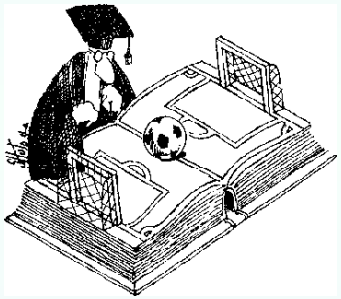
Robust fitting of soccer prediction models



Dimitris Karlis and Ioannis Ntzoufras

Department of Statistics, Athens University of Economics and Business

2009 @ Gronigen, Netherlands, 2nd International Conference on Mathematics in Sport

Outline

- Motivating Example
- Existing Statistical Models
- Robustness issues
- Weighted ML estimation
- Application

Existing models

There is a wealth of potential models for the number of goals scored by each team .
Such as:

- Double Poisson
- Negative Binomial
- Bivariate Poisson
- Inflated model
- Copula based
- etc

But what is their behavior if the data contain outliers, i.e. unexpected large scores?

A Motivating Example

Consider the following data of the first group of Champions league for 2008–9.

Team	Avg.	Avg.	Avg.	Probability (%)			
Team	Points	GF	GA	1st	1-2	2.5-3	3.5-4
AS Roma	11.8 (12)	11.9	6.0	49.2	87.1	10.4	2.5
Chelsea	11.1 (11)	9.0	5.1	36.9	82.7	13.7	3.6
Bordeaux	4.7 (7)	5.0	11.0	1.0	7.4	33.0	59.6
CFR 1907 Cluj	5.6 (4)	5.0	8.9	1.8	13.5	42.6	43.9

Motivating Example

Consider the following data of the first group of Champions league for 2008–9.
Let's now change the result of the game

Chelsea - Cluj = 2–1 $\Rightarrow$ 5–1

Team	Avg.	Avg.	Avg.	Probability (%)			
Team	Points	GF	GA	1st	1-2	2.5-3	3.5-4
AS Roma	11.7 (12)	12.2	6.0	40.4	90.1	8.9	1.0
Chelsea	12.4 (11)	11.9	5.0	50.1	92.1	7.2	0.7
Bordeaux	4.9 (7)	5.0	11.1	0.8	6.7	43.4	49.9
CFR 1907 Cluj	4.6 (4)	5.1	12.1	0.5	4.6	34.7	60.7

Motivating Example

Consider the following data of the first group of Champions league for 2008–9.
Let's now change the result of the game

Chelsea - Cluj = 2–1 ⇒ 5–1

Differences

Team	Avg.	Avg.	Avg.	Probability (%)			
Team	Points	GF	GA	1st	1-2	2.5-3	3.5-4
AS Roma				−8.8	3.0	−1.5	−1.5
Chelsea	+1.3	+2.9		+13.2	9.4	−6.5	−2.9
Bordeaux						+10.4	−9.7
CFR 1907 Cluj	−1.0		+3.2	−1.3	−8.9	−7.9	+16.8

($|\text{changes}| < 0.3$ are omitted)

What do we learn?

- Existing models mostly try to fit the number of goals and this can be misleading especially when few games are used.
- The ML methods usually reproduces well the scoring ability, so accidentally scoring a lot of goals in one game will increase the power of a team.
- While in general this is soccer the model can loose their robustness as they may be influenced from some scores

Need to improve on this robustness issue

Robustness

Robustness in estimation prevents from

- Outliers: an unexpected high score may influence a lot the results
- Model deviations: If the assumed model is not well specified we need to protect against it.

Robust estimators are not necessary efficient and hence a trade-off between robustness and efficiency is usually needed. There are various robustness approaches in the literature.

Existing Robustness issues

Approaches

- Down weight some observations
- Use some kind of robust estimators like M-estimators or some minimum distance estimators

Lindsay (1994) has shown that in discrete data models robustness and efficiency can be achieved almost at the same time, i..e by appropriately defining distances that in some sense down weight some observations

Robust Estimates using Weighted log-Likelihood

We propose a rather simple and easy to fit approach. Theoretical results will be announced elsewhere.

The key idea: Games with large and unexpected scores must be downweighted.

Use a weighted likelihood of the form

$$L_w = \sum w_i \log f_i(x_i, y_i; \theta)$$

where w_i is the weight given in the i-th game, x_i, y_i represent the number of goals scored in this game, f_i is the assumed model and θ denotes the parameters of interest

Robust Estimates using Weighted log-Likelihood

We separate two cases

- $w_i = w(x_i, y_i)$, i.e. the weight depends on the score only but perhaps it is even independent of the score representing for example that the game was indifferent for the two teams or any other factor. The standard approach assumes $w_i = 1$.
- $w_i = w(x_i, y_i, \theta)$ i.e the weight given in each observation depends on the assumed model

The first is easier and perhaps somebody with good knowledge can select weights. For example to represent the indifference of some teams etc. Note also the use of the weights to downweight matches played a long time ago (Dixon and Coles, 1997)

Robust Estimates using Weighted log-Likelihood

We define the weights based on the difference in the score.

$$w_i = \begin{cases} 1 & \text{if } |x_i - y_i| < m_0 \\ p & \text{otherwise} \end{cases}$$

for $p < 1$. i.e. we assume that for a score with large score difference larger than m_0 we put less trust on this

Very simple to handle

Model Based weighting

The idea is to fit the model with standard ML approach and then by looking at the fitted values one may downweight observations that had small probability to occur.

