Regression Analysis of Salary

For Major League Baseball Players

By Cameron Swanson and Marisol Hernandez

Moneyball

“People in both fields operate with beliefs and biases. To the extent you can eliminate both and replace them with data, you gain a clear advantage.”

— Michael Lewis, Moneyball: The Art of Winning an Unfair Game

Interpretation: Instead of relying on generational biases, let the data tell the story.

Case Study: San Francisco Giants

Problem: Since 2016, the Giants have seen a significant decline in the team's performance. With only two players left from their 2014 championships, the team has struggled to improve their roster.

Goal: We built a model based on performance statistics, as well as outside variables, that would project a player's salary. Given their budget for acquiring new players, the Giants can use the model to maximize the stats of the players they can afford.

Data Collection

Source #1: *FanGraphs Baseball*

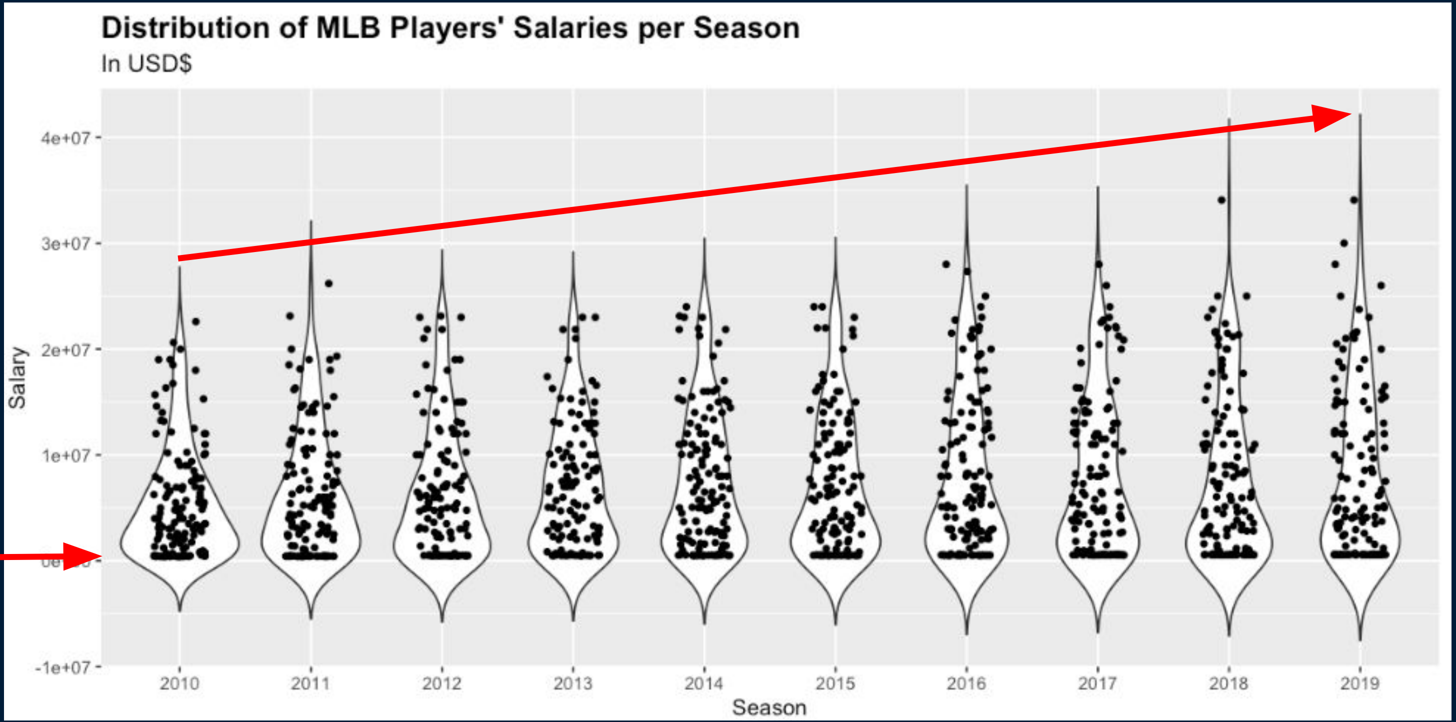
- A website that provides statistics for every player in MLB history
- We collected data from the past decade (2009 - 2019)

Source #2: *USA Today's baseball salaries database*

- Contains year-to-year listings of salaries for MLB players

Merging the two resulted in a data frame that lists one player's performance statistics over a course of a year, as well as their salary.

