UNITED STATES MILITARY ACADEMY
WEST POINT®

MA388 Sabermetrics WPR 1 — February 28, 2020

LTC Kevin Cummiskey

Notes

- You are authorized to use your textbook, computer, notes, code from class exercises and homeworks, and other materials I have provided to you for this course.
- You are not authorized to use the internet except to access the above references.
- This exam is an individual event.
- You have 55 minutes to complete this exam.
- This exam is worth 100 points.

Name: _____

1. (20 points) In the late 1960's, baseball officials were concerned the sport's popularity was decreasing. Many people believed a lack of offense (hitting) was responsible. Therefore, in 1969, the league approved several rule changes to increase offense. Specifically, they lowered the maximum height of the pitching mound from 15 inches to 10 inches and reduced the size of the strike zone. Use the Lahman database to answer the following questions comparing the 1968 and 1970 seasons. *Write your code and your answers in the space below.*

   (a) Using the *Pitching* data frame, how many pitchers in 1968 with at least 20 starts (variable: **GS**) had an earned run average (variable: **ERA**) of less than 2.00? How many such pitchers were there in 1970?

```
Pitching %>%
 filter(yearID == 1968,
        GS >= 20,
        ERA < 2)
```

There were 6 pitchers in 1968 with at least 20 starts and an ERA of less than 2.

```
Pitching %>%
 filter(yearID == 1970,
        GS >= 20,
        ERA < 2)
```

There were 0 pitchers in 1970 with at least 20 starts and an ERA of less than 2.

   (b) Using the *Batting* data frame, determine who had the highest batting average ($AVG = H/AB$) of players with at least 500 at bats (variable: **AB**) in the American League (**lgID == "AL"**) in 1968 and 1970. Report both players' IDs and their batting averages.

```
Batting %>%
 filter(AB >= 500,
        yearID == 1968,
        lgID == "AL") %>%
 mutate(AVG = H/AB) %>%
 arrange(-AVG) %>%
 head(1)
```

In 1968, yastrca01 (Carl Yazstremski) won the AL batting title with a 0.300 average.

```
Batting %>%
 filter(AB >= 500,
        yearID == 1970,
        lgID == "AL") %>%
 mutate(AVG = H/AB) %>%
 arrange(-AVG) %>%
 head(1)
```

In 1970, johnsa101 (Alex Johnson) won the AL batting title with a 0.329 average.

   (c) Using the *Teams* data frame and the *summarize* and *group_by* functions, calculate the overall major league baseball batting average ($AVG = H/AB$) in 1968 and 1970.

```
Batting %>%
 filter(yearID %in% c(1968,1970)) %>%
 group_by(yearID) %>%
 summarize(AVG = sum(H)/sum(AB))
```

The league batting averages were 0.237 and 0.254 in 1968 and 1970, respectively.

2. (30 Points) In this problem, you will again consider effects of the 1969 rule changes. You decide to fit the Pythagorean model to teams in the 1960s and the 1970s to determine if the relationship between season scoring (runs and runs allowed) and expected win percentage is different between these decades. Recall the Pythagorean model is:

$$Wpct = \frac{R^k}{R^k + RA^k} + \epsilon \quad \epsilon \sim \text{Normal}(0, \sigma^2)$$

where *Wpct* is win percentage, *R* is runs scored, and *RA* is runs allowed.

(a) List one advantage of using the Pythagorean model instead of the simple linear regression model, $Wpct = \beta_0 + \beta_1 RD + \epsilon$.

The linear regression model assumes the relationship between run differential and win percentage is the same for all runs scored/allowed combinations with the same run differential. The Pythagorean model relaxes this assumption.

The data frame **myteams** contains a row for every team by season from 1960-1979. *logR* is the log ratio of runs scored to runs allowed and *logW* is the log ratio of wins to losses. The first three rows of the data frame are:

```
myteams %>% head(3)
```

```
##   teamID yearID   R  RA   RD  W  L  Wpct    logW    logR decade
## 1    BAL   1960 682 606   76 89 65 0.578  0.3142  0.1181   1960
## 2    BOS   1960 658 775 -117 65 89 0.422 -0.3142 -0.1637   1960
## 3    CHA   1960 741 617  124 87 67 0.565  0.2612  0.1831   1960
```

Next, you fit the Pythagorean model for each decade.

```
#1960s
model_pyt_1960 <- lm(logW ~ 0 + logR,
                     data = myteams %>% filter(decade == 1960))
coef(model_pyt_1960)
```

```
##     logR
## 1.905517
```

```
#1970s
model_pyt_1970 <- lm(logW ~ 0 + logR,
                     data = myteams %>% filter(decade == 1970))
coef(model_pyt_1970)
```

```
##     logR
## 1.739909
```

