Project For Applied linear regression

Maxime Cathala

11/4/2022

Loading Data

```
library("faraway")
path <- "/Users/Maxime/Desktop/NBA_Salaries/2017-18_NBA_salary.csv"
full_df <- read.csv(path, header=TRUE, stringsAsFactors=FALSE)
nba <- subset(full_df, select = c(Salary,MP,TS.,X3PAr,FTr,TRB.,AST.,STL.,
                                  BLK.,TOV.,USG.))
nba <- na.omit(nba)
head(nba)
```

```
##     Salary   MP   TS. X3PAr   FTr TRB. AST. STL. BLK. TOV. USG.
## 1   815615   87 0.303 0.593 0.370 11.7  1.5  1.1  6.8 18.2 19.5
## 2  3477600  937 0.608 0.004 0.337 18.5 15.4  1.9  1.3 19.3 17.2
## 3 12307692 1508 0.529 0.193 0.140 15.0 14.9  1.4  0.6 12.5 27.6
## 4  3202217  656 0.499 0.346 0.301  7.7 18.6  1.8  0.5  9.7 29.5
## 5  3057240  979 0.487 0.387 0.146 11.7  7.3  0.8  2.5 15.6 15.5
## 6  1312611 2238 0.543 0.489 0.141  6.1 13.3  1.4  0.3  9.1 17.0
```

Full Model Ana

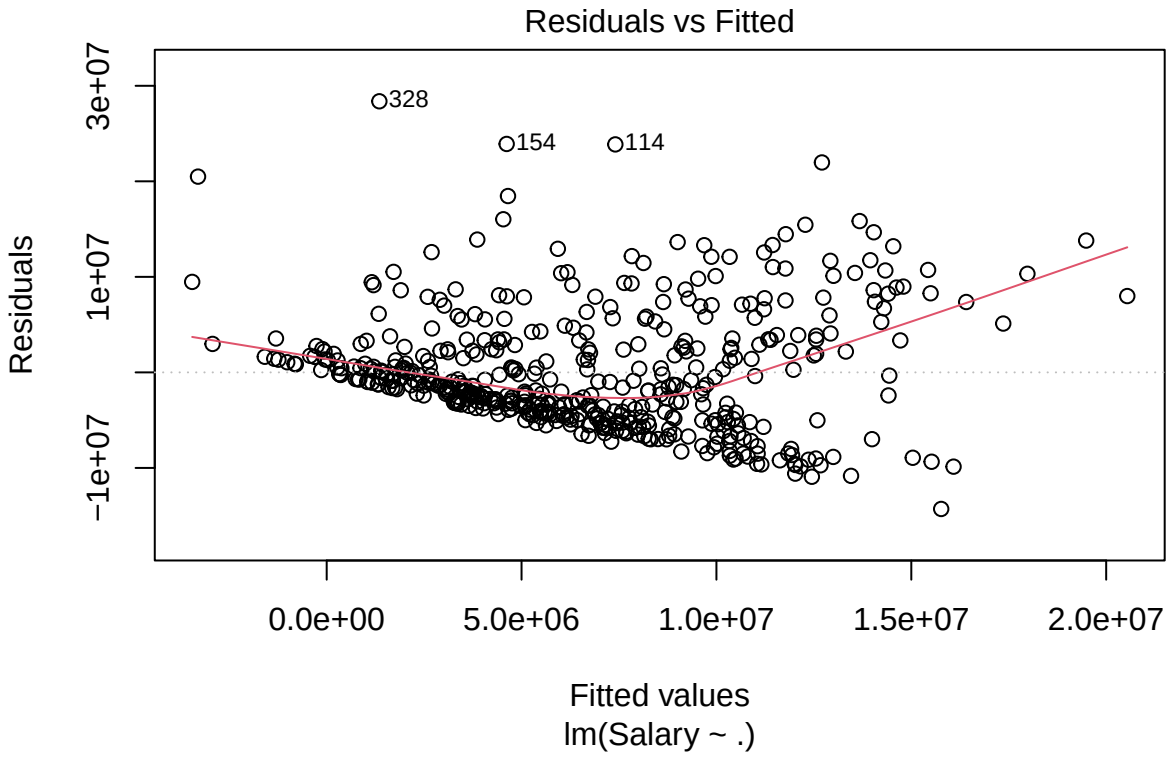
Create Full Model (No modif, All predictors vs Salary)

```
full <- lm(Salary ~ .,data = nba)
```

