STAT 535 - Bayesian Regression Project

By Rocky Rowell

Overview

This project tests what statistics/variables best explain winning percentages for NBA teams. The data used for this project consists of traditional and advanced season average statistics for all 30 NBA teams over the span of 4 seasons (20-21 through 23-24) making a sample size of $n = 120$. Not all variables are used

Variable Selection

To start off this regression project, I picked 14 variables that I thought would have a valuable impact towards winning percentage. Those variables are:

> Points, Defensive Rebounds, Offensive Rebounds, Blocks, Steals, Assists, Turnovers, Personal Fouls, Offensive Rating, Defensive Rating, Field Goal Percentage, Three Point Percentage, Free Throw Percentage and Pace

Prior

Utilizing the "stan_glm" function, I specified a prior on the intercept as well as the coefficients of all variables included. These priors were based off of my personal beliefs and understanding of basketball, however I gave all the priors a variance of $\sigma = 5$ to be non-informative to an extent. These priors are used constantly through the whole project and never change despite multiple regressions being done in the code. Those priors are:

Intercept ~ $N(0, 15)$
Points ~ $N(2, 5)$
Defensive Rebounds ~ $N(1, 5)$
Offensive Rebounds ~ $N(1, 5)$
Blocks ~ $N(1, 5)$
Steals ~ $N(1, 5)$
Assists ~ $N(1, 5)$
Turnovers ~ $N(-1, 5)$
Personal Fouls ~ $N(-1, 5)$
Offensive Rating ~ $N(2, 5)$
Defensive Rating ~ $N(2, 5)$
Field Goal Percentage ~ $N(2, 5)$
Three Point Percentage ~ $N(2, 5)$
Free Throw Percentage ~ $N(1, 5)$
Pace ~ $N(1, 5)$

Distribution

The distribution for this regression is normal (or gaussian as stated in the stan_glm function). Despite most of the statistics used typically being considered count data, since they are averages across the season they become discrete. Some variables are percentages and therefore have a support from 0 to 1 or 0 to 100, however normal regression is still used here regardless

Model Selection Process

I had an extensive model selection process with many steps:

1. My first step was to run regression on the all predictor model to get results. This regression was done using "stan_glm" (as well as all others) with the priors and distribution mentioned before. The results of the 80% confidence interval and beta coefficients showed that only Offensive Rating, Defensive Rating and Blocks are significant (***see image 1***)

2. My second step was to test if any predictor variables had an unclear relationship to the response variable (win percentage) in a basic plot and make models to test based off the information
    - Four variables had unclear trends in the plots: Offensive Rebounds, Blocks, Personal Fouls and Free Throw Percentage
    - However, Blocks was a significant variable with an 80% confidence interval in regression for the model containing all predictors (***see image 1***)
    - My 3 models to test:

    Model 1 - All Predictors
    Model 2 - All Predictors except OREB, BLK, PF and FT%
    Model 3 - All Predictor except OREB, PF and FT%

    - Testing the ELPD values of these 3 models (***see image 5***), we get:

    Model 1 ELPD = 201.12
    Model 2 ELPD = 203.43
    Model 3 ELPD = 203.97

- This shows that Model 3 is the best model here and we should continue with that being our current model
- Running the regression for this model, both the 95% and 80% confidence intervals contain more significant variables (***see image 2***) but there were still more issues to address to slim down the model.

3. The third step was to address the correlation issue. There was clear strong correlation between some of the predictors (***see image 3***). The most concerning correlations were:

   Points and Offensive Rating | Points and Field Goal Percentage
   Offensive Rating and Field Goal Percentage | Three Point Percentage

   - To address the correlation issues, I created multiple models and tested their ELPD values against the current model to determine the best model available:

   Model 1 - No Points
   Model 2 - No Field Goal Percentage
   Model 3 - No Points and Field Goal Percentage
   Model 4 - No Offensive Rating
   Model 5 - No Three Point Percentage

   - All ELPD values (***see image 6***) are:

   Current Model ELPD = 203.97
   Model 1 ELPD = 204.91
   Model 2 ELPD = 202.1
   Model 3 ELPD = 203.18
   Model 4 ELPD = 201.42
   Model 5 ELPD = 204.09

   - This shows that the best model available is Model 1 (no points) and it is now our current model.
   - Looking at the regression for this model, the result is the same significant variables as the last model (***see image 4***)

