How are baseball players' salary justified by their performance

Group 32

Derek Sun (z5160085), Hugo Perron (z5114824), Louis Ye (z5120556), Yujie He (z5161677)

MATH 3821 Assignment 2

August 4, 2019

Overview

Introduction

The goal of this statistical analysis is to understand

- how salaries is reflected by the performance of baseball players
- how different measurements like #home runs contributes to the overall performance
- is it possible to predict salaries and see if a player is overpaid or underpaid, through regression

The data displayed 1986 salaries and performance of North American Major League Baseball players, including hitters, pitchers and team data.

Understand the Data - Hitters

Response

- 1987 annual salary on opening day in thousands of dollars

Predictors

- hitter's name
- #times at bat in 1986
- #hits in 1986
- #home runs in 1986
- #runs in 1986
- #runs batted in in 1986
- #walks in 1986
- #years in the major leagues
- #times at bat during his career
- #hits during his career
- #home runs during his career
- #runs during his career
- #runs batted in during his career
- #walks during his career
- player's league at the end of 1986
- player's division at the end of 1986
- player's team at the end of 1986
- player's position(s) in 1986
- #put outs in 1986
- #assists in 1986
- #errors in 1986
- player's league at the beginning of 1987
- player's team at the beginning of 1987

Understand the Data - Pitchers

Response

- 1987 annual salary on opening day in thousands of dollars

Predictors

- pitcher's name
- player's team at the end of in 1986
- player's league at the end of 1986
- #wins in 1986
- #losses in 1986
- earned run average in 1986
- #games in 1986
- #innings pitched in 1986
- #saves in 1986
- #years in the major leagues
- #wins during his career
- #losses during his career
- earned run average during his career
- #games during his career
- #innings pitched during his career
- #saves during his career
- player's league at the beginning of 1987
- player's team at the beginning of 1987

Understand the Data - Teams

Response

- 1987 average salary

Predictors

- league
- division
- position in final league standings in 1986
- team
- #wins in 1986
- #losses in 1986
- attendance for home games in 1986
- attendance for away games in 1986

Preparing Data for Analysis

1. Retrieve data from the website
2. Import hitters, pitchers and teams csv respectively into R
3. Apply changes (initial data has been revised)
4. Remove NA entries

hitters' name	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary	NewLeague
-Alan Ashby	315	81	7	24	38	39	14	3449	835	69	321	414	375	N	W	632	43	10	475	N
-Alvin Davis	479	130	18	66	72	76	3	1624	457	63	224	266	263	A	W	880	82	14	480	A
-Andre Dawson	496	141	20	65	78	37	11	5628	1575	225	828	838	354	N	E	200	11	3	500	N
-Andres Galarraga	321	87	10	39	42	30	2	396	101	12	48	46	33	N	E	805	40	4	91.5	N
-Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	19	501	336	194	A	W	282	421	25	750	A
-Al Newman	185	37	1	23	8	21	2	214	42	1	30	9	24	N	E	76	127	7	70	A

Note the response in all 3 data-sets are numerical, logistic regression is not applicable here