Model Assumptions:

Residuals vs Fitted:
Quite a random scatter.
If not looking where no points are below

```
plot(full, which = 1)
```

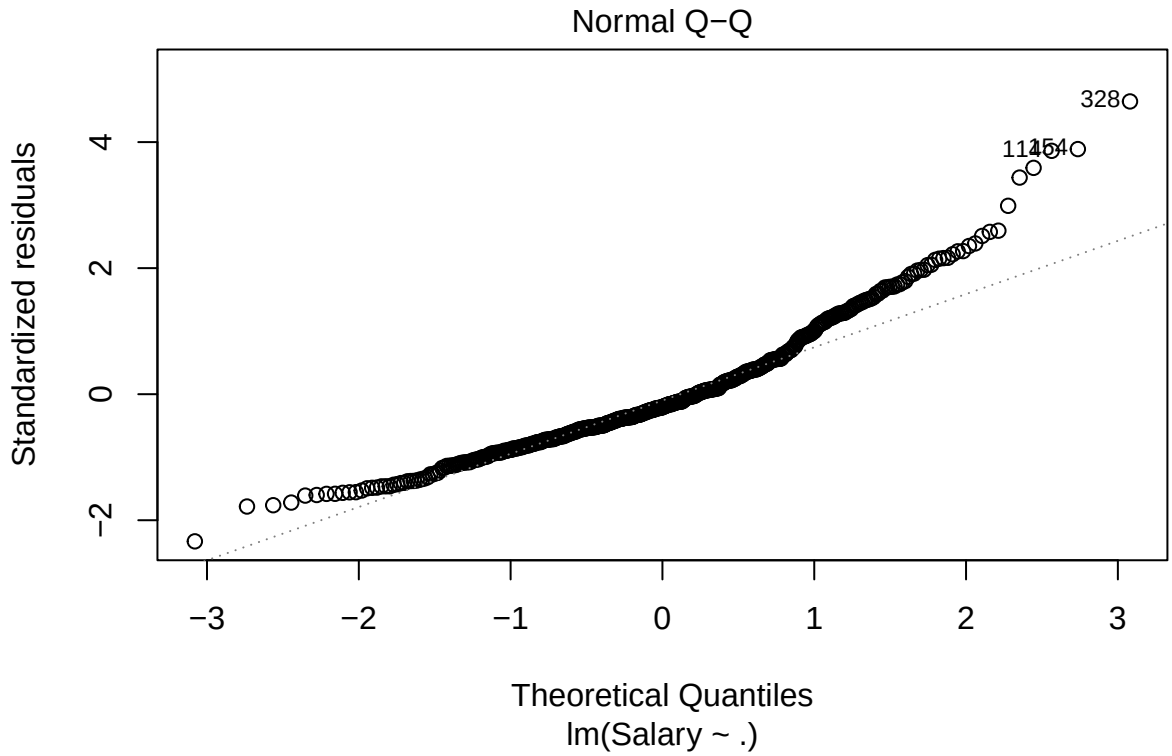


Normality:

Pb with the higher quantiles

Points should follow the line (Normal distribution)

```
plot(full, which = 2)
```



