RedSox2018

Kevin Cummiskey

April 16, 2020

Chapter 6.1 Comparing Proportions - Are the Red Sox better at Fenway Park?

I'm a huge Boston Red Sox fan. Today we are going to investigate: *Are the Red Sox better at Fenway Park?*

Background

The Major League Baseball season consists of 162 games. Each team plays 81 games at home and 81 games on the road. The Red Sox play their home games at Fenway Park.

Let's consider the 2018 Red Sox (*why?*). You can find more baseball data than you could ever analyze at www.retrosheet.org. For today, we are going to use 2018 Retrosheet game logs. This data set contains one row for each game (2,431 total) played in the 2018 season with over 150 variables recorded for each game.

Get data from Retrosheet.org

If you want to follow along in R, go to https://www.retrosheet.org/gamelogs/index.html, download the 2018 file, unzip the file, and move it to your working directory in R. You can then run the code below.

```r
library(tidyverse)
library(knitr)

#---------------get retrosheet game log data----------------------#

website = "https://raw.githubusercontent.com/maxtoki/baseball_R/"
file = "master/data/game_log_header.csv"
glheaders <- read_csv(file = paste(website,file, sep = ""))
gamelogs2018 <- read_csv("GL2018.TXT",
                 col_names = names(glheaders),
                 na = character())

#Here is a small sample of the data.
```

```r
gamelogs2018 %>%
  select(Date, VisitingTeam, VisitorRunsScored, HomeTeam, HomeRunsScore) %>%
  head(10)
```

```
## # A tibble: 10 x 5
##        Date VisitingTeam VisitorRunsScored HomeTeam HomeRunsScore
##       <dbl> <chr>                    <dbl> <chr>            <dbl>
##  1 20180329 COL                          2 ARI                  8
##  2 20180329 PHI                          5 ATL                  8
##  3 20180329 SFN                          1 LAN                  0
##  4 20180329 CHN                          8 MIA                  4
##  5 20180329 SLN                          4 NYN                  9
##  6 20180329 MIL                          2 SDN                  1
##  7 20180329 MIN                          2 BAL                  3
##  8 20180329 CHA                         14 KCA                  7
##  9 20180329 ANA                          5 OAK                  6
## 10 20180329 CLE                          1 SEA                  2
```

Data Wrangling

Let's do a little data wrangling.

```r
#  Add a variable called WinningTeam.
gamelogs2018 %>%
  mutate(WinningTeam = case_when(HomeRunsScore > VisitorRunsScored ~
                                   HomeTeam,
                                 VisitorRunsScored > HomeRunsScore ~
                                   VisitingTeam)) -> gamelogs2018

# Let's look at only games the Red Sox were involved in.
gamelogs2018 %>%
  filter(HomeTeam == "BOS" | VisitingTeam == "BOS" ) -> redsox

# add Result (W/L) and Field (Home/Away) variables
redsox %>%
  mutate(Result = ifelse(WinningTeam == "BOS", "W","L"),
         Field = ifelse(HomeTeam == "BOS", "Home", "Away"),
         RedSoxRunsScored = ifelse(HomeTeam == "BOS",
                                   HomeRunsScore,
                                   VisitorRunsScored),
         RedSoxStartingPitcherName = ifelse(HomeTeam == "BOS",
                                            HomeStartingPitcherName,
                                            VisitorStartingPitcherName)) -> redsox

# view sample of the redsox data frame.
redsox %>%
  select(Date, VisitingTeam, VisitorRunsScored, HomeTeam, HomeRunsScore, Result, Field) %>%
  head(10) %>%
  kable(caption = "First 10 games of the Red Sox 2018 season")
```

Table 1: First 10 games of the Red Sox 2018 season

Date	VisitingTeam	VisitorRunsScored	HomeTeam	HomeRunsScore	Result	Field
20180329	BOS	4	TBA	6	L	Away
20180330	BOS	1	TBA	0	W	Away
20180331	BOS	3	TBA	2	W	Away
20180401	BOS	2	TBA	1	W	Away
20180402	BOS	7	MIA	3	W	Away
20180403	BOS	4	MIA	2	W	Away
20180405	TBA	2	BOS	3	W	Home
20180407	TBA	3	BOS	10	W	Home
20180408	TBA	7	BOS	8	W	Home
20180410	NYA	1	BOS	14	W	Home

Study Design

Recall our research question: *Are the Red Sox better at Fenway Park?* In other words, is there an association between playing at Fenway Park and the Red Sox winning?

