Journal of Modern Applied Statistical Methods

Volume 15 | Issue 2 Article 17

11-1-2016

Evaluation of the Addition of Firth's Penalty Term to the Bradley-Terry Likelihood

Paul Meyvisch
Janssen Pharmaceutica NV, Beerse, Belgium, pmeyvisc@its.jnj.com

Recommended Citation
Meyvisch, Paul (2016) "Evaluation of the Addition of Firth's Penalty Term to the Bradley-Terry Likelihood," *Journal of Modern Applied Statistical Methods*: Vol. 15 : Iss. 2 , Article 17.
DOI: 10.22237/jmasm/1478002500

Evaluation of the Addition of Firth's Penalty Term to the Bradley-Terry Likelihood

Cover Page Footnote

The author wishes to acknowledge Prof. Dr. Geert Molenberghs for his advise and support.

Journal of Modern Applied Statistical Methods
November 2016, Vol. 15, No. 2, 224-237.
doi: 10.22237/jmasm/1478002500

ISSN 1538 − 9472

Evaluation of the Addition of Firth's Penalty Term to the Bradley-Terry Likelihood

Paul Meyvisch
Janssen Pharmaceutica NV
Beerse, Belgium

A major shortcoming of the Bradley-Terry model is that the maximum likelihood estimates are infinite-valued in the presence of separation, and may be unreliable when data are nearly separated. A well-known solution consists of the addition of Firth's penalty term to the log-likelihood function, and solving this penalized likelihood through logistic regression. The maximum likelihood estimates with and without Firth's penalty are compared in a large and heterogeneous population of table tennis players, showing that exact penalized maximum likelihood estimates can be reasonably approximated using a well-chosen Minorization-Maximization (MM) algorithm.

Keywords: Bradley-Terry, Firth, MM algorithm, table tennis

Introduction

Consider the evaluation of the addition of Firth's penalty term to the Bradley-Terry likelihood function, with an application to a large dataset of table tennis players. The problem of rating table tennis players falls into the topic of binary paired comparison modeling, provided the victory margin is ignored. A binary paired-comparison experiment is used to assess the relative worth of t objects even though they can only be compared two at a time, and when the result of such a comparison can only be that one of the objects is preferred to the other. Zermelo (1929) is generally credited with being the first to address the problem of estimating the strengths of players. The model and various parts of the theory have been rediscovered over the intervening years and were first described in detail by Bradley & Terry (1952).

Suppose there are m players and define $\pi = (\pi_1, \ldots, \pi_m)'$ to be the vector of the player's strengths. The Bradley-Terry model assumes that the probability p_{ij} of player i defeating player j is:

Paul Meyvisch is a Biostatistics Director at Janssen Pharmaceutica NV, an affiliate of Johnson & Johnson, and a doctoral student at I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium. Email him at pmeyvisc@its.jnj.com.

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \tag{1}$$

Any constant multiple of the strengths π_i estimates also satisfy (1), so they can be scaled to satisfy an additional constraint such as $\sum_i \pi_I = 1$ or $\pi_I = 1$ for sake of identifiability.

If each pair of players i and j plays n_{ij} games against each other, with player i winning v_{ij} times and losing d_{ij} times, and all games assumed independent, the likelihood takes the form:

$$L(\pi) = \left[\prod_{i=1}^{m} \prod_{j:j\neq i} \frac{\pi_i^{v_{ij}} \pi_{ij}^{d_{ij}}}{\left(\pi_i + \pi_j\right)^{n_{ij}}} \right]^{\frac{1}{2}}, \tag{2}$$

where $v_{ij} = d_{ji}$ and $n_{ij} = n_{ji}$.

As noted by Ford (1957), if it is possible to partition the set of players into two groups A and B such that there are never any intergroup comparisons, then there is no basis for rating any player in A with respect to any player in B (Hunter, 2004). It is therefore assumed that the tournament is completely connected, i.e., there is a chain of matches which links any given pair of players. In order for the maximum likelihood estimates of the strengths to exist, a second condition is required which will be further denoted as Ford's Assumption: In every possible partition of the players into two nonempty subsets, some player in the second set beats some player in the first set at least once (Ford, 1957). As a special case, Ford's Assumption is not satisfied if group A consists of only one player who has lost or won all games. The maximum likelihood estimate for this player will be infinite-valued.