Variable selection (AIC-Methods on Full Model)

```
full_AIC <- step(full, direction = "backward",k = 2)
```

```
## Start:  AIC=15118.67
## Salary ~ MP + TS. + X3PAr + FTr + TRB. + AST. + STL. + BLK. +
##     TOV. + USG.
##
##         Df  Sum of Sq        RSS   AIC
## - STL.   1 7.6966e+10 1.8125e+16 15117
## - BLK.   1 1.8143e+12 1.8127e+16 15117
## - TOV.   1 5.7155e+12 1.8130e+16 15117
## - TS.    1 9.3564e+12 1.8134e+16 15117
## - FTr    1 1.1660e+13 1.8136e+16 15117
## - X3PAr  1 2.6081e+13 1.8151e+16 15117
## <none>                1.8125e+16 15119
## - AST.   1 2.2618e+14 1.8351e+16 15123
## - TRB.   1 4.1439e+14 1.8539e+16 15128
## - USG.   1 4.2326e+14 1.8548e+16 15128
## - MP     1 3.6907e+15 2.1815e+16 15206
##
## Step:  AIC=15116.67
## Salary ~ MP + TS. + X3PAr + FTr + TRB. + AST. + BLK. + TOV. +
##     USG.
##
##         Df  Sum of Sq        RSS   AIC
## - BLK.   1 1.7737e+12 1.8127e+16 15115
## - TOV.   1 5.8914e+12 1.8131e+16 15115
## - TS.    1 9.2796e+12 1.8134e+16 15115
## - FTr    1 1.1586e+13 1.8136e+16 15115
## - X3PAr  1 2.6061e+13 1.8151e+16 15115
## <none>                1.8125e+16 15117
## - AST.   1 2.3845e+14 1.8363e+16 15121
## - TRB.   1 4.1652e+14 1.8541e+16 15126
## - USG.   1 4.2662e+14 1.8551e+16 15126
## - MP     1 3.6924e+15 2.1817e+16 15204
##
## Step:  AIC=15114.72
## Salary ~ MP + TS. + X3PAr + FTr + TRB. + AST. + TOV. + USG.
##
##         Df  Sum of Sq        RSS   AIC
## - TOV.   1 6.0678e+12 1.8133e+16 15113
## - TS.    1 8.2695e+12 1.8135e+16 15113
## - FTr    1 1.2223e+13 1.8139e+16 15113
## - X3PAr  1 2.9062e+13 1.8156e+16 15114
## <none>                1.8127e+16 15115
## - AST.   1 2.4773e+14 1.8374e+16 15119
## - USG.   1 4.2561e+14 1.8552e+16 15124
## - TRB.   1 4.6901e+14 1.8596e+16 15125
## - MP     1 3.7031e+15 2.1830e+16 15202
##
## Step:  AIC=15112.88
## Salary ~ MP + TS. + X3PAr + FTr + TRB. + AST. + USG.
##
```

```
##         Df  Sum of Sq        RSS   AIC
## - TS.    1 7.6471e+12 1.8140e+16 15111
## - FTr    1 1.0447e+13 1.8143e+16 15111
## - X3PAr  1 3.4032e+13 1.8167e+16 15112
## <none>                1.8133e+16 15113
## - AST.   1 2.4839e+14 1.8381e+16 15117
## - USG.   1 4.6243e+14 1.8595e+16 15123
## - TRB.   1 4.6416e+14 1.8597e+16 15123
## - MP     1 3.9040e+15 2.2037e+16 15205
##
## Step:  AIC=15111.08
## Salary ~ MP + X3PAr + FTr + TRB. + AST. + USG.
##
##         Df  Sum of Sq        RSS   AIC
## - FTr    1 1.4724e+13 1.8155e+16 15110
## - X3PAr  1 3.4273e+13 1.8175e+16 15110
## <none>                1.8140e+16 15111
## - AST.   1 2.4718e+14 1.8388e+16 15116
## - USG.   1 4.6857e+14 1.8609e+16 15121
## - TRB.   1 4.7444e+14 1.8615e+16 15122
## - MP     1 4.2739e+15 2.2414e+16 15211
##
## Step:  AIC=15109.47
## Salary ~ MP + X3PAr + TRB. + AST. + USG.
##
##         Df  Sum of Sq        RSS   AIC
## - X3PAr  1 2.8166e+13 1.8183e+16 15108
## <none>                1.8155e+16 15110
## - AST.   1 2.5016e+14 1.8405e+16 15114
## - USG.   1 4.6860e+14 1.8624e+16 15120
## - TRB.   1 4.8987e+14 1.8645e+16 15120
## - MP     1 4.2593e+15 2.2414e+16 15209
##
## Step:  AIC=15108.22
## Salary ~ MP + TRB. + AST. + USG.
##
##        Df  Sum of Sq        RSS   AIC
## <none>               1.8183e+16 15108
## - AST.  1 2.2511e+14 1.8408e+16 15112
## - USG.  1 4.6531e+14 1.8649e+16 15118
## - TRB.  1 5.5041e+14 1.8734e+16 15121
## - MP    1 4.3719e+15 2.2555e+16 15210
```

