Logistic Modelling with Baseball Data

Abstract

This paper investigates the application of statistical methods to analyze the performance of professional baseball teams. Logistic regression was used to model a binomial response variable, if a baseball team made it to the playoffs or didn't make it to the playoffs at the end of the regular season. While several factors were investigated initially through the course of this study it was found that the most influential were the number of runs and the number of strikeouts pitched per team. This model could be useful for baseball fans who wish to make estimates on a whether a team will go to the playoffs or not.

Introduction

The project studied the likelihood of a professional baseball team making it to the major league baseball playoffs. There are 30 Major League Baseball (MLB) teams in the United States and typically each team plays 162 games in a season. At the end of the regular season 10 teams will advance to the playoffs and the top two go on to play in the World Series. The actual scheme in which baseball teams make it to the playoffs is somewhat complicated, and is outside the scope of this statistical analysis project, so the working assumption was that the 10 best teams will proceed to the playoffs. The model constructed for this project used several qualitative parameters to predict the likelihood of a baseball team making it to the playoffs. The factors investigated were the total number of hits, the number of runs, the number of home runs, the number of errors committed and the total number of strike outs pitched per team for an entire season. Data from the 2015 baseball season was utilized for this study [1].

Methodology

Winning a baseball game requires that one team outperforms the other and ultimately scores more points than the opposing team. Scoring points requires hits and runs. To investigate how the previously stated factors influence the likelihood of a baseball team advancing to the playoffs data was gathered from every baseball team at mlb.com. From here the author proceeded to model the response, y, if the baseball team makes it to the playoffs, as a qualitative variable with two levels, also known as a binary variable. This approach requires the response to be modelled as a dummy variable which was set to $y = 1$ if the team made it to the playoffs in 2015 and $y = 0$ if the team did not make it to the playoffs in 2015 [2].

A general logistic model was created using the total number of hits, runs, home runs, errors committed and strike outs pitched as factors affecting the response variable. The statistical significance of the logistic model was determined by running the Chi square test to compare the general logistic model to the intercept model. The hypothesis test corresponding to the Chi square test is shown here: $Ho: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ and Ha: at least one $\beta$is different than zero. The Chi square test resulted in rejection of Ho and therefore at least one parameter has a significant effect on the response variable. However, after analyzing the general logistic model further it was found that none of the factors were statistically significant when$\alpha = 0.05$. Even though all of these factors could provide relevant information to a baseball team's ability to make it to the playoffs further investigation was needed.

Model selection was conducted by comparing the different AIC scores, using the backward elimination procedure, of all imbedded models within the set of the general logistic model. The model with the lowest AIC score was selected for further statistical analysis. As a result, some factors were eliminated from the logistic model, and the remaining factors were found to be statistically significant, as demonstrated by a low p-value less than 0.05 for both parameter coefficients resulting from the t-test. The refined model is much simpler, and is relatively easy to use which could make it more popular with baseball fans. The chosen model will use the total number of runs and number of strikeouts pitched in a season as quantitative parameters to predict the probability of the baseball team going to the playoffs.

Analysis

The logistic modelling used a binomial response variable as whether the baseball team makes it to the playoffs ($yes = 1, no = 0$) with the contributing factors as the total number of runs and the total number of strike outs pitched during the regular season [2]. This is the model selected based on the backward elimination procedure mentioned previously. The number of runs and the number of strikeouts pitched for the season were both found to be statistically significant with the t-test as verified by the low p-values of 0.0033 and 0.033 respectively. Therefore it is safe to conclude that both factors significantly affect the response variable and will be included with the model.

The model equation is shown below: $\log\left(\frac{\pi}{1-\pi}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$with coefficient estimates $\log\left(\frac{\pi}{1-\pi}\right) = -4.611 + 0.0042x_1 + 0.0016x_2$ where $x_1$is the number of runs and $x_2$is the number of strikeouts pitched. From interpretation of the coefficients it is possible to see that there is no realistic interpretation of the intercept as there are no baseball teams with zero runs or strikeouts pitched for an entire season. For every additional run achieved by the baseball team during the season the odds of reaching the playoffs increase by 0.42%, holding the number of strikeouts pitched constant. And, for every additional strikeout pitched by the baseball team during the season the odds of reaching the playoffs increase by 0.16%, holding the number of runs constant.

