Outcome uncertainty and attendance demand in sport: the case of English soccer

Forrest, D., & Simmons, R. (2002)
Journal of the Royal Statistical Society

Presenter: Sarah Kim
2019.01.25
Introduction

- **Uncertainty of outcome:**
  A situation where a given contest within a league structure has a degree of unpredictability about the result.

- Using betting odds by bookmakers, we set up a measure of uncertainty of outcome.

- Given suitable controls, we find that soccer match attendances are indeed maximized where the uncertainty of outcome is greatest.
Data

1. Attendance data
   - We collected data for all matches played on Sat. between 1997.10 and 1998.05 and excluded the August and September period because we intended to use as regressors.
   - We consider only the 872 matches played in divisions 1, 2 and 3 of the Football League.

2. Betting data
   - **Fixed odds betting**: British bookmakers set the odds of soccer bets several days before a match and then these remain unaltered through the betting period.
   - For each match in our sample, we collected the odds for a home team win, draw and away team win.
Probability model for match outcomes

First using an ordered probit model, we regress match outcomes (home win, 0; draw, 1; away win, 2) on BOOKPROB(H) and DIFFATTEND:

- **BOOKPROB(H):** $\frac{\text{podds}(H)}{\sum_{e \in \{H, D, A\}} \text{podds}(e)}$ for $e \in \{H, D, A\}$, and podds is the probability odds (e.g. 3 : 1 becomes 0.25).
- **DIFFATTEND:** (the mean home club home attendance for the previous season)—(the mean away club home attendance for the previous season).

We have a latent regression given by

$$y^* = \beta_1 \text{BOOKPROB(H)} + \beta_2 \text{DIFFATTEND} + \epsilon,$$

where $y^*$ is an unobserved latent variable, and $\epsilon$ is a normally distributed error term.
Probability model for match outcomes

- We observe
  - RESULT = 0 if \( y^* \leq 0 \),
  - RESULT = 1 if \( 0 < y^* \leq \mu \),
  - RESULT = 2 if \( \mu \leq y^* \),

where \( \mu \) is a threshold parameter to be estimated.

- We have the following probabilities:
  - \( \text{Prob}(\text{RESULT} = 0) = 1 - \Phi(\beta'x) \)
  - \( \text{Prob}(\text{RESULT} = 1) = \Phi(\mu - \beta'x) - \Phi(-\beta'x) \)
  - \( \text{Prob}(\text{RESULT} = 2) = 1 - \Phi(\mu - \beta'x) \),

where \( x = [\text{BOOKPROB}(H), \text{DIFFATTEND}] \).
Probability model for match outcomes

- The ordered probit regression equation was used to generate estimated probabilities of home wins and away wins.
- In the 872 matches, the predicted probability of an away win exceeded that of a home win in only 72 (8.2%) cases.
- Let “PROBRATIO” be the estimated ratio of the probability of a home win to the probability of an away win.
- PROBRATIO is our measure of match uncertainty of outcome.
Attendance demand model

- Denote dependent variable LOGATTENDANCE by $A_i$ where $i$ is a home team identifier.

- **Attendance demand model**

$$A_i = \alpha + \gamma_1 \text{PROBRATIO}_i + \gamma_2 \text{PROBRATIO}^2_i$$

$$+ \gamma_3 \text{HOMEPOINTS}_i + \gamma_4 \text{AWAYPOINTS}_i + \gamma_5 \text{DIST}_i + \gamma_6 \text{DIST}^2_i$$

+ month dummies + error,

where DIST is the distance between grounds of competing teams.

