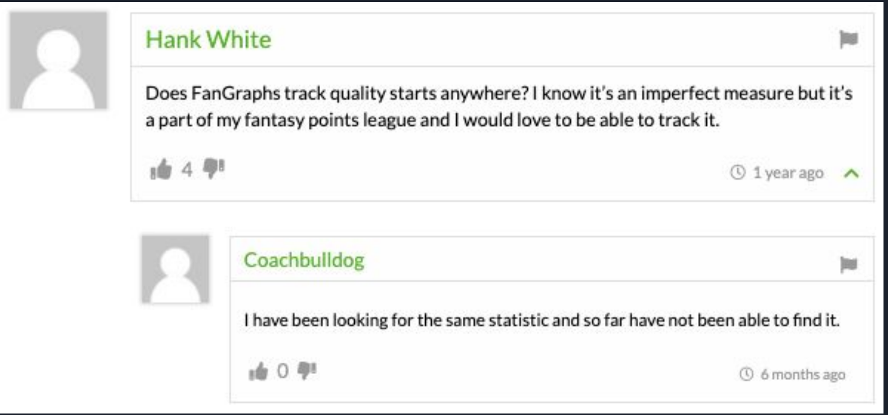
Predicting Quality Starts Using Linear Regression

Metis Project #2
Angeline Protacio
April 17, 2020

Why do we care? What is a quality start?

- Statistic used in fantasy baseball
- Not included in season stat projections
- Quality start projections would help with fantasy baseball draft
- Starting pitcher pitches at least six innings without giving up more than three runs



Methods

Data sources:

- Fangraphs ZiPS Projection Data 2017-2019
  - 4000 observations, 21 features
- Baseball Reference Season Data 2016-2019
  - 1500 observations, 33 features

Libraries:

- Requests
- BeautifulSoup
- Selenium
- Pandas/matplotlib/seaborn
- Scikit-Learn

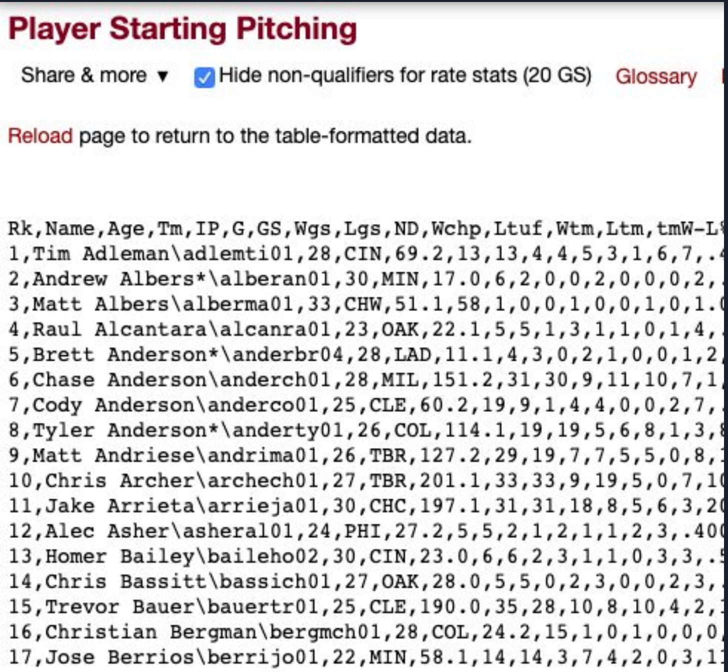
Scraping Fangraphs

- Teams are on different pages
- Year are on different pages
- Three tables to scrape
- Each table has a different numbers of columns
- Wrote a function to do this!

Pitchers, Counting Stats

Player	T	Age	G	GS	IP	K	BB	HR	H	R	ER
Madison Bumgarner	L	27	32	32	211.7	227	44	22	179	71	66
Johnny Cueto	R	31	30	30	207.7	182	48	18	186	75	70
Jeff Samardzija	R	32	29	29	188.3	161	44	19	178	79	74
Matt Moore	L	28	22	22	125.0	116	51	13	115	57	53
Mark Melancon	R	32	70	0	65.0	63	13	5	55	18	17
Jake Peavy	R	36	25	21	123.3	98	36	15	125	60	56

