Professional Baseball Pitchers' Performance and its Effect on Salary

Charles Hills and Marshall Gregory

April 11, 2014

Abstract

In this study we identify factors that affect a Major League Baseball (MLB) pitcher's salary. We are interested in knowing whether ability is a good indicator of compensation. To test this we created a model to predict the salaries of pitchers in the MLB.

1 Introduction

Money is a major driving factor in professional baseball and a major consideration for team managers looking to make changes to their rosters. Baseball is not a fair game: in most professional sports, teams are limited to a salary cap (e.g., the NFL has a salary cap of $133 million per team [1]). In baseball, however, there are no such limitations; team payrolls are limited only by their owners' willingness to pay.

These payrolls may be determined by the amount of money generated by ticket sales or by the sale of team paraphernalia and royalties. There is no set amount required for ticket sales by the MLB, therefore each team can choose to charge as much or as little as they want for tickets. Popular teams with large fanbases are generally able to charge more for tickets or sell tickets in greater volume than less popular teams. Additionally, team payroll may be correlated by the market size of their home city [**?**].

This leads to major discrepancies in the amount teams are able to pay their players and the caliber of players they are able to recruit. In 2013, the Houston Astros had the lowest payroll in baseball at $26.1 million. The New York Yankees, the highest paying team in the league, paid out a staggering $228.1 million - over eight times as much. Alex Rodriguez, the highest paid player on the Yankees and in the league, earned $28 million in 2013: more than every player on the Astros team combined!

With this in mind, it is clear that low-budget teams (often called "small market teams") should be seeking out players who will play for a lower salary but still perform. Contrastingly, large market teams should only accept the best players, and will lure them in with exorbitant salaries. We evaluated the salaries of 345 pitchers in 2013 to see if this is true.

2 Methods

2.1 Sampling

We used a sample of 345 pitchers for this study. Pitchers were chosen as they play a crutial role on the team and are relatively easy to compare to one another. We only considered pitchers who have been playing for three consecutive years (i.e. 2011, 2012, and 2013), as the salaries of rookie pitchers cannot be predicted without statistics from prior years. We also removed from consideration pitchers who make less than \$700,000; these players' salaries are dictated by the MLB price floor, not their (not-so-great) performances.

2.2 Software Used

All statistics and plots were done in R. Charts and formatting were done in LaTeX. Data was gathered and arranged in Microsoft Excel.

3 Results

Since we are want to know if ability is the driving factor behind a pitcher's salary, we chose earned run average (ERA) to quantify this. ERA is defined as

$$ERA = 9 \times \frac{\text{Earned runs allowed}}{\text{Innings pitched}} \tag{1}$$

where “earned runs" are runs not scored on a fielding error [7]. A lower score signifies a pitcher who allows less runs, so a lower ERA is better. Since runs are all that matter at the end of a game, this is a good indicator of a pitcher's performance. It is also worth noting that this is not a count statistic, since it an average per nine innings. The only major downfall here is ERA does not take quality of opponents into account, but since every pitcher are pitching to hundreds of opponents across different teams, this isn't significant.

We defined our testing hypotheses using ERA:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{2}$$

where β_1 is the coefficient for ERA predicting salary.

In order to test these hypotheses, we created by choosing from 36 predictors, including ERA. We choose from all possible models with 5 or fewer predictors based on their Bayesian information criterion (BIC). This selection method helped us deal with multicolinearity and the computational time required to deal with a large number of potential models. BIC is defined as

$$\mathrm{BIC} = -2 \cdot \ln(\hat{L}) + k \cdot (\ln(n) + \ln(2\pi)). \tag{3}$$

where n is the number of data points (in our case 345), k is the number of regressors (this penalizes models using many regressors), and $\hat{L}$ is the maximum value of the likelihood function for the model. Minimizing BIC, we found several good candidates:

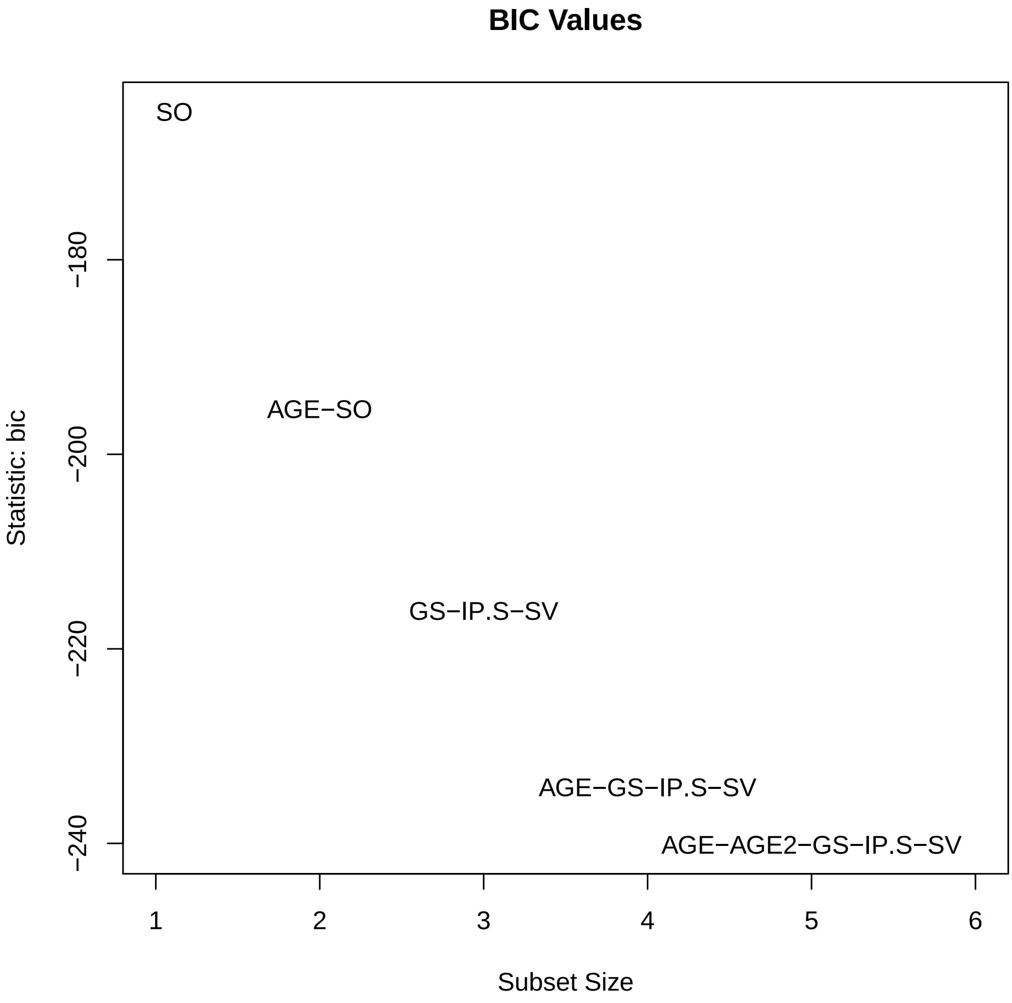

Figure 1: Model candidates and their BIC values. ERA was not chosen in any of the models. Note that models were generated using an exhaustive algorithm (i.e. during each step all models were considered).

Based on this critereon, we chose the predictor variables age (AGE), age squared

(AGE^2), games starting (GS), innings pitched as a starter (IP.S), and saves (SV). In order to test our hypothesis, include ERA.:

Predictor	Min	Mean	Max	Std. Dev.	Obs.
ERA	1.44	3.75	8.11	1.04	345
AGE	21	29	41	3.71	345
AGE^2	441	854.9	961	224.50	345
GS	0	12.4	34	7.856	345
IP.S	0	75.83	236	80.29	345
SV	0	3.518	46	7.86	345

Table 1: Predictor summary statistics.

