResult1718[1:164]==Result[1:164])
[1] 0.2195122
> #shows table of number of correctly and incorrectly predicted
> #wins and losses; model predicted 106 wins,89 of which were actually wins
> table(PredictedResult1718,Result[1:164])

PredictedResult1718 Loss Win
               Loss   39  19
               Win    17  89
```

Conclusion

```
> spursLR=glm(Result~FGA+REB+AST+O3PM+O.AST+Salary.Diff, data=spursdata, subset=1:82,family=binomial)
> summary(spursLR)

Call:
glm(formula = Result ~ FGA + REB + AST + O3PM + O.AST + Salary.Diff,
    family = binomial, data = spursdata, subset = 1:82)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4058  -0.1432   0.1005   0.3960   1.5425

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.04885    4.80031   0.427  0.66951
FGA         -0.22451    0.08133  -2.761  0.00577 **
REB          0.29797    0.10159   2.933  0.00336 **
AST          0.50737    0.13014   3.899 9.67e-05 ***
O3PM        -0.26518    0.13114  -2.022  0.04316 *
O.AST       -0.22610    0.10645  -2.124  0.03367 *
Salary.Diff -0.11330    0.04097  -2.766  0.00568 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93.305  on 81  degrees of freedom
Residual deviance: 46.836  on 75  degrees of freedom
AIC: 60.836

Number of Fisher Scoring iterations: 7
```

```
> spursLR5=glm(Result~REB+AST+ThreePtPerc+STL+TO+OFGPerc+O3PM+OFTPerc+O.AST+
spursdata,subset=1:82,family=binomial)
> summary(spursLR5)

Call:
glm(formula = Result ~ REB + AST + ThreePtPerc + STL + TO + OFGPerc +
    O3PM + OFTPerc + O.AST + Salary.Diff, family = binomial,
    data = spursdata, subset = 1:82)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.20248  -0.02399   0.06172   0.24040   1.32882

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  18.67321   10.82811   1.725  0.08462 .
REB           0.03316    0.07945   0.417  0.67636
AST           0.40854    0.14682   2.783  0.00539 **
ThreePtPerc  17.09797    8.45393   2.022  0.04313 *
STL           0.14758    0.20736   0.712  0.47664
TO           -0.02789    0.16887  -0.165  0.86881
OFGPerc     -48.02905   19.24481  -2.496  0.01257 *
O3PM         -0.24550    0.18512  -1.326  0.18478
OFTPerc      -8.93942    5.36891  -1.665  0.09591 .
O.AST        -0.23514    0.14288  -1.646  0.09981 .
Salary.Diff  -0.08504    0.05562  -1.529  0.12625
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 93.305  on 81  degrees of freedom
Residual deviance: 36.187  on 71  degrees of freedom
AIC: 58.187
```