COLLEGE OF SCIENCE
Mathematics

Final Presentation:
Justin Verlander Pitch Performance Analysis

Kylie Mattox, Hailey Schleining, Sarah Stallman, Kayla Wilkins

Background & Introduction

Background Information

- Pitcher: the player who throws the baseball from the pitcher's mound towards the catcher to start each play with the intent of making it difficult for the batter to hit
- Pitching is considered the most important part of defense in baseball
- To cause the most problem for batters, pitchers look to change the:
    - Pitch type
    - Speed
    - Control (the ability to pitch to specific points in the strike zone)

Project Description

We are analyzing the pitcher Justin Verlander over time with 2 different teams, and thus coaches, to determine if there are changes in his pitching performance and/or patterns.

- Four types of pitches
  - 4-Seam Fastball, Curveball, Changeup, Slider
- Which pitch type yields the best results (most outs)?
- How do the following affect results?
  - Pitch MPH
  - Exit Velocity
  - Launch Angle
  - Distance
- Does he pitch differently when on a different team with a different coach?
- Build predictive models using statistically significant variables

Motivation

Project Specific Motivation

The results of our analysis could:

- Benefit Verlander’s coaching staff - provide them with an analysis of his pitching performance & patterns
- Help identify the most effective pitch - thus identify which pitch to practice
- Place more of an emphasis on pitching coaches & their impact on the game
    - Lead to an increased pay/recognition for pitching coaches
- Determine if a pitcher can improve as he ages and is at more risk for injury

Overview of Results

Overview of Significant Results

- Strikeout probabilities showed that Houston has a significantly higher probability of 0.335 versus Detroit's 0.270
  - Z-value indicates a rejection of our null hypothesis and statistically significant difference between the teams
- Increased the probability of sliders thrown and probability of strikeouts resulting from sliders nearly doubled when Verlander played for Houston.
- Linear and logistic regressions showed significance between LA, EV, and Distance when compared to the results
  - LA was removed due to collinearity
  - Pitch type was only significant for 3 out of 4 models
- Z-values indicated no significant difference between the EV, LA, and Distance variable averages on each team

Code

Overview of Process

Data Collection: game logs from BaseballSavant.MLB.com (MLB.com's clearinghouse for Statcast data)

Data Processing: initially sorted data based on the team (Detroit & Houston), then sorted by result, & then sorted into various data frames for each variable (EV, LA, distance, & MPH), eliminating all NA values to avoid affecting the averages

Analysis:

- Probabilities: of each result, each pitch type, and the strikeout probability
- Linear Model: used On Base plus Slugging values
- Logistic Model: 1 represents any out, 0 represents non-outs
- Analyzed each of the 5 variables - looked for statistically significant ones
- Built predictive models using linear & logistic regression

Code Demonstration: Probabilities

```
### Detroit Probability (Plays)

```{r}
#determining the probabilities of each result for Detroit
probs1 <- data.frame(summary(detroit$Result))
colnames(probs1) <- "Number"
probs1$probability <- probs1$Number/2003
probs1
```

From looking at the initial probabilities of each result for Detroit we see that field out is the most likely at 0.401 and then strikeout follows at 0.270. Both of these outcomes are beneficial for Detroit.

Houston Probability (Plays)

```{r}
#determining the probabilities of each result for Houston
probs2 <- data.frame(summary(houston$Result))
colnames(probs2) <- "Number"
probs2$probability <- probs2$Number/2739
probs2
```

From looking at the initial probabilities of each result for Houston we see that field out is the most likely at 0.387 and then strikeout follows at 0.335. Both of these outcomes are beneficial for Houston.

Comparing Probabilities of Strikeouts (Z-test)

```{r}
#isolating the strikeout probability from the previous probability data frame
#detroit strikeouts probability
probs1["Strikeout",2]

#houston strikeouts probability
probs2["Strikeout",2]

