The Four Factors: Statistics that Make College Basketball Teams Win

Jake Singleton

11/12/2020

Motivation: Dean Oliver and the NBA

In the early 2000s, statistician Dean Oliver published his findings on the "four-factor model," a system used to evaluate the performance of NBA teams. He found that the four factors do a very nice job of revealing teams' strengths and weaknesses in addition to being good predictors of wins. They are as follows:

1. Effective Field Goal Percentage (EFG)
2. Turnovers Committed per Possession (TPP)
3. Offensive Rebounding Percentage (ORP)
4. Free Throw Rate (FTR)

EFG is calculated as (all field goals made + 0.5(3-point field goals made)) / (all field goal attempts). The great thing about EFG is that it weights 3-pointers appropriately, as 3-pointers are 50% more valuable than 2-pointers. Thus, it captures a team's shooting ability far better than standard FG%.

TPP is simply (turnovers committed) / (possessions). Smaller TPP is better, as it means a team gives up the ball less often.

ORP is (offensive rebounds) / (offensive rebounds + opponent's defensive rebounds). Obviously, high ORP is good and means the team gets second chances at making baskets.

FTR is (free throw attempts) / (field goal attempts) and measures a team's propensity to get to the free throw line (a good thing)!

Note that all of the numbers needed to calculate the four factors can be found in the box score. (Kind of... possessions are not in the box score, but analysts have come up with good formulas that estimate the number of possessions in a game very closely.) Not only does this makes our task easier, but it's also interesting that a team's win total can be explained so well by simple box score statistics (obviously these are modifications of box score stats, but they're nothing fancy). While Oliver called these statistics the four factors, he calculated them from a defensive point of view as well, yielding eight factors. The defensive factors are:

1. Opponent's Effective Field Goal Percentage (OPP_EFG), i.e. the EFG a team yields to their opponents. Lower is better.
2. Defensive Turnovers Caused per Possession (DTPP). Higher is better.
3. Defensive Rebounding Percentage (DRP). Higher is better.
4. Opponent's Free Throw Rate (OPP_FTR). Lower is better.

These are computed analogously to their offensive counterparts.

So, how important are these factors? Oliver's analysis led to the following rankings:

1. Shooting differential, EFG - OPP_EFG (40% importance)
2. Turnover differential, TPP - DTPP (25%)
3. Rebounding differential, ORP - DPP (20%)
4. Free throw rate differential, FTR - OPP_FTR (15%)

Note that other analysis online (e.g. here) and in Chapter 28 of Wayne Winston's *Mathletics* has shown Oliver's rankings overvalue free throw rate and rebounding and undervaue turnovers and shooting. Specifically, Winston found that EFG_diff explains 71% of variance in NBA team wins, TO_diff 15%, REB_diff 6%, and FTR_diff essentially 0%. I should also remind you that Oliver's work was done in the early-mid 2000s before the NBA's "3-point revolution" of the mid-late 2010s, and so I believe EFG_diff is absolutely more valuable now than what it was.

In this analysis, I will analyze college basketball data from the 2019-2020 season to see how the college game differs from the pro game.

The Data

There are a lot of Division 1 college basketball games in a season–too many, in fact, to go through every single one. So, I looked at the 10 "high-major" conferences: the Power 5 plus the Big East, Mountain West, Atlantic 10, American Athletic, and West Coast Conferences. From each conference, I took the champion, a middling team, and bottom of the barrel, yielding a total of 30 teams. These teams are: (Big12: Kansas, Texas Tech, KState), (WCC: Gonzaga, Pepperdine, Portland), (A10: Dayton, UMass, Fordham), (ACC: FSU, Syracuse, Wake Forest), (MountainW: SDSU, Colorado State, Wyoming), (SEC: Kentucky, A&M, Vandy), (Pac12: Oregon, Colorado, Washington), (Big10: Wisconsin, Penn State, Nebraska), (BigE: Creighton, Marquette, DePaul), and (AAC: Houston, SMU, Tulane). I employed this approach in order to include a wide variety of team talent. I limited the games scraped for each team to 30 games, yielding a total of 900 games (there was no overlap). All data comes from the 19-20 season.

