# Probabilistic Forecasting with Discrete Choice Models: Evaluating Predictions with Pseudo-Coefficients of Determination

Ming-Chien Sung*, David C.J. McDonald, Johnnie E.V. Johnson

*Centre for Risk Research, Southampton Business School, University of Southampton,*
*Southampton, SO17 1BJ, UK.*

## Abstract

Probabilistic forecasts from discrete choice models, which are widely used in marketing science and competitive event forecasting, are often best evaluated out-of-sample using pseudo-coefficients of determination, or pseudo-$R^2s$. However, there is a danger of misjudging the accuracy of forecast probabilities of event outcomes, based on observed frequencies, because of issues related to pseudo-$R^2$s. First, we show that McFadden's pseudo-$R^2$ varies predictably with the number of alternatives in the choice set. Then we evaluate the relative merits of two methods (bootstrap and asymptotic) for estimating the variance of pseudo-$R^2$s so that their values can be appropriately compared across non-nested models. Finally, in the context of competitive event forecasting, where the accuracy of forecasts has direct economic consequence, we derive new $R^2$ measures that can be used to assess the economic value of forecasts. Throughout, we illustrate using data drawn from UK and Ireland horse race betting markets.

*Keywords:* Forecasting, Decision analysis, Finance, Discrete choice models, Horseracing

*Corresponding author. Tel.: +44 23 8059 8974. Fax: +44 23 8059 3844.
*Email addresses:* m.sung@soton.ac.uk (Ming-Chien Sung), d.mcdonald@soton.ac.uk (David C.J. McDonald), jej@soton.ac.uk (Johnnie E.V. Johnson)

## 1. Introduction

Decision makers often face choices between different alternatives $A_i$, $i = 1, \ldots, n$, with payoffs $x_{ij}$ depending on the future state of the world $W_j$, $j = 1, \ldots, m$. Normative models of decision making indicate that the decision maker should select the option with the highest expected utility,

$$EU(A_i) = \sum_{j=1}^{m} p(W_j)U(x_{ij}), \tag{1}$$

where $p(W_j)$ is the probability of future state $W_j$ occurring and $U(x_{ij})$ is the utility of outcome $x_{ij}$ if alternative $A_i$ were selected and state $W_j$ occurred. Clearly, to determine which alternative is associated with the maximum expected utility it is important to develop accurate estimates of the utility values $U(x_{ij})$. In addition, rather than simply predicting which state of the world is most likely to occur, it is important to accurately forecast the probabilities of the different states of the world, $p(W_j)$. It is the means of improving such probabilistic forecasts derived from discrete choice models that forms the focus of this paper.

The importance that organizations attach to accurate probabilistic forecasts is highlighted by the significant growth in the adoption of prediction markets. These are essentially internal betting markets that attempt to tap into the dispersed 'wisdom of the crowd' to produce probability forecasts for a range of possible future events. They have been used by many companies such as Hewlett Packard, Eli Lilly, General Electric, and Google to predict a variety of uncertain outcomes, from the likelihood of the success of new products to the probability of meeting project deadlines (Plott & Chen, 2002; Cowgill et al., 2009). Forecasts derived from prediction markets can be combined with other information (e.g., data concerning the characteristics of products that have been successful in the past or previous completion times) in an attempt to derive accurate probabilistic forecasts. Discrete choice models, such as conditional logit (CL) and multinomial probit, are ideal for performing this function and are widely applied in probabilistic forecasting. Their primary use has been in predicting individuals' choices from a range of alternatives, so they have been employed in consumer choice and marketing (Lin & Sibdari, 2009) and econometrics (Maddala, 1983), but have also been adopted in such diverse fields as epidemiology (Breslow & Day, 1994), operations research (Cheng & Stough, 2006), and the forecasting of competitive events (Smith & Vaughan Williams, 2010).

