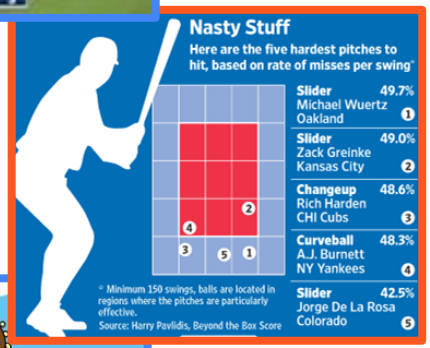
Statistics in Baseball Pitching

Aaron Wu, Ruchi Tiwari, Edwin Muñoz, Shakty Jarez

Project Motivations

- Oldest professional American sport
- 8-Figure Industry: ticket sales, merchandise purchases, **sports betting**
- Sports Analytics looks at
  - Optimizing Player Performances
  - Predicting Wins/Championships
  - Drafting/Trading for Players



Understanding Key Parts

- Two Types of Players: Hitters/**Pitchers**
- Primary Dependent Variable: **ERA** (Earned Run Average)
- Look at everything from how hard a ball is hit, the speed of a pitch, direction the pitch spins, types of pitches,etc

Our goal is to gain a competitive advantage when trying to make predictions regarding baseball

Dataset Introduction

106 Regressors

"last_name..first_name"	"player_id.x"	"year.x"	"player_age"
"p_game"	"p_formatted_ip"	"pa"	"k_percent"
"bb_percent"	"xba"	"xslg"	"woba"
"xwoba"	"solidcontact_percent"	"poorlyweak_percent"	"hard_hit_percent"
"whiff_percent"	"swing_percent"	"groundballs_percent"	"flyballs_percent"
"linedrives_percent"	"popups_percent"	"pitch_hand"	"n"
"n_ff_formatted"	"ff_avg_speed"	"ff_avg_spin"	"ff_avg_break_x"
"ff_avg_break_z"	"ff_avg_break"	"ff_range_speed"	"n_sl_formatted"
"sl_avg_speed"	"sl_avg_spin"	"sl_avg_break_x"	"sl_avg_break_z"
"sl_avg_break"	"sl_range_speed"	"n_ch_formatted"	"ch_avg_speed"
"ch_avg_spin"	"ch_avg_break_x"	"ch_avg_break_z"	"ch_avg_break"
"ch_range_speed"	"n_cu_formatted"	"cu_avg_speed"	"cu_avg_spin"
"cu_avg_break_x"	"cu_avg_break_z"	"cu_avg_break"	"cu_range_speed"
"n_si_formatted"	"si_avg_speed"	"si_avg_spin"	"si_avg_break_x"
"si_avg_break_z"	"si_avg_break"	"si_range_speed"	"n_fc_formatted"
"fc_avg_speed"	"fc_avg_spin"	"fc_avg_break_x"	"fc_avg_break_z"
"fc_avg_break"	"fc_range_speed"	"n_fs_formatted"	"fs_avg_speed"
"fs_avg_spin"	"fs_avg_break_x"	"fs_avg_break_z"	"fs_avg_break"
"fs_range_speed"	"n_kn_formatted"	"kn_avg_speed"	"kn_avg_spin"
"kn_avg_break_x"	"kn_avg_break_z"	"kn_avg_break"	"kn_range_speed"
"n_st_formatted"	"st_avg_speed"	"st_avg_spin"	"st_avg_break_x"
"st_avg_break_z"	"st_avg_break"	"st_range_speed"	"n_sv_formatted"
"sv_avg_speed"	"sv_avg_spin"	"sv_avg_break_x"	"sv_avg_break_z"
"sv_avg_break"	"sv_range_speed"	"n_fo_formatted"	"fo_avg_speed"
"fo_avg_spin"	"fo_avg_break_x"	"fo_avg_break_z"	"fo_avg_break"
"fo_range_speed"	"n_sc_formatted"	"sc_avg_speed"	"sc_avg_spin"
"sc_avg_break_x"	"sc_avg_break_z"	"sc_avg_break"	"sc_range_speed"
"player_id.y"	"year.y"	"p_win"	"p_loss"
"p_era"			

Why did we pick baseball?

- Hundreds of regressors to pick from
- Every single stat has been tracked for decades
- Once defined, simple to understand most statistics

Understanding the Dataset

- Total of 106 Regressors
- Immediately eliminate about half:
  a. Too niche
  b. Not enough data
  c. 0 correlation

player_id.x <int>	year.x <int>	player_age <int>	p_game <int>	p_formatted_ip <dbl>	pa <int>	k_percent <dbl>
645261	2023	27	28	184.2	762	19.8
671106	2023	24	24	125.1	537	22.2
542881	2023	33	27	141.0	629	18.9
668933	2023	25	26	145.2	624	17.8
605135	2023	34	33	200.0	826	22.5
678394	2023	24	28	157.0	668	19.8

Wrapper Method

Wrapper Method: Focuses on finding which combination of predictors works best to explain or predict the target outcome.