The author does concede that the percentages appear to be low, however, the reader must recall that there are 162 baseball games in a regular season. There will be at least several hundred hits and pitches per season. The average number of runs and strikeouts pitched from the gathered data are 688.267 and 1248.2 respectively [1]. Confidence intervals computed for the coefficients show that with 95% confidence the coefficient for the number of runs is between 0.001669 and 0.006889 as the coefficient for the number of strikeouts pitched is between 0.000199 and 0.00300. This result coincides with model equation including coefficient interpretations mentioned previously.

Results

Generally, the variables that had the highest influence on a baseball team's likelihood of going to the playoffs were runs and strikeouts pitched. The other variables considered at the beginning of the project were number of hits, number of errors committed and the number of home runs made per season. However, when all these factors were included in the model none of these factors were found to be statistically significant. Only after selecting a model by using the backward elimination process, which happened to eliminate three of the factors, was it possible to conclude that runs and strikeouts pitched are statistically significant.

Using the logistic model it is possible to predict the probability of any baseball team going to the playoffs by entering values for the number of runs and the number of strikeouts. From this method the author was able to make predictions for the likelihood of a baseball team with 900 runs and 1650 strikeouts pitched for the season. Using these parameter values the probability that the theoretical baseball team will go to the playoffs is 0.8679 or about 87%. Intuitively this estimate makes sense and may be somewhat conservative as both of these values exceed the averages found in the data.

The logistic model created for this project used data gathered from the 2015 season, recorded by the league [1]. MLB documents the performance of every player, every team and every season by recording data on several variables such as the number of home runs. This is a treasure trove of statistical data going back over 100 years. Future study could be done utilizing data from several seasons or even several decades. There is no doubt that many other factors could be investigated for future statistical analysis of baseball. While more complicated models may be used in the baseball industry the logistic model constructed for this project is simple and easy to use. This could be a fun way for baseball fans to predict how their favorite team will finish the season.

References

1. "Sortable Player Stats." *Cincinnati Reds*. N.p., n.d. Web. 05 May 2016.
2. "MLB Playoffs 2015: Bracket, Schedule, Scores and More." *SBNation.com*. N.p., 02 Nov. 2015. Web. 05 May 2016.

Appendix

R-Studio code

```
#Enter baseball data 2015 season
b<-Baseball
#null model
null.model<-glm(b$Playoffs~1,data=b,family = binomial)
#fit model including all factors investigated
fit<-glm(b$Playoffs~b$Runs+b$Hits+b$Home.Runs+b$Errors+b$Strikeouts..Pitched.)
anova(null.model,fit,test="Chisq")
#The model is useful as can be seen by low p-value
summary(fit)
#Even though the model is useful there aren't any significant factors
#will ry to select a better model using backward elimination
#this will simplify the model and one or more factors that remain may be significant
summary(step(fit,direction="backward"))
#New model as selected with backward elimination
fit2<-glm(b$Playoffs~b$Runs+b$Strikeouts..Pitched.)
summary(fit2)
#All remaining factors are statistically significant with new model
#investigate coefficients
exp(coef(fit2))
#compute confidence intervals
confint.default(fit2)
#display averages
mean(b$Runs)
mean(b$Strikeouts..Pitched.)
predict(fit,newdata=data.frame(Runs=800,Strikeouts..Pitched.=1500),type="response")
#create a log plot
plot(b$Runs,b$Playoffs, xlab="Runs",ylab="Probability of Playoffs")
plot(b$Strikeouts..Pitched.,b$Playoffs,xlab="Strikeouts Pitched",ylab="Probability of Playoffs")
#make predictions
pp<-coef(fit2)[1]+coef(fit2)[2]*800+coef(fit2)[3]*1500
pred<-exp(pp)/(1+exp(pp))
pred
pp1<-coef(fit2)[1]+coef(fit2)[2]*900+coef(fit2)[3]*1650
pred1<-exp(pp1)/(1+exp(pp1))
pred1
```