What are the explanatory and response variables? Categorize each as categorical or quantitative.

How does this study differ from simple linear regression and two-sample *t*-tests?

```r
summary = redsox %>%
  group_by(Field)%>%
  count(Result) %>%
  spread(key = Field, value = n)
kable(summary, caption = "Results of the Red Sox 2018 Season")
```

Table 2: Results of the Red Sox 2018 Season

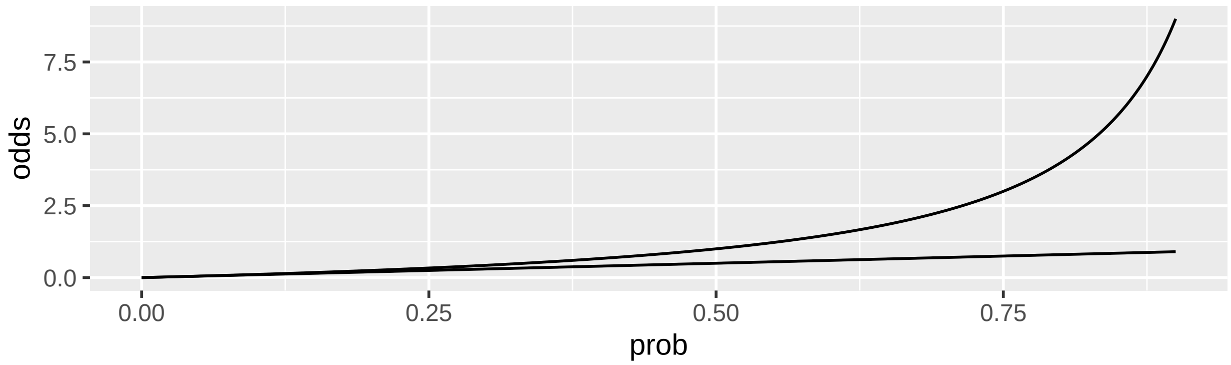
Result	Away	Home
L	30	24
W	51	57

Calculate the *conditional proportions* of wins for home games and away games. (You will also hear conditional proportions referred to as *chances, likelihood, risk*).

Calculate the *odds* of winning at home and away.

Describe the relationship between odds and probabilty. The graph below might help.

```
measures = data.frame(prob = seq(0,.90, by = 0.001))
measures = measures %>% mutate(odds = prob/(1-prob))
measures %>% ggplot(aes(x = prob, y = odds)) +
  geom_line() +
  geom_line(aes(y = prob))
```



Measures of Association

Next, we will calculate various measures of association between playing at Fenway Park and winning. We will discuss four measures of association: *risk difference, relative risk, odds ratio, log odds ratio.*

- Calculate the difference in conditional proportions (also called *risk difference*) comparing home games to away games.

- Calculate the *relative risk* for a win comparing home and away games. How does the risk difference and relative risk tell us something different?

- Calculate the *odds ratio* for wins comparing home and away games.

- Calculate the *log odds ratio* for winning comparing home and away games.

For each measure of association above, describe the range of possible values.

- Risk difference
- Relative risk
- Odds ratio
- Log odds ratio

Inference on Difference in Proportions

Write the null and alternative hypotheses for this test.

What is the statistic of interest for this test?

Theory-based test (two sample z-test)

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Given a sufficiently large sample, z is approximately standard normal under the null hypothesis.

```
# two-sample z-test
phat_home = 57/81
phat_away = 51/81
phat = 108/162
#standardized statistic (pg 420)
z = (phat_home - phat_away)/sqrt(phat*(1-phat)*(1/81 + 1/81))
#p-value
2*(1-pnorm(z,0,1))
```

```
## [1] 0.3173105
```

Theory-based test (χ^2 test)

Fill in the expected values in the table below if home/away has no effect and the Red Sox won 108 games.

Result	Away	Home	Total
W			108
L			54
Total	81	81	182

The χ^2 test compares the observed counts in each cell to the expected counts.