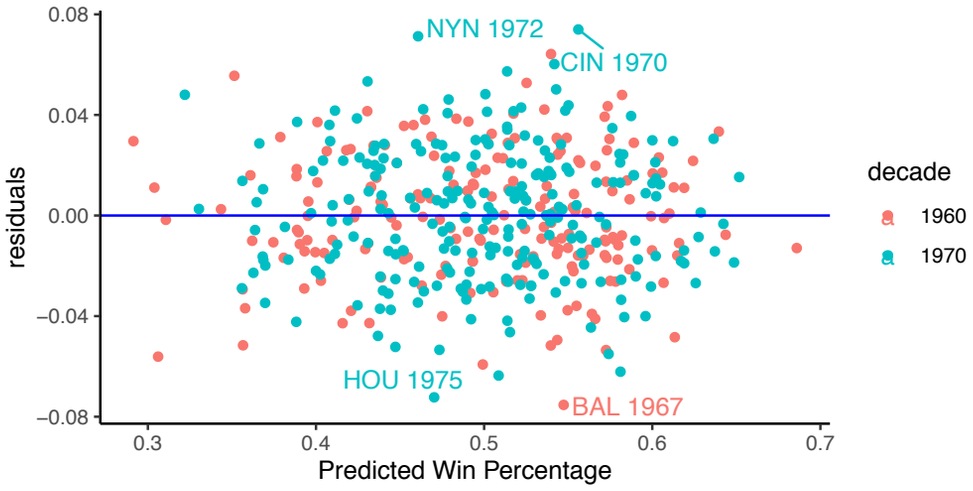

Figure 1: Pythagorean residuals

(b) Using the results above, calculate and compare the expected win percentage in the 1960s to the 1970s for a team that scores 800 runs and allows 700 runs.

1960s: k = 1.905. Therefore, the expected win percent is 800^(1.905)/(800^(1.905) + 700^(1.905)) = 0.563
1970s: k = 1.740. Therefore, the expected win percent is 800^(1.740)/(800^(1.740) + 700^(1.740)) = 0.558

(c) Typically, the value of scoring additional runs in terms of expected win percentage is greater in low scoring environments than in high scoring environments. Is there evidence of that here? (Note there were a lot more runs scored in the 1970s than the 1960s).

The 1970s was a higher run scoring environment, so we would expect additional runs to be less valuable. We see some evidence of that here because the predicted win percent in 1970 for the same run differential was slightly higher. However, the predictions were fairly close.

(d) Figure 1 is a plot of the residuals from the Pythagorean models. Discuss one conclusion you would make from the residuals.

There does not appear to be a trend in the residuals. This suggests the form of the model is appropriate for the data. There are some outliers we might want to investigate further (Baltimore, 1967). However, these outliers are not likely to exert much leverage over the regression estimates.

(e) Based on the residuals, which teams outperformed their expected win percentage the most? Explain one reason a team might outperform their expected win percentage.

The 1972 New York Mets (NYN 72) and the 1970 Cincinnati Red (CIN 70) outperformed their predicted win percentage. Typically, this occurs when a team wins a lot of close games or loses a couple games by wide margins. Having a good closer or just getting lucky are likely reasons.

Lastly, you run the following code:

```
myteams %>%
  group_by(decade) %>%
  summarize(my.stat = sqrt(sum(residuals^2)))
```

```
## # A tibble: 2 x 2
##   decade my.stat
##   <fct>    <dbl>
## 1 1960     0.345
## 2 1970     0.408
```

(f) What do we typically call **my.stat** above? What units is it in? What does it tell us about these models?

I removed this question. It should have been the root mean squared error (RMSE). It is in the units of the response (win percent) and gives us an idea of typically how much error we should expect when making predictions with this model.

3. (20 Points) Let's again consider the effects of the 1969 rule changes. Table 1 contains the Run Expectancy Matrices for 1968 (left table) and 1970 (right table).

Table 1: Run Expectancy Matrix (1968 and 1970)

Runners	Outs0	Outs1	Outs2	Runners	Outs0	Outs1	Outs2
000	0.382	0.201	0.075	000	0.486	0.262	0.097
001	1.152	0.799	0.331	001	1.285	0.893	0.368
010	0.919	0.568	0.280	010	1.064	0.660	0.320
011	1.689	1.166	0.536	011	1.863	1.291	0.591
100	0.717	0.417	0.172	100	0.854	0.509	0.210
101	1.487	1.015	0.428	101	1.653	1.139	0.482
110	1.254	0.784	0.377	110	1.432	0.907	0.433
111	2.024	1.382	0.633	111	2.231	1.538	0.705

(Source: https://legacy.baseballprospectus.com/)

(a) In 1968, when Runners = "000" and there are no outs, the run expectancy is 0.382. Explain how to interpret 0.382.

