

Are Position Players with High On-Base Percentage Undervalued Free Agents in Major League Baseball?

Examining On-Base Percentage and Other Variables to Predict Salaries

Executive Summary

We found that on-base percentage accounts significantly for a baseball player's salary. There may be position players who have high on-base percentage who are undervalued in major league baseball. However, the data would suggest that the marketplace for free agent baseball players is fairly efficient.

Key Takeaways

- Free agent salaries can be explained by a statistical model. This may mean that baseball executives have a general framework from which they compensate players, or that the market is efficient in valuing players according to their ability.
- Executives may use on-base percentage from a position player's free agency year performance to predict a considerable portion of a major league baseball position players' salary.
- Many of the offensive statistics in baseball are heavily correlated with one another. It is possible that even if executives do not focus exclusively on on-base percentage, because of the high correlations between offensive statistics, they compensate players with high on-base percentage with high salaries.

Our Model

- We have created a regression model whose equation is: $\text{Next Year Free Agency Salary} = 14.6 + 3.92 \cdot \text{OBP} - .0845 \cdot \text{Age} + .00972 \cdot \text{TB} - .00557 \cdot \text{SO}$.
 - Executives are compensating players for both their ability to get on base as well as the number of bases a player has gained with hits.
 - Executives penalize players for their age and strikeouts.
- Our model accounts for 71.1% of the variability of the dependent variable.

Business Case

"Heresy was good; heresy meant opportunity. A player's ability to get on base – especially when he got on base in unspectacular ways – tended to be dramatically under priced in relation to other abilities. Never mind fielding skills and foot speed. The ability to get on base – to avoid making outs – was under priced compared to the ability to hit with power." - Michael Lewis in Moneyball: The Art of Winning an Unfair Game, pp. 128

In the book Moneyball, Michael Lewis writes how despite playing in a small market with limited financial resources, the Oakland Athletics have consistently reached the post-season. Lewis claims that the success of the Oakland Athletics stems from the emphasis that the Athletics place upon the statistic on-base percentage in acquiring and developing players. Lewis remarks that the Athletics' management believes in two core principles when evaluating position players:

- On-base percentage is the most important variable in determining the value for a position player
- Other teams do not value on-base percentage but rather use archaic metrics (e.g., batting average, RBIs) in determining a baseball player's value.

Our analysis centers around what baseball executives pay for in position player free agents. If executives are paying for performance aside from on-base percentage, it is possible for teams to purchase "under-valued" players, i.e. players whose pay is not commensurate with their ability to get on-base.

Methodology

We did not intend this analysis to be a comprehensive explanation of what factors go into predicting position player free agent salaries. Rather, we attempt to show that the executives are compensating baseball players for capabilities beyond that of on-base percentage. To demonstrate this, we:

- Researched previous attempts to predict baseball player salaries;
- Collected data;
- Created assumptions based on previous attempts to predict salary and data collected;
- Analyzed distribution of dependent and independent variable;
- Developed a basic model demonstrating on-base percentage as a predictor for post-free agency salary;
- Determined whether we should use the natural log of salary versus the listed salary;
- Ran a full model using all available data;
- Created a correlation matrix comparing all variables;
- Ran a subset regression with the most correlated variables to the natural log of post-free agency salaries;
- Ran a step-wise regression with the most correlated variables to the natural log of post-free agency salaries; and
- Developed our own regression model.

Literature Search

We reviewed scholarly articles found in ProQuest's ABI Inform and Internet searches.

In the 1970's, Gerald Scully developed a framework for measuring an individual player's marginal revenue contribution – from this analysis he determined that players are underpaid in comparison to the marginal revenue contribution they create. His technique involved looking at a team's aggregate performance, market size of team, and other variables that are not associated with a player's on-field performance. His analysis differs from our in that we focus on compensating a player strictly for performance.

In "How Much is Home Run Really Worth?" (1998), David Lang develops a model that attempts to determine which variables are central in determining salaries. Lang hypothesizes that factors such as on-base percentage, batting average, career batting average, games played squared, free agency status, and home runs added to runs batted in are the main predictors for position player salary. Lang's study indicates that the most significant drivers of salary are on-base percentage, home runs plus runs batted in, and career games played squared. Lang concluded that on-base percentage was the most accurate measure of offensive performance and was also the most statistically significant. He also noted that while home runs plus runs batted in was statistically significant, he believed that there are other factors surrounding this statistic such as the fact that players who hit many home runs also draw ticket sales and therefore receive higher salaries. In addition, the variable career games played squared is significant because players theoretically gain in skill and popularity over time, but at a decreasing rate. Lastly, Lang noted that contrary to intuition, batting average and career batting average are not statistically significant in driving salary.

David Hoaglin and Paul Velleman in "A Critical Look At Some Analyses of Major League Baseball Salaries" (1995) employed the help of seasoned statisticians with little knowledge of baseball to come up with an objective means of explaining salary in baseball. This study revealed that certain players skew the data significantly due to abnormally large contracts which may or may not correspond accurately with performance. There also arises the issue that certain major markets for baseball have more money in their budget than minor market teams. Because baseball does not have a salary cap in place, teams such as the New York Yankees potentially have the ability to pay players above the normal rates that the player could hope to have with most of the rest of the league's teams; this has the possibility to skew the data. The author also addresses league management issues surrounding free agent status in conjunction with tenure. He points out that there are different rules surrounding contract negotiation depending on if you have played less than 2 years, more than 2 years but less than 6, or more than 6 years in the league, which may greatly influence contract size. The

study also pointed out the usefulness of using the log of salary.

M.R. Yilmaz and Sangit Chatterjee in “Salaries, Performance, and Owner’s Goals in Major League Baseball: A View Through Data” (2003) also look at players’ salary and performance. We are most interested in how they set up the experiment. Takeaways for us was that they use the previous year’s performance to indicate future compensation. They use 100 at bats in a season as the cutoff value to exclude players – but they look at ALL major league ball players. They divide players by those who make salaries below \$1m and salaries above \$1m – there are significant differences in correlation of the natural log of salary and offensive performance statistics in those two groups. In regard to their findings, they concluded that the most accurate predictor of player salary was home runs, walks, and batting average – these variables explained 32.2% of the variability in log salary.

Data Collection Process

We downloaded a database containing player’s performance statistics and salaries at www.baseball1.com.

We found a listing of free agents for a particular season from roadsidephotos.com/baseball/index.htm, a website run by Doug Pappas, chairman of the Society For American Baseball Research ‘s (SABR) Business of Baseball Committee.

Assumptions

This model attempts to make a first stab at understanding whether there is a business opportunity to purchase “undervalued” baseball players in their free agency year. Given time constraints, we made a number of assumptions which may/may not be valid or need to be further refined as next steps. The assumptions include:

- **Post-free agent salaries are proportional to net present value of player salaries.** Player salaries typically include a signing bonus, salary for a year, and incentives. Salaries may include deferred payments. While we could not easily locate the net present value of salaries throughout a contract, we could locate a “next

year salary” as provided in our baseball database. We assume that these salaries do not include incentives – while we could introduce real option theory to determine the value of these incentives, we choose not to for this exercise.