Let $\hat{f}(x_i, y_i; \hat{\theta})$ be the estimated probability for a match based on the ML estimate $\hat{\theta}$ derived in the usual way we may define the weights as a fuction of this probability

$$w_i(x_i, y_i, \hat{\theta}) = h\left(\hat{f}(x_i, y_i; \hat{\theta})\right)$$

Model Based weighting (2)

A simple possible choice for $h(x)$ is $h(x) = f(x)^q$ for $q \geq 0$.

- Observations that seems to be not relevant to the model are down weighted.
- q controls the volume of weighting.
  - For $q = 0$ we have no weighting (all weights equal to one; usual MLE).
  - As q increases we tend to give more weight to central values and less to outliers resulting in a robust estimate.

More advanced weighting scheme may use also the observed frequency information and downweight observations that occur more frequently than they would be expected.

Actually the method relates to minimum distance estimation by appropriately adjusting the weighting function.

Fitting

We have selected an easy to use weighting scheme. For real data applications one must know that:

- Standard packages can be used for fitting the model.
- Maximization can be easily achieved numerically
- Even an EM type algorithm can be easily extended.
- Use of IRLS (Iterated Reweighted Least Squares, Wang, 2007)

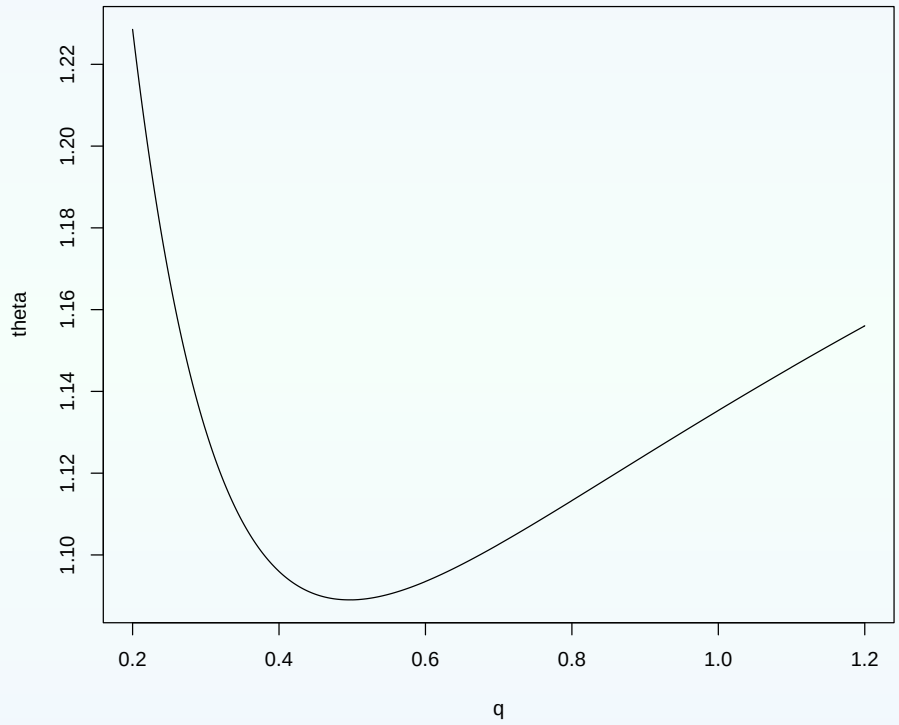
A Simple Toy Example

Consider the following toy example with only 9 observations

$$0, 0, 0, 1, 1, 1, 2, 3, 10$$

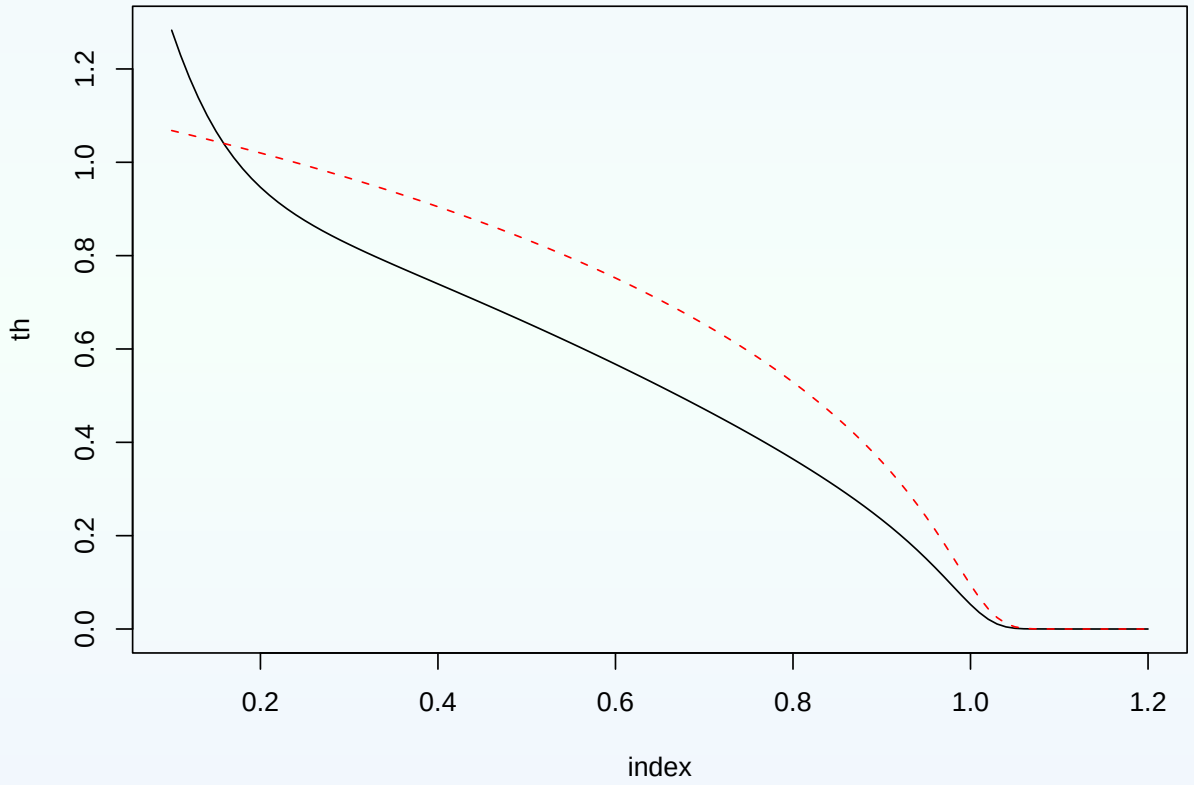
- The last observation clearly looks as an outlier
  - Estimated Sample Mean = 2
  - Estimated Sample Mean (excl. last observation) = 1
- If we use weighting with $w_i = f(x_i, \hat{\theta})^q$ i.e. observations that seems to be not relevant to the model are down weighted.
- q controls the volume of weighting.
  - For $q = 0$ we have no weighting (all weights equal to one; usual MLE).
  - As q increases we tend to give more weight to central values and less to outliers resulting in a robust estimate.

The following plot demonstrated the behavior of the estimated θ for each value of q (with one iteration).



Estimated Sample Mean (excl. last observation) $= 1 \Rightarrow$ i.e. for $q \in (0.4, 0.6)$ we have sample mean close to the one observed without the outlier.

