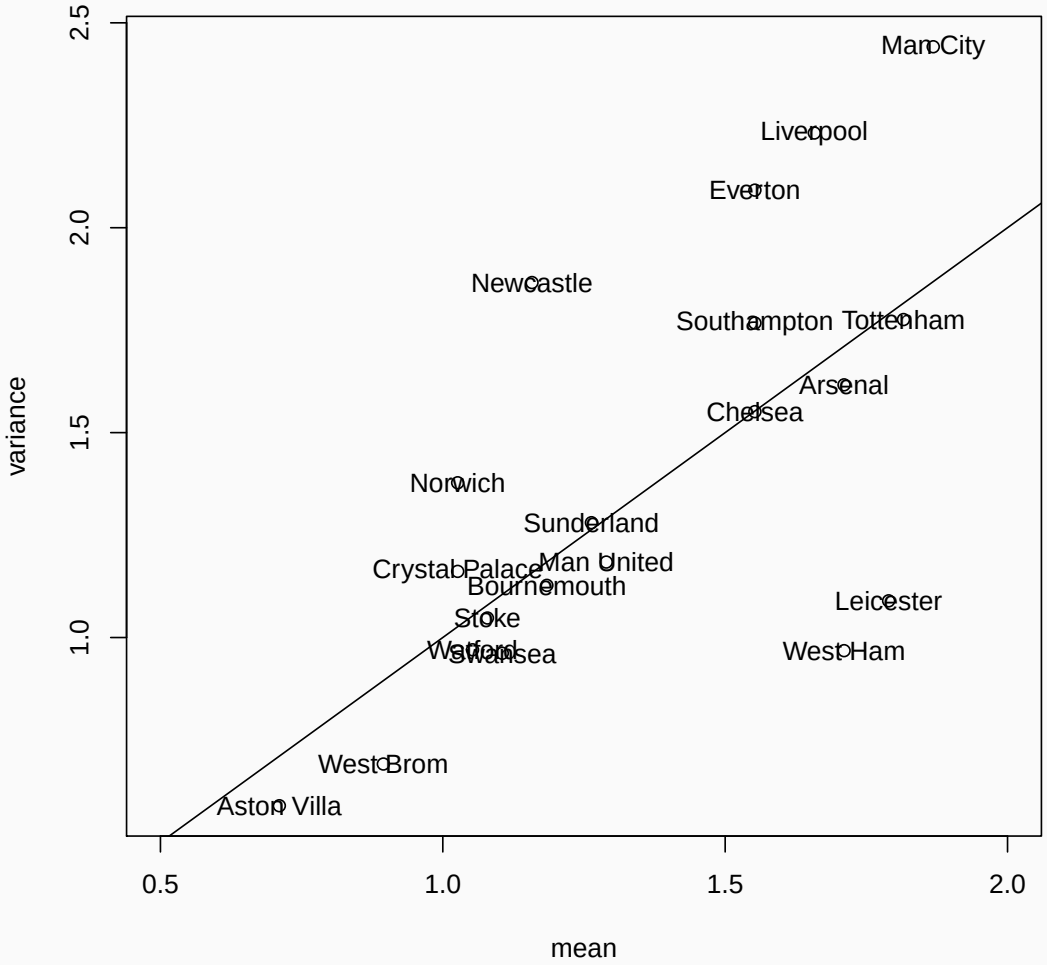
Inplay Model based predictions for football

Dimitris Karlis

Department of Statistics
Sports Analytics Group
AUEB
karlis@aueb.gr

Tzougas and Karlis presentation in brief

Premier League Data 2015-2016



The paper in brief

- We need a model to capture overdispersion
- But overdispersion is not the same for all teams
- We need to model the overdispersion in a neat way
- A solution can be to model both the mean and the overdispersion.varince parameter
- We also need a flexible distribution to do so (perhaps allowing underdispersion
- Among other we apply P-LN motivated by the shape (but we admit that this does not allow for underdispersion)

Contributions

- Flexible model for mean and variance of the PLN
- EM algorithm to fr the model
- Consider several models for the overdispersion (e.g. common for all teams, grouping of tams based on flexible methods etc)
- Comparison with other models

Main findings

- Better fit that NBI regression model
- We need to group teams with different overdispersion
- Improve predictions

Modelling the outcome of a football game - a quick overview

- Model win-loss (no score included)
  - Paired comparison models
  - Logistic and ordinal regression
  - Artificial intelligence models
- Model score
  - Double Poisson model and variants
  - Bivariate models
  - Inflated models
  - Advanced models
- Modelling the difference