- We include month dummy variables to capture the effects of weather, alternative seasonal attractions.
### Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Absolute t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>8.56</td>
<td>123.67</td>
<td>0.000</td>
</tr>
<tr>
<td>PROBRATIO</td>
<td>-0.067</td>
<td>5.26</td>
<td>0.000</td>
</tr>
<tr>
<td>PROBRATIO$^2$</td>
<td>0.0046</td>
<td>5.14</td>
<td>0.000</td>
</tr>
<tr>
<td>HOMEPOINTS</td>
<td>1.00</td>
<td>8.25</td>
<td>0.000</td>
</tr>
<tr>
<td>AWAYPOINTS</td>
<td>0.01</td>
<td>0.13</td>
<td>0.897</td>
</tr>
<tr>
<td>DIST</td>
<td>-0.0026</td>
<td>8.88</td>
<td>0.000</td>
</tr>
<tr>
<td>DIST$^2$</td>
<td>0.00000060</td>
<td>6.74</td>
<td>0.000</td>
</tr>
<tr>
<td>OCTOBER</td>
<td>0.0128</td>
<td>0.58</td>
<td>0.559</td>
</tr>
<tr>
<td>DECEMBER</td>
<td>-0.073</td>
<td>2.89</td>
<td>0.004</td>
</tr>
<tr>
<td>JANUARY</td>
<td>0.045</td>
<td>2.08</td>
<td>0.038</td>
</tr>
<tr>
<td>FEBRUARY</td>
<td>0.016</td>
<td>0.75</td>
<td>0.454</td>
</tr>
<tr>
<td>MARCH</td>
<td>0.036</td>
<td>1.68</td>
<td>0.094</td>
</tr>
<tr>
<td>APRIL/MAY</td>
<td>0.12</td>
<td>5.54</td>
<td>0.000</td>
</tr>
<tr>
<td>$R^2$ (within)</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

†Dependent variable: LOGATTENDANCE.
Results

- The absolute quality of the home team in the season influences the match attendance.
- The quadratic specification of distance captures the curvature of the relationship between attendance and distance (the turning point is at 350 km).
- For the month dummy variables, soccer attracts least support in December when Christmas, whereas interest peaks in April and May when promotion and play-off issues.
- By the coefficient of PROBRATIO, as uncertainty decreases, so also does attendance.
On determining probability forecasts from betting odds

Štrumbelj, E. (2014)

*International journal of forecasting*
Introduction

- There is substantial empirical evidence that betting odds are the most accurate publicly-available source of probability forecasts for spots.
- There are two issues:
  1. Which method should be used to determine probability forecasts from raw betting odds?
  2. Does it make a difference as to which bookmaker or betting exchange we choose?
Determining outcome probabilities from betting odds

1. Basic normalization

- Let \( o = (o_1, \ldots, o_n) \) be the quoted odds for a match with \( n \geq 2 \) possible outcomes, and let \( o_i > 1 \) for all \( i = 1, \ldots, n \).

- For each \( i \), define an inverse odds \( \pi_i = \frac{1}{o_i} \).

- Let \( \beta = \sum_{i=1}^{n} \pi_i \) be the booksum. Dividing by the booksum, \( p_i = \frac{\pi_i}{\beta} \) can be interpreted as outcome probabilities.

- We refer to this as **basic normalization**.
Determining outcome probabilities from betting odds
Assumptions of Shin’s model

▶ Shin’s model is based on the assumption that bookmakers odds which maximize their expected profit in the presence of uninformed bettors and a known proportion of insider traders.

▶ The bookmaker and the uninformed bettors share the probabilistic beliefs $p = (p_1, \ldots, p_n)$, while the insiders know the actual outcome.

▶ W.L.O.G., assume that the total volume of bets is 1, of which $1 - z$ comes from uninformed bettors and $z$ from insiders.
2. Shin’s model

- Conditional outcome \(i\) occurring, the expected volume bet on the \(i\)th outcome is \(p_i(1 - z) + z\).

- If the bookmaker quotes \(o_i = \frac{1}{\pi_i}\) for outcome \(i\), the expected liability for the outcome \(\frac{1}{\pi_i}(p_i(1 - z) + z)\).

- By assuming that the bookmaker has probabilistic beliefs \(p\), the bookmaker’s unconditional expected liabilities is \(\sum_{i=1}^{n} \frac{p_i}{\pi_i}(p_i(1 - z) + z)\), and the total expected profit

\[
T(\pi, p, z) = 1 - \sum_{i=1}^{n} \frac{p_i}{\pi_i}(p_i(1 - z) + z).
\]

- The bookmaker sets \(\pi\) to maximize the expected profit, subject to \(0 \leq \pi_i \leq 1\).
Determining outcome probabilities from betting odds

3. Regression analysis

▶ Use a statistical model to predict the outcome probabilities from odds.

