4th October 2023

Pythagorean Expectation and other Machine Learning applications in EuroLeague

Aristotelis Michailidis
Ioannis Ntzoufras
Christos Marmarinos

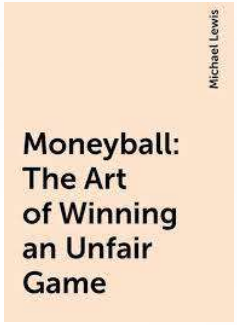
DeptEcon ResearchSeminars
UNIVERSITY OF CRETE

About me

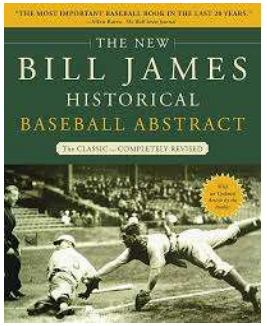
- B.Sc. Mathematics, UoA
- M.Sc. Business Analytics, AUEB
- Business Analyst, Unisys

Aristotelis Michailidis
arismichaild@gmail.com

Origin



SABERMETRICS BASEBALL



Pythagorean Expectation

What is Pythagorean Expectation?

$$PythagoreanW\% = \frac{Points\ Scored^{x}}{Points\ Scored^{x} + Points\ Allowed^{x}}$$

- The % of games a team should have won based on the points.
- Bill James
- Evaluation metric for the teams' performances
- Predictive ability at half season
  - (Baysal and Yildiztepe, 2019, and Miller, 2004)

Example 2021-22

Team	Points Scored	Points Allowed	Actual Wins	Pythagorean percentage	Pythagorean Expected Wins	Performance
Panathinaikos	2089	2235	9	0.28	8	Overperformed
Olympiacos	2222	2045	19	0.76	21	Underperformed

**Total Games: 28*

$$Pyth\ W\% = \frac{Points\ Scored^{13.91}}{Points\ Scored^{13.91} + Points\ Allowed^{13.91}}$$

Aim and objectives

- To predict the final standing (end-season) based on the half-season.
- To define the best exponent value of PE formula in EuroLeague
- To compare PE with other Machine Learning applications

Models

Pythagorean Expectation

Regression (PS and PA)

Regression (Boxscore)

Logistic (PS and PA)

Logistic (Boxscore)

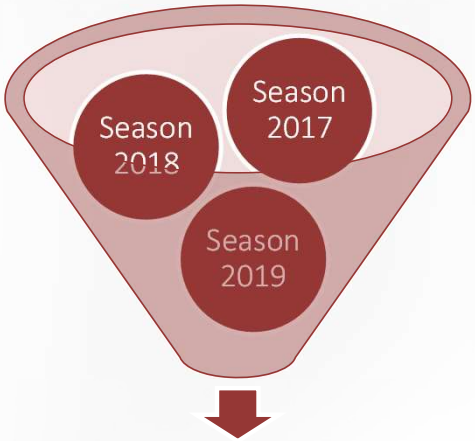
Ranking Level

Game Level

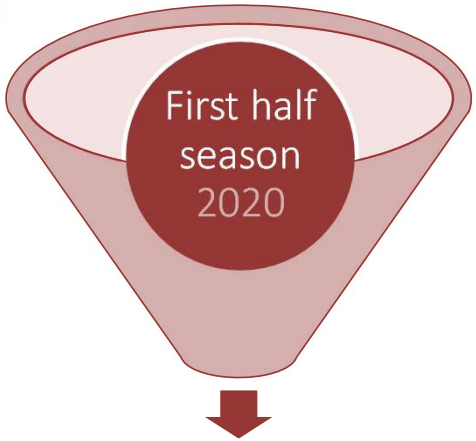
Approach

22 **=** ***10*** **+** ***12***

Final Wins (End Season) = Actual Wins (Half Season) + Predicted Wins (Half Season)



Train Data



Test Data

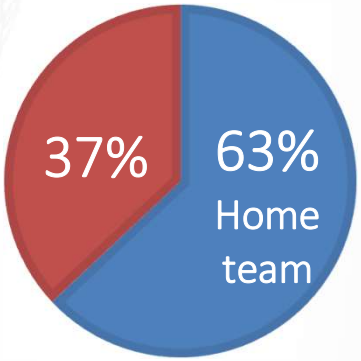
Data Structure

- 6K rows per season
- 918 games
- 25 different teams
- 4 seasons

JSON files → R

Boxscore statistics from 2016-17 to 2019-20

Last six fixtures of season 2019-2020 did not take place due to Covid-19 pandemic



Each game per team and per player

Irrelevant variables were omitted

44 variables out of 54

Ranking level VS Game level

Ranking level		Game level
Mean values of team's attributes per season	Information	Each row now corresponds to each game
$W\_per = \frac{Wins}{Total\ Games}$	Target variable	Winner $\in \{0,1\}$
16 teams 25 variables	Train data	240 games 46 variables
18 teams 25 variables	Test data	198 games 46 variables

Sample of Ranking and Game level

Ranking Level

Team	W	L	PS	PA	Assists	Rebound	…	W%
Real Madrid 2017	23	7	86.16	78.44	20.61	36.65	…	0.77
CSKA Moscow 2017	22	8	87.30	79.56	20.36	33.45	…	0.73
…	…	…	…	…	…	…	…	…
Real Madrid 2018	19	11	86.23	79.83	19.46	35.70	…	0.63
…	…	…	…	…	…	…	…	…

