Which factors can affect to payroll of NBA player? - using LASSO

SAK LEE

May 7, 2016

1 Salaries of NBA players

The National Basketball Association(NBA) is considered to be the No. 1 men's professional basketball leagues in the world. NBA league is the one of the big four leagues in North America. The big four leagues includes Major League Baseball(MLB), National Football League(NFL), National Hockey League(NHL), and NBA. NBA leagues is consists of thirty teams: 29 teams from the United States of America, 1 team from Canada, and the total number of players who is registered in the league is 585 in 2015-2016 season.

Although the NBA takes the third place among the four leagues with respect to the revenue of the leagues, the average salaries of the players is the highest among the four leagues. The average salary of NBA players for the 2013-14 season, \$4.9 million, is higher than \$3.82 million of MLB players which is the second highest average salary among the four.[1] At the same time, the payroll gaps between the players also huge; Kobe Bryant had received the most highest salary, \$30.45 million for the 2013-2014 season while the smallest one was \$490, 180.[1] Since the most of the players in NBA usually sign a contract with their club for several years, a contract could give an owner of a club a huge financial losses. Therefore, it is important for the owners of the clubs to have a clear decision rule of player's salary based on the ability of the players.

2 Previous research

Lyons et al. (2015)[3] try to analyze which factors can affect the salaries of the NBA player through the multiple regressions. Based on the study, points per game, rebounds, and personal fouls contributes to a player's salary[3]. Berri et al. (2007)[2] consider the player's position as the one of the independent variables explaining the salaries in the multiple regression, this study still use the common variables; PTS[1], REB[2], BLK[3], and Games played. However, NBA basket ball teams have a five positions whose main functions are slightly different from each other. For example, a point guard position in the basketball team is expected to distribute the ball to the other team player so that the offense of

[1]Points scored, per game
[2]Rebounds, per game
[3]Block shots, per game

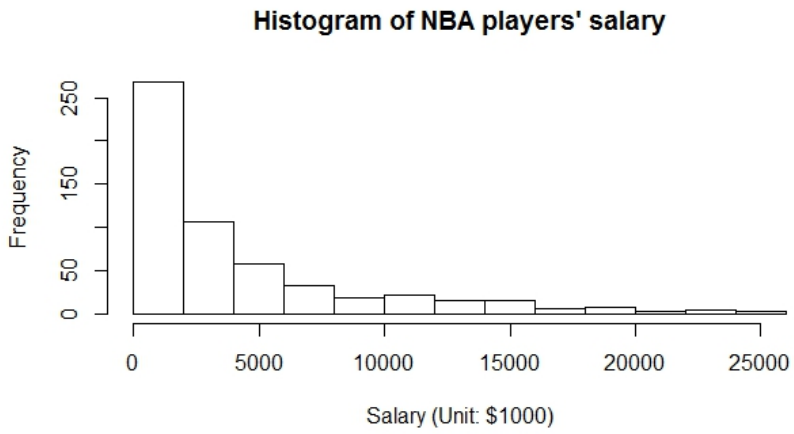

Figure 1: Histogram of the salary of NBA players

the team become more effective. Therefore, it could be more reasonable to decide to give a high salary when the point guard has more assists, steal, and scoring rather than the rebound. For these reasons, this study analyzes the data to find the determinants of NBA players salaries depending on the positions of the players in basketball.

3 Data & Methodology

3.1 Used Data

In 2015-2016 season, there are 558 players who have an active contract in the NBA. Since NBA has a regulation of salaries for the rookies, some players who have entered in the league within 3 years are forced to receive under a rookies salary cap (Lyons et al., 2015)[3]. For these reason, some studies exclude the NBA players who is under the rookies salary cap from their data set. However, the reason NBA has a payroll system controlling the player's salary is that the league can provide financial supports to their players relatively more stable than when they have the free agency contract. Therefore, this study considers this aspect of the data by including player's age to the explanatory variable. Since the age is highly correlated with the experiences of players, the age of player is expected to demonstrate the salary system of NBA in the regression setting. Therefore, all 558 player's data is used in the analysis.