The likelihood can alternatively be expressed as a function of $\theta = (\theta_1, \ldots, \theta_m)'$ with $\theta_i = \log(\pi_i)$, $\forall\ i : 1, \ldots, m$. Using (1), the probability p_{ij} then becomes:

$$p_{ij} = \frac{\exp\left(\theta_i - \theta_j\right)}{1 + \exp\left(\theta_i - \theta_j\right)}, \tag{3}$$

The Bradley-Terry model can hence be solved using logistic regression (Agresti, 2002). Details as to how this model can practically be fit are provided by

So (1995). The non-existence of maximum likelihood estimates is a well-known and understood problem in logistic regression models and has been denoted by Albert & Anderson (1984) as separation.

The log-likelihood takes the form:

$$l(\theta)=\tfrac{1}{2}\sum_{i=1}^{m}\sum_{j:j\neq i}\left[v_{ij}\theta_i+d_{ij}\theta_j-n_{ij}\log\left(\exp\theta_i+\exp\theta_j\right)\right], \tag{4}$$

Extensions to the Bradley-Terry model have been proposed in the literature but are not considered here. Hunter (2004) provides an interesting review.

Firth’s penalty term

The phenomenon of separation or monotone likelihood is observed in the fitting process of a logistic model if at least one parameter estimate diverges to ±∞. It is believed that separation is unpredictable because it is primarily caused by random variation as it may depend on the outcome of a few matches. Furthermore, it is demonstrated by Heinze (2006) that maximum likelihood estimation by logistic regression may give questionable results in the presence of so-called nearly separated data. This situation occurs when the existence of the maximum likelihood estimates depends on the presence of a few particular observations. A solution proposed by Heinze & Schemper (2002) and Heinze (2006) to separation and near-separation is to penalize the log-likelihood, as described by Firth (1993). The basic idea is to introduce a bias term into the standard likelihood function which itself goes to zero as $n\rightarrow\infty$, but for small n operates to counteract the $O(n^{-1})$ bias present here. The penalty function used is Jeffreys invariant prior (Jeffreys, 1961). One of the advantages of the addition of Firth's penalization term is that no arbitrary data manipulation is involved. It is also justified from the use of Jeffreys prior, in the sense that it is non-informative, thereby implying that maximal weight is given to the data. It should also be noted that the interpretation of the model is not changed in any way. Firth (1993) demonstrated that, for a broad class of generalized linear models, this penalized likelihood is asymptotically consistent and eliminates the usual small-sample bias found in maximum likelihood estimates.

The suggested penalized log-likelihood function takes the following form:

$$l^{*}(\theta)=l(\theta)+\tfrac{1}{2}\log\left|I(\theta)\right|, \tag{5}$$

where $I(\theta)$ is the Fisher Information Matrix of θ.

Case Study

The impact of the addition of Firth's penalty using a motivational (simplified) example will be demonstrated. The evaluation will be done on a large data set of table-tennis players. The data that are used for analysis consist of all recorded match results during the sports season of 2006-2007 of a population of 770 players from the province of Vlaams-Brabant (Belgium). It is shown in Figure 1(a) that the population is highly heterogeneous, both in terms of strengths as number of matches played. It is noted that, in line with existing rating systems, the estimates for θ were linearly transformed to fall roughly between 0 and 3,000 (Glickman, 1995 and 1999) and (Marcus, 2001).

The transformation used was such that a difference of 100 points between 2 players corresponds with odds of 2 for the highest rated player to win. The median (Q1-Q3) number of matches per player equals 61 (35-78). The primary objective is to rate each player in this pool using the penalized and unpenalized maximum likelihood estimates, and to provide Wald-based and profile likelihood 95% confidence intervals. The differences between penalized and unpenalized maximum likelihood estimates will be investigated. Additionally, the differences between both types of confidence intervals will be discussed.