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Calculate the χ^2 statistic.

```
#calculate overall win/loss percentage
summary_overall = redsox %>% group_by(Result) %>%
  count() %>% group_by() %>% mutate(perc = n/sum(n)) %>%
  select(-n)
#calculate expected wins
summary_homeaway = redsox %>%
  group_by(Field,Result) %>%
  count() %>%
  left_join(summary_overall, by = "Result") %>%
  group_by(Field) %>%
  mutate(expected = perc*sum(n))
summary_homeaway
```

```
## # A tibble: 4 x 5
## # Groups:   Field [2]
##   Field Result     n  perc expected
##   <chr> <chr>  <int> <dbl>    <dbl>
## 1 Away  L         30 0.333       27
## 2 Away  W         51 0.667       54
## 3 Home  L         24 0.333       27
## 4 Home  W         57 0.667       54
```

```r
#calculate chi-square statistic
chisq = summary_homeaway %>% group_by() %>%
  summarise(chisq = sum((n - expected)^2/expected))

1- pchisq(chisq$chisq, 1)
```

```
## [1] 0.3173105
```

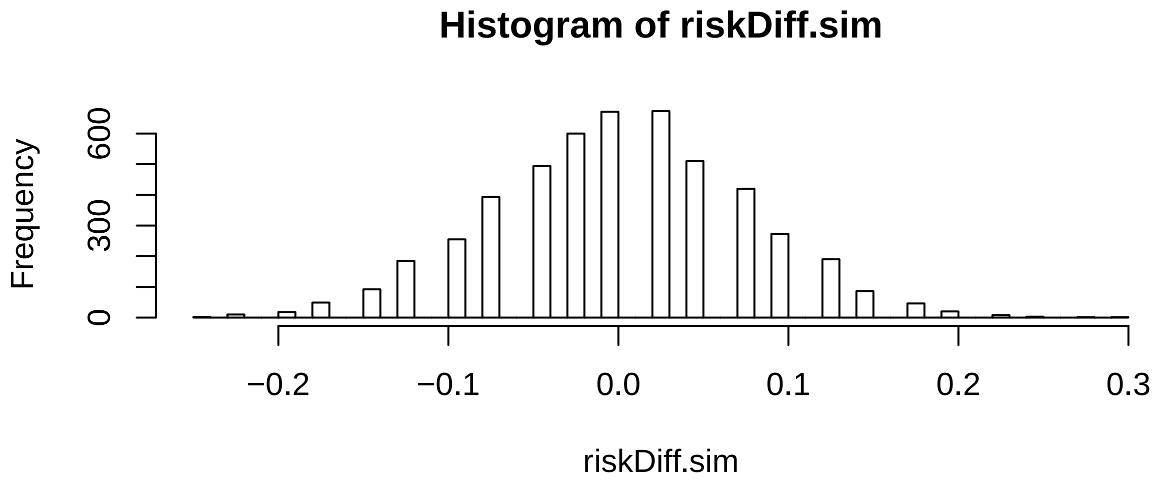
Simulation-based test

Let’s say I gave you 162 playing cards (one for each game) consisting of 81 red cards (home games) and 81 blue cards (away games). Explain how you could conduct a simulation-based test with these cards.

```r
redsox.sim = redsox %>% select(Result, Field)
riskDiff.sim = c()
n.sims = 5000

for(i in 1:n.sims){
  summary.sim = redsox.sim %>%
    mutate(Result.sim = sample(Result)) %>% #shuffle wins
    group_by(Field) %>%
    count(Result.sim) %>% mutate(p = n/sum(n)) #calculate win percentages
  riskDiff.sim[i] = summary.sim$p[3]-summary.sim$p[1]
}

hist(riskDiff.sim, breaks = 50)
```



```r
sum(abs(riskDiff.sim) > (phat_home - phat_away))/n.sims
```

```
## [1] 0.2478
```

What would we conclude from these tests?

Is confounding an issue in this analysis? List variables that are potential confounding variables of the association between playing at Fenway Park and winning.

Intro to Logistic Regression

Here, we are going to repeat the analysis above using logistic regression. For this case, the results will be the same (*why?*). However, when we introduce more complex models, we cannot use the methods above – we have to use logistic regression.