Descriptive Visualization



Correlations

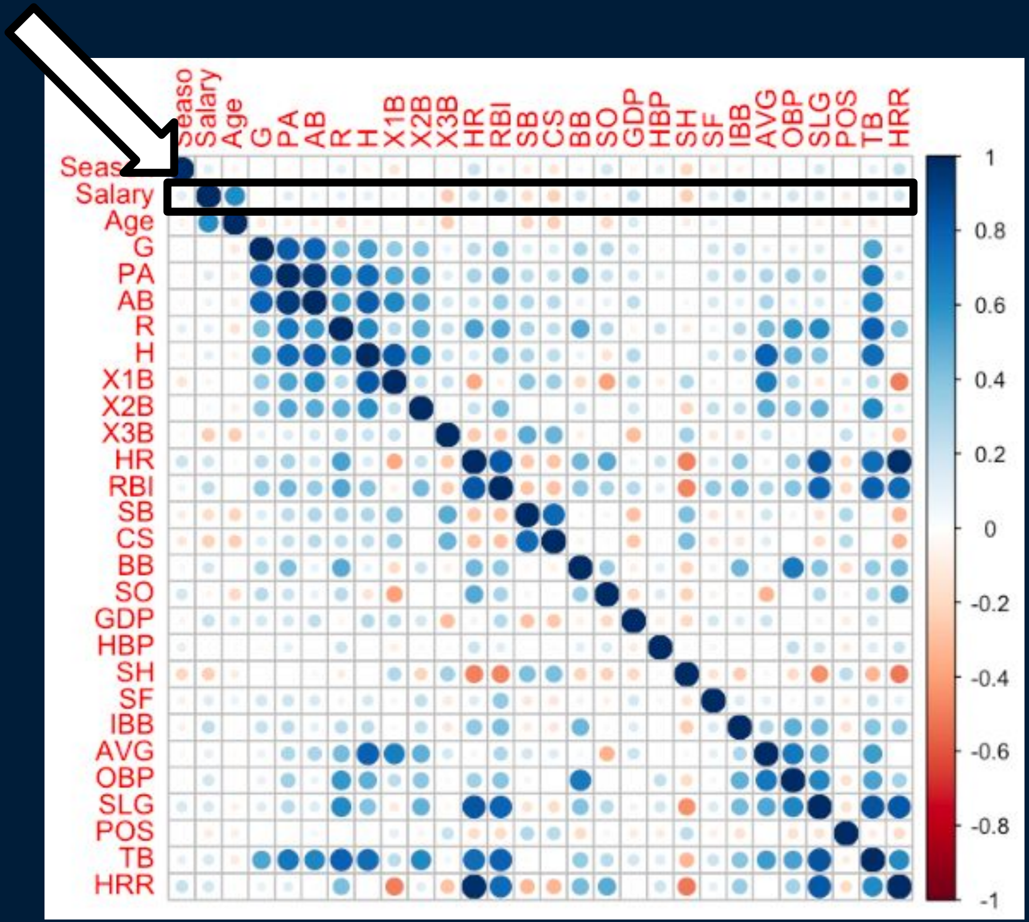
Blue = Positive Correlation

- When one qty. increases, the other increases as well

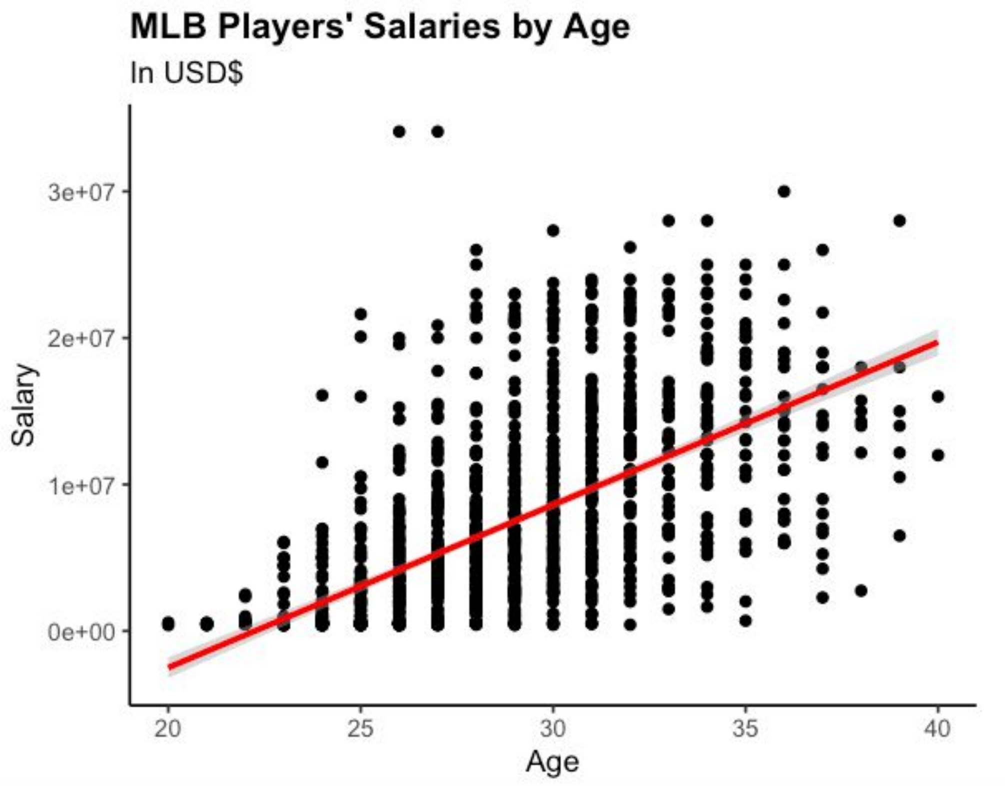
Red = Negative Correlation

- When one qty. increases, the other decreases

Stronger color = stronger correlation



Simple Linear Regression



Simple Linear Regression

- Age vs. Salary
- P-value < 0.05; significant
- Adjusted $R^2 = 0.3778$
- Negative intercept -- prediction weakens outside age range

```
Call:
lm(formula = Salary ~ Age, data = full_stats)

Residuals:
      Min        1Q    Median        3Q       Max
-14741470  -3525664  -1049099   2393272  29918976

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -24711053    1122268  -22.02   <2e-16 ***
Age           1110593      38995   28.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5283000 on 1333 degrees of freedom
Multiple R-squared:  0.3783,   Adjusted R-squared:  0.3778
F-statistic: 811.1 on 1 and 1333 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

- All p-values < 0.05; significant
- Adjusted R^2 = 0.5114
- Compare to SLR - extra variables directly contribute to better fit

```
Call:
lm(formula = Salary ~ Season + Age + G + PA + X2B + RBI + GDP +
    SH + IBB, data = full_stats)

Residuals:
      Min        1Q    Median        3Q       Max
-11561224  -3011989   -569414   2455845  22553705

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -736921763   93329655  -7.896 6.00e-15 ***
Season          353000      46265   7.630 4.47e-14 ***
Age            1088806      35592  30.591  < 2e-16 ***
G              -138625      20598  -6.730 2.52e-11 ***
PA               32978       4008   8.227 4.54e-16 ***
X2B             -63191      20792  -3.039 0.002419 **
RBI              22108       8546   2.587 0.009792 **
GDP              98285      25251   3.892 0.000104 ***
SH             -160172      57867  -2.768 0.005720 **