Consistent with local regulation, a simplified log-likelihood was used to allow the new rating of the i^{th} player, $\forall\ i : 1, \ldots, m$ to depend only on the ratings of each of his/her opponents, which are by way of simplification (naively) considered constant during the season. Therefore, $\forall\ i : 1, \ldots, m$:

$$l(\theta_i) = \sum_{j:j\neq i}\left[v_{ij}\theta_i + d_{ij}\theta_j^{\,c} - n_{ij}\log\left(\exp\theta_i + \exp\theta_j^{\,c}\right)\right], \tag{6}$$

where θ_j^c indicates the (scalar-valued) rating of the j^{th} player.

This log-likelihood (6) can, contrary to (4), not be considered a logistic regression model but has to be optimized using Newton's Method or through an appropriate Minorization Maximization (MM) algorithm. Maximum likelihood estimation using (6) will better allow an evaluation of the impact of separation as it will, unlike model (4), not depend on a linear combination of regressors. It can indeed be verified that monotonicity of the log-likelihood (6) is only to occur when a player loses or wins all matches. It is therefore expected that the phenomenon of near-separation is more simply expressed as a function of the victory rate. Application of (4) to the same data set will be presented in before the conclusion of this article. It can easily be shown that the score function of the penalized log-likelihood can be expressed as:

$$S(\theta_i) = \sum\nolimits_{j:j\neq i}\left[\left(v_{ij} - n_{ij}p_{ij}\right)\right] + \tfrac{1}{2}\left[1 - 2\frac{\sum\nolimits_{j:j\neq i} I_j(\theta_i)\, p_{ij}}{I(\theta_i)}\right], \tag{7}$$

where the Fisher Information $I(\theta_i) = \Sigma_{j:j\neq i}\, n_{ij}p_{ij}(1 - p_{ij})$ is alternatively expressed as $\Sigma_{j:j\neq i}\, I_j(\theta_i)$. It should also be noted that p_{ij} is equal to the expression in (3) with θ_j replaced by θ_j^c.

Rearranging some of the terms and denoting the total number of wins for the i^{th} player as V_i results in

$$V_i + \frac{1}{2} = \sum\nolimits_{j:j\neq i}\left[n_{ij} + \frac{I_j(\theta_i)}{I(\theta_i)}\right]p_{ij}, \tag{8}$$

This expression has a simple interpretation in terms of data adjustments: add ½ match to the players total number of wins and add a fraction of a match to the total number of matches played against the j^{th} player. The fractions to be added depend on the unknown θ_i.

Prior to fitting the data, note Ford's Assumption is not satisfied for about 5% of the players, and hence, the maximum likelihood estimates of these players will be infinite-valued. Removing these players from the data by no means guarantees the maximum likelihood estimates of the remaining players to exist, as some of the latter may have only won matches against those that are removed. To solve this problem, two virtual games for every single player are added, i.e., one win and one loss against a (virtual) player of equal strength. These virtual players are added with their given strengths at the right-hand side of (6). The introduction of virtual matches may dilute the difference between penalized and unpenalized maximum likelihood estimates for every single player; however, given the size and the heterogeneity of the data, the overall relationship between both estimates can still reliably be expressed.

As observed from Figure 1(b), the penalized maximum likelihood (PML) estimates are slightly more conservative, i.e., the estimate is pulled towards the center. Players with a low victory rate, i.e. ≤20%, have a PML estimate which is slightly higher than the ML estimate. The reverse phenomenon is observed for players with a high, perhaps ≥80%, victory rate. The small-sample bias reduction is also evident in the subset of players who have played fewer than 30 matches. The shrinkage towards the mean is more pronounced compared to players on whom more information is available.

