Appl. Statist. (2006)
55, *Part* 2, *pp.* 225–239

Predicting the outcomes of annual sporting contests

Rose Baker and Philip Scarf

University of Salford, UK

[Received July 2004. Revised August 2005]

Summary. Data from 20 sporting contests in which the same two teams compete regularly are studied. Strong and weak symmetry requirements for possible models are identified, and some simple models are proposed and fitted to the data. The need to compute the exact likelihood function and the presence of missing values make this non-trivial. Forecasting match outcomes by using the models can give a modest improvement over a naïve forecast. Significance tests for studying the effect of 'match covariates' such as playing at home or away or winning the toss are introduced, and the effect of these covariates is in general found to be quite large.

Keywords: Akaike information criterion; Cricket; Exact tests; Football; Generalized linear models; Rugby; Sporting contests

1. Introduction

There are very many annual sporting contests in which the same two teams strive to overcome each other year after year. They are sometimes called 'grudge matches'. Such contests occur in a wide range of sports and can endure for many years, the earliest on record dating from the mid-19th century. Other similar contests that are also studied here occur regularly but not strictly annually.

In general, the two protagonists may be rival countries (e.g. for the ashes), universities (e.g. Varsity matches), schools (e.g. the annual Eton–Harrow cricket match), villages (e.g. the annual tug-of-war between East and West Ayton, in Ryedale, Yorkshire), services (e.g. the annual US army–navy football match), and so on.

One of the most famous is the Oxford–Cambridge Boatrace, and an attempt to predict the outcome motivated the present work. It was immediately clear from carrying out a runs test that there were runs of wins for each side, and this fact offered the possibility of modelling the outcome, and so of making a better than chance prediction of the outcome of a forthcoming race. A cursory analysis also showed that a heavier crew was more likely to win (the p-value of a t-test for excess percentage weight between winning and losing teams was 0.001). In general, there are *match covariates*, which are known before the match proper begins, which may affect the outcome.

This paper investigates how well the outcome of sporting contests may be predicted. It addresses such questions as can useful models be identified and can some types of model be ruled out? Do different types of sport require different models? How useful are the models at predicting the outcomes of matches? How important are match covariates, such as the weight of the crew in a boatrace, winning the toss or playing at home, in determining the outcome? How may their effect be tested for, controlling for the effect of other variables?

Address for correspondence: Rose Baker, Centre for Operational Research and Applied Statistics, University of Salford, The Crescent, Salford, M5 4WT, UK.
E-mail: r.d.baker@salford.ac.uk

 0035–9254/06/55225

There are several characteristic features of sporting data that have guided the modelling process. Outcomes are modelled in no more detail than win, lose or draw, so that sport-independent models can be developed. One cannot simplify beyond this point, because draws are common in sports where the score is discrete or playing time is limited, as in football and cricket. It is clear that modelling requires teams to be labelled, so that wins and losses are seen from the viewpoint of team 1 or team 2. For example the weight difference between rowing teams was modelled as the percentage difference

$$200(w_1 - w_2)/(w_1 + w_2)$$

of weights between team 1 and team 2 in an obvious notation.

Past results and present and past match covariates are predictors that must be included. Further, in analysing historical sporting data, there are sometimes gaps of several years due to world wars etc., and also years where results and/or covariates such as who won the toss are missing. Any useful model and inferential procedure must be able to cope with missing data.

Statistically, a sequence of match outcomes is a discrete-valued time series and is a binary series unless draws can occur. In time series analysis Cox (1981) distinguished between observation-driven and parameter-driven models. Observation-driven models use past observations to predict future outcomes. In parameter-driven models, however, there is an underlying or 'hidden' process such as relative team strength that determines outcomes. Hidden Markov chain models are major examples of these; for example see MacDonald and Zucchini (1997). Because of the difficulty of inference for parameter-driven models, and because the actual result that is achieved should have an effect on team and crowd psychology and thus on future outcomes, we explore only observation-driven models here.

The next section introduces some useful models, after which the results of fitting them to data are presented in Section 3.

2. Methodology

McCullagh and Nelder (1989) described models for ordinal data. It is useful to motivate them here by considering a linear function of past observations that models relative team strength y as

$$y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$$

where $\boldsymbol{\beta}$ is a vector of coefficients and the covariates $\mathbf{x}$ include both past results and match covariates that affect the performance of team 1 relative to team 2. The actual relative performance will be $Y + V$, where V is a random variable with distribution function F. Team 1 wins if $Y + V > \theta_{\mathrm{u}}$, loses if $Y + V < \theta_{\mathrm{l}}$, where $\theta_{\mathrm{u}} > \theta_{\mathrm{l}}$, and otherwise draws.

The probability of a team 1 win is therefore

$$\mathrm{Prob}(V > \theta_{\mathrm{u}} - y) = 1 - F(\theta_{\mathrm{u}} - y),$$

giving the following probabilities for a team 1 win, a team 2 win and a draw respectively:

$$\left.\begin{array}{c} 1 - F(\theta_{\mathrm{u}} - y), \\ F(\theta_{\mathrm{l}} - y), \\ F(\theta_{\mathrm{u}} - y) - F(\theta_{\mathrm{l}} - y). \end{array}\right\} \tag{1}$$

Since a constant term can be included in y, we set $\theta_{\mathrm{l}} = 0$ and drop the subscript on θ_{u}. One normally takes F as the distribution function of a normal, logistic or complementary log–

log-distribution. McCullagh and Nelder (1989) generalized the ordinal model in two ways: by a covariate-dependent scaling and by making the two regression lines non-parallel. The full model would give the probability of winning as

$$1 - F\{(\theta - \boldsymbol{\beta}_{\mathrm{u}}^{\mathrm{T}}\mathbf{x})\exp(-\boldsymbol{\tau}^{\mathrm{T}}\mathbf{x})\},$$

and of losing as

$$F\{-\boldsymbol{\beta}_{\mathrm{l}}^{\mathrm{T}}\mathbf{x}\exp(-\boldsymbol{\tau}^{T}\mathbf{x})\},$$

where $\boldsymbol{\tau}$ is another vector of parameters. A problem with this parameterization is the possibility of negative draw probabilities for some covariate values.