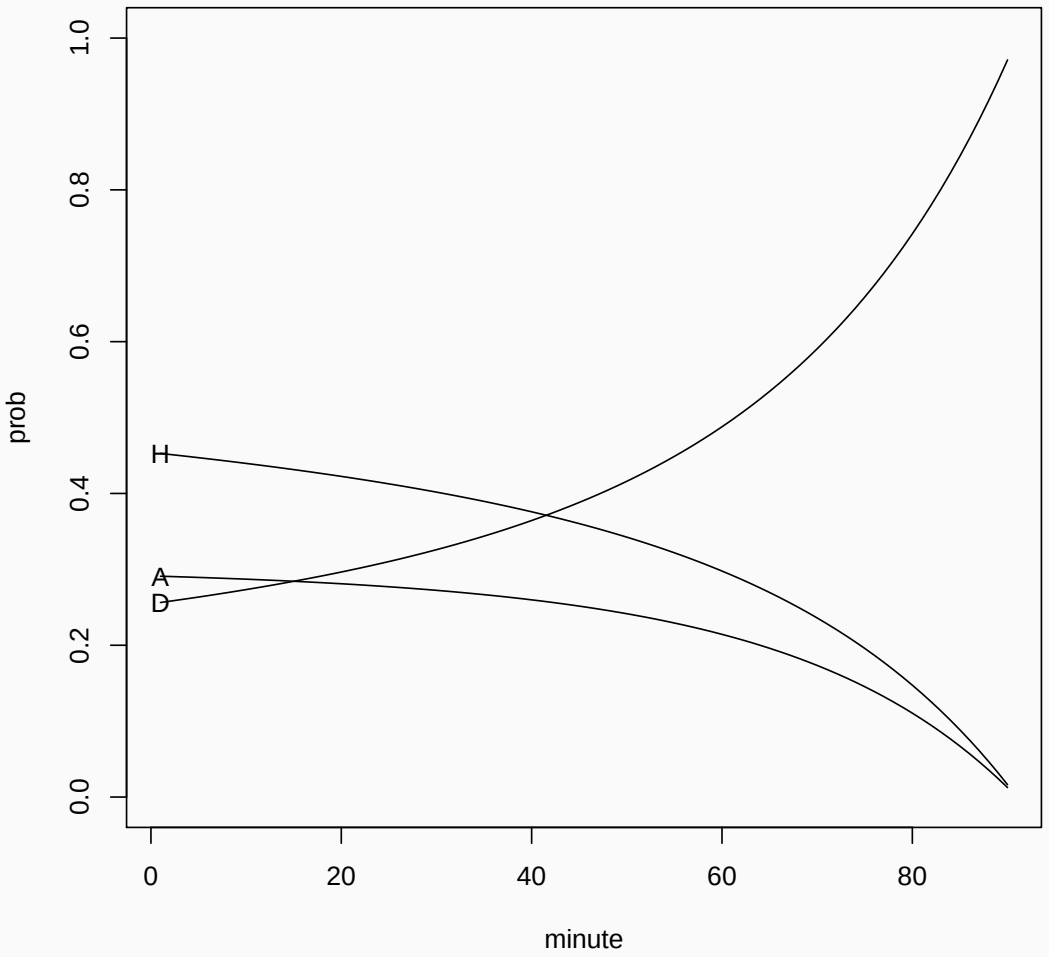
Existing models - before them

- An important aspect lies on the fact that some of the models are used for exploratory usage: i.e. what statistics may influence the score, e.g. is ball possession a predictor?
- But some models are predictive:we care on predicting the outcome for the forthcoming matches. In this case some variables cannot be used as they are not known a priori.
- It is important to separate between them, however as models they may share some common elements.

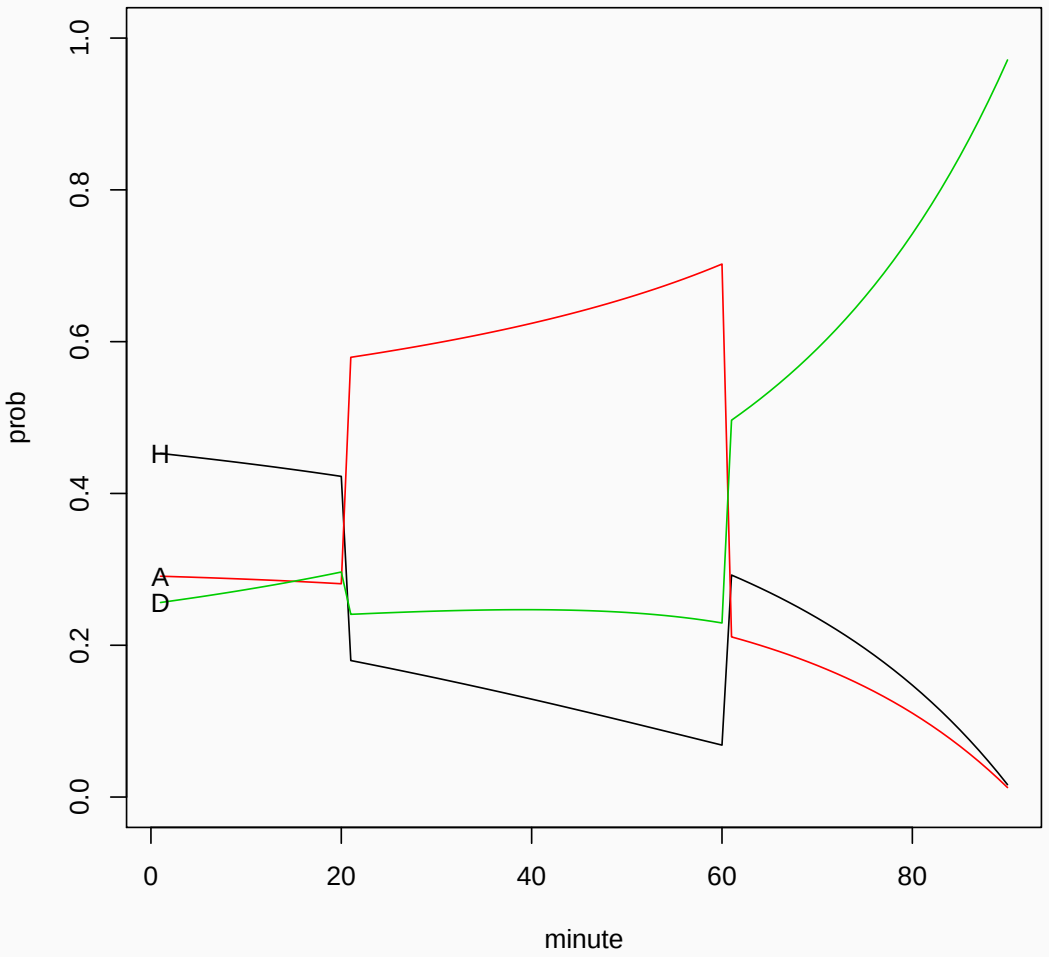
In-Play prediction

- We want to model the final outcome conditional on some information during the game. E.g. what is the probability to win if the score at 20' is 1-0?
- What kind of information could be useful? Is this information available?
- Are the current models models useful for this purpose and how we could amend them?

