

Project - STAT 151A

Aditya Jhanwar

12/8/2019

```
load(url("http://www.stat.berkeley.edu/users/nolan/data/baseball2012.rda"))
baseball = as_tibble(baseball)
```

Data Exploration and Feature Creation

1)

The first step is to clean the `baseball` data by removing unnecessary explanatory variables and entries missing a salary (observed Y_i) value.

```
baseball = baseball %>% dplyr::select(-c("ID", "yearID", "teamID", "lgID", "nameFirst", "nameLast", "G_"))
baseball = baseball %>% drop_na(salary) # remove units with no salary values
```

Next, I followed the author's process in creating new features as described in the textbook.

```
baseball = baseball %>% mutate(AVG = CH/CAB,
                              OBP = 100*(CH + CBB)/(CAB + CBB),
                              CAB.avg = CAB/years,
                              CH.avg = CH/years,
                              CHR.avg = CHR/years,
                              CR.avg = CR/years,
                              CRBI.avg = CRBI/years)
```

Finally, I cleaned the `Position` and `Years` explanatory variables through reimpementing them as dummy variables.

According to Fox's description of his analysis, he mentions **middle infielders** as players who consistently played second base or shortstop so I classified all individuals with either position as such.

```
new.pos = c()
MI = c("12", "1S", "23", "2B", "2S", "3S", "02", "SS")
C = c("C", "C1", "0C")
CF = c("CF")

# Assign new factor assignment for MI (middle infielders), C (catcher), CF (center field), and O (other)
for (i in baseball$POS){
  if (i %in% MI) { new.pos = c(new.pos, "MI") }
  else if (i %in% C) { new.pos = c(new.pos, "C") }
  else if (i %in% CF) { new.pos = c(new.pos, "CF") }
  else { new.pos = c(new.pos, "O") }
}

baseball$POS = relevel(factor(new.pos), "O")
```

```

new.years = c()
neg.cont  = 6
neg.sal   = 3

for (i in baseball$years){
  if      (i >= neg.cont) { new.years = c(new.years, ".cont") }
  else if (i >= neg.sal) { new.years = c(new.years, ".sal")   }
  else                                     { new.years = c(new.years, "other")       }
}

baseball$neg = relevel(factor(new.years), "other")

lm.fit = lm(salary ~ ., data = baseball)
new.baseball = as_tibble(model.matrix(lm.fit)[,-1])
new.baseball$salary = baseball$salary

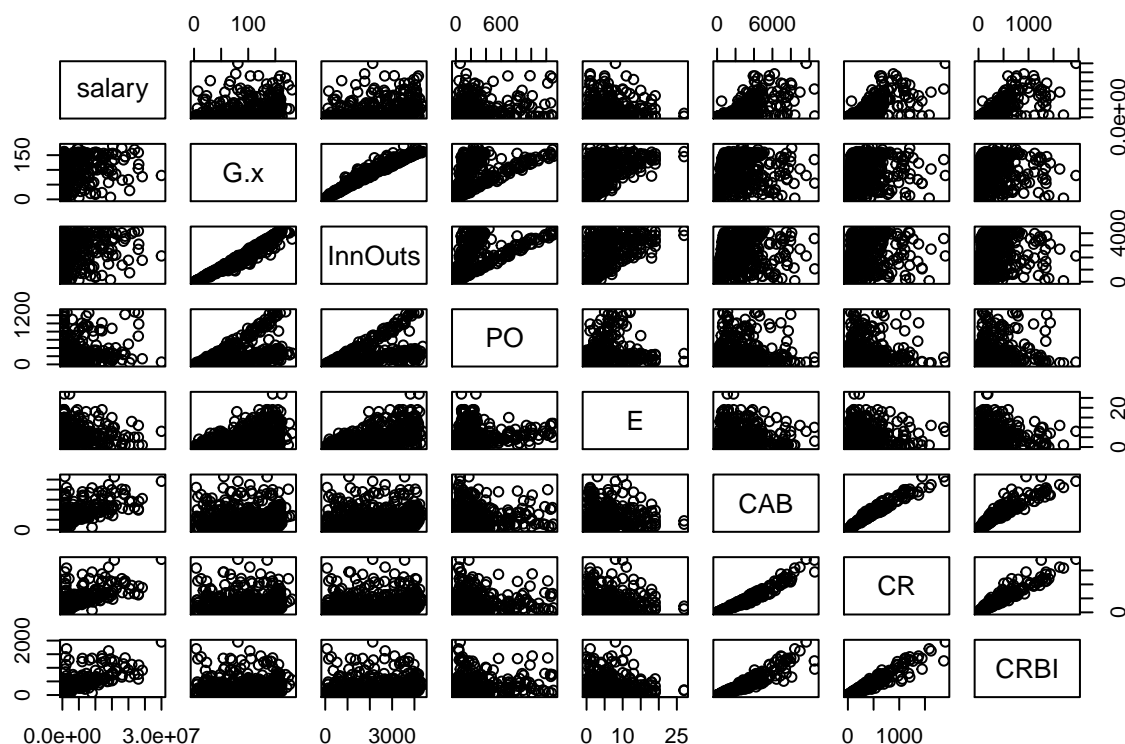
```

Now that we have completed the feature creation process, the next step is to analyze the data itself.

Firstly, I'll look at the structure of the data itself and how the different variables are associated with each other. Since there are a lot of explanatory variables within the data, I will select a few key variables I believe to be the most influential in the model and investigate the structure.

Note: - **G.x** = Position played at specified position - **InnOut** = Time played in the field expressed as outs - **PO** = Putouts - **E** = Errors - **CAB** = Career at bats - **CR** = Career runs - **CRBI** = Career runs batted in

```
pairs(salary ~ G.x + InnOuts + PO + E + CAB + CR + CRBI, data=new.baseball)
```



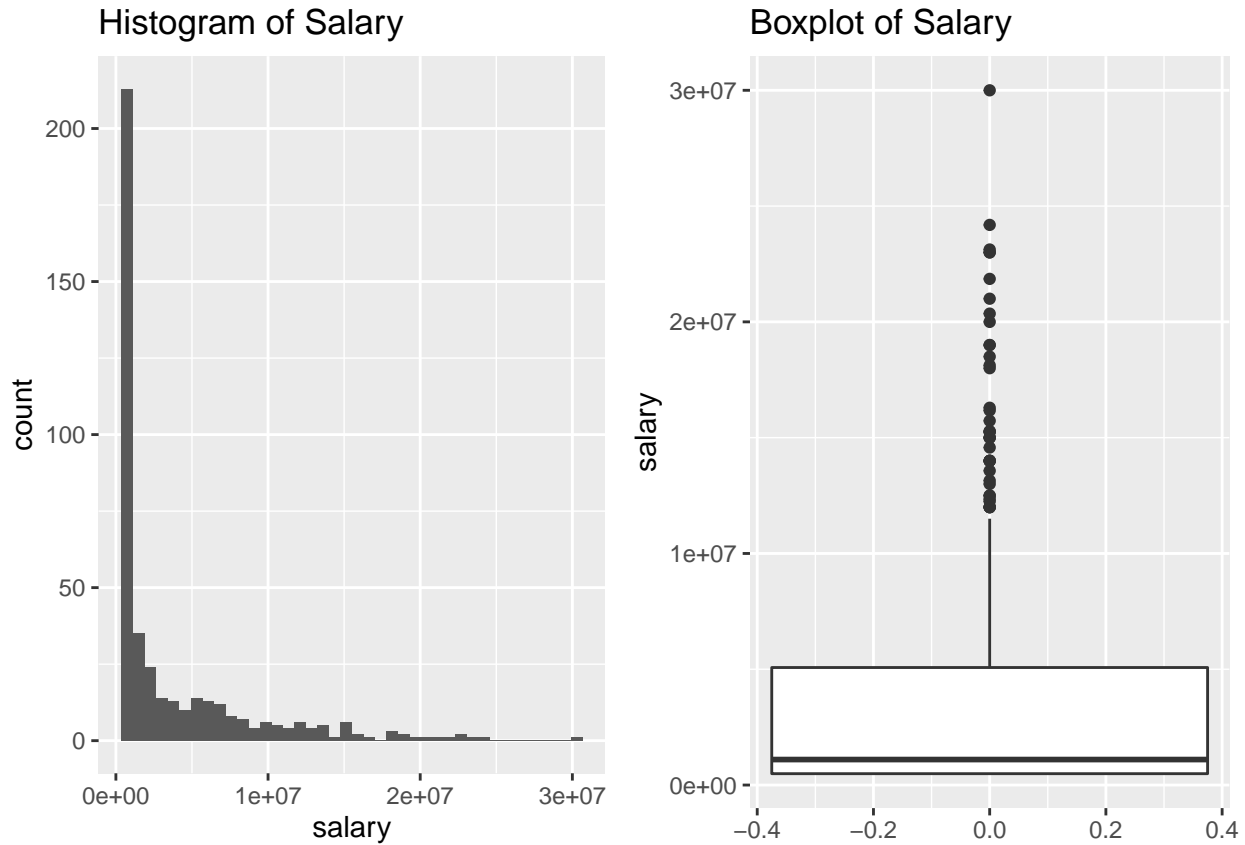
From observing the paired structures of data it is evident that some features are uncorrelated whereas others are strongly correlated. However, this is mostly expected as certain features relate to one another. For example, a player's career at bats would be associated with his career runs or career runs batted in since all tie into a player's capability of scoring bases.