There is an important general principle here: if both teams are similarly organized, and we do not have asymmetric information about the teams, the functional form of the model should not change if the team labels 1 and 2 are swapped. This invariance means that, on fitting the model to data, the predictions do not depend on which team is labelled as team 1. We propose this as the *weak symmetry principle*.

A *strong symmetry principle* follows from assuming further that the two teams behave identically (the stochastic component of the model ensures that not all games result in draws!). This leads to some restrictions on the values of model parameters.

As an example of the restrictions on models arising from these principles, coding a victory in the previous match for team 1 as $X_{t-1} = 1$ and a defeat as $X_{t-1} = 0$ would mean that a covariate such as $X_{t-1}X_{t-2}$ would be 1 if team 1 won both times and 0 otherwise. This would violate the weak symmetry principle if it were the only covariate. To make the parameterization invariant under label swaps, in accordance with weak symmetry, we could replace $\beta_1 X_{t-1}X_{t-2}$ by

$$\beta_1 X_{t-1}X_{t-2} + \beta_2(1 - X_{t-1})(1 - X_{t-2}).$$

Now under a label swap β_1 and β_2 interchange, leaving the model functionally invariant. Strong symmetry would require further that $\beta_1 = \beta_2$.

Further, since on swapping labels we have that $V \to -V$, even under weak symmetry F must be the distribution function of a symmetric distribution, so that $F(-x) = 1 - F(x)$. This rules out the use of the complementary log–log-link function, for example.

It is important that the number of model parameters to be fitted to data be kept small, especially as binary data contain little information. The strong symmetry principle produces parsimonious models, and a sensible modelling strategy is to start with models that are consistent with this principle, adding further terms as needed consistent with only the weak principle. Here, the 'strong' result that each team has long-term probability $\frac{1}{2}$ of winning may be too restrictive, and the addition of a constant term in the modelling of y as the coefficient of $x_1 = 1$ rectifies this. In the Oxford–Cambridge Boatrace, both teams do perform very similarly, and this coefficient is accordingly very small.

Turning to the details of modelling, it can be argued that sometimes a draw does not as in equation (1) result from near equality of team strength. In cricket, a draw results rather from a lack of time for the game to be completed (although this lack of time can sometimes be engineered by a weak team). Regarding a draw *versus* a result as nominal outcomes, we obtain the mixed model with probabilities

$$\left.\begin{array}{c} \{1 - F(-y)\}\{1 - W(z)\}, \\ F(-y)\{1 - W(z)\}, \\ W(z), \end{array}\right\} \tag{2}$$

for a team 1 win, a team 2 win and a draw respectively, where W is a distribution function, $z = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}$ and $\boldsymbol{\gamma}$ is a further vector of coefficients. The models that are used in this paper model y as a linear sum of match covariates and past wins for team 1, where $X_{t-1} = 1$ if team 1 won the previous match, $X_{t-1} = -1$ if team 2 won, $X_{t-1} = 0$ for a draw, and so on. The variable z is modelled as a linear sum of draw variables, such that $D_{t-2} = 1$ if the match before last ended in a draw, and otherwise $D_{t-2} = 0$, and so on. The weak symmetry principle does not preclude the use of X_{t-1} etc. in addition to D in modelling W, but if the two teams were identical in performance the strong principle would require that the coefficient be 0. Hence, in the interests of parsimony, terms containing X_{t-1} etc. are not included *in modelling* W. This shows the usefulness of the strong principle in modelling.

2.1. The likelihood function

With no missing data, the likelihood function can be written down very simply. For example, the boatrace data that are given in Table 3 in Section 3.3 give rise to an observed likelihood function of approximately

$$\mathcal{L} = \prod_{i=3}^{n} G(0.15 + 0.4X_{i-1} + 0.58X_{i-2} + 0.21\xi_i), \tag{3}$$

where n is the number of observations, X is scored as 1 for a Cambridge win and -1 for an Oxford win, and ξ_i is the relative percentage weight difference. Here $G = 1 - F$ for a Cambridge win, and F for an Oxford win, where F is the logistic distribution function.

The general formulation must cover the situation where match results and/or match covariates may be missing. This generality, with the need to calculate the exact likelihood function for a stationary process, rather than conditioning on the first few data values as in equation (3), makes the computations non-trivial.

We specialize to the case where all variables except some current match covariates are discrete. This will be so if only the results of the previous matches are used as predictors. The only continuous variable found in the entire data set was relative crew weight in boatraces. It is then computationally convenient to work in terms of the state of the two-team 'system' as defined only by the sequence of previous match results appearing in the model. If the model includes only $X_{t-1} \ldots X_{t-s}$, and draws can occur, there are $p = 3^s$ possible states. Let the probability of a match at time t causing a transition to state i from state j be $M(\mathbf{c})_{ijt}$, where $\mathbf{c}$ denotes the observed match covariates, such as who won the toss. Clearly, if all covariates are known, the transition probability $P_{ijt} = M(\mathbf{c})_{ijt}$, but otherwise we have that

$$P_{ijt} = \sum M(\mathbf{c})_{ijt} \Pr(\mathbf{c}),$$

where the sum runs over all possible vectors of covariates, which have probability $\Pr(\mathbf{c})$. These probabilities are known for such covariates as coin tossing, which would in general have to be fair.