Scraping Baseball Reference



Putting it all together

	Training Data (n = 214)	Validation Data (n = 208)	Test Data (n = 231)
Features	2016 Season Data	2017 Season Data	2018 Season Data
	2017 Projection Data	2018 Projection Data	2019 Projection Data
Target	Quality Starts in 2017	Quality Starts in 2018	Quality Starts in 2019

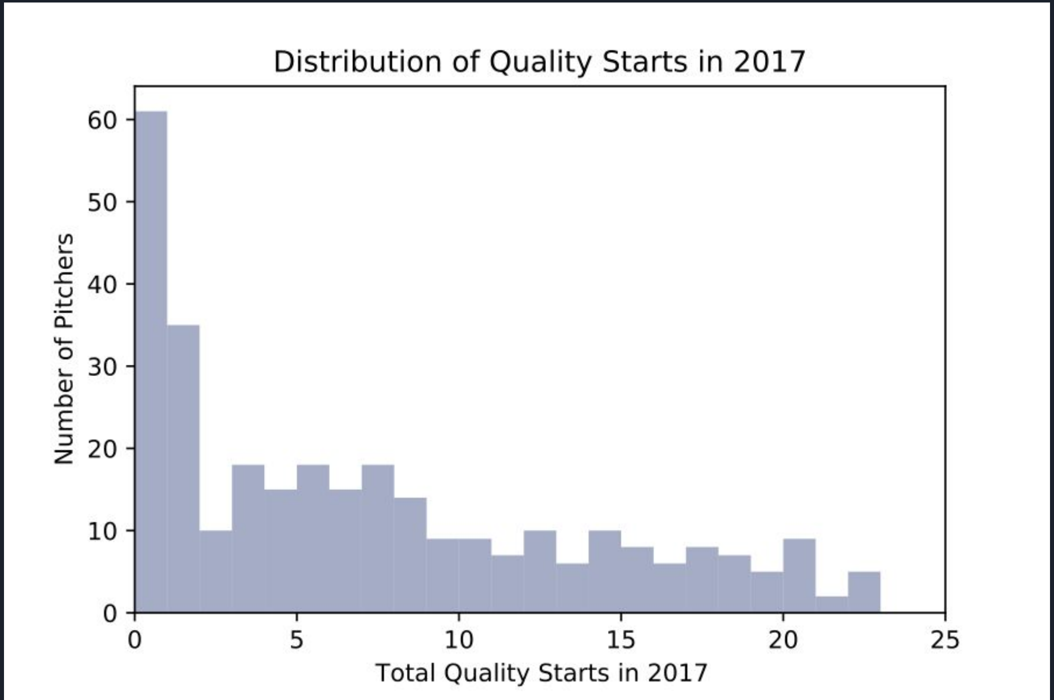
Putting it all together

Angeline
@dataangeline

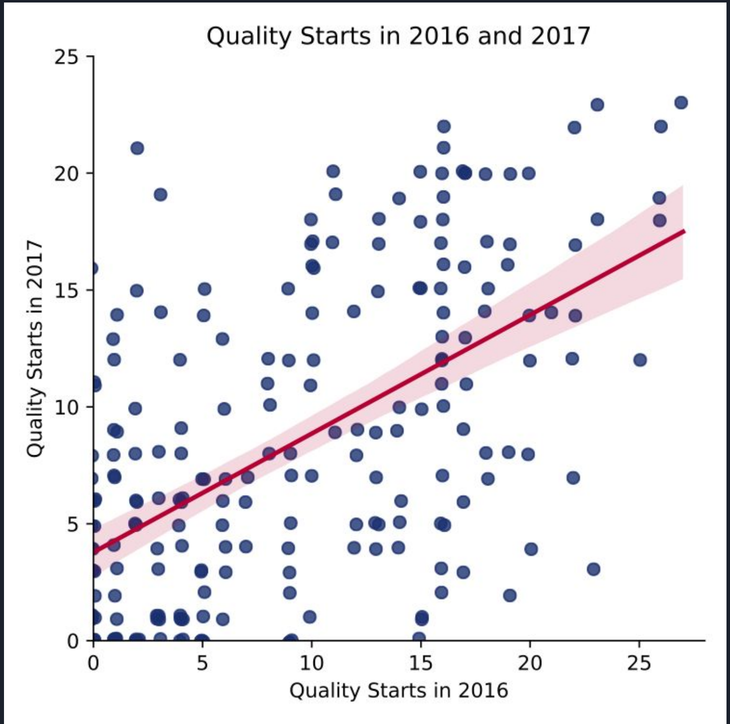
Tonight I watched Will Smith, catcher, make the final out on a dropped third strike thrown by Will Smith, pitcher, while my husband hummed the theme to the Wild Wild West, sung by Will Smith, rapper.