IBB             289396      32338   8.949  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4682000 on 1325 degrees of freedom
Multiple R-squared:  0.5147,	Adjusted R-squared:  0.5114
F-statistic: 156.2 on 9 and 1325 DF,  p-value: < 2.2e-16
```

Model Comparison

```
Call:
lm(formula = Salary ~ Age, data = full_stats)

Residuals:
      Min        1Q    Median        3Q       Max
-14741470  -3525664  -1049099   2393272  29918976

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -24711053    1122268  -22.02   <2e-16 ***
Age           1110593      38995   28.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5283000 on 1333 degrees of freedom
Multiple R-squared:  0.3783,	Adjusted R-squared:  0.3778
F-statistic: 811.1 on 1 and 1333 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Salary ~ Season + Age + G + PA + X2B + RBI + GDP +
    SH + IBB, data = full_stats)

Residuals:
      Min        1Q    Median        3Q       Max
-11561224  -3011989   -569414   2455845  22553705

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -736921763   93329655  -7.896 6.00e-15 ***
Season          353000      46265   7.630 4.47e-14 ***
Age            1088806      35592  30.591  < 2e-16 ***
G              -138625      20598  -6.730 2.52e-11 ***
PA               32978       4008   8.227 4.54e-16 ***
X2B             -63191      20792  -3.039 0.002419 **
RBI              22108       8546   2.587 0.009792 **
GDP              98285      25251   3.892 0.000104 ***
SH             -160172      57867  -2.768 0.005720 **
IBB             289396      32338   8.949  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4682000 on 1325 degrees of freedom
Multiple R-squared:  0.5147,	Adjusted R-squared:  0.5114
F-statistic: 156.2 on 9 and 1325 DF,  p-value: < 2.2e-16
```

Recommendation

- SF Giants should use our model (given their allocated budget) to find the best players they can afford
- Keep for future use -- can account for changes in budget
  - Future data increases model strength