Riazul Islam

December 19th, 2019

Regression and Multivariate Data Analysis

Professor Jeffrey Simonoff

Homework 6

In 2004, Dean Oliver published his book *Basketball on Paper: Rules and Tools for Performance Analysis*, a well-regarded early tome of analytics for basketball.[1] While analytics is now an accepted part of basketball strategy and tactics, at the time analytics were still considered "nerdy" activities that "real" basketball players and coaches generally ignored when looking to improve their performance.

One of Oliver's key findings was the discovery of the basketball "Four Factors" that he believed led to wins: shooting, turnovers, rebounding, and free throws. Each of these are measured using the following metrics:[2]

- Shooting: Effective Field Goal Percentage (eFG%) = (FG + 0.5 * 3P) / FGA
    - FG = field goals made
    - 3P = 3-pointers made
    - FGA = field goals attempted
- Turnovers: Turnover Percentage (TOV%) = TOV / (FGA + 0.44 * FTA + TOV)
    - TOV = turnovers
    - FGA = field goals attempted
    - FTA = free throws attempted
- Rebounding: Rebound Percentage (RB%) = RB / (RB + Opp RB)
    - RB = rebounds
    - Opp RB = opponent rebounds
- Free Throws: Free Throw Rate & Percentage: FT/FGA
    - FT = free throws made
    - FGA = field goals attempted

Dean Oliver believed that these each of these factors influenced the game had the following weights of influence: 40% shooting, 25% turnovers, 20% rebounding, 15% free throws.

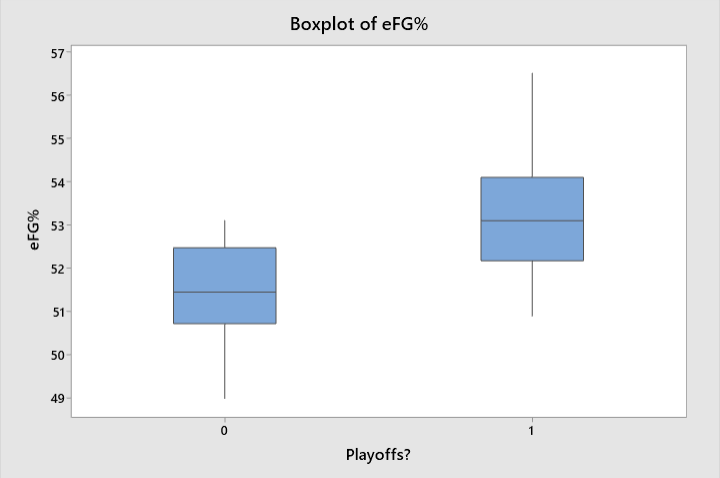
In this analysis, I would like to test whether these four factors could be used to predict whether an NBA team will make the playoffs, focusing specifically on the offensive four factors (defensive will be ignored here). The NBA is currently set up so that 16 out of the 30 teams in the league will enter the playoffs each year. For the 2018-19 NBA season, the following offensive four factor data was captured from Basketball Reference to complete this analysis[3]:

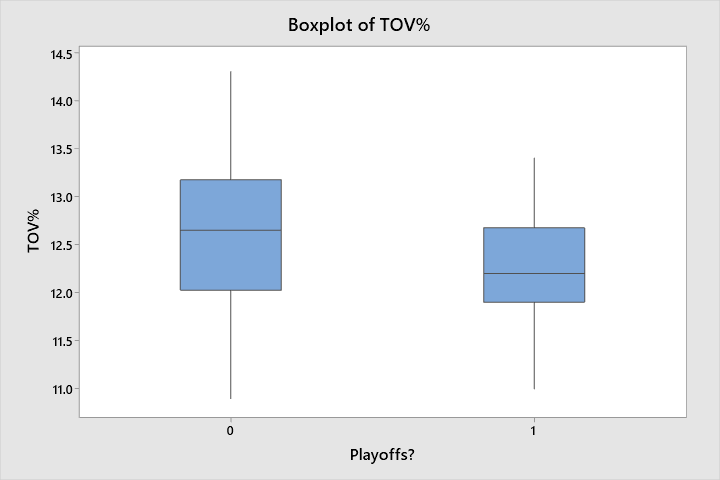
Rk	Team	eFG%	TOV%	ORB%	FT/FGA	Playoffs?
1	Milwaukee Bucks*	0.55	12	20.8	0.197	1
2	Golden State Warriors*	0.565	12.6	22.5	0.182	1
3	Toronto Raptors*	0.543	12.4	21.9	0.198	1
4	Utah Jazz*	0.538	13.4	22.9	0.217	1
5	Houston Rockets*	0.542	12	22.8	0.221	1
6	Portland Trail Blazers*	0.528	12.1	26.6	0.21	1
7	Denver Nuggets*	0.527	11.9	26.6	0.175	1
8	Boston Celtics*	0.534	11.5	21.6	0.173	1

