Bachelor Thesis: Fixed Odds Betting and Real Odds

by Jakub Vojtik

Referee Report

June 14, 2017

This diploma work focuses on the the relationship of real odds with betting odds. The author limits his data analysis to an extensive set of tennis data. The topic itself is original and the results are interesting. It can pass as a bachelor thesis.

Some of the points made in the thesis are worth a deeper discussion. The main point is that the author uses the approach that the same margin is applied to all the existing market selections and proves some results justifying this approach (like Theorem 1.3). In other words, if we denote by B_1 and B_2 the odds on the two selections and by m the margin, we have

$$\frac{1}{B_1} + \frac{1}{B_2} = 1 + m$$

and the relationship to the true probabilities p_1 and p_2 should follow

$$p_i = \frac{1}{B_i(1+m)},$$

or

$$B_i = \frac{1}{p_i(1+m)}.$$

This approach means that the odds are just linearly rescaled inverse probabilities. However, linear scaling of probabilities is not a desirable approach as the odds can get under 1 (for some large selections of p_i and m like $p_i = 0.95$ and $m = 0.07$) and need to be artificially rounded back to 1 (meaning that the selection is not quoted).

My approach to the margin treatment would be somewhat different. Computation of probabilities in regression models is done with logit (or probit) regression. The basic idea is that the estimation is done on the entire real line $\mathbb{R}$, and this estimated number is in turn transformed to an interval $[0,1]$. So it uses a mapping

$$\mathbb{R} \to [0,1].$$

One possible choice of this mapping is a logistic function

$$\text{Logit}(x) = \frac{1}{1+e^{-x}}$$

that is used in the logit regression. As described earlier, the estimation of probability is first done on a linear scale, giving some y_{est} and the estimated probability is then $p_{est} = \text{Logit}(y_{est})$. A possible error ε of the estimate is treated on the linear scale, giving the true value $y = y_{est} + \varepsilon$ and the true probability $p = \text{Logit}(y_{est} + \varepsilon)$. Obviously, we do not see the true values of ε and the resulting y and p, we only statistically estimate them.

Using the principle that the linear estimator is transformed to $[0, 1]$ via a logit transformation, we may adapt this idea for computation of the quotes with the prescribed margin m. As a byproduct, this will also address the question that we may observe the probabilities with some error and the logit transform treats this error proportionally to the probability p. The basic idea is that we find some constant K to both B_1 and B_2 in the following way:

$$B_1 = \frac{1}{\mathrm{Logit}(\mathrm{InvLogit}(p_1) + K)}$$

and

$$B_2 = \frac{1}{\mathrm{Logit}(\mathrm{InvLogit}(p_2) + K)}$$

such that the sum of the inverse odds give the required margin. The function InvLogit(p) defined as

$$\mathrm{InvLogit}(p) = \log\left(\frac{p}{1-p}\right)$$

is the inverse of the Logit function.

As a byproduct of this approach, one can find the function that finds p_i from B_1 and B_2. The formula for the corresponding probability (on the selection corresponding to the quote B_1) is given by

$$p(B_1, B_2) = \begin{cases} \frac{1-B_2+\sqrt{(B_1-1)(B_2-1)}}{B_1-B_2} & B_1 \neq B_2, \\ \frac{1}{2}, & B_1 = B_2. \end{cases}$$

Note that $p(B_1, B_2) + p(B_2, B_1) = 1$, which gives two alternative computations for probability implied on the second selection (i.e., either directly from $p(B_2, B_1)$, or by computing $1 - p(B_1, B_2)$. In contrast, the suggested transformation of odds to probabilities is given by a formula

$$n(B_1, B_2) = \frac{\frac{1}{B_1}}{\frac{1}{B_1} + \frac{1}{B_2}} = \frac{B_2}{B_1 + B_2}.$$

Example: Imbalanced odds combined with a large margin lead to serious estimation discrepancies of the two approaches. For instance, the tennis game Pella vs Polmans quoted $B_1 = 1.04$ and $B_2 = 11.5$. Applying the formula gives $p(1.04, 11.5) = 0.941867$. Probability on the second selection easily follows to be $p(11.5, 1.04) = 0.0581333$. The other inversion gives $n(1.04, 11.5) = 0.917065$ and $n(11.5, 1.04) = 0.0829346$, which is substantially off from the suggested approach.

My guess is that the betting market follows more the logit type margin than the linear type margin and thus we may see some evidence that supports it. I think that the non-identity behavior in graphs on page 25 in the regions around $p = 0$ and $p = 1$ is a consequence of that.

My other points for discussion are:

- The quality of data is somewhat questionable given the unusual distribution of the odds.

- It is impossible to make conclusions about the probabilities implied by betting without seeing the actual profit and loss distribution. The author uses central limit theorem which assumes that we observe identically and independently distributed random variables. However, each game is unique and the assumption that we may apply statistical analysis of a large sample on them may be a bit far fetched. For instance, if we see 100 games with win probability 0.5 and we see that the actual outcomes are indeed statistically in line, it does not mean that each game from this sample was estimated correctly. The knowledge of bettors may in fact exhibit in a limited number of games that is incorrectly estimated.
- I am not sure if the linear regression in part 2.2 is appropriate in contrast to logit or probit regression.

Jan Vecer,
KPMS, MFF UK,
Sokolovska 83
18675 Praha 8
Email: vecer@karlin.mff.cuni.cz