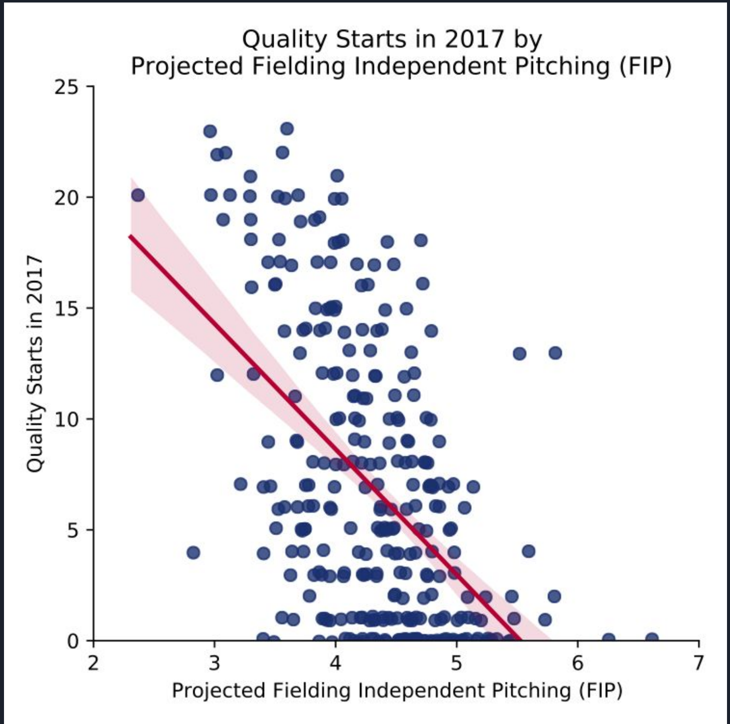
1:43 AM · Sep 7, 2019 · Twitter for Android

Exploratory Data Analysis



Exploratory Data Analysis





Model Results

	R^2 Train/Validation (Projection + Season) n = 214; 208
Selected Features	0.44; 0.35
Selected Features + Polynomial Features	0.80; -4.89
Selected Features + Polynomial Features (LassoCV)	0.35; 0.35

Model Results

	R^2 Train/Validation (Projection + Season) n = 214; 208	R^2 Train/Validation (Projection Only) n = 305; 330
Selected Features	0.44; 0.35	0.42; 0.36
Selected Features + Polynomial Features	0.80; -4.89	0.54; -10
Selected Features + Polynomial Features (LassoCV)	0.35; 0.35	0.39; 0.39