Game Level

Home Team	Away Team	H Points	A Points	H Assists	A Assists	…	Winner
Baskonia	Anadolu Efes	85	84	16	15	…	1
Olympiacos	Anadolu Efes	90	66	18	11	…	1
Anadolu Efes	CSKA Moscow	87	93	27	23	…	0
…	…	…	…	…	…	…	…

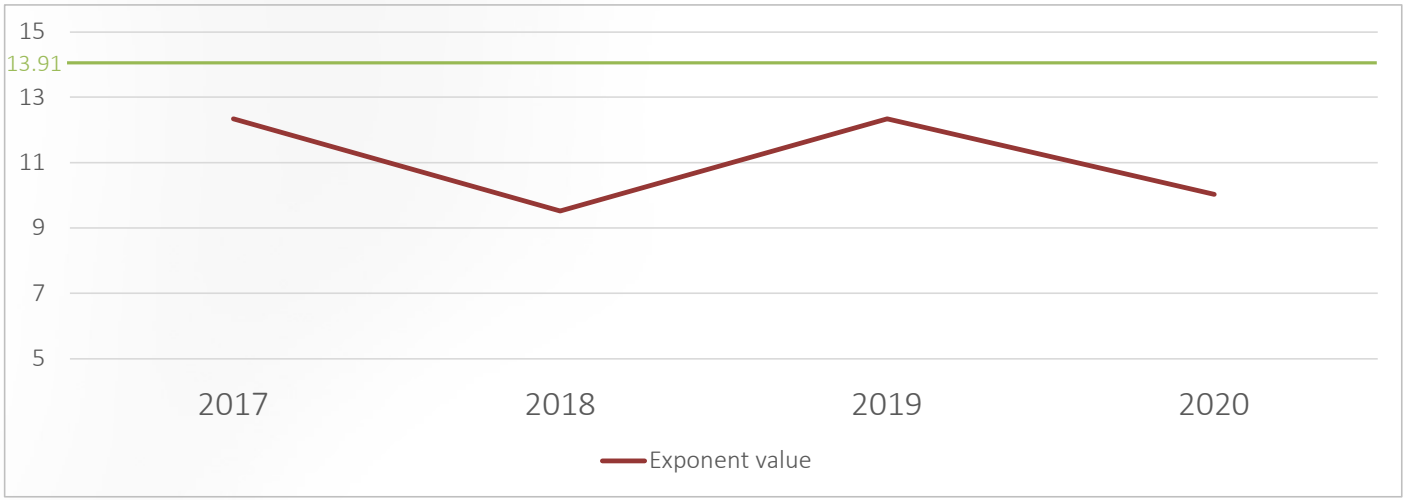
Pythagorean expectation in Basketball and in other sports

- Different exponent value from sport to sport
- In NBA:
  - Morey (13.91)
  - Oliver (14, 16.5)

Sport	Exponent x
Baseball	1.82
Ice hockey	1.92
American Football	2.37
Basketball	13.91

Exponent value (1/2)

- Exponent value of x ➡ (9.5, 12.5) < 13.91 (NBA)
  - RMSE – Half season ➡ (0.81, 1.56)
  - RMSE – End season ➡ (1.36, 2.13)
  - Pearson Correlation coefficients ➡ statistically significant