Moreover, these 558 data are split into the five data set with respect to the player's position because one of the main goals in this study is to determine the main factors of the NBA player's payroll depending on their position in the game. The size of each groups are around 100 which means each portions of the position are equal. These sub groups of data again split into two groups; one for the estimation, another for the validity. For the independent variables, the individual player's performance information data is used which consists of the following 26 variables:

Age	Age	3P	3-Point
G	Games	3PA	3-Point Attempt
GS	Games Started	3P%	3-Point Percent
MP	Minutes Played	2P	2-Point
FG	Field Goals	2PA	2-Point Attempt
FGA	Field Goals Attempt	2P%	2-Point Percent
FG%	Field Goal Percentage	eFG%	Effective Field Goal Percent
FT	Free Throws	STL	Steals
FTA	Free Throws Attempt	BLK	Blocks
FT%	Free Throws Percentage	TOV	Turnovers
ORB	Offensive Rebounds	PF	Personal Fouls
DRB	Defensive Rebounds	PTS	Points
TRB	Total Rebounds		

Table 1: List of independent variables

3.2 Collecting Data from Web

The data used in this study came from the NBA player's contract data from the reference web site (NBA reference). Since the salary data only has information of names of players and their team, each player's stats are should be matched with salary data. The following is the part of the `Rcode` that gather the urls of the each player's individual stat web pages from the html source code using R package named `XML`. In the html code, each player's url is located between $<$ `a href`$\setminus$" and $\setminus$" $>$ and the `substring` function allows R to grab the urls as follows:

```
# Collecting urls for each players
url <- "http://www.basketball-reference.com"
allplayer.url <- vector(mode = "list", length = 558)
for (i in 1:558){
    player.url <- thepage[grep(tab$Player[i] ,thepage)]
    mypattern.start <- '<a href=\"'
    mypattern.end <- '\">'
    s <- regexpr(mypattern.start, player.url)
    e <- regexpr(mypattern.end, player.url)
    allplayer.url[[i]] <- unique(paste(url, substring(player.url,
    s + attributes(s)$match.length, e - 1), sep = ""))
}
allplayer.url[1:3]
[1] "http://www.basketball-reference.com/players/b/bryanko01.html"
[2] "http://www.basketball-reference.com/players/j/johnsjo02.html"
[3] "http://www.basketball-reference.com/players/j/jamesle01.html"
```

After the player's urls obtained, we grabbed the latest season's stats from the each individual's web site in the similar way. The whole `Rcode` can be accessed from the Appendix.

3.3 Lasso(the least absolute shrinkage and selection operator) method

To determine the effective factors on the salary of the players depending on the position, the least absolute shrinkage and selection operator(Lasso) method will be used for the analysis. Lasso was introduced by Tibshirani (1996)[4] which is similar to ordinary regression analysis except the penalty function in the estimation procedure.

Suppose we have n samples of $(y_i, \underline{x}_i)$, where $i = 1, 2, ...n$. The ordinary least square regression estimator $\hat{\beta}$ minimizes $L(\beta)$ where

$$L(\beta) \equiv \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

However, Lasso estimator also minimizes $L(\beta)$ function above under the constrain of L^1 norm of the parameter. Therefore, object function for the minimization $L^{lasso}(\beta)$ is

$$L^{lasso}(\beta) \equiv \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1} |\beta_j| .$$

By giving an absolute value penalty, $\sum_{j=1} |\beta_j|$,to the regression, Lasso method allows us to benefit not only shrinkage effect but also variable selection which have an impact on the y values. In this study, Lasso will be used to determine that which X variables have an impact of the salaries. Lasso estimator usually can be obtained through the many numerical optimization algorithms, however, Friedman et. al.(2007)[5] suggests the coordinate decent algorithm for calculating lasso estimator because of the simplicity and the calculation speed. Donoho and Johnstone (1995)[6] shows that searching the minimizer of the Lasso objective function in a single predictor case is same as getting the value from the threshold function of the least square estimator,

$$\hat{\beta}^{lasso}(\gamma) = S\left(\hat{\beta}, \gamma\right) \equiv sign\left(\hat{\beta}\right)\left(\left|\hat{\beta}\right| - \gamma\right)_+ = \begin{cases} \hat{\beta} - \gamma, & \text{if } \hat{\beta} > 0 \text{ and } \gamma < \left|\hat{\beta}\right| \\ \hat{\beta} + \gamma, & \text{if } \hat{\beta} < 0 \text{ and } \gamma < \left|\hat{\beta}\right| \\ 0, & \text{if } \gamma \geq \left|\hat{\beta}\right| \end{cases}$$

where $\hat{\beta}$ is the ordinary least square estimator. Here is the coordinate decent algorithm for minimization of the Lasso objective function suggested by Friedman et. al.(2007)[5]:

- For given(fixed) λ, updating the beta vector until it converges;

$$\hat{\beta}_j^{lasso}(\gamma) \leftarrow S\left(\sum_{i=1}^{n} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right), \gamma \right),$$

  where $\tilde{y}_i^{(j)} \equiv \sum_{k \neq j} x_{ik} \tilde{\beta}_k(\gamma)$, $j = 1, 2, ..., p, 1, 2, ...$.

it can be shown $\tilde{\beta}_k(\gamma)$ converges to $L^{lasso}(\beta)$. Finally, to draw the Lasso path which means calculating the every coefficients for every possible λ, the following step is added to the coordinate decent algorithm:

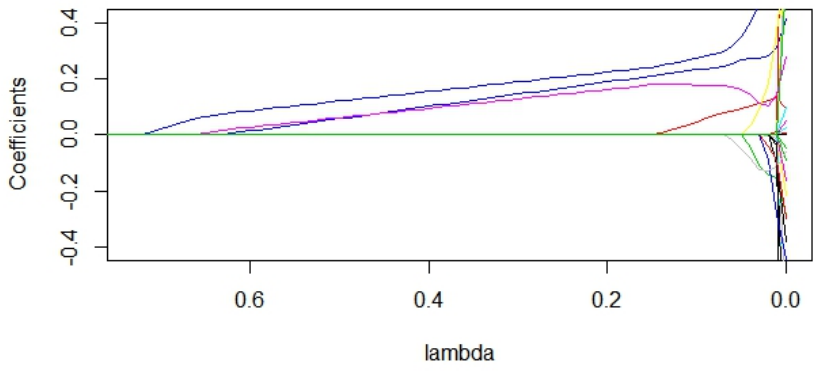

Figure 2: Lasso path for NBA PG(point guard)

- After getting the $L^{lasso}(\beta)$ for given λ, move to the next λ' and used the obtained $\hat{\beta}(\lambda)$ as the initial values for the coordinate decent algorithm for $\hat{\beta}(\lambda')$. This will give a *warm start* to the algorithm.

4 Result and discussion

4.1 Point Guard position

NBA player whose position is PG(point guard) in the basketball plays a roll of distribution of ball and play making. Table 2 shows that selected variables and their coefficients from the Lasso algorithm when given λ is 33. As we expected before, age is selected as an influence factor to players salary which can explain the salary system of the league. The number of games and the number of games as a starting member has a coefficient -0.0459 and 0.2706 respectively. This implies that the player will be likely to get a higher salary as they play more games which coinsides our expectation. However, if they play many game but not a starting member of their team, like a sixman of team, players will have large G and small GS which eventually leads to lower salary than the players who have large G and GS, starting members. Moreover, as PG players tends to more score in a game, they will get a higer salary which agree with the intuition. Note that, however, the coefficient for $2P$ is most highest coefficient among the selected variables which can be seen that among the PG players in the league the market prefer to choose the one is good at not only distributing the ball but also scoring. The number of assist and steals are also positivley corelated to the PG players' salary. The negetive coefficient of the number of block seems weired but it can be ignored since most PG players have few number of blocks in the season.

Variables	Age	G	GS	2P	AST	STL	BLK
Coefficients	0.1015	-0.0459	0.2706	0.4029	0.1415	0.0336	-0.0859

Table 2: Selected variables & Coefficients for PG

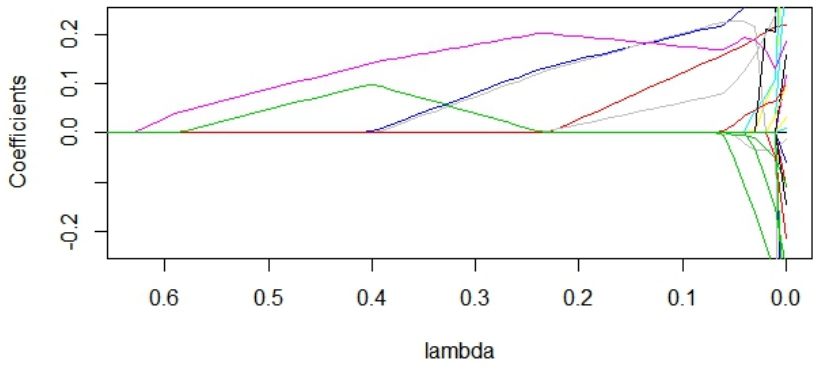

Figure 3: Lasso path for NBA SG(shooting guard)