Let Y_i be whether or not the Red Sox win game i such that $Y_i \sim \text{Bernoulli}(\pi_i)$ be the probability the Red Sox win game i.

Here is our model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\text{Field}_i \tag{1}$$

where Field_i is whether game i was played on the home or away field.

Describe π_i in words.

Solve Equation 1 for π_i.

What is another name for the left side of Equation 1?

In regression models, we include a random error term (ϵ) in the linear model. Equation 1 does **not** have a random error term. *Why?*

According the the model, what are the log odds of the Red Sox winning on the road?

According to the model, what are the log odds ofthe Red Sox winning at Fenway Park?

How do we interpret β_1 in the model?

What values of β_1 would indicate the Red Sox are more likely to win at Fenway Park then on the road?

```r
#reverse factor levels for result
#so win is 1 and loss is 0
redsox$Result = factor(redsox$Result,
                       levels = c("L","W"))
model_homeaway = glm(Result ~ Field,
                     data = redsox,
                     family = "binomial")
summary(model_homeaway)
```

```
##
## Call:
## glm(formula = Result ~ Field, family = "binomial", data = redsox)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5597  -1.4094   0.8383   0.9619   0.9619
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5306     0.2301   2.306   0.0211 *
## FieldHome     0.3344     0.3349   0.998   0.3181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 206.23  on 161  degrees of freedom
## Residual deviance: 205.23  on 160  degrees of freedom
## AIC: 209.23
##
## Number of Fisher Scoring iterations: 4
```

Report estimates of the following:

- log odds ratio for winning comparing games at Fenway Park to road games.
- odds ratio for winning comparing games at Fenway Park to road games.
- probability of the Red Sox winning on the road.
- probability of the Red Sox winning at Fenway Park.

Have we seen these estimates before?

What would we conclude from these results?

Runs Scored and Red Sox winning

In this section, we will investigate: *how does scoring affect the Red Sox's probability of winning?*.

What are the explanatory and response variables? Categorize each as categorical or quantitative.

How does this study differ from the investigation of home/away and winning?

Consider the following model: let Y_i be whether or not the Red Sox win game i such that $Y_i \sim \text{Bernoulli}(\pi_i)$ be the probability the Red Sox win game i.

Here is our model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{RunsScored}_i \qquad (2)$$

where RunsScored_i is the number of runs the Red Sox scored in game i.

According to the model, what are the log odds of a Red Sox win if they score 5 runs? 4 runs?

According to the model, what is the probability the Red Sox win if they score 5 runs?

How do we interpret β_1? β_0?

Let's fit the model.

```r
#restrict analysis to between 1 and 10 runs scored
model.runsScored <- glm(Result == "W" ~ RedSoxRunsScored,
                        family = "binomial",
                        data = redsox)
summary(model.runsScored)
```

```
##
## Call:
## glm(formula = Result == "W" ~ RedSoxRunsScored, family = "binomial",
##     data = redsox)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1884  -0.6537   0.2388   0.7732   1.8156
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.0568     0.4644  -4.428 9.49e-06 ***
## RedSoxRunsScored   0.6222     0.1061   5.862 4.57e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 206.23  on 161  degrees of freedom
## Residual deviance: 142.21  on 160  degrees of freedom
## AIC: 146.21
##
## Number of Fisher Scoring iterations: 6
```

Estimate the log odds of a Red Sox win if they score 5 runs? 4 runs?

Estimate the probability the Red Sox win if they score 5 runs?

```
winProbs = tibble(RedSoxRunsScored = 0:15)
winProbs %>% mutate(pred.win.perc = predict(model.runsScored,
                                           newdata = winProbs,
                                           type = "response")) -> winProbs
winProbs
```

```
## # A tibble: 16 x 2
##    RedSoxRunsScored pred.win.perc
##               <int>         <dbl>
##  1                0         0.113
##  2                1         0.192
##  3                2         0.307
##  4                3         0.453
##  5                4         0.606
##  6                5         0.742
##  7                6         0.842
##  8                7         0.909
##  9                8         0.949
## 10                9         0.972
## 11               10         0.985
## 12               11         0.992
## 13               12         0.996
## 14               13         0.998
## 15               14         0.999
## 16               15         0.999
```

What do we conclude based on this model?

