Project Draft

Tom Bash and Conor Keating

4/20/2021

Abstract

Title: The Relationship Between Baseball Statistics and First Year Arbitration Salary

This study will attempt to find a relationship between a Major League Baseball player's hitting and fielding statistics and their first year arbitration salary. We will be constructing a multiple linear regression model to find which statistics are the best and most accurate when it comes to predicting a baseball player's salary. Upon conclusion of the study, we will have a better understanding of what the most influential variables are for predicting MLB salaries and what part of a player's game they should focus on to make the most money.

Via BaseballReference, Spotrac, MLB.com, FanGraphs, and MLBTradeRumors.

Section 1 – Introduction Research Question: What are the most important and influential variables when trying to predict a Major League Baseball position players' first year arbitration salary.

Important details to understand:

Players receive Major League service time for each day spent on the 26-man Major League roster. Service time is used to determine when players are eligible for salary arbitration. Each Major League regular season consists of 187 days and each day spent on the active roster or injured list earns a player one day of service time. A player is deemed to have reached one year of Major League service upon accruing 172 days in a given year.

All players with between three and six years of Major League service time become eligible for salary arbitration. They can earn substantial raises relative to the Major League minimum salary. Additionally, Major League Baseball each year identifies the group of players that ended the prior season with between two and three years of Major League service and at least 86 days of Major League service in that season and designates the top 22 percent – in terms of service time – as arbitration eligible. Those in the top 22 percent, called Super Two Players, are also eligible for salary arbitration despite having less than three years of Major League service. If the club and player have not agreed on a salary by a deadline in the middle of January, the club and player must exchange salary figures for the upcoming season. After the figures are exchanged, a hearing is scheduled in February. If no settlement can be reached by the hearing date, the case is brought to a panel of arbitrators. After hearing arguments from both sides, the panel selects the salary figure of either the player or the club, not one in between, as the player's salary for the upcoming season.

In order to create an accurate model, we will just be trying to predict salary for first year arbitration eligible players. Multiple linear regressor

Our data is collected from a combination of different websites including BaseballReference, Spotrac, MLB.com, and MLBTradeRumors.

The goal is to be able to build a model in order to calculate what a player's salary in the future should be.

Variables Identified:

Batting Average, On Base Percentage, Home Runs, Runs Batted In, Runs Scored, Slugging Percentage, On Base + Slugging Percentage, Stolen Bases, Plate Appearances, Strikeout Percentage, Walk Percentage, and Defensive Runs Saved (put on excel)

```r
library(ggplot2)
```

```r
baseball = read.csv('DS Proposal - Sheet1.csv')
baseball
```

```
##                Player Salary First.Arbitration.Year    BA   OBP  HR RBI   R
## 1          Jose Abreu  10.83                   2017 0.299 0.360  91 308 235
## 2     George Springer   3.90                   2017 0.258 0.356  65 174 220
## 3     Cesar Hernandez   2.55                   2017 0.281 0.350   8  88 154
## 4     Tuffy Gosewisch   0.64                   2017 0.199 0.237   5  30  24
## 5      Derek Dietrich   1.70                   2017 0.251 0.338  31 106 140
## 6  Jackie Bradley Jr.   3.60                   2017 0.237 0.316  40 170 200
## 7          Sandy Leon   1.30                   2017 0.254 0.319   8  43  53
## 8        Caleb Joseph   0.70                   2017 0.213 0.271  20  77  67
## 9      Jake Marisnick   1.10                   2017 0.225 0.268  18  81 113
## 10        Jesus Sucre   0.63                   2017 0.209 0.246   2  20  18
## 11        Tim Beckham   0.89                   2017 0.238 0.288  14  54  50
## 12     Ehire Adrianza   0.60                   2017 0.220 0.292   3  26  27
## 13    Kevin Kiermaier   2.98                   2017 0.258 0.313  32 112 152
## 14        Kris Bryant  10.85                   2018 0.288 0.388  94 274 319
## 15      Maikel Franco   2.95                   2018 0.247 0.300  63 219 183
## 16           Ryan Rua   0.87                   2018 0.246 0.305  17  55  78
## 17    Addison Russell   3.20                   2018 0.240 0.312  46 192 179
## 18     Yolmer Sanchez   2.35                   2018 0.242 0.286  21 116 124
## 19        Matt Szczur   0.95                   2018 0.237 0.318  11  55  69
## 20       Devon Travis   1.45                   2018 0.292 0.331  24 109 114
## 21       Byron Buxton   1.75                   2019 0.237 0.292  38 145 185
## 22        Curt Casali   0.95                   2019 0.223 0.302  23  65  63
## 23      Brandon Drury   1.30                   2019 0.264 0.314  32 134 108
## 24      Austin Hedges   2.06                   2019 0.210 0.258  35 104  80
## 25   Travis Jankowski   1.17                   2019 0.242 0.319   8  42 117
## 26         Max Kepler   3.13                   2019 0.233 0.313  56 190 199
## 27       Nomar Mazara   3.30                   2019 0.258 0.320  60 242 184
## 28        Jose Peraza   2.78                   2019 0.282 0.319  22 121 163
## 29     Kevin Plawecki   1.14                   2019 0.218 0.308  14  75  68
## 30       Trevor Story   5.00                   2019 0.268 0.333  88 262 223
## 31      Blake Swihart   0.91                   2019 0.256 0.314   8  54  85
## 32        Trea Turner   3.73                   2019 0.289 0.346  44 159 236
## 33       Tony Wolters   0.96                   2019 0.226 0.322   6  73  76
## 34     Cody Bellinger  11.50                   2020 0.278 0.368 111 288 292
## 35      Johan Camargo   1.70                   2020 0.269 0.328  30 135 124
## 36         David Dahl   2.48                   2020 0.297 0.346  38 133 140
## 37       JaCoby Jones   1.58                   2020 0.211 0.276  25  75 110
## 38       Andrew Knapp   0.71                   2020 0.223 0.327   9  36  57
## 39     Hunter Renfroe   3.30                   2020 0.235 0.294  89 204 176
## 40   Daniel Robertson   1.03                   2020 0.231 0.340  16  72  91
## 41   Giovanny Urshela   2.48                   2020 0.269 0.313  29 113 119
## 42       J.P. Crawford   2.05                   2021 0.231 0.325  12  88 101
## 43          J.D. Davis   2.10                   2021 0.268 0.346  33  88 108
## 44      Clint Frazier   2.10                   2021 0.258 0.331  24  82  80
## 45       Carson Kelly   1.70                   2021 0.221 0.305  23  76  64
## 46 Isiah Kiner-Falefa   2.00                   2021 0.260 0.319   8  65  94
## 47  Anthony Santander   2.10                   2021 0.252 0.292  32  99  79
```