This indicates a possible issue in inference of coefficients through linear modeling since the standard error calculation will be grossly inflated.

In addition, I noticed some of the variables have a stronger correlation with the salary than that of other variables. For example, `G.x` and `InnOuts` do not seem to have a strong association with `salary` whereas `CAB`, `CR`, and `CRBI` have comparatively stronger correlations with `salary`. This indicates some sort of variable selection and model pruning may be of benefit.

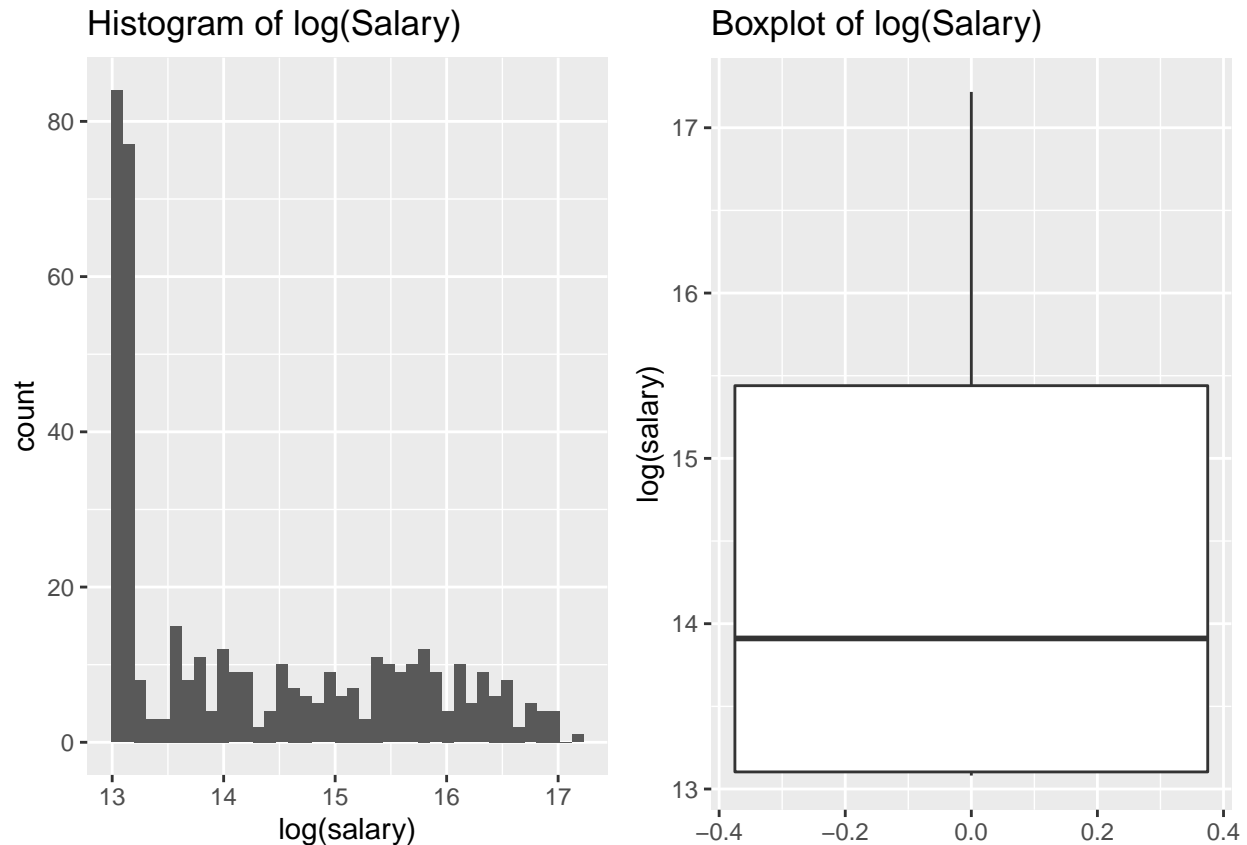
Next, I'd like to look into whether the data is distributed normally as per an assumption of gaussian distributed errors in linear modelling.

```
sal1 = ggplot(data=new.baseball, aes(x=salary)) + geom_histogram(bins=40) + ggtitle("Histogram of Salary")
sal2 = ggplot(data=new.baseball, aes(y=salary)) + geom_boxplot() + ggtitle("Boxplot of Salary")
grid.arrange(sal1, sal2, nrow=1)
```



The histogram above shows that the observed outcome values are not distributed normally at all. Hence, some sort of transformation of the data is necessary in order to use linear modelling.

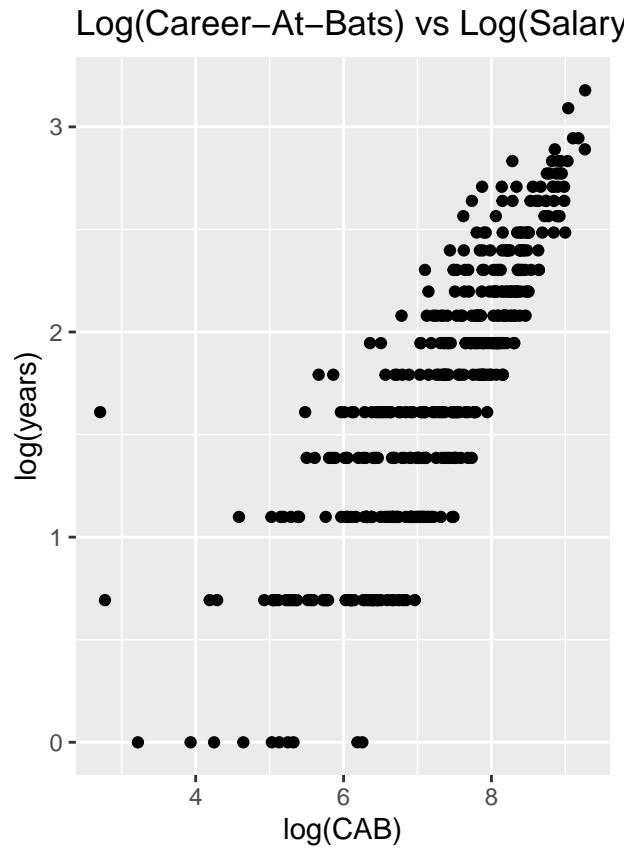
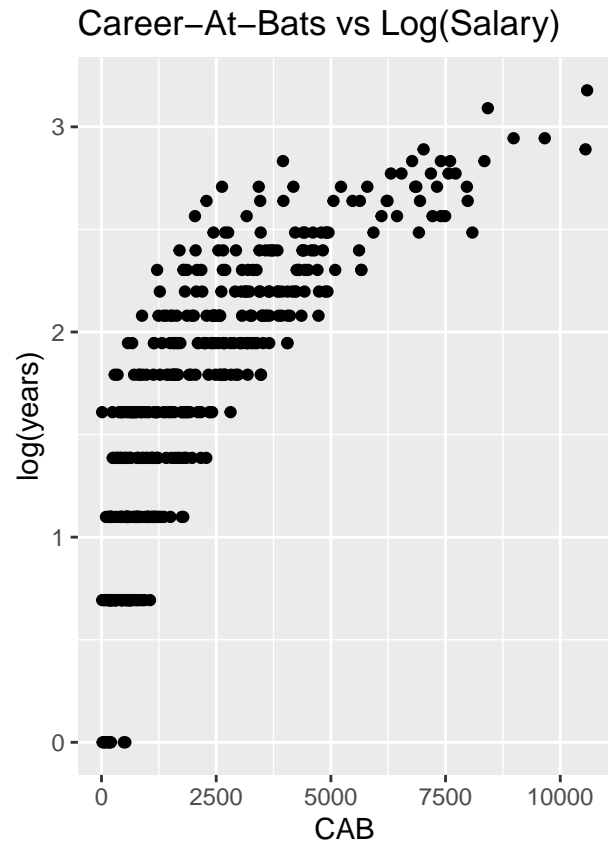
```
log.sal1 = ggplot(data=new.baseball, aes(x=log(salary))) + geom_histogram(bins=40) + ggtitle("Histogram of log(Salary)")
log.sal2 = ggplot(data=new.baseball, aes(y=log(salary))) + geom_boxplot() + ggtitle("Boxplot of log(Salary)")
grid.arrange(log.sal1, log.sal2, nrow=1)
```



Fox mentions log transforming the `salary` data in his linear modelling analysis and this is in line with the observed histograms above. The histogram of the original salaries is right skewed whereas the histogram of the log transformed salaries is somewhat more stabilized and appears more so normally distributed in some sense.