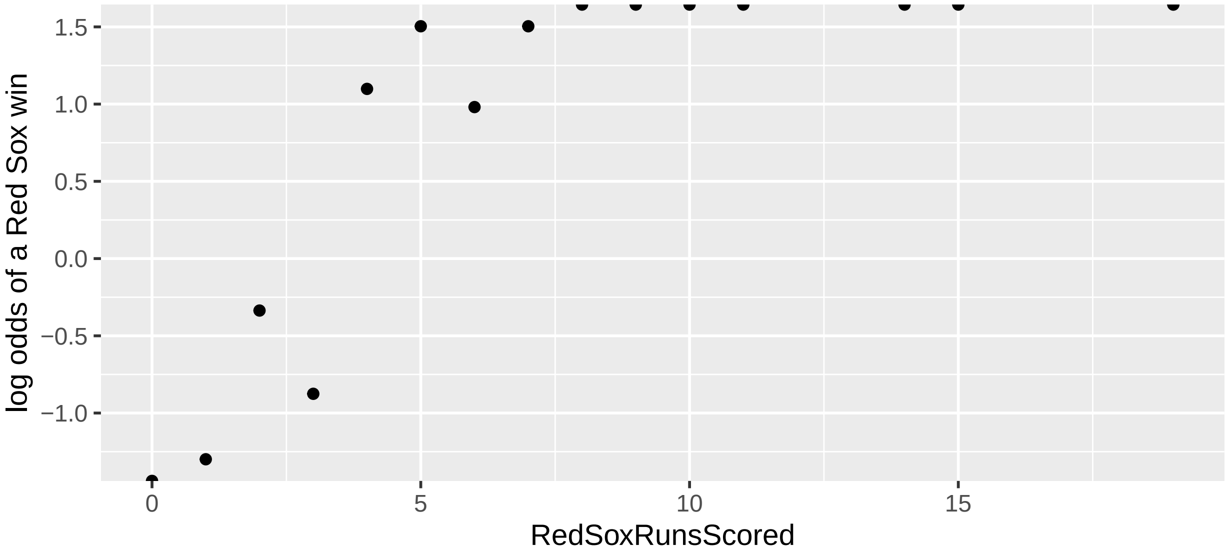
Let's investigate this model further. This model says that the log odds of a Red Sox win are linear in Red Sox runs scored. In other words, the log odds increase linearly with runs scored. Let's see if that's actually the case in the data.

```
#calculate log odds by runs scored
redsox %>%
  group_by(RedSoxRunsScored) %>%
  summarize(Wins = sum(Result == "W"),
            Games = n()) %>%
  mutate(perc = Wins/Games,
         logOdds = log(perc/(1 - perc))) -> runs.summary

runs.summary %>% head(10)
```

```
## # A tibble: 10 x 5
##    RedSoxRunsScored  Wins Games  perc  logOdds
##               <dbl> <int> <int> <dbl>    <dbl>
##  1                0     0     7 0     -Inf
##  2                1     3    14 0.214   -1.30
##  3                2     5    12 0.417   -0.336
##  4                3     5    17 0.294   -0.875
##  5                4    15    20 0.75     1.10
##  6                5    18    22 0.818    1.50
##  7                6    16    22 0.727    0.981
##  8                7     9    11 0.818    1.50
##  9                8     8     8 1       Inf
## 10                9     9     9 1       Inf
```

```
# a plot of the log odds
runs.summary %>%
  ggplot(aes(x = RedSoxRunsScored,
             y = logOdds)) +
  geom_point() +
  labs(y = "log odds of a Red Sox win")
```



Is a linear model for the log odds appropriate?

If the logs odds are linear in runs scored, are the odds linear in runs scored?

Which pitcher is most likely to result in a Red Sox win?