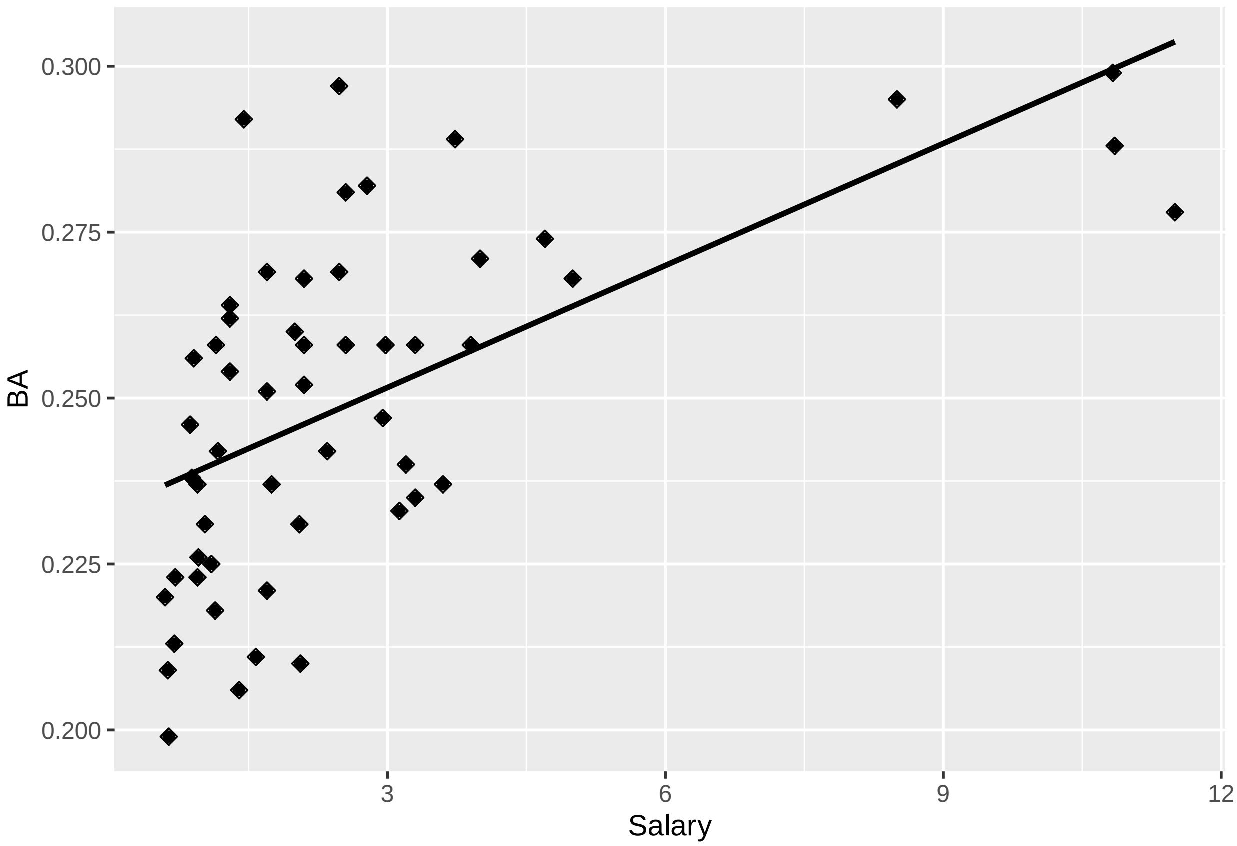
```
## 48      Austin Slater   1.15                 2021 0.258 0.346  14  67  74
## 49      Dominic Smith   2.55                 2021 0.258 0.317  35 104  93
## 50          Juan Soto   8.50                 2021 0.295 0.415  69 217 226
## 51    Jacob Stallings   1.30                 2021 0.262 0.327   9  41  44
## 52     Gleyber Torres   4.00                 2021 0.271 0.340  65 183 167
## 53   Daniel Vogelbach   1.40                 2021 0.206 0.332  40 107  98
## 54          Luke Voit   4.70                 2021 0.274 0.363  62 168 161
##      SLG   OPS  SB   PA  Kper BBper DRS
## 1  0.515 0.875   3 1985 0.200 0.069 -14
## 2  0.460 0.816  30 1540 0.260 0.115  12
## 3  0.361 0.711  37 1330 0.196 0.093 -17
## 4  0.286 0.522   2  416 0.185 0.043  -1
## 5  0.422 0.760   3 1117 0.218 0.071 -20
## 6  0.409 0.726  22 1421 0.256 0.092  34
## 7  0.362 0.681   0  518 0.243 0.077   8
## 8  0.342 0.614   0  771 0.219 0.066  28
## 9  0.339 0.607  48 1038 0.272 0.046  53
## 10 0.276 0.522   0  264 0.167 0.038   4
## 11 0.431 0.720   5  446 0.305 0.061  -9
## 12 0.313 0.605   4  331 0.181 0.070  -1
## 13 0.425 0.738  44 1313 0.183 0.066  77
## 14 0.527 0.915  28 2014 0.239 0.123   3
## 15 0.426 0.726   2 1646 0.162 0.066 -18
## 16 0.388 0.693  12  608 0.293 0.066   1
## 17 0.408 0.719  11 1506 0.249 0.084  37
## 18 0.366 0.652  11 1221 0.212 0.051   9
## 19 0.368 0.686   4  583 0.187 0.098  -2
## 20 0.462 0.792  11  868 0.194 0.052  -3
## 21 0.414 0.706  60 1369 0.298 0.065  31
## 22 0.401 0.704   0  622 0.280 0.090   6
## 23 0.434 0.748   2 1124 0.206 0.061 -11
## 24 0.378 0.637   7  921 0.279 0.057  35
## 25 0.321 0.640  60  953 0.236 0.097  16
## 26 0.417 0.730  16 1633 0.187 0.098  28
## 27 0.425 0.746   3 1720 0.206 0.078 -22
## 28 0.381 0.700  70 1482 0.122 0.039 -15
## 29 0.330 0.638   1  804 0.218 0.095  15
## 30 0.530 0.862  42 1626 0.301 0.081  33
## 31 0.364 0.678  10  597 0.258 0.077 -17
## 32 0.456 0.803 124 1555 0.182 0.075   1
## 33 0.321 0.643   6  712 0.198 0.112  18
## 34 0.559 0.928  39 1841 0.220 0.124  38
## 35 0.438 0.765   2 1028 0.197 0.076   2
## 36 0.521 0.867  14  921 0.257 0.067 -12
## 37 0.369 0.645  26  982 0.319 0.061   9
## 38 0.336 0.663   2  579 0.314 0.126 -19
## 39 0.494 0.788  10 1450 0.281 0.072  27
## 40 0.352 0.692   5  831 0.252 0.116  -2
## 41 0.422 0.735   1  975 0.182 0.054  -8
## 42 0.359 0.683  14  853 0.212 0.111   7
## 43 0.448 0.795   4  863 0.234 0.096 -31
## 44 0.475 0.806   5  589 0.289 0.090  -9
## 45 0.396 0.701   0  625 0.205 0.099   3
## 46 0.351 0.670  18  846 0.169 0.066  12
```

```
## 47 0.467 0.759   2  709 0.198 0.049  11
## 48 0.388 0.735  16  648 0.276 0.102  -1
## 49 0.494 0.811   1  728 0.254 0.070 -12
## 50 0.557 0.972  23 1349 0.192 0.169 -11
## 51 0.372 0.699   1  425 0.224 0.085  21
## 52 0.493 0.834  12 1248 0.224 0.090 -15
## 53 0.409 0.741   0  840 0.266 0.154  -9
## 54 0.527 0.891   0 1029 0.262 0.109  -9
```