4.2 Shooting Guard position

NBA player whose position is SG(shooting guard) in the basketball plays a roll of scoring. Table 3 shows the selected variables and thier coefficients for SG players. Note that variable FT and $FT\%$ are newly picked by Lasso comparing to the case of PG players. FT means scored by free throws which players receive when they got faul from other players while the players was making shooting. $FT\%$ means the success probability. These two factors have 0.2199 and 0.225 respectively. The player who has a large FT implies that he is an agressive attacker at the sametime they are hard to guard. Therefore, players who are good at getting fauls from opponents and have higher success probability in free throw highly contribute to their team's winning.

Variables	Age	G	GS	FT	FT%	AST	BLK
Coefficients	0.1587	-0.0045	0.2199	0.225	0.0038	0.1683	0.0812

Table 3: Selected variables & Coefficients for SG

4.3 Small Forward position

NBA player whose position is SF(Small forward) in the basketball also plays a roll of scoring. Table 4 shows the selected variables and thier coefficients for SF players. Similar to the SG players, FT and $FT\%$ are selected. Moreover, players who have more aggresive play stlye, which indicates high FGA and $2PA$, are likely to have a high salary. Note that the coefficient of AST variable for SF players are larger than that for SG players. Since SF players are usually plays in the inner part of the court and they have to closely play with big mans. Therefore, simillar to the PG cases, among the SF players in the league the market prefer to choose the one is good at not only scoring the ball but also assist which directly leads to the team's scoring. Lastly, PF has a negative coefficient to the salary which implies a faul-manage skill is also important in the evaluation of player.

Table 4: Selected variables & Coefficients for SF

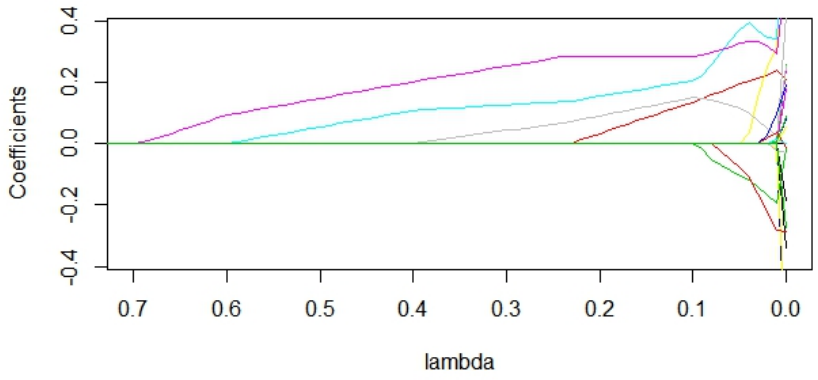

Figure 4: Lasso path for NBA SF(small forward)

Variables	Age	G	FGA	2PA	FT	FT%	AST	PF
Coefficients	0.2148	-0.1382	0.153	0.3688	0.0763	0.0013	0.3325	-0.1632

4.4 Power Forward position & Center forward position

NBA player whose position is PF(Power forward) & C(center forward) in the basketball plays a roll of play maker. As a bigman in the team, they usually play together with other players using screen plays. An interesting selected variables for bigmans from the Lasso result are $eFG\%$ and $3P\%$. $eFG\%$ measures the adjusted percentage of feild goal which will give more weight to the 3 point feild goal. These results reflectes the trend in the NBA league that teams are starting to use their bigmans in the 3 point offensive play in the game. For these reason, if the bigman players has higher percentage of 3 point, their team has one more option to score, which alternatively leads to the winning.