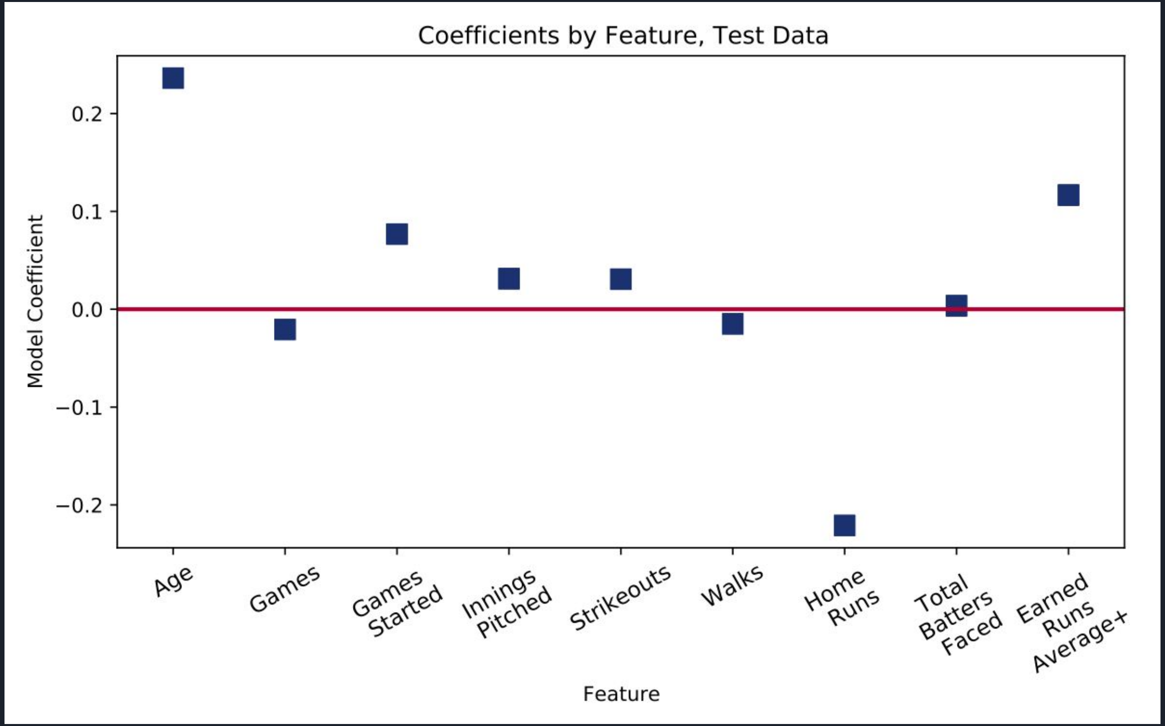
Model Results

	R^2 Train/Validation (Projection + Season) n = 214; 208	R^2 Train/Validation (Projection Only) n = 305; 330
Selected Features	0.44; 0.35	0.42; 0.36
Selected Features + Polynomial Features	0.80; -4.89	0.54; -10
Selected Features + Polynomial Features (LassoCV)	0.35; 0.35	0.39; 0.39
Selected Features (LassoCV)		0.40, 0.39

Model Results

	R^2 Train/Validation (Projection + Season) n = 214; 208	R^2 Train/Validation (Projection Only) n = 305; 330	R^2 Test (Projection Only) n = 359
Selected Features	0.44; 0.35	0.42; 0.36	-0.2
Selected Features + Polynomial Features	0.80; -4.89	0.54; -10	-62807
Selected Features + Polynomial Features (LassoCV)	0.35; 0.35	0.39; 0.39	0.33
Selected Features (LassoCV)	0	0.40; 0.39	0.31