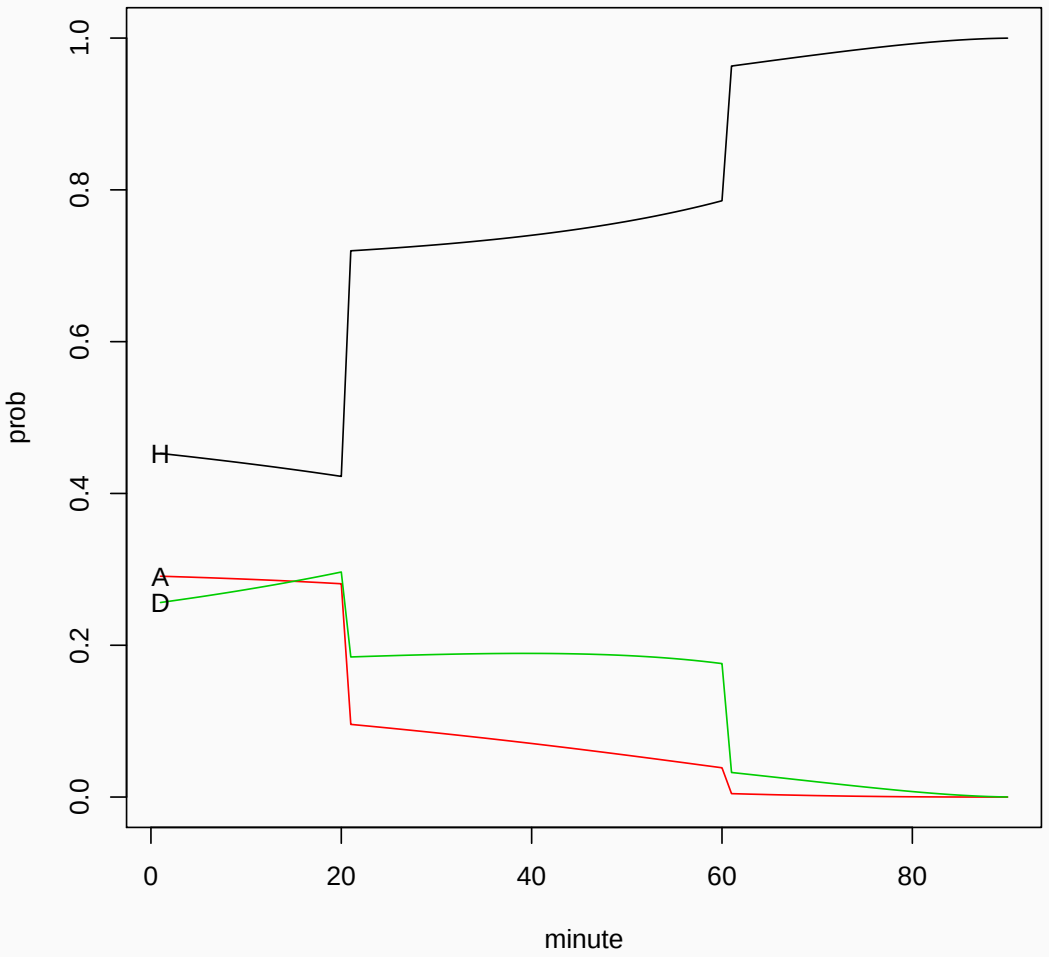
Probability H-A-D as time passes, 0-0



Probability H-A-D as time passes, 1-1



Probability H-A-D as time passes, 2-0

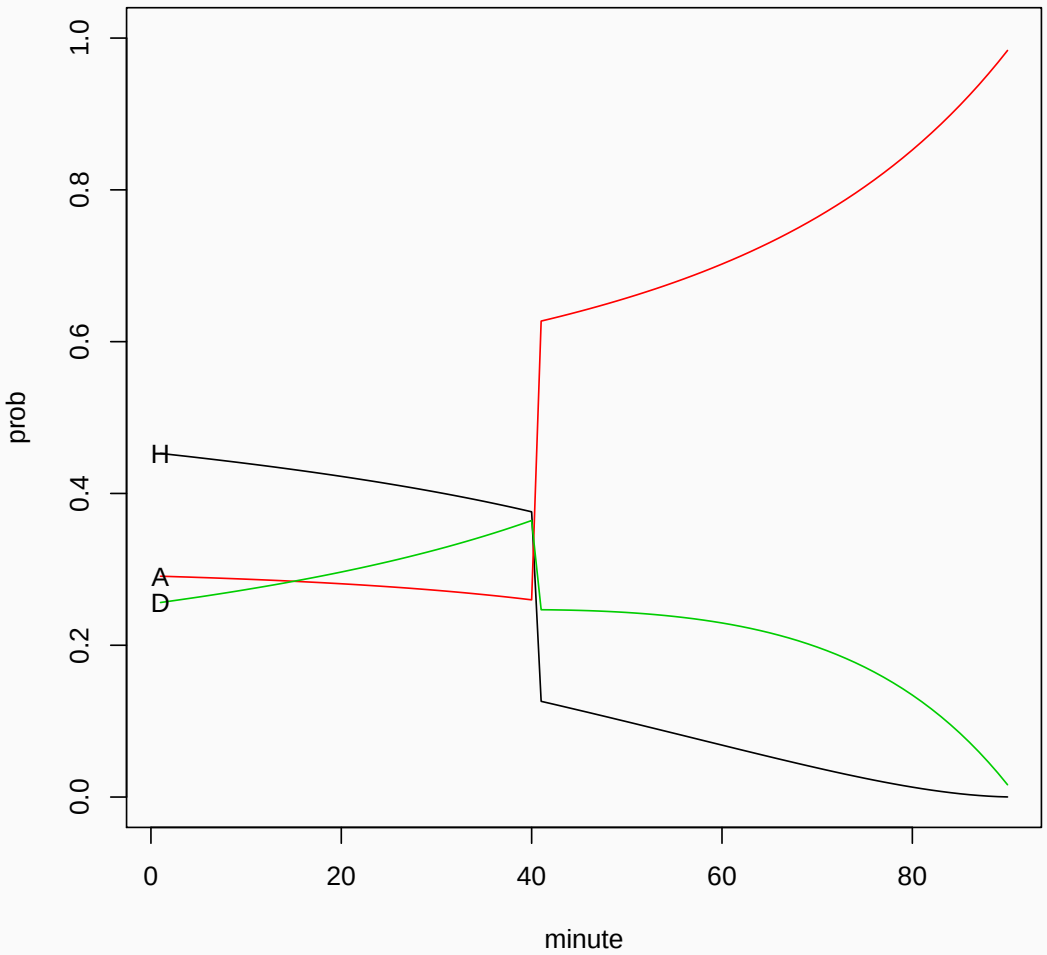


But what if the probabilities change

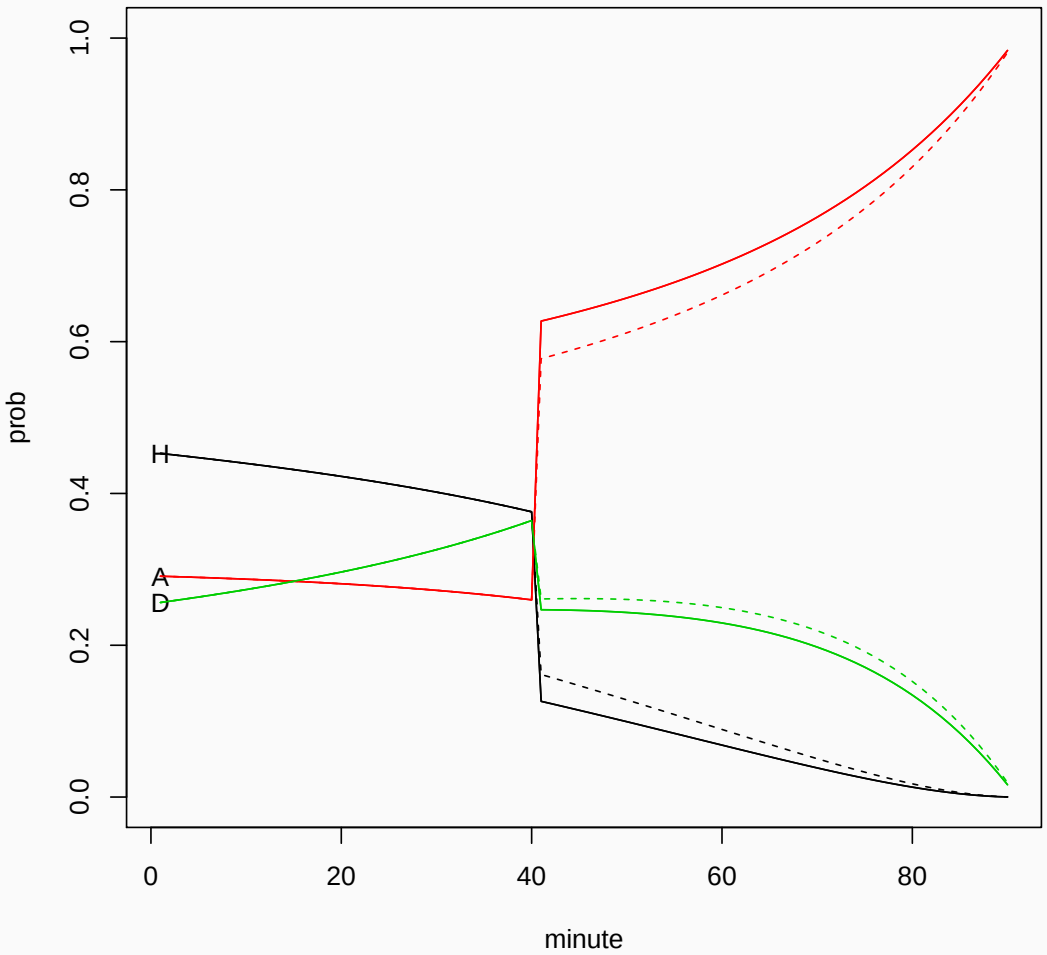
So far

- The probabilities change just because the team has to score more.
- The scoring ability remains constant across time, irrespective the change sin the game
- We will now alter the probabilities by
  - Team behind the score need to play more offensive hence increasing the scoring ability
  - Team increases its scoring ability with time (a favorite that need to win as time passes)

