

Baseball And Linear Regression

***Predicting Wins
Scott Lew***



The Goal

Predict a baseball team's wins in a season using a variety of batting & pitching statistics.

Precedent for Predicting Wins: Bill James

- ‘Godfather’ of Sabermetrics
- Statistician and baseball historian
- “Pythagorean Theorem of Baseball”
Predicts Win% using Runs scored
& Runs allowed



photo:Keith Philpott for TIME magazine

Methodology

- Problem: Predict baseball team wins using batting & pitching stats
- Web scraping to obtain statistics from 1876-2018. Linear Regression Models to predict wins.
- Evaluate Regression Models
- Conclusions: What are some important predictors of wins?

Resources & Tools

- BeautifulSoup: web scraping
- www.baseball-reference.com: batting & pitching statistics
- scikit learn: regression model
- statsmodel: regression model

Some Potential Batting Stats To Predict Wins

- Hits (H): Hits produce runs
- Runs scored: Runs help win games
- Batting Average (BA): Good batting produces hits
- On Base Percentage (OBP): Getting on base is helpful

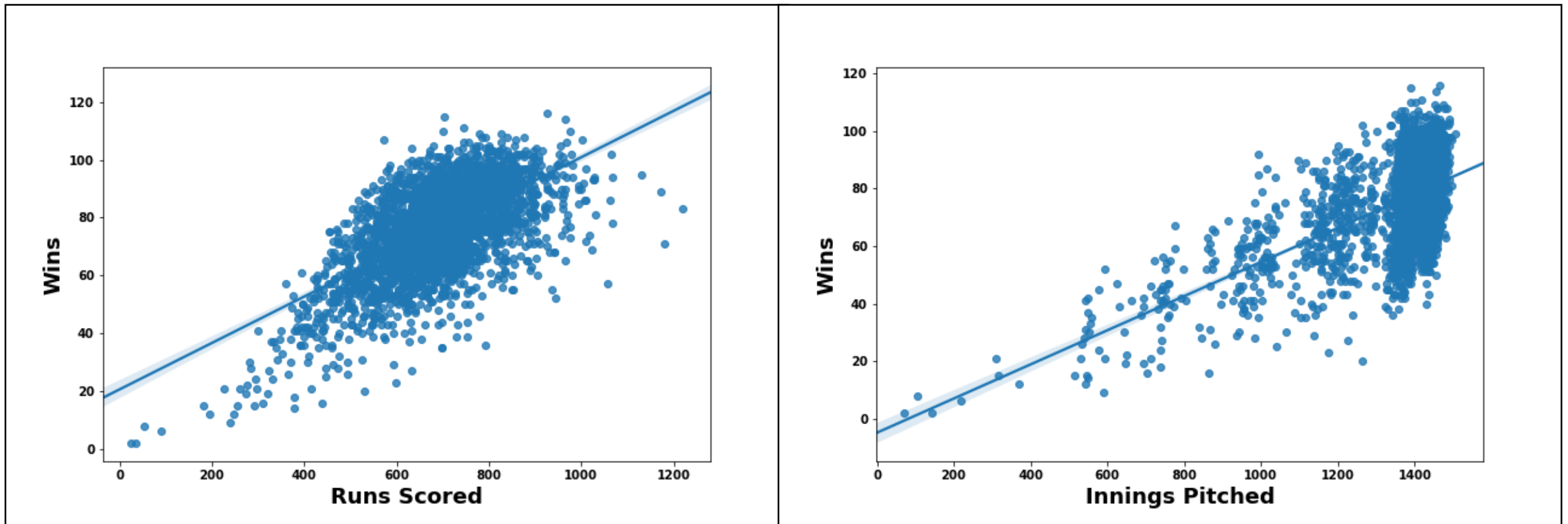
Potential Pitching Stats To Predict Wins

- Earned Runs Average (ERA): metric for pitching, the lower the better.... fewer runs allowed
- Save (SV): A good metric for relief pitchers performance
- Innings Pitched (IP): Another metric for good pitching