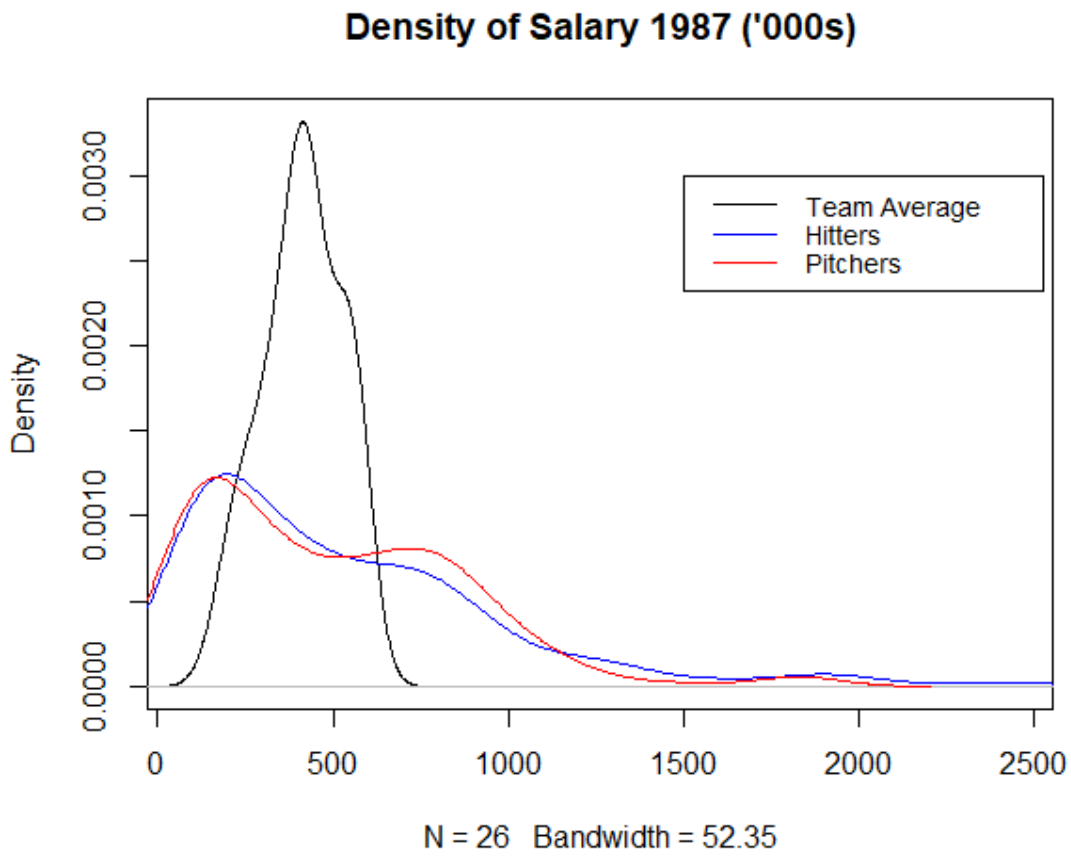
Data Exploration

There are many columns in the hitters and pitchers data, it is important to explore the relationship between variables and construct an accurate model that is not over-complex.