The likelihood function for this Markov chain is

$$\mathcal{L} = \prod_{t=1}^{n} \sum_{i \subset S_t} \sum_{j=1}^{p} P_{ijt} q_{j,t-1}, \tag{4}$$

where $q_{j,t-1}$ is the probability that the system is in the jth state at time $t-1$, and S_t is the set of states at time t that comprise the observed match outcome, for example, that team 1 either drew or won. In practice, for the data sets that are studied, this set is either a single observed outcome or the set of all possible outcomes, i.e. the result was unknown.

Explicitly, the q_{it} are propagated as

$$r_{it} = \sum_{j=1}^{p} P_{ijt} q_{j,t-1},$$

$$q_{it} = r_{it} \Big/ \sum_{j \subset S_t} r_{jt}$$

for $i \subset S_t$; otherwise $r_{it} = 0$.

The exact likelihood function that is used from equation (4) differs from that in equation (3) in that the system is assumed to be in steady state initially, so that the q_{it} obey $\mathbf{q} = \mathbf{Pq}$. When match results are known, $\mathbf{q}$ collapses but, if results are not known for a period, $\mathbf{q}$ gradually reverts to its steady state value.

Use of the full likelihood allows model fitting even when there were gaps in the data, as when matches were not played during the First and Second World Wars, as well as when, as sometimes occurred, matches were played, but results are unknown, or where a match was not played for a single year.

After wars there must have been a great change in the composition of teams and it might seem naïve to treat such gaps in the same way as other missing data. We could instead restart the likelihood function calculation using steady state values of the vector of state probabilities $\mathbf{q}$. However, these gaps of at least 4 years are sufficiently long for $\mathbf{q}$ to revert virtually to its steady state value, so in practice either of these choices would make very little difference to parameter estimation.

Where non-annual data are considered, the number of matches that were missing because of wars etc. is problematical, but fitted parameter values are again very insensitive to the number that is assumed.

Another small problem concerns stationarity. Stationarity allows calculation of exact likelihood functions, and the need for competitive balance (e.g. Szymanski (2001)) suggests that there will not be a long-term drift in match results in increasing favour of one team, and hence that stationary models should work well. To ensure stationarity, however, the distribution of match covariates must also be stationary. Clearly random covariates such as who wins the toss must be stationary, but covariates such as the weight of the crew in a boatrace may not be. It is assumed that such covariates, when transformed to be relative, will be stationary. This is again plausible, given the requirement of competitive balance.

2.2. Computation

Calculating the exact likelihood function requires computation of the initial state probabilities q_{i1} for $1 \leqslant i \leqslant p$, and their subsequent propagation to time n. When match covariates $\mathbf{c}$ are missing, the probabilities $\Pr(\mathbf{c})$ must be found.

States were assigned a number from 1 to p by counting upwards in binary (no draws allowed) or ternary (draws allowed). Sets of missing discrete covariates at each time point were also enumerated and vectorized by counting. The state probabilities q were propagated from times 1 through to time n, conditioning on the observed state at each time point.

The initial state q_{i1} was found by using the device of prepending a 'missing' match with missing match covariates before time 1. Having calculated $\mathbf{P}$ for this match, $\mathbf{q}$ is the solution of $\mathbf{q} = \mathbf{Pq}$, the steady state. Since $\Sigma_{i=1}^{p} q_{it} = 1$, this equation would normally be solved by removing q_p, and solving the resulting linear equations

$$\sum_{j=1}^{p-1} (P_{ij} - P_{ip} - \delta_{ij}) q_j = -P_{ip},$$

where $1 \leqslant i < p$ and δ_{ij} is the Kronecker delta function. Numerically, this procedure was sometimes found to be unstable, resulting in negative probabilities, and it was better to initialize $\mathbf{q}$ as a constant vector $\mathbf{q}_0$ and to approximate $\mathbf{q} \propto \mathbf{P}^n \mathbf{q}_0$, where $n = 8$ was found to be ample.

Each likelihood maximization was repeated several (four) times from a randomized starting-point, to ensure that the true global function minimum had been reached. Because of numerical difficulties arising from fitting the overparameterized models that were explored *en route* to the preferred model, the slower simplex method was found to be preferable to the conjugate gradient method. To reduce the fitting of spurious models, model parameters such as the coefficients of lagged results X_{t-j} and lagged draw variables D_{t-j} were constrained to be positive.

When models are fitted with an excessive number of parameters, these are very weakly constrained by the data, so very large positive or negative values arise, as the function minimizer attempts to find the maximum likelihood point. This causes exponents to exceed the maximum allowed value, for example, as well as causing divide overflows. Attempting to make code robust against these problems is a time-consuming task. We can at least detect implausible parameter values and abandon computation of the negative log-likelihood function, replacing it instead by a very large 'penalty' figure. This, however, in turn causes problems for conjugate gradient or Newton–Raphson-type function minimizers, which require smooth functions. Fortunately, the slow but sure simplex method coped well in such cases.

Computations were done using a purpose-written Fortran 95 program, which is available with the data sets that were used from

`http://www.blackwellpublising.com/rss`

The Numerical Algorithms Group library function minimizers `E04UCF` and `E04CCF` were used for likelihood maximization.

2.3. Goodness of fit

In general, an adequate model can hopefully be found by embedding the chosen model in a larger family of models, such as generalizing an exponential distribution to a Weibull or gamma distribution. Here, the strategy of choosing a minimum Akaike information criterion estimate (MAICE) model from a family of models automatically includes this approach to assuring goodness of fit. The goodness of fit of the chosen model must still, however, be evaluated.