Using these variables we predict salary using a regression in the form

$$\ln(\text{SALARY}) = \beta_0 + \beta_2\text{AGE} + \beta_3\text{AGE}^2 + \beta_4\text{GS} + \beta_5\text{IP.S} + \beta_6\text{SV} \tag{4}$$

We use ln(SALARY) to deal with heteroscedasticity.

There are no negative salaries, so we expect β_0 to be positive.

As previously discussed, a low ERA indicates a more skilled pitcher, so we expect β_1 to be negative

Since professional athletes tend to get better after their rookie year up until a “peak", and then decline with age, we expect the age predictors will create a concave-down parabola peaking somewhere in the mid to late twenties. Therefore, we expect β_2 to be positive and β_3 to be negative.

Valuable pitchers will start more games, so we expect β_4 to be positive. This will also increase the value of IP.S (along with the stamina required to pitch more innings per game), so we predict β_5 will be positive.

A pitcher who finishes a game records a records a save if at least one of three conditions are satisfied:

- his team is ahead by less than four runs when he enters the game and he pitches for an entire inning
- he enters the game when the enemy team has the potential to tie the game with the next at-bat
- he pitches for at least three innings

A pitcher cannot record a win and a save in the same game [7]. Since a good relief pitcher will rack up more saves and have more opportunities to do so, β_6 is positive.

Running the regression, we found the following coefficient values:

Predictor	Coefficient	Std. Error	t-value	p-value
Intercept	27.254145	1.933914	14.093	$<$ 2e-16
ERA	0.042576	0.028177	1.551	0.13172
AGE	-0.736468	0.131323	-5.608	4.26e-08
AGE^2	0.010664	0.002214	4.817	2.21e-06
GS	0.042898	0.038828	1.105	0.270020
IP.S	-0.017646	0.006203	-2.845	0.004715
SV	-0.068462	0.006258	-10.940	$<$ 2e-16

Table 2: Coefficients for the regression of the model given in (4).

Here, β_3 went against our intuition. Stranger yet, β_3 and β_4 have opposite signs, although IP.S is directly dependent on GS. This combination of indicators provides and interesting metric that values pitchers who pitch many innings with few starts - these are pitchers either have the endurance to pitch deeper into the game or are not pulled early from the game as often.

We also predicted β_6 incorrectly. This is likely negative because savers earn less on average than starting pitchers.

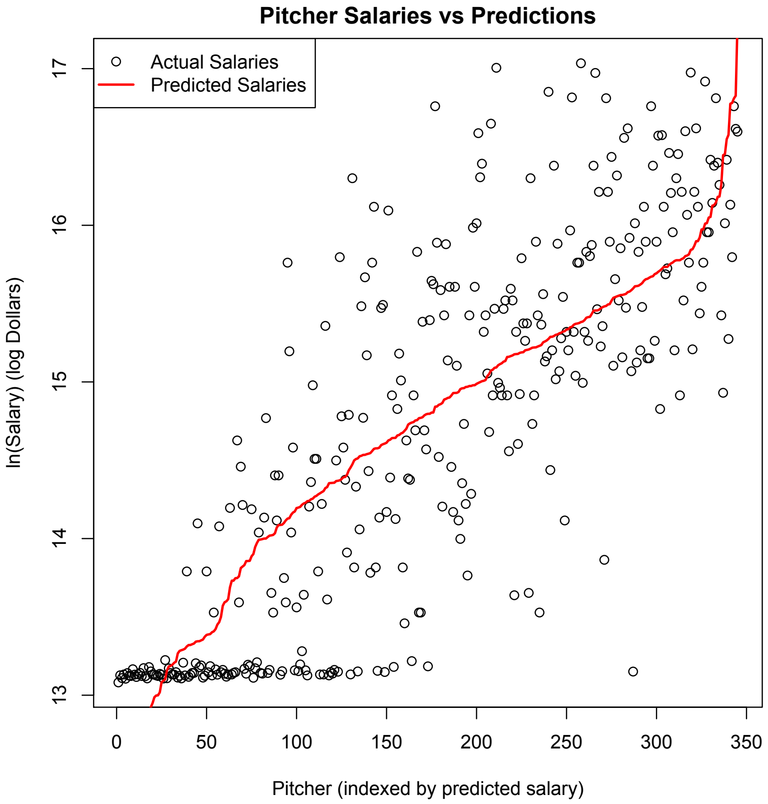

Figure 2: Plot of the regression.