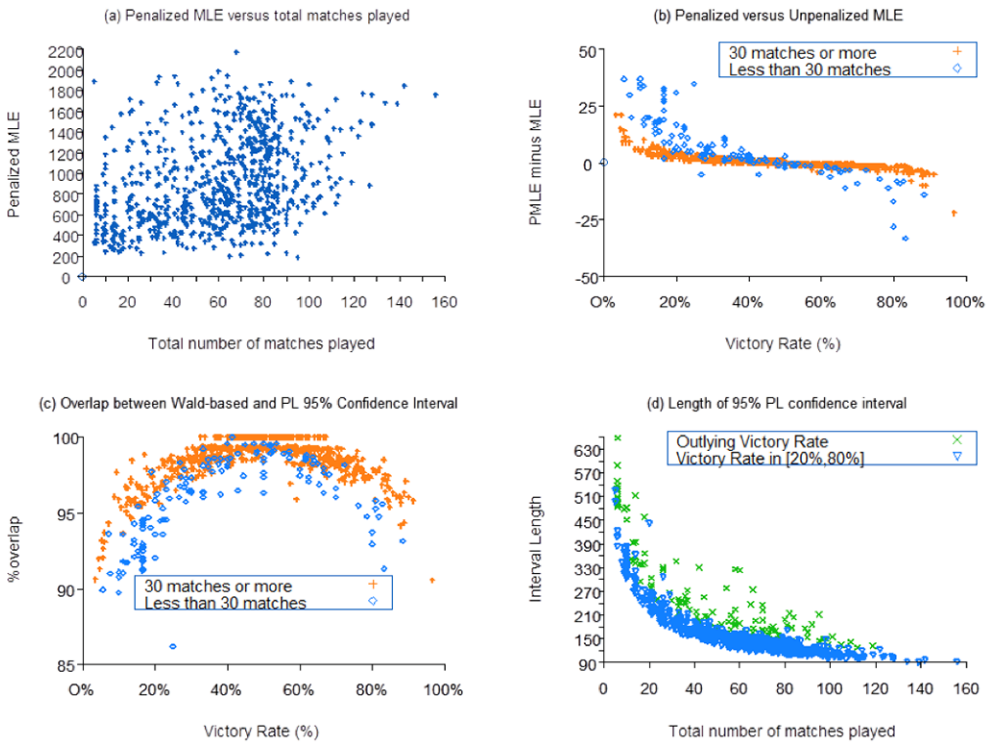

Figure 1. Supporting figures of Case Study

Although the symmetry of the profile likelihood may be enhanced by the addition of a penalization term, it is important to bear in mind that the resulting profile likelihood may still be asymmetric, in particular in the presence of near-separation. Heinze & Schemper (2002) therefore advise against the use of Wald-based confidence intervals and propose the profile penalized likelihood confidence interval as a more suitable solution. The discrepancy between Wald and profile likelihood 95% confidence intervals is graphically presented in Figure 1(c). For this purpose, the percent overlap of both confidence intervals is defined as the length of the intersecting interval, divided by the length of its union. It shows that both confidence intervals match very well when victory rates are close to 50%. However, as the victory rate is an indicator of the likelihood's asymmetry, it is not surprising that the discrepancy is increased with increasing victory or defeat rate. It is also shown that the discrepancy is more pronounced for players on whom less data is available. Compared to Wald-based confidence limits, profile likelihood confidence limits are slightly shifted towards higher values for players with a high victory rate. The reverse phenomenon is observed for players with a low victory rate. Finally, it is seen from Figure 1(d) that the length of the profile likelihood confidence interval is not only dependent on the number of matches played but also on the victory/defeat rate. It may not come as a surprise that the precision of the estimates is lowest for extreme victory rates.

Optimizing the penalized Bradley-Terry log-likelihood

It was shown by Firth (1993) and Heinze & Schemper (2002) that maximum penalized likelihood estimates in logistic regression models are obtained by splitting each original observation *i* into two new observations having response values y_i and $1-y_i$ with iteratively updated weights $1+h_i/2$ and $h_i/2$ respectively (using their notation). It is also argued that the splitting of each original observation into a response and non-response guarantees finite estimates. It is further shown that the h_i's are obtained from the i^{th} diagonal elements of the hat matrix whose elements are refreshed at every iteration. Mathematical details are provided by Firth (1993) and Heinze & Schemper (2002).

This led to the development of software to allow calculation of Firth-type estimates. Direct implementation of the methodology in a SAS macro, S-plus library and R package owes to Heinze & Ploner (2004). An additional R package to fit the Bradley-Terry logistic model was developed by Firth (2005). Implementation in logXact version 8 by Cytel (Cytel, n.d.) has become available in 2005. As of 2008, users of SAS version 9.2 can apply Firth's correction as an option to the LOGISTIC procedure.