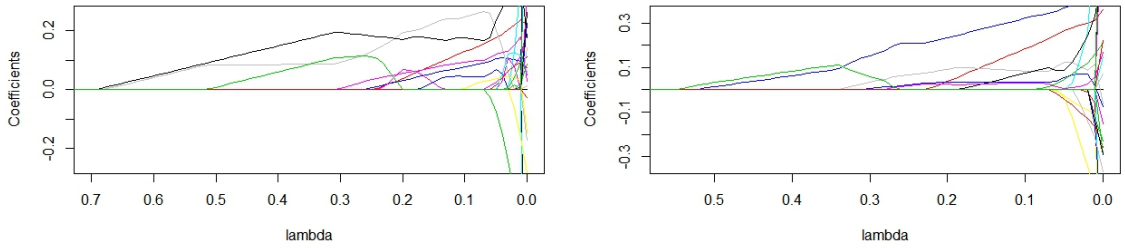

Figure 5: Lasso path for NBA PF(power forward) & C(center forward)

Variables(PF)	Age	GS	2P	eFG%	FT	FTA	AST
Coefficients	0.1375	0.0814	0.0452	0.0138	0.2523	0.1728	0.0958
Variables(C)	Age	GS	3P%	2P	FT	FTA	AST
Coefficients	0.2135	0.0335	0.005	0.3376	0.0814	0.0911	0.0294

Table 5: Selected variables & Coefficients for PF, C

References

[1] Badenhausen, K. (2015). Average MLB player salary nearly double NFL's, but still trails NBA's. ***Forbes***. Retrieved from http://www.forbes.com

[2] Berri, D. J., Brook S. L., & Schmidt, M. B. (2007). Does one simply need to score to score?. *International Journal of Sport Finance, 2*, 142-148.

[3] Lyons, R., Jackson, E. N., & Livingston, A. (2015). Determinants of NBA player salaries. *U.S. Sports Academy - The Sport Journal*. http://thesportjournal.org

[4] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* (1996): 267-288.

[5] Friedman, Jerome, et al. "Pathwise coordinate optimization." *The Annals of Applied Statistics* 1.2 (2007): 302-332.

[6] Donoho, David L., and Iain M. Johnstone. "Adapting to unknown smoothness via wavelet shrinkage." Journal of the american statistical association 90.432 (1995): 1200-1224.

5 Appendix

Rcode for collecting data from web

```
library(XML)
library(beepr)

# Collecting Data from Web
url <- "http://www.basketball-reference.com/contracts/players.html"
thepage <- readLines(url)

# Read table from Web
tab <- readHTMLTable(url, stringsAsFactors = FALSE)[[1]]

# Data cleaning salary, player remove
tab <- tab[-grep("Salary", tab$Player),]
tab <- tab[-grep("Player", tab$Player),]

# Collecting urls for each players
url <- "http://www.basketball-reference.com"
allplayer.url <- vector(mode = "list", length = length(tab$Player))
for (i in 1:length(tab$Player)){
        player.url <- thepage[grep(tab$Player[i] ,thepage)]
        mypattern.start <- '<a href=\"'
        mypattern.end <- '\">'
```

```
        s <- regexpr(mypattern.start, player.url)
        e <- regexpr(mypattern.end, player.url)
        allplayer.url[[i]] <- unique(paste(url, substring(player.url,
                s + attributes(s)$match.length, e - 1), sep = ""))
}
allplayer.url <- unlist(allplayer.url)
allplayer.url <- allplayer.url[grep("html", allplayer.url)]

# Pasting individual stat to salary data
bigtable <- data.frame(matrix(nrow = 585, ncol = 32), stringsAsFactors = FALSE)
total <- length(allplayer.url)
for (i in 1:total){     ability <- readHTMLTable(allplayer.url[i],
                                stringsAsFactors = FALSE)[[1]]
    season <- dim(ability)[1]
    if (season == 1 & length(ability) < 8) {
        bigtable[i, ] <- cbind(tab[i, c(1,2,4)], t(rep(NA, 29)))
    } else{
        bigtable[i, ] <- cbind(tab[i, c(1,2,4)], ability[season, -1])
    }
}

names(bigtable) <- names(cbind(tab[1, c(1,2,4)], ability[season, -1]))
bigtable.NBA <- bigtable[bigtable$Lg == "NBA",]
index <- bigtable.NBA$Pos == ""
bigtable.NBA <- bigtable.NBA[!is.na(bigtable.NBA$`2015-16`) & !index,]

# Dollar to numeric
bigtable.NBA$`2015-16` <- substr(bigtable.NBA$`2015-16`, 2,
                                 nchar(bigtable.NBA$`2015-16`))
bigtable.NBA$`2015-16` <- as.numeric(gsub('\\,', '',
                                 as.character(bigtable.NBA$`2015-16`)))

# change all to numeric using below two line
bigtable.NBA <- as.matrix(bigtable.NBA)
bigtable.NBA <- as.data.frame(bigtable.NBA)

bigtable.NBA[,-c(2,5,6,7)] <- lapply(bigtable.NBA[,-c(2,5,6,7)],
                         function(x) as.numeric(as.character(x)))
bigtable.NBA <- transform(bigtable.NBA, Player = as.character(Player))
bigtable.NBA <- transform(bigtable.NBA, Tm = as.factor(Tm))
bigtable.NBA <- transform(bigtable.NBA, Pos = as.factor(Pos))

# Write CSV in R
# write.csv(bigtable.NBA, file = "bigtableNBA.csv", row.names = FALSE, na = "")
```

```
bigtable.NBAPG <- bigtable.NBA[bigtable.NBA$Pos == "PG",]
bigtable.NBASG <- bigtable.NBA[bigtable.NBA$Pos == "SG",]
bigtable.NBASF <- bigtable.NBA[bigtable.NBA$Pos == "SF",]
bigtable.NBAPF <- bigtable.NBA[bigtable.NBA$Pos == "PF",]
bigtable.NBAC  <- bigtable.NBA[bigtable.NBA$Pos == "C",]
```