The model fits the data well and has an R^2 value of 0.64, but there are multiple problems which are illustrated by the residual plots:

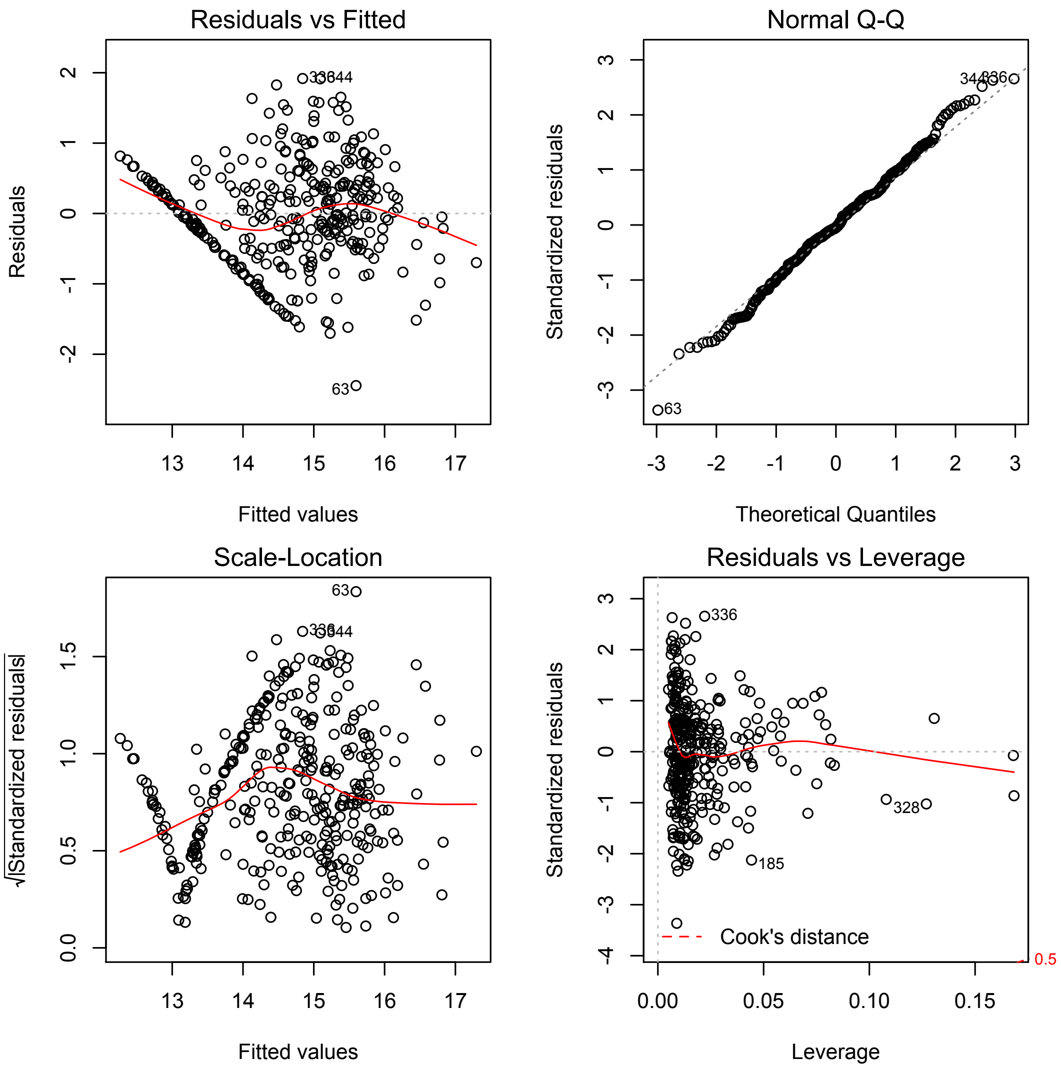

Figure 3: Residual analysis for (4). Note the straight line on the residual plot and the "V shape" on the scale-location plot. Both indicate systematic residuals.