1. Salary Density
2. Relationships between variables
3. Multicollinearity

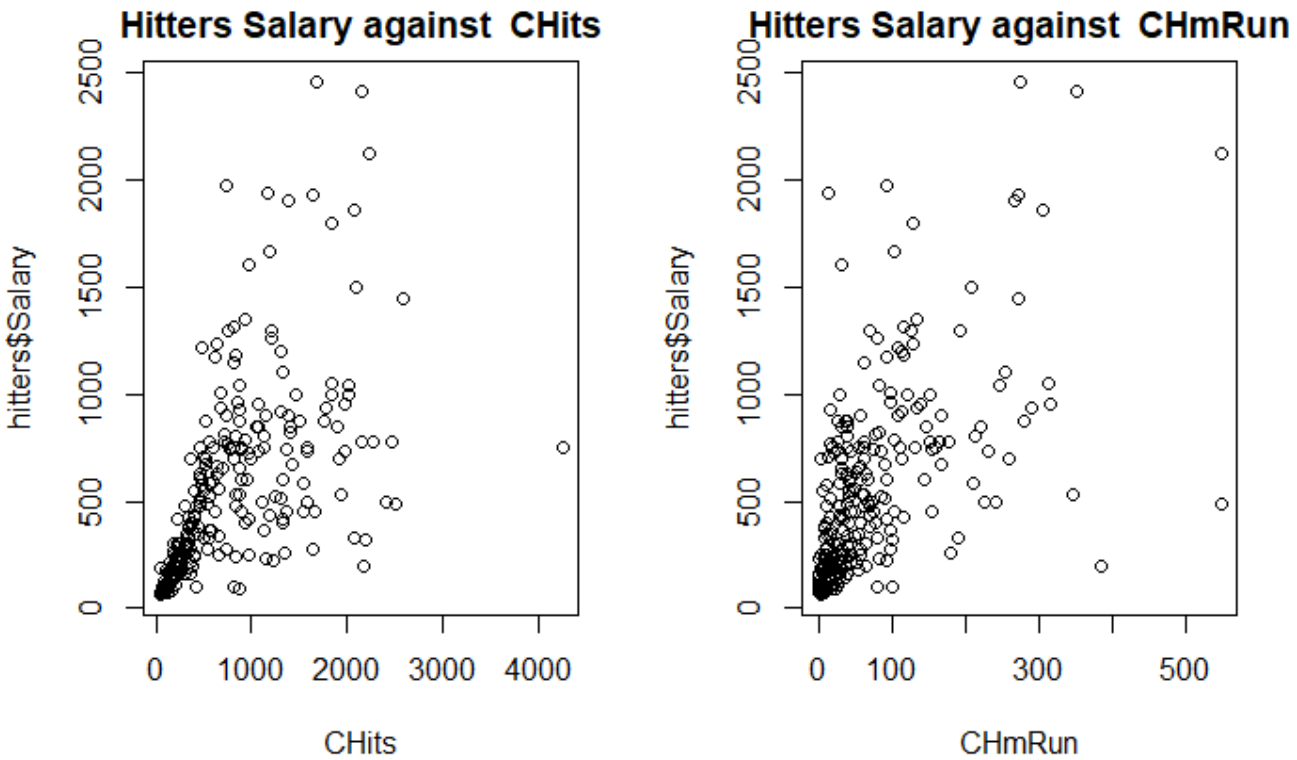
Salary Density

- hitters and the pitchers generally follow similar salary pattern and amount
- teams' average salary is less deviated and roughly symmetric