4. Lastly, I tested ELPD values of 3 models to determine the final model. These models are based off the significant predictors from the 95% and 80% confidence intervals of the current model (***see image 4***)

   Model 1 - 95% CI (OFFRTG, DEFRTG, FG%)
   Model 2 - 80% CI (DREB, BLK, OFFRTG, DEFRTG, FG%, 3P%)

- Looking at these ELPD values (***see image 7***), we see:

  Current Model ELPD = 204.91
  Model 1 ELPD = 206.52
  Model 2 ELPD = 206.57

- The difference between Model 1 and 2 is minimal but there is a clear difference between them and the current model. Therefore, Model 2 with 6 variables is the best available.

5. I also attempted using a gibbs sampler to go through all possible models but it did not end up being very helpful due to the immense amount of models. All the models had a similar probability and most at the top had no intercept which isn't possible. Therefore, the model found in part 4 is the final model

6. Lastly, I performed MCMC checks on the final model.

   Trace Plot (***see image 8***):
   - The trace plots have the desired shape for all variables
   - There are concerning spikes but they are short

   Density Overlay (***see image 9***):
   - All of the density overlays appear good and accurate

   ACF Plot (***see image 10***):
   - Little to no autocorrelation after the initial stretch in these plots

Results

The most significant variables (of the ones used) for predicting win percentage of an NBA team consists of: Defensive Rebound, Blocks, Offensive Rating, Defensive Rating, Field Goal Percentage, and Three Point Percentage.

Of these variables, Blocks and Defensive Rating are the only predictors with negative coefficients. Defensive Rating makes sense because a higher rating implies a worse defense, but Blocks having a negative effect is a surprise and seems counter-intuitive.

Offensive Rating and Defensive Rating by far have the strongest effect on win percentage (based on size of coefficient) and ended up being significant in every model tested which makes sense due to their strong relationships to points and points allowed for each game. Blocks also has one of the bigger effects while Defensive Rebounds, Field Goal Percentage and Three Point Percentage .

The final model (***see image 11***) for this project:

$$Y = 0.287 + 0.002X_1 - 0.01X_2 + 0.025X_3 - 0.027X_4 + 0.006X_5 + 0.004X_6$$

Where $X_1 = DREB,\ X_2 = BLK,\ X_3 = OFFRTG,\ X_4 = DEFRTG,\ X_5 = FG\%,\ X_6 = 3P\%$

This model is relatively good at predicting teams within the years of this data as well as before the data was taken. Those predictions are:

23-24 Houston Rockets: (in the data set)
Predicted: WIN% = **0.5521** which is **42.81** games
Actual: WIN% = **0.5** which is **41** games

23-24 Indiana Pacers: (in the data set)
Predicted: WIN% = **0.582** which is **47.72**
Actual: WIN % = **0.573** which is **47** games

18-19 Toronto Raptors: (2 years prior to data set)
Predicted: WIN% = **0.667** which is **54.73** games
Actual: WIN% = **0.707** which is **58** games

97-98 Chicago Bulls: (23 years prior to data set)
Predicted: WIN% = **0.701** which is **57.51** games
Actual: WIN% = **0.756** which is **62** games

Conclusion

The best variables for predicting an NBA team's winning percentage are Defensive Rebounds (positive), Blocks (negative), Offensive Rating (positive),

Defensive Rating (negative), Field Goal Percentage (Positive) and Three Point Percentage (positive).

Appendix

All Abbreviations:

PTS = Points | DREB = Defensive Rebounds | OREB = Offensive Rebounds | BLK = Blocks | STL = Steals | AST = Assists | TOV = Turnovers | PF = Personal Fouls | OFFRTG = Offensive Rating | DEFRTG = Defensive Rating | FG_PER or FG% = Field Goal Percentage | THREES_PER or 3P% = Three Point Percentage | FT_PER or FT% = Free Throw Percentage | PACE = Pace

Images:

Image 1

```
> tidy(nba_model, conf.int = TRUE, conf.level = 0.8)
# A tibble: 15 × 5
   term          estimate std.error  conf.low conf.high
   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)  0.335       1.18    -1.19       1.85
 2 PTS         -0.0000716   0.00992 -0.0128     0.0128
 3 DREB         0.00540     0.00528 -0.00164    0.0123
 4 OREB        -0.00305     0.00538 -0.00998    0.00384
 5 BLK         -0.00998     0.00575 -0.0173    -0.00247
 6 STL          0.0101      0.00852 -0.000821   0.0210
 7 AST         -0.00155     0.00284 -0.00518    0.00213
 8 TOV         -0.00344     0.00527 -0.0103     0.00338
 9 PF          -0.00418     0.00405 -0.00927    0.00105
10 OFFRTG       0.0250      0.0102   0.0116     0.0380
11 DEFRTG      -0.0245      0.00289 -0.0283    -0.0209
12 FG_PER       0.00718     0.00617 -0.000736   0.0151
13 THREES_PER   0.00444     0.00425 -0.00104    0.00975
14 FT_PER      -0.00106     0.00175 -0.00330    0.00121
15 PACE        -0.00283     0.0115  -0.0179     0.0121
```

Image 2

```
> tidy(nba_model.3, conf.int = TRUE, conf.level = 0.95)
# A tibble: 12 × 5
   term        estimate std.error conf.low conf.high
   <chr>          <dbl>     <dbl>    <dbl>     <dbl>
 1 (Intercept) -0.0580    1.14    -2.29      2.18
 2 PTS         -0.00203   0.00984 -0.0213    0.0170
 3 DREB         0.00761   0.00459 -0.00144   0.0166
 4 BLK         -0.00955   0.00567 -0.0207    0.00162
 5 STL          0.00961   0.00791 -0.00605   0.0252
 6 AST         -0.00120   0.00276 -0.00661   0.00416
 7 TOV         -0.00564   0.00453 -0.0148    0.00333
 8 OFFRTG       0.0249    0.0102   0.00513   0.0449
 9 DEFRTG      -0.0239    0.00264 -0.0292   -0.0187
10 FG_PER       0.0106    0.00464  0.00143   0.0197
11 THREES_PER   0.00476   0.00361 -0.00238   0.0119
12 PACE        -0.00174   0.0114  -0.0239    0.0206
> tidy(nba_model.3, conf.int = TRUE, conf.level = 0.8)
# A tibble: 12 × 5
   term        estimate std.error   conf.low conf.high
   <chr>          <dbl>     <dbl>      <dbl>     <dbl>
 1 (Intercept) -0.0580    1.14    -1.51        1.39
 2 PTS         -0.00203   0.00984 -0.0147      0.0105
 3 DREB         0.00761   0.00459  0.00175     0.0135
 4 BLK         -0.00955   0.00567 -0.0168     -0.00231
 5 STL          0.00961   0.00791 -0.000676    0.0196
 6 AST         -0.00120   0.00276 -0.00476     0.00230
 7 TOV         -0.00564   0.00453 -0.0115      0.000170
 8 OFFRTG       0.0249    0.0102   0.0120      0.0381
 9 DEFRTG      -0.0239    0.00264 -0.0273     -0.0205
10 FG_PER       0.0106    0.00464  0.00463     0.0166
11 THREES_PER   0.00476   0.00361  0.0000976   0.00935
12 PACE        -0.00174   0.0114  -0.0164      0.0127
```