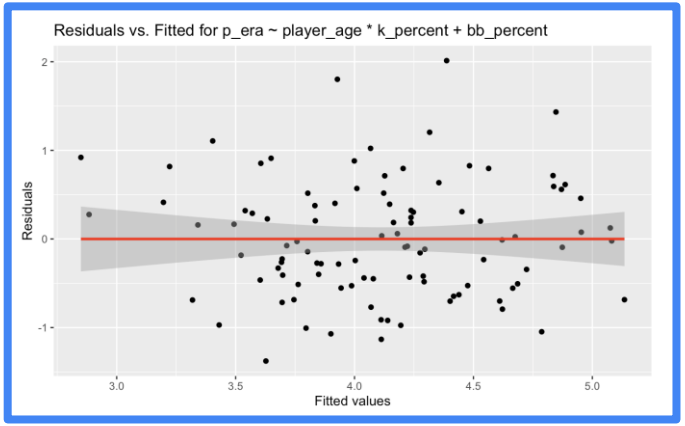
Models Tested:

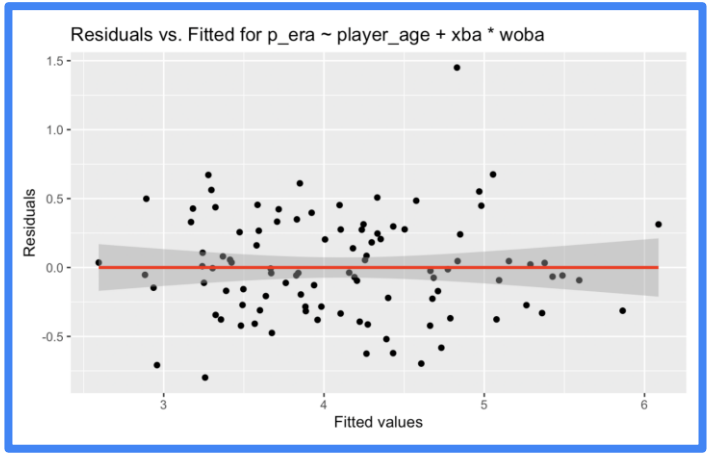
1. **k_percent** (strikeout rate) and **bb_percent** (walk rate) (adj R^2 = 0.2939)
2. **woba** (weighted on-base average) and whiff_percent (swing-and-miss rate). (adj R^2 = 0.7965)
3. player_age, interaction of xba (expected batting average) and woba. (adj R^2 = 0.7994)
4. swing_percent, hard_hit_percent, and **whiff_percent** (adj R^2 = 0.1634)
5. Interaction of **player_age** and **k_percent**, plus **bb_percent** (adj R^2 = 0.3554)

*bold - indicates if a variable was significant

Validation - Residual Plots

- **Residual vs. Fitted Plots:**
  - Shows if there are patterns in the model's errors: Non-Random Pattern, Heteroscedasticity, Outliers, Clusters or Groups
- **Analysis:**
  - All plots showed no clear random patterns, suggesting the models captured key relationships.
  - Some outliers and influential points were present, did not pose any serious concern at this stage





Validation - AIC

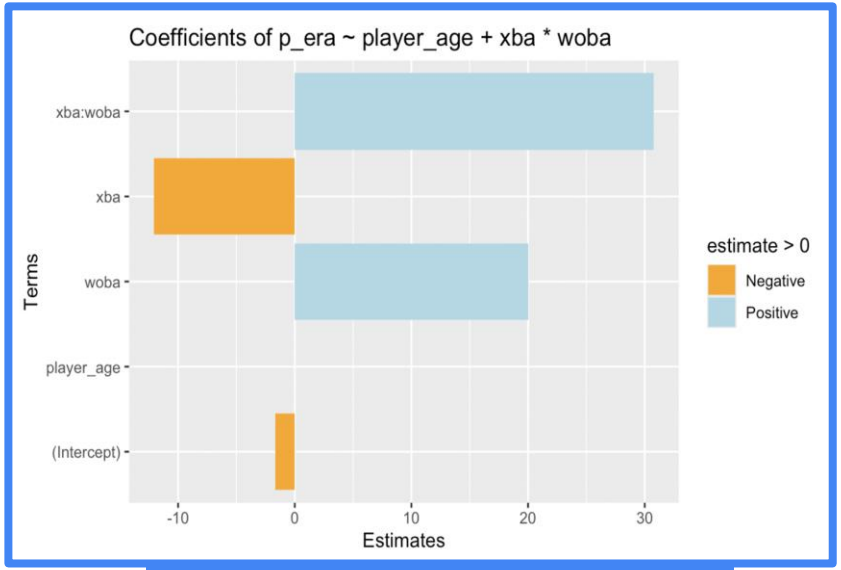
- Models were compared using **AIC (Akaike Information Criterion):**
  - Helps us avoid making models too complex just to fit the data better.
  - Lower AIC means a better balance between fit and simplicity