#determining whether there is a significant difference
p_bar <- (probs1["Strikeout",1] + probs2["Strikeout",1])/(2003+2739)
z <- (probs1["Strikeout",2] - probs2["Strikeout",2])/sqrt((p_bar*(1-p_bar))*((1/2003)+(1/2739)))
z
```

 [1] 0.2700949
 [1] 0.3351588
 [1] -4.795067
```

Code Demonstration: Categorizing for Regression

```
### Setting up Linear Regression (Detroit)

```{r}
linreg_det <- detroit
linreg_det$Result <- as.character(linreg_det$Result)
linreg_det$Pitch.Type <- as.character(linreg_det$Pitch.Type)
linreg_det[linreg_det$Result == "Strikeout", 6] <- 0
linreg_det[linreg_det$Result == "Field Out", 6] <- 0
linreg_det[linreg_det$Result == "Force Out", 6] <- 0
linreg_det[linreg_det$Result == "Field Error", 6] <- 0
linreg_det[linreg_det$Result == "Double Play", 6] <- 0
linreg_det[linreg_det$Result == "Grounded Into Double Play", 6] <- 0
linreg_det[linreg_det$Result == "Hit By Pitch", 6] <- 1
linreg_det[linreg_det$Result == "Strikeout Double Play", 6] <- 0
linreg_det[linreg_det$Result == "Fielders Choice Out", 6] <- 0
linreg_det[linreg_det$Result == "Walk", 6] <- 1
linreg_det[linreg_det$Result == "Sac Bunt", 6] <- 1
linreg_det[linreg_det$Result == "Sac Fly", 6] <- 1
linreg_det[linreg_det$Result == "Fielders Choice", 6] <- 1
linreg_det[linreg_det$Result == "Single", 6] <- 2
linreg_det[linreg_det$Result == "Double", 6] <- 3
linreg_det[linreg_det$Result == "Triple", 6] <- 4
linreg_det[linreg_det$Result == "Home Run", 6] <- 5
linreg_det[linreg_det$Pitch.Type == "4-Seam Fastball", 12] <- 0
linreg_det[linreg_det$Pitch.Type == "Changeup", 12] <- 1
linreg_det[linreg_det$Pitch.Type == "Curveball", 12] <- 2
linreg_det[linreg_det$Pitch.Type == "Slider", 12] <- 3
linreg_det$Result <- as.numeric(linreg_det$Result)
linreg_det$Pitch.Type <- as.numeric(linreg_det$Pitch.Type)
```
```

```
### Setting up Logistic Regression (Detroit)

```{r}
logreg_det <- detroit
logreg_det$Result <- as.character(logreg_det$Result)
logreg_det$Pitch.Type <- as.character(logreg_det$Pitch.Type)
logreg_det[logreg_det$Result == "Strikeout", 6] <- 1
logreg_det[logreg_det$Result == "Field Out", 6] <- 1
logreg_det[logreg_det$Result == "Force Out", 6] <- 1
logreg_det[logreg_det$Result == "Field Error", 6] <- 1
logreg_det[logreg_det$Result == "Double Play", 6] <- 1
logreg_det[logreg_det$Result == "Grounded Into Double Play", 6] <- 1
logreg_det[logreg_det$Result == "Hit By Pitch", 6] <- 0
logreg_det[logreg_det$Result == "Strikeout Double Play", 6] <- 1
logreg_det[logreg_det$Result == "Fielders Choice Out", 6] <- 1
logreg_det[logreg_det$Result == "Walk", 6] <- 0
logreg_det[logreg_det$Result == "Sac Bunt", 6] <- 0
logreg_det[logreg_det$Result == "Sac Fly", 6] <- 0
logreg_det[logreg_det$Result == "Fielders Choice", 6] <- 0
logreg_det[logreg_det$Result == "Single", 6] <- 0
logreg_det[logreg_det$Result == "Double", 6] <- 0
logreg_det[logreg_det$Result == "Triple", 6] <- 0
logreg_det[logreg_det$Result == "Home Run", 6] <- 0
logreg_det[logreg_det$Pitch.Type == "4-Seam Fastball", 12] <- 0
logreg_det[logreg_det$Pitch.Type == "Changeup", 12] <- 1
logreg_det[logreg_det$Pitch.Type == "Curveball", 12] <- 2
logreg_det[logreg_det$Pitch.Type == "Slider", 12] <- 3
logreg_det$Result <- as.numeric(logreg_det$Result)
logreg_det$Pitch.Type <- as.numeric(logreg_det$Pitch.Type)
```
```