With grouped data, it is straightforward to assess goodness of fit. Here, however, construction of a deviance that has a χ^2-distribution is not possible. To assess the adequacy of the MAICE model, it is necessary to group the data somehow. For binary data, Hosmer *et al.* (1997) suggested grouping outcomes into deciles by predicted probability.

Another approach that is applicable here is to group according to the covariate patterns. Hence goodness of fit was assessed by examining observed and predicted numbers of match outcomes following various types of covariate event, such as a draw following a 'win–lose' sequence. With draws possible, there are nine possibilities for two consecutive outcomes, which with a series of length over 100 give reasonable numbers in each category. For longer series, the 27 possible triple events preceding an outcome were examined.

Predicted outcome probabilities may depend not only on the previous two or three outcomes but also on outcomes that were further back in time, or on match covariates. Hence from each instance of the first part of a sequence, such as win–lose, expected numbers were calculated by summing the calculated outcome probabilities for the final match. These statistics were summarized as a χ^2-statistic. Because model parameters were fitted, and fitted by maximizing a likelihood rather than by minimizing a χ^2-statistic, the exact number of degrees of freedom that

should be used is lower than the value that is quoted. It was also necessary to group cells with low predicted numbers in the usual way.

Finally, the explanatory power of a linear model may be measured by the coefficient of determination, R^2. For instance, Cox and Snell (1989) and others proposed a pseudo-R^2 for general models given by

$$-\log(1-R^2)=\frac{2}{n}\{l(\hat{\boldsymbol{\beta}})-l(\mathbf{0})\},$$

where l denotes log-likelihood, $\boldsymbol{\beta}$ denotes a vector of parameters and $\boldsymbol{\beta}=\mathbf{0}$ denotes the 'null' model. Nagelkerke (1991) proposed a correction, since the maximum value of R^2 that is attainable may be less than 1. The correction required normalizing R^2 to its maximum value of $1-\exp\{2n^{-1}\,l(\mathbf{0})\}$, and we use this corrected value of R^2. The null model is here taken as the naïve model, in which the probability of an outcome is estimated as the corresponding proportion. The naïve or null model is a special case of the general model, where match covariate and other covariate coefficients are all 0.

2.4. Statistical tests

We may wish to test whether the effect of a match covariate with parameter ϕ such as who wins the toss or whether the game is played at home or away is significant. We might also wish to test whether other model parameters are non-zero. The asymptotic properties of the log-likelihood l under hypothesis H_0, that $2\{l(\hat{\phi})-l(0)\}$ is distributed as a random variable from a χ^2-distribution with 1 degree of freedom, are well known. The score test based on derivatives of l has test statistic Z, where

$$Z=\frac{\partial l/\partial\phi|_{\phi=0}}{(-E\,\partial^2 l/\partial\phi^2|_{\phi=0})^{1/2}}$$

is asymptotically a standard normal random variate under $H_0:\phi=0$. However, we can generate exact tests that do not rely on asymptotic properties, by permuting match covariates between matches to generate the reference distribution of either $l(\hat{\phi})$ or of $\partial l/\partial\phi|_{\phi=0}$ under hypothesis H_0. Of these, the score statistic is computationally preferable in requiring only one function minimization to estimate values for all other model parameters, instead of maximizing the whole likelihood function afresh for each permutation. The latter test is, however, more powerful.

In this study, the full test based on $l(\hat{\phi})$ was used, and the permutation distribution of the match covariate coefficient was found. Fig. 1 shows an example. For each permutation of match covariates, the likelihood function was remaximized and the covariate coefficient estimate found.

3. Results

3.1. Data sets and exploratory analysis

A sample of data from famous annual and otherwise regular sporting contests was assembled for analysis. Years when the match was not played, largely owing to war, were coded as missing data, and match covariates were also not always recorded. The methodology that was developed can cope with this. In general, data sets with at least about 100 matches played were sought.

A preliminary analysis was done by applying runs tests, both for the total number of runs of like outcomes and also with runs broken down into runs of wins and losses, and runs of draws. Runs were continued across missing years. Table 1 shows significant evidence of departure from randomness in seven cases, with an additional 'possible'. Sometimes, as in the case of

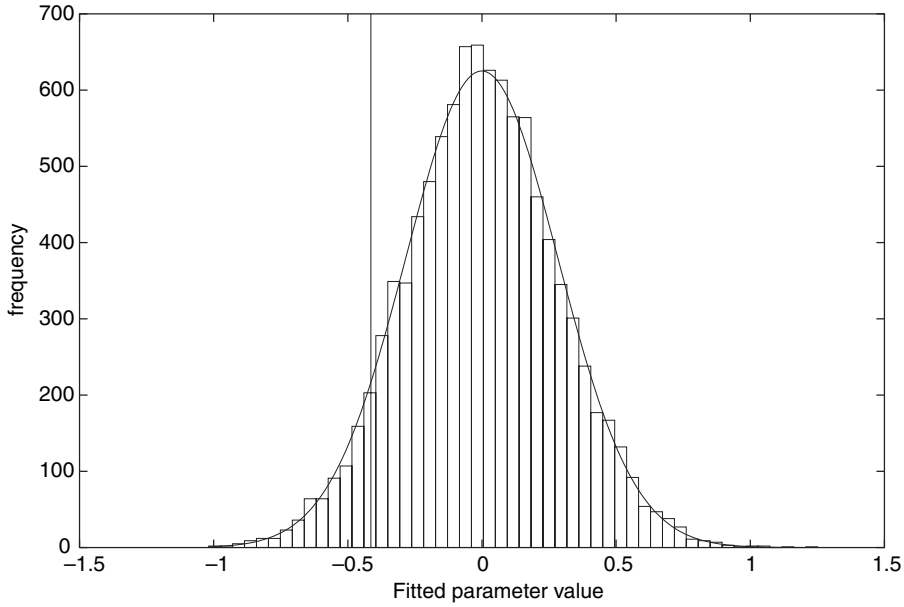

Fig. 1. Distribution of the England–Australia home–away effect coefficient for the ashes data, showing the observed coefficient value of −0.414: the data were obtained by permuting the home–away variable.