Fit Reduced model with selected predictors in AIC

```
reduced <- lm(Salary ~ AST. + USG. + TRB. + MP, data = nba)
summary(reduced)
```

```
##
## Call:
## lm(formula = Salary ~ AST. + USG. + TRB. + MP, data = nba)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -13935004  -4058565  -1230742   2689042  28737162
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4842241.8  1134506.7  -4.268 2.38e-05 ***
## AST.           87432.9    35942.1   2.433 0.015357 *
## USG.          186335.2    53277.4   3.497 0.000514 ***
## TRB.          225252.8    59217.3   3.804 0.000161 ***
## MP              3959.9      369.4  10.720  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6168000 on 478 degrees of freedom
## Multiple R-squared:  0.3108, Adjusted R-squared:  0.305
## F-statistic: 53.88 on 4 and 478 DF,  p-value: < 2.2e-16
```

```
vif(reduced)
```

```
##     AST.     USG.     TRB.       MP
## 1.353386 1.220988 1.077895 1.133232
```

Is reduced model better or same then full?

The hypotheses for the F-test performed in the anova are:

H0: The additional terms in the full model are 0.
HA: At least one of the additional terms is non 0.
Yes: p val = 0.9577

```
anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: Salary ~ AST. + USG. + TRB. + MP
## Model 2: Salary ~ MP + TS. + X3PAr + FTr + TRB. + AST. + STL. + BLK. +
##     TOV. + USG.
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1    478 1.8183e+16
## 2    472 1.8125e+16  6 5.8455e+13 0.2537 0.9577
```

Exploring other methods Full with sqrt(response)