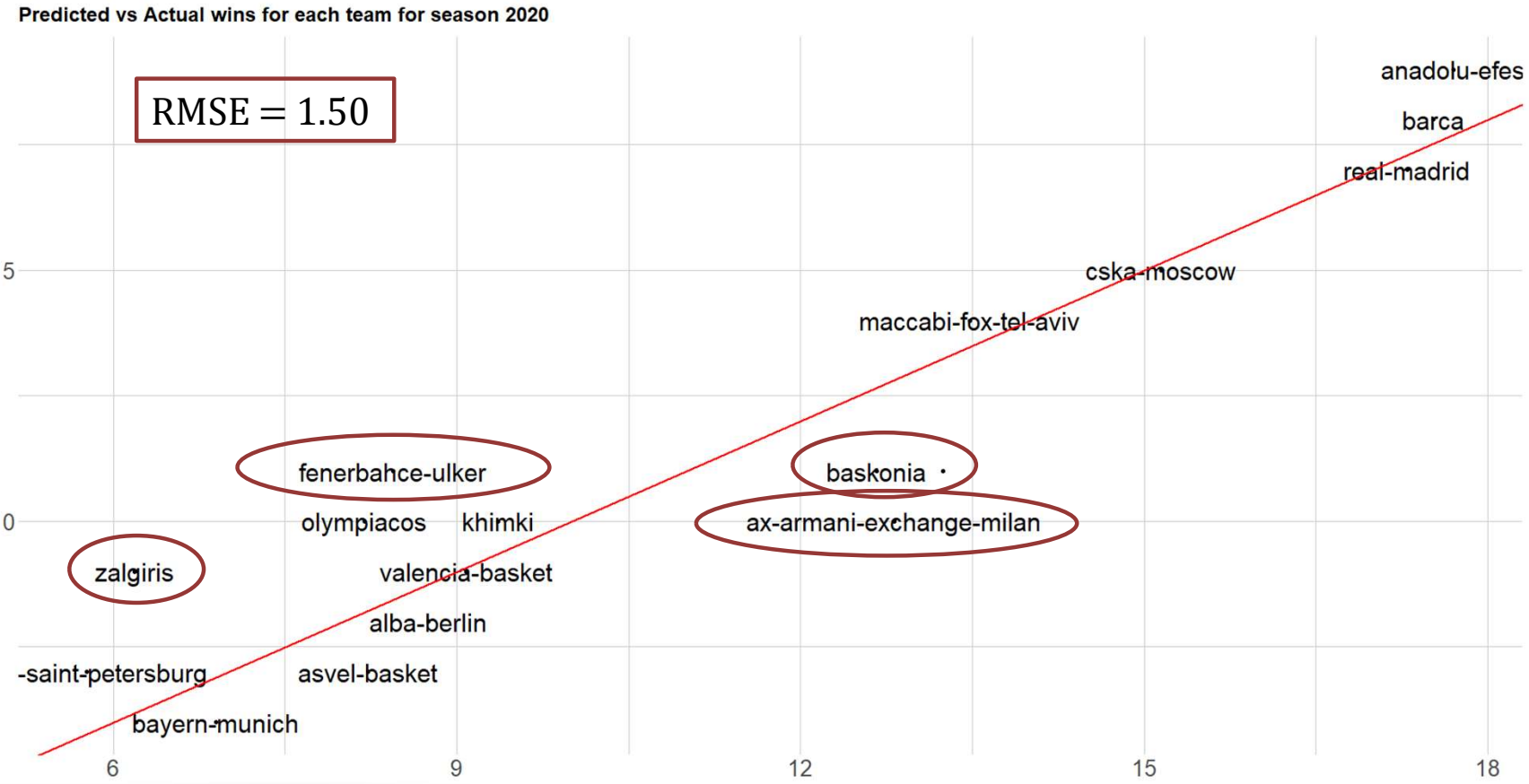


Exponent value (2/2)

- Best value → 11.19
- Check the stability of the exponent value → Bootstrapping

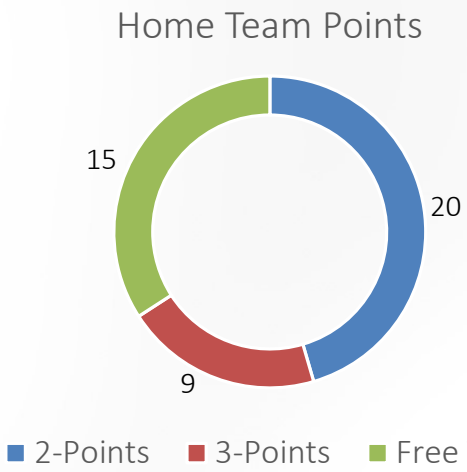
Method	Confidence Interval 95%
Normal	(10.03, 12.33)
Basic	(9.98, 12.32)
Percentile	(10.06, 12.40)
Bias Corrected Accelerated	(10.05, 12.37)

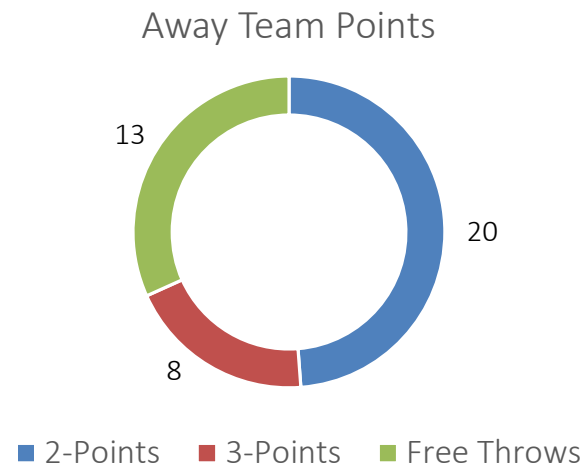
Results


Predicted vs Actual wins for each team for season 2020
RMSE = 1.50
anadolu-efes
barca
real-madrid
cska-moscow
maccabi-fox-tel-aviv
fenerbahce-ulker
olympiacos
khimki
baskonia
ax-armani-exchange-milan
zalgiris
valencia-basket
alba-berlin
-saint-petersburg
asvel-basket
bayern-munich
15
10
6
9
12
15
18