- **Listed salaries are actual salaries.** Based on a review of the salaries, it seems clear that the salaries are at best estimates (given that they are all rounded) and not actuals.
- **Salaries can be estimated using the final year of a player’s contract and/or career statistics.** We do not include variables that account for a 2 year or 3 year average prior to a player’s free agency.
- **Injury and/or history of injury is irrelevant to determining salary.**
- **The market for free agents did not change in 2001-2002.** In our model we use those who declared free agency and were signed after the 2001 and 2002 season. During the 1990s baseball salaries continually were driven upward. We assume that the market for 2001 and 2002 free agents was the same.

Experimental Design

The dependent variable in this experiment is the player’s first year salary the year after he becomes a free agent. The independent variables, unless otherwise noted, are the free agent year statistics, i.e. the performance of a position player the year that he will file for free agency.

We chose players who filed for free agency in 2001 and 2002 that had played a minimum of 70 games – we consider this the minimum to being a “full time” player.

In one respect, the type of experiment that we wish to conduct will always be fundamentally flawed in that the observations do not represent random samples from the population. In 2001 and 2002, only 72 players filed for free agency and had played a minimum of 70 games that were eventually signed by a team. Because the market value of baseball players changes from year to year, we are limited in the amount of data that we have access to. As such, we recognize that our model may not be as theoretically sound as one would hope.

We used variables given to us in the baseball database – additionally we created variables commonly used in baseball circles. For a list and explanation of all variables used, please see *Exhibit 1*.

Because of the problems associated with analyzing free agents, our independent variables and dependent variables are not evenly distributed (see *Exhibit 2A for the dependent variable Next Year Salary and Exhibit 2B for an example of one independent variable, on-base percentage*). The reason they are not evenly distributed is rather simple – in baseball, talent is neither uniformly distributed nor normally distributed. Talent in baseball takes on more of a pyramid shape with many low caliber players and few exceptional players (e.g., there can be only one Barry Bonds in a generation who files for free agency).

Our First Try – On-base percentage predicting Next Year Salary Model

Our first objective was to understand the ability of on-base percentage in predicting next year salaries.

We ran a simple regression between on-base percentage and next year salary. The regression is statistically significant, with an adjusted R Square of 48.9% (*Exhibit 3*). The histogram of the residuals does not appear normally distributed but rather is skewed, in large part due to outliers such as Barry Bonds. The normal probability plot of the residuals reveals a relatively straight line, meaning that the residuals are somewhat normally distributed; however, at the tail end of the plot the residuals no longer are linear but rather seem to be tailing off.

The residuals versus the fitted values plot are not evenly distributed. The residual plot flares out as the independent variable gets larger. This pattern is an indication of heteroscedasticity, which is a violation of the assumption of constant variance for error terms. Heteroscedasticity is caused by nonnormality of one of the variables, an indirect relationship between variables, or to the effect of a data transformation. Heteroscedasticity is not fatal to our analysis -- in our case, this regression allowed us to recognize the need to use the normal log for next year salary variable.

Our Second Try – Using On-base Percentage and Normal Log of Next Year Salary

Our second try yielded much more favorable results. While our adjusted R Square actually declined by more than 15% (*Exhibit 4*), our various plots (e.g., residuals vs. fitted lines, normal probability plot of the residuals, and histogram of the residuals) look better. However, the data is still not normally distributed. We suspected that the reason for this is found in the nature of the statistic we used – on-base percentage. This is only a percentage – what we really want to compensate baseball players is for their ability to get on base, both as a percentage of at bats and as an actual number of times the player gets on base. As such, we hypothesize that it is critical to use AB as a predictor to have an explanatory base model.

Our Third Try Looks Good as a Base Model– Using On-base Percentage and At Bats to predict Normal Log of Next Year Salary

We finally calculated a good predictive base model that does not have noticeable errors when reviewing the residuals. This model has a 60.8% adjusted R-Square (*Exhibit 5*). The normal probability plot of the residuals appears relatively plotted along the line with slight deviations at the tail ends. The histogram of the residuals appears much more normally distributed than the other regressions we ran. The residuals versus the fitted values plot looks fairly good – however, the right side of the plot does not have as much variation across the fitted line as we would hope. Overall, we are comfortable with using this as our first stab base model.

An interesting side note – we ran the same model but replaced OBP with SLG – the adjusted R-Square was 61.4%. Michael Lewis argued that on-base percentage was significantly undervalued in comparison to hitting with power (baseball analysts traditionally say that this is what SLG represents). The cursory evidence we provide here would suggest otherwise.

Full Model

We dumped all of the variables that we had collected to predict the natural log of next year salaries – our model was statistically significant at the 95% confidence level, with an adjusted R-Square of 71% (*Exhibit 6A*). However, many variables were not statistically significant at a 95% confidence level on a last-in basis.

We also dumped all variables that we had collected excluding career numbers to predict the natural log of next year salaries – our model was statistically significant at the 95% confidence level, with an adjusted R-Square of roughly 66% (*Exhibit 6B*). Again, many variables were not statistically significant at the 95% confidence level on a last-in basis.

A problem with both models is that we are attempting to create a model that rewards players for contributions on the field and penalizes for poor performance. Variables whose coefficients should be negative include striking out, caught stealing, and grounding into a double play. Age can be either a positive or negative (we have no opinion of whether an older player will be more or less compensated). Otherwise, all other variables should have positive coefficients as they represent offensive performance that should be rewarded.

Correlation Table

We constructed two correlation tables – one that includes free agency year statistics (*Exhibit 7A*) and one that includes career statistics (*Exhibit 7B*). From these tables a number of conclusions can be made:

- There appears to be little if any correlation between the natural log of next year salaries and career statistics. The highest correlations range from -.25 for HBP to .21 for Career SB%. This data would suggest that career statistics most likely will not be used in our final regression model.
- Free agency year statistics are heavily correlated with the natural log of next year player salaries. The highest correlated statistic is RC at .78, with

the next being TB at .78. Surprisingly, OBP only has a .57 correlation.

- Many of the free agency year statistics are heavily correlated with one another. Given that one of the assumptions in regression analysis is that there are independent variables, this is of great concern to us. This means that we will have to be careful in our model creation.

Best Sub-set Regression

Due to the restrictive nature of our statistical package (MINITAB), we were unable to run a subset regression on all of our variables. As such, we ran the subset regression on the variables that have the highest correlations and the variables we thought would potentially have an impact on salaries. As models should be simplistic, we thought that we would look at subsets with a maximum of 6 variables.

The first time we ran this model, we did not request any variables to be included in all models (*Exhibit 8A*). When included two or more variables in the model, the adjusted R-Square ranged from 68.1% to 71.3%. The variables that were most consistently chosen included Age, SO, and TB.

The second time we ran this model, we requested that OBP be included in all models (*Exhibit 8B*). When included two or more variables in the model, the adjusted R-Square ranged from 69.2% to 71.1%. Age, SO, and TB were again the most chosen variables in developing these models.

Best Stepwise Regression

The stepwise regression provided additional insight as to what variables should be chosen in the model. We ran two step-wise regressions.

The first step-wise regression used all variables in the dataset, with statistical significance at 95% (*Exhibit 9A*). The range of adjusted R-Square with three variables in the model ranged from 67.1% to 76.0%. However, a problem with this step-wise regression is that four of the regressions include career HBP as a negative number – getting a man on base for being hit by a pitch should be compensated salary wise. The highest adjusted R-Square with

logical predictor values is 70.1% -- this model includes OBP, Age, and TB.

The second step-wise regression uses the most highly correlated variables in the dataset, with statistical significance at 95% (*Exhibit 9B*). The step-wise regression only has up to three variables, those again being OBP, TB, and Age.