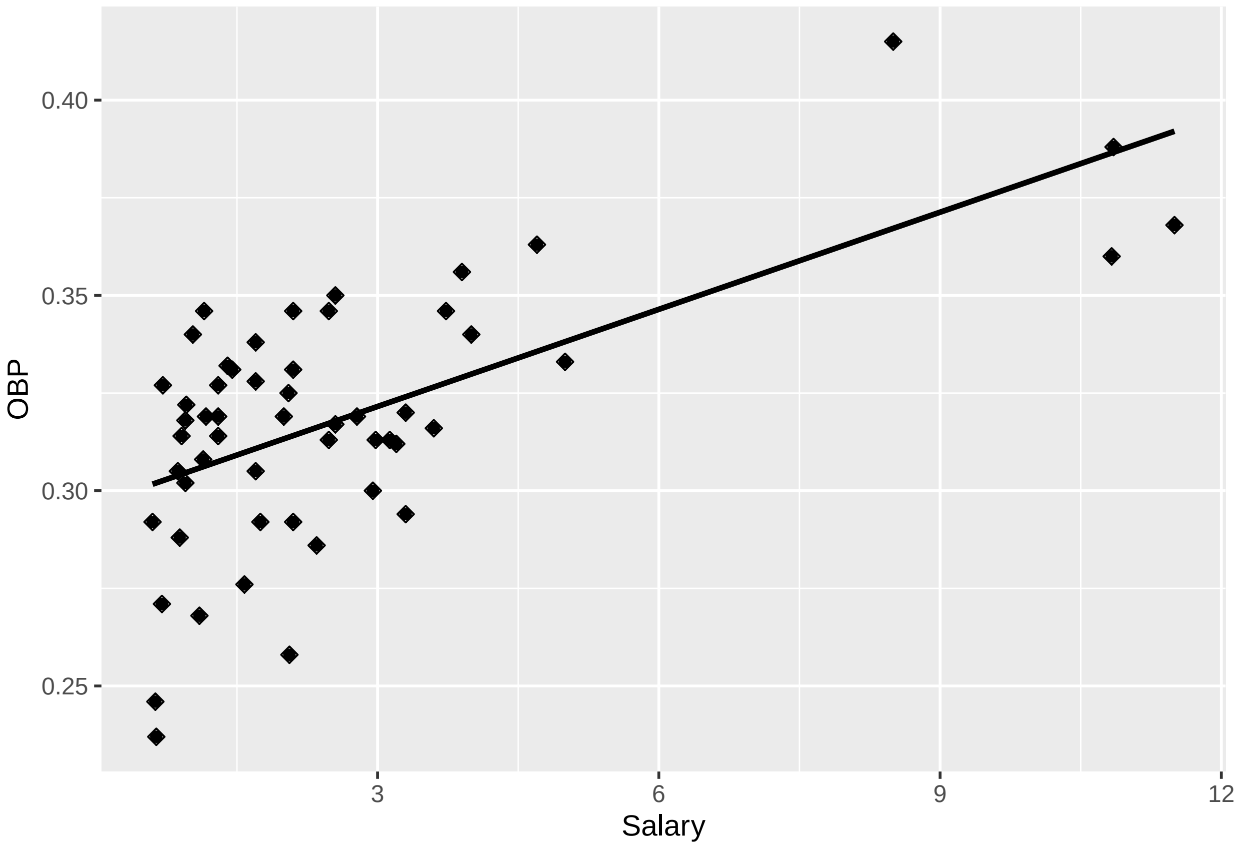
```
summary(baseball)
```

```
##     Player              Salary       First.Arbitration.Year       BA
##  Length:54          Min.   : 0.600   Min.   :2017           Min.   :0.1990
##  Class :character   1st Qu.: 1.143   1st Qu.:2018           1st Qu.:0.2310
##  Mode  :character   Median : 2.025   Median :2019           Median :0.2515
##                     Mean   : 2.642   Mean   :2019           Mean   :0.2494
##                     3rd Qu.: 3.092   3rd Qu.:2020           3rd Qu.:0.2680
##                     Max.   :11.500   Max.   :2021           Max.   :0.2990
##       OBP                HR              RBI               R
##  Min.   :0.2370   Min.   :  2.0   Min.   : 20.00   Min.   : 18.0
##  1st Qu.:0.3028   1st Qu.: 14.0   1st Qu.: 68.25   1st Qu.: 76.5
##  Median :0.3190   Median : 27.0   Median :104.00   Median :111.5
##  Mean   :0.3186   Mean   : 33.7   Mean   :118.81   Mean   :126.0
##  3rd Qu.:0.3367   3rd Qu.: 43.0   3rd Qu.:165.75   3rd Qu.:173.8
##  Max.   :0.4150   Max.   :111.0   Max.   :308.00   Max.   :319.0
##       SLG               OPS               SB               PA
##  Min.   :0.2760   Min.   :0.5220   Min.   :  0.00   Min.   : 264.0
##  1st Qu.:0.3625   1st Qu.:0.6787   1st Qu.:  2.00   1st Qu.: 663.2
##  Median :0.4090   Median :0.7230   Median :  6.50   Median : 937.0
##  Mean   :0.4127   Mean   :0.7314   Mean   : 16.17   Mean   :1026.2
##  3rd Qu.:0.4590   3rd Qu.:0.7910   3rd Qu.: 21.00   3rd Qu.:1364.0
##  Max.   :0.5590   Max.   :0.9720   Max.   :124.00   Max.   :2014.0
##       Kper             BBper              DRS
##  Min.   :0.1220   Min.   :0.03800   Min.   :-31.000
##  1st Qu.:0.1963   1st Qu.:0.06600   1st Qu.:-10.500
##  Median :0.2220   Median :0.07700   Median :  1.500
##  Mean   :0.2300   Mean   :0.08256   Mean   :  5.389
##  3rd Qu.:0.2615   3rd Qu.:0.09775   3rd Qu.: 15.750
##  Max.   :0.3190   Max.   :0.16900   Max.   : 77.000
```

