What Predicts Salary?

Ruixie Fang

At first, I tried to figure out what affect the baseball player's salary most. However, when I pay more attention to the baseball, I found that playing baseball is not just an easy sport the player simply run or bat ball, and every step's change in a performance might cause the difference. Based on the curiosity about the baseball, this study examines the salaries of Major League Baseball (MLB) players and which part of performance predicts the player's salary efficiently. This report discusses my finding.

Method

I tended to find the significant difference of mean between variables to find what predicts salary. I decided to investigate whether the number of errors, batting average or the number of running base tends to react more obviously to affect the player's salary. To test this, I used the baseball dataset about the salary of 337 players in the 1991 and 1992 season. I then used these data to calculate for the mean difference between variables.

Results

Because the basic statistical measures indicated great dispersion in the salary, I transferred the salary to normal distribution and made it easier for me to find the relationship between salary and other variables. After that, the variance of normalized salary, 1.38425, was much less than the variance of the original variable salary, 1537633, In this case, it made better sense to attempt to predict the normalized salary value of a player instead of the original salary. With the same

reason, I also did same transformation with several variables. The following study based on the normalized variables.

After finished those preparation, simple linear regression was carried out to investigate the relationship between salary and other variables. Salary was the dependent variable and the number of errors, the number of running base and batting average were the predictor. Then, all outputs showed a significant relationship between salary and other variables ($p < 0.001$) (Table1,2,3) . The *P*-values for the *t*-tests appearing in the table of estimates showed that the difference in means was statistically significantly different from 0 and that we should reject the null hypothesis. That is, the value of those variables would be useful in predicting salary.

Table 1. P-value: salary&RBI

Parameter Estimate					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	4.49138	0.13144	34.17	<.0001
rt_RBI	1	0.33035	0.01981	16.68	<.0001

Table 2. P-value: salary&errors

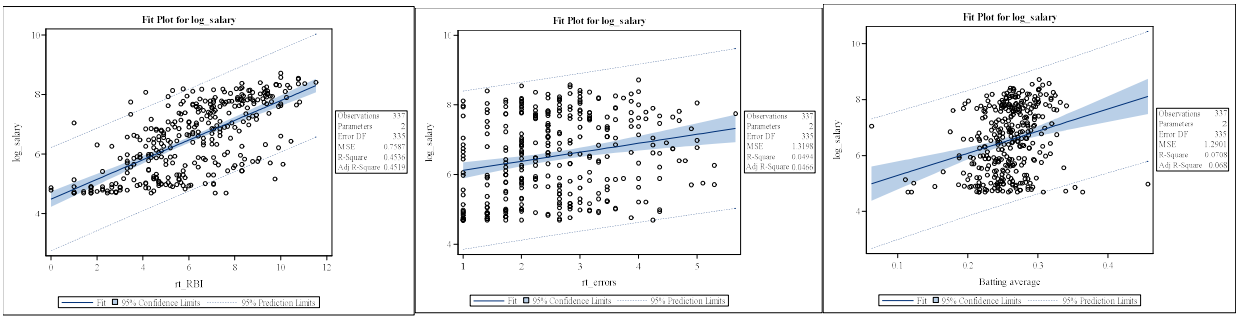
Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	5.86707	0.17197	34.12	<.0001
rt_errors	1	0.25740	0.06169	4.17	<.0001

Table 3. P-value: salary&avg

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	4.49432	0.40869	11.00	<.0001
avg	Batting average	1	7.91662	1.56687	5.05	<.0001

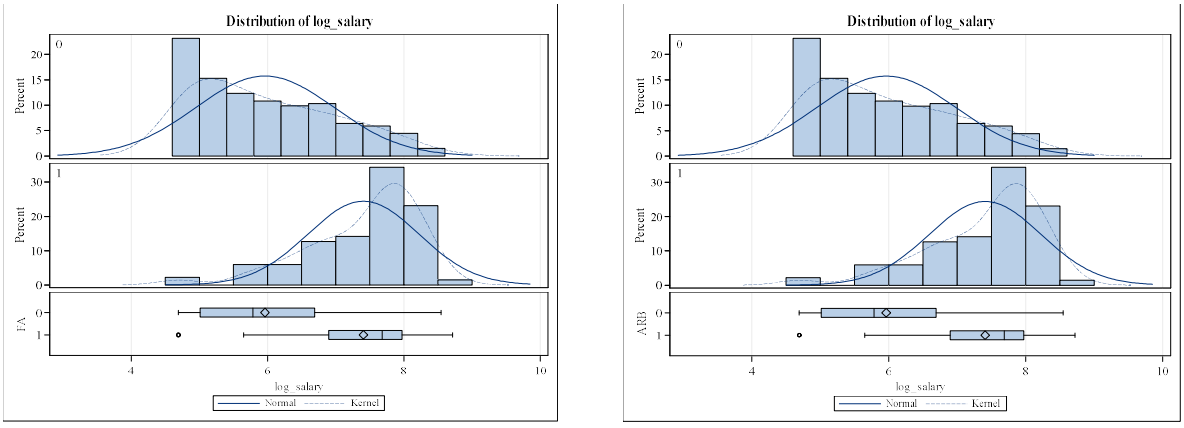
In addition to getting the regression table, it can be useful to see a scatterplot of the predicted and outcome variables with the regression line plotted. The following scatterplot showed that the model for the chart on the left is very accurate (Figure 1), there was a strong correlation between the model's predictions and its actual results. The models for the rest chart were not very good (Figure 2,3). Combined with the conclusion we got above, it is possible to argue that the number of running base (RBI) would be more useful in predicting salary.