A typical game

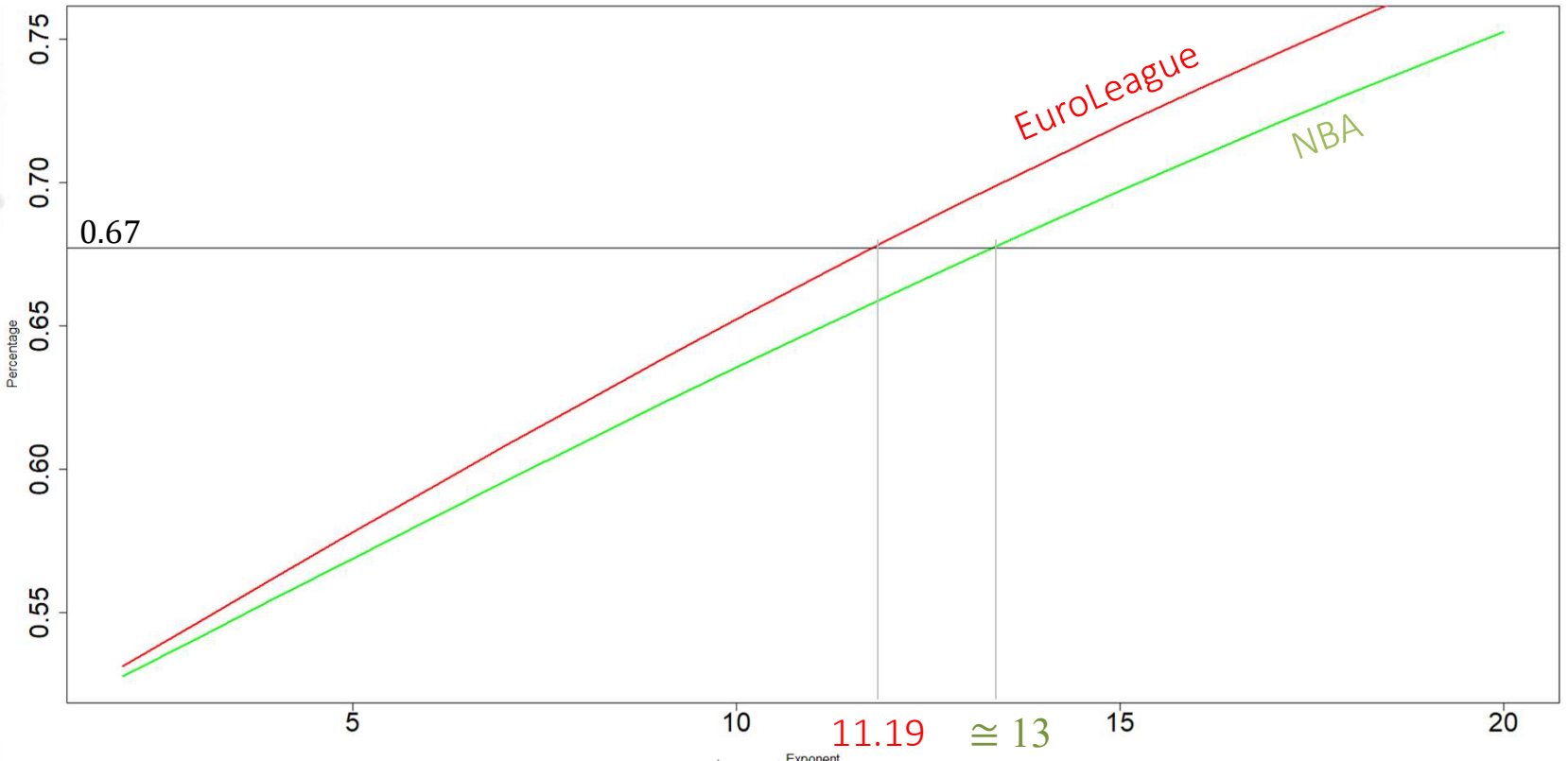
	Points	Assists	Rebounds	Steals	Blocks	Fouls
Home Team	82	18	34	7	3	20
Away Team	77	17	33	6	2	21
Differences	+5	+1	+1	+1	+1	-1





NBA vs EuroLeague

NBA Season 2017-2020 ⟹ Home team – Away team: 111 - 105



Machine Learning Applications

1. OLS with Boxscore Statistics
2. OLS with PS & PA
3. Binomial Logistic with Boxscore Statistics
4. Binomial Logistic with PS & PA
5. Binomial Logistic at Game Level

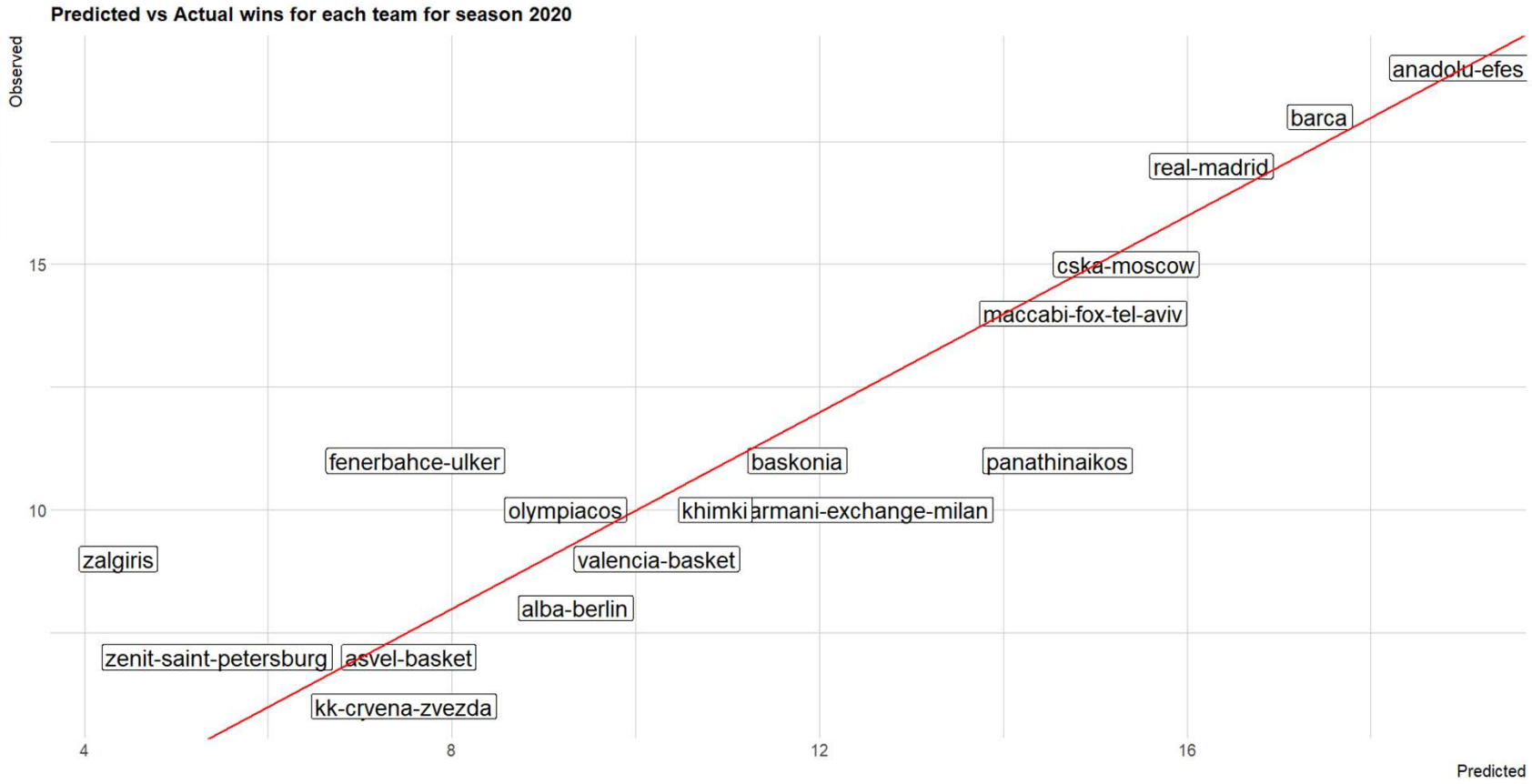
OLS with Boxscore Statistics (1/2)