The development of novel and sophisticated discrete choice methods, particularly to improve the accuracy of forecast probabilities, has received a great deal of attention (e.g., Liu, 2011; Abe, 1999). In general terms, probabilistic forecasts are regarded as being accurate when the relative frequencies of observed events match the forecast probabilities (Maddala, 1983). However, there has been relatively little consideration in the literature to the manner in which probabilities derived from discrete choice models are evaluated out-of-sample, yet this is essential for maximizing the accuracy of probabilistic forecasts. A key property of any means of evaluating the accuracy of forecast probabilities is its comparability across empirical models (Kvålseth, 1985). Otherwise, the researcher cannot be certain whether differences in the evaluation arise because of changes in the model's predictive power or because of confounding factors, such as properties of the data. Comparability is also reliant on being able to assign degrees of uncertainty to forecast point estimates, in order to ensure that conclusions drawn from evaluating measures of accuracy are statistically significant and hence robust to randomness.

In linear models, the coefficient of determination $R^2$ is widely used as a measure of a model's ability to explain variation in the data, and thus the accuracy of the model's forecasts. The properties of $R^2$ and when and how it should be employed are now well-understood (e.g., Kvålseth, 1985; Draper & Smith, 1998), although with some caveats in the case of out-of-sample forecasting (Armstrong, 2001). However, in this paper, we are concerned with the out-of-sample forecasting accuracy of choice models, which are nonlinear. Measures such as information criteria would be useful for assessing predictive accuracy in this case, but only if one were interested in the accuracy of specific event prediction, such as binary win/loss predictions. This is not our focus. Rather, we are concerned with the calibration of probability forecasts. In such settings, pseudo-$R^2$s, which are equivalent measures to $R^2$ for nonlinear models, are generally recommended (e.g., Greene, 2012; Maddala, 1983). In fact, in any environment where one seeks to use probability predictions to make pecuniary gain (e.g., in options, futures, spread trading and betting markets), it can be shown that there is a direct link between increases in pseudo-$R^2$ values and out-of-sample returns (e.g., Benter, 1994; Lessmann et al., 2012).

Despite the literature's focus on the use of pseudo-$R^2$s to assess the out-of-sample forecasting accuracy of choice models, there is little consensus regarding the properties and appropriate application of pseudo-$R^2$s for this task. To begin with, there are significant differences between $R^2$s and pseudo-$R^2$s. So, while pseudo-$R^2$s are commonly reported (Cheng & Stough, 2006; Schnytzer et al., 2010), their usage is seldom justified (Veall & Zimmerman, 1996), and there are still many unresolved issues associated with them. First, unlike $R^2$, there is no single definition of pseudo-$R^2$ that is universally employed. Rather, a variety of measures has been proposed, which may have different interpretations (Menard, 2000). Second, they are not necessarily comparable across different datasets. Finally, the distributions of $R^2$s are complex and depend on unknown parameters (Ohtani, 2000). For pseudo-$R^2$s, this issue is exacerbated because not only are the distributional properties of pseudo-$R^2$s different to those of $R^2$, they also depend on the particular definition of pseudo-$R^2$ employed *and* the choice of model. Consequently, while pseudo-$R^2$s are often reported, they are seldom accompanied by standard errors (Press & Zellner, 1978), meaning significance tests are often ignored. The above considerations have serious consequences for the development of accurate discrete choice models because they impact effective evaluation of the forecasting accuracy of these models. Since these issues are rarely examined, there is a significant danger of selecting a sub-optimal model or misinterpreting the relative forecasting ability of different discrete choice models.

One of the many applications for discrete choice models is in the forecasting of competitive event (CE) outcomes (e.g., Lessmann et al., 2009, 2012). A CE is a contest between at least two rival participants, where (generally) one winner is declared and the outcome is uncertain, such as political elections or sporting events. Probabilistic forecasting in this context involves estimating the probability of the various competitors winning. Often, these events are associated with markets for betting or trading on their outcome, e.g., betting markets in the case of sporting events (Sung et al., 2012), or prediction markets for political contests or for outcomes associated with business policies (Wolfers & Zitzewitz, 2006). Since the outcomes of CEs are of particular interest for economic (in the case of sporting events or business policies) or policy reasons (elections), the forecasting of CEs is a prominent subject in the literature (e.g., Schnytzer et al., 2010).