Both of these problems can be explained by the large group of players seen in Figure 2 on the bottom end of salaries. These players are earning the the MLB minimum wage ($500000) [8]. Since these players' salaries are not dictated by the salary price floor and not skill, we removed them and reran the model:

Predictor	Coefficient	Std. Error	t-value	p-value
Intercept	8.293866	1.784664	4.647	5.51e-06
ERA	-0.051700	0.039049	-1.324	0.186755
AGE	0.394701	0.119750	3.296	0.001126
AGE^2	-0.005752	0.001978	-2.907	0.003979
GS	-0.062597	0.033955	-1.844	0.066467
IP.S	0.019358	0.005356	3.614	0.000366
SV	0.047817	0.005014	9.537	$<$ 2e-16

Table 3: Coefficients for the regression of the model given in (4) with minimum wage players excluded.

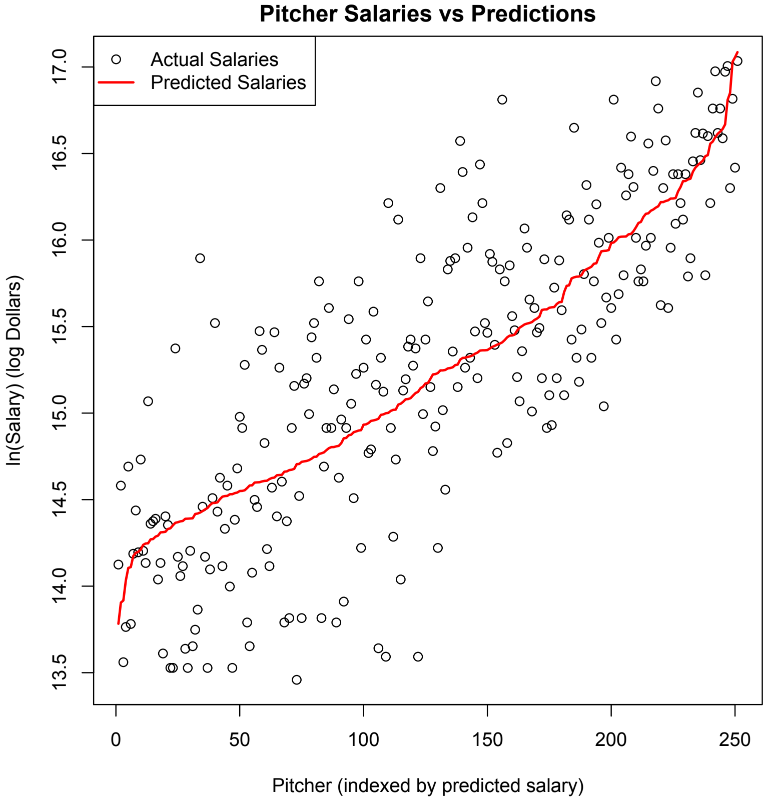

Figure 4: Plot of the regression with minimum wage players excluded.