Fox also suggests log transforming some feature variables (years in the majors, career-at-bats) through preliminary examination and so I will do the same to carry forward analysis in a similar manner. An argument for why this may be beneficial is that it might garner a stronger linear relationship between the seemingly most influential explanatory variables and the salary and thus improving the model's predictive capability overall.

```
cab.plot1 = ggplot(data=new.baseball, aes(x=CAB, y=log(years))) + geom_point() + ggtitle(label = "Career")
cab.plot2 = ggplot(data=new.baseball, aes(x=log(CAB), y=log(years))) + geom_point() + ggtitle(label = "Log Career")
grid.arrange(cab.plot1, cab.plot2, nrow=1)
```



```
yrs.plot1 = ggplot(data=new.baseball, aes(x=years, y=log(years))) + geom_point() + ggtitle(label = "Yea
yrs.plot2 = ggplot(data=new.baseball, aes(x=log(years), y=log(years))) + geom_point() + ggtitle(label =
grid.arrange(yrs.plot1, yrs.plot2, nrow=1)
```



```
new.baseball$log.CAB    = log(new.baseball$CAB)
new.baseball$log.years  = log(new.baseball$years)
new.baseball$log.salary = log(new.baseball$salary)
new.baseball = new.baseball %>% dplyr::select(-c(CAB, years, salary))
```

Data Analysis

1)

For the first of the project, I will be fitting a simple model that predicts `log(salary)` from the **dummy variables** for years in majors and `log(career runs)`, allowing for an interactio between the feature variables.

```
dat1 = new.baseball %>% dplyr::select(log.salary, CR, neg.sal, neg.cont)
simple.model = lm(log.salary ~ log(1+CR)*(neg.cont + neg.sal), data=dat1)
simple.model
```

Call:

```
lm(formula = log.salary ~ log(1 + CR) * (neg.cont + neg.sal),
    data = dat1)
```

Coefficients:

(Intercept)	log(1 + CR)	neg.cont
12.88064	0.08743	-3.96431
neg.sal	log(1 + CR):neg.cont	log(1 + CR):neg.sal
-1.06871	0.94036	0.24940

2)

Although I have fitted the simple model above, I want to check for any outliers, high leverage points, and influential observations for further evaluation of the simple model. All criteria in determining such observations will be in line with what Fox suggests using.

```
hat.vals = hatvalues(simple.model)
stud.res = studres(simple.model)
cook.dis = cooks.distance(simple.model)

measures = tibble(Hat.Values=hat.vals, Studentized.Residuals=stud.res, Cooks.Distance=cook.dis)
```

First, I'd like to take a look at the **high leverage** points, which are observations with explanatory variables markedly different from that of the average. In terms of numerical cutoffs for diagnostic statistics, *hat values exceeding twice the average hat value $(k+1)/n$ are noteworthy.*

```
h.3 = 3*length(simple.model$coefficients)/nrow(new.baseball)
high.leverage = measures[hat.vals > h.3,]
high.leverage
```

```
# A tibble: 17 x 3
  Hat.Values Studentized.Residuals Cooks.Distance
    <dbl>             <dbl>             <dbl>
1  0.0477             1.87             0.0289
2  0.0486             0.525            0.00235
3  0.0565             0.0299            0.00000898
4  0.0428            -0.305            0.000695
5  0.0733             0.0664            0.0000582
6  0.0507            -0.267            0.000637
7  0.0428            -0.305            0.000694
8  0.0733             0.0664            0.0000582
9  0.0437             0.475             0.00172
10 0.0554             0.606             0.00359
11 0.126              0.163             0.000643
12 0.0565             0.0284            0.00000807
13 0.0489             0.533             0.00244
14 0.0462             0.504             0.00206
15 0.0573             0.564             0.00323
16 0.245              0.334             0.00606
17 0.148              1.51             0.0663
```

There appears to be **17** data points which have a relatively high leverage.

In addition to high leverage points, I'll analyze discrepant observations to detect outliers within the data through utilizing **studentized residuals** with a numerical cutoff of `|t-test statistic| > 2`

```
outliers = measures[abs(stud.res) > 2,]
outliers %>% head(n=5)
```

```
# A tibble: 5 x 3
  Hat.Values Studentized.Residuals Cooks.Distance
    <dbl>             <dbl>             <dbl>
1  0.0258             2.41             0.0254
```

2	0.00468	-2.18	0.00370
3	0.0274	4.12	0.0768
4	0.00804	-3.01	0.0120
5	0.00476	-2.54	0.00510

There are **26** observations which are determined to be outliers.

Although I have determined observations that have high leverage or are outliers, what I am most concerned about are the subset of these points which have an influence on the determined coefficients of the model. Such points greatly alter the predictive capability of the simple model and thus cannot be overlooked.

Through recommendation by Fox, the criterion I will be using to determine highly influential points is $D_i > 4/(n-k-1)$

```
cook.cutoff = 4/(nrow(new.baseball)-length(simple.model$coefficients))
influential.points = measures[cook.dis > cook.cutoff,]
influential.points %>% arrange(desc(Cooks.Distance))
```

```
# A tibble: 16 x 3
  Hat.Values Studentized.Residuals Cooks.Distance
    <dbl>          <dbl>          <dbl>
1  0.0274          4.12          0.0768
2  0.148           1.51          0.0663
3  0.0189         -4.27          0.0563
4  0.0211         -3.63          0.0462
5  0.0149         -3.53          0.0306
6  0.0477          1.87          0.0289
7  0.0268          2.50          0.0285
8  0.0122         -3.64          0.0264
9  0.0258          2.41          0.0254
10 0.0255          2.40          0.0249
11 0.0272          2.22          0.0227
12 0.0112         -3.06          0.0173
13 0.0149         -2.49          0.0155
14 0.00804        -3.01          0.0120
15 0.0256          1.64          0.0118
16 0.0189          1.86          0.0111
```

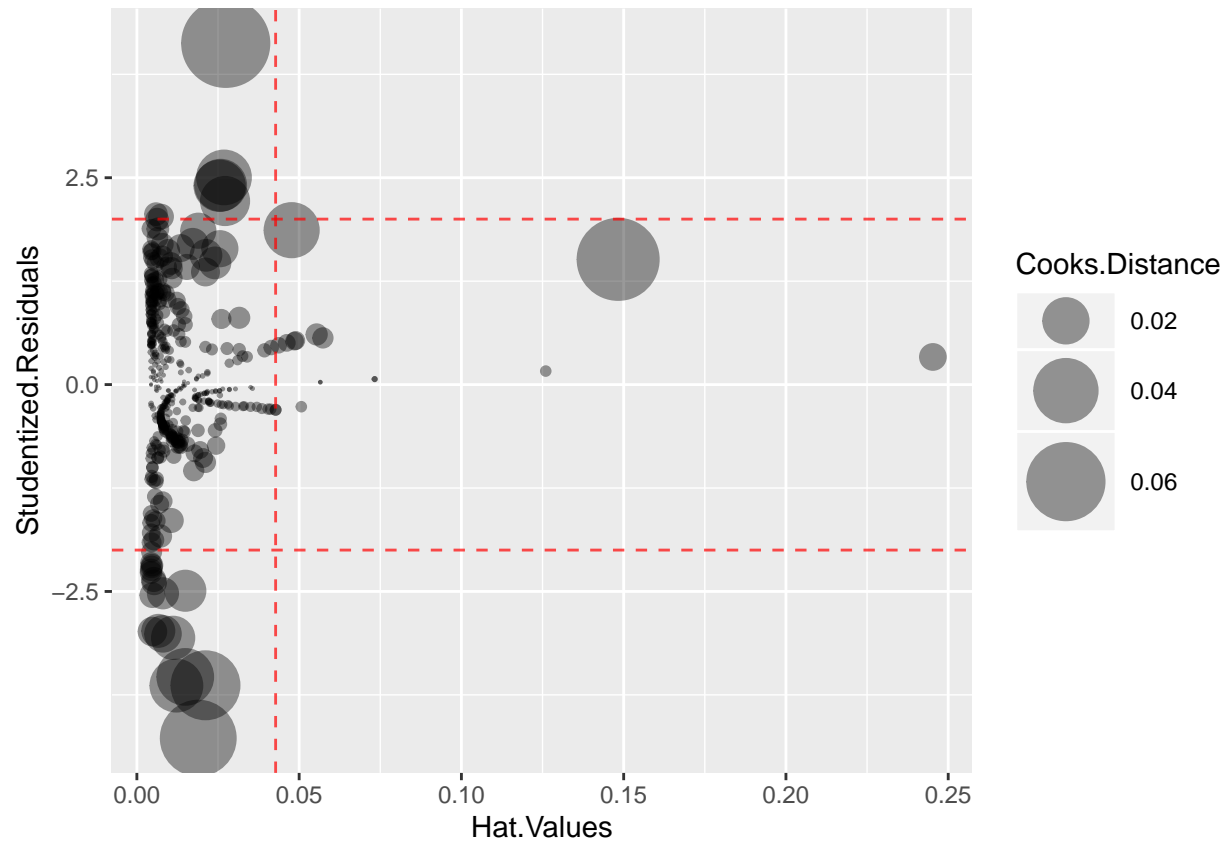
It appears there are **16** influential points within the dataset. I'm not very surprised as players' baseball data is incredibly varied and prone to uniquely performing individuals, thus causing there to be influential observations.