Image 3

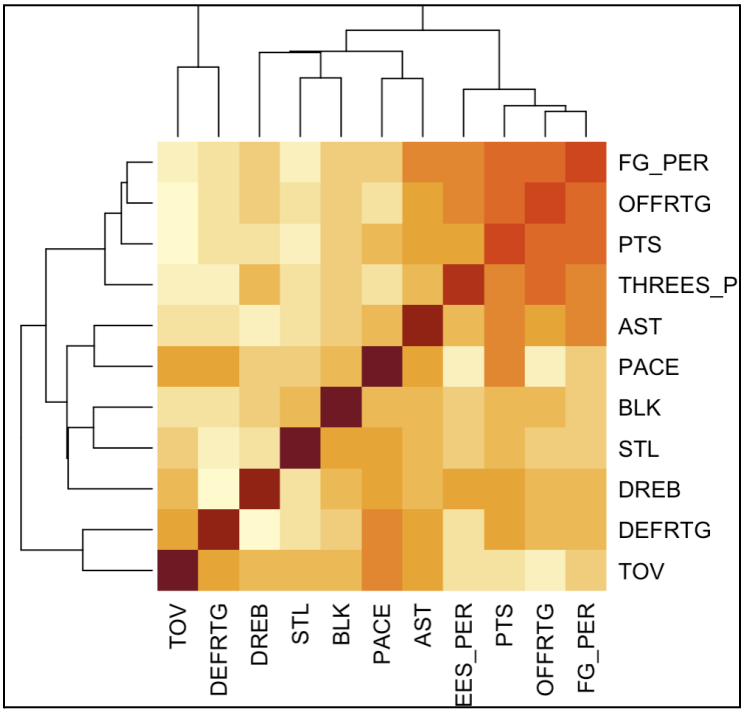

Image 4

```
> tidy(nba_model.no_pts, conf.int = TRUE, conf.level = 0.95)
# A tibble: 11 × 5
   term        estimate std.error conf.low conf.high
   <chr>          <dbl>     <dbl>    <dbl>     <dbl>
 1 (Intercept)  0.169     0.380   -0.580     0.914
 2 DREB         0.00768   0.00455 -0.00134   0.0166
 3 BLK         -0.00953   0.00567 -0.0208    0.00164
 4 STL          0.00967   0.00795 -0.00587   0.0254
 5 AST         -0.00117   0.00276 -0.00658   0.00435
 6 TOV         -0.00578   0.00450 -0.0146    0.00300
 7 OFFRTG       0.0229    0.00219  0.0186    0.0272
 8 DEFRTG      -0.0239    0.00264 -0.0291   -0.0187
 9 FG_PER       0.0105    0.00457  0.00146   0.0196
10 THREES_PER   0.00495   0.00370 -0.00216   0.0120
11 PACE        -0.00408   0.00345 -0.0107    0.00257
> tidy(nba_model.no_pts, conf.int = TRUE, conf.level = 0.8)
# A tibble: 11 × 5
   term        estimate std.error  conf.low  conf.high
   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
 1 (Intercept)  0.169     0.380   -0.319     0.653
 2 DREB         0.00768   0.00455  0.00176   0.0135
 3 BLK         -0.00953   0.00567 -0.0168   -0.00228
 4 STL          0.00967   0.00795 -0.000558  0.0198
 5 AST         -0.00117   0.00276 -0.00470   0.00237
 6 TOV         -0.00578   0.00450 -0.0114    0.0000183
 7 OFFRTG       0.0229    0.00219  0.0201    0.0257
 8 DEFRTG      -0.0239    0.00264 -0.0272   -0.0205
 9 FG_PER       0.0105    0.00457  0.00463   0.0164
10 THREES_PER   0.00495   0.00370  0.000196  0.00961
11 PACE        -0.00408   0.00345 -0.00845   0.000350
```

Image 5

Model 1

	Estimate	SE
elpd_loo	201.11747	7.756636
p_loo	15.34136	1.935285
looic	-402.23494	15.513272

Model 2

	Estimate	SE
elpd_loo	203.43321	8.064443
p_loo	11.33626	1.566554
looic	-406.86641	16.128886

Model 3

	Estimate	SE
elpd_loo	203.96689	7.945430
p_loo	12.19391	1.609104
looic	-407.93378	15.890860

Image 6

Current Model

	Estimate	SE
elpd_loo	203.96689	7.945430
p_loo	12.19391	1.609104
looic	-407.93378	15.890860

Model 1

	Estimate	SE
elpd_loo	204.91172	7.977884
p_loo	11.43071	1.566116
looic	-409.82344	15.955769

Model 2

	Estimate	SE
elpd_loo	202.10308	7.633925
p_loo	11.23561	1.482854
looic	-404.20616	15.267849

Model 3

	Estimate	SE
elpd_loo	203.18060	7.695557
p_loo	10.44711	1.434629
looic	-406.36119	15.391115

Model 4

	Estimate	SE
elpd_loo	201.41593	7.922206
p_loo	11.52665	1.534147
looic	-402.83186	15.844413

Model 5

	Estimate	SE
elpd_loo	204.08597	7.879051
p_loo	11.33657	1.535737
looic	-408.17194	15.758103

Image 7

Model 1

	Estimate	SE
elpd_loo	206.515327	8.0765847
p_loo	5.267298	0.8746655
looic	-413.030655	16.1531695