The transformation did not significantly improve our R^2 value, but this is a much more valid model:

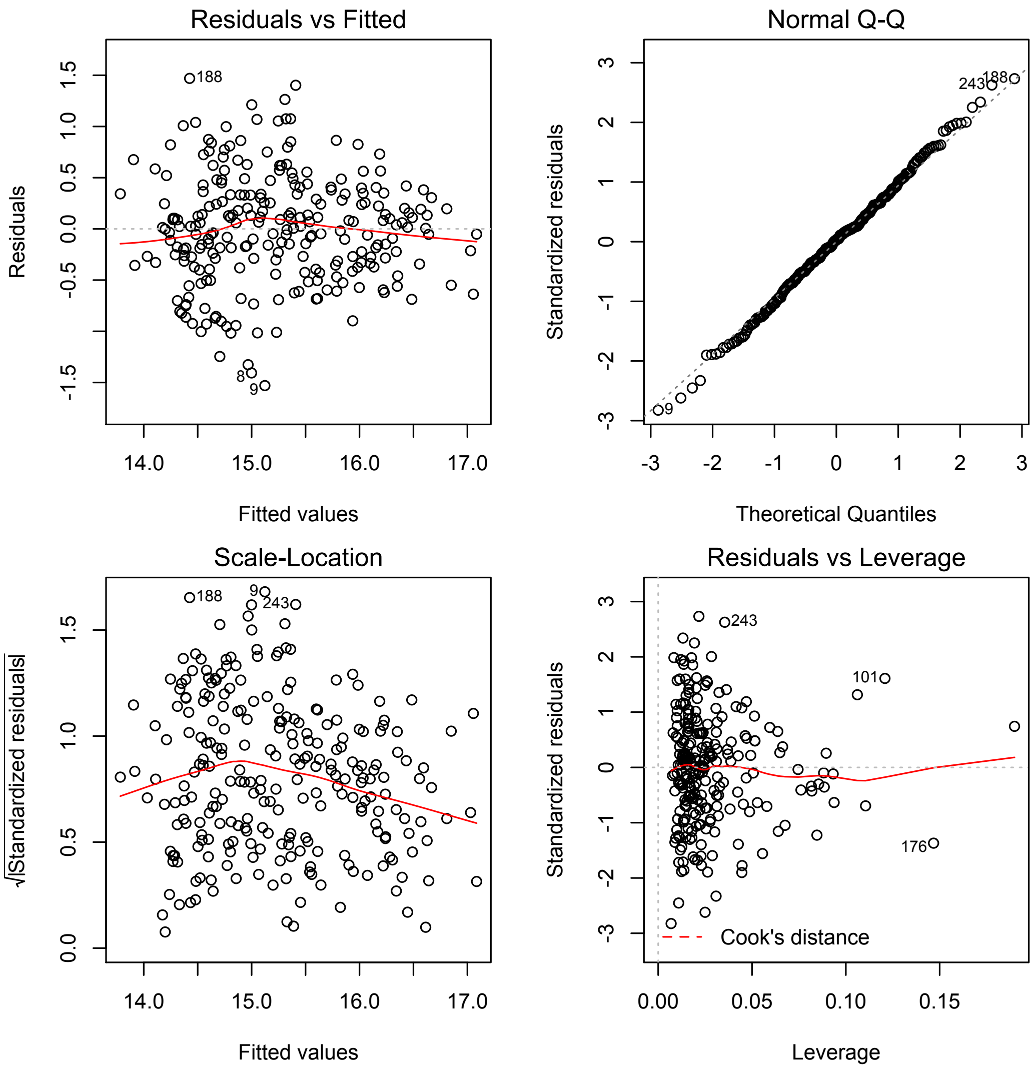

Figure 5: Residual analysis after removing minimum wage players. The only potential problem is the slight movement in standardized residuals, but this is not a big issue considering the data. A transformation other than log might fix this.

We failed H_0 at a 0.05 significance level in this model for both data sets.

4 Conclusion

ERA was not chosen as a strong predictor in our BIC predictor selection, and it was not found to be significant in our model. Every other predictor selected is either a count statistic or age. The count statistics are related to the amount a player is chosen to play, instead of directly measuring his ability; while a better player is certainly likely to play more often, there is could an additional effect of coaches trying to “get their money’s worth" out of highly-paid players.

We hypothesized that high ability should yield a high salary, but this is not strictly the case for the data we observed. One explanation for this could be contract restrictions: contracts can sometimes block players from being paid a salary deserve.

