Final Project

Binary Logistic Regression Analysis White Sox

David Xu & Andy Chu

March 14, 2022

```
whitesox <- read_csv("C:/Users/David/Downloads/whitesox3.csv")
```

```
## Rows: 162 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (8): WL, R, RA, Hits, TB, SO, HomeAway, DayNight
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(whitesox)
```

Introduction

Our project goal is to be able to predict how well the Chicago Cubs and Chicago White Sox should perform during the 2021 season. Our logistic regression model will focus on predicting either of the Chicago teams winning a game, based on their game statistic performance during the season. To predict the response variable of the team winning a game,we plan to use what we think were the most suitable predictors, including runs scored, runs allowed, hits per game, errors per game, base on balls per game, if the game was played home or away, and if the game was during day or night. We will include all the 162 game statistics of both teams during the 2021 baseball season. Our data set will be collected from the MLB official website and the Baseball-Reference website.

Simple Logistic Regression

a. The predictor we chose was R(runs). Because it showed the best improvement in decrease of residual deviance.

```
modR <- glm(WL ~ R, family = "binomial", data = whitesox)
modRA <- glm(WL ~ RA, family = "binomial", data = whitesox)
modHits <- glm(WL ~ Hits, family = "binomial", data = whitesox)
modTB <- glm(WL ~ TB, family = "binomial", data = whitesox)
modSO <- glm(WL ~ SO, family = "binomial", data = whitesox)
modHA <- glm(WL ~ HomeAway, family = "binomial", data = whitesox)
modDN <- glm(WL ~ DayNight, family = "binomial", data = whitesox)
summary(modR)
```

```
##
## Call:
## glm(formula = WL ~ R, family = "binomial", data = whitesox)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3308  -0.6952   0.1291   0.5679   2.0867
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.8169     0.4949  -5.691 1.26e-08 ***
## R             0.7600     0.1231   6.172 6.73e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 221.01  on 161  degrees of freedom
## Residual deviance: 131.29  on 160  degrees of freedom
## AIC: 135.29
##
## Number of Fisher Scoring iterations: 6
```

```
#summary(modRA)
#summary(modHits)
#summary(modTB)
#summary(modHA)
#summary(modDN)
```

b. The slope, measuring the change in log(odds) for every unit change in Runs Scored, is 0.7600.

After calculating for the odds ratio, we see that the odds of winning a game increase by 2.138276 for every additional runs scored. From the regression output above, we see that the p-value is really small, therefore, we reject the null hypothesis of change in log(odds) being zero.

```
exp(0.7600)
```

```
## [1] 2.138276
```

c. The 95% CI for the slope is from 0.5424383 to 1.028789.

Also, the 95% confidence interval for the odds ratio suggests that we are 95% confident that the odds of winning increase by a factor of between 1.72019604 and 2.7976752 with an additional run scored. This 95% CI does not include 1, which means the coefficient of Runs scored would not be zero, winning a game depend on the amount of runs scored for the White Sox.

```
confint(modR)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %    97.5 %
## (Intercept) -3.8748837 -1.919570
## R            0.5424383  1.028789
```

```
exp(confint(modR))
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %    97.5 %
## (Intercept) 0.02075675 0.1466699
## R           1.72019604 2.7976752
```

d. Fitted model:

$$\widehat{log(Odds)} = -2.8169 + 0.7600(R)$$

We decided to plug in the following values for runs as we found them to be good intervals for testing the change of odds: 0, 5, 10, 15

$$\widehat{log(Odds)} = -2.8169 + 0.7600 * 0 = -2.8169$$
$$\frac{\pi}{1-\pi} = e^{-2.8169} = 0.05979101$$
$$\pi = \frac{e^{-2.8169}}{1+e^{-2.8169}} = 0.05641773$$

$$\widehat{log(Odds)} = -2.8169 + 0.7600 * 5 = 0.9831$$
$$\frac{\pi}{1-\pi} = e^{0.9831} = 2.672729$$
$$\pi = \frac{e^{0.9831}}{1+e^{0.9831}} = 0.7277229$$

$$\widehat{log(Odds)} = -2.8169 + 0.7600 * 10 = 4.7831$$
$$\frac{\pi}{1-\pi} = e^{4.7831} = 119.4741$$
$$\pi = \frac{e^{4.7831}}{1+e^{4.7831}} = 0.9916995$$

$$\widehat{log(Odds)} = -2.8169 + 0.7600 * 15 = 8.5831$$
$$\frac{\pi}{1-\pi} = e^{8.5831} = 5340.636$$
$$\pi == \frac{e^{8.5831}}{1+e^{8.5831}} = 0.9998128$$