Table 1. *p*-values of one-sided runs tests of H_0 that the number of runs is no fewer than expected from chance†

Sport	*Teams*	*Significance*	*Significance‡ of wins or losses*	*Significance‡§ of draws*
Boatrace	Oxford *versus* Cambridge	0.0002	0.0002	NA
Chess	Oxford *versus* Cambridge	0.0046	0.0021	NS
Cricket	Oxford *versus* Cambridge	0.0212	0.095	0.0198
	North *versus* South Argentina	NS	NS	NS
(Ashes)	England *versus* Australia	0.0003	0.0032	0.0022
Rugby	Oxford *versus* Cambridge	NS	NS	NS
(Calcutta Cup)	England *versus* Scotland	NS	NS	NS
Football (USA)	US Army *versus* US Navy	NS	NS	NS
	Harvard *versus* Yale	NS	NS	NS
	Alabama *versus* Auburn	NS	NS	NS
	Ohio State *versus* Michigan	NS	NS	NS
	Texas *versus* Texas A&M	NS	NS	NS
	Georgia *versus* Florida	0.0201	0.01978	NS
	Oregon *versus* Oregon State	0.074	0.0932	NS
	Michigan *versus* Michigan State	0.0161	0.024	NS
Football	England *versus* Scotland	NS	NS	NS
	Kilmarnock *versus* Partick Thistle	NS	NS	NS
	Arsenal *versus* Spurs	NS	NS	NS
	Penarol *versus* Nacional	0.0355	0.11	0.0347
	Barcelona *versus* Real Madrid	NS	NS	NS

†The last three columns show *p*-values for runs of any type, then *p*-values for runs of wins and losses and finally *p*-values for runs of draws only. Runs of all types are seen.
‡NS, not significant.
§NA, not applicable.

Table 2. MAICE models fitted to the data sets and fit statistics (pseudo-R^2, χ^2 and degrees of freedom)†

Sport	*Teams*	*Match covariate*	*Model*‡	R^2	χ^2	f
Boatrace	Oxford *versus* Cambridge	Weight difference	1(2)	0.255	0.082	2
Chess	Oxford *versus* Cambridge	—	2(1,0)	0.083	0.46	6
Cricket	Oxford *versus* Cambridge	Toss	2(0,3)	0.126	4.61	6
	North *versus* South Argentina	—	2(0,2)	0.047	5.89	6
(Ashes)	England *versus* Australia	Home or away	2 ES(1)	0.069	1.56	6
Rugby	Oxford *versus* Cambridge	—	1(0)	0.000	2.88	6
(Calcutta Cup)	England *versus* Scotland	Home or away	2(0,0)	0.027	3.79	3
Football (USA)	US Army *versus* US Navy	—	1(0)	0.000	1.30	2
	Harvard *versus* Yale	—	1(0)	0.000	1.51	2
	Alabama *versus* Auburn	—	1(0)	0.000	1.18	2
	Ohio State *versus* Michigan	—	1(0)	0.000	0.31	2
	Texas *versus* Texas A&M	—	1(2)	0.072	0.26	2
	Georgia *versus* Florida	—	1 ES	0.096	0.59	2
	Oregon *versus* Oregon State	—	1 ES	0.051	0.13	2
	Michigan *versus* Michigan State	—	1(2)	0.073	1.14	2
Football	England *versus* Scotland	Home or away	1(0)	0.005	4.61	6
	Kilmarnock *versus* Partick Thistle	—	2(0,2)	0.011	11.7	6
	Arsenal *versus* Spurs	—	1(0)	0.000	7.84	6
	Penarol *versus* Nacional	—	2(0,3)	0.019	3.10	6
	Barcelona *versus* Real Madrid	Home or away	1(0)	0.151	3.29	6

†The model description notation that is used is explained in Section 3.1.
‡ES, exponential smoothing.

the American football match between Georgia and Georgia State, only runs of wins or losses occur, with no runs of draws, and sometimes, as in the Penarol–Nacional football match, only runs of draws are seen.

This finding shows that the data are not entirely random.

Best models were fitted 'automatically' to the 20 data sets, by finding the MAICE model among the following:

(a) 'type 1' or autoregressive models as given in equation 1, with covariates taken as lagged results X_{t-j}, coded as 1, 0 or −1—these models are identified in Table 2 as '1(l)' models, where l is the number of lagged terms;

(b) as above, but with the 'exponential smoothing' term

$$\mathrm{ES} = \beta_s \sum_{i=1}^{\infty} r^j X_j \tag{5}$$

replacing a linear sum of lagged scores—this covariate, which requires two parameters, β_s and r, is intended to give a smoothed measure of recent team 1 performance as contrasted with that of team 2; these models are identified as '1 ES' in Table 2;

(c) type 2 models as given in equation (2), with lagged covariates as before and with $W(z)$ also being a function of 'draw' covariates, which are 1 if the previous or earlier match ended in a draw and 0 otherwise—these models are coded as $2(l, m)$, where l is the number of lagged X terms and m the number of lagged draw terms;

(d) type 2 models but with the X-terms replaced as before by an exponential smoothing term—these models are identified as 2 ES(m), where m is the number of lagged draw terms.