Let's look at the Red Sox record by starting pitcher.

```r
redsox %>%
  group_by(RedSoxStartingPitcherName) %>%
  summarize(W = sum(Result == "W"),
            L = sum(Result == "L"),
            Games = W + L,
            logOddsW = log(W/L),
            RunSupport = mean(RedSoxRunsScored)) %>%
  arrange(-Games) -> redsox.bypitcher
redsox.bypitcher %>%
  kable(caption = "Results of Red Sox games by starting pitcher", digits = 1)
```

Table 4: Results of Red Sox games by starting pitcher

RedSoxStartingPitcherName	W	L	Games	logOddsW	RunSupport
Rick Porcello	22	11	33	0.7	5.8
David Price	22	8	30	1.0	5.4
Chris Sale	18	9	27	0.7	5.0
Eduardo Rodriguez	19	4	23	1.6	5.7
Brian Johnson	9	4	13	0.8	6.5
Drew Pomeranz	5	6	11	-0.2	4.8
Nathan Eovaldi	5	6	11	-0.2	5.1
Hector Velazquez	5	3	8	0.5	5.1
Steven Wright	3	1	4	1.1	5.5
Jalen Beeks	0	1	1	-Inf	2.0
William Cuevas	0	1	1	-Inf	0.0

Why might we want to adjust for run support when assessing the relationship between pitcher and the Red Sox winning?

Let's consider only pitchers with at least 20 starts. (*why?*)

```r
redsox.bypitcher %>%
  filter(Games > 20 ) %>%
  pull(RedSoxStartingPitcherName) -> redsox.20plusStarters

redsox.20plusStarters
```

```
## [1] "Rick Porcello"      "David Price"        "Chris Sale"
## [4] "Eduardo Rodriguez"
```

Here is a model for pitcher and Red Sox winning adjusting for RunsScored.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{RedSoxRunsScored}_i + \beta_2 \text{Price}_i + \beta_3 \text{Rodriguez}_i + \beta_4 \text{Porcello}_i \tag{3}$$

where RunsScored_i is the number of runs the Red Sox scored in game i and the other variables are whether or not that pitcher started the game.

```r
model.pitcher <- glm(Result == "W" ~ RedSoxRunsScored + RedSoxStartingPitcherName,
                     data = filter(redsox, RedSoxStartingPitcherName %in% redsox.20plusStarters),
                     family = "binomial")
summary(model.pitcher)
```

```
##
## Call:
## glm(formula = Result == "W" ~ RedSoxRunsScored + RedSoxStartingPitcherName,
##     family = "binomial", data = filter(redsox, RedSoxStartingPitcherName %in%
##         redsox.20plusStarters))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4440  -0.7243   0.3312   0.7349   1.7126
##
## Coefficients:
##                                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                  -1.7663     0.6989  -2.527   0.0115
## RedSoxRunsScored                              0.5880     0.1347   4.364 1.28e-05
## RedSoxStartingPitcherNameDavid Price          0.5854     0.6985   0.838   0.4020
## RedSoxStartingPitcherNameEduardo Rodriguez    0.5263     0.7802   0.675   0.4999
## RedSoxStartingPitcherNameRick Porcello       -0.6138     0.6570  -0.934   0.3502
##
## (Intercept)                                 *
## RedSoxRunsScored                            ***
## RedSoxStartingPitcherNameDavid Price
## RedSoxStartingPitcherNameEduardo Rodriguez
## RedSoxStartingPitcherNameRick Porcello
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 134.68  on 112  degrees of freedom
## Residual deviance: 100.24  on 108  degrees of freedom
## AIC: 110.24
##
## Number of Fisher Scoring iterations: 6
```