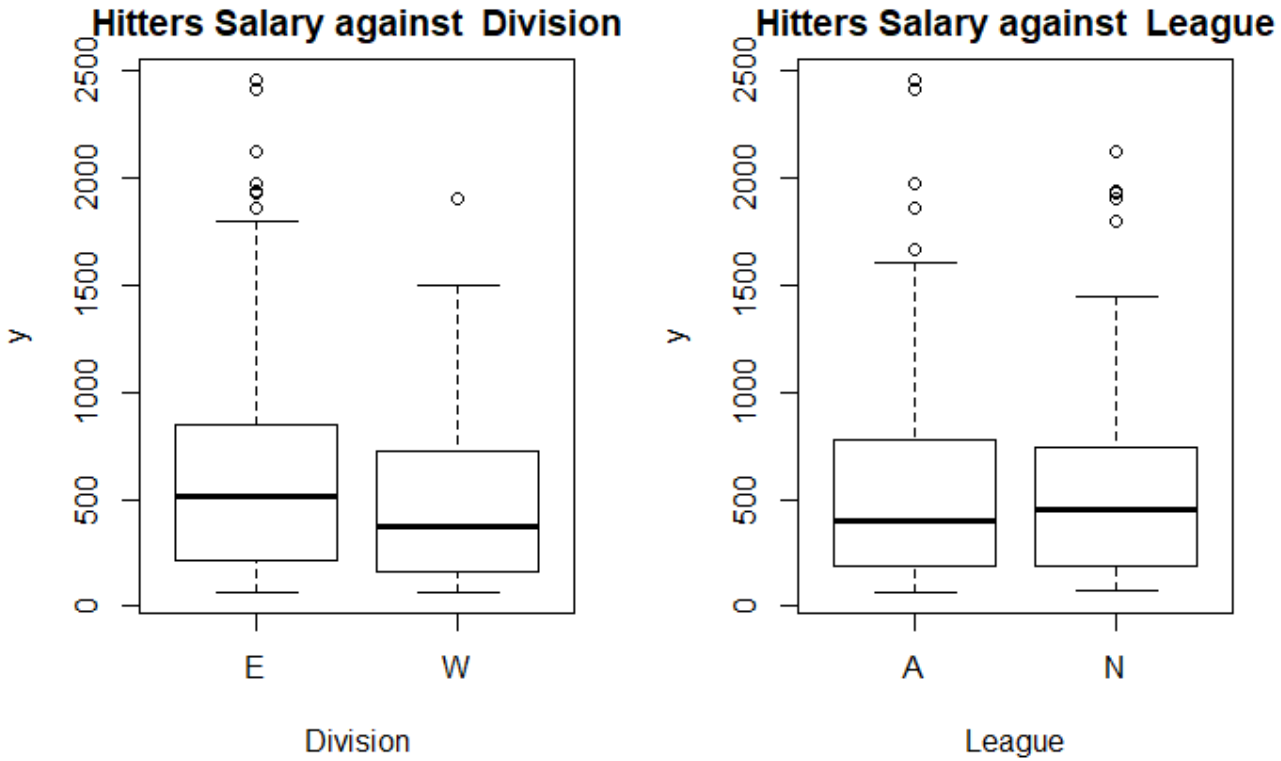
Relationships between variables

There are clear positive relationship with many quantitative predictors, but variance gets larger, a penalized regression might be better.



