Modeling and Predicting English Premier League Goal Scoring

Quang Nguyen[1]*
[1] Loyola University Chicago, Chicago, Illinois, USA
[*] Work completed at Wittenberg University, Springfield, Ohio, USA, under *Dr. Douglas Andrews*

Abbreviated abstract: The first purpose of this research was to verify the consistency between goal scoring in the English Premier League and the Poisson process; specifically, the relationships between the number of goals and the Poisson distribution, the time between goals and the exponential distribution, and the time location of goals in a match and the continuous uniform distribution. It turned out to be that the Poisson process and the three probability distributions do perform a great job of describing Premier League goal scoring. In addition, Poisson regression was utilized to predict a Premier League season's results, using different sets of season data and with a large number of simulations being involved.

Related links:
– https://bookdown.org/theqdata/honors_thesis
– https://github.com/qntkhvn/eplgoals

UCSAS
2020

Purpose, Background, and Data

- <u>Poisson Process</u> is used to model the occurrences of events over a continuous interval, with key distributions:
  - Poisson (number of events)
  - Exponential (time between events)
  - Uniform (time position of events)
→ Q: Can goal scoring be modeled by a Poisson Process?

- <u>Poisson Regression</u> is used to model responses that are counts and follow a Poisson distribution (like number of goals)
→ Determine teams' expected scoring rate
→ Simulate and examine results

- Data: **(a)** all EPL final scores from 1992 to 2018 and **(b)** Manchester United's goal scoring times for 2018-19 season

(a)

Season	HomeTeam	AwayTeam	Home.Goals	Away.Goals
2018-2019	Man United	Cardiff	0	2
2018-2019	Southampton	Huddersfield	1	1
2018-2019	Tottenham	Everton	2	2
2018-2019	Watford	West Ham	1	4

(b)

Min	Matchweek	H1_stoppage	H2_stoppage	TimeBetween
3	1	2	5	0
83	1	2	5	82
34	2	5	6	46
95	2	5	6	66

UCSAS 2020

Goal Scoring and the Poisson Process

- Analyze and determine the consistency between:
  - Poisson and number of goals **(1)**
  - Exponential and time between goals **(2)**
  - Uniform and goal times (re-scaled) **(3)**

- Conduct goodness-of-fit tests: Chi-square for **(1)** and Kolmogorov-Smirnov for **(2)** and **(3)**
→ Large p-values → insufficient evidence to conclude that the data do not fit the distributions.
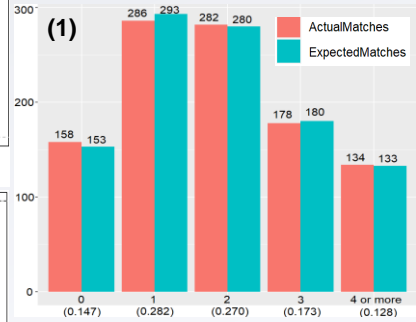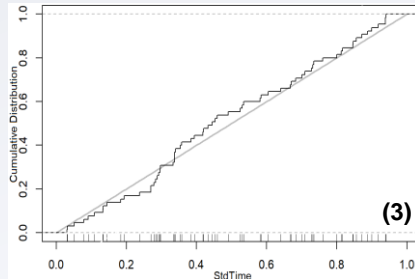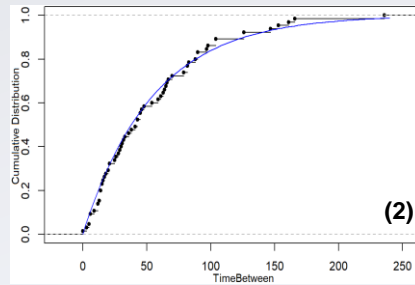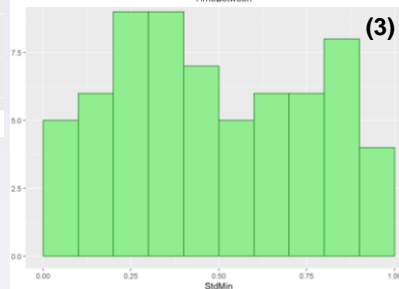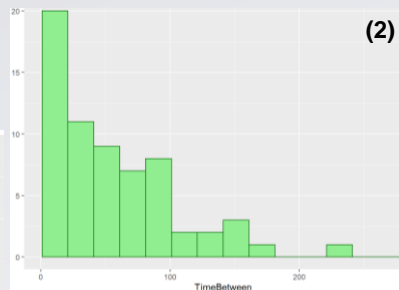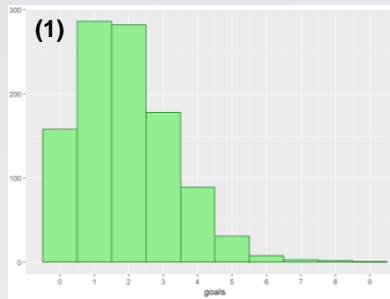
Using Poisson Regression to Predict Season Results

- Fit Poisson regression models (Goals ~ Teams) to get expected scoring rate for home and away games from the models' coefficients

HomeTeam	HomeRate	AwayTeam	AwayRate	HomeScore	AwayScore
Arsenal	2.040462	Bournemouth	1.1447368	1	0
Arsenal	2.040462	Brighton	0.6842105	1	1
Arsenal	2.040462	Burnley	0.8947368	4	2
Arsenal	2.040462	Cardiff	0.6578947	3	2

Teams' scoring rates and randomly generated scores

- Generate 10000 simulated seasons
  - Keep track of all final scores, teams' goals and points
  - Do this for 3 sets of data (all seasons, the 2010s, all seasons but put more weight toward recent years)
  - Compare different metrics (champions, top 4, bottom 3, etc.)

Rank	Team	FinalPoints	GD
1	Tottenham	83	33
2	Liverpool	77	35
3	Everton	68	19
4	Man United	66	19
5	Newcastle	63	18
6	Arsenal	62	7
7	Man City	58	16

Sample simulated season

Team	All Seasons	2010s	Assign Weight
Arsenal	19.68	15.05	14.07
Chelsea	14.28	11.70	9.03
Liverpool	12.50	10.96	17.61
Man City	7.00	41.71	38.09
Man United	36.99	10.47	10.53
Tottenham	3.91	7.38	8.34

Chances (%) of winning the 2018-19 season

UCSAS 2020