The standard modeling approach is to view competitors as alternatives in a choice set with the winner being the participant whose attributes lead it to being 'preferred', hence, the suitability of discrete choice models. While the typical motivation for pseudo-$R^2$ in out-of-sample forecast evaluation is as a measure of improvement from the null model (where each alternative is considered equally likely) to the fitted model (e.g., Benter, 1994; Franck et al., 2010), a more useful measure would be improvement of the fitted model from a model based on the forecasts of prices from the associated market. In the case of prediction markets, this would be the degree to which the fitted model (incorporating the prediction market prices together with information from other data sources, such as the probability of meeting project deadlines in the past) improves on the probability forecasts derived directly from the prediction market prices. Such a measure, which we call *relative* pseudo-$R^2$, would be of value because there would be a direct link between relative pseudo-$R^2$s and the economic value of the forecast probabilities derived from the fitted model. Most importantly, this metric would be comparable across different market settings.

In this paper, we address the above unresolved issues related to pseudo-$R^2$s and illustrate these points empirically with data drawn from CEs. Throughout the paper we refer to the CL model, which is the most widely-used in this context, although our findings readily extend to other discrete choice models. We show that at least one commonly reported pseudo-$R^2$ measure is not robust to changes in the number of alternatives in each event. Consequently, in order for discrete choice models to be comparable using these measures across non-nested models or across models evaluated on different datasets, we suggest a rescaling of the pseudo-$R^2$. We also demonstrate, with an example from CEs, a potential misinterpretation that may arise from the bias if this rescaling is not employed.

We also describe means of conducting significance tests for comparing pseudo-$R^2$ values from different forecasting models. In particular, we describe two methods for obtaining estimates of the variance of pseudo-$R^2$ measures, the bootstrap and asymptotic methods, and find that they both produce estimated variances that are reasonably close. Consequently, either method could be used to conduct significance tests for comparing pseudo-$R^2$ values.

In addition, we define relative pseudo-$R^2$s that measure the improvement of a fitted model from a model based on the forecast prices in the associated betting or prediction market (in the context of out-of-sample forecasting of CEs). This provides a comparable metric across different market settings, and shows that there is a relationship between relative pseudo-$R^2$s and the economic value of forecast probabilities, a finding that has important implications for assessing the efficiency of financial markets.

Throughout the paper we illustrate the value of the approaches and measures we introduce using data drawn from horserace forecasting. These data contain the essential features of all complex choice modelling problems, including different numbers of alternatives (horses) in different choice sets (races). In addition, they offer the advantage, for the purpose of illustrating the applicability of the techniques we suggest, that a certain point in time (the end of the race) all uncertainty is resolved and we are able to assess the accuracy of forecasts. In addition, the associated betting markets provide us with a means of assessing the economic value of forecast probabilities.

## 2. The conditional logit model, pseudo-$R^2$s, and dependence on the number of alternatives

### *2.1. The conditional logit model*

The conditional logit (CL) model (McFadden, 1974) is employed to estimate or forecast the probability of each alternative being chosen (based on attributes of the alternatives) in situations where a decision maker or 'nature' selects a specific alternative from a number of competing alternatives. The utility of each alternative $i$ in event $j$ is given by

$$W_{ij} = \beta^\top x_{ij} + \epsilon_{ij}, \tag{2}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_m)^\top$ are the coefficients of the attributes $x_{ij}$, and $\epsilon_{ij}$ is an independent error term. There are likely to be many independent factors that contribute to the error term and, consequently, the central limit theorem can be used to justify the assumption that the $\epsilon_{ij}$ are normally distributed, resulting in a probit specification. However, to make the forecast probabilities tractable, it is generally assumed that the error term has a double exponential distribution $f(x) = \exp[-x - \exp(-x)]$ (Maddala, 1983); in practice, the difference between logit and probit models is small (Judge et al., 1985). The probabilities are then given by

$$\begin{aligned} p_{ij} &= Pr(W_{ij} > W_{kj}, k = 1, 2, \ldots, n_j, k \neq i) \qquad (3)\\ &= \frac{\exp(\beta^\top x_{ij})}{\sum_{k=1}^{n_j} \exp(\beta^\top x_{kj})}, \end{aligned}$$

where $n_j$ is the number of alternatives. The coefficients are estimated by maximum likelihood:

$$\ln L = \sum_{j=1}^{N} \sum_{i=1}^{n_j} y_{ij} \ln p_{ij}, \tag{4}$$

where $y_{ij} = 1$ if alternative $i$ is chosen, $y_{ij} = 0$ otherwise, and $N$ is the total number of choice problems. We denote the maximized log-likelihood function by $\ln L(\hat{\beta})$.