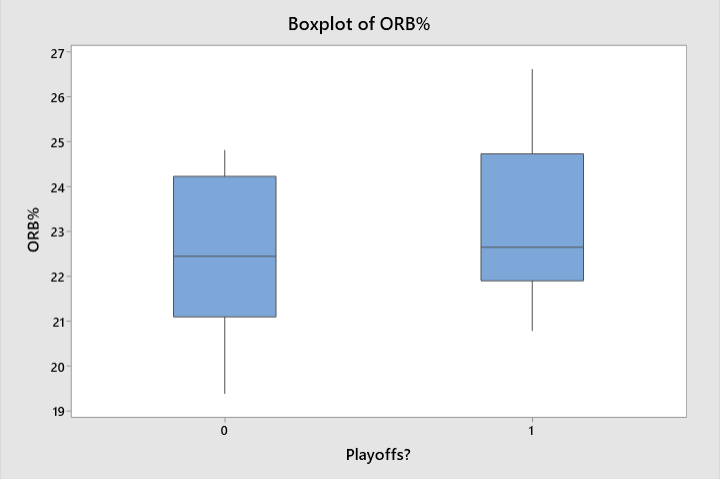
[1] https://www.amazon.com/gp/product/1574886886/qid=1135475833/sr=8-1/ref=pd_bbs_1/002-2579531-9906406?n=507846&s=books&v=glance
[2] https://www.basketball-reference.com/about/factors.html
[3] https://www.basketball-reference.com/leagues/NBA_2019.html#all_misc_stats

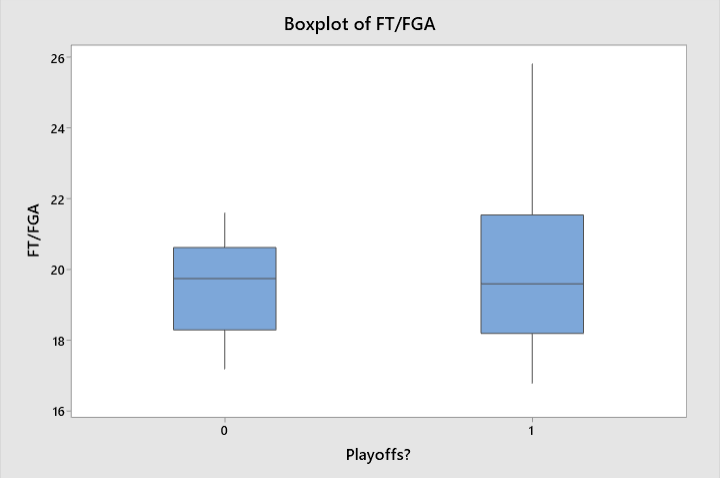
9	Oklahoma City Thunder*	0.514	11.7	26	0.19	1
10	Indiana Pacers*	0.53	12.4	21.9	0.182	1
11	Philadelphia 76ers*	0.532	12.9	24.5	0.241	1
12	San Antonio Spurs*	0.534	11	21	0.194	1
13	Los Angeles Clippers*	0.529	12.7	22	0.258	1
14	Orlando Magic*	0.518	11.9	22	0.168	1
15	Brooklyn Nets*	0.52	13	23.8	0.211	1
16	Miami Heat	0.515	13.1	24.8	0.172	0
17	Detroit Pistons*	0.509	12.3	24.8	0.195	1
18	Sacramento Kings	0.524	11.5	23.1	0.177	0
19	Dallas Mavericks	0.519	12.7	22.7	0.216	0
20	Minnesota Timberwolves	0.511	11.4	24.6	0.21	0
21	New Orleans Pelicans	0.529	12.6	24.1	0.193	0
22	Charlotte Hornets	0.514	10.9	21.7	0.205	0
23	Los Angeles Lakers	0.527	13.4	22.2	0.18	0
24	Memphis Grizzlies	0.508	12.9	20	0.21	0
25	Washington Wizards	0.531	12.3	21.3	0.204	0
26	Atlanta Hawks	0.522	14.3	24.7	0.192	0
27	Chicago Bulls	0.505	12.7	19.4	0.184	0
28	Phoenix Suns	0.514	13.8	20.5	0.202	0
29	New York Knicks	0.49	12.4	22.1	0.205	0
30	Cleveland Cavaliers	0.503	12.2	23.7	0.187	0

An initial look at boxplots of the data can help us better understand the predictive power of each of these offensive four factors by testing whether there is separation between the playoffs and non-playoffs groups. While these do not account for joint effects, this can be helpful for visualizing whether there is possible predictive power in these offensive four factors.









Effective Field Goal Percentange (eFG%) seems to exhibit the most separation among these four factors, which indicates that it could have some useful predictive power. The other factors do not seem to have much separation, with only outliers really creating the minor separation in the Turnover Percentage

(TOV%) and Offensive Rebounding (ORB%). The Free Throw Rate (FT/FGA) is even worse, with the non-playoff teams FT/FGA metrics seemingly fitting neatly within the larger playoff teams range. This seems to suggest that FT/FGA is not useful in predicting whether a team will make the playoffs. While these all have somewhat long tails, there are no obvious outliers on any of the offensive four factors here that are magnitudes above or below the bulk of the data points, and so logging of any factors will not be done at this time.