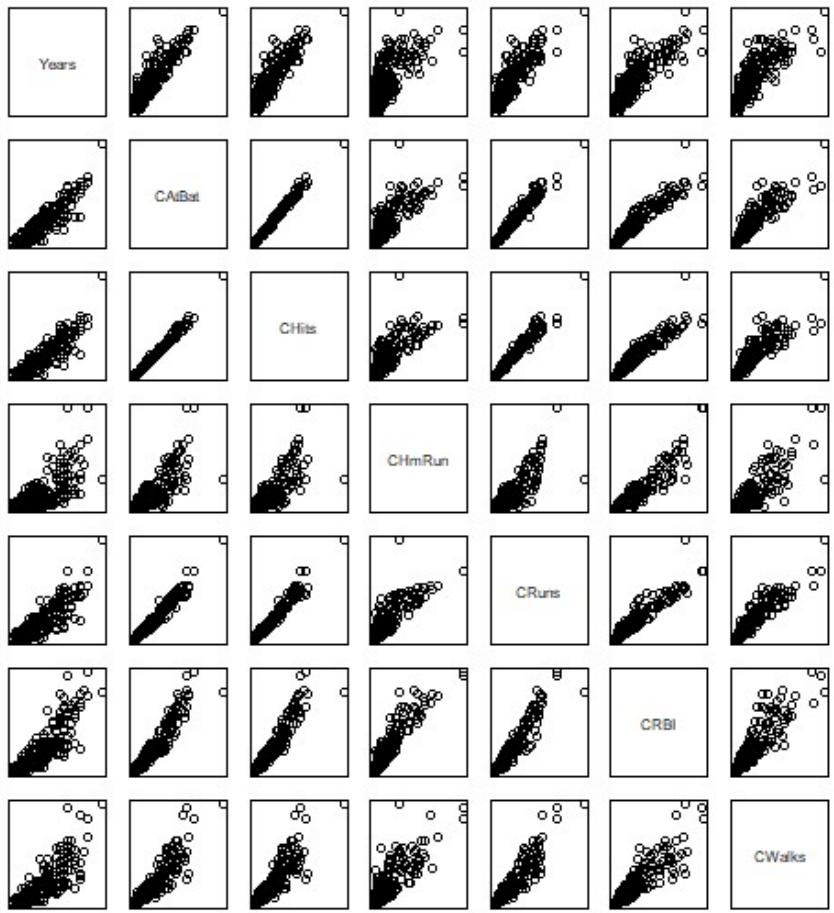
Relationships between variables

Qualitative variables also seem to have significant impact on expected salary income