This is the expected number of runs a team will score in the remainder of the inning when there are no runners on base and no outs.

(b) Based on an inspection of the run expectancy matrices, does it appear the rule changes resulted in more runs? Explain.

Yes, the run expectancies are higher for every state in 1970 than they were in 1968.

(c) Let's say there are runners on first and second with no outs. In terms of run value, was a successful sacrifice bunt more valuable in 1968 or 1970? (Note a successful sacrifice bunt would typically result in runners on second and third with one out.)

1968: RunValue = 1.166 - 1.254 = -0.088 runs
1970: RunValue = 1.291 - 1.432 = -0.141 runs

In both years, a bunt has a negative run value. In 1970, the run value of a bunt was less, so 1968 is more valuable.

4. (30 Points) Next, we investigate whether the rule changes resulted in umpires actually shrinking their strike zones. Unfortunately, PITCHf/x data didn't become available until 2005. However, you have some very industrious interns working for you that manually record the location of 1000 pitches (balls and called strikes only) in each of the 1968 and 1970 seasons. Here are the first few lines of their data set and a summary table of their results.

```
pitches %>% head(5)
```

```
## # A tibble: 5 x 6
##   year  start_speed end_speed     px    pz type
##   <fct>       <dbl>     <dbl>  <dbl> <dbl> <chr>
## 1 1968         81        75.5 -1.41   2.74 B
## 2 1968         86.7      79.3  0.290  1.53 B
## 3 1968         92.2      85.4  1.05   3.18 B
## 4 1968         95.2      88.3 -0.32   1.49 B
## 5 1968         84.1      78.7  0.71   0.98 B
```

Table 2: Number of balls and called strikes by year.

type	1968	1970
Balls	661	721
Called Strikes	339	279

(a) Calculate the proportion of pitches in 1968 and 1970 that were called strikes.

1968: 339/(661+339) = 0.339
1970: 279/(279+721) = 0.279

(b) Calculate the odds ratio for a called strike comparing 1970 to 1968. (Note the odds ratio is the odds of a called strike in 1970 divided by the odds of a strike in 1968.)

odds ratio = (279/721)/(339/661) = 0.755

(c) Is there evidence umpires were calling less strikes in 1970? Briefly explain.

Called strikes were less likely in 1970 in this sample, so that is some evidence.

(d) The analysis above does not adjust for pitch location. Briefly explain why an analysis adjusting for pitch location is better.

Ideally, we would like to see if there was a difference in called strike probability on similar pitches. We can accomplish this in a statistical model by adjusting for pitch location.

You decide to use the following model to adjust for pitch location:

$$\text{Strike}_i \sim \text{Bernoulli}(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + f(px_i, pz_i) + \beta_1 year_i$$

where $f(px, pz)$ is a smooth function of vertical and horizontal pitch location.

(e) Interpret β_1 in this model.

B1 is the log odds ratio for a called strike comparing 1970 to 1968 adjusted for the pitch location.

In R, you fit the model above:

```
model <- gam(type == "C" ~ s(px,pz) + year,
         family = "binomial",
         data = pitches)
summary(model)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## type == "C" ~ s(px, pz) + year
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -14.5190     3.5157  -4.130 3.63e-05 ***
## year1970     -0.1472     0.1673  -0.879    0.379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df Chi.sq p-value
## s(px,pz) 28.6  28.94  310.8  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.651   Deviance explained = 62.9%
## UBRE = -0.51048  Scale est. = 1         n = 1999
```

```
exp(-0.1472)
```

```
## [1] 0.8631213
```

(f) Based on the output above, is there evidence umpires changed their strike zone? Explain.

No, we see that B1hat is negative indicating that there was a reduced probability of a called strike in 1970 compared to 1968 after adjusting for pitch location. However, the p-value (0.379) indicates this is not statistically significant. In other words, if there were no difference in the population, it would not have been that unusual to observe results this extreme in our sample.

Let's look at high strikes in particular.

```
#1968
predict(model, type = "response",
        newdata = data.frame(px = 0, pz = 3.5, year = 1968))
```

```
##         1
## 0.3233473
```

```
#1970
predict(model, type = "response",
        newdata = data.frame(px = 0, pz = 3.5, year = 1970))
```

```
##         1
## 0.2920211
```

(g) Explain what the above code is doing? Do the results provide evidence umpires have changed their strike zone? Explain.

The code above calculates the predicted probability of a called strike in 1968 and 1970 for a pitch in the middle of the plate and 3.5 feet off the ground. There is a small decrease in the probability of a called strike for this pitch.

(h) Briefly explain how you would improve upon the analysis in this problem.

We may want to adjust for other variables including the type of pitch, handedness of the batter/pitcher, and count. In addition, it would be interesting to see if the rule changes influences pitchers approach to at bats.