Model 2

	Estimate	SE
elpd_loo	206.568463	7.562575
p_loo	7.851342	1.117860
looic	-413.136927	15.125150

Image 8

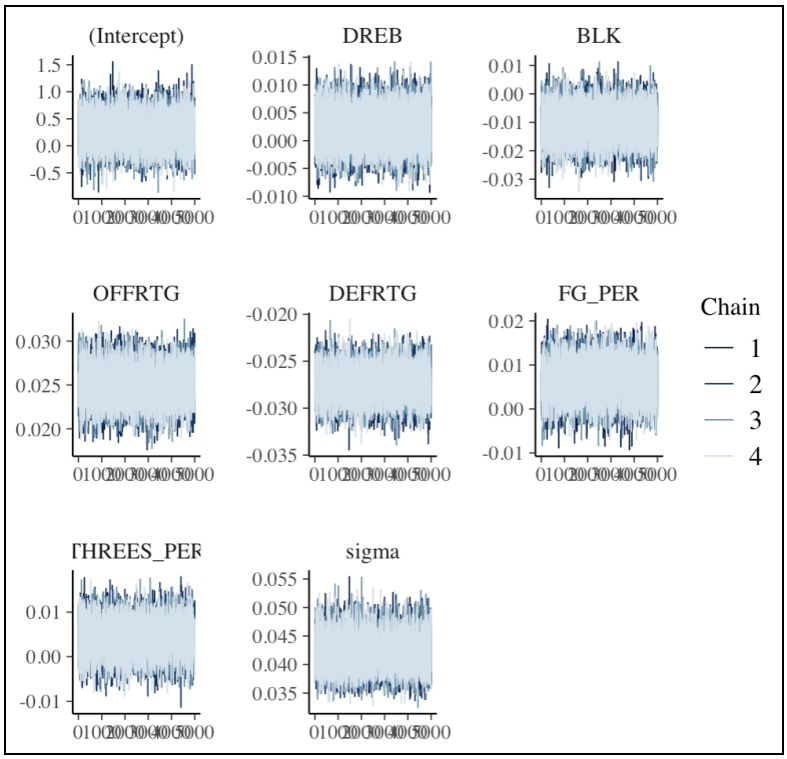
(Intercept)
DREB
BLK
OFFRTG
DEFRTG
FG_PER
THREES_PER
sigma
Chain
1
2
3
4

Image 9

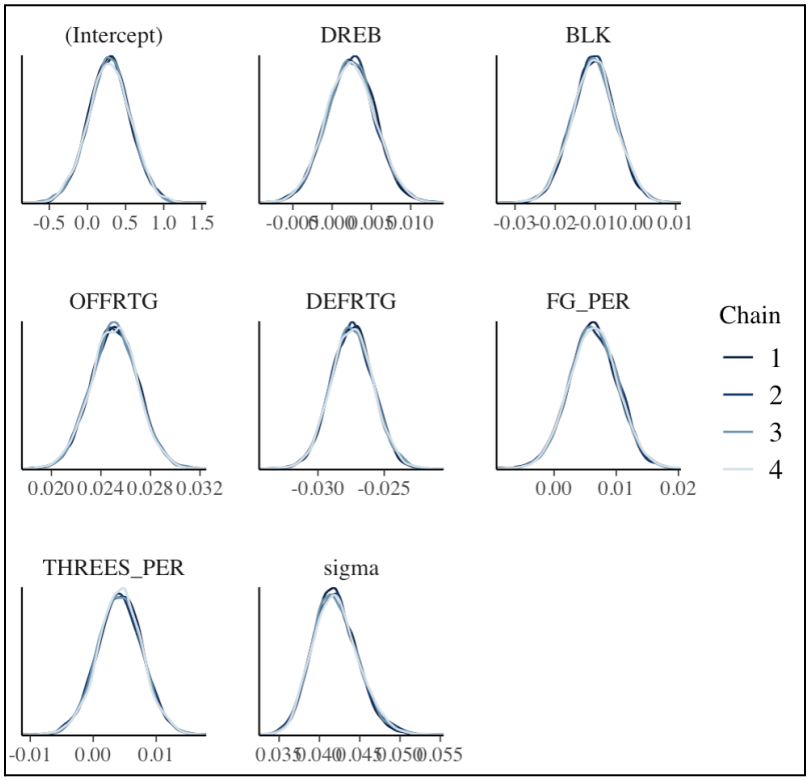
(Intercept)
DREB
BLK
OFFRTG
DEFRTG
FG_PER
THREES_PER
sigma
Chain
1
2
3
4