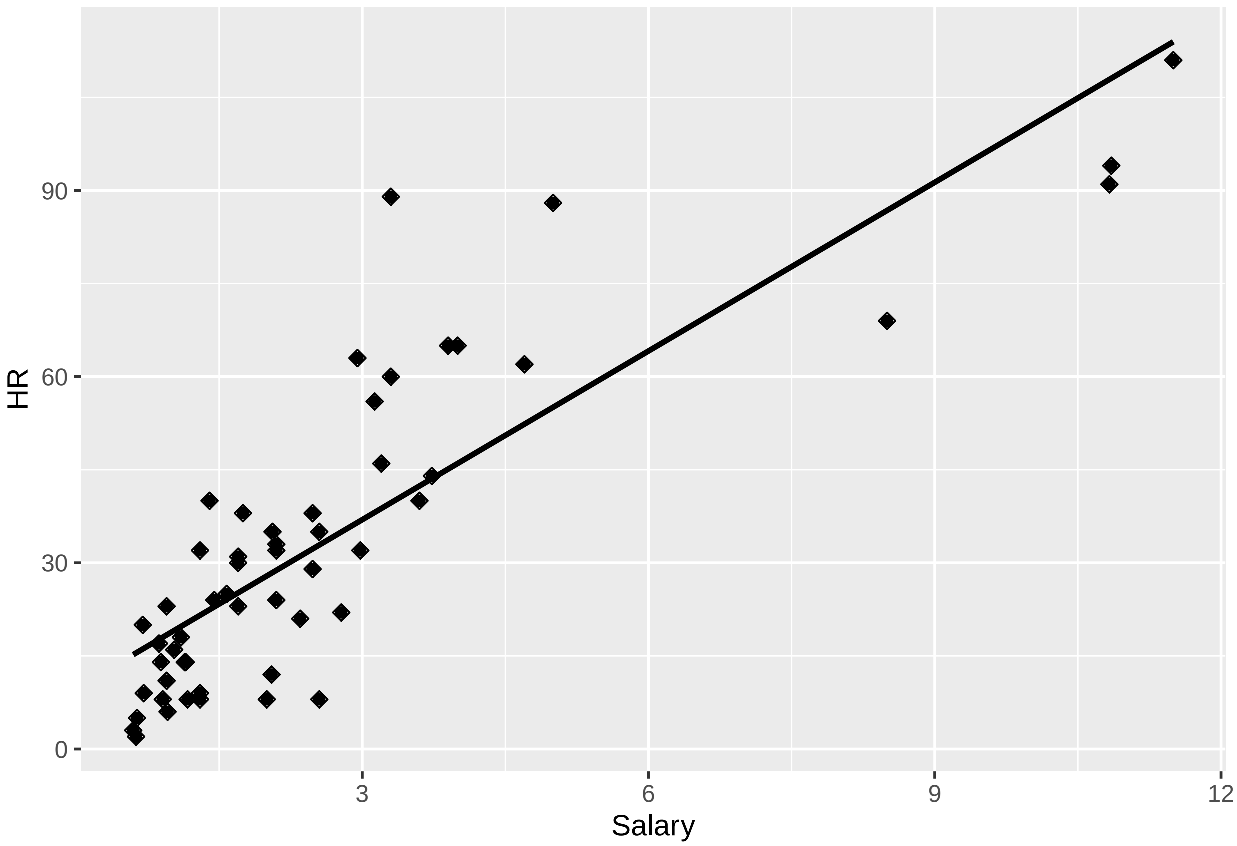
```
ggplot(data = baseball, aes(Salary, BA)) +
  geom_point(size=2, shape=23) +
          geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
          geom_point()
```



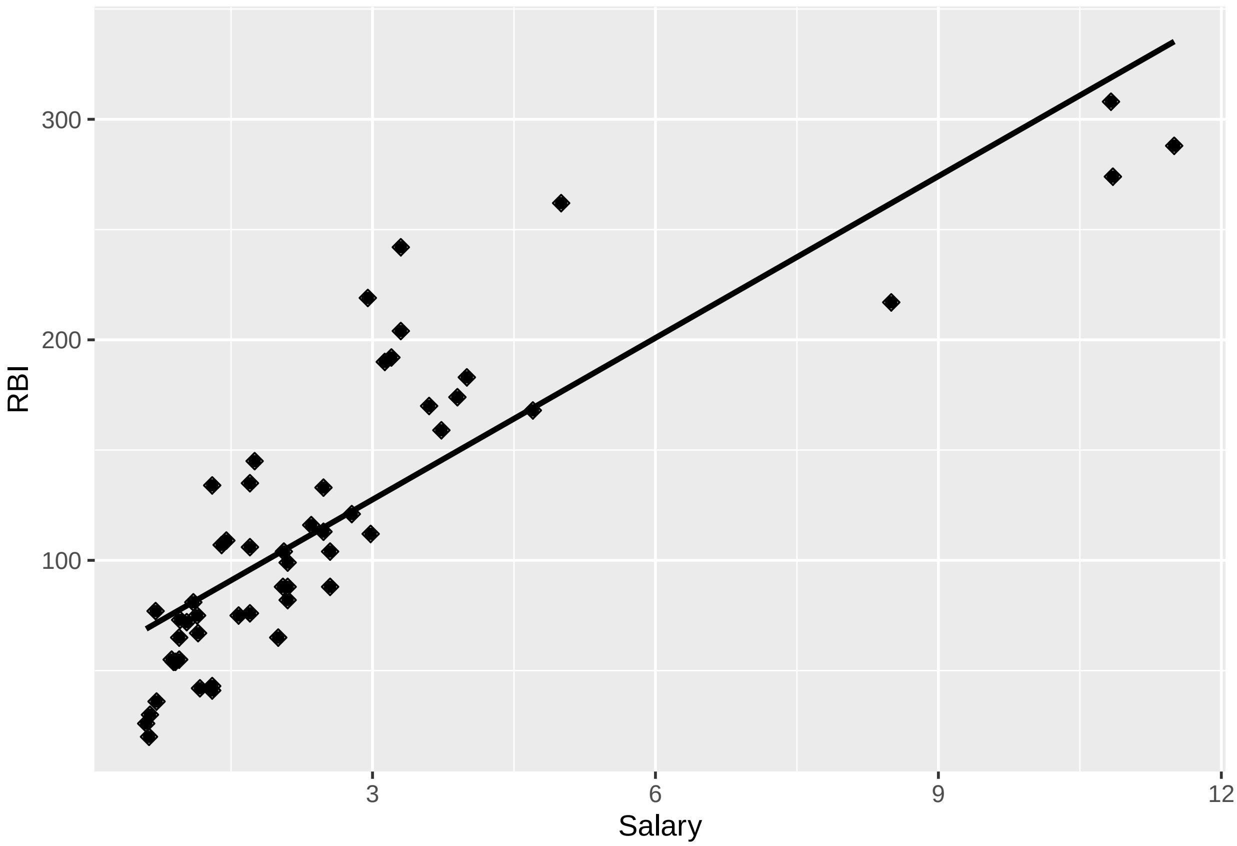

```r
ggplot(data = baseball, aes(Salary, OBP)) +
  geom_point(size=2, shape=23) +
           geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
           geom_point()
```