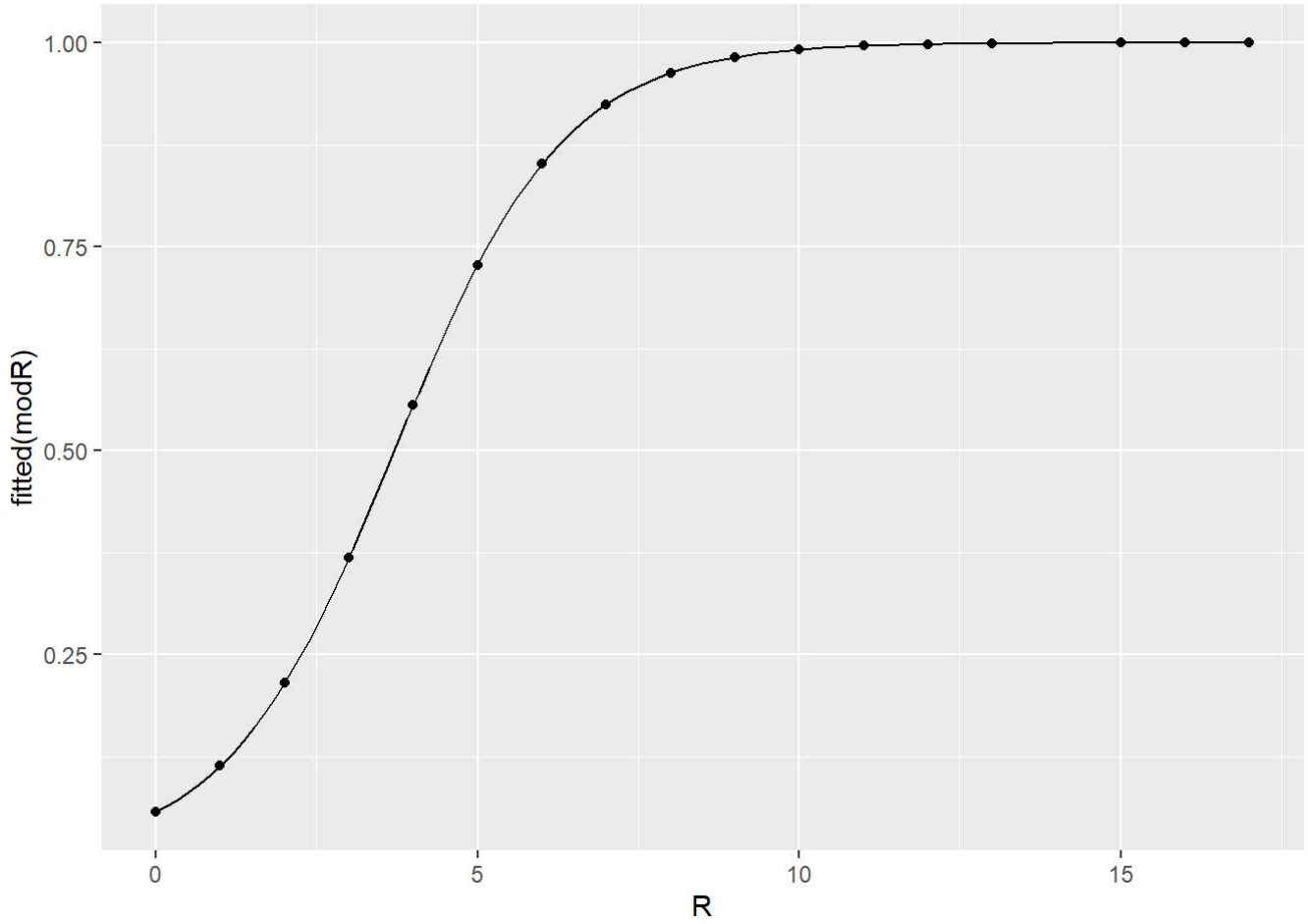
Comment:

We see that the probability of winning increases as the runs scored goes up. Interestingly, the winning probability is already at around 72% if the White Sox has 5 runs scored.

e.

From the graph we can see that the probability of winning a game increases drastically from scoring 1 to scoring 7. If White Sox has more than 7 runs, the probability of winning is more than 90%.

```
MODR_fun<- makeFun(modR)
gf_point(fitted(modR)~R, data = whitesox)%>%
  gf_fun(MODR_fun(R)~R, color = "Black")
```



f. The output below gives us the G-statistic of 89.72 with an extremely small P-value. Therefore, we can conclude that the model is very useful in predicting the response variable, in this case which is winning or losing a game.

```
anova(modR, test ="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: WL
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                   161     221.01
## R     1    89.72       160     131.29 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Logistic Regression

a.

The models with “Runs” as a base could not give us strong improvement after adding other predictor variables to the model. This is as expected, because there is strong multicollinearity between Runs and Hits or Total Base.

The multiple regression model we decided to use was the RA + TB + HomeAway. This model gives us one of the smallest residual deviance and the p-values for the predictor variables were all small to indicate all 3 variables to be significant.