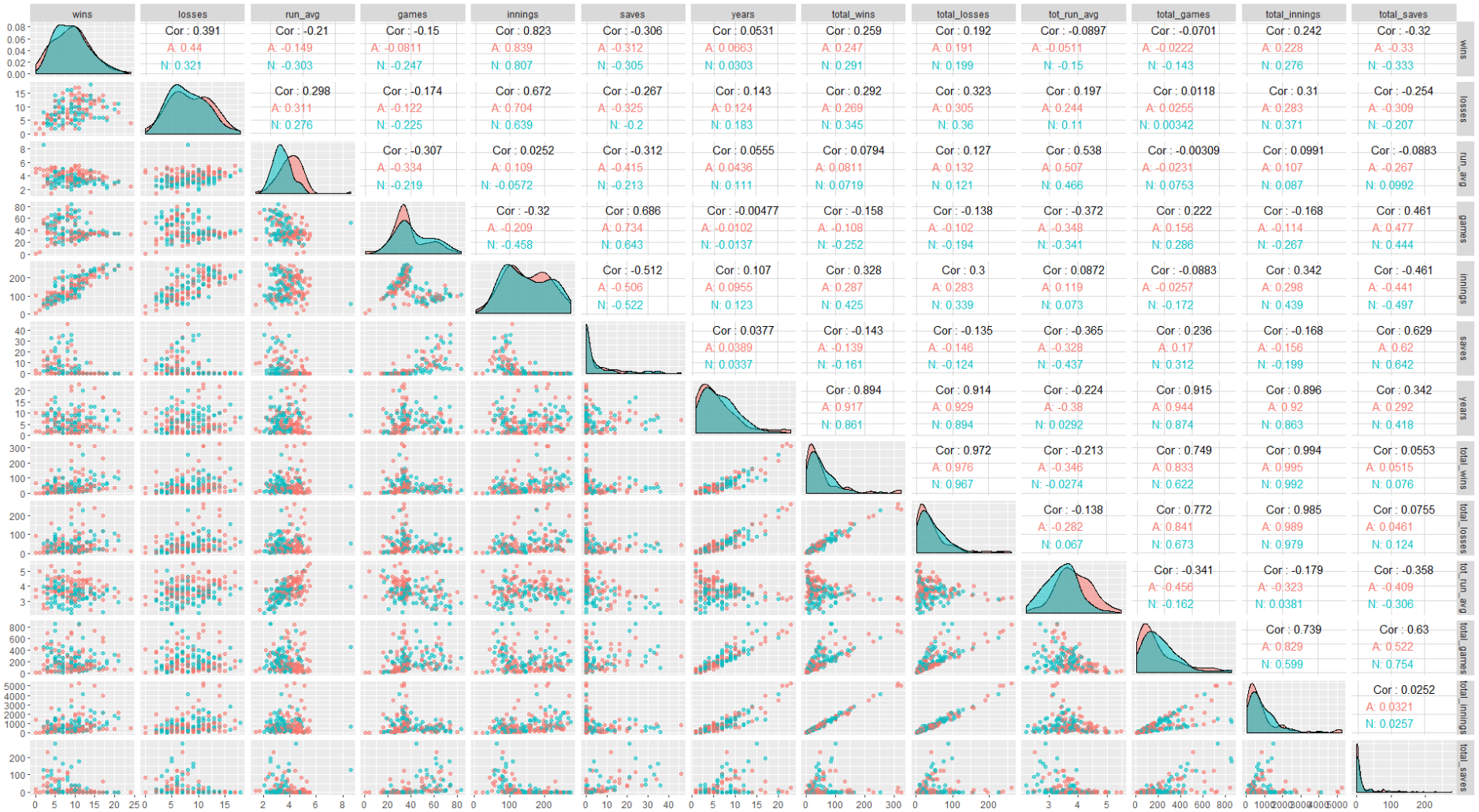
Multicollinearity

Some measures are highly correlated and positive proportional, like number of bats in career and number of hits in career for hitters. Can perhaps simplify models by just choosing one out of the two.



Multicollinearity - Pitchers

Likewise, we repeat the process for Pitchers and Teams. Note that that Pitchers also displays the issue of heavily left-skewed data.



Model Selection and Fitting - Generalized Linear Model

- Used R to fit and test models.
- Initially used all predictors, later implemented interactions/polynomials (e.g. $\frac{\#\text{Hits}}{\#\text{Number of Times at Bat}}$, and $\frac{\#\text{Wins}}{\#\text{Games}}$).
- Used anova() to remove unneeded predictors.

	Estimate	Std. Error
(Intercept)	-29.60969135	94.68916289
AtBat	-1.37766195	0.64580660
Hits	7.17582224	2.17271067
HmRun	1.91919939	5.46732080
RBI	-1.88604373	2.58857571
Walks	3.69070555	1.32463172
Years	-4.76808107	13.24329833
CAtBat	-0.08776789	0.13762401
CHits	0.53552848	0.43553995
CHmRun	1.07086832	0.53723217
DivisionW	-59.31541854	43.43269813
PutOuts	0.35202796	0.08465343

Figure: Hitters GLM $MSE = 100029.9$

	Estimate
(Intercept)	378.2980
LeagueN	66.9828
Wins	16.8755
Saves	0.8250
Cwins	3.0835
CRunAverage	-74.1340
CGames	0.6205
CInnings	-0.1061

Figure: Pitchers GLM $MSE = 60480.22$

Model Selection and Fitting - Penalized Regression (Lasso)

- Used glmnet() to fit Penalized Regression.
- Obtains a lower MSE for both.

```
12 x 1 sparse Matrix of class "dgCMatri
                      1
(Intercept) -84.4664127
(Intercept)   .
AtBat         .
Hits          2.5275481
HmRun         .
RBI           .
Walks         2.6855683
Years         .
CAtBat        0.0526253
CHmRun        1.1353218
DivisionW   -90.4915883
PutOuts       0.1909965
```