### *2.2. Pseudo-$R^2$s*

The coefficient of determination $R^2$ is a popular goodness-of-fit measure in linear models. Varying between 0 and 1, it has multiple interpretations: (i) the proportion of variation explained by the model, (ii) the square of the correlation between predicted and observed values, and (iii) the improvement from a null model (with no independent variables) to the fitted model. For nonlinear models, a range of alternative pseudo-$R^2$ measures have been proposed. The motivation for these is primarily interpretation (iii) above, i.e., improvement from null to fitted model. CL is an example of a model estimated by maximum likelihood, and in fact, for *any* model estimated by maximum likelihood, pseudo-$R^2$s that satisfy this criterion can be defined.

The most widely used measure (e.g., Franck et al., 2010) is the McFadden (1974) pseudo-$R^2$, which is given by

$$R_M^2 = 1 - \frac{\ln L(\hat{\beta})}{\ln L(0)}, \tag{5}$$

where $\ln L(0)$ is the $\ln L$ of the null model, where each alternative is assigned the same probability of being chosen:

$$\ln L(0) = \sum_{j=1}^{N} \ln(1/n_j). \tag{6}$$

An alternative is the Maddala (1983) pseudo-$R^2$, given by

$$R_D^2 = 1 - \exp\{-(2/N)[\ln L(\hat{\beta}) - \ln L(0)]\}. \tag{7}$$

The McFadden pseudo-$R^2$ has a maximum value of 1, while the maximum value of Maddala's is $R_D^2 = 1 - \exp\{(2/N)\ln L(0)\}$. Nagelkerke (1991) proposed that Maddala's definition be rescaled so that it takes a maximum of 1, but the original definition actually has the desirable property of alternatives independence (see the next section), so we would not recommend this rescaling. In this paper we consider only these two definitions of pseudo-$R^2$, which are the two most popular, but our results are readily extended to other versions that have been proposed (e.g., Nagelkerke, 1991).

A major concern in probabilistic forecasting is that of evaluating the accuracy of forecast probabilities. We believe that pseudo-$R^2$s are the most fundamentally important tool for assessing the performance of discrete choice models designed to forecast the probabilities of future events. Maximizing pseudo-$R^2$ is equivalent to maximum likelihood, where the criterion essentially chooses the set of parameters which maximizes the probability of observing the particular set of alternatives that are observed ex post. It is especially important in choice models to maximize the model probability for the alternative that is eventually chosen. For example, from the perspective of an organization, it is important when estimating the probability of success of various products, to maximize the model probability of success for the product that turns out to be successful (from a range of possible products), as this will increase the chance that this product has the highest expected utility and is then the one that is selected. In addition, the forecast probability of success of this product is likely to affect the level of investment from the organization.

Equally, in probabilistic forecasting of CEs, it is important to maximize the probability for the eventual winning competitor, since a decision of whether or not and how much to bet on that competitor will depend on the predicted 'edge' to a bet on that competitor $p_j/q_j$, where $q_j$ is some baseline probability that the model is being compared against (the null probability later in this section, and the market probability in section 4). Indeed, this is an approach advocated by Benter (1994), who is reported to be the most successful bettor of all time. He recommends the application of CL models (incorporating a range of explanatory variables) to develop forecasts of the winning probability for each competitor and emphasizes the role of pseudo-$R^2$ as a measure of the 'explanatory power' of a model. He indicates that pseudo-$R^2$ is the best means he has found for comparing the efficacy of alternative models. In addition, it has been shown that probabilities estimated by maximum likelihood yield maximum in-sample return (on investments based on these forecast probabilities) to a log utility investor (Johnstone, 2011). Moreover, increases in out-of-sample returns generally result directly from an increase in pseudo-$R^2$ (Lessmann et al., 2012). Note that the emphasis here is on the accuracy of the *probabilities* derived from discrete choice models, rather than a binary classification of win/lose (chosen / not chosen). It is a common misconception in forecasting of CEs that the goal above all else is to pick winners, whereas it is easy to demonstrate that maximizing edge is the key to success. Therefore, traditional measures of forecasting accuracy such as percentage of correct classifications are significantly less relevant in this context.