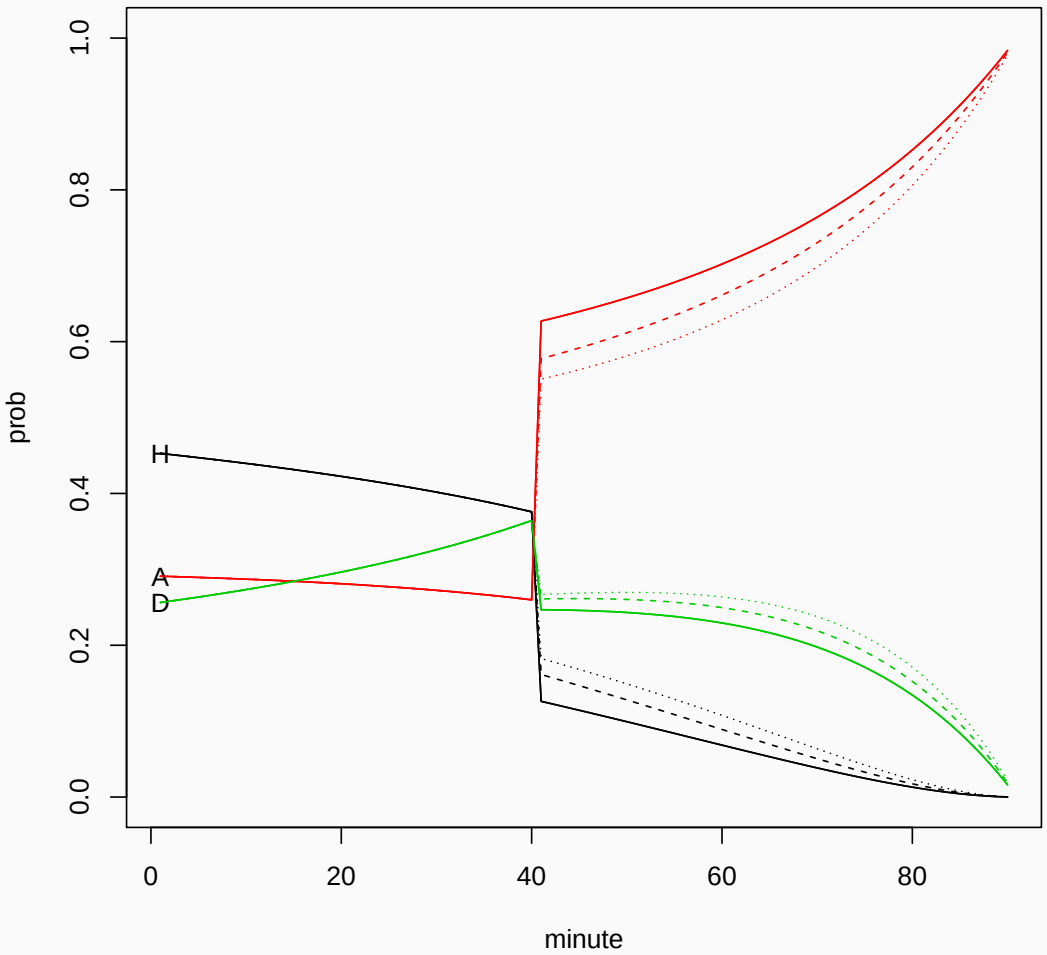
Probability H-A-D as time passes



Home team plays more offensive



Also increasing its scoring ability as time passes



Lessons learned

- Current models assume a constant goal rate
- This is not realistic
- We need to investigate the factors that affect this during the game
- Altering the scoring ability we end up with different probabilities
- All we need to assume is that the scoring rate λ f a team depends on the time, the score and some events insidte the game, being as $\lambda(t; z)$.

Existing literature

- There is an increase on this kind of prediction, mainly due to betting purposes
- Online betting is an important fraction of the current business, expected to increase
- Small published work on this: mainly n issues about market efficiency
- Half time-Full time score prediction based on a 4-variate model (Poisosn with copulas)
- Dobson and Goddard (2017) using survival models
- Mainly models based on stochastic process and time to event ideas

Time to event models

- Model he time till next goal (see Ntzoufras and Karlis, 2015)
- We assume that certain events during the match alter the expected time of the next goal
- We may assume different type of processes to model this
- Note: what kind of information we need?
- Dobson and Goddard (2017) using survival models with some covariates information

Some covariates

They used as covariates

- a measure of the relative quality of the two competing teams, (e.g. based on the betting prices for the match result prior to the start of the match)
- the number of minutes of the match currently elapsed,
- dummy variables indicating the current goal difference between the two teams,
- any difference between the numbers of players on the pitch owing to red cards already incurred

A new Approach

- Count data model are not appropriate.
- Recall that we can approximate them (Poisson as an approximation of a binomial)
- We propose a new approach. We split the game in intervals of one minute nd we model directly the probability of scoring a goal at that minute
- Current models assume constant probability

Binomial vs Poisson approach

Probabilities from a Poisson($\lambda = 2$) and a Binomial with $p = 2/90$ and $n = 90$.

Goals	Binomial	Poisson
0	0.132	0.135
1	0.271	0.271
2	0.274	0.271
3	0.182	0.180
4	0.090	0.090
5	0.035	0.036
> 5	0.016	0.017

Model

Assume the standard logistic regression approach. Split the game to a sequence of 90+ minutes, we have 0-1 outcome based on whether a goal is scored by team j against team k.

A goal can alter this probability. for example

- the team which is behind at the score may increase the probability of scoring a goal
- The time played can be also a factor (fatigue)
- Red cards?
- Other events inside the game
- "Good" teams have a record of scoring the last minutes