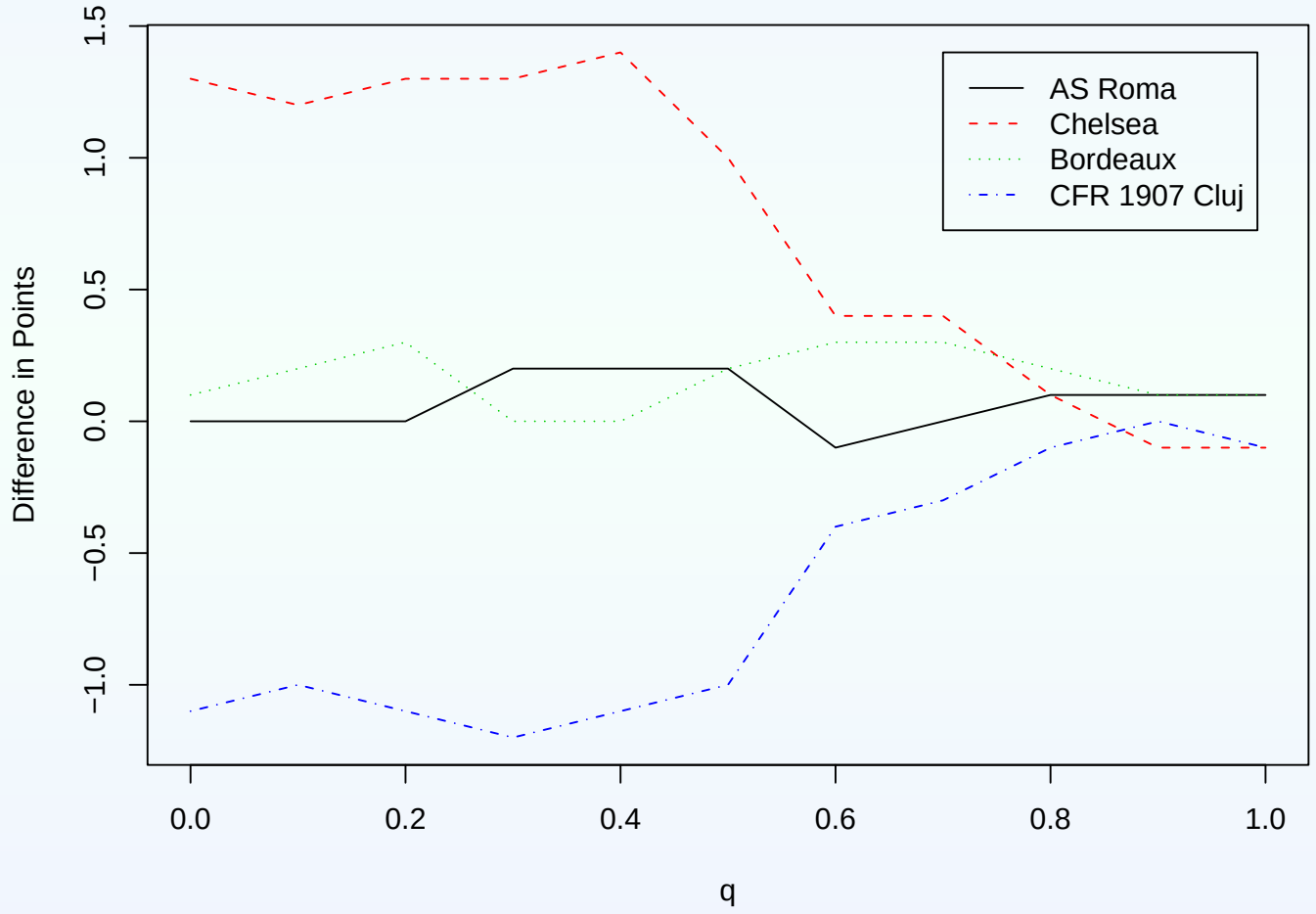
The following plot demonstrated the behavior of the estimated θ for each value of q (with many iterations). Red line indicates estimates for data with $10 \rightarrow 3$.