### *2.3. Dependence on the number of alternatives*

When comparing two or more competing models, a problem may arise when the models are compared over different data. Specifically, the datasets might differ in their underlying characteristics. This could arise when subsets of data are sampled, or when data is split into training and holdout samples. For example, Hyndman & Koehler (2006) give the example from time series of scale-dependent measures being compared across datasets with different scales. Menard (2000) shows that the Maddala pseudo-$R^2$ has a dependence on the base rate in logistic regression models. A specific example in discrete choice modelling is that, depending on how the data are sampled, the average number of alternatives available to each subject may vary. For example, in CEs, we might seek to analyze variations in the predictability of horseraces depending on the number of horses in each race. Here, we sample alternative datasets depending on the number of runners in each race. Consequently, the average number of competitors will be different in each dataset. This presents us with a problem, as we cannot adequately assess the predictability of these events by comparing $R^2$s over these samples using the McFadden pseudo-$R^2$. We now demonstrate this.

Recall that the motivation behind pseudo-$R^2$s is that they measure the degree of improvement from the null to the fitted model. Null model probabilities are $1/n_j$, since without predictors we cannot make any distinction between alternatives. Suppose that, for each set of alternatives $j$, our model assigns a probability of $p_j = f_j/n_j$ to the eventual chosen alternative, i.e., model probabilities have an 'edge' $f_j \leq n_j$ over the null probabilities. In this way, pseudo-$R^2$s can be evaluated for their dependence on (or independence from) the number of alternatives. These dependencies are given in the following proposition (for a proof, see Appendix B).

**Proposition 1.** *If the model probabilities assigned to the eventual chosen alternative are $p_j = f_j/n_j$, then the McFadden and Maddala pseudo-$R^2$s are given by*

$$\begin{aligned} R^2_M &= \frac{\ln \tilde{f}}{\ln \tilde{n}}, \\ R^2_D &= 1 - 1/\tilde{f}^2, \end{aligned} \tag{8}$$

*respectively, where $\tilde{n} = \left(\prod_{j=1}^{N} n_j\right)^{1/N}$ and $\tilde{f} = \left(\prod_{j=1}^{N} f_j\right)^{1/N}$ are the geometric means of the number of alternatives and of $f_j$, respectively.*

Here, $\tilde{f}$ can be viewed as the part of $R^2$ that measures the accuracy of probabilities derived from the model. In each case, as $\tilde{f}$ increases, so do the $R^2$s. However, the McFadden version has a predictable dependence on the number of alternatives: as $n_j$ increases, $R^2_M$ decreases in proportion to $\ln \tilde{n}$. Hence, in order to define an unbiased, rescaled McFadden pseudo-$R^2$, we multiply $R^2_M$ by $\ln \tilde{n}$, i.e.,

$$\tilde{R}^2_M = (\ln \tilde{n}) \left[ 1 - \frac{\ln L(\hat{\beta})}{\ln L(0)} \right]. \tag{9}$$

Note that this definition now has a maximum of $\ln \tilde{n}$, rather than 1. The Maddala pseudo-$R^2$ is already independent of $n_j$.

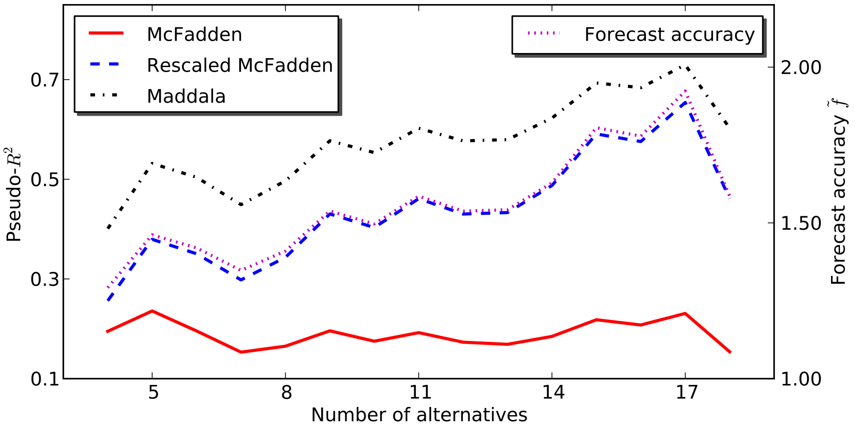
We now show empirically that the Maddala and rescaled McFadden $R^2$s are unbiased by fitting CL models on subsets of horserace betting data categorized by the number of runners in each race. The data employed are final bookmaker odds (market prices) from 6064 horserace betting markets in the UK and Ireland in 2009 and 2010 (for a description of UK betting markets, see Appendix A). These models have just one independent variable, which is the log of winning probability as implied by market prices; the coefficient of this variable is given by $\beta$. The results are presented in Table 1 and Figure 1.