In the interests of parsimony, these models obey the strong symmetry principle, except for the constant term in the modelling of y, which allows unequal probabilities of winning.

Table 2 shows the best model that was fitted to each data set.

3.2. Model choice

MAICE models are shown in Table 2. Even using such a criterion, the problems of overfitting are too well known to bear repetition here, especially when a large number of models are fitted. Small sample corrections to the AIC partially ameliorate but do not solve this problem. There is an AIC correction for binary logistic models (Yanagihara *et al.*, 2003) but this cannot be used for the more general models of the current problem.

To study the probability that the correct model had been fitted by using the MAICE criterion, the sequence of match results was randomized, and the model choice procedure was repeated. For the randomized data sets, the correct model was the two-parameter model, coded as 1(0) in Table 2. This model fitted only about 60% of the permutations for a sample size of 100 matches, increasing for longer series such as the ashes to 70%. This finding is consistent with Table 2, where each of the eight matches with significant or suggestive runs test results fitted a non-trivial model, whereas, of the 12 matches showing no evidence of runs, three of the 12 matches had non-trivial models fitted.

For studying the type of non-random behaviour that was encountered, we conclude that those matches showing no evidence of runs may be genuinely completely random, though we should bear in mind when examining the remaining eight that the model that was fitted may not be the correct one.

3.3. Models for various sports

The type 2 models fitted best in both of the cricket matches, showing non-randomness but, surprisingly, the Penarol–Nacional football match data also fitted best to a type 2 model, a result which is in agreement with the runs test result in Table 1. This is far from an annual contest, with the two teams playing several times a year under a variety of auspices, sometimes only days

Table 3. Parameters of the MAICE model fitted to the Oxford–Cambridge Boatrace data, and (under) goodness-of-fit data†

Variable	*Coefficient*	*95% confidence interval*
Constant (1)	0.1523	(0.0607, 0.2438)
X_{t-1}	0.3986	(0.3068, 0.4904)
X_{t-2}	0.5766	(0.4837, 0.6695)
% weight difference	0.2082	(0.1519, 0.2646)
Sequence	Observed	Predicted
Lose after lose–lose	33	31.02
Win after lose–lose	10	11.98
Lose after lose–win	11	12.58
Win after lose–win	14	12.42
Lose after win–lose	11	12.34
Win after win–lose	14	12.66
Lose after win–win	14	13.01
Win after win–win	37	37.99

†The coefficients are for the model given in equation (1).

Table 4. MAICE model fitted to the ashes data and (under) goodness-of-fit data†

Variable	*Coefficient*	*95% confidence interval*
Constant (1)	0.4518	(0.3689, 0.5347)
θ draws offset	−1.1968	(−1.2747, −1.1188)
β_s	0.3249	(0.2890, 0.3653)
r	0.7063	(0.6721, 0.7374)
D_{t-1}	0.8040	(0.7188, 0.8993)
Home or away	−0.4143	(−0.4787, −0.3500)
Sequence	Observed	Predicted
Lose after lose	58	61.59
Draw after lose	32	28.54
Win after lose	33	32.87
Lose after draw	29	28.67
Draw after draw	35	34.66
Win after draw	22	22.67
Lose after win	34	33.23
Draw after win	19	21.58
Win after win	38	38.19

†The coefficients are for the model that is given in equation (2). β_s and r are parameters for the smoothed performance in equation (5). The home–away variable is coded as $\frac{1}{2}$.

apart. The draws do not always have the same total number of goals scored and probably do reflect an equality of ability of the two teams.

It is only feasible to discuss a sample of our results. Table 3 shows the fitted coefficient values and 95% confidence interval for the MAICE model fitted to the Oxford–Cambridge Boatrace data, and some information on model fit. The positive X_{t-1}- and X_{t-2}-terms show that the probability of winning depends directly on the outcomes of the two previous races. Crew weight is also a significant variable.

Table 4 shows results for the ashes series of test-matches. The probability of winning depends on previous smoothed performance, and the probability of drawing is greater if the last game was drawn. There is potential for more elaborate modelling here, as in the present treatment the information that matches are grouped into series has not been used.

Table 5 shows results for the Georgia–Florida American football match. Again, an exponentially smoothed estimate of performance is the best predictor of who will win.

None of the model types that were developed can be ruled out as irrelevant. In three of the eight matches where the runs test gave a significant or suggestive result, exponential smoothing models fitted best, whereas, of the seven of these contests in which draws were possible, three type 1 and four type 2 models fitted best.

It can be seen from an examination of the χ^2-statistics in Table 2, and the detailed statistics in Tables 3–5, that, whereas the naïve model often gave unrealistic expected numbers of such events, the fit of the preferred model was usually adequate.

In general, the explanatory power of the models relative to a naïve model, as measured by pseudo-R^2, is low as can be seen from Table 2. The mean R^2 is 0.054. In several cases, no improvement over the naïve model was possible, and $R^2 = 0$. The presence of match covariates naturally increases R^2, the mean value then being 0.106. The data that led to this study, the Oxford–Cambridge Boatrace, has the highest R^2 of the data sets that were studied, i.e. 0.255.

Table 5. MAICE model fitted to the Georgia–Florida American football match data and (under) goodness-of-fit data†

Variable	*Coefficient*	*95% confidence interval*
Constant (1)	0.1625	(0.0585, 0.2664)
θ draws offset	0.1112	(0.0593, 0.1631)
β_s	0.3766	(0.3336, 0.4250)
r	0.7403	(0.7102, 0.7677)
Sequence	Observed	Predicted
Lose after lose	19	19.04
Draw after lose	1	0.86
Win after lose	14	14.11
Lose after draw	1	0.96
Draw after draw	0	0.06
Win after draw	1	0.98
Lose after win	15	12.43
Draw after win	1	0.93
Win after win	21	26.65

†The coefficients are for the model that is given in equation (1). β_s and r are parameters for the smoothed performance in equation (5).