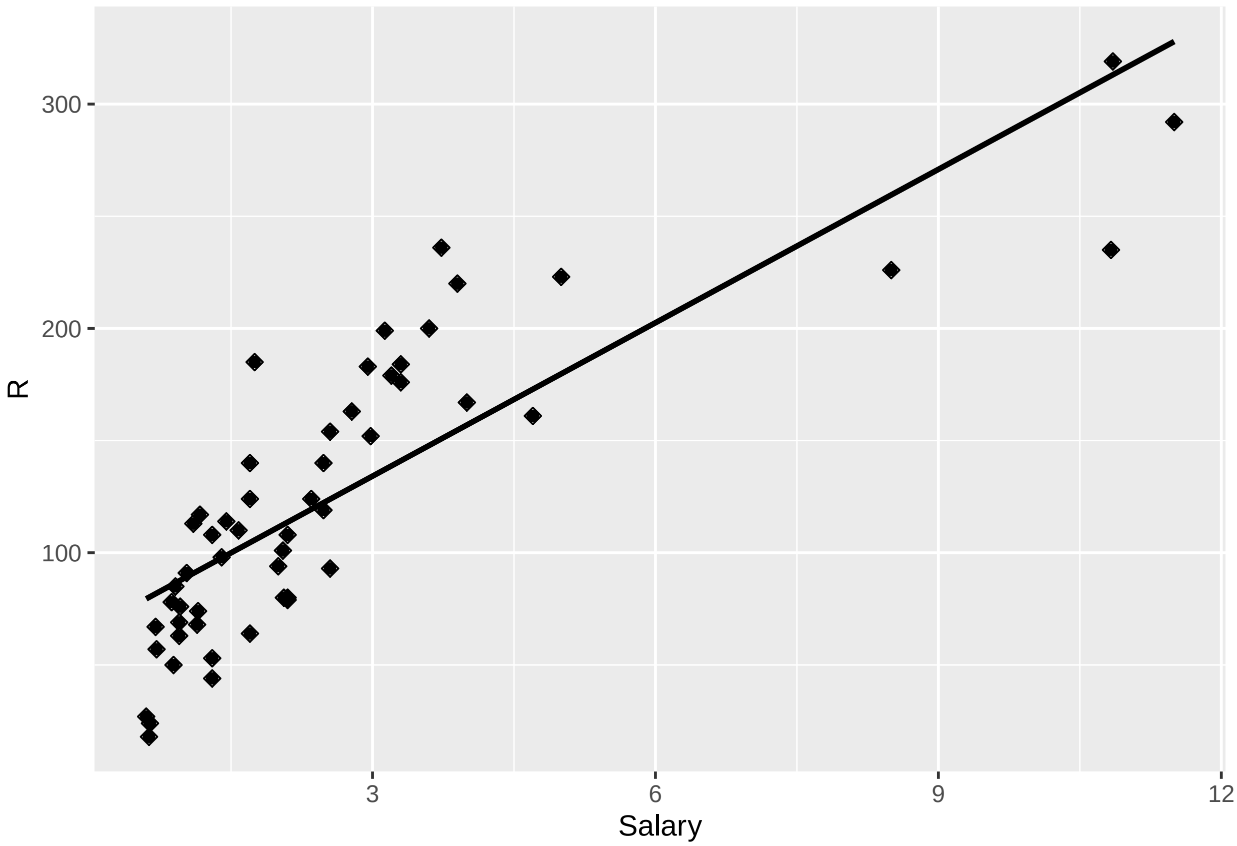

```
ggplot(data = baseball, aes(Salary, HR)) +
  geom_point(size=2, shape=23) +
            geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
            geom_point()
```



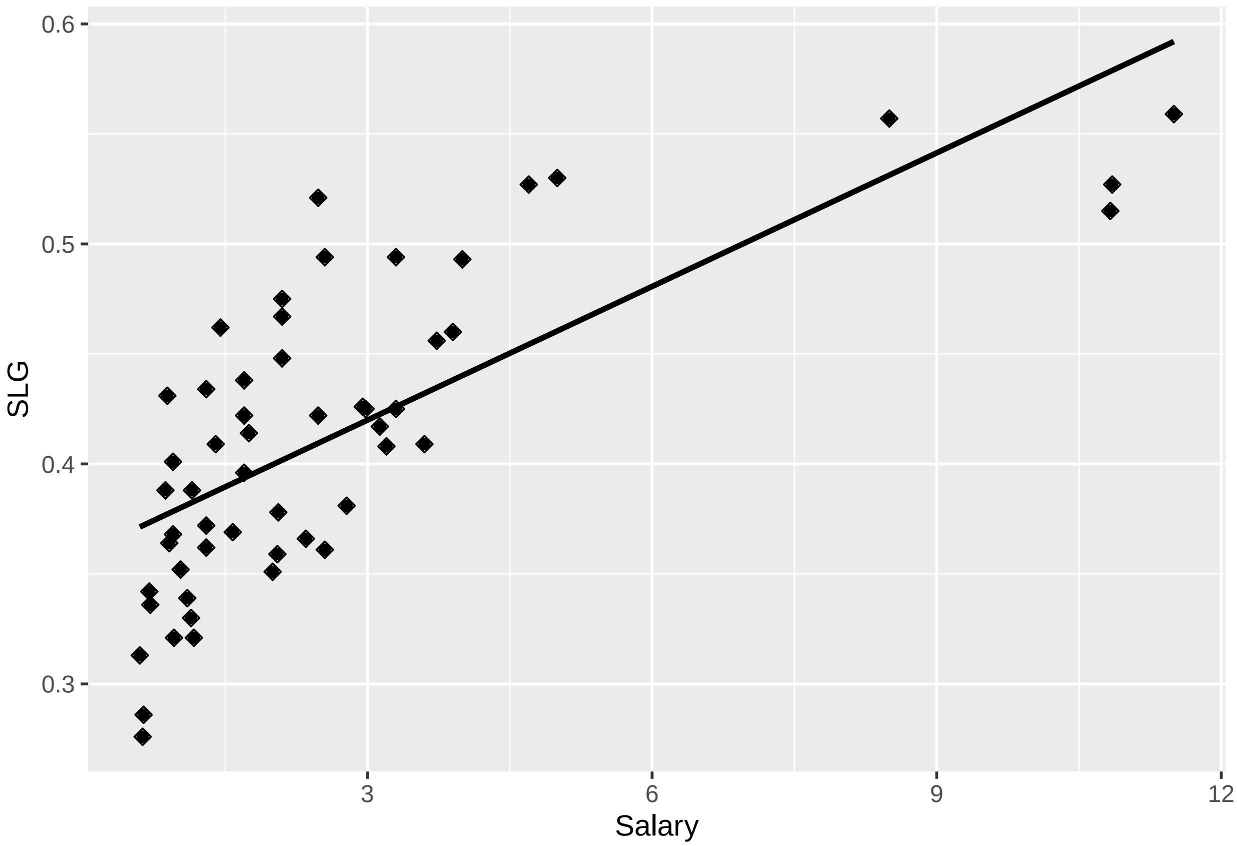

```r
ggplot(data = baseball, aes(Salary, RBI)) +
  geom_point(size=2, shape=23) +
            geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
            geom_point()
```