For $q \approx 0.2$ we have sample mean close for both data.

Comparison of Results for Champions League Group 1

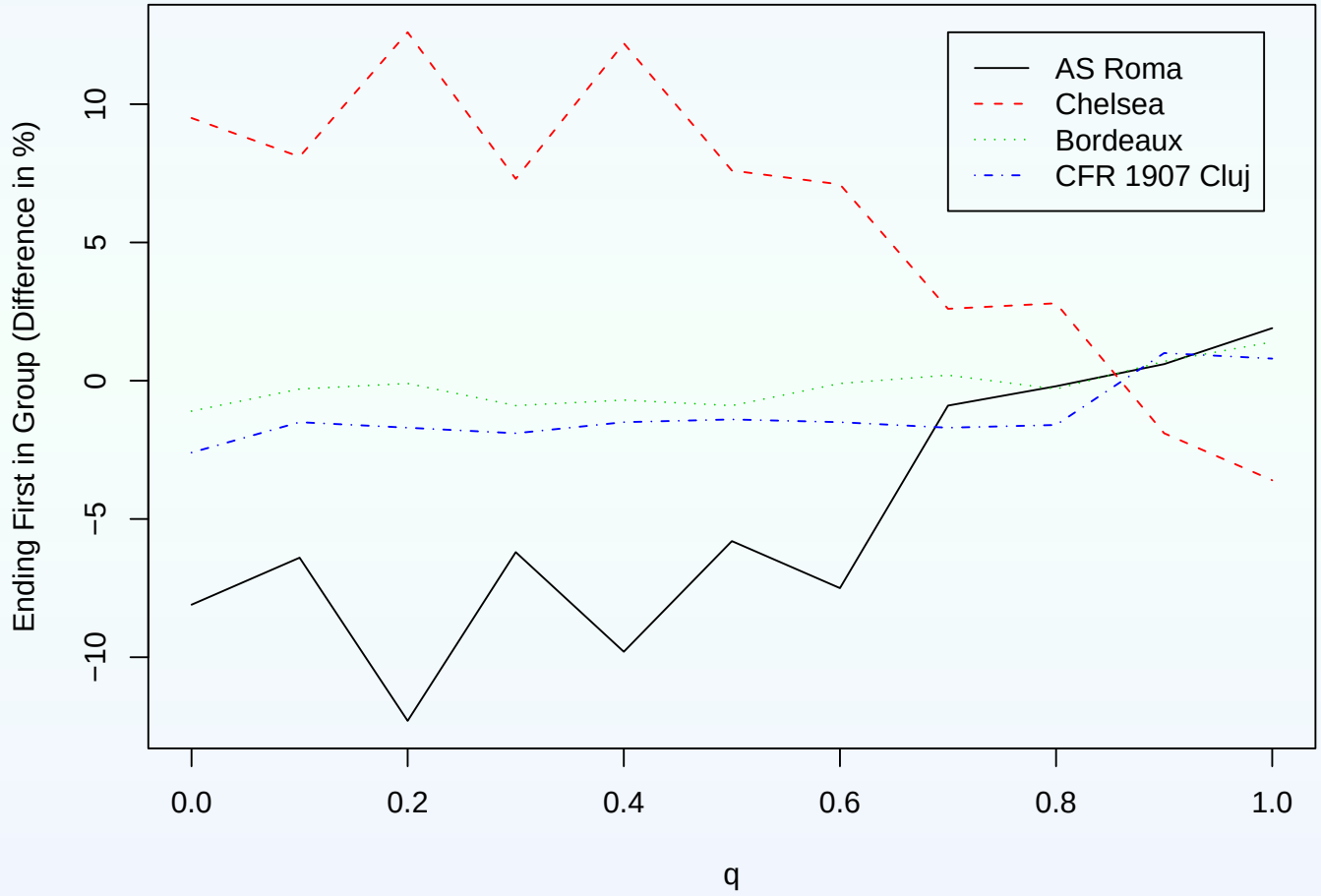
(Data with outlier vs. Original data)