3.4. Link functions

The distribution function in equation (1) can be Gaussian (probit link), logistic (logistic link), etc. Whether using the logistic or probit link, in all cases the MAICE model was identical in all other respects, with the AIC varying typically by 0.1–0.2 between the two links. In the eight cases where the runs test showed non-randomness, the logistic link gave a lower AIC in six cases. The logistic distribution is very similar to the t-distribution with 9 degrees of freedom (Mudholkar and George, 1978), and so an exploration of various link functions including the approximation to the logistic link was done by using a t-distribution as link and stepping through the number of degrees of freedom. The variation in AIC was again small, typically 0.1–0.2. As by the weak symmetry principle the distribution must be symmetric, ruling out many possible link functions, the indication is that no improvement in fit can be attained by varying the link function, and that the use of the logistic link is marginally preferable.

3.5. Match covariates

Match covariates were available in six cases. The toss variable was coded as ± 1 and home or away as 1 or 2. Table 6 shows estimated match covariate coefficients, and p-values for the permutation likelihood ratio test of no effect. The increase Δp of the probability of winning as the match covariate value becomes favourable (e.g. on winning rather than losing the toss) is also shown. All these probability increases were calculated for the team showing the higher probability of winning, assuming that other covariates had no effect and, in matches where a draw could occur, the calculated increase is conditional on the outcome not being a draw.

In two out of four cases in Table 6, playing at home had a statistically significant positive effect on the probability of winning. Fig. 1 shows an example of the permutation distribution of the coefficient of the home or away covariate for the ashes data. This distribution is close to normal. The p-value of 0.07 for the ashes data does not attain the magic 5%-level, but it is suggestive.

Table 6. Match covariate point estimates, increase in probability of winning on changing from unfavourable to favourable covariate values and exact *p*-values of significance tests of the effect†

Sport	*Teams*	*Match covariate*	*Coefficient*	Δp	*p-value*‡
Boatrace	Oxford *versus* Cambridge	% weight difference	0.208	0.48	0.0005
Cricket	Oxford *versus* Cambridge	Toss	0.2661	0.13	NS (0.76)
Cricket	England *versus* Australia	Home or away	−0.414342	0.1	0.07
Rugby	England *versus* Scotland	Home or away	−0.918872	0.21	0.0183
Football	England *versus* Scotland	Home or away	−0.468938	0.13	0.15
Football	Barcelona *versus* Real Madrid	Home or away	−1.52964	0.43	<0.0001

†For the boatrace, the change is taken from 5% less than average weight to 5% above.
‡NS, not significant.

The increase in the probability of winning caused by the home or away effect can be very large. In the Barcelona–Real Madrid football match, there were 66 games where Barcelona won at home, 57 where they lost away and only 48 where they either lost at home or won away. This is clearly a significant effect. Although, in the case of the ashes, playing on the other side of the world in unfamiliar turf conditions might be expected to produce a slight drop in performance, the large effect in some of the ball games must surely be partly crowd induced and hence psychological.

Winning the toss (which was available in one cricket match) had a non-significant positive effect. We would expect the benefit to vary from sport to sport. More data with that covariate present is needed, but in the Varsity cricket match, although the effect is statistically not significant, it is of the right sign and could be responsible for a worthwhile increase in probability of winning of over 0.1.

Having a heavier crew in the boatrace has a very large positive effect, which can be seen by inspection of Fig. 2. A crew that is 5% heavier than the opponent's would increase the probability of winning very drastically, by 0.48.

3.6. Forecasting

The obvious naïve forecast for comparison is always to assume that the next outcome will be the most frequent outcome to date. Use of the models that were fitted here should give a modest improvement where runs of wins and losses or runs of draws often occur, and this is seen in examining the percentage of correct forecasts. For example, for the Oxford–Cambridge Boatrace, the naïve success rate is 52% correct, and the model gives nearly 67% of correct predictions.

However, this is an in-sample success rate, and the only trustworthy measure of forecasting success is to examine out-of-sample performance. The best solution seems to be to use an automatic model fitting procedure using only data up to time t to choose the MAICE model and to predict the next outcome. Table 7 shows results for such out-of-sample one-step-ahead forecasts. Clearly, fitting such models can indeed give a modest improvement in out-of-sample forecasting accuracy.

4. Conclusions

Although the outcome of many sporting contests appears completely random, a significant minority show significant non-randomness. There may be runs of wins by a team, runs of draws or runs of both types. This allows modelling of the outcome as a function of previous results,