Because of the recent advancements in software development for logistic regression, maximum likelihood estimation using a Minorization-Maximization (MM) algorithm seems to be of lesser use from a practical point of view. In addition, an MM algorithm method to obtain the maximum penalized likelihood estimates has so far not been developed. However, it is important to note that some of the extensions to the Bradley-Terry model cannot be fitted using logistic regression (Hunter, 2004) and the MM algorithm may need to be used here as an alternative. In the next sections, the approximate score equations and an MM algorithm for approximate maximum penalized likelihood estimation will be presented.

Approximating the penalized score equation

From (4) it follows that

$$\frac{\partial l(\theta)}{\partial \theta_i} = \sum\nolimits_{j:j \neq i} \left(v_{ij} - n_{ij} p_{ij} \right) \tag{9}$$

because $\frac{\partial p_{ij}}{\partial \theta_i} = p_{ij}\left(1 - p_{ij}\right)$ and $\frac{\partial p_{ij}}{\partial \theta_j} = -p_{ij}\left(1 - p_{ij}\right)$, the information matrix $I(\theta)$ has diagonal elements

$$I(\theta)_{ii} = -\frac{\partial^2 l(\theta)}{\partial^2 \theta_i} = \sum\nolimits_{j:j \neq i} \left[n_{ij} p_{ij} \left(1 - p_{ij}\right) \right] \tag{10}$$

and off-diagonal elements

$$I(\theta)_{ij} = -\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} = -n_{ij} p_{ij} \left(1 - p_{ij}\right) \tag{11}$$

Differentiation of log$|I(\theta)|$in (5) requires derivatives of a log determinant with respect to the vector $\boldsymbol{\theta}$. To avoid that optimization of the penalized score equation would require major matrix operations at every iteration, lengthening the computational process and likely making it less stable, suggesting an approximate rather that an exact approach. The approximation consists of imposing the score function to be of a similar structure as (7) to obtain:

$$S_{\text{approx}}(\theta) = \frac{\partial l(\theta)}{\partial \theta_i} = \sum\nolimits_{j:j \neq i} \left[\left(v_{ij} - n_{ij} p_{ij} \right) \right] + \tfrac{1}{2} \left[1 - 2 \frac{\sum\nolimits_{j:j \neq i} I_j(\theta)_{ii}\, p_{ij}}{I(\theta)_{ii}} \right] \tag{12}$$

The term $I_j(\theta)_{ii}$ in the numerator is the jth contribution to $I(\theta)_{ii}$ and is equal to $n_{ij}p_{ij}(1 - p_{ij})$. Setting this expression (12) to zero and rearranging some of the terms results in:

$$V_i + \frac{1}{2} = \sum\nolimits_{j:j\neq i} \left[n_{ij} + \frac{I_j(\theta)_{ii}}{I(\theta)_{ii}} \right] p_{ij} \tag{13}$$

The same reasoning as Firth (1993) is applied, i.e., that each original observation x_{ij} (i.e., a win or a loss of the i^{th} player against the j^{th} player) can be split into 2 new observations having response values x_{ij} and $1 - x_{ij}$ with iteratively updated weights $1 + g_{ij}/2$ and $g_{ij}/2$ respectively. Note that the weights g_{ij} are an approximation to the diagonal elements of the hat matrix introduced earlier if we were to express (5) as a logistic regression model. The weights are updated at each iteration and depend on the unknown θ. It can then be verified that the approximation to the likelihood function $l^*(\theta)$ can alternatively be expressed as:

$$l^*_{\text{approx}}(\theta) = \tfrac{1}{2}\sum_{i=1}^{m}\sum\nolimits_{j:j\neq i} \left[\begin{array}{r} \left[v_{ij} + n_{ij}\dfrac{g_{ij}}{2} \right]\theta_i + \left[d_{ij} + n_{ij}\dfrac{g_{ij}}{2} \right]\theta_j \\ -\left[n_{ij} + n_{ij} g_{ij} \right]\log\left(\exp\theta_i + \exp\theta_j\right) \end{array} \right] \tag{14}$$

Optimizing (14) for θ_i, it is easily verified from (8) that

$$\sum\nolimits_{j:j\neq i} n_{ij} g_{ij} = \sum\nolimits_{j:j\neq i} \frac{I_j(\theta)_{ii}}{I(\theta)_{ii}} = 1 \tag{15}$$

Expressions (14) and (15) will allow construction of a MM-algorithm.