What becomes clear from both the sub-set regression and step-wise regression is that in creating our final predictive model we should probably include OBP, TB, and Age.

Our Final Model

In our final model we included the following variables: OBP, Age, TB, and SO (*Exhibit 10*).

The equation is: $14.6 + 3.92*OBP - .0845*Age + .00972*TB - .00557 SO$. We interpret this equation as follows: Baseball players do get compensated for on-base percentage. As baseball players age, they receive less pay – we presume that there is a high correlation between an increase in age and chance of injury. Moreover, contract length may play a factor – more investigation is needed on this point. Players are compensated for hitting for power, as defined by TB. A player who achieves getting on base beyond first base has the potential for two things: driving in a runner and placing himself in a greater position to score. Finally, players are penalized for striking out – this makes sense, as a strike out means that there is less of an opportunity for a team to score given the maximum 27 outs in a game. Additionally, a strike out does not allow the player to advance a potential runner on base.

The F significance of the model is statistically significant at a 95% confidence level. However, on a last-in basis, two variables are not statistically significant at a 95% confidence level: OBP and SO. OBP has a p-value of 5.7, while SO has a p-value of 6.9. We are comfortable with these values, as a reasonable story can be told from the variables serving as predictors.

Below is a correlation table for all predictor variables.

	OBP	TB	Age
TB	0.56		
Age	-0.13	-0.17	
SO	0.26	0.71	-0.02

There are two high correlations between the predictor variables: TB and SO; and OBP and TB. The high correlation between OBP and TB is not surprising, as both statistics incorporate the number of hits a player accumulates in a season. The high positive correlation between TB and SO is more shocking. Perhaps players who hit for power have the tendency to strike out.

We are concerned with the variance inflation factor for TB – it has a value of 3. Our rule of thumb is that values of VIF greater than 5 indicate serious trouble. We do not have serious trouble here, but as we further develop and refine this model we probably should explore methods to make this number decrease.

How do we compare our final model with both our base model and the full model? Our final model actually has a higher adjusted R-Square than either model. We are a bit perplexed as to why this is the case – surely the full model with all of the included variables should better predict our dependent variable than the reduced model. As for our base model, it would seem that there are other factors baseball executives may use when compensating players. However, in our final model on-base percentage still significantly drives the overall salary. At best, one can interpret the model as showing that baseball executives undervalue baseball players with high on-base percentage and high strike outs. Because the strikeouts variable does not significantly add to the adjusted R-Square, we cannot make this assertion until we refine this model further.

Next Steps and Unresolved Issues

Next steps for an analysis of this nature should include:

- **Fielding statistics.** At this stage in this analysis, we can only hypothesize that including fielding statistics would allow the regression line to better explain the variability of the residuals.

- Two-year and three-year average. It's possible that a player may have had an off year in his free-agency year.
- Injury dummy variable. In our current model, players are penalized for the age variable – we assume that age and prone to injury are highly correlated. However, this may not be the case. We assume that the higher the probability of injury, the lower the base salary the player will receive.
- Player popularity variable. It may be possible that players are compensated for popularity/general goodwill they have with fans.
- Clutch hitting. Baseball analysts heatedly debate whether clutch hitting exists in baseball. Nonetheless, executives may compensate players for “clutch” performance, which may include statistics like OBP and SLG when there are players already on base.
- Free agent year players who do not file for free agency but resign with their existing team.
- Net present value of player contracts.
- The number of years in a contract.
- The affect of a player's ball park on his performance.
- The impact of a player's team contributions on the player's performance.
- The relative value of a player in comparison to league averages and position offensive performance averages.

Exhibits

EXHIBIT 1: LIST AND DESCRIPTION OF VARIABLES USED IN ANALYSIS

EXHIBIT 2: DISTRIBUTION OF DEPENDENT AND INDEPENDENT VARIABLE (OBP)

EXHIBIT 3: OUTPUT OF OBP AND NEXT YEAR SALARY REGRESSION

EXHIBIT 4: OUTPUT OF OB% AND NATURAL LOG OF NEXT YEAR SALARY REGRESSION

EXHIBIT 5: OUTPUT OF OB% + AB AND NATURAL LOG OF NEXT YEAR SALARY REGRESSION

EXHIBIT 6A: OUTPUT OF FULL MODEL USING ALL VARIABLES

EXHIBIT 6B: OUTPUT OF FULL MODEL NOT USING CAREER STATISTICS

EXHIBIT 7A: CORRELATION MATRIX – INCLUDES NATURAL LOG OF SALARY AND FREE AGENCY YEAR STATISTICS

EXHIBIT 7B: CORRELATION MATRIX – INCLUDES NATURAL LOG OF SALARY AND CAREER STATISTICS

EXHIBIT 8A: SUBSET REGRESSION, USING MOST CORRELATED VARIABLES TO NATURAL LOG OF SALARIES, W/O VARIABLE REQUIRED IN ALL MODELS

EXHIBIT 8B: SUBSET REGRESSION, USING MOST CORRELATED VARIABLES TO NATURAL LOG OF SALARIES, W ON-BASE PERCENTAGE VARIABLE REQUIRED IN ALL MODELS

EXHIBIT 9B: OUTPUT OF STEPWISE REGRESSION USING MOST HIGHLY CORRELATED VARIABLES W/ ALPHA AT .05

EXHIBIT 10: FINAL MODEL REGRESSION OUTPUT

EXHIBIT 1: LIST AND DESCRIPTION OF VARIABLES USED IN ANALYSIS

- **G.** The number of games played.
- **AB.** The number of at bats.
- **Runs.** The number of runs scored.
- **H.** Hits.
- **2B.** Doubles
- **3B.** Triples
- **HR.** Home Runs.
- **RBI.** Runs batted in.
- **SB.** Stolen bases.
- **CS.** Caught stealing.
- **BB.** Walks.
- **SO.** Strike outs.
- **IBB.** Intentional walks.
- **HBP.** Hit by pitch.
- **SH.** Sacrifice hits or bunts.
- **SF.** Sacrifice flies.
- **GIDP.** Ground into double plays.
- **TB.** Total bases. Formula is:
(Singles+2*2B+3*3B+4*HR)
- **BA.** Batting average. Formula is: H/AB.
- **OBP.** On-base percentage. Formula is:
(H+BB+HBP)/(AB+BB+SF+HBP)
- **SLG.** Slugging Percentage. Formula is: TB/AB.
- **OPS.** OBP+SLG.
- **HR+RBI.** Home runs + RBIs.
- **RC.** Runs created. A run estimator attempts to quantify the entire contribution of a player's statistics to a team's total runs scored. It typically involves some positive values for things like hits, walks, steals, home runs, and negative values for outs, caught stealing and GIDP (explanation from www.baseball-reference.com). There are a number of versions of this statistic. We use: $(H+BB-CS)*((TB+.55*SB)/(BB+AB))$
- **Age.**
- **Old Age.** Age greater than 34 years of age.
- **US Native Born.** Dummy variable where 1=born in the US.
- **Next year salary/1,000,000.** We divide by 1,000,000 to make the number smaller and more manageable in creating the linear equation.
- **Log of next year salary.**
- **Career G.**
- **Career AB.**
- **Career H.**
- **Career 2B.**
- **Career 3B.**
- **Career HR.**
- **Career RBI.**
- **Career SB.**
- **Career CS.**
- **Career BB.**
- **Career SO.**
- **Career IBB.**
- **Career HBP.**
- **Career SH.**
- **Career SH.**
- **Career SF.**
- **Career GIDP.**
- **Career TB.**
- **Career OBP.**
- **Career SLG.**
- **Career OPS.**
- **Career HR+RBI.**
- **Career RC.**