Code Demonstration: Linear & Logistic Regression

```
### Linear regression for detroit
```{r}
linreg_det_ev <-  lm(linreg_det$Result ~ linreg_det$EV..MPH., data =
linreg_det)
summary(linreg_det_ev)
linreg_det_dist <-  lm(linreg_det$Result ~ linreg_det$Dist..ft., data =
linreg_det)
summary(linreg_det_dist)
linreg_det_mph <-  lm(linreg_det$Result ~ linreg_det$Pitch..MPH., data
= linreg_det)
summary(linreg_det_mph)
linreg_det_la <-  lm(linreg_det$Result ~ linreg_det$LA, data =
linreg_det)
summary(linreg_det_la)
```

Building a logistic regression, with different independent
variables. Using the detroit dataframe.

```{r}
dist_log_det = glm(Result~Dist..ft., data = logreg_det, family =
binomial)
summary(dist_log_det)   #distance is a significant variable

ev_log_det = glm(Result~EV..MPH., data = logreg_det, family = binomial)
summary(ev_log_det)  #exit velocity is significant

la_log_det = glm(Result~LA, data = logreg_det, family = binomial)
summary(la_log_det)  #launch angle is significant

pitch_mph_log_det = glm(Result~Pitch..MPH., data = logreg_det, family =
binomial)
summary(pitch_mph_log_det)  #not significant

det_log = glm(Result~LA + EV..MPH. + Dist..ft., data = logreg_det,
family = binomial)
summary(det_log)
```
```

```
### Linear regression model for houston
```{r}
linreg_hou_ev <-  lm(linreg_hou$Result ~ linreg_hou$EV..MPH., data =
linreg_hou)
summary(linreg_hou_ev)
linreg_hou_dist <-  lm(linreg_hou$Result ~ linreg_hou$Dist..ft., data =
linreg_hou)
summary(linreg_hou_dist)
linreg_hou_mph <-  lm(linreg_hou$Result ~ linreg_hou$Pitch..MPH., data
= linreg_hou)
summary(linreg_hou_mph)
linreg_hou_la <-  lm(linreg_hou$Result ~ linreg_hou$LA, data =
linreg_hou)
summary(linreg_hou_la)
```

Building a logistic regression, with different independent
variables. Using the houston dataframe.

```{r}
dist_log_hou = glm(Result~Dist..ft., data = logreg_hou, family =
binomial)
summary(dist_log_hou)   #distance is a significant variable

ev_log_hou = glm(Result~EV..MPH., data = logreg_hou, family = binomial)
summary(ev_log_hou)  #exit velocity is significant

la_log_hou = glm(Result~LA, data = logreg_hou, family = binomial)
summary(la_log_hou)  #launch angle is significant

pitch_mph_log_hou = glm(Result~Pitch..MPH., data = logreg_hou, family =
binomial)
summary(pitch_mph_log_hou)  #pitch mph is significant

hou_log = glm(Result~LA + EV..MPH. + Dist..ft., data = logreg_hou,
family = binomial)
summary(hou_log)
```
```