Data example

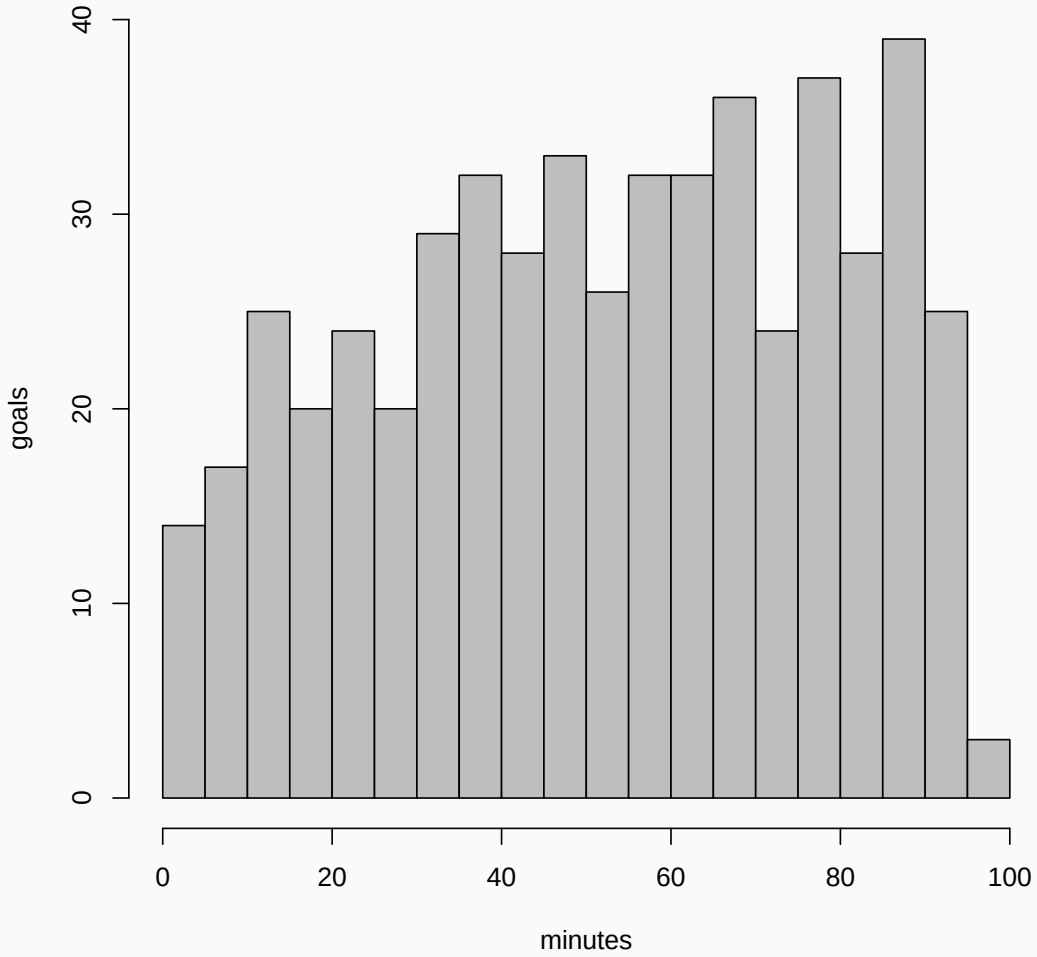
Suppose that A plays against B at his own home. the score is 3-1 with goals scored at 12 (1-0), 15 (1-1), 45 (2-1), 76 (3-1).

O1	O2	minutes	goal	diff	home
A	B	12	1	0	1
B	A	12	0	-0	0
B	A	3	1	-1	0
A	B	3	0	1	1
A	B	30	1	0	1
B	A	30	0	-0	0
A	B	31	1	1	1
B	A	31	0	-1	0
A	B	17	0	2	1
B	A	17	0	-2	0