EXHIBIT 2: DISTRIBUTION OF DEPENDENT AND INDEPENDENT VARIABLE (OBP)

EXHIBIT 2A

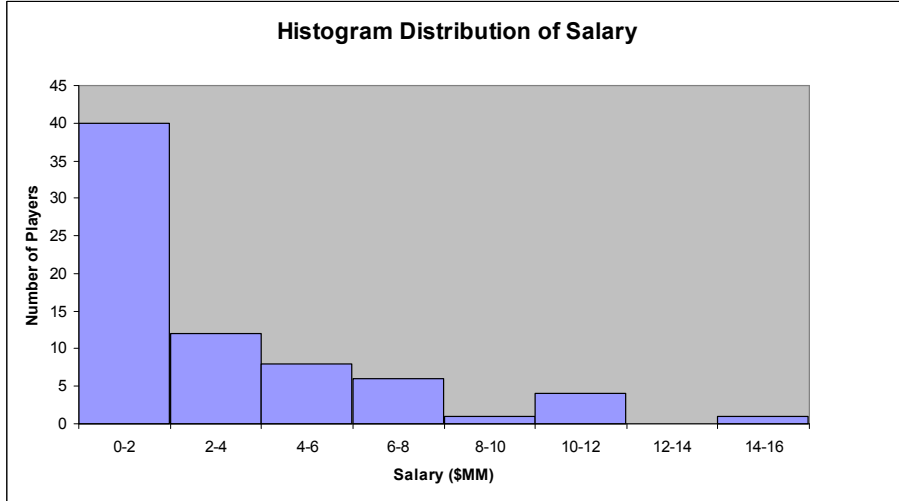


EXHIBIT 2B

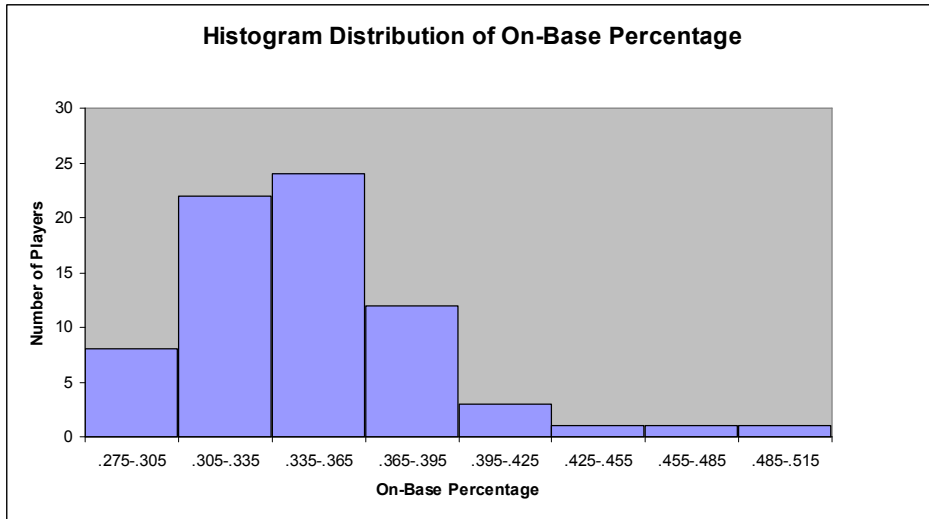


EXHIBIT 3: OUTPUT OF OBP AND NEXT YEAR SALARY REGRESSION

Regression Analysis: Next Year Salary/1,000,000 versus OBP

The regression equation is
 Next Year Salary/1,000,000 = - 16.2 + 55.2 OBP

Predictor	Coef	SE Coef	T	P
Constant	-16.197	2.318	-6.99	0.000
OBP	55.165	6.638	8.31	0.000

S = 2.25607 R-Sq = 49.7% R-Sq(adj) = 48.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	351.58	351.58	69.07	0.000
Residual Error	70	356.29	5.09		
Total	71	707.87			

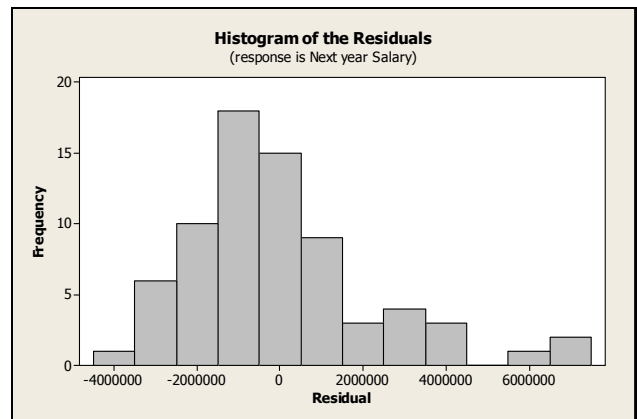
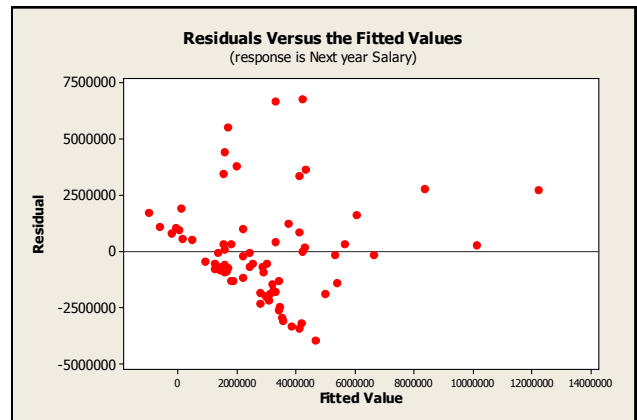
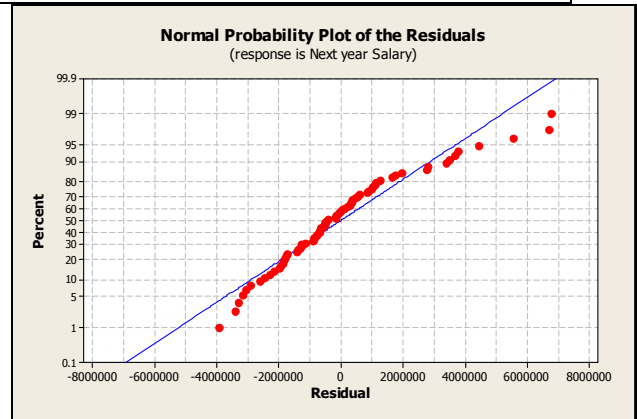


EXHIBIT 4: OUTPUT OF OB% AND NATURAL LOG OF NEXT YEAR SALARY REGRESSION

Regression Analysis: Log of Next Years Salary versus OBP

The regression equation is
 Log of Next Years Salary = 9.35 + 14.7 OBP

Predictor	Coef	SE Coef	T	P
Constant	9.3452	0.8756	10.67	0.000
OBP	14.697	2.507	5.86	0.000

S = 0.852096 R-Sq = 32.9% R-Sq(adj) = 32.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	24.953	24.953	34.37	0.000
Residual Error	70	50.825	0.726		
Total	71	75.778			