To better grasp the idea behind the information produced above, the following is a plot of the **hat** values representing the leverage with relation to the **studentized residuals**. Each circle represents an observation with it's area proportional to it's calculated Cook's Distance.

Note: The horizontal line represents 3 times the average hat value and the 2 vertical lines mark t-test statistics of -2 and 2.

```
measures = tibble(Hat.Values=hat.vals, Studentized.Residuals=stud.res, Cooks.Distance=cook.dis)

ggplot(aes(x=Hat.Values, y=Studentized.Residuals, size=Cooks.Distance), data=measures) +
  geom_point(alpha=0.4) + scale_size(range=c(0, 15)) +
  geom_vline(xintercept = 3*6/421, color='red', alpha=.7, linetype = "dashed") +
  geom_hline(yintercept = -2, color='red', alpha=.7, linetype = "dashed") +
  geom_hline(yintercept = 2, color='red', alpha=.7, linetype = "dashed")
```



Lastly, I'd like to take a look at these influential points and observe the feature values.

3)

```
all.fit = lm(log.salary ~ ., data = new.baseball)
summary(all.fit)
```

Call:

```
lm(formula = log.salary ~ ., data = new.baseball)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.63762	-0.31246	0.06423	0.31947	1.51943

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	14.2702482	0.5240182	27.232	< 2e-16	***
POSC	-0.0426637	0.1518434	-0.281	0.778886	
POSCF	-0.0268705	0.1494435	-0.180	0.857403	
POSMI	0.0780726	0.1027617	0.760	0.447882	
G.x	0.0039647	0.0046887	0.846	0.398321	
GS	-0.0062069	0.0109901	-0.565	0.572567	
InnOuts	0.0001452	0.0004900	0.296	0.767175	
PO	0.0002185	0.0003197	0.683	0.494768	
A	0.0003503	0.0008990	0.390	0.697023	
E	0.0026430	0.0104617	0.253	0.800687	
DP	-0.0042697	0.0029331	-1.456	0.146306	
G.y	-0.0118007	0.0033350	-3.538	0.000453	***
AB	0.0065267	0.0016068	4.062	5.92e-05	***
R	-0.0093742	0.0063621	-1.473	0.141465	
H	-0.0065391	0.0048988	-1.335	0.182729	
X2B	-0.0063059	0.0070018	-0.901	0.368367	
X3B	0.0006723	0.0184037	0.037	0.970877	
HR	-0.0102874	0.0121474	-0.847	0.397599	
RBI	0.0039993	0.0057956	0.690	0.490580	
SB	-0.0013153	0.0058889	-0.223	0.823382	
CS	-0.0276728	0.0185481	-1.492	0.136549	
BB	0.0056713	0.0039142	1.449	0.148192	
SO	-0.0012819	0.0019363	-0.662	0.508337	
IBB	0.0093704	0.0136217	0.688	0.491937	
HBP	0.0036525	0.0110282	0.331	0.740680	
SH	0.0093553	0.0161609	0.579	0.563012	
SF	0.0014463	0.0204969	0.071	0.943783	
GIDP	-0.0086423	0.0093884	-0.921	0.357889	
CH	0.0021802	0.0010424	2.091	0.037157	*
CHR	0.0076434	0.0038205	2.001	0.046147	*
CR	-0.0022325	0.0016753	-1.333	0.183480	
CRBI	-0.0026771	0.0018308	-1.462	0.144502	
CBB	-0.0011839	0.0005733	-2.065	0.039609	*
AVG	0.5156185	3.3598388	0.153	0.878113	
OBP	-0.0025942	0.0291738	-0.089	0.929190	
CAB.avg	-0.0042031	0.0030973	-1.357	0.175589	
CH.avg	-0.0133749	0.0133111	-1.005	0.315641	
CHR.avg	-0.0709547	0.0361205	-1.964	0.050219	.
CR.avg	0.0534385	0.0151076	3.537	0.000455	***

CRBI.avg	0.0469204	0.0166307	2.821	0.005036	**
neg.cont	1.2416258	0.2315305	5.363	1.43e-07	***
neg.sal	0.0541843	0.1522766	0.356	0.722168	
log.CAB	-0.3434284	0.1713702	-2.004	0.045783	*
log.years	0.2802810	0.2404271	1.166	0.244447	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5807 on 377 degrees of freedom
Multiple R-squared: 0.8082, Adjusted R-squared: 0.7863
F-statistic: 36.95 on 43 and 377 DF, p-value: < 2.2e-16

From former analysis I am aware of collinearity within the explanatory variables and hence inflated standard errors. However, it appears several of the coefficients are statistically significant and that the omnibus F-statistic for all explanatory variables in the model having coefficients of 0 is also statistically significant. In addition, the adjusted R-squared value is significantly lower than R-squared. This all leans toward the argument of possible variable selection that may improve the model's regression ability.

4)

Next, I'll be finding the best **10** models for each model size using forward selection.

```
best10 = regsubsets(log.salary ~ ., data=new.baseball, nvmax=43, nbest=10, method="forward")
summary(best10)
```

Subset selection object

Call: regsubsets.formula(log.salary ~ ., data = new.baseball, nvmax = 43,
nbest = 10, method = "forward")

43 Variables (and intercept)

	Forced in	Forced out
POSC	FALSE	FALSE
POSCF	FALSE	FALSE
POSMI	FALSE	FALSE
G.x	FALSE	FALSE
GS	FALSE	FALSE
InnOuts	FALSE	FALSE
P0	FALSE	FALSE
A	FALSE	FALSE
E	FALSE	FALSE
DP	FALSE	FALSE
G.y	FALSE	FALSE
AB	FALSE	FALSE
R	FALSE	FALSE
H	FALSE	FALSE
X2B	FALSE	FALSE
X3B	FALSE	FALSE
HR	FALSE	FALSE
RBI	FALSE	FALSE
SB	FALSE	FALSE
CS	FALSE	FALSE
BB	FALSE	FALSE
SO	FALSE	FALSE
IBB	FALSE	FALSE
HBP	FALSE	FALSE
SH	FALSE	FALSE
SF	FALSE	FALSE
GIDP	FALSE	FALSE
CH	FALSE	FALSE
CHR	FALSE	FALSE
CR	FALSE	FALSE
CRBI	FALSE	FALSE
CBB	FALSE	FALSE
AVG	FALSE	FALSE
OBP	FALSE	FALSE
CAB.avg	FALSE	FALSE
CH.avg	FALSE	FALSE
CHR.avg	FALSE	FALSE
CR.avg	FALSE	FALSE
CRBI.avg	FALSE	FALSE
neg.cont	FALSE	FALSE
neg.sal	FALSE	FALSE
log.CAB	FALSE	FALSE