- Best predictors:
  - Stepwise regression based on AIC + LASSO regression

$$W_{per} = 0.20 + 0.02 \times FT + 0.09 \times Three_P - 0.06 \times FGA + 0.06 \times Two_p + 0.11 \times STL - 0.05 \times TOV + 0.06 \times TRB + \varepsilon, \varepsilon \sim N(0,0.08)$$

R-squared Adjusted	0.77
RMSE (half season)	1.46
RMSE (end season)	1.81
Pearson Correlation	0.90

OLS with Boxscore Statistics (2/2)



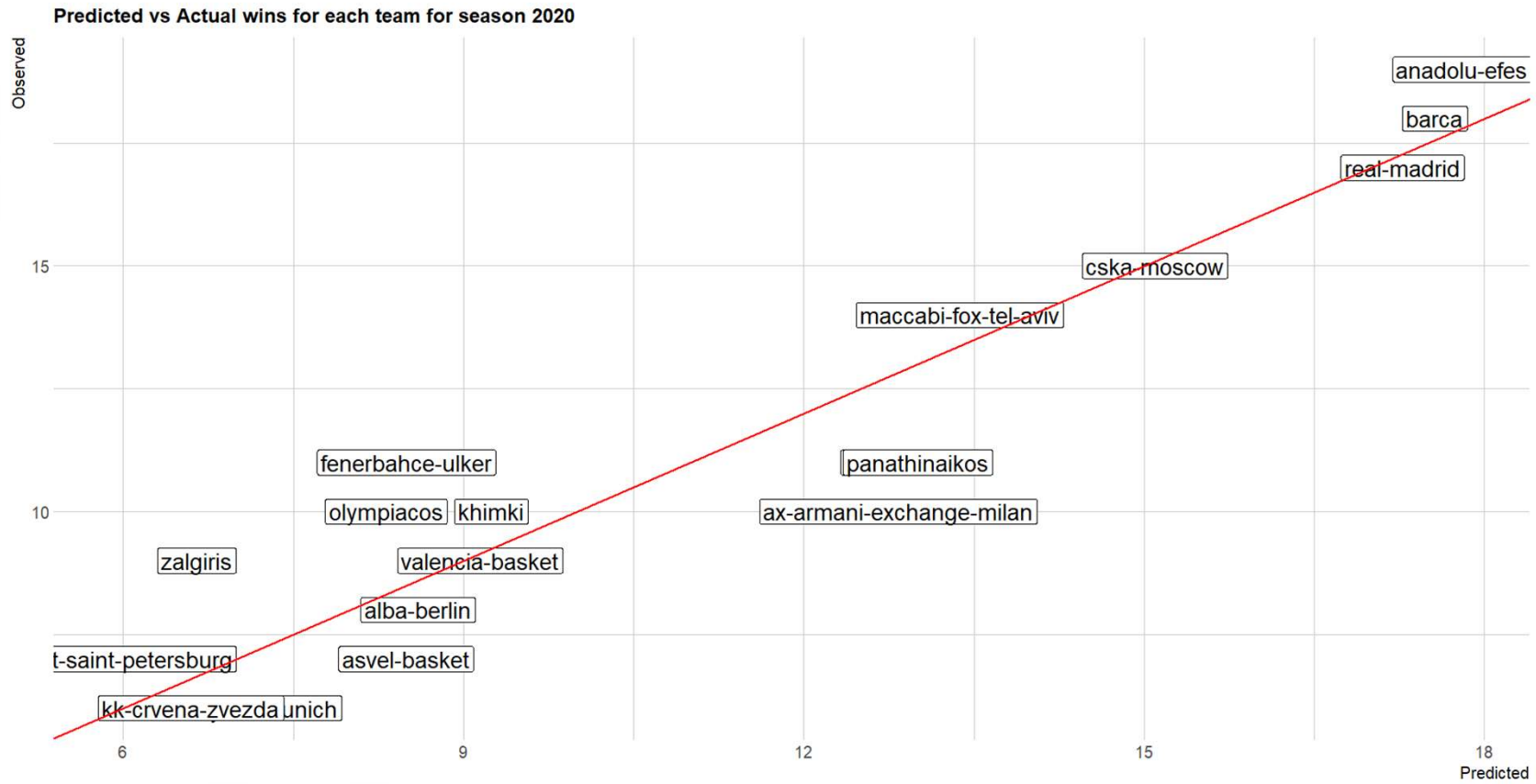
OLS with PS & PA (1/2)