lm_1	lm_2	lm_3	lm_4	lm_5
216.28381	91.88718	94.42964	235.24084	211.17271

- **lm_2** is the best model (AIC = 91.89), effectively predicting ERA using woba and whiff_percent.
- **lm_3 (AIC = 94.43)** performs slightly worse, using player_age and interaction of xba with woba.
- **lm_4, lm_1, and lm_5** have much higher AIC values, making them less suitable, we need to further explore variables

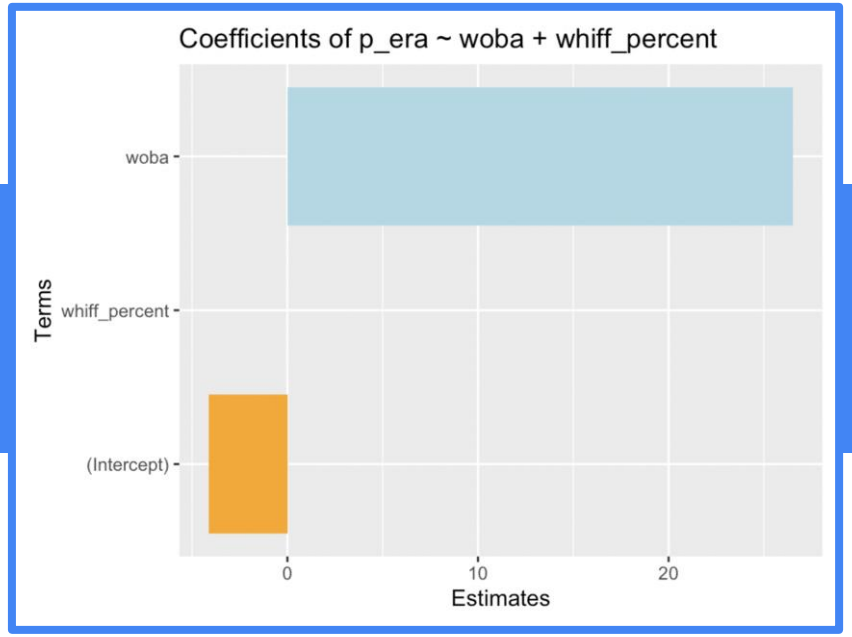
Coefficient Plots

- **Visualizing coefficients:** Makes it clear which variables matter most and how they impact ERA.
    - Bar plots show the strength and direction of each variable's impact:
        - **Positive coefficients** (e.g., weighted on-base average (woba) increasing ERA).
        - **Negative coefficients** (e.g., expected batting average (xba) reducing ERA).

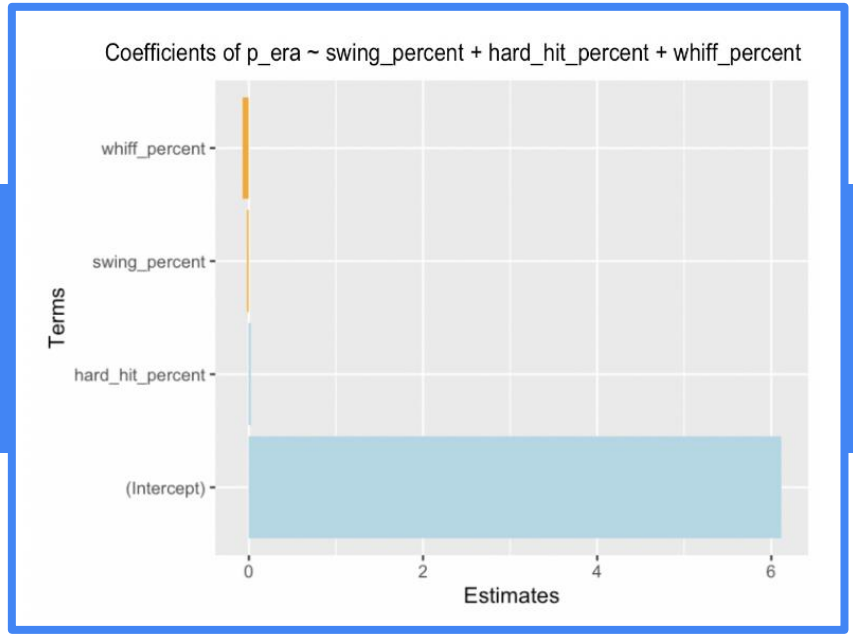


Im 3 playerage not impactful

More Coefficient Plots



lm 2 woba highly influential
whiff_percent relatively less



lm 4 whiff_percent relatively
more impactful

Initial Model

- **Introducing Initial Model**
  - Predictors included: Significance, Theoretical Relevance, Interactions, Relative Comparison Predictors

```
p_era ~ k_percent + bb_percent + whiff_percent + woba + xba:woba
```