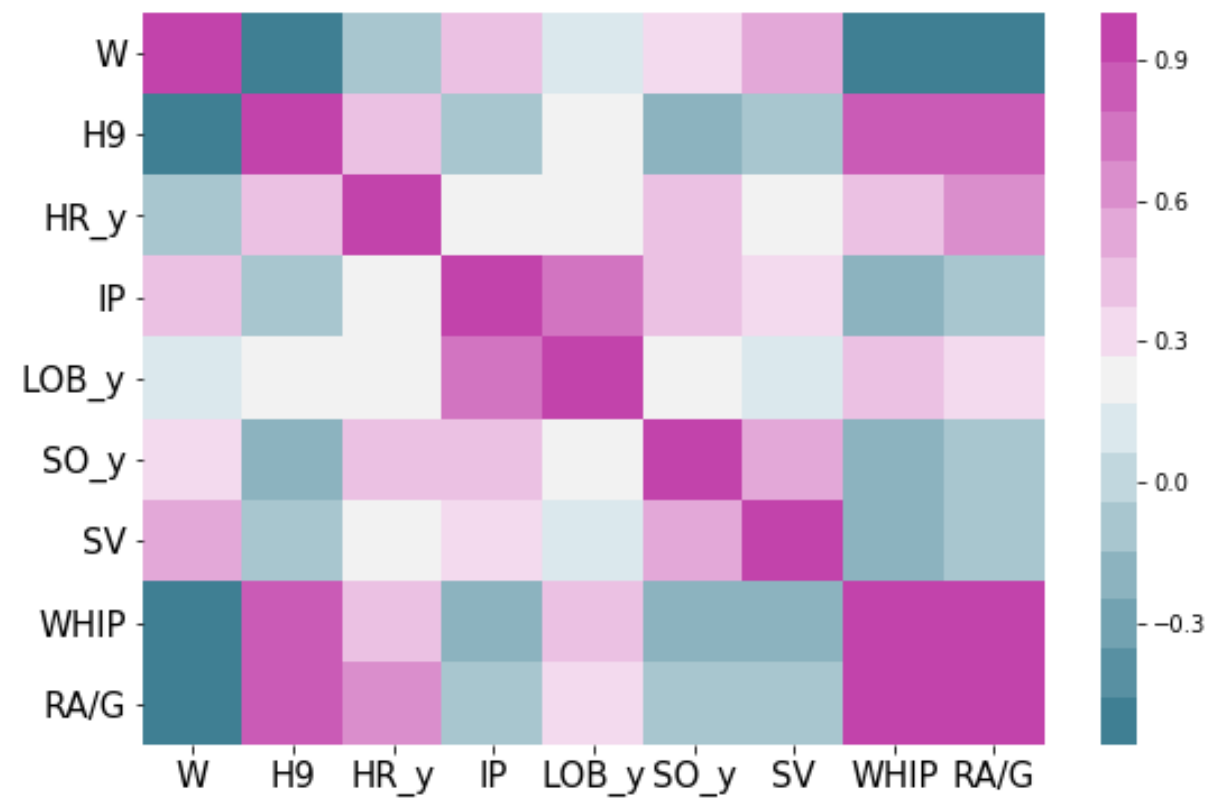