Unusual Observations

Obs	OBP	Log of Next Years Salary	Fit	SE Fit	Residual	St Resid
7	0.515	17.000	16.914	0.433	0.086	0.12 X
15	0.324	16.000	14.107	0.116	1.893	2.24R
23	0.477	16.000	16.355	0.341	-0.355	-0.46 X
44	0.322	16.000	14.077	0.118	1.923	2.28R
47	0.329	16.000	14.180	0.110	1.820	2.15R
53	0.368	13.000	14.753	0.113	-1.753	-2.08R
65	0.445	16.000	15.885	0.265	0.115	0.14 X

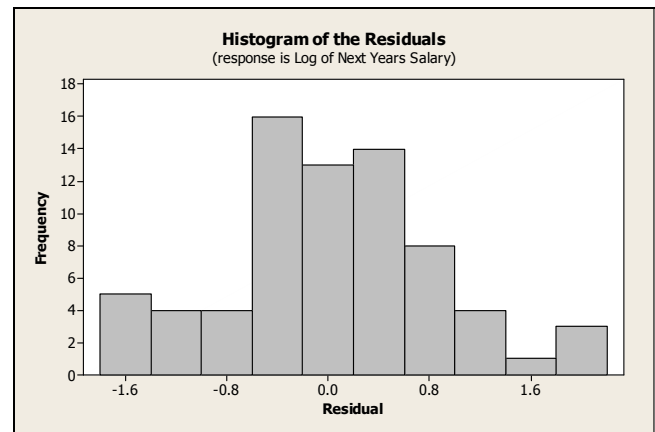
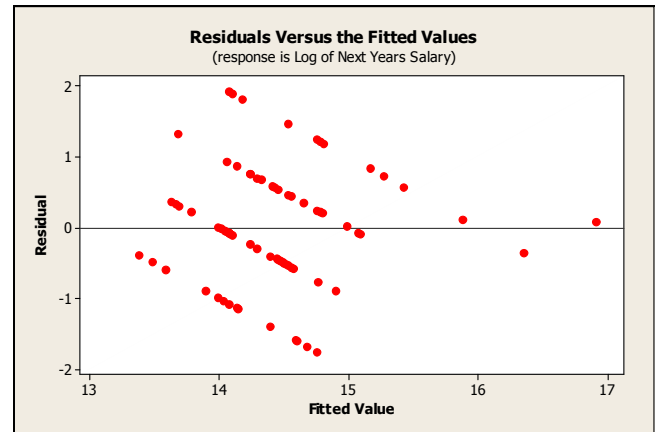
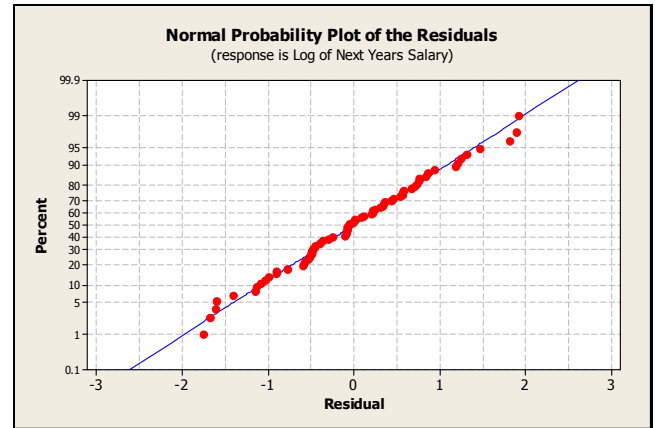


EXHIBIT 5: OUTPUT OF OB% + AB AND NATURAL LOG OF NEXT YEAR SALARY REGRESSION

Regression Analysis: Log of Next Years Salary versus OBP, AB

The regression equation is

$$\text{Log of Next Years Salary} = 8.76 + 11.7 \text{ OBP} + 0.00421 \text{ AB}$$

Predictor	Coef	SE Coef	T	P
Constant	8.7600	0.6696	13.08	0.000
OBP	11.715	1.947	6.02	0.000
AB	0.0042056	0.0005805	7.24	0.000

S = 0.646824 R-Sq = 61.9% R-Sq(adj) = 60.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	46.909	23.455	56.06	0.000
Residual Error	69	28.868	0.418		
Total	71	75.778			

Source	DF	Seq SS
OBP	1	24.953
AB	1	21.956

Unusual Observations

Obs	OBP	Log of Next Years Salary	Fit	SE Fit	Residual	St Resid
7	0.515	17.0000	16.7948	0.3291	0.2052	0.37 X
20	0.357	13.0000	14.3636	0.0847	-1.3636	-2.13R
23	0.477	16.0000	16.5347	0.2601	-0.5347	-0.90 X
44	0.322	16.0000	14.3741	0.0987	1.6259	2.54R
55	0.353	16.0000	14.6111	0.0778	1.3889	2.16R
62	0.358	13.0000	14.3837	0.0848	-1.3837	-2.16R

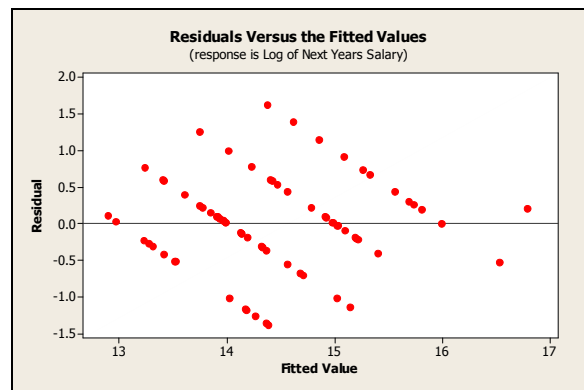
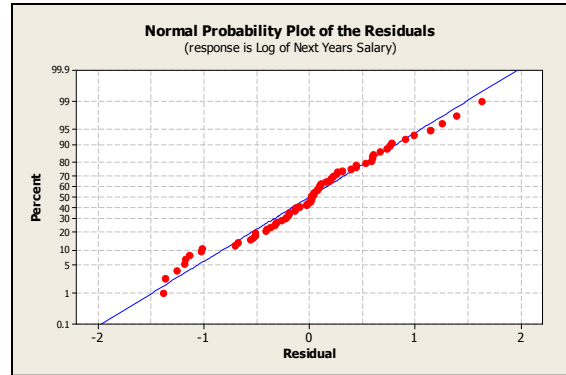


EXHIBIT 6A: OUTPUT OF FULL MODEL USING ALL VARIABLES

The regression equation is

$$\begin{aligned} \text{Log of Next Years Salary} = & 14.7 - 0.0102 G + 0.0052 AB - 0.0074 R + 0.0193 H \\ & - 0.0435 2B - 0.109 3B - 0.061 HR - 0.0023 RBI \\ & + 0.0225 SB - 0.0177 CS - 0.0182 BB - 0.00291 SO \\ & - 0.0361 IBB - 0.0487 HBP - 0.0072 SH + 0.120 SF \\ & - 0.0357 GIDP - 21.1 BA + 49 OBP + 46 SLG - 34 OPS \\ & - 0.566 SB\% + 0.0167 RC - 0.0450 \text{ Age} \\ & + 0.110 \text{ US Native?} - 0.00093 \text{ Career G} \\ & + 0.0050 \text{ Career AB} + 0.00719 \text{ Career R} \\ & - 0.0192 \text{ Career H} - 0.0112 \text{ Career 2B} \\ & - 0.0595 \text{ Career 3B} - 0.0293 \text{ Career HR} \\ & - 0.00418 \text{ Career RBI} + 0.0109 \text{ Career SB} \\ & - 0.0446 \text{ Career CS} - 0.0061 \text{ Career BB} \\ & - 0.00577 \text{ Career SO} - 0.0117 \text{ Career IBB} \\ & - 0.0364 \text{ Career HBP} + 0.0335 \text{ Career SH} \\ & - 0.0267 \text{ Career SF} - 0.0140 \text{ Career GIDP} \\ & - 29.9 \text{ Career BA} - 303 \text{ Career OBP} - 295 \text{ Career SLG} \\ & + 305 \text{ Career OPS} + 0.40 \text{ Career SB\%} + 0.0294 \text{ Career RC} \\ & + 0.266 \text{ Old Age} \end{aligned}$$