Code Demonstration: Fin Model

```
### Final Models

```{r}
final_lm_det <- (lm(Result ~ EV..MPH. + Dist..ft., data=linreg_det))
summary(final_glm_hou)
final_lm_hou <- (lm(Result ~ EV..MPH. + Dist..ft., data=linreg_hou))
final_glm_det <- (glm(Result ~ Pitch.Type + EV..MPH. + Dist..ft., data=logreg_det, family=binomial))
final_glm_hou <- (glm(Result ~ Pitch.Type + EV..MPH. + Dist..ft., data=logreg_hou, family=binomial))
```

Logistic Model Accuracy (Detroit)

```{r}
probs <- predict(final_glm_det, type = "response")
pred <- rep('Not Out', length(probs))
pred[probs > .6] <- "Out"
table(pred, logreg_det$Result)
```

pred   0   1
  Not Out 115 143
  Out   484 1261

The percentage of correct predictions for a threshold of .6 on the training data is (115+1261)/2003, or about 68%.
This means the training error rate is about 31%. For outs, the model is correct about 90% of the time, while for not
outs, the model is right about 19% of the time.

Logistic Model (Houston)

```{r}
prob2 <- predict(final_glm_hou, type = "response")
pred2 <- rep('Not Out', length(prob2))
pred2[prob2 > .6] <- "Out"
table(pred2, logreg_hou$Result)
```

pred2   0   1
  Not Out 138 182
  Out   549 1870

The percentage of correct predictions for a threshold of .6 on the training data is (138+1870)/2739, or about 73%.
This means the training error rate is about 26%. For outs, the model is correct about 91% of the time, while for not
outs, the model is right about 20% of the time.
```

Code Demonstration: Z-Values & Averages

```
### comparing the means of our different predictive variables to determine if there are significant differences between Houston and Detroit
```{r}
#removing NA values for exit velocity
detroit_fixed_ev <- subset(detroit, subset=(detroit$EV..MPH. != 0))
houston_fixed_ev <- subset(houston, subset=(houston$EV..MPH. != 0))
#removing NA values for launch angle
detroit_fixed_la <- subset(detroit, subset=(detroit$LA != -180))
houston_fixed_la <- subset(houston, subset=(houston$LA != -180))
#removing NA values for distance
detroit_fixed_dist <- subset(detroit, subset=(detroit$Dist..ft. != 0))
houston_fixed_dist <- subset(houston, subset=(houston$Dist..ft. != 0))

zEV <- (mean(detroit_fixed_ev$EV..MPH.)-mean(houston_fixed_ev$EV..MPH.))/sqrt((var(detroit_fixed_ev$EV..MPH.)/1296)+(var(houston_fixed_ev$EV..MPH.)
/1644))
zEV

zLA <- (mean(detroit_fixed_la$LA)-mean(houston_fixed_la$LA))/sqrt((var(detroit_fixed_la$LA)/1296)+(var(houston_fixed_la$LA)/1644))
zLA

zDist <- (mean(detroit_fixed_dist$Dist..ft.)-mean(houston_fixed_dist$Dist..ft.))/sqrt((var(detroit_fixed_dist$Dist..ft.)/1296)+(var(houston_fixed_d
ist$Dist..ft.)/1644))
zDist
```
the z value for the EV is 0.988
the z value for the LA is 0.119
the z value for the distance is 1.016
```

Results

Results: Result Probabilities

Initially we looked at the probabilities of each result:

Detroit		
Result	Amount	Probability
Field out	803	0.401
Strikeout	541	0.270
Single	261	0.130
Walk	150	0.075
Double	80	0.040
Homerun	64	0.032

Houston		
Result	Amount	Probability
Field out	1059	0.387
Strikeout	918	0.335
Single	295	0.108
Walk	151	0.055
Double	93	0.034
Homerun	93	0.034

**These charts only depict the more important results

Results: Comparing Strikeout Probabilities

Then we compared probabilities of strikeouts by performing a Z-test:

- Z-value: 4.795
- Can reject the null hypothesis
- Conclude that the probabilities are significantly different, with a higher probability for strikeouts when Verlander played for Houston at 0.335 versus 0.270 for when he played for Detroit.