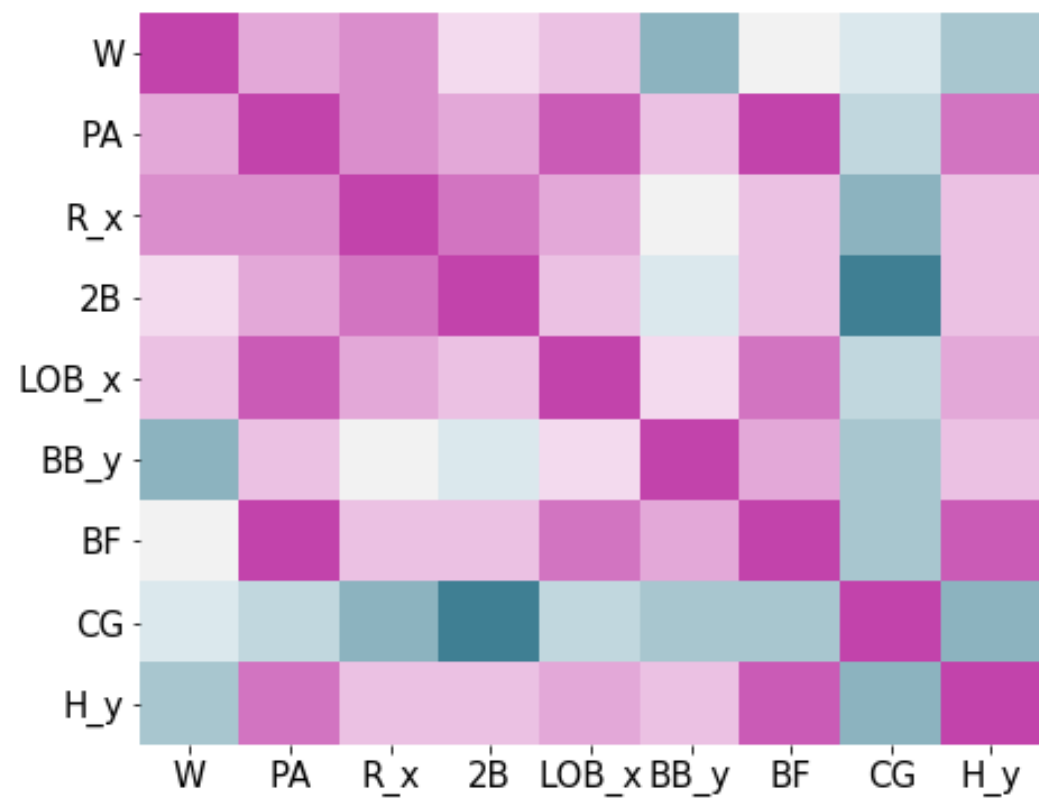
Dataset