Difference in Points ≈ 0 for $q \geq 0.8$.

Comparison of Results for Champions League Group 1

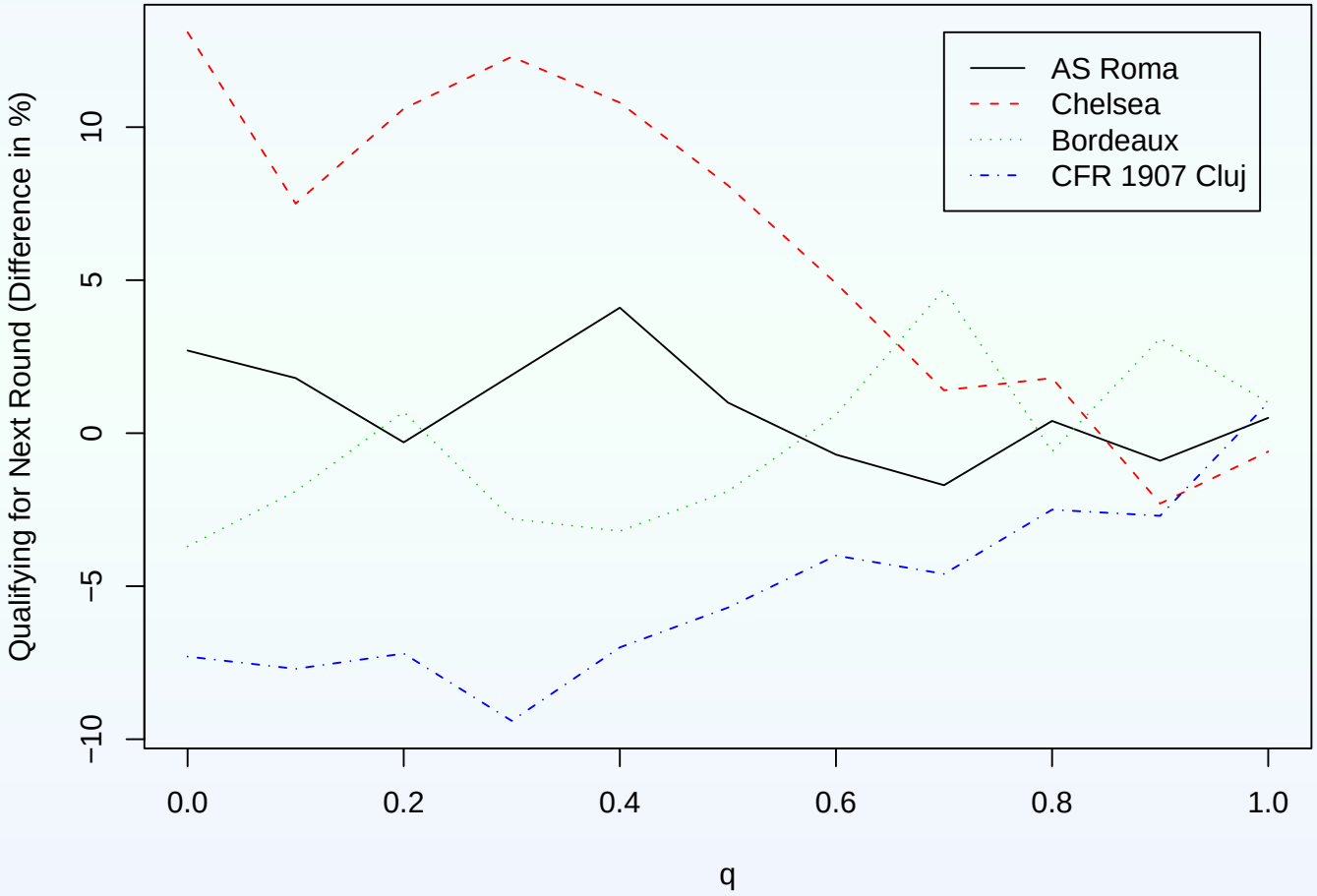
(Data with outlier vs. Original data)



Difference in Probability of First Place (%) ≈ 0 for $q \geq 0.8$.