```
log.years      FALSE      FALSE
10 subsets of each size up to 43
Selection Algorithm: forward
```

[illegible]

[illegible]

17

[illegible]

[illegible]

[illegible]

23

[illegible]

[illegible]

26	(1)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	*	"	"	*	"	*	"	"	*	"
26	(2)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	*	"	"	"	"	*	"	*	"	*
26	(3)	"	"	*	"	*	"	"	"	*	"	*	"	*	"	"	*	"	"	"	"	*	"	*	"
26	(4)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	*	"	"	"	"	*	"	*	"
26	(5)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	*	"	"	"	"	*	"	*
26	(6)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	*	"	"	"	"	*	"	*
26	(7)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	*	*	"	*	"	"	"	"	*	"
26	(8)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	*	"	"	"	"	*	"	*
26	(9)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	"
26	(10)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	"
27	(1)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	*	"	"	"	"	*	"	*
27	(2)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	"
27	(3)	"	"	*	"	*	"	"	"	*	"	*	"	*	"	"	"	"	*	"	"	"	"	*	"
27	(4)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	"
27	(5)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	"
27	(6)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	"
27	(7)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	"
27	(8)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	*	*	"	*	"	"	"	*	"
27	(9)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	"
27	(10)	"	"	*	"	*	"	*	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	"
28	(1)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	*
28	(2)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	*
28	(3)	"	"	*	"	*	"	"	"	*	"	*	"	*	"	"	"	"	*	"	"	"	"	*	*
28	(4)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	*
28	(5)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	*
28	(6)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	*
28	(7)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	*
28	(8)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	*
28	(9)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	*	*	"	*	"	"	"	"	*	*
28	(10)	"	"	*	"	*	"	*	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	*
29	(1)	"	"	*	"	*	"	"	"	*	"	*	"	"	"	"	"	"	*	"	"	"	"	*	*

[illegible]

36	(9)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*
37	(1)	"	"	"*	"*	"	"	"*	"*	"*	"*	"*	"	"	"*	"*	"*
37	(2)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
37	(3)	"	"	"*	"*	"	"	"*	"*	"*	"*	"*	"	"	"*	"*	"*
37	(4)	"	"	"*	"*	"	"	"*	"*	"*	"*	"*	"	"	"*	"*	"*
37	(5)	"	"	"*	"*	"	"	"*	"*	"*	"*	"*	"	"	"*	"*	"*
37	(6)	"	"	"*	"*	"	"	"*	"*	"*	"*	"*	"	"	"*	"*	"*
37	(7)	"*	"*	"*	"	"	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
37	(8)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
38	(1)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
38	(2)	"	"	"*	"*	"	"	"*	"*	"*	"*	"	"	"*	"*	"*	"*
38	(3)	"	"	"*	"*	"	"	"*	"*	"*	"*	"	"	"*	"*	"*	"*
38	(4)	"	"	"*	"*	"	"	"*	"*	"*	"*	"	"	"*	"*	"*	"*
38	(5)	"	"	"*	"*	"	"	"*	"*	"*	"*	"	"	"*	"*	"*	"*
38	(6)	"*	"*	"*	"	"	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
38	(7)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
39	(1)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
39	(2)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
39	(3)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
39	(4)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
39	(5)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
39	(6)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
40	(1)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
40	(2)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
40	(3)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
40	(4)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
40	(5)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
41	(1)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
41	(2)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
41	(3)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
41	(4)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
42	(1)	"	"	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
42	(2)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
42	(3)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
43	(1)	"*	"*	"*	"*	"*	"*	"*	"*	"*	"*	"	"	"*	"*	"*	"*
		CBB	AVG	OBP	CAB.avg	CH.avg	CHR.avg	CR.avg	CRBI.avg	neg.cont	neg.sal						
1	(1)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
1	(2)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
1	(3)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
1	(4)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
1	(5)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
1	(6)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
1	(7)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
1	(8)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
1	(9)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
1	(10)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
2	(1)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
2	(2)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
2	(3)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
2	(4)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
2	(5)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
2	(6)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
2	(7)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"
2	(8)	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"

[illegible]

[illegible]

13	(7)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
13	(8)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
13	(9)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
13	(10)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(1)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(2)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(3)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(4)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(5)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(6)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(7)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(8)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(9)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
14	(10)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(1)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(2)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(3)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(4)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(5)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(6)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(7)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(8)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(9)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
15	(10)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
16	(1)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
16	(2)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
16	(3)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
16	(4)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
16	(5)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
16	(6)	"*	"	"	"	"	*	"	"	"	*	"	"	"	"	*	"	"
16	(7)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
16	(8)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
16	(9)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
16	(10)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
17	(1)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
17	(2)	"*	"	"	"	"	*	"	"	"	*	"	"	"	"	*	"	"
17	(3)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
17	(4)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
17	(5)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
17	(6)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
17	(7)	"*	"	"	"	"	*	"										

33

24	(5)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
24	(6)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
24	(7)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
24	(8)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
24	(9)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
24	(10)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(1)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(2)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(3)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(4)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(5)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(6)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(7)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(8)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(9)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
25	(10)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
26	(1)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
26	(2)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
26	(3)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
26	(4)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
26	(5)	"*	"	"	"	"	*	"	"	*	"	*	"	"	"	*	"	"
26	(6)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
26	(7)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
26	(8)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
26	(9)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
26	(10)	"*	"	*	"	"	*	"	"	"	"	*	"	"	"	*	"	"
27	(1)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
27	(2)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
27	(3)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
27	(4)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
27	(5)	"*	"	"	"	"	*	"	"	*	"	*	"	"	"	*	"	"
27	(6)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
27	(7)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
27	(8)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
27	(9)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
27	(10)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
28	(1)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
28	(2)	"*	"	"	"	"	*	"	"	*	"	*	"	"	"	*	"	"
28	(3)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
28	(4)	"*	"	"	"	"	*	"	"	"	"	*	"	"	"	*	"	"
28	(5)	"*	"	"	"	"	*	"	"</									

[illegible]

```

35 ( 3 ) "*" " " " " "*" "*" "*" "*" "*" " "
35 ( 4 ) "*" " " " " "*" "*" "*" "*" "*" " "
35 ( 5 ) "*" " " " " "*" "*" "*" "*" "*" " "
35 ( 6 ) "*" "*" " " "*" "*" "*" "*" "*" " "
35 ( 7 ) "*" " " " " "*" "*" "*" "*" "*" " "
35 ( 8 ) "*" " " "*" "*" "*" "*" "*" "*" "*" " "
35 ( 9 ) "*" " " " " "*" "*" "*" "*" "*" " "
35 ( 10 ) "*" "*" "*" "*" " " " " " " " " " "
36 ( 1 ) "*" " " " " "*" "*" "*" "*" "*" "*"
36 ( 2 ) "*" " " " " "*" "*" "*" "*" "*" "*"
36 ( 3 ) "*" " " " " "*" "*" "*" "*" "*" "*"
36 ( 4 ) "*" " " " " "*" "*" "*" "*" "*" "*"
36 ( 5 ) "*" "*" " " "*" "*" "*" "*" "*" "*"
36 ( 6 ) "*" " " " " "*" "*" "*" "*" "*" "*"
36 ( 7 ) "*" " " "*" "*" "*" "*" "*" "*" "*" "*"
36 ( 8 ) "*" " " " " "*" "*" "*" "*" "*" "*"
36 ( 9 ) "*" "*" "*" "*" "*" " " " " " " " "
37 ( 1 ) "*" " " " " "*" "*" "*" "*" "*" "*"
37 ( 2 ) "*" " " " " "*" "*" "*" "*" "*" "*"
37 ( 3 ) "*" "*" " " "*" "*" "*" "*" "*" "*"
37 ( 4 ) "*" " " " " "*" "*" "*" "*" "*" "*"
37 ( 5 ) "*" " " "*" "*" "*" "*" "*" "*" "*" "*"
37 ( 6 ) "*" " " " " "*" "*" "*" "*" "*" "*"
37 ( 7 ) "*" " " " " "*" "*" "*" "*" "*" "*"
37 ( 8 ) "*" "*" "*" "*" "*" "*" " " " " " "
38 ( 1 ) "*" " " " " "*" "*" "*" "*" "*" "*"
38 ( 2 ) "*" " " " " "*" "*" "*" "*" "*" "*"
38 ( 3 ) "*" "*" " " "*" "*" "*" "*" "*" "*"
38 ( 4 ) "*" " " " " "*" "*" "*" "*" "*" "*"
38 ( 5 ) "*" " " "*" "*" "*" "*" "*" "*" "*" "*"
38 ( 6 ) "*" " " " " "*" "*" "*" "*" "*" "*"
38 ( 7 ) "*" "*" "*" "*" "*" "*" " " " " " "
39 ( 1 ) "*" " " " " "*" "*" "*" "*" "*" "*"
39 ( 2 ) "*" "*" " " "*" "*" "*" "*" "*" "*"
39 ( 3 ) "*" " " " " "*" "*" "*" "*" "*" "*"
39 ( 4 ) "*" " " "*" "*" "*" "*" "*" "*" "*" "*"
39 ( 5 ) "*" " " " " "*" "*" "*" "*" "*" "*"
39 ( 6 ) "*" "*" "*" "*" "*" "*" " " " " " "
40 ( 1 ) "*" "*" " " "*" "*" "*" "*" "*" "*"
40 ( 2 ) "*" " " " " "*" "*" "*" "*" "*" "*"
40 ( 3 ) "*" " " "*" "*" "*" "*" "*" "*" "*" "*"
40 ( 4 ) "*" " " " " "*" "*" "*" "*" "*" "*"
40 ( 5 ) "*" "*" "*" "*" "*" "*" "*" "*" " "
41 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
41 ( 2 ) "*" "*" " " "*" "*" "*" "*" "*" "*"
41 ( 3 ) "*" "*" " " "*" "*" "*" "*" "*" "*"
41 ( 4 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
42 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
42 ( 2 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
42 ( 3 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
43 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"