Clearly, the Maddala and rescaled McFadden $R^2$s vary in a consistent manner as the number of alternatives is changed, while the standard McFadden

Table 1: Conditional logit models with log of odds-implied probability as the single variable fitted to different subsets of the data depending on the number of competitors $n$ in each event ($N$ denotes number of events).

| $n$ | $N$ | $\tilde{n}$ | $\hat{\beta}$ | $\ln L(0)$ | $\ln L(\hat{\beta})$ | $R^2_M$ | $\tilde{R}^2_M$ | $R^2_D$ | $\tilde{f}$ | ROC area | Brier |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2–4 | 201 | 3.7 | 1.10 | -264.2 | -212.7 | 0.195 | 0.256 | 0.401 | 1.29 | 0.774 | 0.158 |
| 5 | 335 | 5 | 1.24 | -539.2 | -412.0 | 0.236 | 0.379 | 0.532 | 1.46 | 0.801 | 0.127 |
| 6 | 448 | 6 | 1.19 | -802.7 | -645.7 | 0.196 | 0.351 | 0.504 | 1.42 | 0.775 | 0.119 |
| 7 | 578 | 7 | 1.05 | -1124.7 | -952.4 | 0.153 | 0.298 | 0.449 | 1.35 | 0.748 | 0.109 |
| 8 | 611 | 8 | 1.11 | -1270.5 | -1060.7 | 0.165 | 0.343 | 0.497 | 1.41 | 0.760 | 0.098 |
| 9 | 646 | 9 | 1.24 | -1419.4 | -1141.3 | 0.196 | 0.430 | 0.577 | 1.54 | 0.783 | 0.086 |
| 10 | 585 | 10 | 1.17 | -1347.0 | -1111.1 | 0.175 | 0.403 | 0.554 | 1.50 | 0.773 | 0.080 |
| 11 | 572 | 11 | 1.23 | -1371.6 | -1107.8 | 0.192 | 0.461 | 0.602 | 1.59 | 0.785 | 0.073 |
| 12 | 533 | 12 | 1.17 | -1324.5 | -1095.1 | 0.173 | 0.430 | 0.577 | 1.54 | 0.774 | 0.069 |
| 13 | 450 | 13 | 1.21 | -1154.2 | -959.3 | 0.169 | 0.433 | 0.580 | 1.54 | 0.773 | 0.064 |
| 14 | 345 | 14 | 1.24 | -910.5 | -742.3 | 0.185 | 0.488 | 0.623 | 1.63 | 0.776 | 0.059 |
| 15 | 222 | 15 | 1.46 | -601.2 | -470.0 | 0.218 | 0.591 | 0.693 | 1.81 | 0.809 | 0.055 |
| 16 | 189 | 16 | 1.38 | -524.0 | -415.2 | 0.208 | 0.576 | 0.684 | 1.79 | 0.806 | 0.053 |
| 17 | 83 | 17 | 1.72 | -235.2 | -180.9 | 0.231 | 0.654 | 0.730 | 1.92 | 0.830 | 0.050 |
| 18–40 | 266 | 20.0 | 1.29 | -797.2 | -674.5 | 0.154 | 0.461 | 0.603 | 1.59 | 0.777 | 0.044 |
| All | 6064 | 9.6 | 1.21 | -13686.1 | -11194.8 | 0.182 | 0.411 | 0.560 | 1.51 | 0.794 | 0.076 |

Figure 1: A comparison of pseudo-$R^2$s across subsets of choice data categorized by the number of alternatives in each event.