Figure: Hitters $MSE = 80332.44$

```
9 x 1 sparse Matrix of class "dgCM
                       1
(Intercept) 440.4432786
(Intercept)   .
LeagueN      44.1606079
Wins         14.7045672
Saves         .
Cwins         1.2307638
CRunAverage -79.0459742
CGames        0.5988865
CInnings      .
```

Figure: Pitchers $MSE = 61908.06$

Model Selection and Fitting - AIC Forward Selection

- Used regsubsets() to perform AIC forward selection on models
- Obtains a higher MSE.

```
##  (Intercept)        AtBat         Hits        Walks         CRBI
##  -62.7082652   -1.4012164    7.0983883    3.7541639    0.7240970
##    DivisionW      PutOuts
## -138.0145332    0.2813861
```

Figure: Hitters $MSE = 89638.63$

```
(Intercept)      Losses     Innings       Cwins CRunAverage      CSaves
 559.795204    5.276936    1.474735    2.108245 -131.575452    2.236095
```

Figure: Pitchers $MSE = 58647.45$

Final Models

$Salary_{Hitter} = -84 + 2.527Hits + 2.685Walks + 0.053CAtBat + 1.135CHmRun - 90.491DivisionW + 0.1909965PutOuts$

$Salary_{Pitcher} = 559.795 + 5.28Losses + 1.474Innings + 2.108Cwins - 131.57CRunAverage + 2.236CGames$

Model Selection and Fitting - Other Models

Other Models Considered/Possible:

- Model with interactions
  ($Salary \sim Hits/AtBat + HmRun/Runs + RBI/Walks + Years + CHmRun/CRuns + ICHits/CAtBat + PutOuts + Assists$) - MSE was too high.
- Polynomial/logarithmic transformations
- Simple Generalized Additive Models with 'simple' performance measures, e.g. $Salary\ AtBat + Hit$ and Winrate - has been explored elsewhere. (Lackritz 1990), (Wiseman 1997)

Model Assessment - Predictions

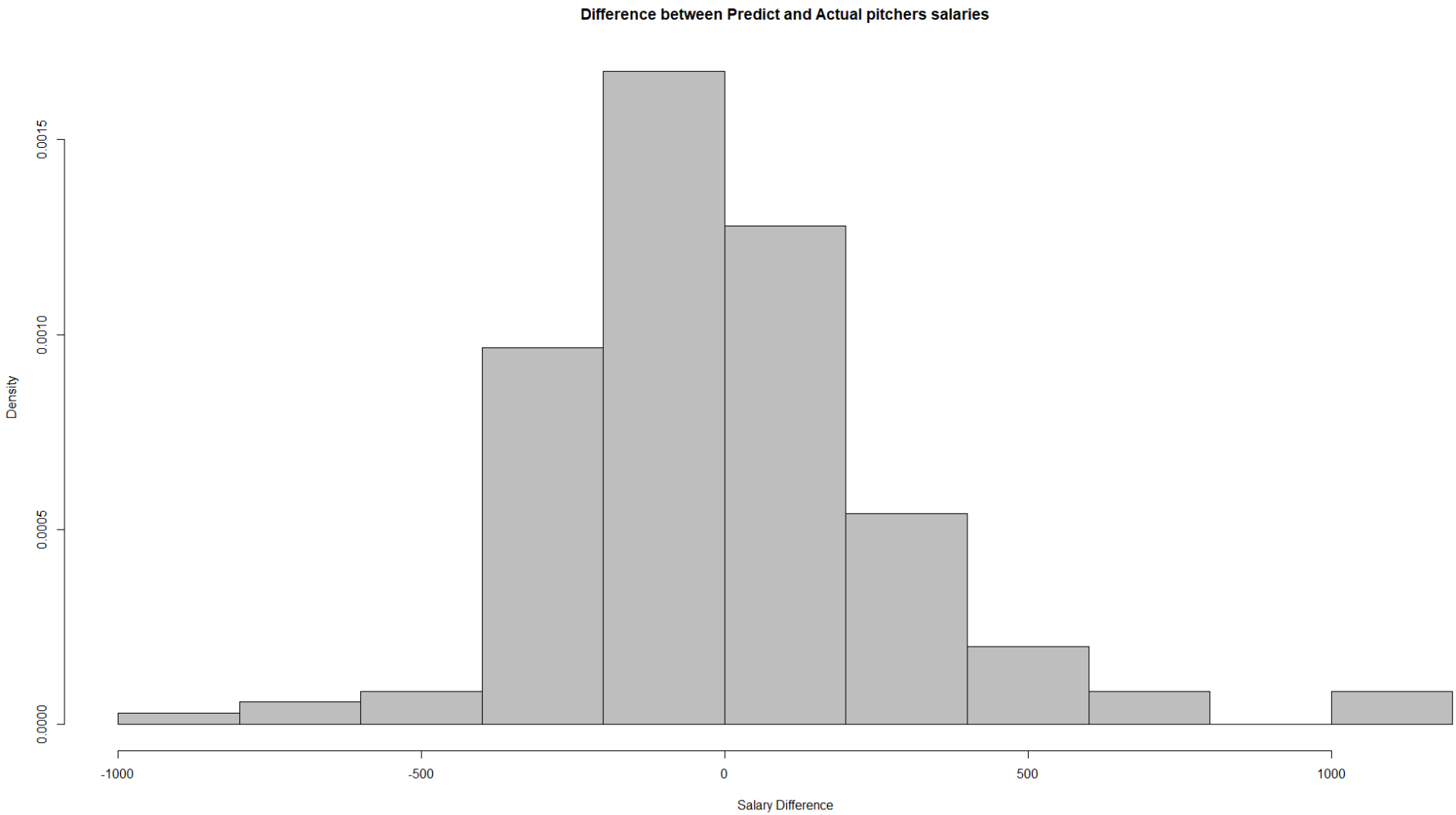
Actual salary with predicted salary for players