log.CAB log.years
1 ( 1 ) " " " "
1 ( 2 ) " " " "

```

1	(3)	" "	" "
1	(4)	"*"	" "
1	(5)	" "	" "
1	(6)	" "	" "
1	(7)	" "	" "
1	(8)	" "	" "
1	(9)	" "	" "
1	(10)	" "	" "
2	(1)	" "	" "
2	(2)	" "	" "
2	(3)	" "	" "
2	(4)	" "	" "
2	(5)	" "	" "
2	(6)	" "	" "
2	(7)	" "	" "
2	(8)	" "	" "
2	(9)	" "	" "
2	(10)	" "	" "
3	(1)	" "	" "
3	(2)	" "	" "
3	(3)	" "	" "
3	(4)	" "	" "
3	(5)	" "	" "
3	(6)	" "	" "
3	(7)	" "	" "
3	(8)	" "	" "
3	(9)	" "	" "
3	(10)	" "	" "
4	(1)	" "	" "
4	(2)	" "	" "
4	(3)	" "	" "
4	(4)	" "	" "
4	(5)	" "	" "
4	(6)	" "	" "
4	(7)	" "	" "
4	(8)	" "	" "
4	(9)	" "	" "
4	(10)	" "	" "
5	(1)	"*"	" "
5	(2)	" "	" "
5	(3)	" "	" "
5	(4)	" "	" "
5	(5)	" "	" "
5	(6)	" "	"*"
5	(7)	" "	" "
5	(8)	" "	" "
5	(9)	" "	" "
5	(10)	" "	" "
6	(1)	"*"	" "
6	(2)	"*"	" "
6	(3)	"*"	" "
6	(4)	"*"	" "
6	(5)	"*"	" "
6	(6)	"*"	" "

6	(7)	"*"	" "
6	(8)	"*"	" "
6	(9)	"*"	" "
6	(10)	"*"	" "
7	(1)	"*"	"*"
7	(2)	"*"	" "
7	(3)	"*"	" "
7	(4)	"*"	" "
7	(5)	"*"	" "
7	(6)	"*"	" "
7	(7)	"*"	" "
7	(8)	"*"	" "
7	(9)	"*"	" "
7	(10)	"*"	" "
8	(1)	"*"	"*"
8	(2)	"*"	"*"
8	(3)	"*"	"*"
8	(4)	"*"	"*"
8	(5)	"*"	"*"
8	(6)	"*"	"*"
8	(7)	"*"	"*"
8	(8)	"*"	"*"
8	(9)	"*"	"*"
8	(10)	"*"	"*"
9	(1)	"*"	"*"
9	(2)	"*"	"*"
9	(3)	"*"	"*"
9	(4)	"*"	"*"
9	(5)	"*"	"*"
9	(6)	"*"	"*"
9	(7)	"*"	"*"
9	(8)	"*"	"*"
9	(9)	"*"	"*"
9	(10)	"*"	"*"
10	(1)	"*"	"*"
10	(2)	"*"	"*"
10	(3)	"*"	"*"
10	(4)	"*"	"*"
10	(5)	"*"	"*"
10	(6)	"*"	"*"
10	(7)	"*"	"*"
10	(8)	"*"	"*"
10	(9)	"*"	"*"
10	(10)	"*"	"*"
11	(1)	"*"	"*"
11	(2)	"*"	"*"
11	(3)	"*"	"*"
11	(4)	"*"	"*"
11	(5)	"*"	"*"
11	(6)	"*"	"*"
11	(7)	"*"	"*"
11	(8)	"*"	"*"
11	(9)	"*"	"*"
11	(10)	"*"	"*"

12	(1)	"*"	"*"
12	(2)	"*"	"*"
12	(3)	"*"	"*"
12	(4)	"*"	"*"
12	(5)	"*"	"*"
12	(6)	"*"	"*"
12	(7)	"*"	"*"
12	(8)	"*"	"*"
12	(9)	"*"	"*"
12	(10)	"*"	"*"
13	(1)	"*"	"*"
13	(2)	"*"	"*"
13	(3)	"*"	"*"
13	(4)	"*"	"*"
13	(5)	"*"	"*"
13	(6)	"*"	"*"
13	(7)	"*"	"*"
13	(8)	"*"	"*"
13	(9)	"*"	"*"
13	(10)	"*"	"*"
14	(1)	"*"	"*"
14	(2)	"*"	"*"
14	(3)	"*"	"*"
14	(4)	"*"	"*"
14	(5)	"*"	"*"
14	(6)	"*"	"*"
14	(7)	"*"	"*"
14	(8)	"*"	"*"
14	(9)	"*"	"*"
14	(10)	"*"	"*"
15	(1)	"*"	"*"
15	(2)	"*"	"*"
15	(3)	"*"	"*"
15	(4)	"*"	"*"
15	(5)	"*"	"*"
15	(6)	"*"	"*"
15	(7)	"*"	"*"
15	(8)	"*"	"*"
15	(9)	"*"	"*"
15	(10)	"*"	"*"
16	(1)	"*"	"*"
16	(2)	"*"	"*"
16	(3)	"*"	"*"
16	(4)	"*"	"*"
16	(5)	"*"	"*"
16	(6)	"*"	"*"
16	(7)	"*"	"*"
16	(8)	"*"	"*"
16	(9)	"*"	"*"
16	(10)	"*"	"*"
17	(1)	"*"	"*"
17	(2)	"*"	"*"
17	(3)	"*"	"*"
17	(4)	"*"	"*"

17	(5)	"*"	"*"
17	(6)	"*"	"*"
17	(7)	"*"	"*"
17	(8)	"*"	"*"
17	(9)	"*"	"*"
17	(10)	"*"	"*"
18	(1)	"*"	"*"
18	(2)	"*"	"*"
18	(3)	"*"	"*"
18	(4)	"*"	"*"
18	(5)	"*"	"*"
18	(6)	"*"	"*"
18	(7)	"*"	"*"
18	(8)	"*"	"*"
18	(9)	"*"	"*"
18	(10)	"*"	"*"
19	(1)	"*"	"*"
19	(2)	"*"	"*"
19	(3)	"*"	"*"
19	(4)	"*"	"*"
19	(5)	"*"	"*"
19	(6)	"*"	"*"
19	(7)	"*"	"*"
19	(8)	"*"	"*"
19	(9)	"*"	"*"
19	(10)	"*"	"*"
20	(1)	"*"	"*"
20	(2)	"*"	"*"
20	(3)	"*"	"*"
20	(4)	"*"	"*"
20	(5)	"*"	"*"
20	(6)	"*"	"*"
20	(7)	"*"	"*"
20	(8)	"*"	"*"
20	(9)	"*"	"*"
20	(10)	"*"	"*"
21	(1)	"*"	"*"
21	(2)	"*"	"*"
21	(3)	"*"	"*"
21	(4)	"*"	"*"
21	(5)	"*"	"*"
21	(6)	"*"	"*"
21	(7)	"*"	"*"
21	(8)	"*"	"*"
21	(9)	"*"	"*"
21	(10)	"*"	"*"
22	(1)	"*"	"*"
22	(2)	"*"	"*"
22	(3)	"*"	"*"
22	(4)	"*"	"*"
22	(5)	"*"	"*"
22	(6)	"*"	"*"
22	(7)	"*"	"*"
22	(8)	"*"	"*"

22	(9)	"*"	"*"
22	(10)	"*"	"*"
23	(1)	"*"	"*"
23	(2)	"*"	"*"
23	(3)	"*"	"*"
23	(4)	"*"	"*"
23	(5)	"*"	"*"
23	(6)	"*"	"*"
23	(7)	"*"	"*"
23	(8)	"*"	"*"
23	(9)	"*"	"*"
23	(10)	"*"	"*"
24	(1)	"*"	"*"
24	(2)	"*"	"*"
24	(3)	"*"	"*"
24	(4)	"*"	"*"
24	(5)	"*"	"*"
24	(6)	"*"	"*"
24	(7)	"*"	"*"
24	(8)	"*"	"*"
24	(9)	"*"	"*"
24	(10)	"*"	"*"
25	(1)	"*"	"*"
25	(2)	"*"	"*"
25	(3)	"*"	"*"
25	(4)	"*"	"*"
25	(5)	"*"	"*"
25	(6)	"*"	"*"
25	(7)	"*"	"*"
25	(8)	"*"	"*"
25	(9)	"*"	"*"
25	(10)	"*"	"*"
26	(1)	"*"	"*"
26	(2)	"*"	"*"
26	(3)	"*"	"*"
26	(4)	"*"	"*"
26	(5)	"*"	"*"
26	(6)	"*"	"*"
26	(7)	"*"	"*"
26	(8)	"*"	"*"
26	(9)	"*"	"*"
26	(10)	"*"	"*"
27	(1)	"*"	"*"
27	(2)	"*"	"*"
27	(3)	"*"	"*"
27	(4)	"*"	"*"
27	(5)	"*"	"*"
27	(6)	"*"	"*"
27	(7)	"*"	"*"
27	(8)	"*"	"*"
27	(9)	"*"	"*"
27	(10)	"*"	"*"
28	(1)	"*"	"*"
28	(2)	"*"	"*"

28	(3)	"*"	"*"
28	(4)	"*"	"*"
28	(5)	"*"	"*"
28	(6)	"*"	"*"
28	(7)	"*"	"*"
28	(8)	"*"	"*"
28	(9)	"*"	"*"
28	(10)	"*"	"*"
29	(1)	"*"	"*"
29	(2)	"*"	"*"
29	(3)	"*"	"*"
29	(4)	"*"	"*"
29	(5)	"*"	"*"
29	(6)	"*"	"*"
29	(7)	"*"	"*"
29	(8)	"*"	"*"
29	(9)	"*"	"*"
29	(10)	"*"	"*"
30	(1)	"*"	"*"
30	(2)	"*"	"*"
30	(3)	"*"	"*"
30	(4)	"*"	"*"
30	(5)	"*"	"*"
30	(6)	"*"	"*"
30	(7)	"*"	"*"
30	(8)	"*"	"*"
30	(9)	"*"	"*"
30	(10)	"*"	"*"
31	(1)	"*"	"*"
31	(2)	"*"	"*"
31	(3)	"*"	"*"
31	(4)	"*"	"*"
31	(5)	"*"	"*"
31	(6)	"*"	"*"
31	(7)	"*"	"*"
31	(8)	"*"	"*"
31	(9)	"*"	"*"
31	(10)	"*"	"*"
32	(1)	"*"	"*"
32	(2)	"*"	"*"
32	(3)	"*"	"*"
32	(4)	"*"	"*"
32	(5)	"*"	"*"
32	(6)	"*"	"*"
32	(7)	"*"	"*"
32	(8)	"*"	"*"
32	(9)	"*"	"*"
32	(10)	"*"	"*"
33	(1)	"*"	"*"
33	(2)	"*"	"*"
33	(3)	"*"	"*"
33	(4)	"*"	"*"
33	(5)	"*"	"*"
33	(6)	"*"	"*"

33	(7)	"*"	"*"
33	(8)	"*"	"*"
33	(9)	"*"	"*"
33	(10)	"*"	"*"
34	(1)	"*"	"*"
34	(2)	"*"	"*"
34	(3)	"*"	"*"
34	(4)	"*"	"*"
34	(5)	"*"	"*"
34	(6)	"*"	"*"
34	(7)	"*"	"*"
34	(8)	"*"	"*"
34	(9)	"*"	"*"
34	(10)	"*"	"*"
35	(1)	"*"	"*"
35	(2)	"*"	"*"
35	(3)	"*"	"*"
35	(4)	"*"	"*"
35	(5)	"*"	"*"
35	(6)	"*"	"*"
35	(7)	"*"	"*"
35	(8)	"*"	"*"
35	(9)	"*"	"*"
35	(10)	" "	" "
36	(1)	"*"	"*"
36	(2)	"*"	"*"
36	(3)	"*"	"*"
36	(4)	"*"	"*"
36	(5)	"*"	"*"
36	(6)	"*"	"*"
36	(7)	"*"	"*"
36	(8)	"*"	"*"
36	(9)	" "	" "
37	(1)	"*"	"*"
37	(2)	"*"	"*"
37	(3)	"*"	"*"
37	(4)	"*"	"*"
37	(5)	"*"	"*"
37	(6)	"*"	"*"
37	(7)	"*"	"*"
37	(8)	" "	" "
38	(1)	"*"	"*"
38	(2)	"*"	"*"
38	(3)	"*"	"*"
38	(4)	"*"	"*"
38	(5)	"*"	"*"
38	(6)	"*"	"*"
38	(7)	" "	" "
39	(1)	"*"	"*"
39	(2)	"*"	"*"
39	(3)	"*"	"*"
39	(4)	"*"	"*"
39	(5)	"*"	"*"
39	(6)	" "	" "

40	(1)	"*"	"*"
40	(2)	"*"	"*"
40	(3)	"*"	"*"
40	(4)	"*"	"*"
40	(5)	" "	" "
41	(1)	"*"	"*"
41	(2)	"*"	"*"
41	(3)	"*"	"*"
41	(4)	" "	" "
42	(1)	"*"	"*"
42	(2)	"*"	"*"
42	(3)	"*"	" "
43	(1)	"*"	"*"

5)