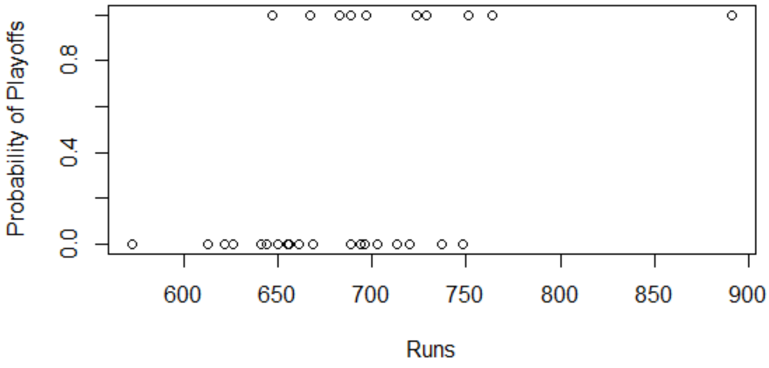
R-Studio output

```
> anova(null.model,fit,test="Chisq")
Analysis of Deviance Table
Model 1: b$Playoffs ~ 1
Model 2: b$Playoffs ~ b$Runs + b$Hits + b$Home.Runs + b$Errors + b$Strikeouts
..Pitched.
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        29     38.191
2        24      4.336  5   33.855 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

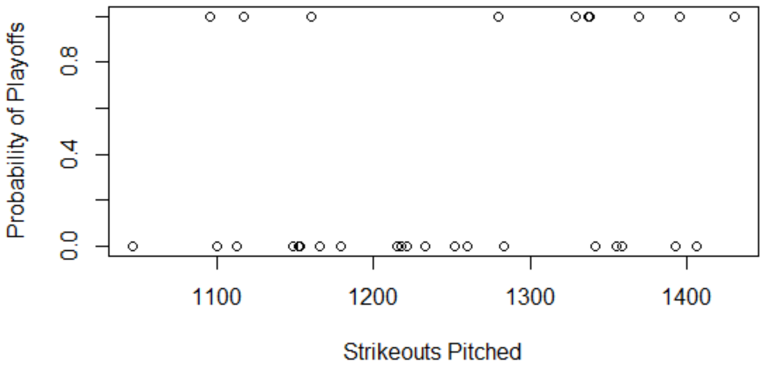
```
> summary(fit)
Call:
glm(formula = b$Playoffs ~ b$Runs + b$Hits + b$Home.Runs + b$Errors + 
    b$Strikeouts..Pitched.)
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.5771  -0.2968  -0.1072   0.2794   0.6972  
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)  
(Intercept)            -3.2767584  3.0951904  -1.059   0.3003  
b$Runs                  0.0058365  0.0028346   2.059   0.0505 .
b$Hits                 -0.0015552  0.0021910  -0.710   0.4847  
b$Home.Runs            -0.0023510  0.0047241  -0.498   0.6232  
b$Errors                0.0031678  0.0060621   0.523   0.6061  
b$Strikeouts..Pitched.  0.0014915  0.0008327   1.791   0.0859 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 0.1806481)
    Null deviance: 6.6667  on 29  degrees of freedom
Residual deviance: 4.3356  on 24  degrees of freedom
AIC: 41.106
Number of Fisher Scoring iterations: 2
> fit2<-glm(b$Playoffs~b$Runs+b$Hits+b$Errors+b$Strikeouts..Pitched.)
> summary(fit2)
Call:
glm(formula = b$Playoffs ~ b$Runs + b$Hits + b$Errors + b$Strikeouts..Pitched
.)
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.5536  -0.3081  -0.1008   0.3133   0.7372  
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)   
(Intercept)            -3.7883092  2.8752651  -1.318  0.19960   
b$Runs                  0.0046503  0.0015113   3.077  0.00501 **
b$Hits                 -0.0008981  0.0017220  -0.522  0.60656   
b$Errors                0.0042491  0.0055736   0.762  0.45298   
b$Strikeouts..Pitched.  0.0014266  0.0008099   1.761  0.09040 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 0.1752119)
    Null deviance: 6.6667  on 29  degrees of freedom
Residual deviance: 4.3803  on 25  degrees of freedom
AIC: 39.414
Number of Fisher Scoring iterations: 2
> summary(step(fit,direction="backward"))
Start:  AIC=41.11
b$Playoffs ~ b$Runs + b$Hits + b$Home.Runs + b$Errors + b$Strikeouts..Pitched
.