Comparison of Results for Champions League Group 1

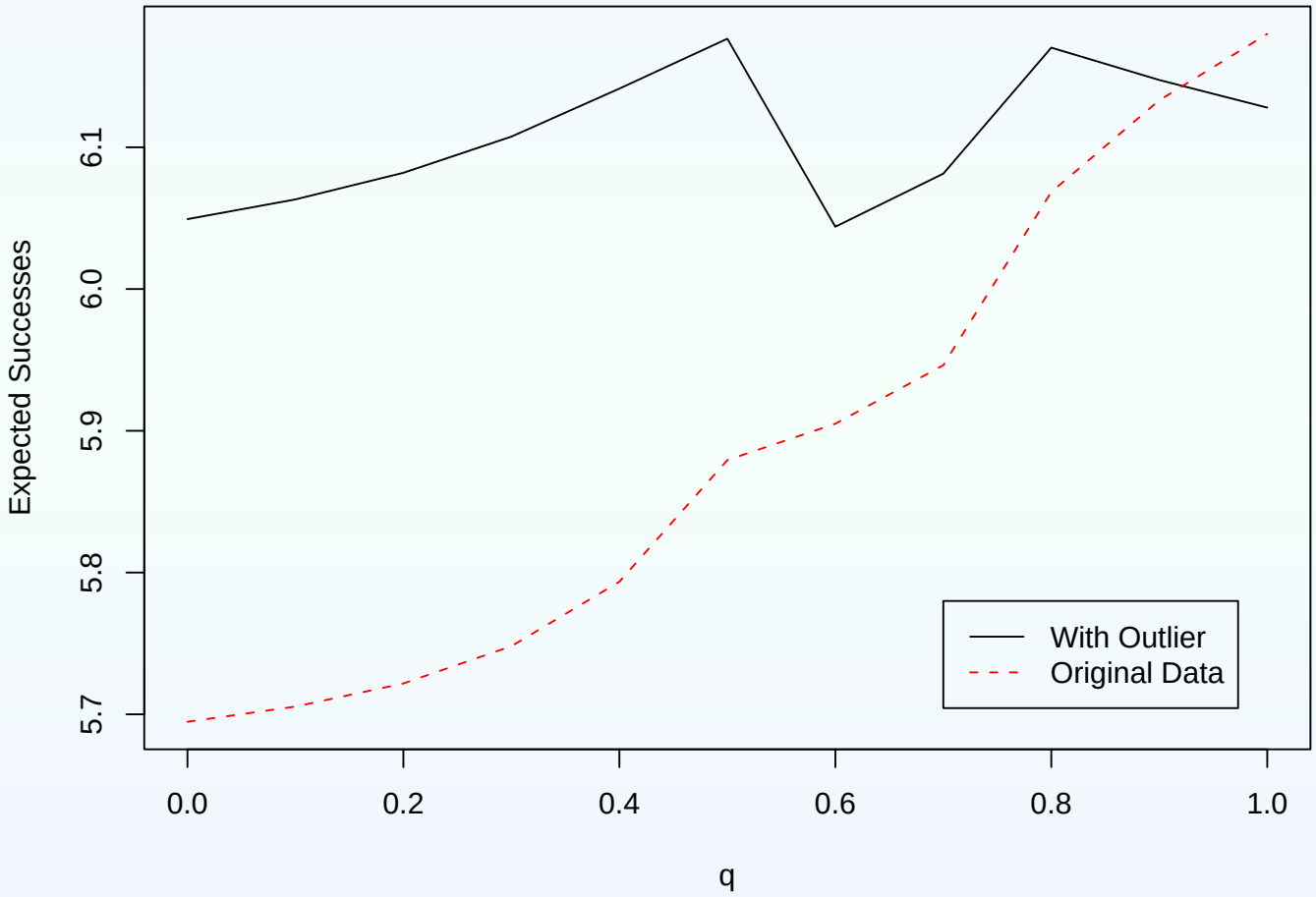
(Data with outlier vs. Original data)



Difference in Probability of Qualifying to the Next Round (%) ≈ 0 for $q \geq 0.8$.

Comparison of Results for Champions League Group 1

(Data with outlier vs. Original data)



Number of expected successes is close for two datasets for $q \geq 0.8$.