Predictor	Coef	SE Coef	T	P
Constant	14.695	5.880	2.50	0.020
G	-0.010243	0.009553	-1.07	0.295
AB	0.00518	0.01637	0.32	0.755
R	-0.00737	0.01872	-0.39	0.698
H	0.01927	0.07406	0.26	0.797
2B	-0.04345	0.05047	-0.86	0.399
3B	-0.1094	0.1054	-1.04	0.310
HR	-0.0608	0.1435	-0.42	0.676
RBI	-0.00232	0.01793	-0.13	0.898
SB	0.02254	0.03353	0.67	0.508
CS	-0.01767	0.07246	-0.24	0.810
BB	-0.01818	0.03777	-0.48	0.635
SO	-0.002914	0.006602	-0.44	0.663
IBB	-0.03606	0.04495	-0.80	0.431
HBP	-0.04869	0.04065	-1.20	0.244
SH	-0.00721	0.05316	-0.14	0.893
SF	0.12047	0.06584	1.83	0.081
GIDP	-0.03568	0.02932	-1.22	0.236
BA	-21.13	23.47	-0.90	0.378
OBP	49.0	233.9	0.21	0.836
SLG	46.5	234.1	0.20	0.845
OPS	-34.3	233.1	-0.15	0.884
SB%	-0.5661	0.3726	-1.52	0.143
RC	0.01665	0.08615	0.19	0.848
Age	-0.04498	0.06663	-0.68	0.507
US Native?	0.1100	0.2844	0.39	0.703
Career G	-0.000929	0.002027	-0.46	0.651
Career AB	0.00501	0.01019	0.49	0.628
Career R	0.007187	0.006268	1.15	0.264
Career H	-0.01917	0.05634	-0.34	0.737

Career 2B	-0.01118	0.02727	-0.41	0.686
Career 3B	-0.05951	0.05533	-1.08	0.294
Career HR	-0.02930	0.07835	-0.37	0.712
Career RBI	-0.004175	0.006597	-0.63	0.533
Career SB	0.01094	0.01640	0.67	0.512
Career CS	-0.04459	0.04891	-0.91	0.372
Career BB	-0.00606	0.02325	-0.26	0.797
Career SO	-0.005769	0.002942	-1.96	0.063
Career IBB	-0.01171	0.01604	-0.73	0.473
Career HBP	-0.03641	0.01706	-2.13	0.044
Career SH	0.03350	0.01655	2.02	0.055
Career SF	-0.02670	0.03385	-0.79	0.439
Career GIDP	-0.01397	0.01106	-1.26	0.220
Career BA	-29.89	25.61	-1.17	0.256
Career OBP	-303.2	206.4	-1.47	0.156
Career SLG	-295.1	214.1	-1.38	0.182
Career OPS	304.8	212.3	1.44	0.165
Career SB%	0.401	1.333	0.30	0.766
Career RC	0.02936	0.07689	0.38	0.706
Old Age	0.2660	0.4205	0.63	0.533

S = 0.556305 R-Sq = 91.0% R-Sq(adj) = 71.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	49	68.9693	1.4075	4.55	0.000
Residual Error	22	6.8085	0.3095		
Total	71	75.7778			

EXHIBIT 6B: OUTPUT OF FULL MODEL NOT USING CAREER STATISTICS

* TB is highly correlated with other X variables
 * TB has been removed from the equation.

* HR+RBI is highly correlated with other X variables
 * HR+RBI has been removed from the equation.

The regression equation is

$$\begin{aligned} \text{Log of Next Years Salary} = & 14.0 - 0.00167 G + 0.0079 AB - \\ & 0.0133 R - 0.0130 H \\ & + 0.0010 2B - 0.0623 3B + 0.015 HR + 0.0039 \\ \text{RBI} & \\ & + 0.0160 SB - 0.0306 CS - 0.0175 BB - \\ 0.00728 \text{ SO} & \\ & - 0.0555 IBB - 0.0106 HBP + 0.0323 SH - \\ 0.0062 \text{ SF} & \\ & - 0.0200 \text{ GDP} - 7.8 \text{ BA} - 46 \text{ OBP} - 62 \text{ SLG} + \\ 63 \text{ OPS} & \\ & - 0.213 \text{ SB}\% + 0.0263 \text{ RC} - 0.145 \text{ Age} \\ & - 0.039 \text{ US Native?} + 0.643 \text{ Old Age} \end{aligned}$$

Predictor	Coef	SE Coef	T	P
Constant	14.048	3.448	4.07	0.000
G	-0.001675	0.006996	-0.24	0.812
AB	0.00794	0.01351	0.59	0.560
R	-0.01334	0.01362	-0.98	0.332
H	-0.01298	0.06016	-0.22	0.830
2B	0.00103	0.03931	0.03	0.979
3B	-0.06231	0.09067	-0.69	0.495

HR	0.0152	0.1068	0.14	0.888
RBI	0.00387	0.01364	0.28	0.778
SB	0.01604	0.02605	0.62	0.541
CS	-0.03065	0.06216	-0.49	0.624
BB	-0.01752	0.03001	-0.58	0.562
SO	-0.007276	0.004650	-1.56	0.125
IBB	-0.05548	0.03485	-1.59	0.118
HBP	-0.01062	0.03338	-0.32	0.752
SH	0.03231	0.04041	0.80	0.428
SF	-0.00623	0.04825	-0.13	0.898
GIDP	-0.02001	0.02368	-0.85	0.403
BA	-7.77	15.03	-0.52	0.608
OBP	-46.4	202.8	-0.23	0.820
SLG	-62.4	202.5	-0.31	0.759
OPS	63.3	201.8	0.31	0.755
SB%	-0.2128	0.3310	-0.64	0.523
RC	0.02629	0.06929	0.38	0.706
Age	-0.14550	0.04202	-3.46	0.001
US Native?	-0.0391	0.2114	-0.19	0.854
Old Age	0.6430	0.3414	1.88	0.066

S = 0.603526 R-Sq = 78.4% R-Sq(adj) = 65.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	26	59.3868	2.2841	6.27	0.000
Residual Error	45	16.3910	0.3642		
Total	71	75.7778			