Results: Pitch Type Probabilities

Then we looked at the probabilities of each pitch type:

Detroit		
Pitch Type	Amount	Probability
4-Seam Fastball	1059	0.529
Slider	458	0.229
Curveball	305	0.152
Changeup	181	0.090

Houston		
Pitch Type	Amount	Probability
4-Seam Fastball	1327	0.484
Slider	902	0.329
Curveball	421	0.154
Changeup	89	0.032

Results: Strikeouts Based on Pitch Type

4 Main Pitch Types Resulting in Strikeout				
	4-Seam Fastball	Change up	Curveball	Slider
Detroit	0.5176	0.0573	0.1848	0.2403
Houston	0.3802	0.0316	0.1841	0.4041

Results: Statistically Significant Variables

Analyzed statistically significant variables from our models

Linear Models		
	Parameters Compared to Results	P-Value
Detroit	Pitch Type	0.0602
	Pitch MPH	1.000
	Launch Angle	<2e^-16
	Distance	<2e^-16
	Exit Velocity	<2e^-16
Houston	Pitch Type	0.000357
	Pitch MPH	0.00945
	Launch Angle	<2e^-16
	Distance	<2e^-16
	Exit Velocity	<2e^-16

Logistic Models		
	Parameters Compared to Results	P-Value
Detroit	Pitch Type	0.00145
	Pitch MPH	0.1487
	Launch Angle	0.000109
	Distance	1.88e^-11
	Exit Velocity	3.22e^-12
Houston	Pitch Type	1.91e^-5
	Pitch MPH	0.0015
	Launch Angle	1.85e^-15
	Distance	<2e^-16
	Exit Velocity	<2e^-16

Results: Linear & Logistic Models

Linear:

Result_Detroit = .1529 + .0036(EV) + .0023(Distance)

Result_Houston = .1 + .0027(EV) + .0029(Distance)

Logistic:

$$Probability.Det = \frac{e^{1.22 + .117(Pitch\ Type) - .0059(EV) - .001(Distance)}}{1 + e^{1.22 + .117(Pitch\ Type) - .0059(EV) - .001(Distance)}}$$

$$Probability.Hou = \frac{e^{1.65 + .078(Pitch\ Type) - .0065(EV) - .002(Distance)}}{1 + e^{1.65 + .078(Pitch\ Type) - .0065(EV) - .002(Distance)}}$$