Comparison of two datasets with q=0.8

Original Data

	Av.Pts	Av.Goals1	Av.Goals2	Pr.1st	Pr.Q	Pr.2.5-3	Pr.3.5-4
AS Roma	8.2	4.0	3.5	11.2	54.2	31.2	14.6
Chelsea	11.5	9.3	1.5	74.0	95.0	4.4	0.6
Bordeaux	6.4	3.1	8.2	2.4	22.7	37.7	39.6
CFR 1907 Cluj	5.7	1.7	5.0	1.4	13.5	29.0	57.5

Data with Outlier

	Av.Pts	Av.Goals1	Av.Goals2	Pr.1st	Pr.Q	Pr.2.5-3	Pr.3.5-4
AS Roma	8.2	4.1	3.6	8.6	56.9	30.6	12.5
Chelsea	11.7	9.5	1.4	77.7	97.0	2.7	0.3
Bordeaux	6.3	3.0	8.3	2.3	20.3	37.4	42.3
CFR 1907 Cluj	5.6	1.7	5.0	0.5	11.7	31.6	56.7

Application: 2008-9 Champions League Data

For ith game with home team HT_i against AT_i we have expected counts λ_{1i} and λ_{2i} respectively.

Usual Model Structure:

$$\begin{aligned} \log \lambda_{1i} &= \mu + home + att_{HT_i} + def_{AT_i} \\ \log \lambda_{2i} &= \mu + \qquad\quad att_{AT_i} + def_{HT_i} \end{aligned}$$

will be appropriate for each group separately but not for prediction in the next knock-out round.

Application: 2008-9 Champions League Data

Reasons for the failure of the above model: The design is incomplete, teams of different groups are isolated leading to non-identifiable parameters.

Solution: Find variables that connect teams of different groups.

Proposed Solution: Use common attacking and defensive for teams of same countries (may be also random effects).
Use UEFA ranking and scores to discriminate between different teams.

Advantage: Makes the model identifiable since it carries information from different groups.

Disadvantage: UEFA score is based on the performance of the previous year.

Application: 2008-9 Champions League Data

For ith game with home team HT_i against AT_i we have expected counts λ_{1i} and λ_{2i} respectively.

Proposed structure:

$$\begin{aligned}\log\lambda_{1i} &= \mu + home + co.att_{CH_i} + co.def_{CA_i} + \beta_1 UEFA_{HT_i} + \beta_2 UEFA_{AT_i}\\ \log\lambda_{2i} &= \mu + co.att_{CA_i} + co.def_{CH_i} + \beta_1 UEFA_{AT_i} + \beta_2 UEFA_{HT_i}\end{aligned}$$

- $co.att_k$, $co.def_k$: team and defensive parameters for countries coming from country k.
- CH_i, CA_k: country for home and away team (respectively) in game i.
- $UEFA_k$: UEFA score ranking for team k.
- β_1, β_2: Attacking and defensive parameters related to uefa ranking
- HT_i, AT_i: home and away team (respectively) in game i.

Application: 2008-9 Champions League Data

Comparison of the two models:

- AIC/BIC indicate the proposed model as better.
- R^2 type of statistic shows that goodness of fit is similar for the two models.

	Usual Model	Proposed Model
AIC	761.9	733.6
BIC	981.2	858.9
R2	38.6%	30.2%

CLD 2008-9: Double Poisson - 1st Approach

Scale down results with high goal difference.

- Model 1 - Usual (Unweighted) Maximum likelihood model.
- Model 2 - WL with Weights: $w_i = 0.5$ if $|d_i| \geq 3$ and $w_i = 1$ otherwise.
- Model 3 - WL with Weights: $w_i = 0.5$ if $|d_i| = 3$, $w_i = 0.5$ if $|d_i| \geq 4$ and $w_i = 1$ otherwise.

where

$$d_i = goals_{HT_i} - goals_{AT_i}$$

is the goal difference and

$$y_i = I(d_i > 0) + 2 * I(d_i = 0) + 3 * I(d_i < 0)$$

is the final outcome of the game (1:home wins, 2:draw and 3:home looses).

Predicted outcome probabilities are presented in three tables

- First Knock out round (Phase of 16 teams) using data from groups.
- Quarter-finals using previous data (from groups and first KO round).
- Semi-Finals using previous data and prediction of the winner.

In each phase we use the measure

$$\sum_i \sum_{k=1}^{3} p_{ik} I(k = y_i)$$

which denotes the success rate of each model.

CLD 2008-9: Double Poisson

Results for 1st KO round (phase of 16) using Groups resutls.

Model		Expected Successes	% Relative to Saturated
1.	ML	6.26	57.8
2.	WML1 ($w_i = 0.75, 0.5, 1$)	5.96	55.1
3.	WML2 ($w_i = 0.5, 1$)	5.92	54.7
4.	WML3 ($w_i = \sqrt{f(x_i)}$)	5.72	52.8
6.	Saturated	10.82	100.0

- Expected number of successes is similar but we now have a robust model.

CLD 2008-9: Double Poisson

Results for Q-Finals using results of previous rounds.

Model		Expected Successes	% Relative to Saturated
1.	ML	2.72	63.1
2.	WML1 ($w_i = 0.75, 0.5, 1$)	2.71	62.8
3.	WML2 ($w_i = 0.5, 1$)	2.70	62.6
4.	WML3 ($w_i = \sqrt{f(x_i)}$)	3.05	70.7
6.	Saturated	4.31	100.0

- Expected number of successes is similar but we now have a robust model.