- At this stage, since we are using as covariates only the PS and PA, there is no need to perform any variable selection for this model.

$$W_{per} = 0.76 + 0.03 \times PS - 0.03PA + \varepsilon, \varepsilon \sim N(0,0.06)$$

R-squared Adjusted	0.88
RMSE (half season)	1.03
RMSE (end season)	1.45
Pearson Correlation	0.93

OLS with PS & PA (2/2)



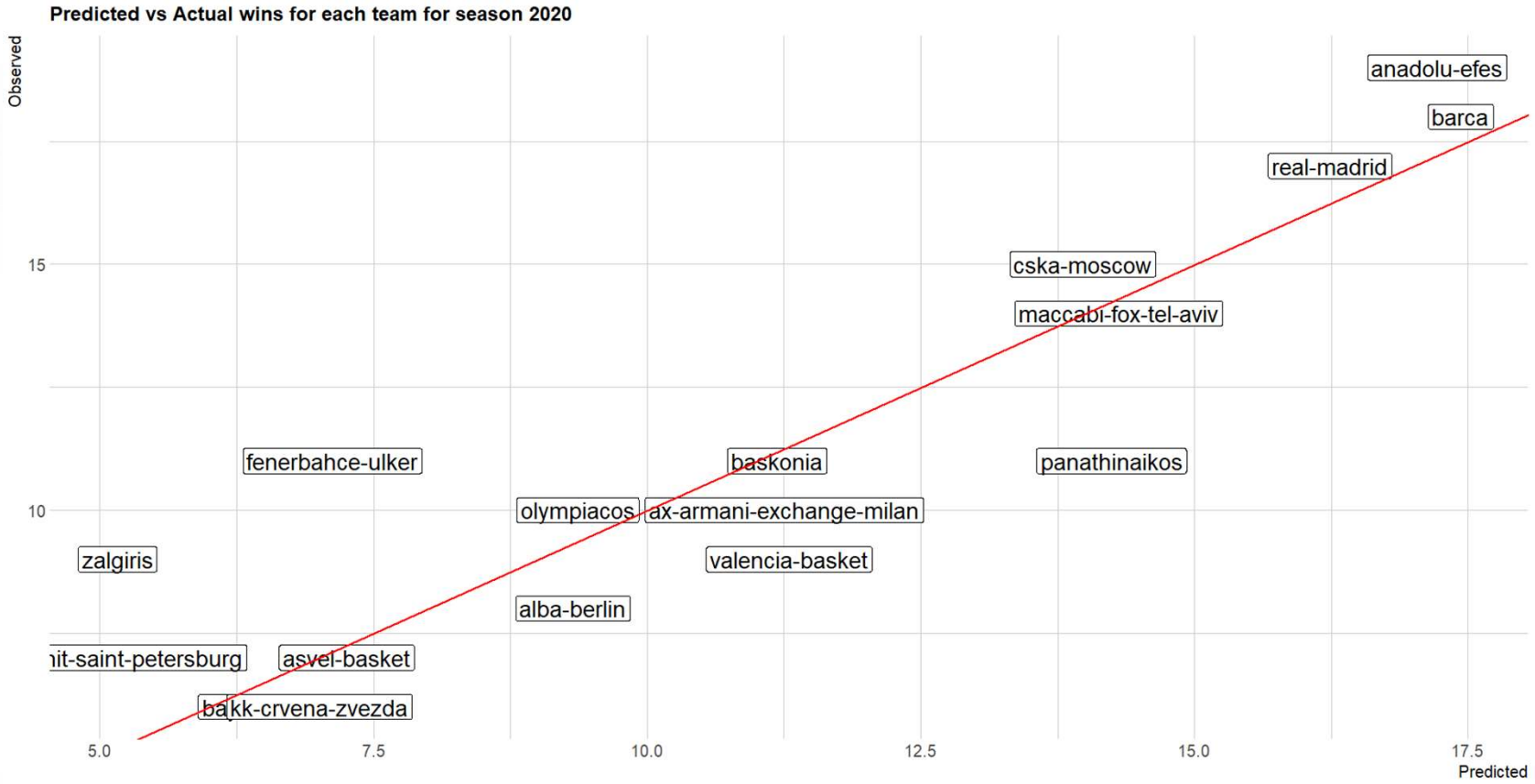
Binomial Logistic with Boxscore Statistics (1/2)