▶ For sports with three outcomes (home, draw, away), we use an ordered logistic regression model with (inverse) betting odds as input variables.
Comparison

- We compare three different methods for determining probabilities from betting odds.

- Let \( p = (p_1, \ldots, p_n) \) be our probability estimates and \( a \) the vector indicating the actual outcome.

- The Brier score of a single forecast is defined as

\[
\text{BRIER}(p, a) = \frac{1}{n} \|p - a\|^2
\]

and RPS as

\[
\text{RPS}(p, a) = \frac{1}{n} \|C(p) - C(a)\|^2,
\]

where \( C(x) = (C_1(x), \ldots, C_n(x)) \), \( C_i(x) = \sum_{j=1}^{i} x_i \) is the cumulative distribution.
## Comparison

<table>
<thead>
<tr>
<th>Sport</th>
<th>Method</th>
<th>BRIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basketball</td>
<td>Shin</td>
<td>0.1908</td>
</tr>
<tr>
<td></td>
<td>Basic</td>
<td>0.1912</td>
</tr>
<tr>
<td></td>
<td>LogitR</td>
<td>0.1926</td>
</tr>
<tr>
<td>Handball</td>
<td>Shin</td>
<td>0.1320</td>
</tr>
<tr>
<td></td>
<td>Basic</td>
<td>0.1325</td>
</tr>
<tr>
<td></td>
<td>LogitR</td>
<td>0.1332</td>
</tr>
<tr>
<td>Hockey</td>
<td>Shin</td>
<td>0.2057</td>
</tr>
<tr>
<td></td>
<td>Basic</td>
<td>0.2057</td>
</tr>
<tr>
<td></td>
<td>LogitR</td>
<td>0.2061</td>
</tr>
<tr>
<td>Soccer</td>
<td>Shin</td>
<td>0.1998</td>
</tr>
<tr>
<td></td>
<td>Basic</td>
<td>0.1999</td>
</tr>
<tr>
<td></td>
<td>LogitR</td>
<td>0.2003</td>
</tr>
</tbody>
</table>

**Figure**: Comparison of three models using the Brier scores
Comparison

<table>
<thead>
<tr>
<th>Bookmaker</th>
<th>Basic</th>
<th></th>
<th>Shin</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>bet.at.home</td>
<td>0.2114</td>
<td>0.1857</td>
<td>0.2109</td>
<td>0.1818</td>
</tr>
<tr>
<td>bet365</td>
<td>0.2110</td>
<td>0.1837</td>
<td>0.2106</td>
<td>0.1820</td>
</tr>
<tr>
<td>Betclic</td>
<td>0.2116</td>
<td>0.1882</td>
<td>0.2110</td>
<td>0.1841</td>
</tr>
<tr>
<td>Betsafe</td>
<td>0.2112</td>
<td>0.1861</td>
<td>0.2107</td>
<td>0.1828</td>
</tr>
<tr>
<td>bwin</td>
<td>0.2113</td>
<td>0.1833</td>
<td>0.2109</td>
<td>0.1803</td>
</tr>
<tr>
<td>DOXXbet</td>
<td>0.2113</td>
<td>0.1895</td>
<td>0.2107</td>
<td>0.1858</td>
</tr>
<tr>
<td>Expekt</td>
<td>0.2115</td>
<td>0.1870</td>
<td>0.2110</td>
<td>0.1837</td>
</tr>
<tr>
<td>Interwetten</td>
<td>0.2123</td>
<td>0.1963</td>
<td>0.2113</td>
<td>0.1895</td>
</tr>
<tr>
<td>NordicBet</td>
<td>0.2112</td>
<td>0.1871</td>
<td>0.2107</td>
<td>0.1842</td>
</tr>
<tr>
<td>Betfair</td>
<td>0.2101</td>
<td>0.1833</td>
<td>0.2100</td>
<td>0.1810</td>
</tr>
</tbody>
</table>

**Figure**: Comparison of bookmakers using the mean and median RPS scores