The Code

To scrape the relevant data, they key part was collecting all of the necessary ESPN game IDs. For example, California's first game of the season was against Pepperdine, and the box score is at this link: https://www.espn.com/mens-college-basketball/matchup?gameId=401170553. The last 9 digits of the link comprise the game ID, and I had to collect each one manually. If anyone knows a better way of getting these IDs, please let me know.

I'm not going to show all of the code since it's pretty long, but I will display all of the important outputs. You can find the R Markdown file in its full form in the code directory of this repository.

Team	FGM	3FGM	FGA	FTA	FTM	TO	OR	DR	OPP_FGM	OPP_3FGM
KU	23	4	50	26	16	28	10	30	23	8
DUKE	23	8	64	23	14	16	11	19	23	4
MONM	18	4	60	23	17	15	11	23	37	14
KU	37	14	66	31	24	7	11	32	18	4
ETSU	23	9	60	15	8	16	11	22	30	1
KU	30	1	54	18	14	15	5	27	23	9
KU	38	12	69	11	5	14	13	20	23	7
CHAM	23	7	50	14	10	27	6	22	38	12
KU	29	4	61	15	9	12	9	26	22	9
BYU	22	9	54	4	3	20	5	27	29	4

The above table shows a slice of the relevant box score statistics for each game that we'll use to calculate

the four factors. For example, the first two rows represent the Kansas vs. Duke game, one of the first of the season.

```
# Group data by team and aggregate (we use sum of course)
grouped = all_game_df %>%
  group_by(Team) %>%
  summarize(across(.cols = everything(), sum)) %>%
  filter(Team %in% teams_of_interest)

kable(grouped[1:5, 1:11], "simple") %>%
  kable_styling(font_size = 5)
```

Team	FGM	3FGM	FGA	FTA	FTM	TO	OR	DR	OPP_FGM	OPP_3FGM
COLO	890	272	2070	732	534	501	384	997	882	230
CREI	1038	357	2189	658	485	416	285	988	977	286
CSU	921	272	2017	790	535	479	333	969	937	320
DAY	1013	284	1962	656	469	431	271	947	828	220
DEP	891	190	2087	709	472	504	412	870	854	289

The code above aggregates the data for each team of interest by summing each column.

We see that this data frame includes each of the box score statistics necessary to calculate our factors. For example, Colorado made 890 field goals out of 2,070, and they gave up 882 successful field goals to their opponents. 272 of their 890 were 3-point field goals, and of their opponents' 882 makes, 230 were 3-pointers.

Team	Wins	EFG	TPP	ORP	FTR	OPP_EFG	DTPP	DRP	OPP_FTR
COLO	19	0.4956522	0.1996748	0.3089300	0.3536232	0.4707271	0.1914657	0.7681048	0.2615675
CREI	23	0.5557332	0.1594163	0.2307692	0.3005939	0.4867449	0.1801892	0.7052106	0.2338114
CSU	18	0.5240456	0.1907910	0.2709520	0.3916708	0.5172089	0.1781407	0.7708831	0.2555398
DAY	28	0.5886850	0.1787907	0.2563860	0.3343527	0.4685315	0.1864669	0.7363919	0.3111888
DEP	14	0.4724485	0.2023316	0.3083832	0.3397221	0.5002505	0.2228732	0.7044534	0.3662325
FOR	8	0.4481567	0.1889301	0.2049252	0.2523041	0.4687500	0.2076005	0.7317073	0.3349057
FSU	25	0.5179688	0.1908363	0.3409506	0.2968750	0.4690165	0.2372893	0.7013575	0.3547470
GONZ	28	0.5608169	0.1589040	0.3322581	0.3720826	0.4775684	0.1845429	0.7667660	0.2135487
HOU	22	0.4802299	0.1776675	0.3884298	0.3181609	0.4516466	0.1864761	0.7468750	0.4082593
KSU	9	0.4730954	0.2208161	0.3009868	0.3622802	0.5055188	0.2442662	0.6768402	0.3907285
KU	27	0.5388016	0.1911158	0.3307888	0.3595285	0.4441540	0.1932620	0.7360294	0.2433460
MARQ	18	0.5099432	0.1934985	0.2971071	0.3787879	0.4704277	0.1511445	0.7427356	0.3557780
MASS	13	0.4820391	0.1917171	0.2566984	0.2958267	0.5119617	0.2011746	0.7051971	0.3269537
NEB	6	0.4676734	0.1617666	0.2190408	0.2840963	0.5290308	0.1868190	0.6771714	0.2081286
ORE	23	0.5310395	0.1716237	0.3227158	0.2916266	0.4749504	0.2004339	0.6822581	0.2986111
PEPP	14	0.4939701	0.1766483	0.2690763	0.3174144	0.5245902	0.1921985	0.7278912	0.3378043
PORT	7	0.4859890	0.2015189	0.2021371	0.3214286	0.5123177	0.1834585	0.6889265	0.3092006
PSU	20	0.4891866	0.1608479	0.2991648	0.3055947	0.4599208	0.1850407	0.7329240	0.3275606
SDSU	28	0.5529563	0.1662649	0.2763401	0.2915167	0.4548668	0.2135021	0.7493671	0.3023382
SMU	19	0.5232849	0.1989808	0.3375959	0.3174762	0.4967549	0.1898454	0.7096513	0.2930604
SYR	16	0.4940599	0.1628018	0.2888147	0.3610537	0.4858174	0.2062157	0.6585564	0.3187210
TA&M	16	0.4647603	0.2165732	0.3091231	0.3939919	0.4933993	0.2270247	0.6803939	0.2634763
TTU	17	0.5084459	0.1984626	0.2800659	0.3373552	0.4713217	0.2301591	0.7037331	0.3680798
TULN	12	0.4822294	0.1745032	0.2199832	0.3521809	0.5289234	0.2281850	0.6883593	0.2945903
UK	24	0.5082579	0.1846714	0.3059490	0.3969100	0.4598485	0.1724378	0.7147360	0.3242424
VAN	9	0.4918033	0.1857368	0.2504078	0.4093072	0.5309917	0.1907720	0.7053010	0.3631198