Figure 1.salary&RBI **Figure 2.salary&errors** **Figure 1.salary&avg**



Then, let's examine the relationship between salary and free agency/arbitration eligibility to see if salary is related to eligibility. The summary panel compared the distribution of salary with free agency eligibility (0-no eligibility 1-with eligibility)(Figure 4,5). The mean value of the players with eligibility is greater than the players without eligibility. That might indicate that the players' eligibility do helpful to predict the salary.

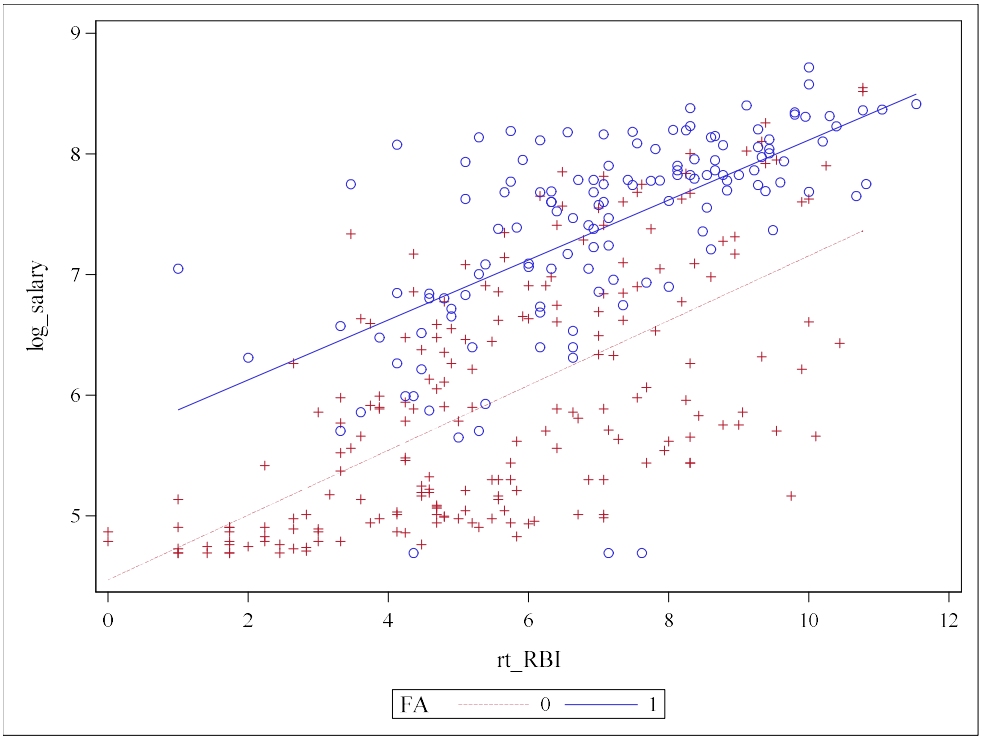
Figure 4.salary&FA **Figure 5.salary&ARB**



Discussion

In conclusion, the correlation coefficient confirmed my conclusion that the number of running base and eligibility may useful in predicting salary. The number of running base had a positive relationship with salary, and the player with eligibility (FA-1) had higher salary than other players without eligibility (Figure 6) .

Figure 6. salary&RBI&FA



Appendix: SAS Code

SAS Code Group 1 : Create A New Library

```
/*create a new library*/
LIBNAME Baseball 'D:\GSU\Study\SAS\New folder';
RUN;
```

SAS Code Group 2 : Normalize or combine variables

```
ods rtf file="D:\GSU\Study\SAS\New folder\output2.rtf";
/*Normalize or combine variables to get better sense to predict salary*/
DATA baseball.baseballnew;
SET baseball.baseball;
log_salary=log(salary);
rt_RBI=sqrt(RBI);
rt_errors=sqrt(errors+1);
FA=0;
IF FA_Eligible=1 or FA_9192=1 then FA=1;
ARB=1;
IF FA_Eligible=0 or FA_Eligible=0 then ARB=0;
TITLE 'Normalizing&Combination';
RUN;
```

SAS Code Group 3 : Linear Regression

```
/*Regression between log_salary and other variables*/
PROC REG DATA=Baseball.baseballnew;
MODEL log_salary=rt_RBI;
TITLE 'Linear Regression log_salary&rt_RBI';
RUN;
PROC REG DATA=Baseball.baseballnew;
MODEL log_salary=rt_errors;
TITLE 'Linear Regression log_salary&rt_errors';
RUN;
PROC REG DATA=Baseball.baseballnew;
MODEL log_salary=avg;
TITLE 'Linear Regression log_salary&rt_avg';
RUN;
```

SAS Code Group 4 : T test

```
/*Ttest between salary and free agency&arbitration eligible*/
PROC TTEST DATA=Baseball.Baseballnew;
CLASS FA;
VAR log_salary;
TITLE 'T test # FA&log_salary';
RUN;
PROC TTEST DATA=Baseball.Baseballnew;
CLASS ARB;
VAR log_salary;
TITLE 'T test # ARB&log_salary';
RUN;
TITLE;
```

SAS Code Group 5 : Correlation salary&RBI&FA_eligibility

```
/*Confirme by correlation among variables*/
PROC CORR DATA=baseball.baseballnew PLOTS=SCATTER;
VAR  rt_RBI rt_errors avg FA ARB;
WITH log_salary;
 TITLE 'Correlation among variables';
RUN;

PROC SGPLOT DATA=Baseball.Baseballnew;
REG x=rt_RBI y=log_salary/Group=FA;
RUN;
PROC SGPLOT DATA=Baseball.Baseballnew;
REG x=rt_RBI y=log_salary/Group=ARB;
RUN;

ods rtf close;
```