With these four factors available, we can look at potential best subsets using ordinary least squares best subsets regression, with making the 0/1 making the playoffs variable as the response variable. While this in no way is technically valid, this OLS best subsets regression can provide some guidance on the number of predictors and which predictors could be a strong predictive model for whether a team makes the playoffs based on the offensive four factors:

FULL DATASET

Best Subsets Regression: Playoffs? versus eFG%, TOV%, ORB%, FT/FGA

Response is Playoffs?

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	eFG%	TOV%	ORB%	FT/FGA
1	31.7	29.3	24.1	4.1	0.42673	X			
1	5.2	1.8	0.0	15.9	0.50291		X		
2	37.0	32.4	26.5	3.8	0.41731	X		X	
2	36.1	31.4	24.3	4.2	0.42025	X	X		
3	41.2	34.5	26.6	3.9	0.41080	X	X	X	
3	38.3	31.1	23.1	5.3	0.42110	X		X	X
4	43.4	34.3	24.6	5.0	0.41128	X	X	X	X

This best subsets analysis points to a model with two predictors with eFG% and ORB%, since the Mallows Cp is lowest here at 3.8, though a three predictor model with eFG%, TOV%, and ORB% come very close with a Mallows Cp of 3.9. We will look at the two predictor model first, with the residuals updated to be Pearson residuals (since these better approximate normality than deviance residuals), and the Likelihood ratio test rather than a Wald test for testing individual slopes using different χ^2 tests:

FULL DATASET

Binary Logistic Regression: Playoffs? versus eFG%, ORB%

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

Method

Link function	Logit
Residuals for diagnostics	Pearson
Rows used	30

Response Information

Variable	Value	Count
Playoffs?	1	16 (Event)
	0	14
	Total	30

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -85.4 + 1.434 eFG% + 0.453 ORB%

Coefficients

Term	Coef	SE Coef	VIF
Constant	-85.4	33.3	
eFG%	1.434	0.560	1.23
ORB%	0.453	0.323	1.23

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
eFG%	4.1955	(1.4002, 12.5711)
ORB%	1.5738	(0.8354, 2.9647)

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC
35.30%	30.47%	32.82	33.75	37.03

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	27	26.82	0.473
Pearson	27	23.00	0.685
Hosmer-Lemeshow	8	5.12	0.744

Analysis of Variance

Source	DF	Adj Dev	Adj Mean	Likelihood Ratio Chi-Square	P-Value
Regression	2	14.633	7.3163	14.63	0.001
eFG%	1	13.398	13.3983	13.40	0.000
ORB%	1	2.507	2.5073	2.51	0.113
Error	27	26.823	0.9934		
Total	29	41.455			

Observed and Expected Frequencies for Hosmer-Lemeshow Test

Group	Event Probability Range	Playoffs? = 1 Observed	Playoffs? = 1 Expected	Playoffs? = 0 Observed	Playoffs? = 0 Expected
1	(0.000, 0.029)	0	0.1	3	2.9
2	(0.029, 0.134)	0	0.3	3	2.7
3	(0.134, 0.272)	2	0.7	1	2.3
4	(0.272, 0.486)	1	1.2	2	1.8
5	(0.486, 0.556)	1	1.6	2	1.4
6	(0.556, 0.627)	2	1.8	1	1.2
7	(0.627, 0.722)	2	2.0	1	1.0
8	(0.722, 0.893)	2	2.6	1	0.4
9	(0.893, 0.915)	3	2.7	0	0.3
10	(0.915, 0.997)	3	2.9	0	0.1

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	192	85.7	Somers' D	0.71
Discordant	32	14.3	Goodman-Kruskal Gamma	0.71
Ties	0	0.0	Kendall's Tau-a	0.37
Total	224	100.0		

Association is between the response variable and predicted probabilities

Fits and Diagnostics for Unusual Observations

Obs	Observed Probability	Fit	Resid	Std Resid	
21	0.000	0.798	-1.986	-2.07	R

R Large residual

An initial reading of this regression indicates that this may be a strong model already. Using the **Analysis of Variance** table, the overall regression test (a null of all slopes equal to 0 versus the alternative of at least one slope equal to 0) on a χ^2 distribution has a test statistic of 14.63 and a p-value of 0.001, indicating that we may strongly reject the null hypothesis that there is no relationship between eFG%, ORB% and whether a team makes the playoffs. In addition, we can test the individual slopes using the likelihood ratio tests (another $\chi 2$ test). Here, we can see the eFG% is strongly statistically significant with a p-value below 0.001; however, the ORB% test statistic's p-value is 0.113, which is marginally statistically not significant.

For this regression, we can use the **Odds Ratios for Continuous Predictors** table to understand how a change in the eFG% or ORB% is associated with whether a team makes the playoffs, holding all else in the model fixed. In this case, the eFG% coefficient says that an increase in one percentage point in eFG% is associated with an increase in the odds of making the playoffs by 320%, while the ORB% coefficient says that an increase of one percentage point in the ORB% for a team is associated with an increase in the odds of making the playoffs by 57.38% (all holding everythine else in the model constant).

There is very little indication of collinearity in this model, with VIFs both at 1.23.