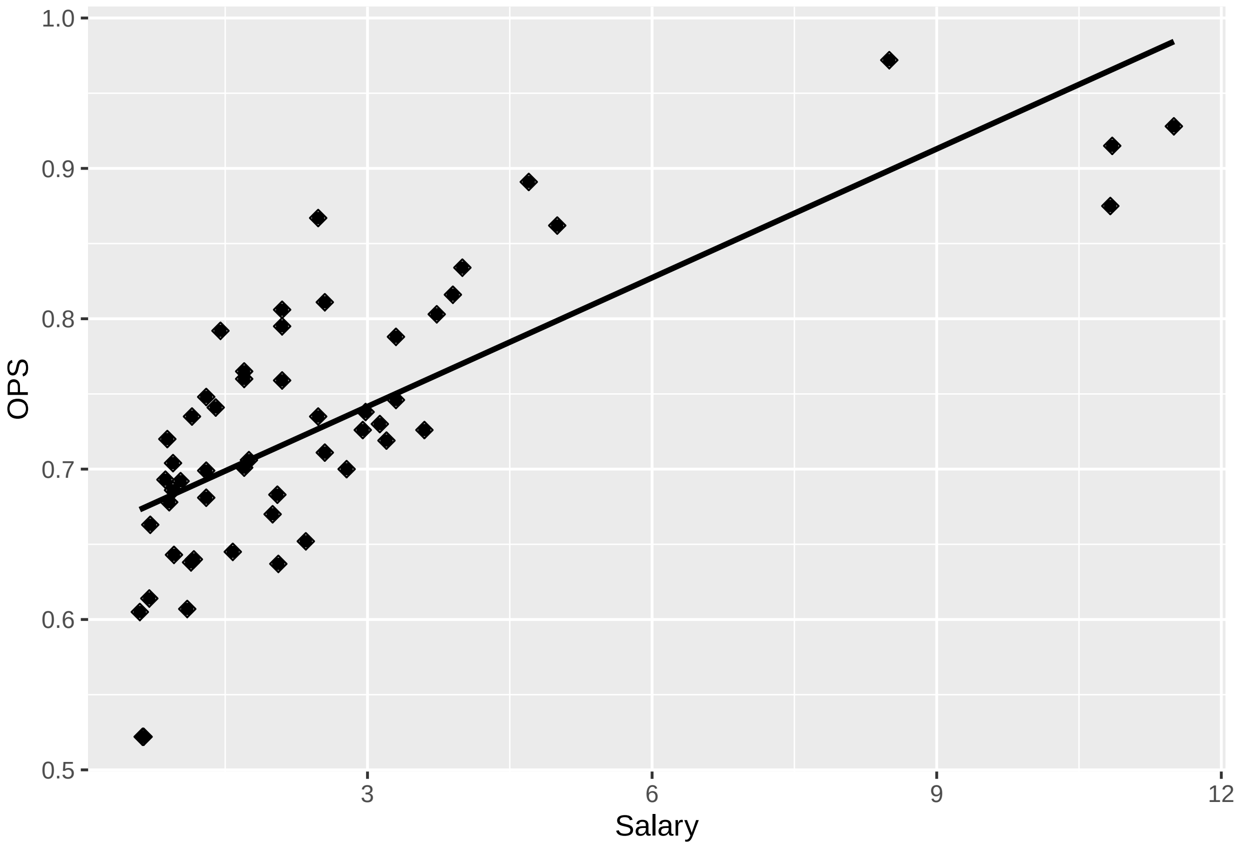

```r
ggplot(data = baseball, aes(Salary, R)) +
  geom_point(size=2, shape=23) +
            geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
            geom_point()
```



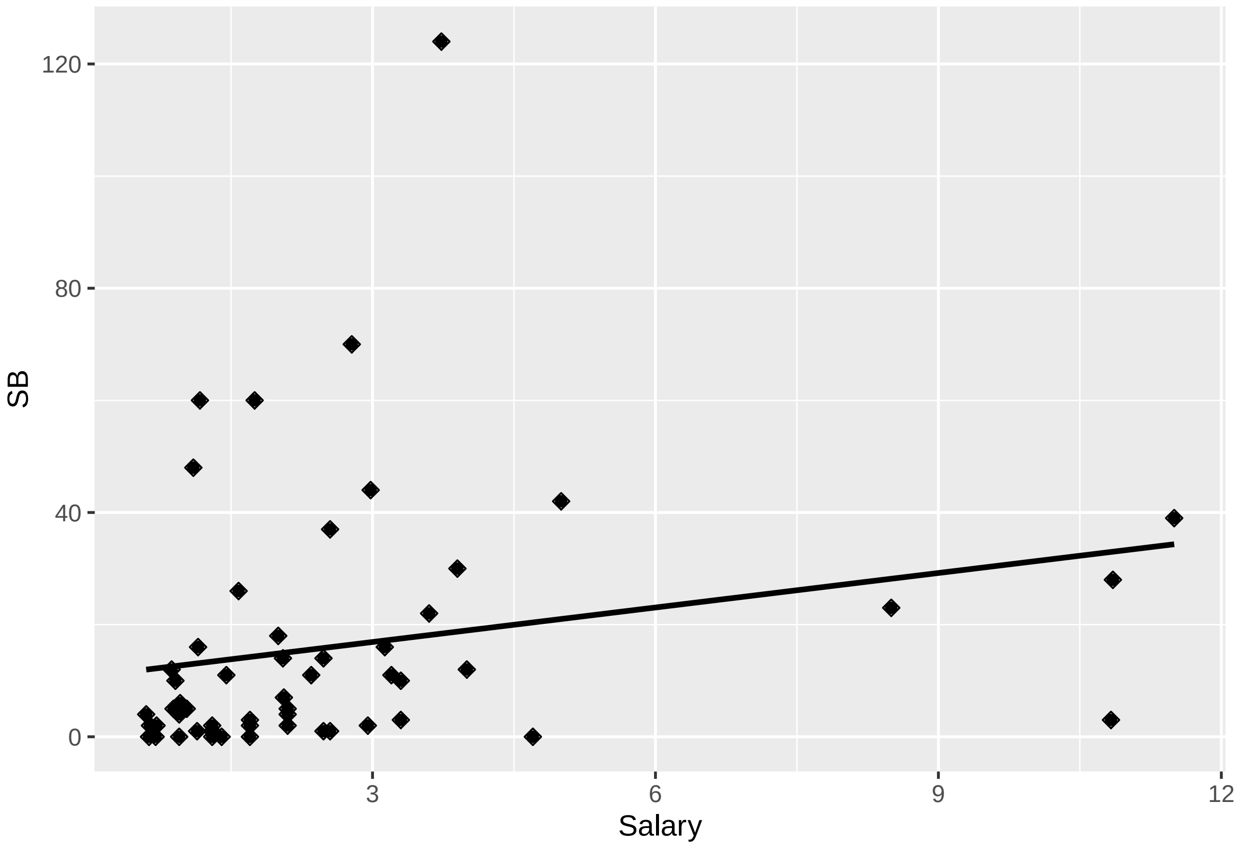

```
ggplot(data = baseball, aes(Salary, SLG)) +
  geom_point(size=2, shape=23) +
            geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
            geom_point()
```