CLD 2008-9: Double Poisson

Results for Semi-Finals

Game	Score	Models 1-3	Model 4
Barcelona - Chelsea	0-0	41.6 22.5 35.9	
		41.4 22.6 36.1	
		40.9 23.0 36.1	29.5 31.2 39.3
Manchester United - Arsenal	1-0	40.4 29.3 30.3	
		39.4 29.2 31.4	
		39.3 29.2 31.5	29.2 49.2 21.6
Chelsea - Barcelona	1-1	51.4 21.4 27.2	
		48.3 22.0 29.8	
		47.8 22.4 29.8	54.7 27.5 17.8
Arsenal - Manchester United	1-3	42.1 29.0 28.9	
		40.7 29.1 30.2	
		40.6 29.1 30.3	30.3 49.0 20.7

CLD 2008-9: Double Poisson

Estimated probabilities for the final.

	Model	Barcelona	Draw	Und
1.	ML	37.9	25.9	36.1
2.	WML1 ($w_i = 0.75, 0.5, 1$)	38.2	25.7	36.1
3.	WML2 ($w_i = 0.5, 1$)	37.8	26.1	36.1
4.	WML3 ($w_i = \sqrt{f(x_i)}$)	25.5	36.2	38.3

CLD 2008-9: Bivariate Poisson (1 iteration)

Results for Semi-Finals

Game	Score	Models 1-3	Model 4
Barcelona - Chelsea	0-0	37.7 23.3 39.0	
		36.4 23.0 40.5	
		36.2 22.7 41.1	35.7 27.9 36.4
Manchester United - Arsenal	1-0	38.5 31.4 30.1	
		37.9 32.0 30.0	
		38.3 31.8 29.9	33.4 40.4 26.2
Chelsea - Barcelona	1-1	55.1 21.2 23.7	
		56.3 20.8 22.9	
		57.8 20.1 22.1	54.5 24.8 20.6
Arsenal - Manchester United	1-3	39.9 31.2 28.9	
		39.1 31.9 29.0	
		39.4 31.7 28.9	34.3 40.3 25.4

CLD 2008-9: Bivariate Poisson (1 iteration)

Estimated probabilities for the final.

	Model	Barcelona	Draw	Und	$\log \lambda_3$
1.	ML	34.5	26.8	38.7	-1.189
2.	WML1 ($w_i = 0.75, 0.5, 1$)	33.1	26.6	40.3	-1.092
3.	WML2 ($w_i = 0.5, 1$)	32.5	26.5	41.0	-1.073
4.	WML3 ($w_i = \sqrt{f(x_i)}$)	30.6	32.7	36.7	-1.34

Extensions

Consider a typical robust estimator, the one based on Minimum Hellinger distance of the form

$$\sum_x \left(d(x)^{1/2} - f_\beta(x)^{1/2} \right)^2$$

where $d(x)$ is the observed relative frequency (i.e. a simple estimate of the probability at x) and $f_\beta(x)$ is the assumed model with parameters of interest β. It turns out that this quantity leads to estimating equations of the form

$$\sum_x \left(\frac{d(x)}{f_\beta(x)} \right)^{1/2} \frac{\partial f_\beta(x)}{\partial \beta} = 0$$

directly comparable to the ML estimating equations

$$\sum_x \frac{d(x)}{f_\beta(x)} \frac{\partial f_\beta(x)}{\partial \beta} = 0$$

which actually implies that we weight the observations differently. The above formula assumes no covariates. If covariates are present actually we have $d(x) = 1$ since each observation can have different covariates

Extensions (2)

The idea generalizes as in Lindsay (1994). Let

$$\delta(x) = \frac{d(x) - f_\beta(x)}{f_\beta(x)}$$

and let $A(\delta)$ an increasing function twice differentiable in $[-1, \infty)$ with $A(0) = 0$ and $A'(0) = -1$ (usually called the Residual Adjustment function) then the estimating equations can take the form

$$\sum_x A(\delta(x)) \frac{1}{f_\beta(x)} \frac{\partial \log f_\beta(x)}{\partial \beta} = 0$$

where β are the parameters of interest.

Since $A(\delta)$ is increasing for δ this means that at points where δ is large, $A(\delta)$ is also large and hence $f_\beta(x)$ must be small, in simpler words for x where the observed data disagree with the assume model we must give less weight.

Discussion

Summary

- We proposed a simple weighted likelihood approach to achieve robustness against unexpectedly large scores
- The approach is simple to use in practice.
- Extension to more complicated weights is straigthforward.

Open problems

- What is the appropriate weight to have robustness and increased success rate?
- If $w = f(x)^q$ what value of q is optimal? In the regression $q = 1/2$ was suggested. In our motivating example $q = 4/5$ was much better providing robust results and increased success rate.
- Can we use weighted regression to weight accordingly past results?