Rcode for Lasso path

```
# for bigtable.NBAPG
xdata <- bigtable.NBAPG[-grep('X2015.16', names(bigtable.NBAPG))]
ydata <- bigtable.NBAPG[ grep('X2015.16', names(bigtable.NBAPG))]

xdata <- xdata[-c(1:2,4:6)]
xdata[is.na(xdata)] <- 0
xdata <- scale(xdata)
ydata <- scale(ydata)
xdata <- cbind(rep(1,nrow(xdata)), xdata)
colnames(xdata)[1] <- "Constant"

set.seed(123)
test <- sort(sample(nrow(bigtable.NBAPG), 20, replace = FALSE))
train.x <- xdata[-test,] train.y <- ydata[-test,]

# beta matrix calculate
#================================================
soft_th <- function(z, r){
    sign(z) * max( abs(z) - r, 0 )
}

lambda <- 1 - seq(0.01, 1, length = 100)
beta_matrix <- matrix(0, nrow = 27, ncol = length(lambda))
total <- length(lambda)
# create progress bar
pb <- txtProgressBar(min = 0, max = total, style = 3)
beta_new <- rep(0, 27)
beta_hat <- c(mean(train.y), rep(0, 26))
for (i in 1:length(lambda)){
    lamb <- lambda[i]
    repeat{
        for (j in 1:27){
            y_hat <- train.x %*% beta_hat
            s <- soft_th( mean(train.x[,j] *
                        (train.y - y_hat)) + beta_hat[j] , lamb)
            beta_hat[j] <- s
```

```
        }
            epsilon <- max(abs(beta_new - beta_hat))

if (epsilon < 2.22e-10){
            break
        }
            beta_new <- beta_hat
    }
    beta_matrix[ , i] <- beta_hat

    Sys.sleep(0.1)
    # update progress bar
    setTxtProgressBar(pb, i)
}
beep()
close(pb)

# Calculate the number of selected variable for given lambda
vr.num <- rep(0, length(lambda))
for (i in 1:length(lambda)){
    vr.num[i] <- length(which(beta_matrix[,i] != 0))
}

col <- which(vr.num == 6)[1]
index <- which(beta_matrix[, col] != 0)
beta_matrix[index, col]
fit.x <- train.x[,index]
colnames(fit.x)
b <- max(beta_matrix[index, col]) + 0.01

# Draw Lasso path
plot(lambda, rep(0,100),
     xlim = rev(c(0, lambda[which(vr.num == 1)[1]] + 0.01)),
     ylim = c(-b, b),
     ylab = "Coefficients",
     col = "black",
     type = 'l')
par(new = TRUE)
for(i in 1:27){
    lines(lambda, beta_matrix[i, ], col = i, xlim = rev(range(lambda)))
}
```