EXHIBIT 7A: CORRELATION MATRIX – INCLUDES NATURAL LOG OF SALARY AND FREE AGENCY YEAR STATISTICS

Log-	Sal	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP	BA	OBP	SLG	TB	OPS	SB%	RC	
G	0.55																									
AB	0.65	0.87																								
R	0.74	0.77	0.87																							
H	0.73	0.83	0.96	0.91																						
2B	0.67	0.66	0.81	0.79	0.84																					
3B	0.15	0.22	0.34	0.35	0.31	0.06																				
HR	0.68	0.58	0.57	0.76	0.65	0.61	-0.07																			
RBI	0.72	0.72	0.75	0.83	0.82	0.74	0.02	0.91																		
SB	0.16	0.16	0.29	0.34	0.24	0.06	0.72	-0.10	-0.04																	
CS	0.15	0.21	0.36	0.38	0.29	0.17	0.64	-0.04	0.01	0.82																
BB	0.61	0.61	0.56	0.74	0.58	0.57	0.06	0.78	0.71	0.14	0.08															
SO	0.42	0.59	0.66	0.61	0.62	0.51	0.13	0.68	0.71	0.07	0.19	0.52														
IBB	0.44	0.33	0.22	0.46	0.31	0.36	-0.16	0.73	0.56	-0.09	-0.12	0.78	0.30													
HBP	0.40	0.37	0.43	0.45	0.43	0.45	-0.04	0.43	0.41	0.09	0.13	0.39	0.37	0.32												
SH	-0.07	0.05	0.12	0.03	0.07	-0.04	0.39	-0.40	-0.31	0.34	0.39	-0.22	-0.29	-0.33	0.08											
SF	0.45	0.56	0.61	0.56	0.64	0.57	0.13	0.38	0.65	-0.02	-0.03	0.38	0.36	0.15	0.29	0.07										
GIDP	0.29	0.50	0.57	0.32	0.57	0.56	-0.10	0.24	0.44	-0.19	-0.16	0.19	0.37	0.00	0.17	0.00	0.51									
BA	0.50	0.10	0.18	0.39	0.42	0.35	0.05	0.40	0.41	-0.01	-0.05	0.24	0.06	0.32	0.10	-0.11	0.25	0.18								
OBP	0.57	0.24	0.21	0.54	0.38	0.39	-0.07	0.67	0.56	0.04	-0.06	0.77	0.26	0.74	0.32	-0.30	0.22	0.03	0.69							
SLG	0.63	0.28	0.31	0.60	0.46	0.52	-0.08	0.88	0.74	-0.12	-0.09	0.64	0.46	0.71	0.34	-0.44	0.24	0.11	0.65	0.80						
TB	0.78	0.79	0.87	0.93	0.93	0.84	0.18	0.88	0.94	0.12	0.18	0.74	0.71	0.55	0.47	-0.14	0.58	0.46	0.45	0.56	0.72					
OPS	0.64	0.28	0.30	0.61	0.46	0.50	-0.08	0.86	0.72	-0.08	-0.09	0.71	0.42	0.75	0.35	-0.42	0.25	0.09	0.69	0.90	0.98	0.70				
SB%	-0.02	-0.06	-0.04	0.06	0.00	-0.06	0.03	0.01	-0.01	0.28	0.00	0.03	-0.12	0.09	0.11	-0.02	-0.05	-0.09	0.12	0.12	0.03	0.00	0.06			
RC	0.78	0.70	0.75	0.90	0.84	0.77	0.14	0.91	0.91	0.11	0.11	0.85	0.62	0.70	0.45	-0.19	0.52	0.37	0.54	0.73	0.81	0.96	0.82	0.05		
Age	-0.42	-0.09	-0.19	-0.19	-0.23	-0.22	-0.11	-0.04	-0.09	-0.16	-0.21	-0.01	-0.02	0.00	-0.19	-0.19	-0.14	0.02	-0.29	-0.13	-0.12	-0.17	-0.13	0.09	-0.14	

EXHIBIT 7B: CORRELATION MATRIX – INCLUDES NATURAL LOG OF SALARY AND CAREER STATISTICS

	Log- Sal	Car. AB	Car. R	Car. H	Car. 2B	Car. 3B	Car. HR	Car. RBI	Car. SB	Car. CS	Car. BB	Car. SO	Car. IBB	Car. HBP	Car. SH	Car. SF	Car. GID	Car. P	Car. BA	Car. OBP	Car. SLG	Car. TB	Car. OPS	Car. SB%	Car. HR+	Car. RB	Car. RC	Car. Old	Car. Age		
Career AB	-0.16																														
Career R	-0.14	0.98																													
Career H	-0.17	0.99	0.98																												
Career 2B	-0.17	0.97	0.96	0.97																											
Career 3B	-0.16	0.81	0.81	0.81	0.77																										
Career HR	-0.04	0.77	0.80	0.75	0.77	0.53																									
Career RBI	-0.15	0.97	0.97	0.96	0.96	0.75	0.88																								
Career SB	0.00	0.80	0.83	0.82	0.76	0.77	0.53	0.74																							
Career CS	-0.05	0.82	0.83	0.83	0.76	0.79	0.51	0.74	0.92																						
Career BB	-0.11	0.91	0.94	0.90	0.87	0.76	0.78	0.91	0.80	0.80																					
Career SO	-0.10	0.92	0.92	0.89	0.89	0.72	0.88	0.94	0.73	0.72	0.92																				
Career IBB	-0.13	0.74	0.69	0.70	0.67	0.54	0.71	0.74	0.50	0.50	0.72	0.73																			
Career HBP	-0.25	0.65	0.67	0.65	0.72	0.44	0.65	0.68	0.41	0.39	0.56	0.64	0.40																		
Career SH	0.01	0.28	0.25	0.25	0.26	0.30	0.10	0.19	0.18	0.23	0.19	0.28	0.00	0.24																	
Career SF	-0.24	0.92	0.90	0.92	0.89	0.72	0.69	0.90	0.68	0.69	0.82	0.82	0.70	0.65	0.22																
Career GDP	-0.19	0.94	0.92	0.95	0.90	0.74	0.64	0.90	0.77	0.80	0.83	0.81	0.64	0.53	0.17	0.86															
Career BA	-0.12	0.22	0.31	0.31	0.33	0.25	0.18	0.27	0.26	0.25	0.21	0.09	-0.01	0.28	-0.11	0.21	0.27														
Career OBP	-0.09	0.08	0.22	0.15	0.17	0.13	0.19	0.17	0.20	0.14	0.30	0.09	0.02	0.22	-0.21	0.07	0.09	0.81													
Career SLG	-0.03	0.19	0.31	0.23	0.30	0.17	0.59	0.38	0.14	0.08	0.29	0.31	0.17	0.39	-0.20	0.15	0.15	0.63	0.67												
Career TB	-0.15	0.99	0.99	0.99	0.97	0.79	0.85	0.99	0.78	0.79	0.91	0.93	0.73	0.69	0.23	0.91	0.91	0.30	0.17	0.34											
Career OPS	-0.05	0.17	0.30	0.22	0.27	0.17	0.49	0.33	0.17	0.11	0.32	0.25	0.12	0.36	-0.22	0.14	0.14	0.76	0.85	0.96	0.31										
Career SB%	0.21	0.27	0.30	0.26	0.28	0.24	0.30	0.27	0.41	0.20	0.24	0.28	0.18	0.29	0.08	0.25	0.18	0.09	0.02	0.20	0.29	0.15									
Career HR+RB	-0.13	0.95	0.95	0.94	0.94	0.72	0.92	1.00	0.71	0.71	0.90	0.95	0.75	0.69	0.18	0.88	0.86	0.25	0.18	0.43	0.98	0.37	0.28								
Career RC	-0.15	0.97	0.99	0.98	0.96	0.79	0.83	0.98	0.81	0.80	0.93	0.92	0.71	0.67	0.20	0.89	0.92	0.36	0.26	0.38	0.99	0.36	0.28	0.97							
Old Age	-0.29	0.36	0.35	0.33	0.38	0.30	0.36	0.37	0.21	0.22	0.41	0.42	0.34	0.17	0.00	0.29	0.26	0.04	0.08	0.17	0.36	0.15	0.05	0.38	0.36						
US Native?	0.19	-0.25	-0.25	-0.26	-0.24	-0.24	-0.17	-0.26	-0.14	-0.13	-0.19	-0.20	-0.13	-0.30	-0.08	-0.28	-0.26	-0.01	0.08	0.00	-0.25	0.03	-0.11	-0.25	-0.25	0.01					