- MSE errors to see how much deviance in the predicted values
- See if models are overvaluing or undervaluing players

Fictional player comparison

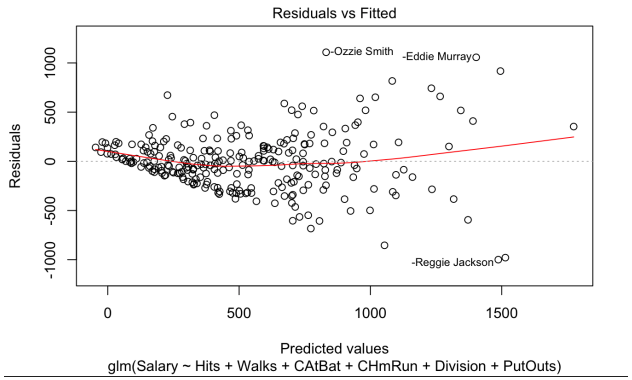
- Create fictional player with mean value in each metric
- Compared predicted salary to mean salary among the players

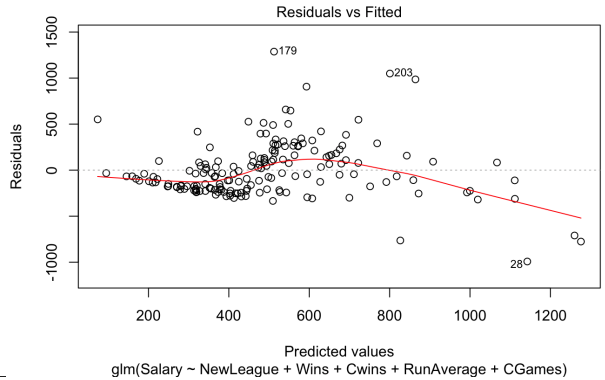
Real vs predicted Salaries



Residual Plots

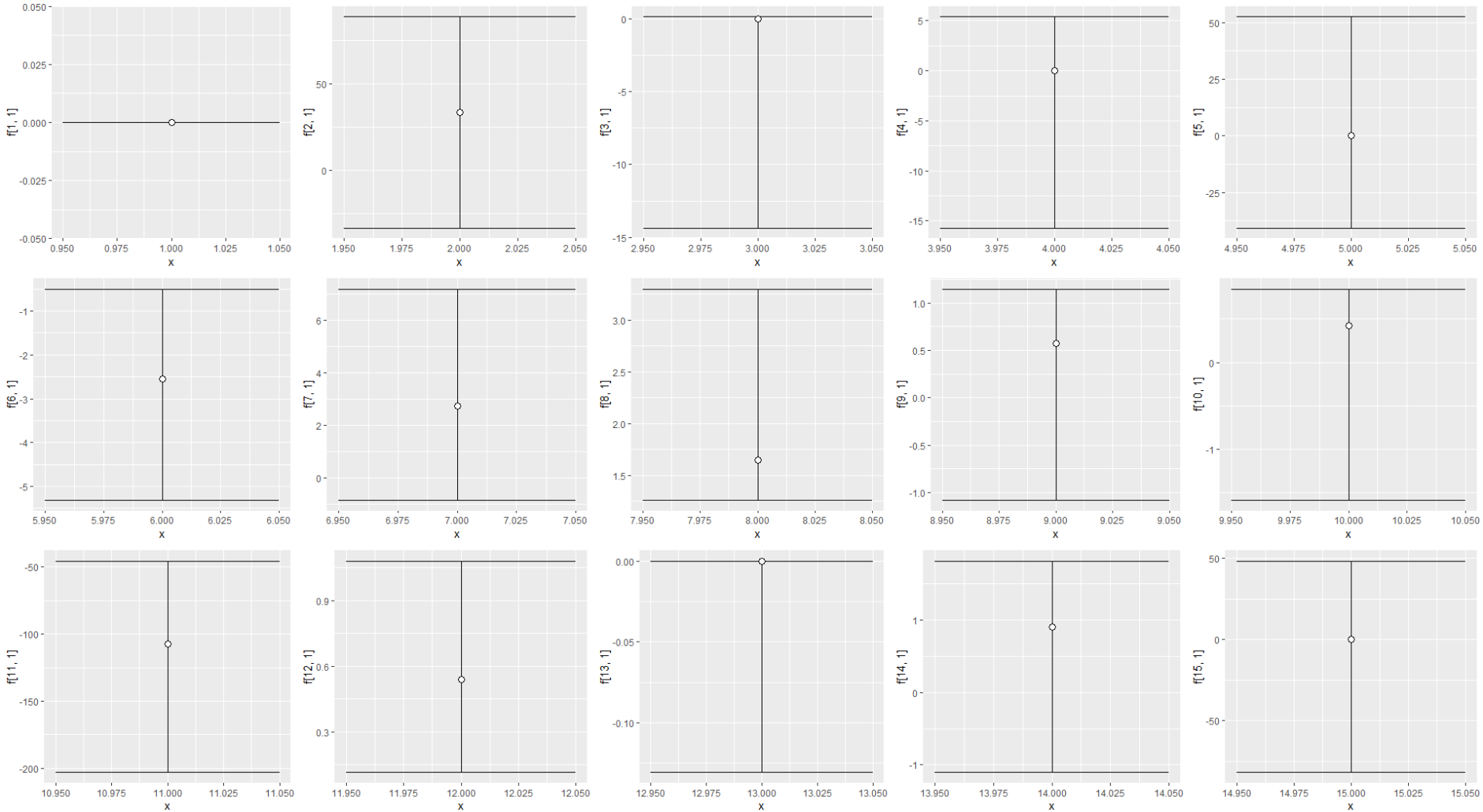
The residuals plot for hitters is quite flat which apart from a few large outliers. This can happen in a sport like baseball where star players are extremely well paid.





More shape in this plot for the pitchers. This illustrates that the residuals are not so random and the linear and the relationship is not entirely linear. This can also be attributed to the fact that there are some players earning a big salary and this is affecting the plot.

Bootstrapped Confidence Intervals



Limitations

- No team salary regulation in the 1980's
- Rookie contracts and minimum salaries
- Assumption that salary is purely determined by on-field performance
- Lack of performance statistics available
- Good team statistics can inflate an individual's worth

Conclusion

- A regularized regression is able to perform best out of all "linear" regressions.
- Some variables display a very significant effect on salary, whilst others appear to have no effect
- More data would be extremely useful in improving the model.

The End

References

Wiseman, F., Chatterjee, S. (1997)

Major League Baseball Player Salaries: Bringing Realism into Introductory Statistics Courses

The American Statistician 51(4), 350-352.

Lackritz, J. (1990)

Salary Evaluation for Professional Baseball Players

The American Statistician 44(1), 4-8.