Minorization-Maximization algorithms

Hunter (2004) demonstrated optimization of the unpenalized log-likelihood function is obtained using a specific case of a general class of algorithms referred to here as Minorization-Maximization (MM) algorithms and shows that convergence is reached provided Ford's Assumption holds.

An MM algorithm operates by creating at each iteration a surrogate function $Q(\theta)$ that minorizes the log-likelihood function $l(\theta)$. This is to say $Q(\theta) \leq l(\theta)$ with equality if and only if $\theta = \theta^{(k)}$. When now the surrogate function is maximized, the log-likelihood is driven uphill. This combination of a minorization and a maximization step is repeated until convergence.

The strict concavity of the logarithm function implies for positive x and y that $-\log(x) \geq 1 - \log(y) - x/y$ with equality if $x = y$. As shown in Hunter (2004), fixing $\theta^{(k)}$ and defining the function

$$Q_k^*(\theta) = \tfrac{1}{2}\sum_{i=1}^{m}\sum\nolimits_{j:j\neq i}\left[\begin{array}{l}\left[v_{ij} + n_{ij}\dfrac{g_{ij}}{2}\right]\theta_i + \left[d_{ij} + n_{ij}\dfrac{g_{ij}}{2}\right]\theta_j \\ \qquad + \left[n_{ij} + n_{ij}g_{ij}\right]\left[\begin{array}{l} 1 - \log\left(\exp\left(\theta_i^{(k)}\right) + \exp\left(\theta_j^{(k)}\right)\right) \\ -\dfrac{\exp\left(\theta_i\right) + \exp\left(\theta_j\right)}{\exp\left(\theta_i^{(k)}\right) + \exp\left(\theta_j^{(k)}\right)}\end{array}\right]\end{array}\right]$$

it can be seen that $Q_k^*(\theta)$ minorizes $l^*_{\text{approx}}(\theta)$ as

$$Q_k^*(\theta) \leq l^*_{\text{approx}}(\theta) \tag{16}$$

with equality if $\theta = \theta^{(k)}$.

Using (15), optimization of $Q_k^*(\theta)$ for θ_i is now straightforward with solution

$$\exp\left(\theta_i^{k+1}\right) = \left(V_i + \tfrac{1}{2}\right)\left[\sum\nolimits_{j:j\neq i}\frac{n_{ij} + \dfrac{I_j(\theta)_{ii}}{I(\theta)_{ii}}}{\exp\left(\theta_i^{(k)}\right) + \exp\left(\theta_j^{(k)}\right)}\right]^{-1} \tag{17}$$

Similarly, minorization and maximization of the unpenalized log-likelihood function $l(\theta)$ is achieved with

$$\exp\left(\theta_i^{k+1}\right) = \left(V_i\right)\left[\sum\nolimits_{j:j\neq i}\frac{n_{ij}}{\exp\left(\theta_i^{(k)}\right) + \exp\left(\theta_j^{(k)}\right)}\right]^{-1} \tag{18}$$

Application

The same data will be used. Approximate maximum penalized likelihood estimates will be produced using (17). In addition equation (17) will be generalized such that the ½ match to the player's total number of wins can be modified at both sides of the equation:

$$V_i + a = \sum\nolimits_{j:j\neq i}\left[n_{ij} + 2a\frac{I_j\left(\theta\right)_{ii}}{I\left(\theta\right)_{ii}}\right]p_{ij} \tag{19}$$

Comparisons with the exact penalized maximum likelihood estimates obtained using logistic regression are compared with the approximate penalized likelihood estimates for $a = 0.3, 0.5, 0.7$ and 1. A comparison between exact penalized likelihood estimates and unpenalized likelihood estimates is presented in Figure 2(a).