Remove outliers (Salary over 25millions)

```
nba2 <- nba[nba$Salary <= 25000000,]
full2 <- lm((Salary) ~ .,data = nba2)
full2_AIC <- step(full2, direction = "backward",k = 2)
```

```
## Start:  AIC=14558.14
## (Salary) ~ MP + TS. + X3PAr + FTr + TRB. + AST. + STL. + BLK. +
##     TOV. + USG.
##
##         Df  Sum of Sq        RSS   AIC
## - BLK.   1 1.1457e+11 1.3541e+16 14556
## - X3PAr  1 2.8001e+11 1.3541e+16 14556
## - FTr    1 1.6512e+12 1.3543e+16 14556
```

```
## - STL.   1 5.6950e+12 1.3547e+16 14556
## - TS.    1 5.8596e+12 1.3547e+16 14556
## - TOV.   1 6.8865e+12 1.3548e+16 14556
## - AST.   1 1.1108e+13 1.3552e+16 14556
## <none>                1.3541e+16 14558
## - USG.   1 1.7542e+14 1.3716e+16 14562
## - TRB.   1 2.1598e+14 1.3757e+16 14564
## - MP     1 3.3902e+15 1.6931e+16 14661
##
## Step:  AIC=14556.14
## (Salary) ~ MP + TS. + X3PAr + FTr + TRB. + AST. + STL. + TOV. +
##     USG.
##
##         Df  Sum of Sq        RSS   AIC
## - X3PAr  1 2.3418e+11 1.3541e+16 14554
## - FTr    1 1.7116e+12 1.3543e+16 14554
## - STL.   1 5.6174e+12 1.3547e+16 14554
## - TS.    1 5.7471e+12 1.3547e+16 14554
## - TOV.   1 6.8444e+12 1.3548e+16 14554
## - AST.   1 1.1605e+13 1.3553e+16 14554
## <none>                1.3541e+16 14556
## - USG.   1 1.7531e+14 1.3716e+16 14560
## - TRB.   1 2.5233e+14 1.3793e+16 14563
## - MP     1 3.3980e+15 1.6939e+16 14659
##
## Step:  AIC=14554.15
## (Salary) ~ MP + TS. + FTr + TRB. + AST. + STL. + TOV. + USG.
##
##        Df  Sum of Sq        RSS   AIC
## - FTr   1 1.9503e+12 1.3543e+16 14552
## - TS.   1 5.7454e+12 1.3547e+16 14552
## - STL.  1 5.9622e+12 1.3547e+16 14552
## - TOV.  1 7.4283e+12 1.3549e+16 14552
## - AST.  1 1.2472e+13 1.3554e+16 14553
## <none>               1.3541e+16 14554
## - USG.  1 1.7694e+14 1.3718e+16 14558
## - TRB.  1 3.7859e+14 1.3920e+16 14565
## - MP    1 3.4039e+15 1.6945e+16 14657
##
## Step:  AIC=14552.22
## (Salary) ~ MP + TS. + TRB. + AST. + STL. + TOV. + USG.
##
##        Df  Sum of Sq        RSS   AIC
## - STL.  1 5.6006e+12 1.3549e+16 14550
## - TS.   1 7.2491e+12 1.3551e+16 14550
## - TOV.  1 8.6391e+12 1.3552e+16 14550
## - AST.  1 1.2597e+13 1.3556e+16 14551
## <none>               1.3543e+16 14552
## - USG.  1 1.7714e+14 1.3720e+16 14556
## - TRB.  1 3.9647e+14 1.3940e+16 14564
## - MP    1 3.4189e+15 1.6962e+16 14656
##
## Step:  AIC=14550.41
## (Salary) ~ MP + TS. + TRB. + AST. + TOV. + USG.
```

```
##
##        Df  Sum of Sq        RSS   AIC
## - TS.   1 6.1786e+12 1.3555e+16 14549
## - TOV.  1 7.4889e+12 1.3556e+16 14549
## - AST.  1 1.7236e+13 1.3566e+16 14549
## <none>               1.3549e+16 14550
## - USG.  1 1.7270e+14 1.3722e+16 14554
## - TRB.  1 3.9548e+14 1.3944e+16 14562
## - MP    1 3.4242e+15 1.6973e+16 14654
##
## Step:  AIC=14548.63
## (Salary) ~ MP + TRB. + AST. + TOV. + USG.
##
##        Df  Sum of Sq        RSS   AIC
## - TOV.  1 8.5608e+12 1.3564e+16 14547
## - AST.  1 1.6641e+13 1.3572e+16 14547
## <none>               1.3555e+16 14549
## - USG.  1 1.7679e+14 1.3732e+16 14553
## - TRB.  1 4.0961e+14 1.3965e+16 14561
## - MP    1 3.7343e+15 1.7289e+16 14661
##
## Step:  AIC=14546.92
## (Salary) ~ MP + TRB. + AST. + USG.
##
##        Df  Sum of Sq        RSS   AIC
## - AST.  1 2.9219e+13 1.3593e+16 14546
## <none>               1.3564e+16 14547
## - USG.  1 1.6864e+14 1.3732e+16 14551
## - TRB.  1 4.5535e+14 1.4019e+16 14560
## - MP    1 3.8275e+15 1.7391e+16 14662
##
## Step:  AIC=14545.93
## (Salary) ~ MP + TRB. + USG.
##
##        Df  Sum of Sq        RSS   AIC
## <none>               1.3593e+16 14546
## - USG.  1 2.4021e+14 1.3833e+16 14552
## - TRB.  1 4.2656e+14 1.4019e+16 14558
## - MP    1 4.1663e+15 1.7759e+16 14669
```

```
reduced2 <- lm(Salary ~ USG. +  TRB. + MP,data = nba2)
anova(reduced2,full2)
```

```
## Warning in anova.lmlist(object, ...): models with response '"(Salary)"' removed
## because response differs from model 1
```

```
## Analysis of Variance Table
##
## Response: Salary
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## USG.        1 8.5283e+14 8.5283e+14  29.174 1.058e-07 ***
## TRB.        1 4.8220e+14 4.8220e+14  16.495 5.723e-05 ***
## MP          1 4.1663e+15 4.1663e+15 142.526 < 2.2e-16 ***
## Residuals 465 1.3593e+16 2.9232e+13
## ---
```