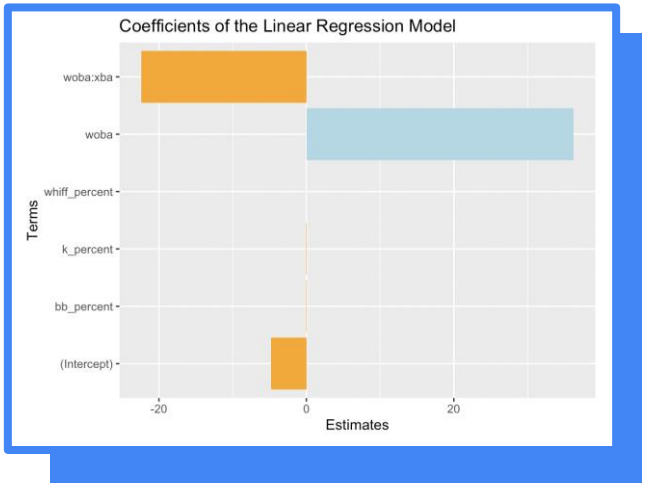
  - ➔ **k_percent**: Strikeout percentage
  - ➔ **bb_percent**: Walk percentage
  - ➔ **whiff_percent**: Swing-and-miss percentage
  - ➔ **woba**: Weighted on-base average
  - ➔ **xba**: Interaction term between expected batting average (**xba**) and **woba**

- **Removed:**
  - **Player_age-** Although age may influence ERA (older players have experience, younger players have more physical ability). The coefficient plot showed a weak, inconsistent effect
  - **Swing_percent -** This variable showed a weak both the coefficient plot and summary statistics. Redundancy with whiff_percent was included, contributing little unique information to the model.
  - **Hard_hit_percent -** Similar to swing_percent, this variable had a weak or insignificant effect on ERA. It overlapped with woba and xba, which better explain batter performance against pitchers, making it unnecessary in the model.

Evaluating our Initial Model

Model Creation:

- We fit a linear model using the predictors: k_percent, bb_percent, whiff_percent, woba, and the interaction term woba:xba



Key Insights from Coefficients:

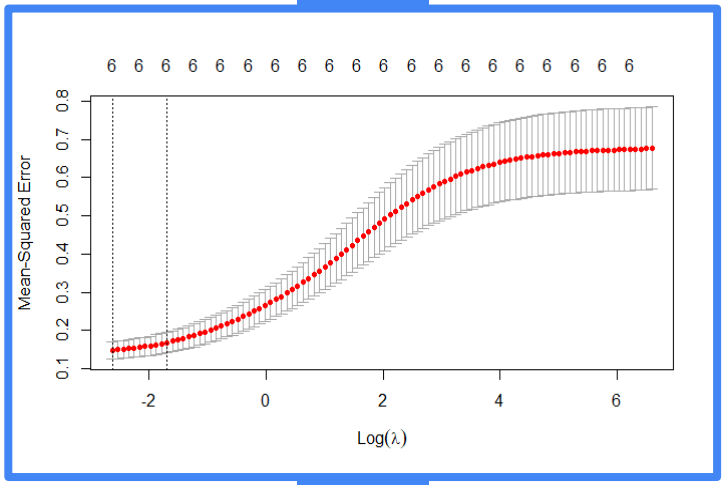
- woba strongly increases ERA.
- Interaction term woba:xba has a significant negative impact on ERA.

Checking for Multicollinearity (Variance Inflation Factor):

- Analysis shows a high VIF for woba, xba, and woba:xba
  - Indicates multicollinearity, suggesting a need for alternative modeling techniques (e.g., Ridge regression)

k_percent	bb_percent	whiff_percent	woba	xba	woba:xba
5.765558	1.665863	4.609108	127.518518	117.732045	383.278966

Optimizing with Ridge Regression



	s0
(Intercept)	-1.587411966
k_percent	-0.009096944
bb_percent	-0.037764256
whiff_percent	0.003090086
woba	20.879342654
xba	-7.850279908
woba:xba	19.597393033

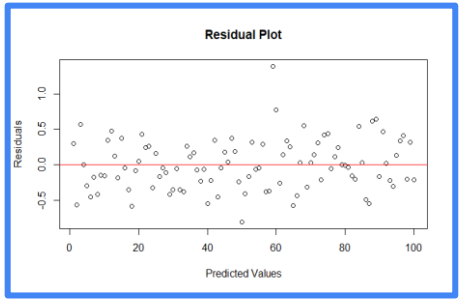
Goal of Ridge Regression: Handle multicollinearity by adding a penalty to reduce large coefficient magnitudes.