Unlike the results shown in Figure 1(b), the differences between both estimates are not only a function of the percentage of wins and of the sample size. This is because separation can occur as a result of a non-trivial linear combination of regressors, which can potentially occur at any sample size or victory rate. Also note the far larger presence of players with low rather than high victory rates in the data. It is further shown in Figure 2(b) that unpenalized estimates obtained using either logistic regression or by the MM algorithm (18) effectively give the same results. An investigation of the effect of the value a for the added match in (19) is presented in Figures 2(c) to 2(f). It is shown in Figure 2(c) that the approximate penalized ML estimates (for $a = 1$) strongly differ from the exact penalized ML estimates.

It is also clear from Figures 2(c) and 2(d) that approximations implied by values of a larger than 0.5 result in a too strong correction of the unpenalized ML estimates, when compared to the exact penalized ML estimates. The reverse phenomenon is observed for $a = 0.3$ (see Figure 2(f)) and for any value of a lower than 0.3 (results not shown). For these small values of a, the comparison with the exact penalized ML estimates will become more and more similar to the pattern observed in Figure 2(a), for $a \to 0$. It is clear from Figure 2(e) that choosing $a = 0.5$ resulted in the best fit. Similar results were obtained through simulations (data not shown). A value of $a = 0.5$ always yielded results that are sufficiently close to the exact values. It was observed that the correction implied by the exact results, both on the real data as on the simulation, was always slightly larger compared to the approximate results. However, differences between the exact and the approximate estimates were always negligible.

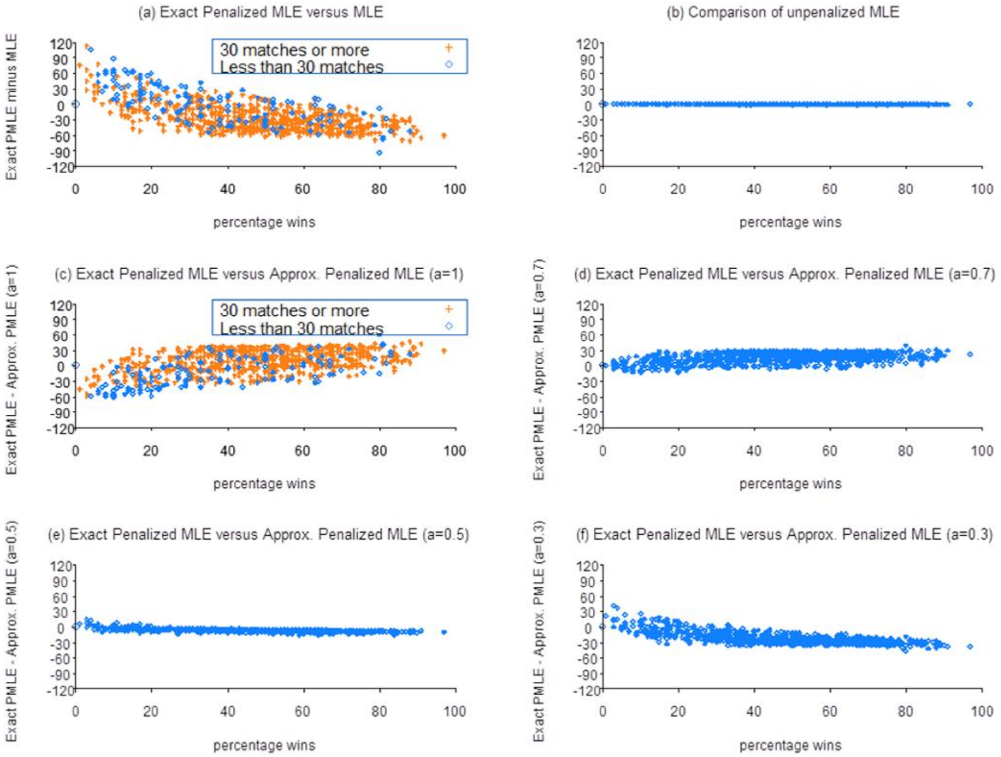

Figure 2. Supporting figures of Case Study

Conclusion

The objective of this work was to evaluate the effect of the addition of Firth's penalty term to the Bradley-Terry log-likelihood. One of the fundamental differences between the current work and earlier applications of strength estimation in the literature, such as in Agresti (2002) and Firth (2005), is due to the size and degree of imbalance of the data. Application of the implied models to a sufficiently large and heterogeneous pool of players allows better characterization of the impact of the penalty term. The differences between penalized and unpenalized ML estimates were generally more pronounced when the number of matches were relatively low or when victory or defeat rates were in the high range. Findings due to Heinze & Schemper (2002) such as the recommended use of profile likelihood confidence intervals over Wald-based confidence limits, in presence of asymmetric likelihood functions, also carry over to the Bradley-Terry model.