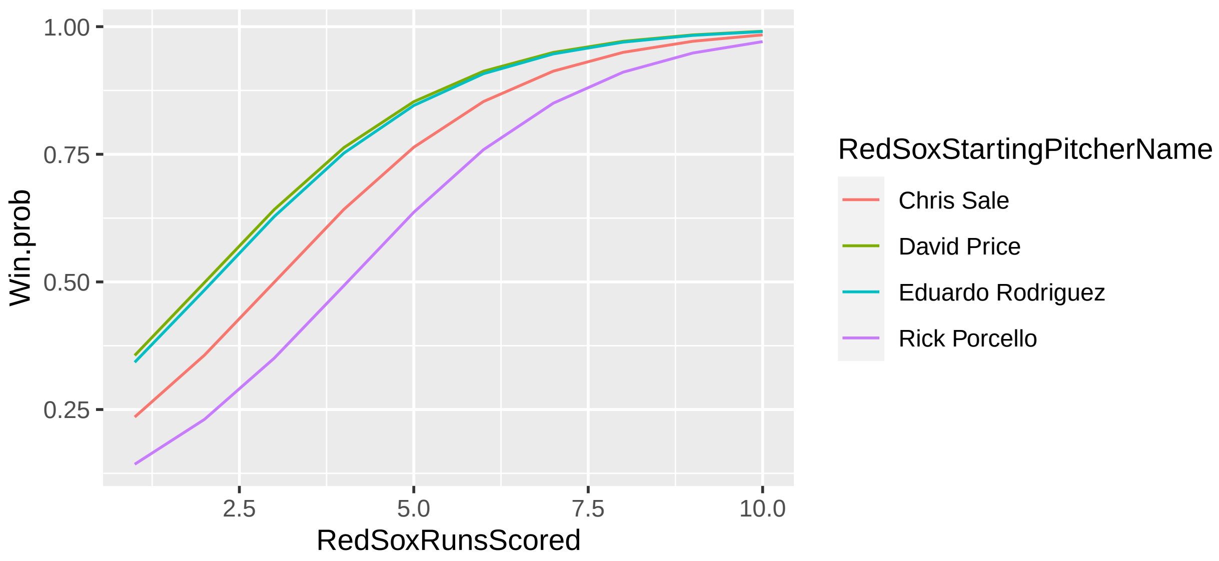
Calculate the probability the Red Sox win a game when they score 5 runs with Chris Sale pitching.

What conclusions should we make from this output?

Here are model predictions (win probabilities) for each pitcher by runs scored. (*why are they not linear?*)

```
runs = 1:10
tibble(expand_grid(runs,redsox.20plusStarters)) %>%
  rename(RedSoxRunsScored = runs,
         RedSoxStartingPitcherName = redsox.20plusStarters) %>%
  mutate(Win.prob = predict(model.pitcher, newdata = .,
                            type = "response")) -> predictions

predictions %>%
  ggplot(aes(x = RedSoxRunsScored,
             y = Win.prob,
             color = RedSoxStartingPitcherName)) +
  geom_line()
```



How does a model with an interaction between RunsScored and pitcher change the interpretation?

Logistic Regression and Classification

A common application for logistic regression is to predict (or classify) a binary outcome. In other words, based on some variables, we want to say whether the outcome will be a 1 or a 0, instead of merely reporting a probability. For example, in the analyses above, we might want to predict whether the Red Sox will win (1) or lose (0) based on how many runs they scored and who was pitching. A common approach is to fit a logistic regression model, obtain predicted probabilities, and establish a cutoff probability.

Based on the model in the last section, would you classify a game that the Red Sox scored 3 runs and Chris Sale was pitching a win? What cutoff did you use?

Ok, now let's do this for all games in 2018 with Porcello, Price, Rodriguez, or Sale starting and see how often our predictions are correct. To do this, we will obtain the *confusion matrix*.

```
cutoff = 0.5
redsox %>%
  filter(RedSoxStartingPitcherName %in% redsox.20plusStarters) %>%
  mutate(Win.prob = predict(model.pitcher,
                            newdata = .,
                            type = "response"),
         Prediction = ifelse(Win.prob >= cutoff, "W","L")) %>%
  select(HomeTeam, VisitingTeam,
         RedSoxStartingPitcherName,
         RedSoxRunsScored, Win.prob,
         Prediction, Result) -> redsox.predictions

redsox.predictions %>%
  select(-RedSoxStartingPitcherName) %>%
  head(10)
```

```
## # A tibble: 10 x 6
##    HomeTeam VisitingTeam RedSoxRunsScored Win.prob Prediction Result
##    <chr>    <chr>                   <dbl>    <dbl> <chr>      <fct>
##  1 TBA      BOS                         4    0.642 W          L
##  2 TBA      BOS                         1    0.356 L          W
##  3 TBA      BOS                         3    0.351 L          W
##  4 MIA      BOS                         4    0.642 W          W
##  5 BOS      TBA                         3    0.642 W          W
##  6 BOS      TBA                        10    0.971 W          W
##  7 BOS      TBA                         8    0.970 W          W
##  8 BOS      NYA                        14    0.998 W          W
##  9 BOS      NYA                         7    0.950 W          L
## 10 BOS      NYA                         6    0.759 W          W
```