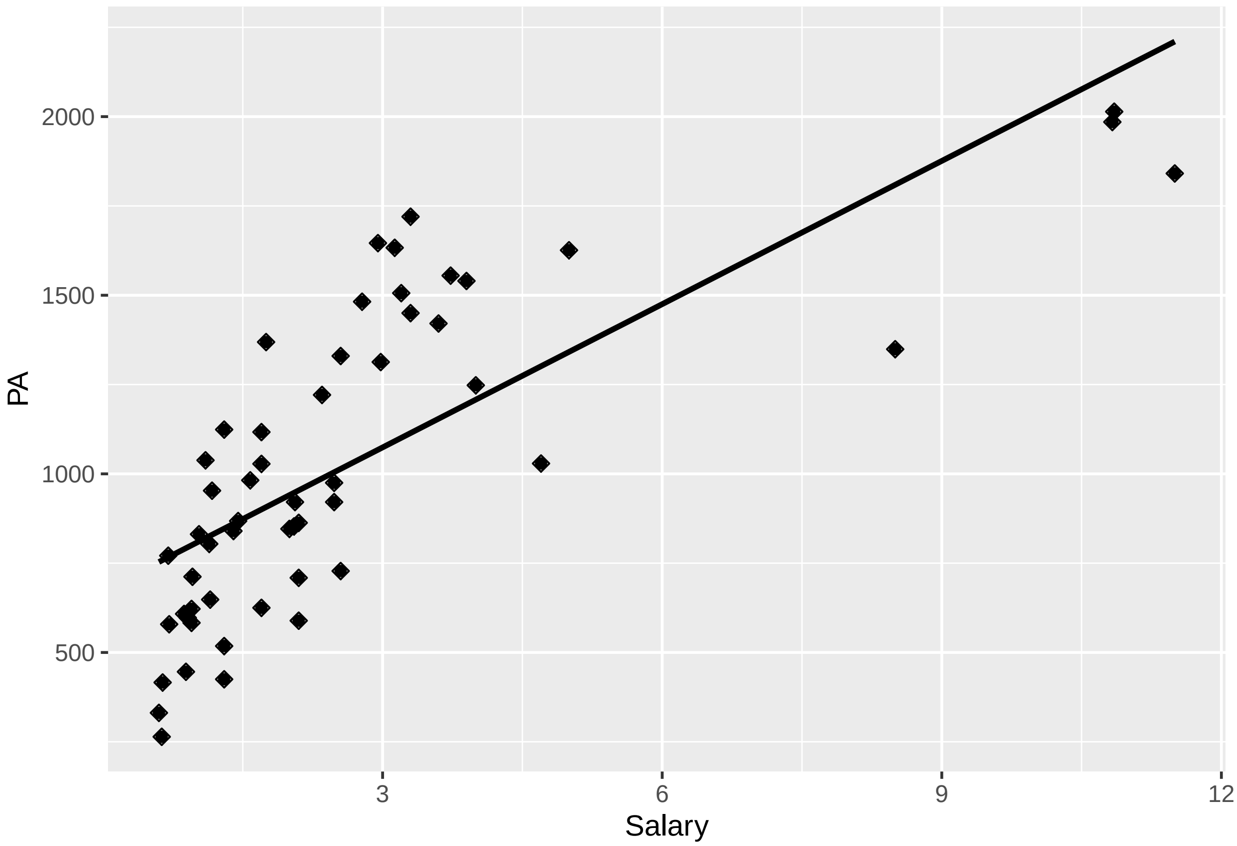

```r
ggplot(data = baseball, aes(Salary, OPS)) +
  geom_point(size=2, shape=23) +
            geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
            geom_point()
```



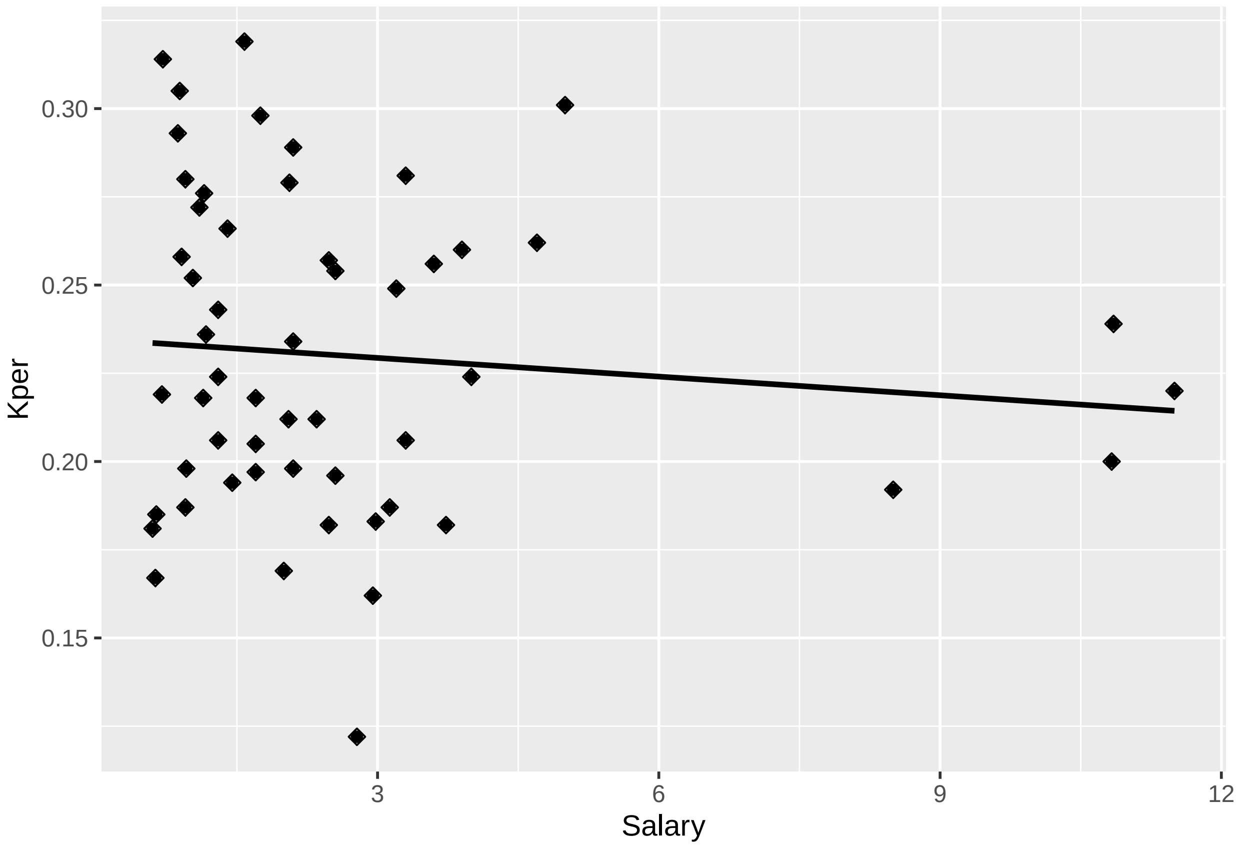

```r
ggplot(data = baseball, aes(Salary, SB)) +
  geom_point(size=2, shape=23) +
            geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
            geom_point()
```