Team	Wins	EFG	TPP	ORP	FTR	OPP_EFG	DTPP	DRP	OPP_FTR
WAKE	12	0.4759132	0.2013155	0.2864322	0.4060349	0.4940060	0.1545006	0.7302100	0.3491508
WASH	13	0.5050829	0.2180559	0.2722273	0.3777421	0.4447381	0.1869748	0.6746988	0.3123498
WISC	20	0.5121053	0.1563216	0.2390543	0.2552632	0.4641736	0.1692777	0.7495798	0.2573222
WYO	7	0.4794953	0.1870324	0.1443806	0.3059937	0.5133367	0.1925503	0.7047540	0.3865123

These are our four factors for each team (or eight, if you like). Note we have each team's number of wins now, which will come into play soon as our dependent variable.

Team	Wins	EFG	TPP	ORP	FTR	OPP_EFG	DTPP	DRP	OPP_FTR
DAY	28	1	12	22	16	8	21	8	14
GONZ	28	2	2	4	8	15	23	3	2
SDSU	28	4	7	17	27	4	7	5	12
KU	27	5	19	5	11	1	12	9	4
FSU	25	9	18	2	24	10	2	22	23
UK	24	13	13	10	3	5	27	14	17
CREI	23	3	3	25	23	17	25	17	3
ORE	23	6	8	6	26	14	11	25	11
HOU	22	23	11	1	18	3	20	6	30
PSU	20	19	4	12	22	6	22	10	19
WISC	20	10	1	24	29	7	28	4	6
COLO	19	15	24	8	12	12	15	2	7
SMU	19	8	23	3	19	20	17	15	9
CSU	18	7	17	19	5	26	26	1	5
MARQ	18	11	21	13	6	11	30	7	24
TTU	17	12	22	16	15	13	3	21	27
SYR	16	16	6	14	10	16	9	30	16
TA&M	16	29	28	7	4	18	5	26	8
DEP	14	27	27	9	14	21	6	20	26
PEPP	14	17	10	20	20	27	14	13	21
MASS	13	22	20	21	25	23	10	18	18
WASH	13	14	29	18	7	2	18	29	15
TULN	12	21	9	26	13	28	4	24	10
WAKE	12	25	25	15	2	19	29	12	22
KSU	9	26	30	11	9	22	1	28	29
VAN	9	18	14	23	1	30	16	16	25
FOR	8	30	16	28	30	9	8	11	20
PORT	7	20	26	29	17	24	24	23	13
WYO	7	24	15	30	21	25	13	19	28
NEB	6	28	5	27	28	29	19	27	1

This data frame is a table of rankings. For example, Dayton was 1st among our 30 teams in EFG, 12th in TPP, 22nd in ORP, 16th in FTR, etc.