- Best predictors:
  - Stepwise regression based on AIC + LASSO regression
- Weights were used in the model which is a vector of the number of total games (30) represented 48 times, as the number of total teams.

$$W_{per} = \frac{1}{1 + e^{-(-1.27 + 0.07 \times FT + 0.34 \times Three_P - 0.27 \times FGA + 0.26 \times Two_p + 0.46 \times STL - 0.23 \times TOV + 0.24 \times TRB)}}$$

Mc Fadden R-squared	0.79
RMSE (half season)	1.43
RMSE (end season)	1.82
Pearson Correlation	0.89

Binomial Logistic with Boxscore Statistics (2/2)



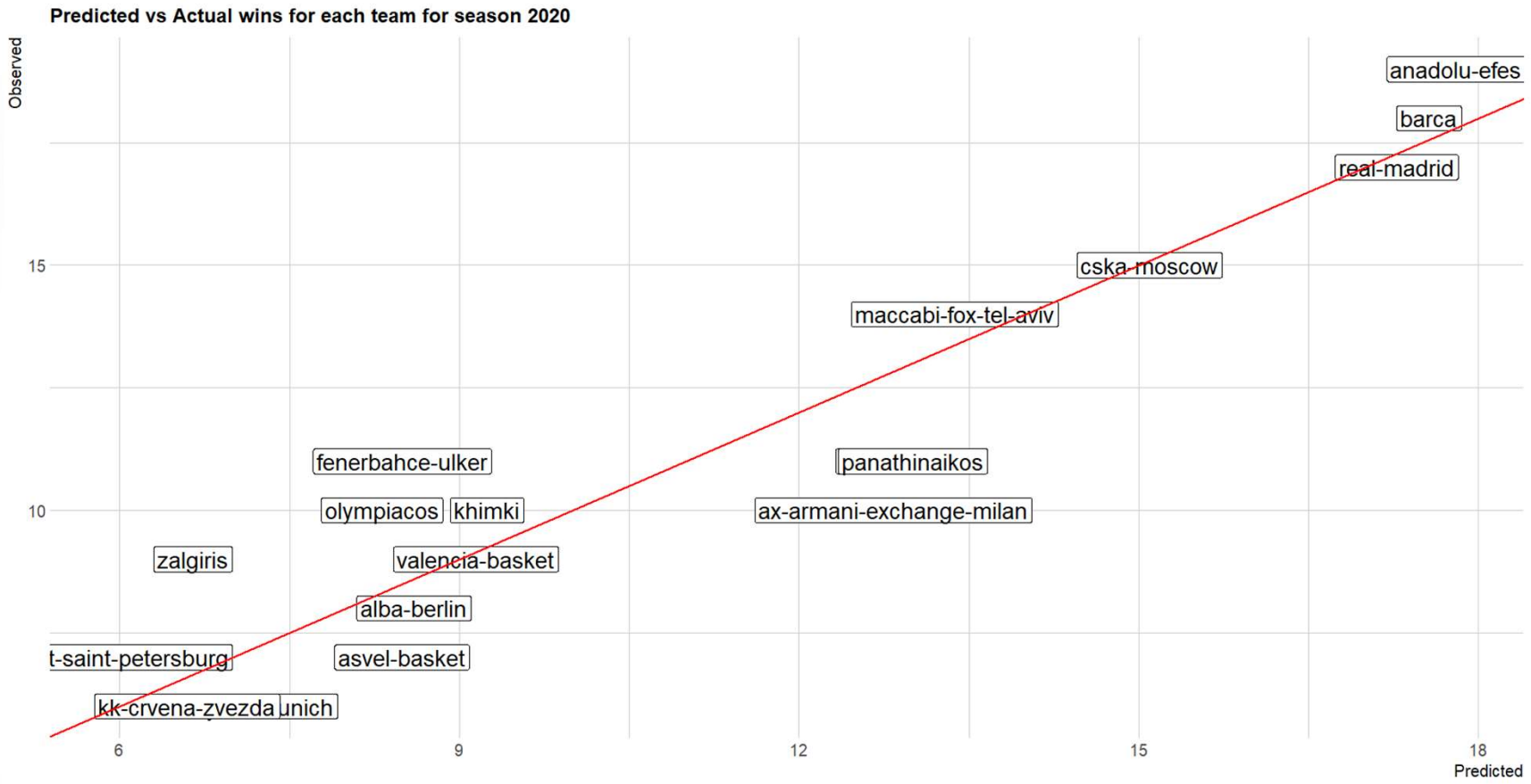
Binomial Logistic with PS & PA (1/2)

- At this stage, since we are using as covariates only the PS and PA, there is no need to perform any variable selection for this model.
- Both features are statistically significant in our model (p-value < .01), except for the intercept.

$$W_{per} = \frac{1}{1 + e^{1.01+0.13\times PS-0.14\times PA}}$$

Mc Fadden R-squared	0.88
RMSE (half season)	1.03
RMSE (end season)	1.45
Pearson Correlation	0.93

Binomial Logistic with PS & PA (2/2)