EXHIBIT 8A: SUBSET REGRESSION, USING MOST CORRELATED VARIABLES TO NATURAL LOG OF SALARIES, W/O VARIABLE REQUIRED IN ALL MODELS

Response is Log of Next Years Salary

Vars	R-Sq	R-Sq(adj)	Mallows	C-p	S	R T	R	B H 2 A P L B B	O S O	A
						C B R H I R B B S G B P G A O e				B S g
1	60.8	60.3	16.5	0.65115	X					
1	60.6	60.0	17.0	0.65323	X					
2	70.5	69.7	-2.4	0.56906	X					X
2	69.0	68.1	0.9	0.58377	X					X
3	71.4	70.1	-2.2	0.56480	X			X		X
3	71.2	70.0	-1.9	0.56618	X					X X
4	72.8	71.1	-3.3	0.55491	X			X		X X
4	72.8	71.1	-3.2	0.55508	X				X	X X
5	73.3	71.3	-2.4	0.55381	X		X	X	X	X X
5	73.1	71.1	-2.0	0.55551	X		X	X	X	X X
6	73.4	70.9	-0.6	0.55705	X		X	X	X	X X
6	73.4	70.9	-0.6	0.55714	X		X	X	X	X X X

EXHIBIT 8B: SUBSET REGRESSION, USING MOST CORRELATED VARIABLES TO NATURAL LOG OF SALARIES, W/O BASE PERCENTAGE VARIABLE REQUIRED IN ALL MODELS

Response is Log of Next Years Salary

The following variables are included in all models: OBP

Vars	R-Sq	R-Sq(adj)	Mallows	C-p	S	R T	R	B H 2 A P L B B	O S O	A
						C B R H I R B B S G B P G A O e				B S g
1	63.4	62.3	13.0	0.63421	X					
1	63.2	62.1	13.5	0.63605	X					
2	71.4	70.1	-2.2	0.56480	X					X
2	70.5	69.2	-0.4	0.57314	X					X
3	72.8	71.1	-3.2	0.55508	X					X X
3	71.4	69.7	-0.4	0.56839	X			X		X
4	72.9	70.8	-1.5	0.55781	X		X			X X
4	72.9	70.8	-1.5	0.55808	X		X		X	X X
5	73.3	70.9	-0.5	0.55753	X		X	X	X	X X
5	73.3	70.9	-0.5	0.55757	X		X	X	X	X X
6	73.5	70.7	1.1	0.55968	X		X	X	X	X X
6	73.5	70.6	1.2	0.56040	X		X	X	X	X X

EXHIBIT 9A: OUTPUT OF STEPWISE REGRESSION USING ALL AVAILABLE VARIABLES W/ ALPHA AT .05

Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.05

Response is Log of Next Years Salary on 53 predictors, with N = 72

Step	1	2	3	4	5	6	7	8
Constant	8.76	11.882	14.143	14.424	14.039	13.205	13.35	14.32
AB	0.00421	0.00386	0.00045					
T-Value	7.24	7.16	0.32					
P-Value	0	0	0.751					
OBP	11.7	11.1	5.4	4.8	4.8	4.4	3.3	
T-Value	6.02	6.2	1.94	2.39	2.47	2.35	1.8	
P-Value	0	0	0.056	0.02	0.016	0.022	0.076	
Age		-0.082	-0.088	-0.09	-0.073	-0.069	-0.063	-0.063
T-Value		-3.78	-4.21	-4.36	-3.37	-3.36	-3.13	-3.08
P-Value		0	0	0	0.001	0.001	0.003	0.003
TB		0.0071	0.0071	0.0079	0.00798	0.00786	0.0099	0.0112
T-Value		2.6	2.6	7.94	8.22	8.52	7.85	10.33
P-Value		0.011	0	0	0	0	0	0
Career HBP					-0.0128	-0.0179	-0.0184	-0.0186
T-Value					-2.15	-3.01	-3.2	-3.19
P-Value					0.035	0.004	0.002	0.002
Career SB%						1.5	1.6	1.7
T-Value						2.89	3.17	3.32
P-Value						0.005	0.002	0.001
SO							-0.0064	-0.0076
T-Value							-2.32	-2.79
P-Value							0.023	0.007
S	0.647	0.592	0.569	0.565	0.55	0.523	0.506	0.515
R-Sq	61.9	68.53	71.42	71.37	73.22	76.22	78.04	76.94
R-Sq(adj)	60.8	67.14	69.71	70.11	71.62	74.42	76.01	75.2

EXHIBIT 9B: OUTPUT OF STEPWISE REGRESSION USING MOST HIGHLY CORRELATED VARIABLES W/ ALPHA AT .05

Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.05

Response is Log of Next Years Salary on 16 predictors, with N = 72

Step	1	2	3
Constant	9.345	11.182	14.424
OBP	14.7	5.2	4.8
T-Value	5.86	2.29	2.39
P-Value	0.000	0.025	0.020
TB		0.0084	0.0079
T-Value		7.57	7.94
P-Value		0.000	0.000
Age			-0.090
T-Value			-4.36
P-Value			0.000
S	0.852	0.634	0.565
R-Sq	32.93	63.38	71.37
R-Sq(adj)	31.97	62.31	70.11

EXHIBIT 10: FINAL MODEL REGRESSION OUTPUT

The regression equation is
 Log of Next Years Salary = 14.6 + 3.92 OBP - 0.0845 Age +
 0.00972 TB
 - 0.00557 SO

Predictor	Coef	SE Coef	T	P	VIF
Constant	14.6059	0.9562	15.28	0.000	
OBP	3.925	2.027	1.94	0.057	1.5
Age	-0.08453	0.02038	-4.15	0.000	1.0
TB	0.009719	0.001388	7.00	0.000	3.0
SO	-0.005566	0.003017	-1.84	0.069	2.2

S = 0.555076 R-Sq = 72.8% R-Sq(adj) = 71.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	55.134	13.784	44.74	0.000
Residual Error	67	20.643	0.308		
Total	71	75.778			

Source	DF	Seq SS
OBP	1	24.953
Age	1	9.015
TB	1	20.118
SO	1	1.049

Unusual Observations

Obs	OBP	Log of Next Years Salary	Fit	SE Fit	Residual	St Resid
7	0.515	17.0000	16.9764	0.3151	0.0236	0.05 X
19	0.414	16.0000	14.8616	0.1830	1.1384	2.17R
32	0.356	14.0000	14.6067	0.3132	-0.6067	-1.32 X
44	0.322	16.0000	14.6005	0.1083	1.3995	2.57R
56	0.322	13.0000	14.1865	0.1090	-1.1865	-2.18R
62	0.358	13.0000	14.3243	0.0936	-1.3243	-2.42R