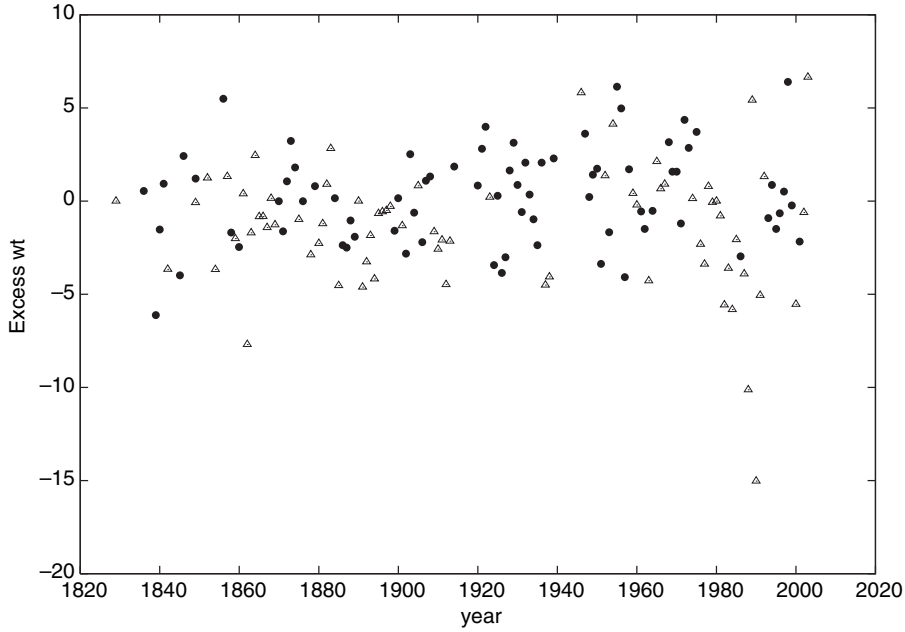

Fig. 2. Boatrace data: excess weight of the Cambridge crew as a percentage of average weight by year of race, with Cambridge wins (●) and Oxford wins (△) indicated

Table 7. Out-of-sample one-step-ahead forecast success rates compared with the 'naïve' method of predicting the most common outcome, with number of forecasts made

Sport	*Teams*	*% correct*	*% correct (naïve method)*	*Number*
Boatrace	Oxford *versus* Cambridge	63	50	131
Chess	Oxford *versus* Cambridge	49	37	108
Cricket	Oxford *versus* Cambridge	45	36	117
(Ashes)	England *versus* Australia	46	41	286
	Georgia *versus* Florida	48	48	61
	Oregon *versus* Oregon State	21	18	67
	Michigan *versus* Michigan State	39	43	82
	Penarol *versus* Nacional	35	34	459

with a resulting modest improvement in outcome prediction over a naïve prediction that the outcome will be the most frequent to date.

In the absence of insider knowledge about one or the other team, models must obey a weak symmetry principle, which restricts the form of possible models considerably. A strong symmetry principle requiring teams to perform identically can be used to ensure parsimony in model development.

What are here called match covariates, such as whether a game is played at home or away, sometimes have a large effect on the outcome of a match. Exact tests for the significance of these have been developed. There should be some interest from sports enthusiasts in this, because anything that reduces the competitive balance and hence the uncertainty in the outcome of a match will lessen its interest for spectators and the resulting gate money.

A theme in this paper has been the attempt to automate the modelling process, and to incorporate model choice into the statistical inference. Future work is needed on model selection, e.g. on small sample corrections to the AIC for ordinal and mixed models.

The treatment throughout has been likelihood based but frequentist. A Bayesian treatment would require a little further development. It would be necessary to assign prior distributions to the model parameters. Computation of their posterior distribution would then require low dimensional integrations of the likelihood function, which is probably best done by Gauss–Hermite integration with iterative scaling. Such an approach would have the advantage that Bayesian model choice criteria would then be available. Proceeding as indicated would be computationally more efficient than using Markov chain Monte Carlo methods.

The methodology has other potential application areas, e.g. results of general elections. A quick examination of the last 43 UK general election results, classified as Conservative wins or other, showed rather surprisingly no significant departure from randomness, as did the series of US Presidents, classified as Democrat or Republican. Extensions to contests of various types where more than two teams compete (such as the Eurovision Song Contest) are possible. Another example is the Rugby Union Six Nations contest, where England, Wales, Scotland, Ireland, France and Italy compete annually, each nation playing matches against all the others. In multiteam contests there are many modelling choices, and available data might consist merely of knowledge of the winning team, or all teams might be ranked. In the Six Nations contest, there is an overall winner, and there may also be a grand slam winner, if one nation beats all the others. The triple crown result is also of interest, where one of England, Wales, Scotland or Ireland beats all the others. Here a runs test showed no significant clustering of overall winner, and intuitively we would perhaps expect departures from randomness to be less evident in multiteam events.

The models that were developed here can also be applied to the performance of surgeons, which is of some topical interest. It is known that the success rate of a surgeon can change with time. Data on individual surgeons are recorded and have been studied, particularly for coronary artery bypass graft operations. Here the outcome is life or death of the patient, whose condition as measured for example by the Parsonnet score or more recently by EuroSCORE plays the role of the match covariate (e.g. Gogbashian *et al.* (2004)).

References

Cox, D. R. (1981) Statistical analysis of time series: some recent developments. *Scand. J. Statist.*, **8**, 93–115.
Cox, D. R. and Snell, E. J. (1989) *The Analysis of Binary Data*. London: Chapman and Hall.
Gogbashian, A., Sedrakyan, A. and Treasure, T. (2004) EuroSCORE: a systematic review of international performance. *Eur. J. Cardthorac. Surg.*, **25**, 695–700.
Hosmer, D. W., Hosmer, T., le Cessie, S. and Lemeshow, S. (1997) A comparison of goodness of fit tests for the logistic regression model. *Statist. Med.*, **16**, 965–980.
MacDonald, I. L. and Zucchini, W. (1997) *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman and Hall.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Mudholkar, G. S. and George, E. O. (1978) A remark on the shape of the logistic distribution. *Biometrika*, **65**, 667–668.
Nagelkerke, N. J. D. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691–692.
Szymanski, S. (2001) Income inequality, competitive balance and the attractiveness of team sports: some evidence and a natural experiment from English soccer. *Econ. J.*, **111**, f69–f94.
Yanagihara, H., Sekiguchi, R. and Fujikoshi, Y. (2003) Bias correction of AIC in logistic regression models. *J. Statist. Planng Inf.*, **115**, 349–360.