```
lowest5.bic = sort(summary(best10)$bic)[1:5]
top5.bic.loc = order(summary(best10)$bic)[1:5]
top5.models = summary(best10)$which[top5.bic.loc,]
models = c()

for(i in 1:nrow(top5.models)){
  cat(sprintf("Model %d:", i))
  cat(sprintf("\nBIC: %f", lowest5.bic[i]))

  models = c(models, list(names(which(top5.models[i,]))[-1]))
  cat(sprintf("\nFeature: %s", models[[i]]))
  cat("\n\n")
}
```

Model 1:
BIC: -586.225215
Feature: GS
Feature: G.y
Feature: AB
Feature: R
Feature: BB
Feature: CH
Feature: CBB
Feature: CAB.avg
Feature: CR.avg
Feature: CRBI.avg
Feature: neg.cont
Feature: log.CAB
Feature: log.years

Model 2:
BIC: -583.405937
Feature: GS
Feature: R
Feature: CH
Feature: CBB
Feature: CAB.avg
Feature: CR.avg
Feature: CRBI.avg
Feature: neg.cont
Feature: log.CAB
Feature: log.years

Model 3:
BIC: -582.502074
Feature: G.x
Feature: GS
Feature: G.y
Feature: AB
Feature: R
Feature: BB

Feature: CH
Feature: CBB
Feature: CAB.avg
Feature: CR.avg
Feature: CRBI.avg
Feature: neg.cont
Feature: log.CAB
Feature: log.years

Model 4:
BIC: -582.402615
Feature: GS
Feature: G.y
Feature: AB
Feature: R
Feature: X2B
Feature: BB
Feature: CH
Feature: CBB
Feature: CAB.avg
Feature: CR.avg
Feature: CRBI.avg
Feature: neg.cont
Feature: log.CAB
Feature: log.years

Model 5:
BIC: -582.248168
Feature: GS
Feature: G.y
Feature: AB
Feature: R
Feature: H
Feature: BB
Feature: CH
Feature: CBB
Feature: CAB.avg
Feature: CR.avg
Feature: CRBI.avg
Feature: neg.cont
Feature: log.CAB
Feature: log.years

6)

Using BIC as a criterion may not be sufficient and hence I will use 10-fold cross validation to re-rank the 5 models determined above.