The Analysis

Interestingly, we see the top 3 teams in wins, Dayton, Gonzaga, and SDSU, are 1st, 2nd, and 4th respectively in EFG ranking. Moreover, 7 of the top 10 teams in wins are top 10 in EFG, and 6 of the top 10 rank inside the top 10 in OPP_EFG. This supports the NBA analysis that has found shooting is the most important factor. If we look at the bottom six teams or so , not only do we see low ranks for EFG, but we also see the

same for ORP. Note that Fordham, Portland, Wyoming, and Nebraska all shot poorly and didn't get second chances either. Fordham is also interesting because they were last in EFG and FTR, yet they played good defense, ranking 9th and 8th in OPP_EFG and DTPP respectively. As for a couple teams in the middle, SMU and Colorado State were good shooting teams (8th and 7th respectively), but poor defending shooters (20th and 26th respectively). Further, they were in the bottom half in turnovers committed on offense (23rd and 17th). Another team I want to point out is Washington, the most unlucky team in the nation last year. They were 14th in EFG (not bad), 7th in FTR, and 2nd(!) in OPP_EFG, but they finished last in the PAC-12 and won only 13 games. Note that they were 29th in TPP, 18th in ORP, and 29th in DRP, though, meaning they turned the ball over a lot and were not a good rebounding team.

Now, we will dig into some of the good statistics stuff, emulating Wayne Winston's analysis from his fantastic book *Mathletics*. We will run a regression to see which factors most influence wins. Our explanatory variables are:

1. Shooting differential: EFG - OPP_EFG
2. Turnover differential: TPP - DTPP
3. Rebounding differential: ORP - DRP
4. Free throw differential: FTR - OPP_FTR

Our response variable is wins.

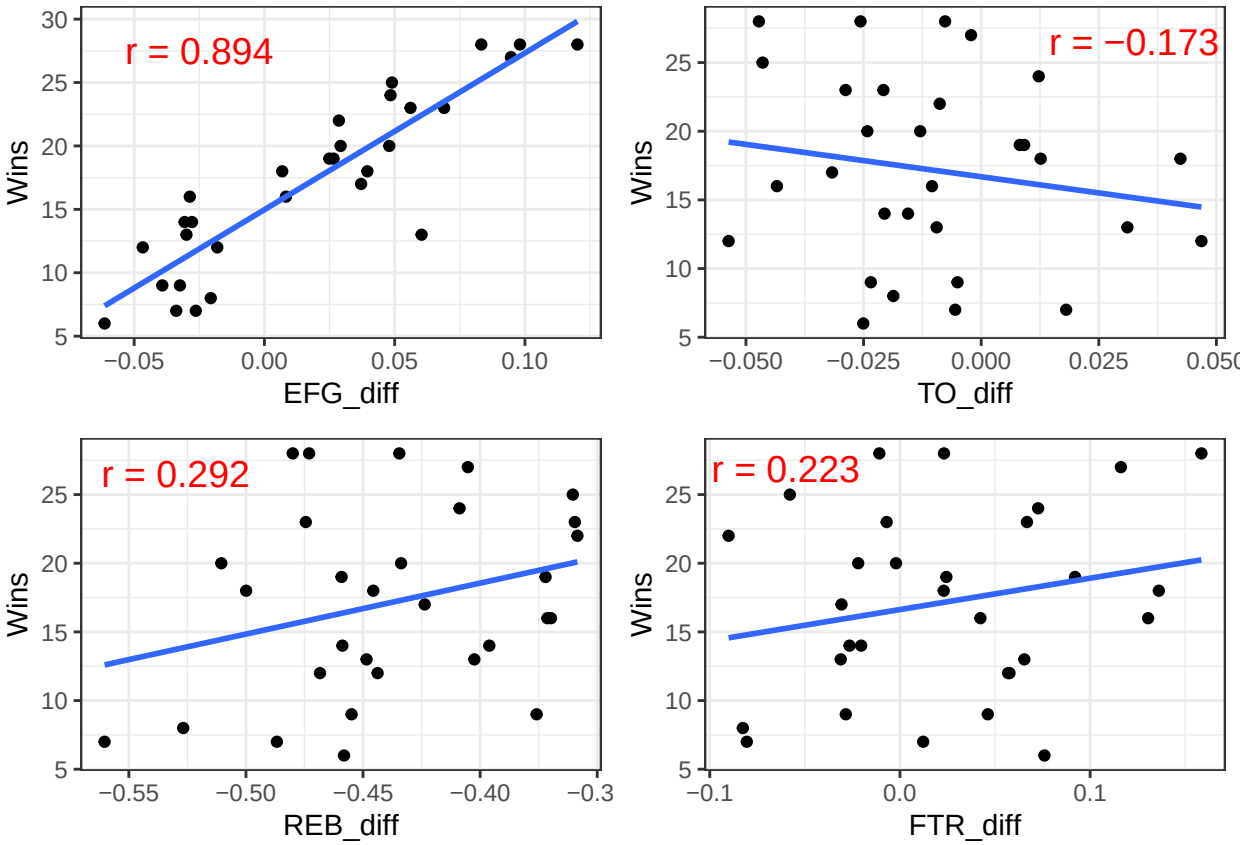
```
# Prepare to run regression by making appropriate variables
deviation_tab = four_factors_df %>%
  mutate(EFG_diff = EFG - OPP_EFG,
         TO_diff = TPP - DTPP,
         REB_diff = ORP - DRP,
         FTR_diff = FTR - OPP_FTR) %>%
  select(Wins, EFG_diff, TO_diff, REB_diff, FTR_diff)

# Correlation matrix
kable(cor(deviation_tab))
```