With the number of data points in this analysis small (only 30 teams in the league) and only one replication per observation, we will move straight to the Hosmer-Lemeshow test under the **Goodness-of-Fit Tests** table to understand goodness-of-fit in this model. With 10 groups in the test, there is mostly a good indication of fit with a p-value of 0.744.

Using the **Measures of Association** table, we can also understand better the proportion of variability. In this analysis, there are $16 \cdot 14$ pairs of observations, since there are 16 playoff teams and 14 non-playoff teams, leading to a total of 224 such pairs. Here, we can see that there are 85.7% are concordant pairs, where the probability of making the playoffs for the team bound for the playoffs is higher than the probability of making the playoffs for the team not bound for the playoffs. This is a fairly strong performance, indicative of quite good practical use. Summarizing the concordancies, we can use Somers' D, with a relatively high value of 0.71 an indication of potential good, but not great separation.

Before moving to an outlier analysis, it may be best to see what the three variable model indicated by best subsets earlier could look like. Using the eFG%, TOV%, and ORB%, a logit regression predicting making the playoffs would look like this:

FULL DATASET

Binary Logistic Regression: Playoffs? versus eFG%, TOV%, ORB%

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

Method

Link function	Logit
Residuals for diagnostics	Pearson
Rows used	30

Response Information

Variable	Value	Count
Playoffs?	1	16 (Event)
	0	14
	Total	30

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -85.3 + 1.585 eFG% - 0.867 TOV% + 0.571 ORB%

Coefficients

Term	Coef	SE Coef	VIF
Constant	-85.3	36.8	
eFG%	1.585	0.633	1.42
TOV%	-0.867	0.634	1.14
ORB%	0.571	0.392	1.42

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
eFG%	4.8817	(1.4113, 16.8860)
TOV%	0.4203	(0.1214, 1.4550)
ORB%	1.7705	(0.8205, 3.8207)

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC
40.33%	33.09%	32.74	34.34	38.34

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	26	24.74	0.534
Pearson	26	21.51	0.715
Hosmer-Lemeshow	8	7.70	0.464

Analysis of Variance

				Likelihood Ratio	
Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	3	16.717	5.5724	16.72	0.001
eFG%	1	13.902	13.9025	13.90	0.000
TOV%	1	2.085	2.0846	2.08	0.149
ORB%	1	2.949	2.9488	2.95	0.086
Error	26	24.738	0.9515		
Total	29	41.455			

Observed and Expected Frequencies for Hosmer-Lemeshow Test

	Event Probability	Playoffs? = 1		Playoffs? = 0	
Group	Range	Observed	Expected	Observed	Expected
1	(0.000, 0.010)	0	0.0	3	3.0
2	(0.010, 0.244)	1	0.3	2	2.7
3	(0.244, 0.290)	1	0.8	2	2.2
4	(0.290, 0.328)	0	0.9	3	2.1
5	(0.328, 0.524)	2	1.3	1	1.7
6	(0.524, 0.704)	2	1.9	1	1.1
7	(0.704, 0.807)	1	2.3	2	0.7
8	(0.807, 0.861)	3	2.6	0	0.4
9	(0.861, 0.957)	3	2.8	0	0.2
10	(0.957, 0.998)	3	2.9	0	0.1

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	196	87.5	Somers' D	0.75
Discordant	28	12.5	Goodman-Kruskal Gamma	0.75
Ties	0	0.0	Kendall's Tau-a	0.39
Total	224	100.0		

Association is between the response variable and predicted probabilities

Fits and Diagnostics for Unusual Observations

Obs	Observed Probability	Fit	Resid	Std Resid	
21	0.000	0.799	-1.995	-2.10	R

R Large residual

Reading through the three factor regression, the **Analysis of Variance** table indicates that the regression is about as strong, with a p-value of 0.001 on a test statistic of 16.72. However, while the eFG% p-value is still below 0.001, the TOV% p-value is 0.149, indicating lack of statistical significance and lack of evidence to reject the null hypothesis. Furthermore, the ORB% is now somewhat statistically significant with a p-value of 0.086, a bit better than before.

For this model, the **Odds Ratios for Continuous Predictors** table indicates that a one percentage point increase in eFG% is associated with a 388% increase in the odds of making the playoffs, much higher than in the previous model, while a one percentage point increase in ORB% is associated with a 77% increase in the odds of making the playoffs. As is expected given the negative impact of offensive turnovers to a team, a one percentage point increase in TOV% is associated with a decrease in the odds of making the playoffs by 58%, quite a significant reduction.

VIFs here, as in the last model, also indicate very little possibility of collinearity in the model.

Unfortunately, the Hosmer-Lemeshow test under the **Goodness-of-Fit Tests** table has gotten much worse, with a p-value of only 0.464, indicating that this model may have very poor goodness-of-fit and that the logit model may not fit the data adequately. The **Measures of Association** table displays a higher percentage of concordant pairs at 87.5%, while we also have a higher Somers' D at 0.75, indicating even better separation than the previous model.