```
paste(models[[1]], collapse=" + ")
```

```
[1] "GS + G.y + AB + R + BB + CH + CBB + CAB.avg + CR.avg + CRBI.avg + neg.cont + log.CAB + log.years"
```

7)

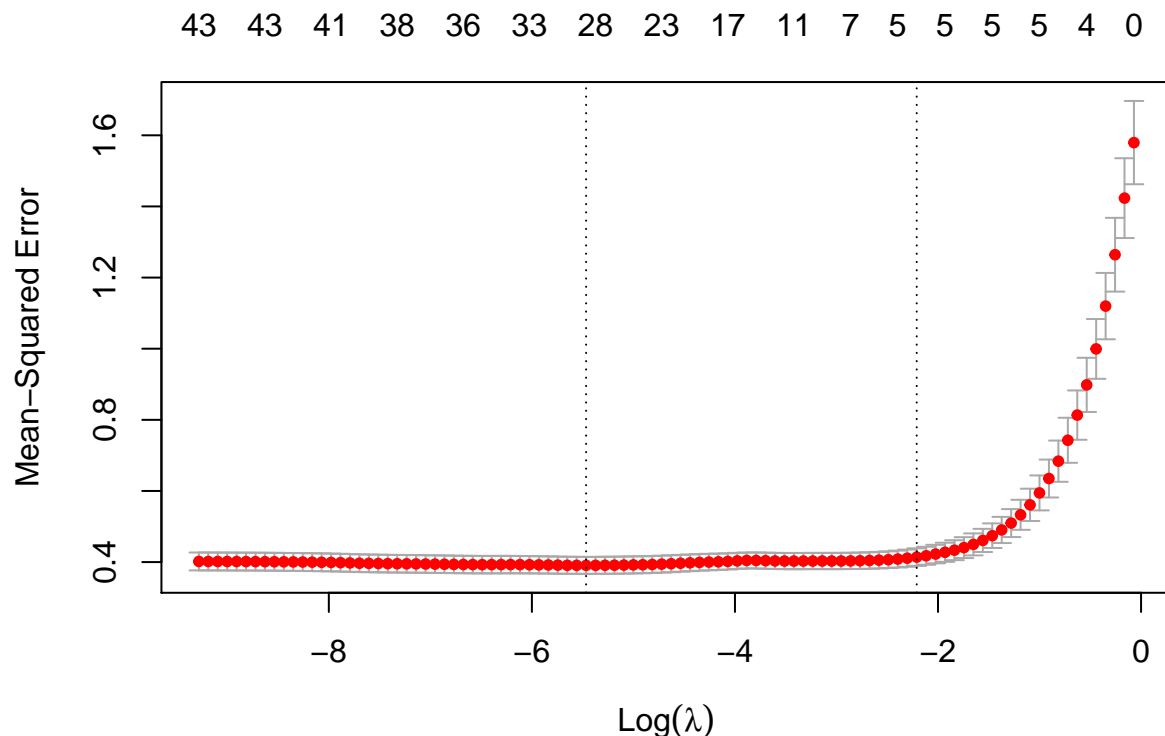
In determining the value of the hyperparameter `lambda` to use for the LASSO model (original feature variables), I will be using 10-fold cross validation with an average MSE for the criterion.

```
X = new.baseball %>% dplyr::select(-c(log.salary)) %>% as.matrix()
y = new.baseball$log.salary

lasso.model = glmnet(x = X, y = y, alpha = 1)
cv.lasso.model = cv.glmnet(x = X, y = y, alpha = 1, nfolds = 10)
```

The following is the plot of the log of the hyperparameter `lambda` with respect to the resulting average MSE.

```
plot(cv.lasso.model)
```



The left-most dashed vertical line in the plot represents the `lambda` corresponding to the minimum MSE whereas the right-most dashed vertical line represents the `lambda` 1 standard error away from the min `lambda`.

I believe the data intrinsically contains a lot of collinear explanatory variables and would like to reduce the number of features used in model prediction. I believe the `lambda` corresponding to the minimum MSE would be suitable as it not only performs variable selection through utilizing just 28 feature variables for regression rather than all 43, it also has the lowest MSE out of all other possible `lambda` hyperparameter values and thus seems to be the best choice in prediction.

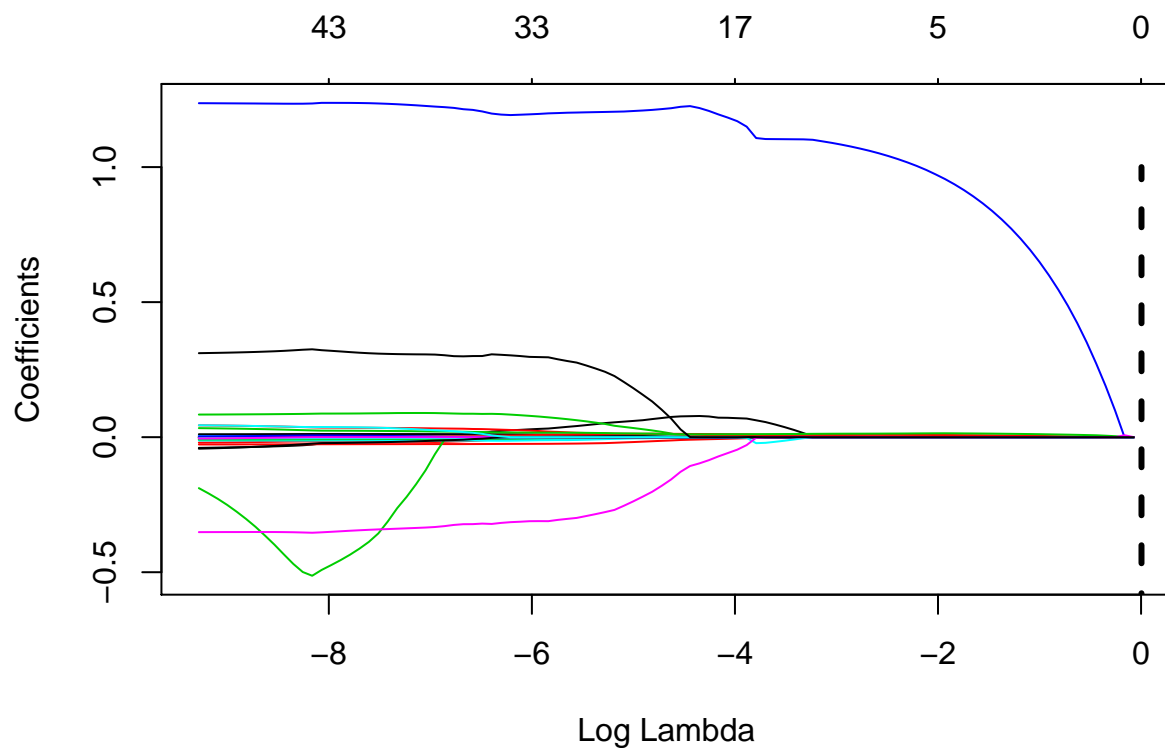
The value of `lambda` I will use for the LASSO model is the following:


```
best.lasso.lambda = cv.lasso.model$lambda.min
best.lasso.lambda
```

```
[1] 0.004228187
```

The following is a visualization of how different values of lambda influence the coefficient values of the regression model:

```
plot(lasso.model, xvar = "lambda", )
lines(c(best.lasso.lambda, best.lasso.lambda), c(-1, 1), lty = "dashed", lwd = 3)
```



Lastly, using the hyperparameter value of lambda specified above, I'd like to analyze which coefficients are retained through implicit variable selection of LASSO.

```
best.lasso.coefs = predict(lasso.model, type = 'coefficients', s = best.lasso.lambda)
best.lasso.coefs
```

```
44 x 1 sparse Matrix of class "dgCMatrix"
```

```
      1
(Intercept) 1.395376e+01
POSC        4.464352e-02
POSCF       .
POSMI       5.860071e-02
G.x         9.792966e-04
```

GS	.
InnOuts	7.228064e-05
PO	3.176194e-05
A	.
E	.
DP	-1.580214e-03
G.y	-6.484793e-03
AB	1.509140e-03
R	.
H	.
X2B	-3.628000e-03
X3B	.
HR	-9.471769e-03
RBI	1.628877e-03
SB	-4.844934e-04
CS	-2.316805e-02
BB	4.166142e-03
SO	.
IBB	8.517895e-03
HBP	1.603347e-03
SH	7.769733e-03
SF	1.004775e-02
GIDP	.
CH	2.917989e-05
CHR	9.569908e-04
CR	8.298561e-04
CRBI	.
CBB	-1.065447e-03
AVG	.
OBP	.
CAB.avg	-5.338189e-04
CH.avg	.
CHR.avg	.
CR.avg	1.543482e-02
CRBI.avg	1.603380e-02
neg.cont	1.202804e+00
neg.sal	.
log.CAB	-2.921016e-01
log.years	2.656102e-01