Results: Final Predictive Model Accuracy

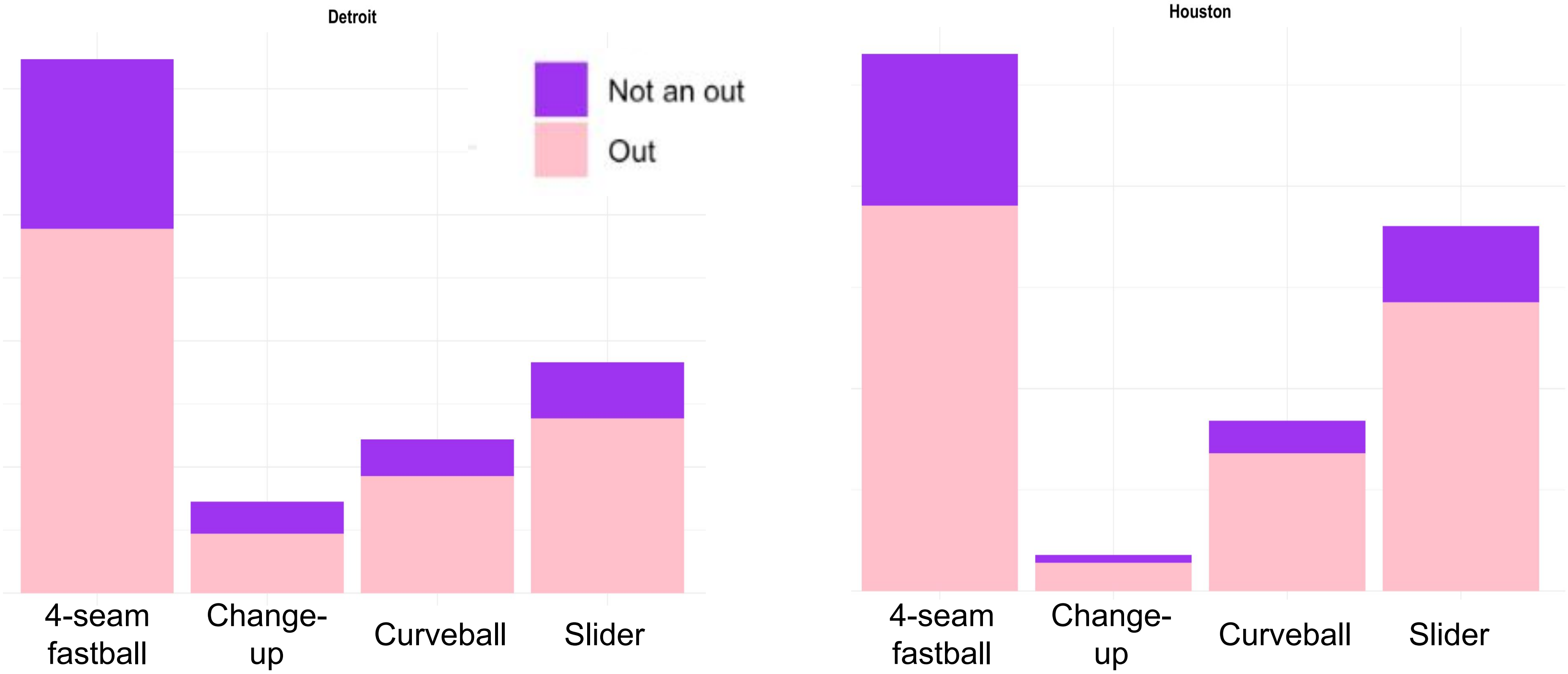
Linear Models:

- R-squared for Detroit: 0.1447
- R-squared for Houston: 0.1814

Logistic Models:

- Detroit: correctly predicts outs 90% of time & 19% for non-outs
- Houston: correctly predicts outs 91% of time & 20% for non-outs

Distribution of Outs vs Not Outs



COLLEGE OF SCIENCE
Mathematics

Results: Comparing P-Values & Averages

Compared p-values for each variable that was statistically significant between each team

Result:

- p-value for exit velocity: 0.1635 - not significant
- p-value for launch angle: 0.4522 - not significant
- p-value for distance: 0.1539 - not significant
- Thus, we can fail to reject the null & conclude that there is no significant difference between these parameters when Verlander was playing for Detroit vs Houston.

Average Values	Exit Velocity	Launch Angle	Distance
Detroit	87.82	17.60	191.31
Houston	88.34	17.49	186.32

Importance of Results

Utilization

- Shows how team and pitching coach affects a pitchers performance regardless of an aging pitcher
- Sliders became more effective at yielding strikeouts than 4-seam fastballs
- Shows that Verlander should continue throwing sliders
- Pitch MPH not significant when compared to results whereas pitch type is
    - Focus on pitching technique rather than speed
- Even though our predictors are very significant, our models are not very accurate
- No significant difference in the averages of EV, LA, and distance for Detroit vs Houston, yet Verlander’s strikeout probability increased significantly
- Useful for his new coaching staff for the Mets

Future Research

- To expand upon this project:
  - Perform this on a larger scale and look at other pitcher's and their performance stats
  - Research various pitchers under certain coaches or teams to see if there is any pattern in the performance regardless of the pitcher
  - Include data based on pitch location in or out of the box
    - Could potentially improve the accuracy of our models
  - Analyze the stats of certain batters to determine which pitch type is most effective to throw against them

Thank you for your time!
Questions?