	Wins	EFG_diff	TO_diff	REB_diff	FTR_diff
Wins	1.0000000	0.8937964	-0.1734100	0.2919753	0.2230980
EFG_diff	0.8937964	1.0000000	-0.0094522	0.1433813	0.1723719
TO_diff	-0.1734100	-0.0094522	1.0000000	-0.1242344	0.2499004
REB_diff	0.2919753	0.1433813	-0.1242344	1.0000000	0.0534000
FTR_diff	0.2230980	0.1723719	0.2499004	0.0534000	1.0000000

Here, note that our independent variables are not highly-correlated with each other. The highest correlation we see is between TO_diff and FTR_diff at 0.25. I postulate this is because teams that shoot a lot of free throws are "bigger" teams that draw more fouls, i.e. their lineups are taller than average, meaning they are worse ball-handlers and therefore commit more turnovers.

Statistically, the fact that these variables are relatively uncorrelated is a good thing, since the standard errors of our coefficient estimates will consequently be low. This means that our coefficient estimates will be reliable, allowing us to infer which variabels most influence wins with confidence. The problem that results when two explanatory varibles are highly correlated is called "collinearity," and here is a simple discussion of the problem that anyone can understand.

Before we run our regression, let's visualize each variable's relationship with wins via scatterplots. I include the correlation of each variable with wins, denoted by "r," and include regression lines.



The Wins vs. EFG_diff plot... textbook. Note the high correlation there as well. Furthermore, we see a negative correlation between wins and TO_diff–since an increase in TO_diff means committing more turners on offense, this should lead to a decrease in wins. So that makes sense. We see positive correlations in the REB_diff and FTR_diff plots, but they are much weaker than that of EFG_diff.

Great! We are now ready to regress wins on these four variables.

```
# Run regression on all four factors
four_factors_fit = lm(Wins ~ EFG_diff + TO_diff + REB_diff + FTR_diff, data = deviation_tab)
summary(four_factors_fit)
```

```
##
## Call:
## lm(formula = Wins ~ EFG_diff + TO_diff + REB_diff + FTR_diff,
##     data = deviation_tab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3228 -1.0203  0.5858  1.5347  3.5094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22.208      4.374   5.077 3.05e-05 ***
## EFG_diff     117.855     10.652  11.064 4.01e-11 ***
## TO_diff      -47.810     21.326  -2.242   0.0341 *
## REB_diff      18.049      9.750   1.851   0.0760 .
## FTR_diff      11.516      8.089   1.424   0.1669
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.742 on 25 degrees of freedom
## Multiple R-squared:  0.8589, Adjusted R-squared:  0.8363
## F-statistic: 38.04 on 4 and 25 DF,  p-value: 2.747e-10
```