- MLB Stats from 1876- 2018
- 28 Batting Stats
- 35 Pitching Stats

EDA: Stats With Correlation To Wins



Heatmap: Correlation of Stats with Wins

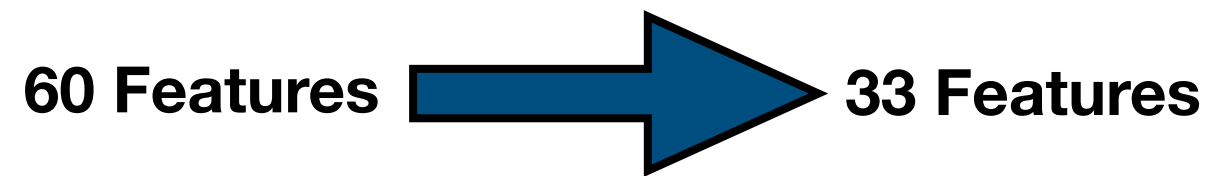


Runs scored (R_x), Innings Pitched (IP), & Save (SV): all have positive correlation with wins

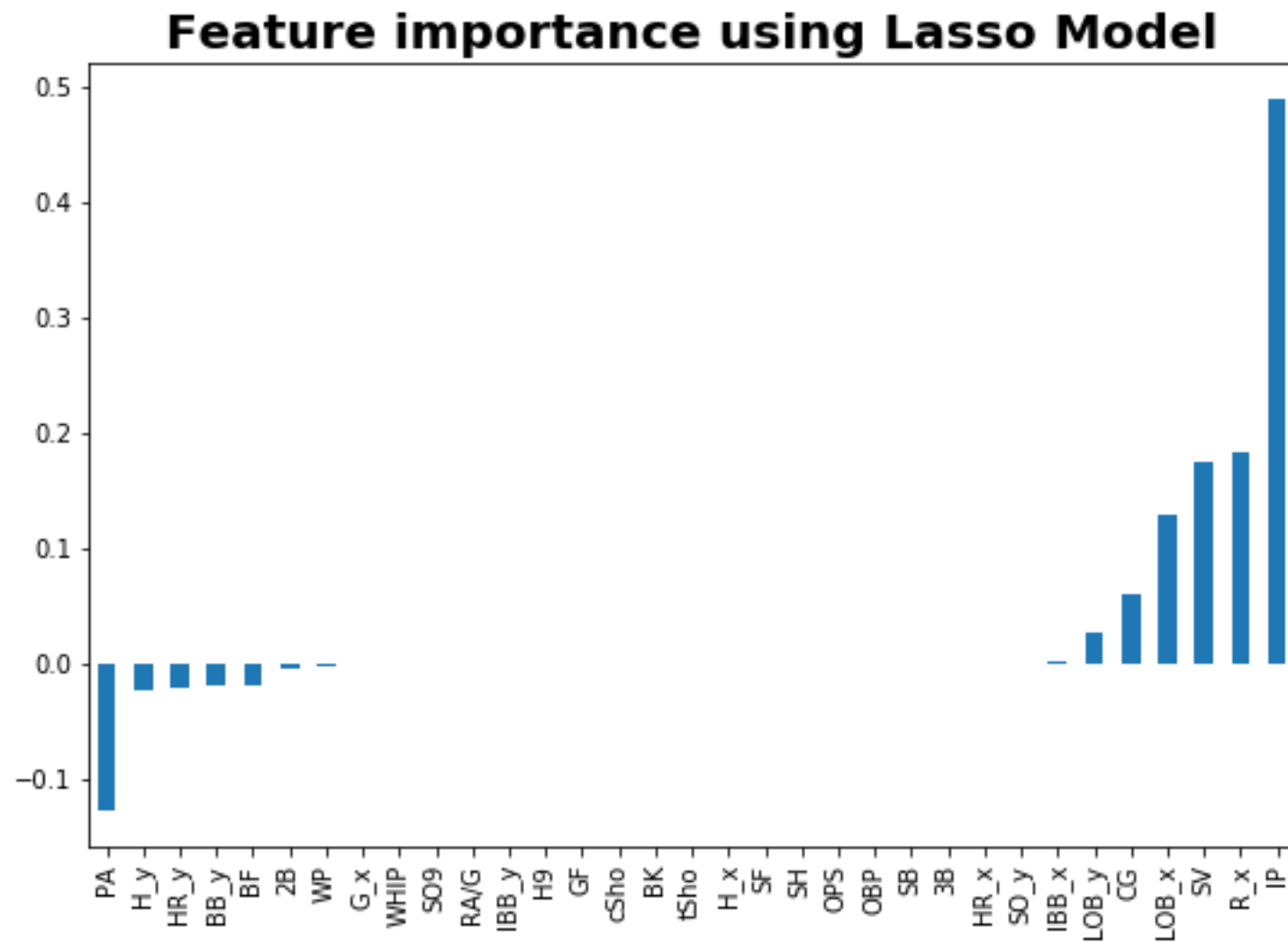
OLS Regression

- Perform multiple linear regression with statsmodels and Scikit Learn using all stats for initial regression model
- Select features with backwards stepwise regression
- Feature selection using LASSO

Backwards Stepwise Regression



LASSO Selection



33 Features



16 Features

Some Stats Used To Predict Wins

=====

	coef	std err	t	P> t	[0.025	0.975]

PA	-0.2743	0.006	-49.435	0.000	-0.285	-0.263
R_x	0.2919	0.004	68.689	0.000	0.284	0.300
LOB_x	0.2771	0.006	48.856	0.000	0.266	0.288
IP	0.6798	0.027	25.344	0.000	0.627	0.732
SV	0.0742	0.012	6.418	0.000	0.052	0.097