Another variable likely to be significant which we ommited is attendance per game. Since ticket sales generate revenue for teams, a more likable or exciting pitcher may be worth more than a highly skilled one. This might help to explain why age was so significant, since older players have had more time to gather a large fanbase.

We can see that the data does indicate some correlation between salary performance, but this effect is not as direct as we had expected.

References

[1] http://www.nfl.com/news/story/0ap2000000329753/article/nfl-salary-cap-makes-nearly-10m-jump-to-133-million

[2] http://www.baseballprospectus.com/compensation/?cyear=2013

[3] http://www.baseball-reference.com/leagues/MLB/2013-standard-pitching.shtml

[4] http://www.baseballprospectus.com/sortable/index.php?cid=1405180

[5] http://data.newsday.com/long-island/data/baseball/mlb-salaries-2013

[6] http://espn.go.com/mlb/team/salaries/_/name/ari/arizona-diamondbacks

[7] http://mlb.mlb.com/mlb/official_info/baseball_basics/abbreviations.jsp

[8] http://mlb.mlb.com/pa/info/faq.jsp#minimum

5 Code

```
#Get predictor variables and salary data:
X <- read.csv("predictors.csv")
d <- read.csv("PitcherData.csv")
SALARY <- d$SALARY

#Choose predictor variables
library(alr3)
library(leaps)
library(car)
ss <- regsubsets(as.matrix(X),Y, nvmax=5)
rs <- summary(ss)
subsets(ss,statistic=c("bic"),legend=FALSE, xlim=c(1,6))
title("BIC values")

#Create linear model
attach(X)
lm <- lm(SALARY ~ AGE + AGE2 + GS + IP.Start + SV)
sink("lmoutput1.txt")
summary(lm)
fit <- lm$coefficients[1]+lm$coefficients[2]*AGE+lm$coefficients[3]*AGE2+lm$coefficients[4]*GS+lm$coefficients[5]*IP.Start+lm$coefficients[6]*SV

#Plot this model
par(mar=c(4,4,2,2))
plot(SALARY[order(fit)], ylim=c(0,25000000), xlab="Pitcher", ylab="Salary (Millions of Dollars)", axes=FALSE)
box()
axis(2, at=seq(0,25000000,5000000),label=seq(0,25,5))
par(new=t)
plot(fit[order(fit)], ylim=c(0,25000000), col="red", type="l", lwd=2, axes=F, ylab="", xlab="")
title("Pitcher Salaries vs Predictions")
legend("topleft", legend=c("Actual Salaries", "Predicted Salaries"), pch=c(1,26), lty=c(0,1), lwd=c(0,2), col=c("black","red"))

#plot residuals
par(mfrow=c(2,2), mar=c(4,4,2,2))
plot(lm)

powerTransform(lm)
bcSALARY <- 4*(Y^0.25-1)
bclm <- lm(bcSALARY ~ AGE + AGE2 + GS + IP.Start + SV)
sink("lmoutput2.txt")
summary(bclm)
bcfit <- bclm$coefficients[1]+bclm$coefficients[2]*AGE+bclm$coefficients[3]*AGE2+bclm$coefficients[4]*GS+bclm$coefficients[5]*IP.Start+bclm$coefficients[6]*SV

#Plot bc model
par(mfrow=c(1,1), mar=c(4,4,2,2))
plot(bcSALARY[order(bcfit)], ylim=c(111,279), xlab="Pitcher", ylab="Salary (Transformed Dollars)", axes=F)
box()
axis(2)
par(new=t)
plot(bcfit[order(bcfit)], ylim=c(111,279), col="red", type="l", lwd=2, axes=F, ylab="", xlab="")
title("Transformed Pitcher Salaries vs Predictions")
legend("topleft", legend=c("Actual Salaries", "Predicted Salaries"), pch=c(1,26), lty=c(0,1), lwd=c(0,2), col=c("black","red"))

#plot bc residuals
par(mfrow=c(2,2), mar=c(4,4,2,2))
plot(bclm)
```