However, it generally looks like TOV% does not add much to the model, and that the 2-factor eFG% and ORB% model is quite strong. However, we may want to consider a 1-factor model with just eFG%, since the previous 2-factor model had a poor test statistic's p-value (0.113) for ORB%. We will test the 1-factor model with only eFG% here:

FULL DATASET

Binary Logistic Regression: Playoffs? versus eFG%

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

Method

Link function	Logit
Residuals for diagnostics	Pearson
Rows used	30

Response Information

Variable	Value	Count
Playoffs?	1	16 (Event)
	0	14
	Total	30

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))
Y' = -65.1 + 1.247 eFG%

Coefficients

Term	Coef	SE Coef	VIF
Constant	-65.1	25.0	
eFG%	1.247	0.479	1.00

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
eFG%	3.4802	(1.3622, 8.8912)

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC
29.25%	26.84%	33.33	33.77	36.13

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	28	29.33	0.396
Pearson	28	25.71	0.589
Hosmer-Lemeshow	8	1.58	0.991

Analysis of Variance

				Likelihood Ratio	
Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	1	12.13	12.125	12.13	0.000
eFG%	1	12.13	12.125	12.13	0.000
Error	28	29.33	1.048		
Total	29	41.46			

Observed and Expected Frequencies for Hosmer-Lemeshow Test

	Event Probability	Playoffs? = 1		Playoffs? = 0	
Group	Range	Observed	Expected	Observed	Expected
1	(0.000, 0.105)	0	0.2	3	2.8
2	(0.105, 0.199)	1	0.5	2	2.5
3	(0.199, 0.265)	1	0.8	2	2.2
4	(0.265, 0.402)	1	1.1	2	1.9
5	(0.402, 0.557)	1	1.5	2	1.5
6	(0.557, 0.674)	2	2.0	1	1.0
7	(0.674, 0.726)	2	2.1	1	0.9
8	(0.726, 0.814)	3	3.2	1	0.8
9	(0.814, 0.931)	3	2.7	0	0.3
10	(0.931, 0.995)	2	2.0	0	0.0

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	186	83.0	Somers' D	0.68
Discordant	34	15.2	Goodman-Kruskal Gamma	0.69
Ties	4	1.8	Kendall's Tau-a	0.35
Total	224	100.0		

Association is between the response variable and predicted probabilities

Fits and Diagnostics for Unusual Observations

Obs	Observed Probability	Fit	Resid	Std Resid	
17	1.0000	0.1621	2.2736	2.38	R

R Large residual

This is single factor model is very interesting, as the **Analysis of Variance** table shows p-values for both the overall regression and the eFG% at below 0.001 (as is expected for a regression with one predictor variable), indicating that we may reject both the full regression and eFG% null hypotheses. The odds ratio here indicates that every one percentage point increase in eFG% is associated with a 248% increase in likelihood of making the playoffs. This is very interesting: the previous two models had higher likelihood of making the playoffs given a one percentage point increase in eFG% with all other model elements held constant. The Hosmer-Lemeshow test statistic leads to a p-value of 0.991, a great indication of goodness-of-fit for this model. However, concordant pairs now only account for 83.0% of total pairs; however, 4 ties do help and their inclusion in the concordant pair metric accounts for 84.8% of all pairs,

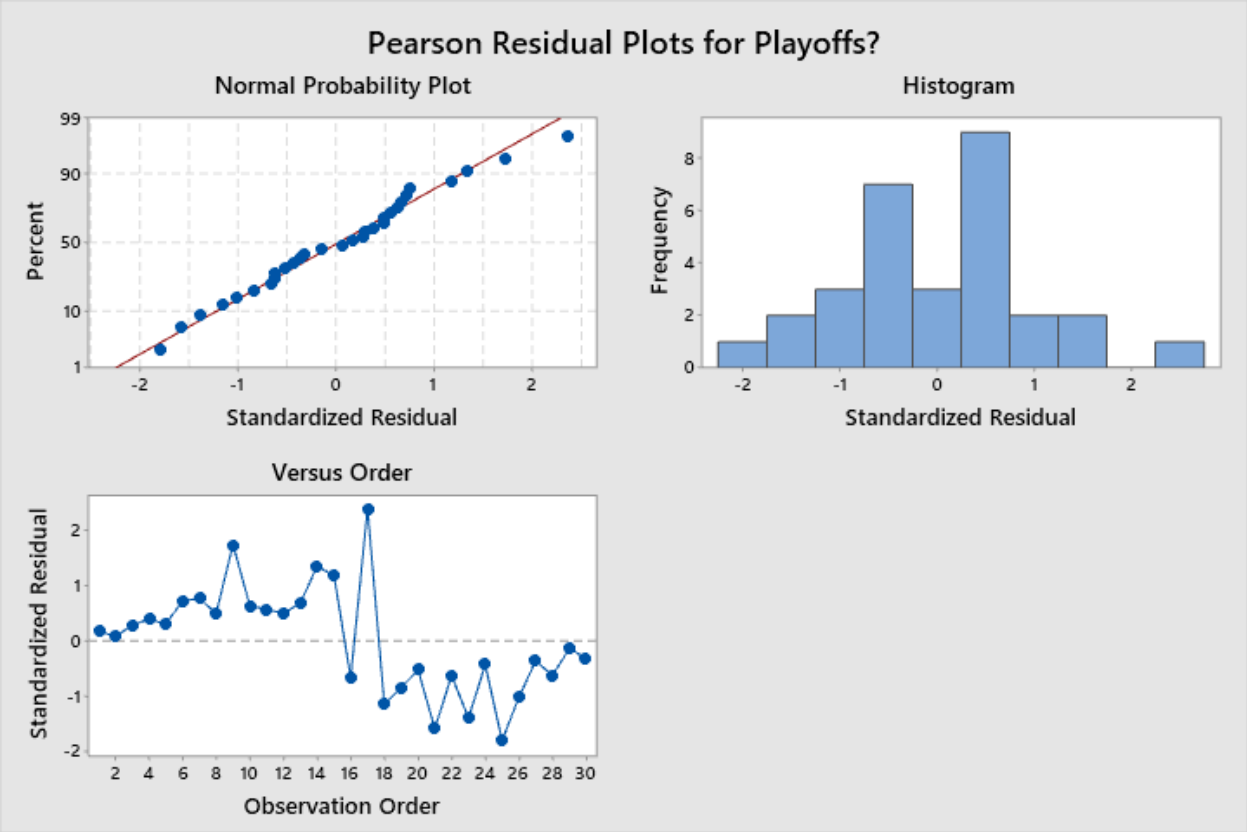
less than 1% less than the 2-factor model initially run. Somers' D also falls to 0.68, which is still good, but not as good as the other two models.