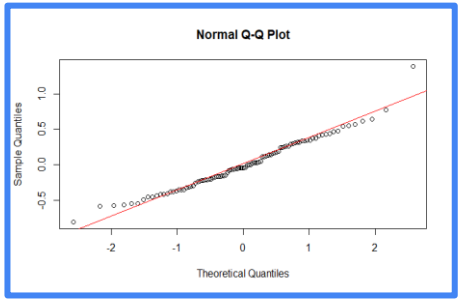
- **We built Ridge Model using:**
  - Predictor variables: k_percent, bb_percent, whiff_percent, woba, and the interaction woba:xba.
  - Range of lambda values tested to determine optimal regularization strength.
  - Cross-validated (10-fold) to find the best lambda value for minimizing Mean Squared Error (MSE).
- **Results:**
  - Optimal lambda value selected through cross-validation.
  - Coefficients indicate reduced multicollinearity effects

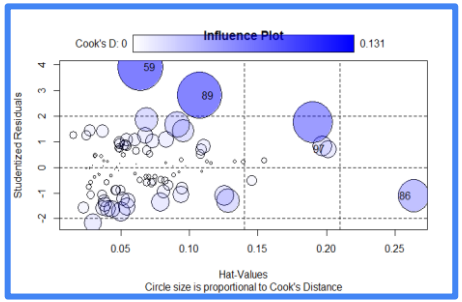
Model Diagnostics

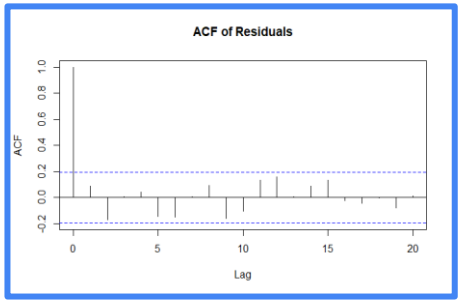
Model Residual Analysis:

- Residual Plot: No strong patterns observed, suggesting a reasonable fit.
- Additional Checks:
  - QQ-Plot for normality of residuals
  - Influence Plot to identify leverage points, Autocorrelation Checks
  - Durbin-Watson Test for autocorrelation
  - Ljung-Box Test for serial correlation










```
        Box-Ljung test

data:  residuals(model)
X-squared = 0.77085, df = 1, p-value = 0.38
```

```
       Durbin-Watson test

data:  model
DW = 1.8042, p-value = 0.1791
alternative hypothesis: true autocorrelation is greater than 0
```

Conclusions:

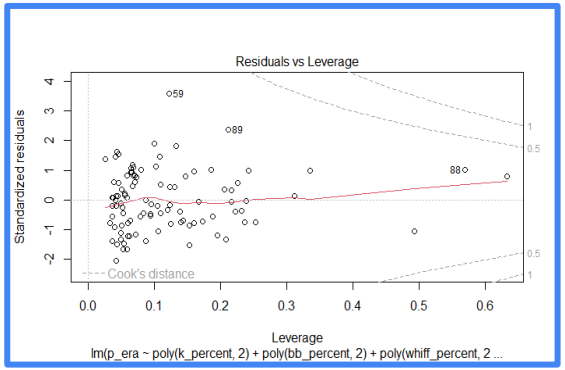
- Ridge model performed well with limited multicollinearity.
- Residual and diagnostic plots indicate a good fit.
- Next steps involve exploring polynomial features and LASSO regression for further optimization.