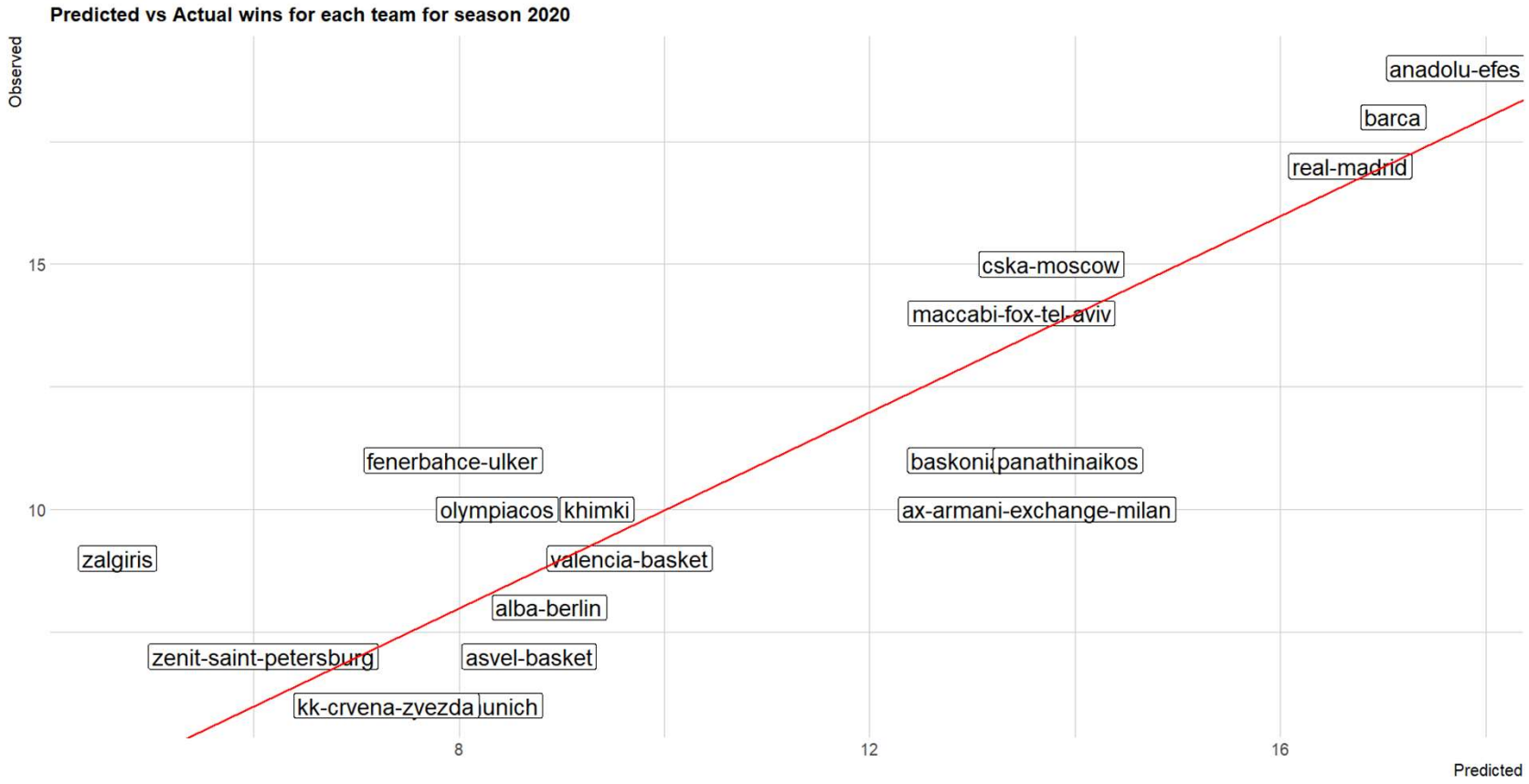
Binomial Logistic at Game Level (1/2)

- LASSO Regression, Stepwise Regression based on AIC and Learning Vector Quantization (LVQ) model were used for feature selection. The ones that provide the highest accuracy and are not correlated with each other were finally selected.
- Target variable is *Winner* which takes values 0 and 1.

$$Winner = \frac{1}{1 + e^{-(0.79 - 0.29\times A_{FG_{per}} + 0.25\times H_{FG_{per}} - 0.07\times A_{3P_{per}} + 0.06\times H_{3P_{per}} + 0.17\times H_{TRB} - 0.16\times A_{TRB} + 0.45\times H_{STL} - 0.48\times A_{STL})}}$$

Mc Fadden R-squared	0.57
RMSE (half season)	0.77
RMSE (end season)	1.99
Pearson Correlation	0.86

Binomial Logistic at Game Level (2/2)



Summary of the models

	Ranking level					Game level
Models	Pythagorean Expectation	Regression (Boxscore)	Regression (PS + PA)	Logistic (Boxscore)	Logistic (PS + PA)	Logistic (Boxscore)
RMSE (half season)	1.71	1.46	1.03	1.43	1.03	0.77
RMSE (end season)	1.50	1.81	1.45	1.82	1.45	1.99
Pearson correlation	0.92	0.90	0.93	0.89	0.93	0.86

Why Pythagorean Expectation?

- Fast and easy to understand
- Uses only Points
- Coaches and team analysts can use these models for inference and prediction
- 1 vs 3 parameters for estimation
- IT ACTUALLY WORKS!
  - Accuracy close to the best ones

Extensions & future work

- Other models to be applied such as Decision Trees
- Vanilla model on game level
- Player's boxscore statistics
- Other metrics to be used as covariates
- A modification to the traditional PE formula

THANK YOU