In order to identify potential unusual observations on this 1-factor model, we can look at the standardized Pearson residuals, leverage values, and Cook's distances. I've included fits here as well:

Rk	Team	eFG%	Playoffs?	FITS_5	SPEARRES_5	HI_5	COOK_5
1	Milwaukee Bucks*	55	1	0.969832	0.181473	0.055455	0.000967
2	Golden State Warriors*	56.5	1	0.995232	0.069931	0.020296	5.07E-05
3	Toronto Raptors*	54.3	1	0.930693	0.283479	0.073328	0.003179
4	Utah Jazz*	53.8	1	0.878023	0.38826	0.078434	0.006415
5	Houston Rockets*	54.2	1	0.922204	0.302007	0.075098	0.003703
6	Portland Trail Blazers*	52.8	1	0.67409	0.716665	0.058659	0.016003
7	Denver Nuggets*	52.7	1	0.646122	0.761769	0.056171	0.017268
8	Boston Celtics*	53.4	1	0.813816	0.497169	0.074434	0.009939
9	Oklahoma City Thunder*	51.4	1	0.265185	1.730814	0.075031	0.121502
10	Indiana Pacers*	53	1	0.726346	0.634538	0.064288	0.013832
11	Philadelphia 76ers*	53.2	1	0.773044	0.561816	0.069859	0.011853
12	San Antonio Spurs*	53.4	1	0.813816	0.497169	0.074434	0.009939
13	Los Angeles Clippers*	52.9	1	0.700872	0.674329	0.061413	0.014876
14	Orlando Magic*	51.8	1	0.37277	1.338431	0.060723	0.057906
15	Brooklyn Nets*	52	1	0.432678	1.177992	0.055114	0.04047
16	Miami Heat	51.5	0	0.290186	-0.66353	0.071442	0.016937
17	Detroit Pistons*	50.9	1	0.162093	2.380217	0.087573	0.271879
18	Sacramento Kings	52.4	0	0.556731	-1.15079	0.051607	0.036032
19	Dallas Mavericks	51.9	0	0.40236	-0.84525	0.05767	0.021862
20	Minnesota Timberwolves	51.1	0	0.198879	-0.52059	0.083988	0.012424
21	New Orleans Pelicans	52.9	0	0.700872	-1.57999	0.061413	0.08167
22	Charlotte Hornets	51.4	0	0.265185	-0.62463	0.075031	0.015824
23	Los Angeles Lakers	52.7	0	0.646122	-1.39086	0.056171	0.057565
24	Memphis Grizzlies	50.8	0	0.14586	-0.43284	0.088513	0.009097
25	Washington Wizards	53.1	0	0.750424	-1.79533	0.067146	0.116002
26	Atlanta Hawks	52.2	0	0.494622	-1.01602	0.05191	0.02826
27	Chicago Bulls	50.5	0	0.105121	-0.35889	0.087957	0.006211
28	Phoenix Suns	51.4	0	0.265185	-0.62463	0.075031	0.015824
29	New York Knicks	49	0	0.017772	-0.13777	0.046714	0.000465
30	Cleveland Cavaliers	50.3	0	0.083862	-0.31631	0.085097	0.004653

Immediately, row 17 stands out with a substantial standardized Pearson's residual of 2.38 and a Cook's distance of 0.27. It seems like the Detroit Pistons' season, where they made the playoffs but had a lower eFG% than 10 teams, is driving this outlier. Interestingly, the Detroit Pistons had one of the top ORB% at 24.8%, ranking 4th in the league. That may have been a driver in the Pistons making the playoffs, featuring center Andre Drummond grabbing 5.4 offensive rebounds per game, an outlier in itself.