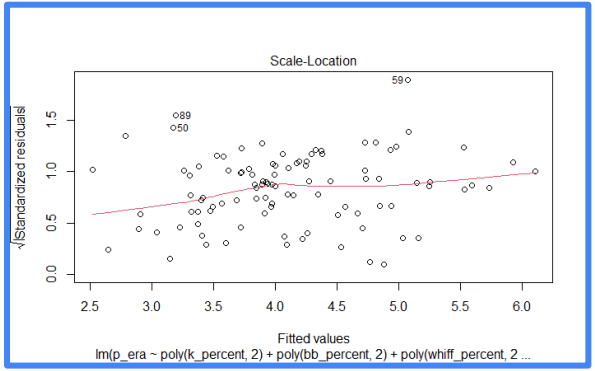
Polynomial Regression

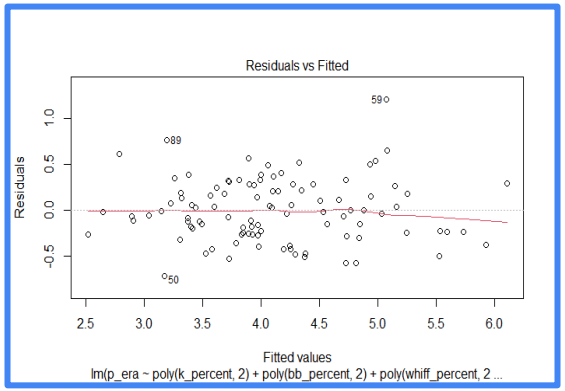
Goal: Capture more complex, non-linear relationships by transforming variables to polynomial terms.

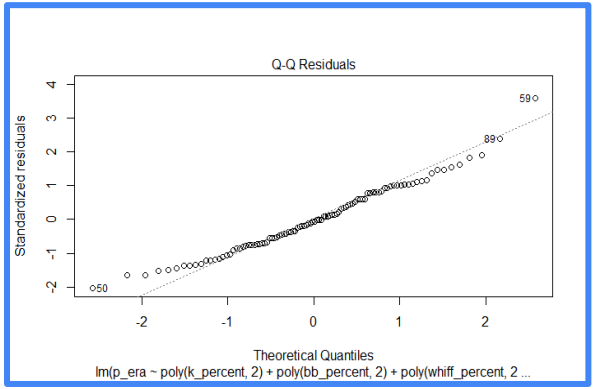
Pros: Slightly improved fit to training data.

Cons: Prone to overfitting, risking poor performance on new data and potential homoscedasticity issues (variance inconsistency).









LASSO Regression

Why LASSO?

- Reduces complexity by shrinking coefficients of less important variables to zero.

Steps:

1. Built LASSO Model using predictor matrix.
2. Identified optimal lambda value using cross-validation.
3. Visualized LASSO results to examine non-zero coefficients.

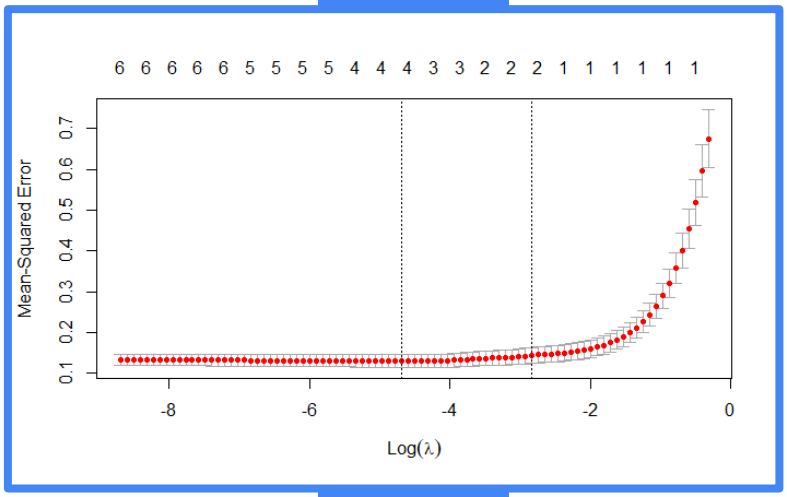
Key Insights from LASSO Model:

Identified two important variables: whiff_percent and interaction term xba:woba.

➔ Suggested possible redundancy or high complexity in certain features.

Next Steps:

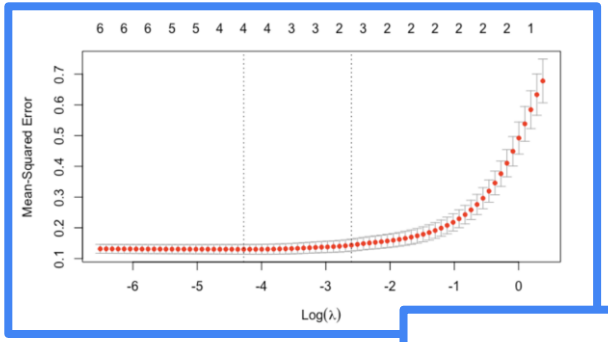
Elastic Net Regression: Combines LASSO and Ridge penalties to balance feature selection and regularization.



	s0
(Intercept)	-3.571597421
k_percent	-0.001962429
bb_percent	-0.053073298
whiff_percent	.
woba	29.391493117
xba	-4.281871487
woba:xba	.