There is a lot to notice here. By looking at R-squared, we see that 86% of the variability in wins can be explained by our four explanatory variables. That's pretty high. Next, we can learn a lot from the coefficients themselves. EFG_diff is highly significant (at the 0.1% level), while TO_diff is significant at the 5% level and REB_diff at the 10% level. FTR_diff is not significant. The coefficient estimates bear this out as well: in order of decreasing magnitude, we have EFG_diff (117.9), TO_diff (47.8), REB_diff (18.0), and FTR_diff (11.5). These results imply that EFG_diff is by far the most important factor in a college basketball team winning games.

Now we'll run a few more regressions. First, we remove FTR_diff since it was the one insignificant variable.

```
# Remove FTR_diff
summary(lm(Wins ~ EFG_diff + TO_diff + REB_diff, data = deviation_tab))
```

```
##
## Call:
## lm(formula = Wins ~ EFG_diff + TO_diff + REB_diff, data = deviation_tab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3106 -1.3277  0.4103  1.7482  4.2439
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22.902      4.431   5.168 2.16e-05 ***
## EFG_diff     120.438     10.702  11.254 1.71e-11 ***
## TO_diff      -39.881     20.989  -1.900   0.0686 .
## REB_diff      18.933      9.921   1.908   0.0674 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.796 on 26 degrees of freedom
## Multiple R-squared:  0.8475, Adjusted R-squared:  0.8299
## F-statistic: 48.15 on 3 and 26 DF,  p-value: 9.389e-11
```

Here, note that even when we exclude FTR_diff, 85% of the variability in wins is still explained by the other three factors as opposed to 86% we saw previously. This confirms our suspicion that FTR_diff is the least important factor in a team's success. Next, we see that the p-value for REB_diff (0.0674) is now lower than that for TO_diff (0.0686). This is the opposite of what we saw in the first regression, and I believe it's due to the correlation of 0.25 between FTR_diff and TO_diff, the highest between the four explanatory variables we saw earlier. This high correlation likely caused the coefficient for TO_diff to be a bit higher than it deserved in the first regression, and removing FTR_diff caused TO_diff's coefficient to stabilize.

The last four regressions we analyze will be single variable, i.e. we regress wins on each of the four factors individually.

```r
# Individual regressions. First EFG_diff.
summary(lm(Wins ~ EFG_diff, data = deviation_tab))
```

```
##
## Call:
## lm(formula = Wins ~ EFG_diff, data = deviation_tab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4311 -1.7145  0.5388  2.3920  4.5656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.9736     0.6018   24.88  < 2e-16 ***
## EFG_diff    123.5804    11.7184   10.55 2.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.093 on 28 degrees of freedom
## Multiple R-squared:  0.7989, Adjusted R-squared:  0.7917
## F-statistic: 111.2 on 1 and 28 DF,  p-value: 2.94e-11
```

EFG_diff alone explains 80% of variance in a team's wins and is highly significant.

```r
# Second TO_diff.
summary(lm(Wins ~ TO_diff, data = deviation_tab))
```

```
##
## Call:
## lm(formula = Wins ~ TO_diff, data = deviation_tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8659  -4.0112   0.3671   4.9427  10.9534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.685      1.344  12.416 6.64e-13 ***
## TO_diff      -47.148     50.603  -0.932    0.359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.793 on 28 degrees of freedom
## Multiple R-squared:  0.03007,    Adjusted R-squared:  -0.004569
## F-statistic: 0.8681 on 1 and 28 DF,  p-value: 0.3594
```

TO_diff alone explains only 3% of variation in wins and is insignificant.

```r
# Third REB_diff.
summary(lm(Wins ~ REB_diff, data = deviation_tab))
```

```
##
```

```
## Call:
## lm(formula = Wins ~ REB_diff, data = deviation_tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4510  -4.8715  -0.6337   4.5207  12.4168
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.41      10.13   3.299  0.00265 **
## REB_diff       37.14      22.99   1.615  0.11744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.597 on 28 degrees of freedom
## Multiple R-squared:  0.08525,    Adjusted R-squared:  0.05258
## F-statistic: 2.609 on 1 and 28 DF,  p-value: 0.1174
```