Here are the residual plots, which clearly shows the Detroit Pistons in the top right corner as an outlier:



If we omit the Detroit Pistons from the dataset, we can restart the analysis by observing the best subsets analysis again:

FULL DATASET EX. DETROIT PISTONS

Best Subsets Regression: Playoffs? versus eFG%, TOV%, ORB%, FT/FGA

Response is Playoffs?

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	eFG%	TOV%	ORB%	FT/FGA

									G A
1	37.9	35.6	30.6	3.5	0.40798	X			
1	5.1	1.6	0.0	18.5	0.50444		X		
2	42.1	37.7	30.9	3.5	0.40145	X	X		
2	41.3	36.8	30.5	3.9	0.40436	X		X	
3	45.4	38.8	30.9	4.1	0.39781	X	X	X	
3	44.4	37.7	29.5	4.5	0.40142	X	X		X
4	47.7	38.9	29.1	5.0	0.39736	X	X	X	X

Here, we can see that the 1-factor and 2-factor models are tied on the Mallows Cp metric (at 3.5). However, while the 1-factor model continues to use eFG% as its sole predictor variable, the 2-factor model has switched from ORB% to TOV%, with the eFG%-ORB% model having a 3.9 Mallows Cp. Let's analyze the 1-factor model without the Detroit Pistons:

FULL DATASET EX. DETROIT PISTONS

Binary Logistic Regression: Playoffs? versus eFG%

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

Method

Link function	Logit
Residuals for diagnostics	Pearson
Rows used	29

Response Information

Variable	Value	Count	
Playoffs?	1	15	(Event)
	0	14	
	Total	29	

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -84.5 + 1.613 eFG%

Coefficients

Term	Coef	SE Coef	VIF
Constant	-84.5	31.7	
eFG%	1.613	0.606	1.00

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
eFG%	5.0171	(1.5306, 16.4449)

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC
37.45%	34.96%	29.12	29.59	31.86

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	27	25.12	0.567
Pearson	27	22.37	0.718
Hosmer-Lemeshow	8	1.36	0.995

Analysis of Variance

				Likelihood Ratio	
Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	1	15.04	15.0432	15.04	0.000
eFG%	1	15.04	15.0432	15.04	0.000
Error	27	25.12	0.9305		

Total	28	40.17

Observed and Expected Frequencies for Hosmer-Lemeshow Test

Group	Event Probability Range	Playoffs? = 1 Observed	Playoffs? = 1 Expected	Playoffs? = 0 Observed	Playoffs? = 0 Expected
1	(0.000, 0.034)	0	0.0	2	2.0
2	(0.034, 0.114)	0	0.2	3	2.8
3	(0.114, 0.173)	1	0.5	2	2.5
4	(0.173, 0.319)	1	0.8	2	2.2
5	(0.319, 0.512)	1	1.3	2	1.7
6	(0.512, 0.667)	2	1.9	1	1.1
7	(0.667, 0.734)	2	2.1	1	0.9
8	(0.734, 0.841)	3	3.2	1	0.8
9	(0.841, 0.950)	2	1.9	0	0.1
10	(0.950, 0.999)	3	2.9	0	0.1

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	182	86.7	Somers' D	0.75
Discordant	24	11.4	Goodman-Kruskal Gamma	0.77
Ties	4	1.9	Kendall's Tau-a	0.39
Total	210	100.0		

Association is between the response variable and predicted probabilities

Fits and Diagnostics for Unusual Observations

Obs	Observed Probability	Fit	Resid	Std Resid	
9	1.0000	0.1731	2.1853	2.29	R

R Large residual

Again, we have a full regression and individual eFG% p-values on the respective $\chi 2$ tests at below 0.001, which leads us to strongly reject the null hypothesis for each test. An increase of one percentage point in eFG% is associated with a 402% increase in the odds of making the playoffs. The Hosmer-Lemeshow test test statstic is stronger than any previous model run, with a p-value of 0.995, indicating extremely strong goodness-of-fit. Concordant pairs, out of the new total number of pairs (15 x 14 = 210), are 86.7% of total pairs, a strong showing, while Somers' D is up to 0.75, stronger than the previous 1-factor model.

We can now construct a classification matrix to see whether the logistic regression model can best predict whether a team goes to the playoffs:

FULL DATASET

Tabulated Statistics: Playoffs?, Predict

Rows: Playoffs? Columns: Predict

	0	1	All
0	10	4	14
	33.33	13.33	46.67
1	4	12	16
	13.33	40.00	53.33
All	14	16	30
	46.67	53.33	100.00

Cell Contents
Count
% of Total

It looks like the model fits 73.33% of observations correctly classified (22 out of 30). This is fairly high, but we could compare to a prediction simply based on predicting whether all observations would have come from the larger group. In this case, C_{max} = max(16/30, 14/30) = 53.33%. This lower bound for what we could expect the observed proportion of correctly classified teams is much lower than the 73.33% observed from the model, supporting the usefulness of the logistic regression. We can also calculate C_{pro}, which would represent the likelihood that we have correctly classified playoff-bound and non-playoff bound teams given the logistic regression having no power to make predictions. In this case, the C_{pro} would be calculated as:

C_{pro} = 1.25*[(53.33)(46.67)+(46.67)(53.33)] = 54.45%

Since the model's correctly observed teams going/not going to the playoffs is 73.33%, significantly higher than C_{pro}, the usefulness of the logistic regression is upheld.