Elastic Net Regression

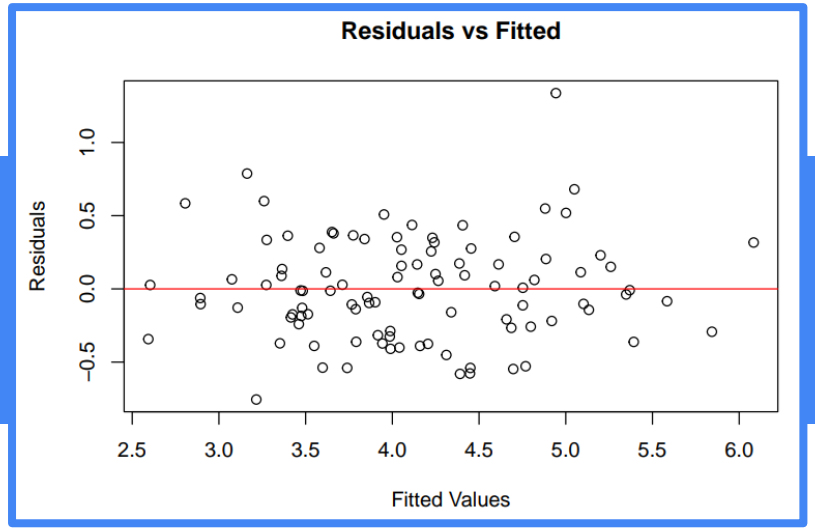
- Our final model uses a net elastic regression model to predict ERA based on the factors of
  - k_percent
  - bb_percent
  - woba
  - xba

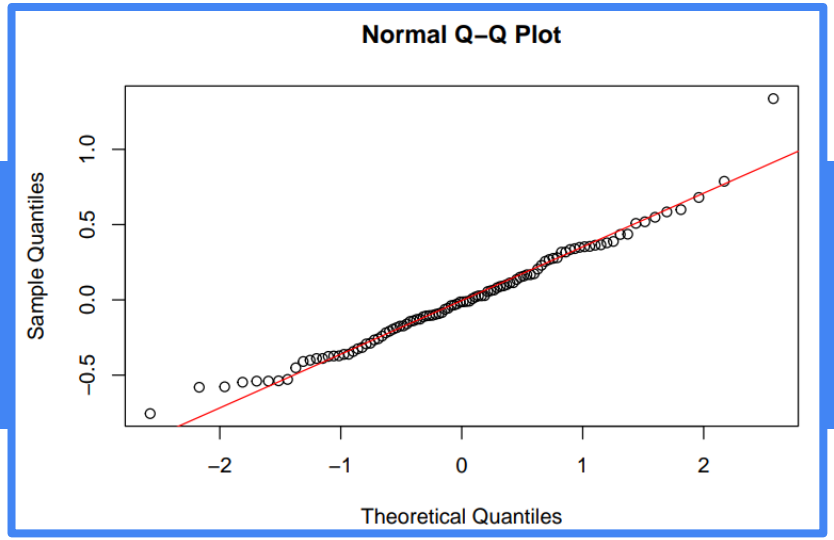


	s0
(Intercept)	-3.354298923
k_percent	-0.004638614
bb_percent	-0.052643783
whiff_percent	.
woba	29.107008140
xba	-4.562269425
woba:xba	.

- Our model found correlation between regressors and ERA with a correlation coefficient of 0.8086
- Overall, data had randomness as baseball overall has significance variance from season to season, game to game, and even pitch to pitch