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

Full with sqrt(response)

```
full3 <- lm(sqrt(Salary) ~ .,data = nba)
full3_AIC <- step(full3, direction = "backward",k = 2)
```

```
## Start:  AIC=6767.92
## sqrt(Salary) ~ MP + TS. + X3PAr + FTr + TRB. + AST. + STL. +
##     BLK. + TOV. + USG.
##
##         Df Sum of Sq       RSS    AIC
## - STL.   1      6324 561847414 6765.9
## - X3PAr  1      9311 561850401 6765.9
## - FTr    1     94396 561935486 6766.0
## - BLK.   1    295875 562136965 6766.2
## - TOV.   1    302578 562143668 6766.2
## - TS.    1    761106 562602196 6766.6
## <none>               561841090 6767.9
## - AST.   1   4577940 566419030 6769.8
## - USG.   1   5671293 567512383 6770.8
## - TRB.   1  12403395 574244486 6776.5
## - MP     1 164862184 726703275 6890.2
##
## Step:  AIC=6765.93
## sqrt(Salary) ~ MP + TS. + X3PAr + FTr + TRB. + AST. + BLK. +
##     TOV. + USG.
##
##         Df Sum of Sq       RSS    AIC
## - X3PAr  1     11037 561858452 6763.9
## - FTr    1     91642 561939057 6764.0
## - TOV.   1    297054 562144468 6764.2
## - BLK.   1    302780 562150195 6764.2
## - TS.    1    778869 562626284 6764.6
## <none>               561847414 6765.9
## - AST.   1   4714750 566562164 6768.0
## - USG.   1   5771804 567619218 6768.9
## - TRB.   1  12545003 574392417 6774.6
## - MP     1 164876882 726724296 6888.2
##
## Step:  AIC=6763.94
## sqrt(Salary) ~ MP + TS. + FTr + TRB. + AST. + BLK. + TOV. + USG.
##
##        Df Sum of Sq       RSS    AIC
## - FTr   1    103428 561961879 6762.0
## - TOV.  1    320815 562179267 6762.2
## - BLK.  1    330365 562188817 6762.2
## - TS.   1    787259 562645711 6762.6
## <none>              561858452 6763.9
## - AST.  1   4782977 566641428 6766.0
## - USG.  1   5760872 567619324 6766.9
## - TRB.  1  15003124 576861576 6774.7
## - MP    1 165329839 727188291 6886.5
##
```

```
## Step:  AIC=6762.03
## sqrt(Salary) ~ MP + TS. + TRB. + AST. + BLK. + TOV. + USG.
##
##        Df Sum of Sq       RSS    AIC
## - BLK.  1    315856 562277735 6760.3
## - TOV.  1    382245 562344124 6760.4
## - TS.   1    708219 562670099 6760.6
## <none>              561961879 6762.0
## - AST.  1   4774707 566736587 6764.1
## - USG.  1   5730156 567692035 6764.9
## - TRB.  1  14975118 576936998 6772.7
## - MP    1 167620373 729582252 6886.1
##
## Step:  AIC=6760.3
## sqrt(Salary) ~ MP + TS. + TRB. + AST. + TOV. + USG.
##
##        Df Sum of Sq       RSS    AIC
## - TOV.  1    415660 562693396 6758.7
## - TS.   1    590455 562868191 6758.8
## <none>              562277735 6760.3
## - AST.  1   5081530 567359265 6762.6
## - USG.  1   5658816 567936551 6763.1
## - TRB.  1  17610867 579888602 6773.2
## - MP    1 168482920 730760655 6884.9
##
## Step:  AIC=6758.65
## sqrt(Salary) ~ MP + TS. + TRB. + AST. + USG.
##
##        Df Sum of Sq       RSS    AIC
## - TS.   1    523742 563217138 6757.1
## <none>              562693396 6758.7
## - AST.  1   4689973 567383369 6760.7
## - USG.  1   6472584 569165980 6762.2
## - TRB.  1  17238843 579932238 6771.2
## - MP    1 180748384 743441779 6891.2
##
## Step:  AIC=6757.1
## sqrt(Salary) ~ MP + TRB. + AST. + USG.
##
##        Df Sum of Sq       RSS    AIC
## <none>              563217138 6757.1
## - AST.  1   4676272 567893409 6759.1
## - USG.  1   6646104 569863241 6760.8
## - TRB.  1  18196192 581413329 6770.5
## - MP    1 196780224 759997362 6899.8
```

```
reduced3 <- lm(sqrt(Salary) ~ AST. + USG. + TRB. + MP,data = nba)
summary(reduced3)
```

```
##
## Call:
## lm(formula = sqrt(Salary) ~ AST. + USG. + TRB. + MP, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2605.5  -804.7  -198.7   731.2  4360.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 238.08948  199.66855   1.192   0.2337
## AST.         12.60177    6.32566   1.992   0.0469 *
## USG.         22.26924    9.37661   2.375   0.0179 *
## TRB.         40.95596   10.42200   3.930 9.76e-05 ***
## MP            0.84012    0.06501  12.923  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1085 on 478 degrees of freedom
## Multiple R-squared:  0.3507, Adjusted R-squared:  0.3453
## F-statistic: 64.54 on 4 and 478 DF,  p-value: < 2.2e-16
```