Understanding the Model Results



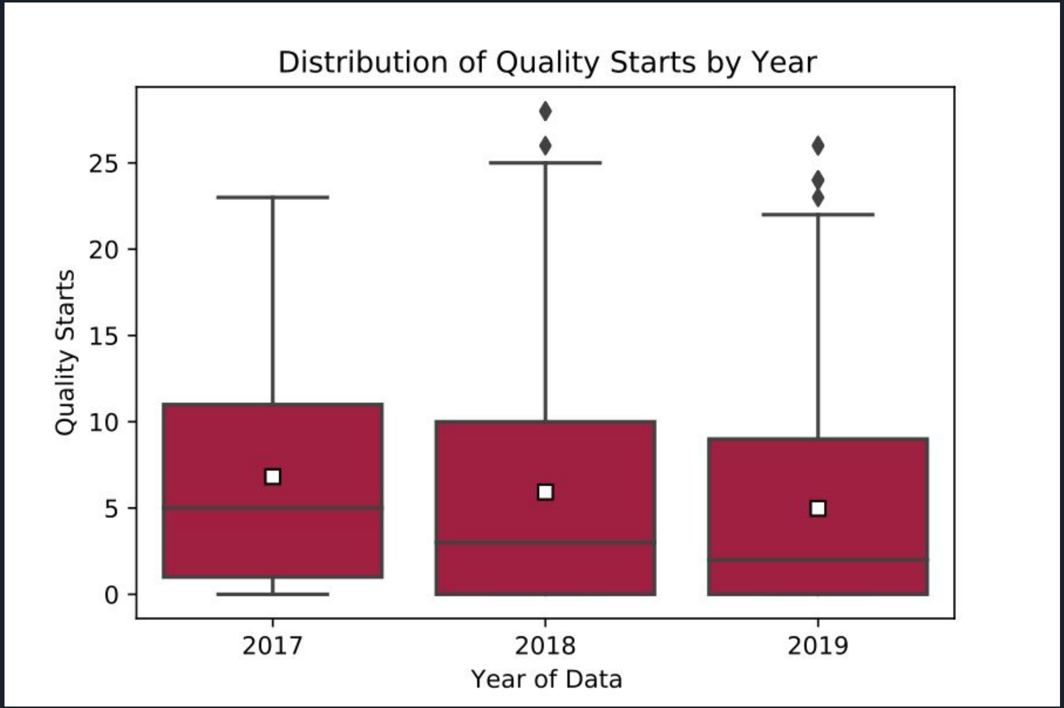
Understanding the Model Results

Full_Name	Age	QS	Predicted QS	Residuals
Hyun-Jin Ryu	32	22	7.064266	14.935734
Lucas Giolito	24	17	5.324824	11.675176
Marco Gonzales	27	19	7.769529	11.230471
Shane Bieber	24	24	13.013365	10.986635
Madison Bumgarner	29	20	9.060545	10.939455

Full_Name	Age	QS	Predicted QS	Residuals
Zack Godley	29	1	11.371593	-10.371593
Chad Green	28	0	11.919603	-11.919603
Carlos Carrasco	32	2	16.352834	-14.352834
Luis Severino	25	0	15.821816	-15.821816
Corey Kluber	33	2	19.703679	-17.703679

Mean Average Error: 4-5 Quality Starts

Understanding the Model Results



Improving the model

- Park factors
- Injury data
- Improve sample size
- Time series analysis

Thank you!

Appendix

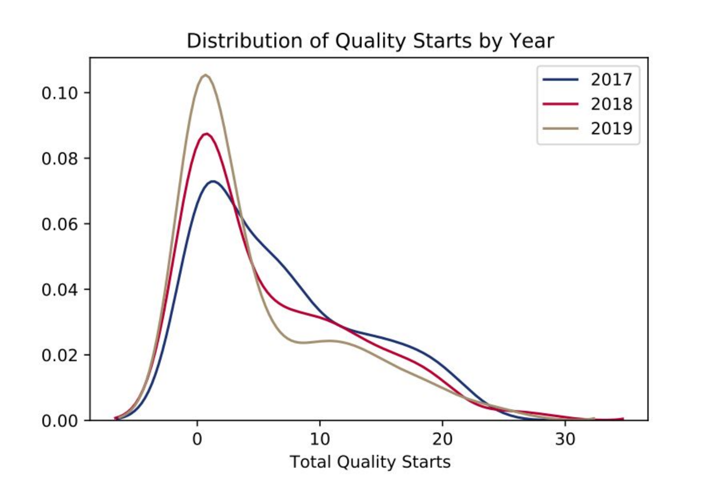
	R^2 Train/Validation (Projection + Season) n = 214; 208	R^2 Train/Validation (Projection Only) n = 305; 330	R^2 Train/Validation (Season Only) n = 214, 208
Selected Features	0.44; 0.35	0.42; 0.36	0.43; 0.32
Selected Features + Polynomial Features	0.80; -4.89	0.54; -10	0.99; -1.69
Selected Features + Polynomial Features (LassoCV)	0.35; 0.35	0.39; 0.39	0.34; 0.32

Appendix

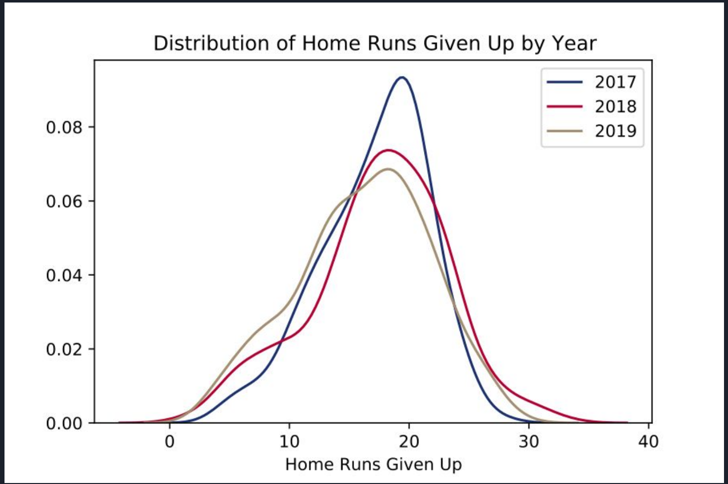
ZiPS Projections

- Developed by Dan Szymborski
- Uses growth and decline curves based on player type to find trends
- Factors trends into past performance to come up with projections
- Uses statistics from prior four years for players 24-38, recent data more heavily weighted
- Younger and older players use prior three years
- Includes velocity, injury data, and play-by-play data into equations