```
modRA
```

```
##
## Call:  glm(formula = WL ~ RA, family = "binomial", data = whitesox)
##
## Coefficients:
## (Intercept)           RA
##      2.0700      -0.4416
##
## Degrees of Freedom: 161 Total (i.e. Null);  160 Residual
## Null Deviance:       221
## Residual Deviance: 177.6     AIC: 181.6
```

```
mod1 <- glm(WL~RA+TB, family = "binomial", data = whitesox)
mod2 <- glm(WL~RA+TB+HomeAway,family="binomial", data = whitesox)
mod3 <- glm(WL~RA+Hits+TB+HomeAway+DayNight,family="binomial", data = whitesox)


anova(modRA, mod1, mod2, mod3, test ="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: WL ~ RA
## Model 2: WL ~ RA + TB
## Model 3: WL ~ RA + TB + HomeAway
## Model 4: WL ~ RA + Hits + TB + HomeAway + DayNight
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       160    177.559
## 2       159     95.496  1   82.062  < 2e-16 ***
## 3       158     90.064  1    5.432  0.01977 *
## 4       156     89.966  2    0.099  0.95190
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#anova(modR, mod4, mod5, mod6, test ="Chisq")
summary(mod2)
```

```
##
## Call:
## glm(formula = WL ~ RA + TB + HomeAway, family = "binomial", data = whitesox)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.44381  -0.34590   0.03977   0.31116   2.57612
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.38890    0.75670  -3.157  0.00159 **
## RA          -1.02105    0.19052  -5.359 8.36e-08 ***
## TB           0.47581    0.08494   5.602 2.12e-08 ***
## HomeAway     1.24919    0.55651   2.245  0.02479 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 221.011  on 161  degrees of freedom
## Residual deviance:  90.064  on 158  degrees of freedom
## AIC: 98.064
##
## Number of Fisher Scoring iterations: 7
```

b.

Fitted model:

$$\widehat{log(Odds)} = -2.38890 - 1.02105RA + 0.47581TB + 1.24919HomeAway$$

We are interested to predict the probability of winning a game if the game is played at home with 5 runs against and 20 total bases. We are curious about this case because 20 total bases should allow the team to score more than 5 and win the game.

$$\widehat{log(Odds)} = -2.38890 - 1.02105 * 5 + 0.47581 * 20 + 1.24919 * 1$$

$$\widehat{log(Odds)} = 3.146969$$

$$\frac{\pi}{1-\pi} = e^{3.146969} = 23.26544$$

$$\pi = \frac{23.26544}{1 + 23.26544} = 0.9587875$$

The probability of winning under this condition is really high at around 0.9587, which is expected.

C.

Statistical and practical Significance

```
exp(-1.02105)
```

```
## [1] 0.3602165
```

```
exp(0.47581)
```

```
## [1] 1.609317
```

```
exp(1.24919)
```

```
## [1] 3.487517
```

```
sd(whitesox$RA)
```

```
## [1] 2.823054
```

```
exp(-1.40129*sd(whitesox$RA))
```

```
## [1] 0.0191411
```

```
sd(whitesox$TB)
```

```
## [1] 6.550038
```

```
exp(0.38474*sd(whitesox$TB))
```

```
## [1] 12.42936
```

To prove the significance of the predictor variables used for our Multi regression model, we could multiply the predictor odds ratio by the standard deviation. By multiplying the standard deviation we can see the change in the predictor variable for example if the Runs Against was higher how it impacted the winning possibility. We can see after calculating each one Runs Against, Total Base, and Home/Away, there was a clear change for each predictor variable, indicating that there is significance.

Prediction Table

a. and b.

```
whitesox2 <- whitesox%>%
  mutate(win_prob = fitted(mod2)) %>%
  mutate(success = ifelse(win_prob > 0.5 & WL == 1 | win_prob < 0.5 & WL == 0, "success", "failu
re"))

View(whitesox2)
```

c.

We accidentally entered one game statistic wrong for Win/Loss. The actual total wins for White Sox last year was 93, but we only received 92 wins in this case.

```
tally(~WL+success, data = whitesox2)
```

```
##    success
## WL  failure success
##   0       9      60
##   1      10      83
```

d. Success Rate, in terms of losses/wins/overalll

```
83/92
```

```
## [1] 0.9021739
```

```
60/70
```

```
## [1] 0.8571429
```

Comment:

The success rate of predicting White Sox wins using our model is 90.21%

The success rate of predicting White Sox loosing using our model is 85.71%

Conclusion

To conclude the White Sox results our Simple regression model and Multi regression model were both useful. We can see from our prediction table that the our model predicting the rate of White Sox winning or loosing is high. To improve on our model we could have included more statistic about pitchers. We found out that strike outs was not a useful predictor to our model. It allows people who are curious about the White Sox winning a game this year.