To further validate the final model, we can test using NBA season data from the 2017-18 season to see whether the regression will accurately predict whether a team makes the playoffs using the eFG% metric. Here is the data for running the validation, including the logit, prob, and Predict calculations:

Rk	Team	eFG%	TOV %	ORB %	FT/FG A	Playoffs ?	logit	prob	Predict
1	Houston Rockets*	55.1	12.7	21.3	23.3	1	4.3763	0.987584	1
2	Toronto Raptors*	53.9	12.1	23	19.8	1	2.4407	0.919879	1
3	Golden State Warriors*	56.9	14.1	21	19.5	1	7.2797	0.999311	1
4	Utah Jazz*	52.7	13.7	21.5	20.2	1	0.5051	0.623657	1
5	Philadelphia 76ers*	53.5	14.6	25.3	19.8	1	1.7955	0.8576	1
6	Oklahoma City Thunder*	51.4	12.4	27.7	19.7	1	-1.5918	0.169131	0
7	Boston Celtics*	51.8	13	21.5	18.8	1	-0.9466	0.279569	0
8	San Antonio Spurs*	50.7	12.2	23.7	18.9	1	-2.7209	0.061751	0
9	Portland Trail Blazers*	51.1	12.3	23.3	19.2	1	-2.0757	0.111481	0
10	Minnesota Timberwolves*	52.3	11.4	24.4	22.5	1	-0.1401	0.465032	0
11	Denver Nuggets	53.6	13.4	25.7	19.8	0	1.9568	0.876186	1

12	New Orleans Pelicans*	54.1	13.3	20	18.3	1	2.7633	0.94066	1
13	Indiana Pacers*	52.5	12.3	22.7	17.3	1	0.1825	0.545499	1
14	Cleveland Cavaliers*	54.7	12.6	20.1	21.4	1	3.7311	0.976594	1
15	Washington Wizards*	52.5	13.3	23.5	19.6	1	0.1825	0.545499	1
16	Los Angeles Clippers	52.7	13.2	23.5	22.2	0	0.5051	0.623657	1
17	Miami Heat*	52	13.3	21.5	17.3	1	-0.624	0.348872	0
18	Charlotte Hornets	50.8	11.4	22.2	23.3	0	-2.5596	0.071784	0
19	Detroit Pistons	51.2	12.3	22.7	16.9	0	-1.9144	0.128487	0
20	Milwaukee Bucks*	53.1	12.9	20.4	22	1	1.1503	0.759566	1
21	Los Angeles Lakers	51.7	13.8	23.6	18.8	0	-1.1079	0.248263	0
22	Dallas Mavericks	51.3	11.6	18	16.6	0	-1.7531	0.147657	0
23	New York Knicks	51	13.3	24.1	17	0	-2.237	0.096477	0
24	Brooklyn Nets	51.4	13.6	21	20.1	0	-1.5918	0.169131	0
25	Orlando Magic	51.2	13.3	20	18	0	-1.9144	0.128487	0
26	Atlanta Hawks	51.2	14.1	21.1	18.5	0	-1.9144	0.128487	0
27	Memphis Grizzlies	50	14	22.4	20.1	0	-3.85	0.020836	0
28	Sacramento Kings	50.2	12.8	21.5	14.3	0	-3.5274	0.028543	0
29	Chicago Bulls	49.7	12.6	20.6	16.4	0	-4.3339	0.012946	0
30	Phoenix Suns	49.5	13.9	22.5	20.3	0	-4.6565	0.00941	0

Based on this data, we can rerun the classification matrix using the final logistic regression model run:

2017-18 DATASET

Tabulated Statistics: Playoffs?, Predict

Rows: Playoffs? Columns: Predict

	0	1	All
0	12	2	14
	40.00	6.67	46.67
1	6	10	16
	20.00	33.33	53.33

All	18	12	30
	60.00	40.00	100.00

Cell Contents
Count
% of Total

We now have an observed successful classification of 73.33%, the same as before. In comparison, our new C_{pro} is:

C_{pro} = [(53.33)(40.00)+(46.67)(60.00)] = 49.33%

With the observed successful classification at 73.33% being much higher than the comparison C_{pro}, we have revalidated the usefulness of the logistic regression.

References:

- http://www.basketballonpaper.com/
- https://www.amazon.com/gp/product/1574886886/qid=1135475833/sr=8-1/ref=pd_bbs_1/002-2579531-9906406?n=507846&s=books&v=glance
- https://www.basketball-reference.com/about/factors.html
- https://www.basketball-reference.com/leagues/NBA_2019.html#all_misc_stats
- https://www.nba.com/thunder/news/factors050127.html
- http://www.rawbw.com/~deano/articles/20040601_roboscout.htm
- https://squared2020.com/2017/09/05/introduction-to-olivers-four-factors/
- https://www.basketball-reference.com/leagues/NBA_2018.html#all_misc_stats