A secondary objective consisted of the development of a MM algorithm for optimization of the penalized log-likelihood. Direct application of the MM algorithm to this type of data may seem inefficient due to the availability of logistic regression software that can easily produce Firth-type maximum likelihood estimates. However, some of the extensions of the Bradley-Terry model cannot be expressed as a logistic regression model and MM algorithms can be used as an alternative as they tend to give fast, simple-to-code iterations, where each iteration moves in the right direction. When applied to the full size of the data, the MM algorithm converged within an acceptable time frame and behaved stably for any set of starting values. Although exact results were not obtained with the proposed MM algorithm, the approximate values were shown to be sufficiently close to the exact values when applied to the data at hand. The applicability of these results may need to be confirmed on other data sets. A favorable feature of the proposed MM algorithm is that it is constructed in such a way that major matrix operations at every single iteration are avoided. As convergence of the algorithm is only obtained after several hundreds of iterations, the gain in processing time is expected to be considerable. In a next step, approximate MM algorithms will need to be constructed on some of the well-known extensions of the Bradley-Terry model. This will be a subject for further research.

References

Agresti A. (2002). *Categorical data analysis* (2nd edition). New York: John Wiley & Sons.

Albert, A. & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71*(1), 1-10. doi: 10.1093/biomet/71.1.1

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika, 39*(3-4), 324-345. doi: 10.1093/biomet/39.3-4.324

Cytel. (n.d.). *LogXact® 11: Exact Inference for Logistic Regression*. Retrieved October 5, 2014, from http://www.cytel.com/software/logxact

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*(1), 27-38. doi: 10.1093/biomet/80.1.27

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software, 12*(1), 1-12. doi: 10.18637/jss.v012.i01

Ford, L. R. (1957). Solution of a ranking problem from binary comparisons. *American Mathematical Monthly, 64*(8.2), 28-33. doi: 10.2307/2308513

Glickman M. E. (1995). Chess rating systems. *American Chess Journal, 3*, 59-102.

Glickman M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society C: Applied Statistics, 48*(3), 377-394. doi: 10.1111/1467-9876.00159

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine, 25*(24), 4216-4226. doi: 10.1002/sim.2687

Heinze, G. & Ploner, M. (2004). *A SAS macro, S-plus library and R package to perform logistic regression without convergence problems* (Technical Report 2/2004). Vienna, Austria: Department of Medical Computer Sciences, Medical University of Vienna.

Heinze, G. & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine, 21*(16), 2409-2419. doi: 10.1002/sim.1047

Hunter, D. R. (2004). MM Algorithms for generalized Bradley-Terry models. *The Annals of Statistics, 32*(1), 386-408. doi: 10.1214/aos/1079120141

Jeffreys, A. (1961). *The theory of probability*. Cambridge, UK: Cambridge University Press.

Marcus, D. J. (2001). New table-tennis rating system. *Journal of the Royal Statistical Society, Series D (The Statistician), 50*(2), 191-208. doi: 10.1111/1467-9884.00271

SAS Institute, Inc. (2009). *SAS/STAT users guide, version 9.2.*: Cary, NC: SAS Institute Inc.

So, Y. (1995). A tutorial on logistic regression. [PDF]. Retrieved from http://www.ats.ucla.edu/stat/SAS/library/logistic.pdf

Zermelo, E. (1929). Die berechnung der Turnier-Ergebnisse als ein maximumproblem der Wahrscheinlichkeitsrechnung. *Mathemathische Zeitschrift, 29*(1), 436-460. doi: 10.1007/BF01180541