=====

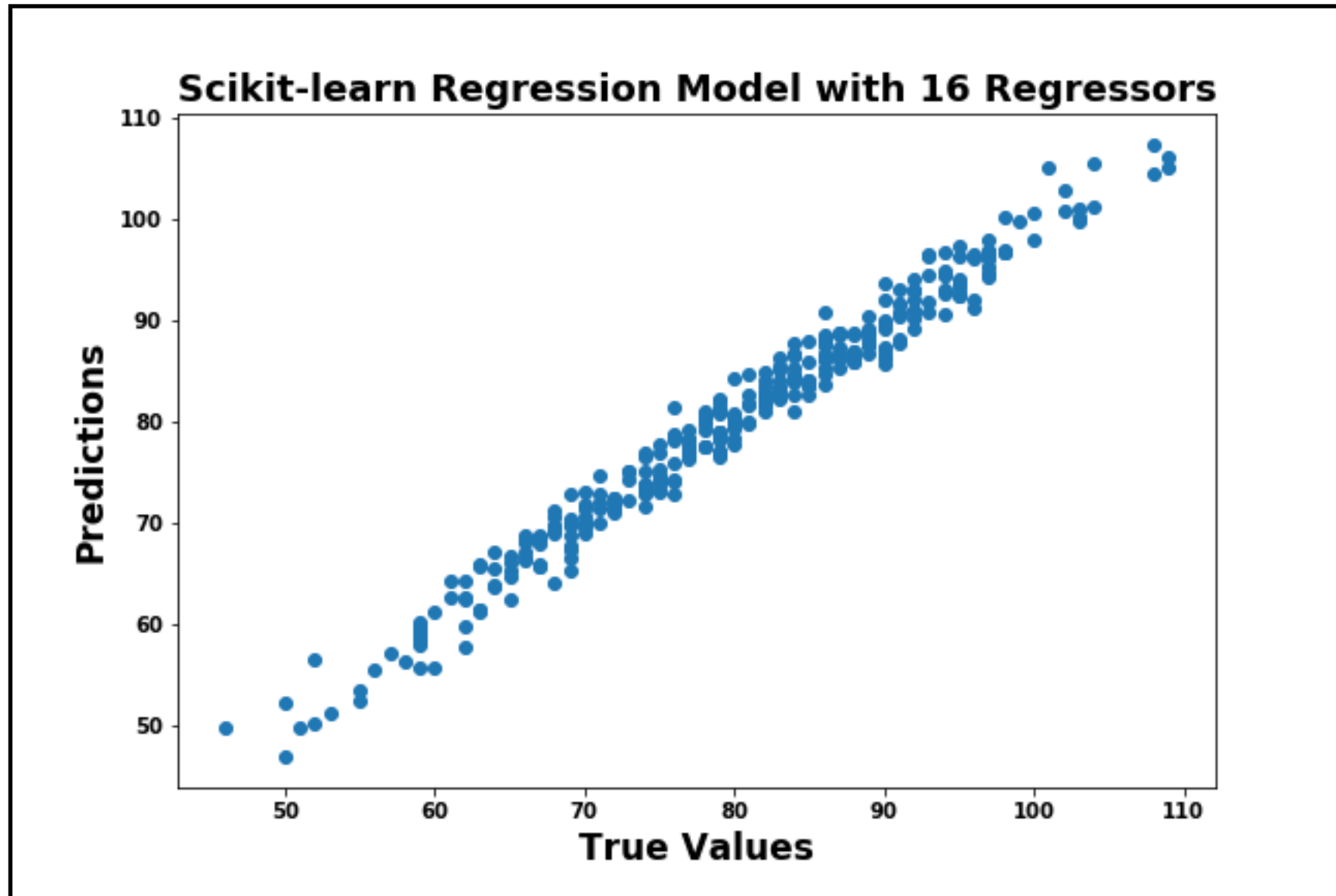
Statsmodel OLS

All the above features are significant in predicting wins

Linear OLS Model

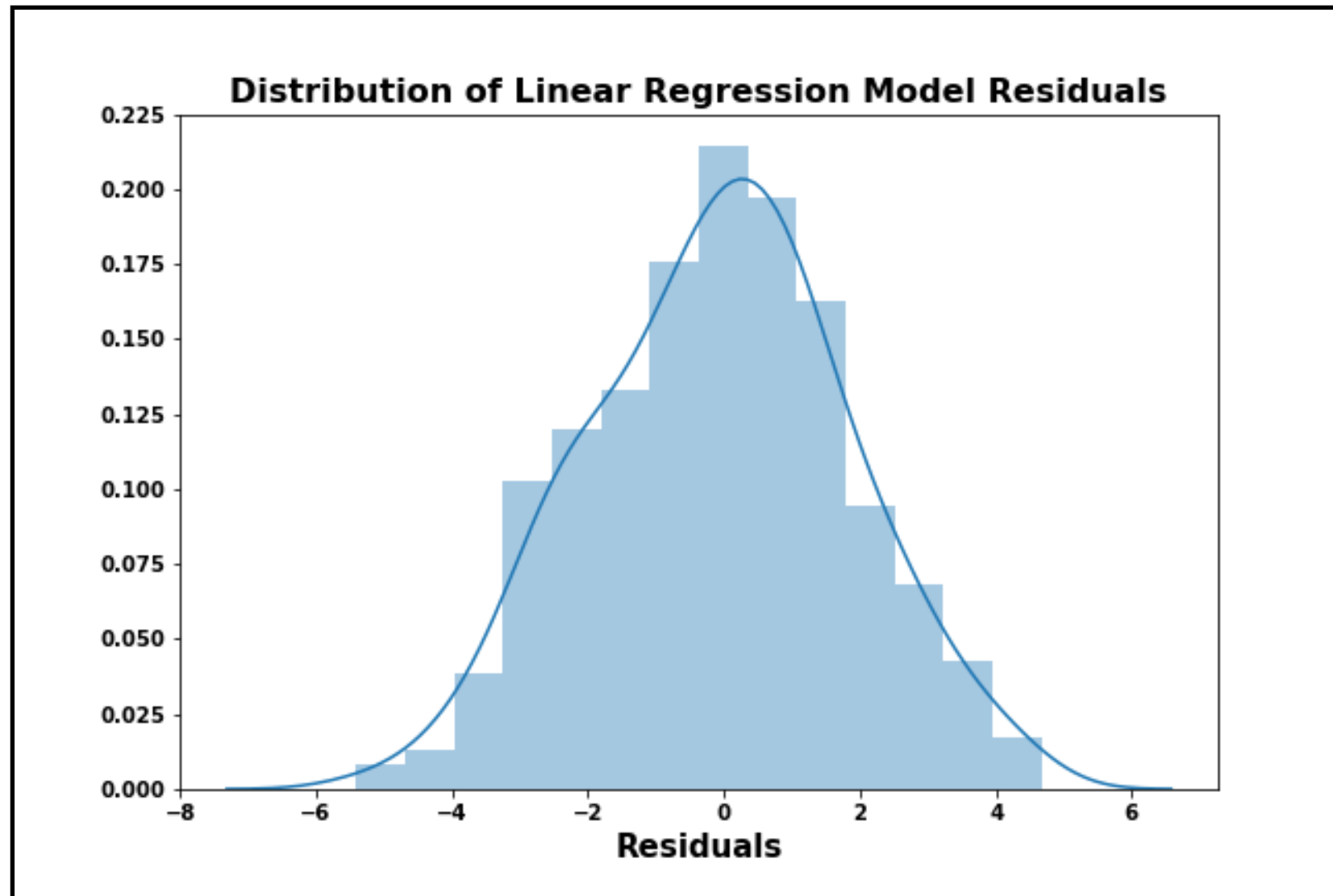
- 16 Batting & Pitching stats used to predict wins.
- RMSE of 1.9 for test data
- R^2 of 99%
- However, multicollinearity exists in the model with these predictors

OLS Model Predictions

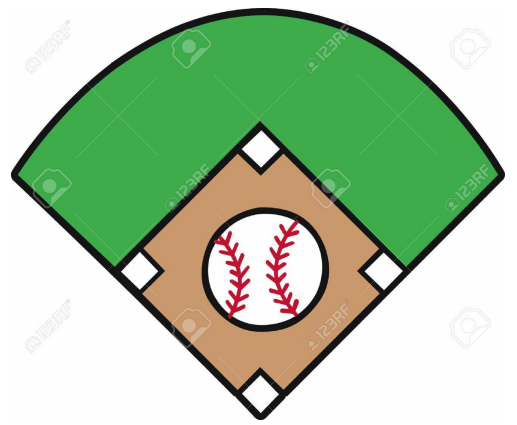


Residual Analysis

Normal Distribution of Residuals



CONCLUSIONS



- *Team Runs scored (R_x), Innings Pitched(IP), LOB_x & SV are good predictors of wins.*
- *Good batting & pitching are always useful in baseball.*
- *Future Work: Incorporate Financial Stats such as average team salary and/or Fielding Stats into regression models.*

THANKS!