Elastic Net Regression

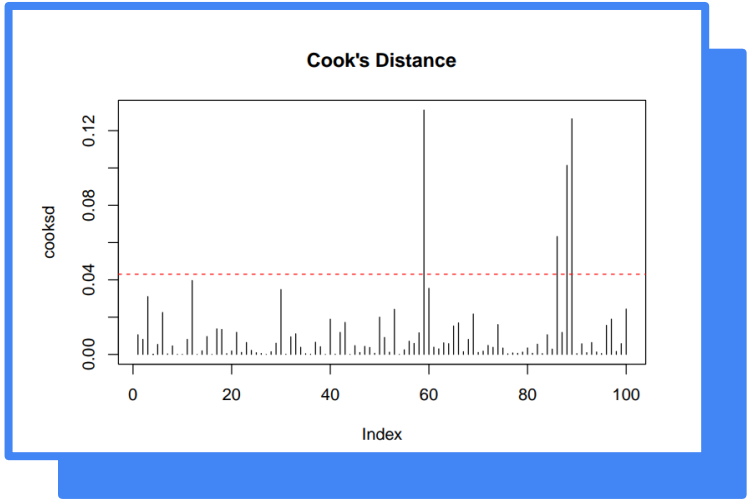


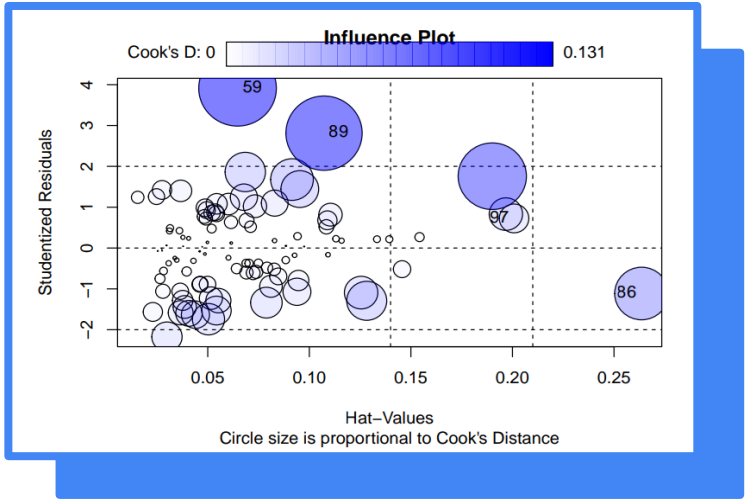


- ➔ Evenly distributed
- ➔ Only a few outliers

- ➔ Better job at capturing fitted values
- ➔ Once again, only a few outliers

Outliers

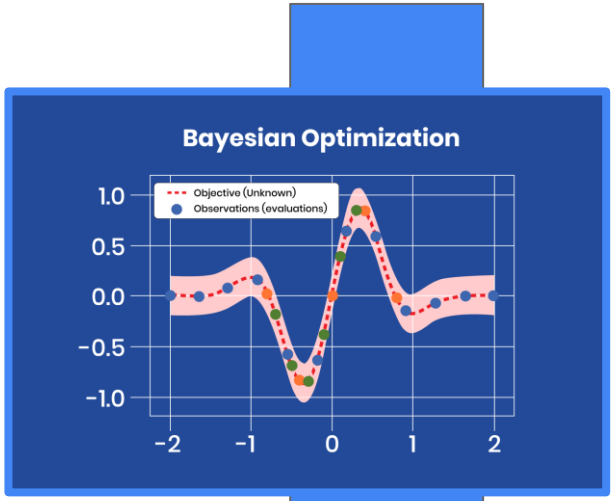


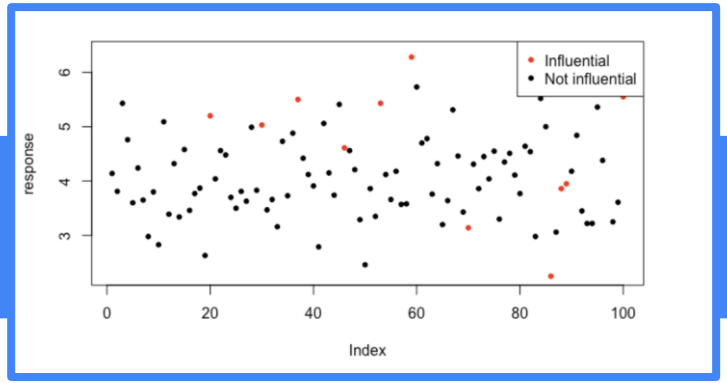


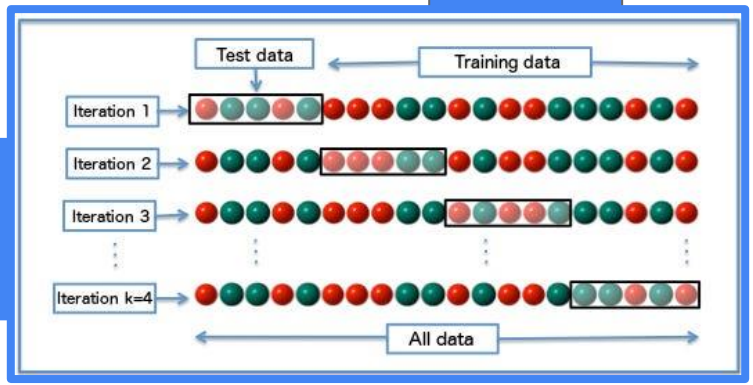
- ➔ As in every sport, there are some elite athletes that outperform most other athletes by a large margin
- ➔ These athletes have a massive influence on the data, making it difficult to create a “perfect” model

Future Work

- Parameter Tuning
- Additional Predictors
- Outlier Observation
- Cross-Validation







References

https://www.kaggle.com/datasets/vivovinco/2023-mlb-player-stats

https://www.blessyouboys.com/2019/1/9/18172095/baseball-stats-for-beginners-earned-run-average-field-independent-pitching-explained

https://www.mlb.com/glossary/standard-stats/walks-and-hits-per-inning-pitched

https://baseballsavant.mlb.com/savant-player/shohei-ohtani-660271?stats=statcast-r-pitching-mlb&playerType=pitcher

Questions?