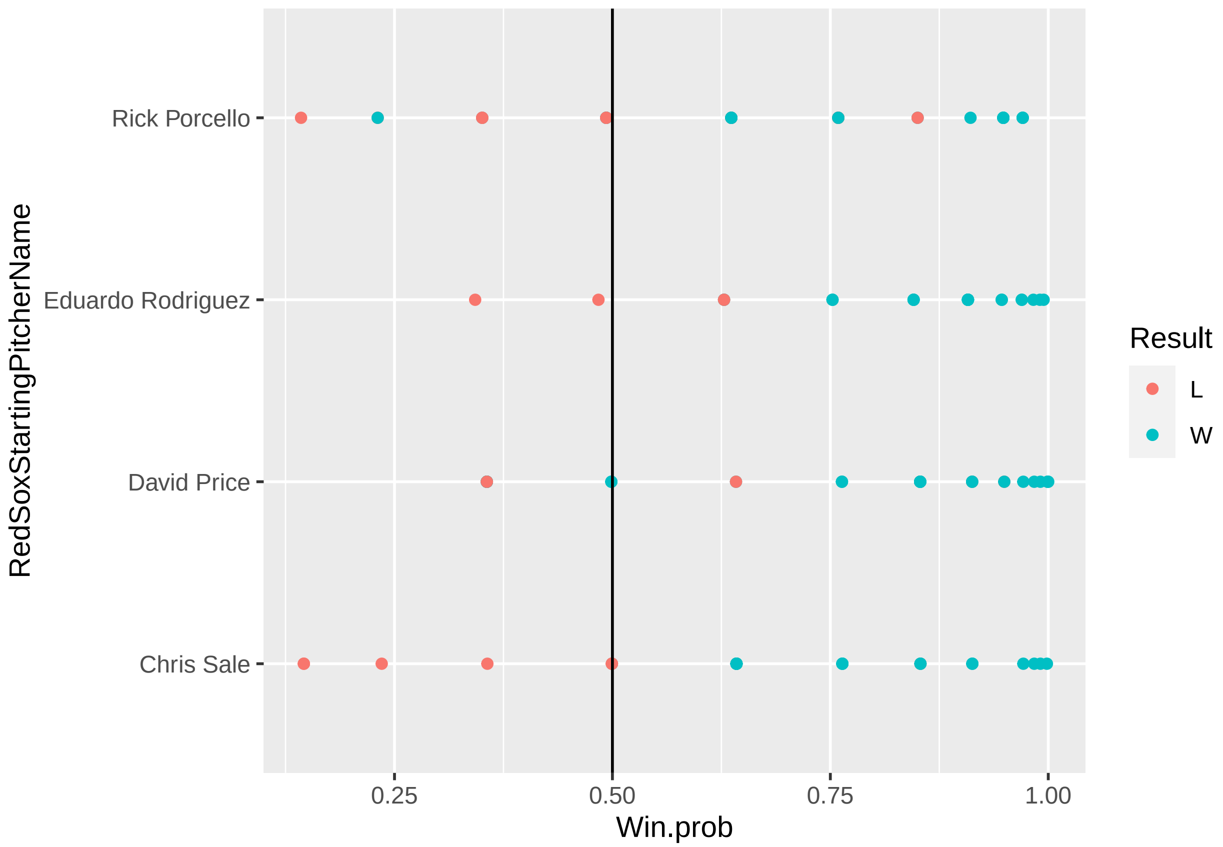
The confusion matrix. . . .

```
redsox.predictions %>%
  count(Result, Prediction) %>%
  pivot_wider(values_from = n,
              names_from = Prediction)
```

```
## # A tibble: 2 x 3
##   Result     L     W
##   <fct>  <int> <int>
## 1 L         19    13
## 2 W         11    70
```

What is the correct classification rate?

```
redsox.predictions %>%
  ggplot(aes(x = RedSoxStartingPitcherName,
             y = Win.prob,
             color = Result)) +
  geom_point() +
  coord_flip() +
  geom_hline(yintercept = cutoff)
```



How could we improve this analysis with cross validation?