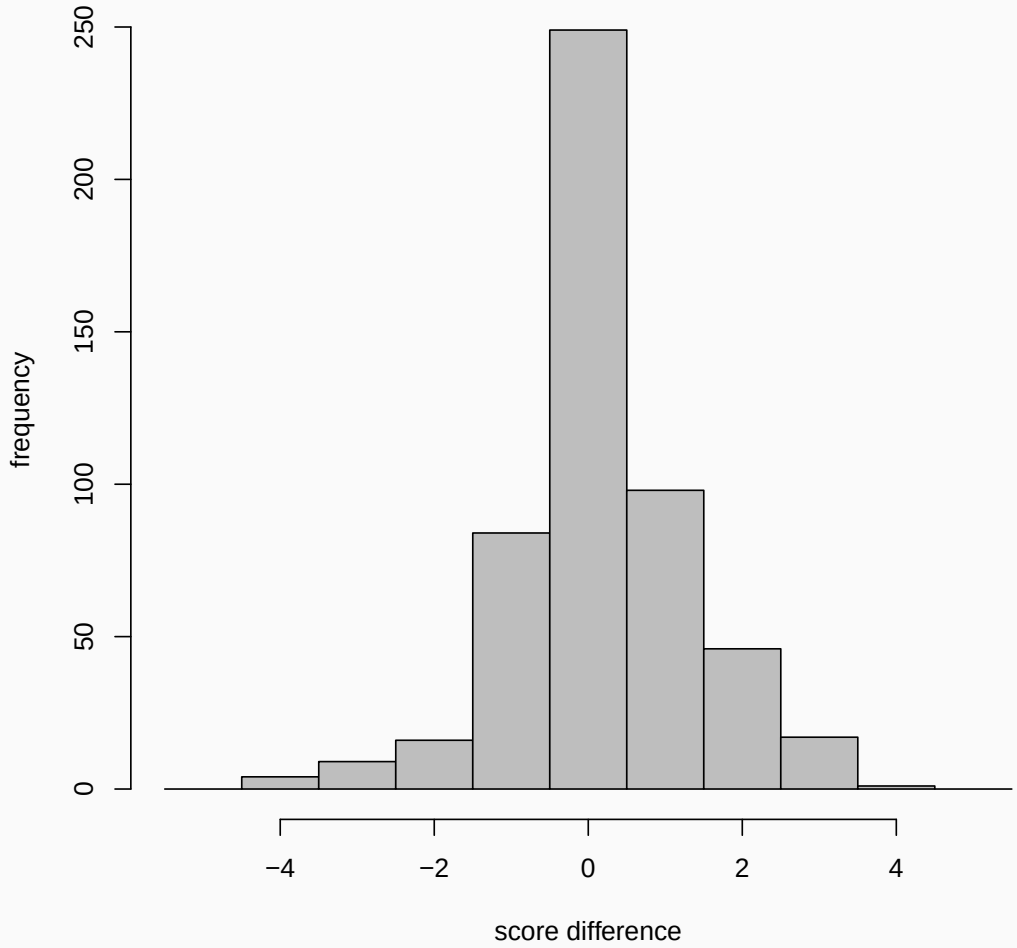
Application - Superleague 2017-2018

- We use the 239 matches of the Superleague 2017-2018.
- Excluded the match that never played (PAOK - Olympiakos) and also we used the data up to the moment of the game played for PAOK-AEK.
- We want to check assumptions like: Is the current score important? Are the red cards important? Last minutes? Other events?
- Data were constructed manually from the SuperLeague web site
- If no info for the extra time was given we used 3 minutes.