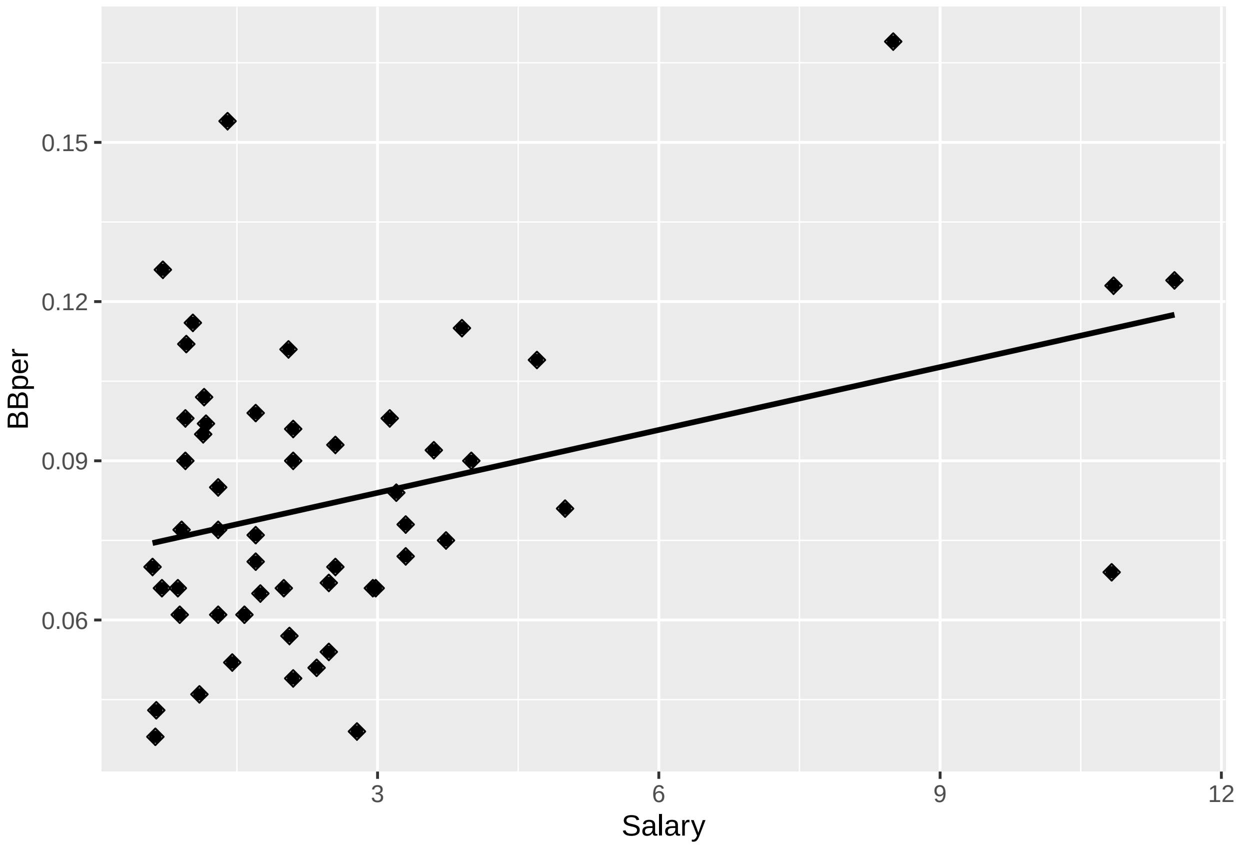

```
ggplot(data = baseball, aes(Salary, PA)) +
  geom_point(size=2, shape=23) +
            geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
            geom_point()
```



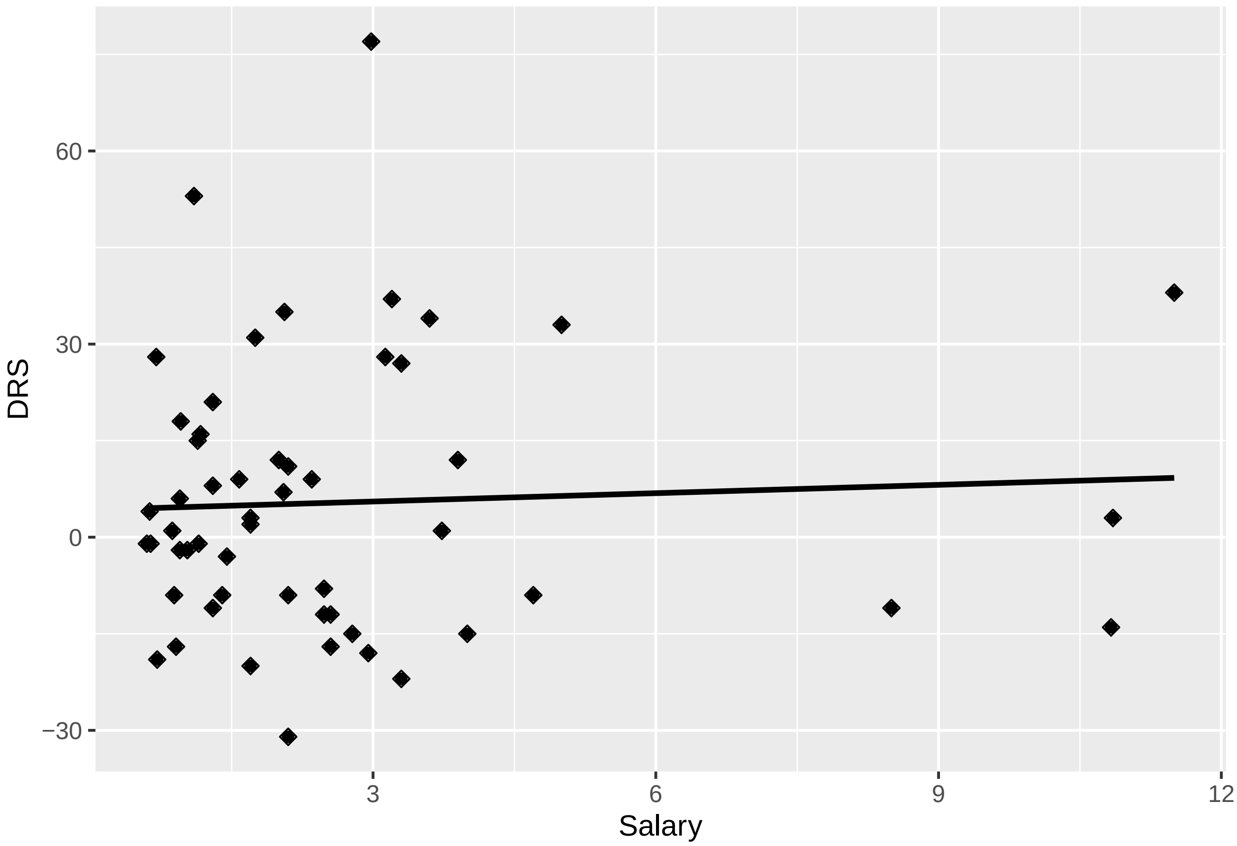

```
ggplot(data = baseball, aes(Salary, Kper)) +
  geom_point(size=2, shape=23) +
          geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
          geom_point()
```




```
ggplot(data = baseball, aes(Salary, BBper)) +
  geom_point(size=2, shape=23) +
            geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
            geom_point()
```




```r
ggplot(data = baseball, aes(Salary, DRS)) +
  geom_point(size=2, shape=23) +
            geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
            geom_point()
```



R-squared values for each graph when compared to Salary:

BA: 0.3459 OBP: 0.4019 HR: 0.7205 RBI: 0.7228 R: 0.7004 SLG: 0.5276 OPS: 0.5726 SB: 0.0494 PA: 0.5641 Kper: 0.0096 BBper:0.1279 DRS: 0.0026

After reviewing the R^2 values of the graphs above, there were a few conclusions that could be reached. We can say the most linear variables are the easiest to use to predict a player's salary. Home Runs, Runs Batted In, and Runs Scored were the variables that can predict salary. Stolen Bases, Strikeout Percent, Walk Percentage, and Defensive Runs Saved were the variables we did not find to have a strong predictive tendency.