                         Df Deviance    AIC
- b$Home.Runs             1   4.3803 39.414
- b$Errors                1   4.3849 39.445
- b$Hits                  1   4.4266 39.729
<none>                        4.3356 41.106
- b$Strikeouts..Pitched.  1   4.9152 42.870
- b$Runs                  1   5.1014 43.986
```

```
Step:  AIC=39.41
b$Playoffs ~ b$Runs + b$Hits + b$Errors + b$Strikeouts..Pitched.

                         Df Deviance    AIC
- b$Hits                  1   4.4280 37.739
- b$Errors                1   4.4821 38.103
<none>                        4.3803 39.414
- b$Strikeouts..Pitched.  1   4.9239 40.923
- b$Runs                  1   6.0392 47.048
Step:  AIC=37.74
b$Playoffs ~ b$Runs + b$Errors + b$Strikeouts..Pitched.
                         Df Deviance    AIC
- b$Errors                1   4.5347 36.453
<none>                        4.4280 37.739
- b$Strikeouts..Pitched.  1   5.2768 41.000
- b$Runs                  1   6.1838 45.759

Step:  AIC=36.45
b$Playoffs ~ b$Runs + b$Strikeouts..Pitched.
                         Df Deviance    AIC
<none>                        4.5347 36.453
- b$Strikeouts..Pitched.  1   5.3760 39.559
- b$Runs                  1   6.2695 44.171
Call:
glm(formula = b$Playoffs ~ b$Runs + b$Strikeouts..Pitched.)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5466  -0.2989  -0.1181   0.3582   0.7139
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -4.6111769  1.4216537  -3.244  0.00314 **
b$Runs                  0.0042798  0.0013317   3.214  0.00338 **
b$Strikeouts..Pitched.  0.0016014  0.0007155   2.238  0.03366 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 0.1679507)
    Null deviance: 6.6667  on 29  degrees of freedom
Residual deviance: 4.5347  on 27  degrees of freedom
AIC: 36.453
Number of Fisher Scoring iterations: 2
> anova(null.model,fit,test="Chisq")
Analysis of Deviance Table
Model 1: b$Playoffs ~ 1
Model 2: b$Playoffs ~ b$Runs + b$Hits + b$Home.Runs + b$Errors + b$Strikeouts
..Pitched.
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        29     38.191
2        24      4.336  5   33.855 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> exp(coef(fit2))
           (Intercept)                b$Runs b$Strikeouts..Pitched.
           0.009940113           1.004289002            1.001602665
```

```
> plot(b$Runs,b$Playoffs, xlab="Runs",ylab="Probability of Playoffs")
```



```
> plot(b$Strikeouts..Pitched.,b$Playoffs,xlab="Strikeouts Pitched",ylab="Prob
ability of Playoffs")
```



```
> pp<-coef(fit2)[1]+coef(fit2)[2]*800+coef(fit2)[3]*1500
> pred<-exp(pp)/(1+exp(pp))
> pred
(Intercept)
  0.7711403
> pp1<-coef(fit2)[1]+coef(fit2)[2]*900+coef(fit2)[3]*1650
> pred1<-exp(pp1)/(1+exp(pp1))
> pred1
(Intercept)
  0.8679498
```