Reduced models from both methods

```
summary(reduced3)
```

```
##
## Call:
## lm(formula = sqrt(Salary) ~ AST. + USG. + TRB. + MP, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2605.5  -804.7  -198.7   731.2  4360.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 238.08948  199.66855   1.192   0.2337
## AST.         12.60177    6.32566   1.992   0.0469 *
## USG.         22.26924    9.37661   2.375   0.0179 *
## TRB.         40.95596   10.42200   3.930 9.76e-05 ***
## MP            0.84012    0.06501  12.923  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1085 on 478 degrees of freedom
## Multiple R-squared:  0.3507, Adjusted R-squared:  0.3453
## F-statistic: 64.54 on 4 and 478 DF,  p-value: < 2.2e-16
```

```
summary(reduced2)
```

```
##
## Call:
## lm(formula = Salary ~ USG. + TRB. + MP, data = nba2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -11268700  -3782080  -1260720   2640140  19346386
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -2687970.4  1011737.6  -2.657 0.008160 **
## USG.          128932.1    44977.5   2.867 0.004337 **
## TRB.          192033.8    50271.1   3.820 0.000152 ***
## MP              3840.6      321.7  11.938  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5407000 on 465 degrees of freedom
## Multiple R-squared:  0.2881, Adjusted R-squared:  0.2835
## F-statistic: 62.73 on 3 and 465 DF,  p-value: < 2.2e-16
```

Is reduced model better or same then full?

The hypotheses for the F-test performed in the anova are:

H0: The additional terms in the full model are 0.
HA: At least one of the additional terms is non 0.
Yes: p val = 0.9788

```
anova(reduced3,full3)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(Salary) ~ AST. + USG. + TRB. + MP
## Model 2: sqrt(Salary) ~ MP + TS. + X3PAr + FTr + TRB. + AST. + STL. +
##     BLK. + TOV. + USG.
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    478 563217138
## 2    472 561841090  6   1376048 0.1927 0.9788
```