REB_diff alone explains only 9% of variation in wins and is insignificant.

```
# Last FTR_diff.
summary(lm(Wins ~ FTR_diff, data = deviation_tab))
```

```
##
## Call:
## lm(formula = Wins ~ FTR_diff, data = deviation_tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3608  -5.7236  -0.6605   5.4982  11.6206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.626      1.306  12.730 3.66e-13 ***
## FTR_diff      22.830     18.851   1.211    0.236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.724 on 28 degrees of freedom
## Multiple R-squared:  0.04977,    Adjusted R-squared:  0.01584
## F-statistic: 1.467 on 1 and 28 DF,  p-value: 0.236
```

As expected, FTR_diff is insignificant and explains only 5% of variance in wins.

Conclusion

To summarize, our regression of wins on all four factors shows we can predict wins with the formula: predicted wins = 22.208 + 117.855(EFG_diff) - 47.810(TO_diff) + 18.049(REB_diff) + 11.516(FTR_diff). However, if we had to select one of the four factors to predict or infer team success, it would undoubtedly have to be EFG_diff. Explicitly, good shooting and/or good defense against shooters is most valuable.

Given these results, let's quantify how college basketball teams can improve. By looking at the coefficient estimates, a 1% improvement in EFG_diff is worth 1.17 wins. This could happen through a team improving

their own EFG by 1%, improving their defense by 1% (decreasing their opponent's EFG by 1%), or a mixture of the two. Next, a 1% increase in in TO_diff is worth -0.47 wins. This make sense, since increasing TO_diff (turnovers committed per possession - defensive turnovers caused per possession by opponent) is obviously bad. Hence, committing 1 more turnover per 100 possessions or forcing 1 fewer turnover per 100 opponent posessions would lead to about 0.5 fewer wins. Next, a 1% increase in REB_diff (ORP - DRP) would lead to 0.18 more wins. This could mean grabbing 1 more offensive rebound per 100 missed shots or grabbing 1 more defensive rebound per 100 missed shots by opponent. Finally, a 1% increase in FTR_diff would lead to 0.12 more wins, which could occur through gaining 1 free throw attempt per 100 field goal attempts or conceding 1 less free throw attempt per 100 field goal attempts by opponent.

To compare to the NBA, recall that Winston found EFG_diff explains 71% of variance in NBA team wins, TO_diff 15%, REB_diff 6%, and FTR_diff essentially 0%. We found corresponding percentages of 80%, 3%, 9%, and 5%. This implies that FTR_diff in particular is more important in college than in the NBA, while TO_diff is more important in the NBA than in the college game. Although EFG_diff in our analysis is 9% higher than Winston's number, I'm hesitant to say it's more important in college than in the NBA because shooting in the NBA has recently become more important than it was at the time of Winston's findings.)

To devise our own weights for college basketball, we can simply divide the absolute value of each coefficient by the sum of the absolute values of all four. Doing so, we find:

1. Shooting differential (60%)
2. Turnover differential (25%)
3. Rebounding differential (9%)
4. Free throw rate differential (6%)

Overall, shooting offense and defense is undoubtedly the king of the four factors in both college and professional basketball. Meanwhile, shooting more free throws and preventing an opponent from doing so is more important in college ball, and causing more turnovers while protecting the ball better is a bit more important in the pro game. Rebounding takes on a similar level of importance at both levels of play. To speculate why we see these differences, I believe free throws are more valuable in college for a couple reasons: 1) the games are shorter, making every point more valuable, and 2) although I don't have the data to support this claim, the variance between two college teams' free throw percentages is greater than what we see between two NBA teams. As for turnovers, I presume possessions in the NBA are more valuable than those in college since 1) college players are simply worse at protecting the ball than professional players, and 2) the playing field is much more level in the NBA, meaning games are closer.

Further Work

The main weakness of this analysis is that it does not adjust for opponent strength. For example, ACC competition is undoubtedly fiercer than WCC competition, meaning a team like Gonzaga may have unfairly influential numbers in certain areas. For instance, the fact that Gonzaga was a particularly good rebounding team in 19-20 (see the ranking table above) could cause the importance of rebounding to be inflated. A more complete work would adjust each team's numbers according to opponent strength.