Image 10

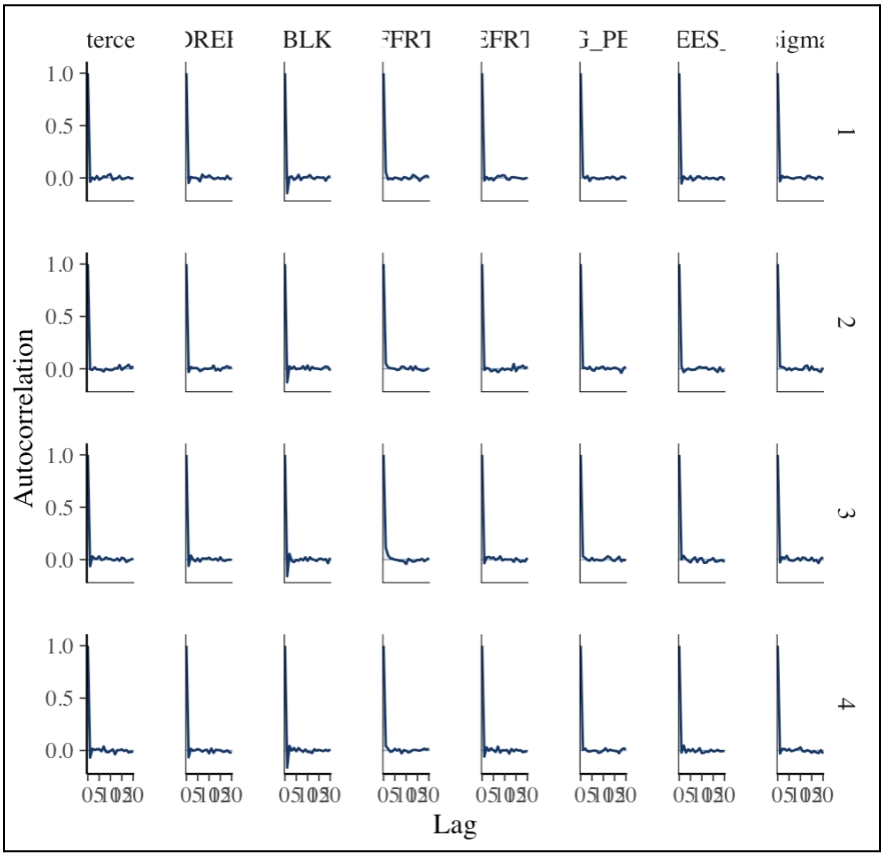

Image 11

```
> tidy(nba_model.DREB_BLK_OFFRTG_DEFRTG_FG_THREE, conf.int = TRUE, conf.level = 0.8)
# A tibble: 7 × 5
  term        estimate std.error  conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)  0.287     0.273   -0.0668     0.642
2 DREB         0.00239   0.00311 -0.00165    0.00638
3 BLK         -0.0104    0.00542 -0.0176    -0.00332
4 OFFRTG       0.0251    0.00190  0.0226     0.0275
5 DEFRTG      -0.0274    0.00160 -0.0295    -0.0254
6 FG_PER       0.00629   0.00389  0.00135    0.0113
7 THREES_PER   0.00424   0.00354 -0.000332   0.00876
> tidy(nba_model.DREB_BLK_OFFRTG_DEFRTG_FG_THREE, conf.int = TRUE, conf.level = 0.95)
# A tibble: 7 × 5
  term        estimate std.error conf.low conf.high
  <chr>          <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)  0.287     0.273   -0.265     0.836
2 DREB         0.00239   0.00311 -0.00376   0.00855
3 BLK         -0.0104    0.00542 -0.0212    0.000449
4 OFFRTG       0.0251    0.00190  0.0213    0.0288
5 DEFRTG      -0.0274    0.00160 -0.0306   -0.0243
6 FG_PER       0.00629   0.00389 -0.00133   0.0139
7 THREES_PER   0.00424   0.00354 -0.00291   0.0113
```