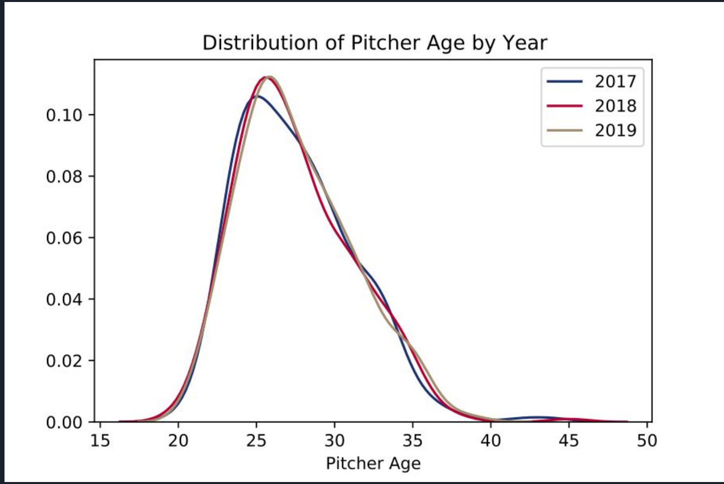
Appendix



Distribution of Quality Starts by Year
2017
2018
2019
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
Total Quality Starts

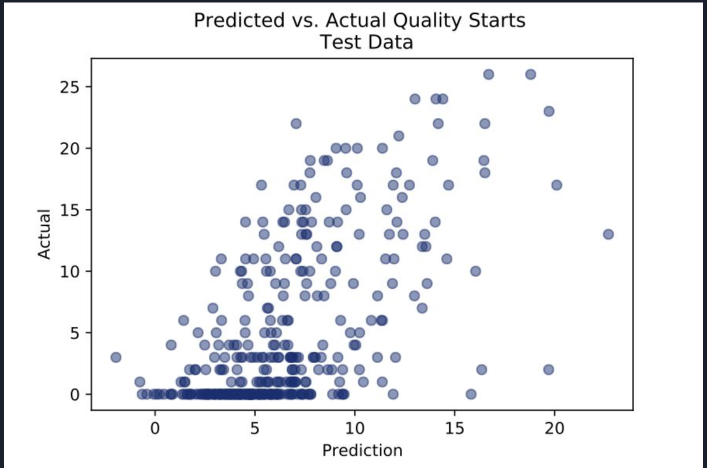


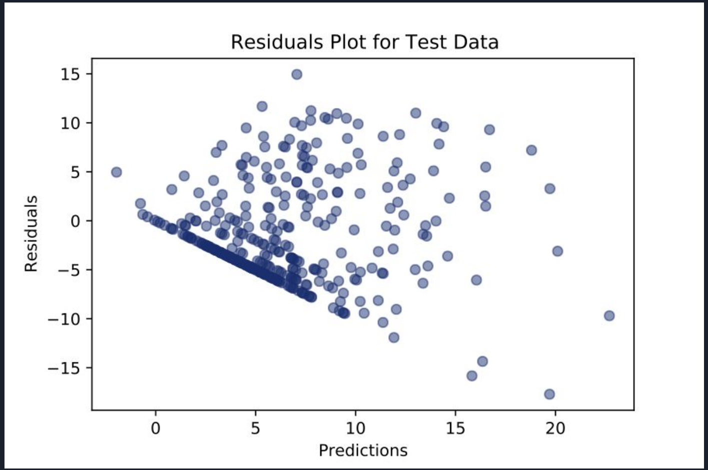
Distribution of Home Runs Given Up by Year
2017
2018
2019
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
Home Runs Given Up



Distribution of Pitcher Age by Year
2017
2018
2019
0.10
0.08
0.06
0.04
0.02
0.00
15
20
25
30
35
40
45
50
Pitcher Age

Appendix


Predicted vs. Actual Quality Starts
Test Data
Actual
Prediction
0
5
10
15
20
25


Residuals Plot for Test Data
Residuals
Predictions
15
10
5
0
−5
−10
−15
20