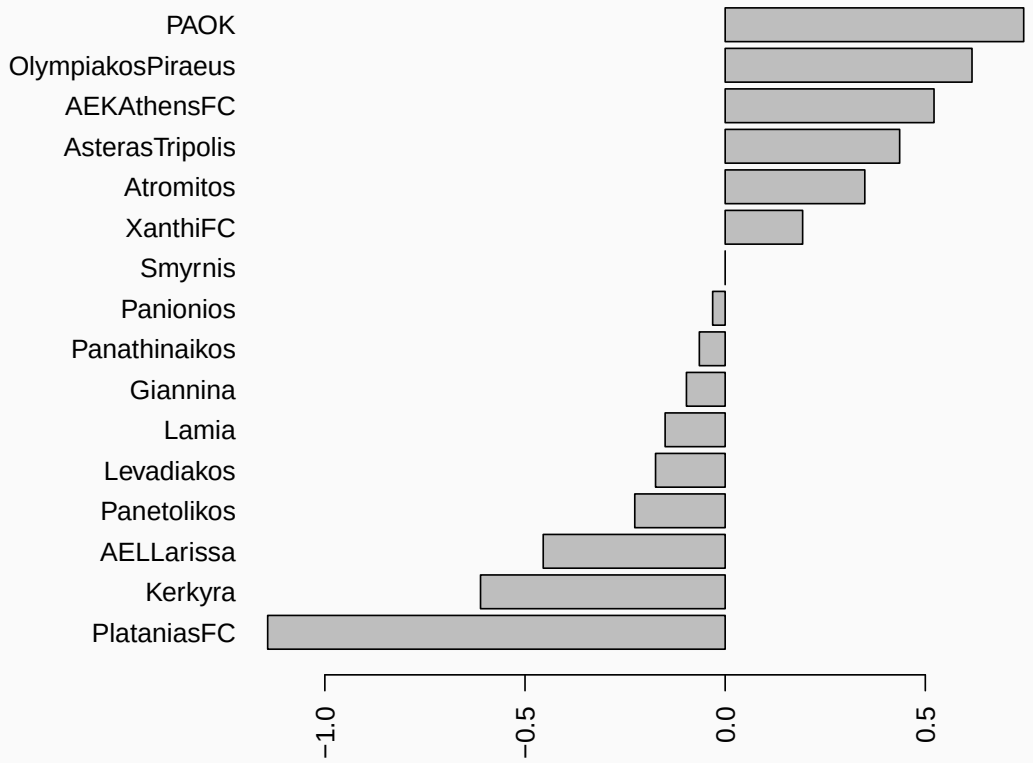
Minutes of goals



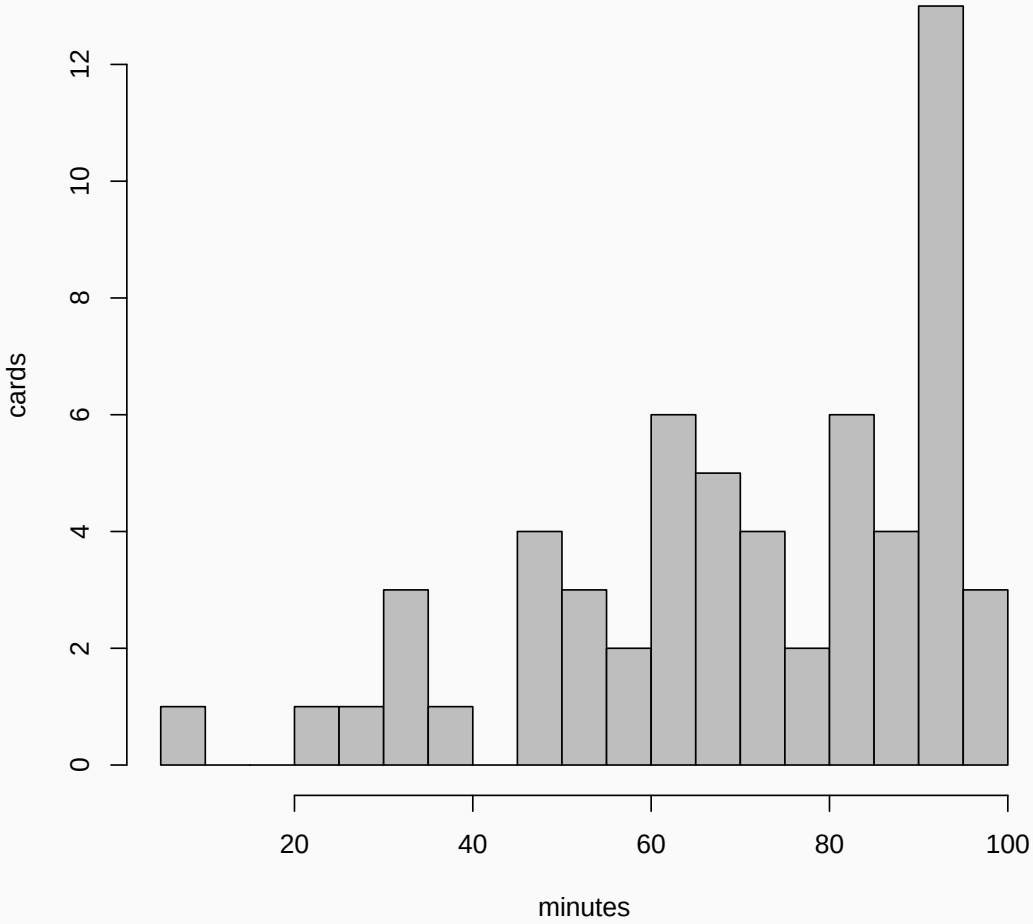
Score Difference when scoring



Score Difference when scoring - per team



Red Cards



Results

Using different covariates in the model.

Basic model: Home + Offensive ability + Defensive Ability + Covariates

Model	Effect	Result
1	Score Diff	Significant increase
2	Score Diff rounded	Significant increase
3	Different Team effect	no difference
4	Red Card	not significant
5	Last 10 minutes	Significant increase only with diff

Best model based on AIC: Model 2

PAOK-AEK

The most controversial match of the year. The score was 0-0 at 90th minute. PAOK scored a goal, the referee cancelled it after some hesitation. The match never continued.

Predictions from the model (up to this week)

	0	1	2	3	4
0	0.215	0.117	0.033	0.005	0.000
1	0.229	0.104	0.026	0.005	0.001
2	0.112	0.054	0.012	0.002	0.001
3	0.042	0.017	0.004	0.000	0.000
4	0.011	0.005	0.001	0.000	0.000
5	0.001	0.001	0.000	0.000	0.000

PAOK-AEK

https://www.youtube.com/watch?v=09SsZtvK_gw

What would have happened if the game was continued?

- Probabilities before the match:
  PAOK : 0.477, Draw: 0.331, AEK: 0.192
  (averaged over 10000 runs)
- Given the score was 0-0 at 90 and assuming 5 minutes extra time:
  PAOK: 0.0536, Draw: 0.9207, AEK: 0.0257
- If the goals was counted
  PAOK: 0.952, Draw: 0.046, AEK: 0.002

Betting

- An obvious application of the model is on online betting
- The model updates the probabilities based on certain events and this can be used to update the odds
- Note that odds are necessary following the probabilities but they may have other business aspects

Further Comments

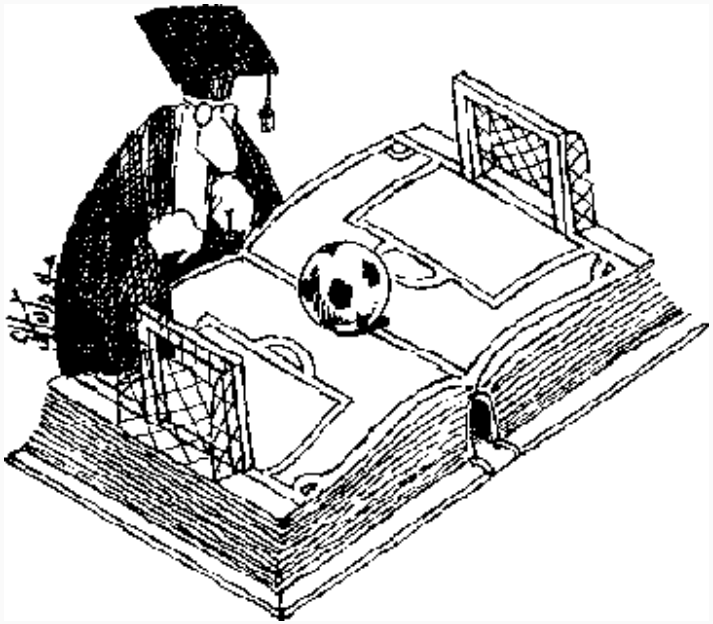
- What are the events that can be considered as adding information? We mean events during the game. E.g. is the ball position at the last minutes such a predictor? Some injuries? Substitutions? Accumulated fatigue? Spatial information? Shots to goal? Corners? what else?
- Are such data available?
- Can the model be improved? E.g. other link functions or/and other assumption (like beta binomial, copulas based models etc)
- Predictions is based on simulating large series, not easy to derive in closed forms
- How extra time is taking into account?

Final Points

- There is an increasing demand on statistical models for soccer (and other sports) prediction, from various sources, not only betting
- Ongoing work relates to a model that can capture many of the interesting characteristics-bets
- Covariate determination is important
- Online prediction has its own interest but it is much more difficult.

Some Literature

- Hoog, E. (2014) Modelling prices of in-play football betting markets
- Dobson, S., & Goddard, J. (2017). Evaluating probabilities for a football in-play betting market. The Economics of Sports Betting, 52.
- Asif, M., & McHale, I. G. (2016). In-play forecasting of win probability in one-day international cricket: A dynamic logistic regression model. International Journal of Forecasting, 32(1), 34-43.
- Divos, P., del Bano Rollin, S., Bihari, Z., & Aste, T. (2018). Risk-Neutral Pricing and Hedging of In-Play Football Bets. Applied Mathematical Finance, 1-21.
- Croxson, K., & James Reade, J. (2013). Information and efficiency: Goal arrival in soccer betting. The Economic Journal, 124(575), 62-91.



THANKS