version is inconsistent. In particular, there appears to be an increasing trend in the forecast accuracy of odds-implied probabilities as $n_j$ is increased, but this trend is not captured by the standard McFadden version (note that $\tilde{f}$ maps almost perfectly onto the rescaled McFadden pseudo-$R^2$). We confirm this in two ways: first, we fit linear regressions of $R^2_D - R^2_M$ and $R^2_D - \tilde{R}^2_M$ on the number of competitors; gradients are given by 0.0174 ($t = 10.67$, $p = 0.00$) and -0.0035 ($t = 0.20$, $p = 0.42$), respectively, i.e., the difference between the Maddala and McFadden $R^2$s increases with the number of alternatives while the difference between the Maddala and rescaled McFadden $R^2$s does not. Second, we compare the trends in pseudo-$R^2$s with two alternative measures of forecasting accuracy, used by Franck et al. (2010), the ROC area and the Brier score. We observed a monotonic relationship between pseudo-$R^2$ and these information criteria, with the ROC area increasing with number of alternatives and the Brier score decreasing (a small Brier score indicates high forecast accuracy). The result here is that, had we interpreted the results only with reference to the McFadden $R^2$, we might have concluded that forecast accuracy does not change with number of alternatives, which is clearly incorrect. In summary, forecasters must take great care when comparing pseudo-$R^2$s between models fitted on different data.

In particular, the specific definition of pseudo-$R^2$ employed must be appropriate. A similar analysis could be carried out to test the consistency of other definitions.

## 3. Comparing probabilistic forecasts from different models: distributional properties of pseudo-$R^2$s

### *3.1. Bootstrapping pseudo-$R^2$s*

A common problem in forecasting is how to compare models that either (a) consist of different predictor variables (are non-nested), or (b) are evaluated on different datasets, or a combination of both. It is straightforward to compare forecasts from nested models (where one of the models includes all the independent variables from the other model) evaluated on the same data (with a likelihood ratio test, for example). However, it is significantly more difficult to compare the accuracy of non-nested models or out-of-sample forecasts evaluated on different data. Such comparisons are important because it is often the case in probability forecasting that one is trying to select the 'best' model from a number of candidate models (e.g., models incorporating different independent variables). One possibility is to directly compare measures of forecast accuracy, such as pseudo-$R^2$s or the ROC area or Brier score mentioned above. Without standard error estimates, a model may be selected over another simply because its pseudo-$R^2$ is apparently higher. However, the model with the higher pseudo-$R^2$ could have a higher standard error (perhaps because it was estimated on a smaller dataset) and statistical tests may indicate that it is not in fact superior. This is particulary important in discrete choice modelling, where the inappropriate selection of one model over another (perhaps containing different predictor variables) may lead the researcher to mis-specify the behavioral factors influencing the consumers' choices. Clearly, a statistical test for differences in the pseudo-$R^2$ values requires their distributions, which are complex and depend on unknown parameters. However, it is possible to carry out significance tests by estimating the distribution of pseudo-$R^2$s, and we demonstrate two methods here. The first is to adopt an $M$-bootstrap approach (Efron, 1979), as recommended by Ohtani (2000) for ordinary $R^2$s. The bootstrap is commonly used when the theoretical distribution of a statistic is complicated, which is the case for the CL model. Suppose we have fitted CL models to two datasets, $D_1$ and $D_2$, consisting of $N_1$ and $N_2$ events, respectively. The $M$-bootstrap method proceeds as follows:

1. Randomly sample $N_1$ events, with replacement, from $D_1$, to form a new dataset $BD_1$. Similarly, randomly sample $N_2$ events, with replacement, from $D_2$, to form a new dataset $BD_2$.
2. Fit CL models on $BD_1$ and $BD_2$ and record the resulting values of pseudo-$R^2$.
3. Perform $M$ iterations of steps 1 and 2.
4. The sample means, $\mu(R_1^2)$ and $\mu(R_2^2)$, and sample variances, $s^2(R_1^2)$ and $s^2(R_2^2)$, of the sets of $M$ pseudo-$R^2$s are used to derive a standard normal test statistic,

$$z[\mu(R^2)] = \left[\mu(R_1^2) - \mu(R_2^2)\right] / \sqrt{s^2(R_1^2) + s^2(R_2^2)}, \tag{10}$$

which can be used to test the alternative hypothesis that the probabilities derived from one model are more accurate than those derived from the other, against the null hypothesis of no difference.

### *3.2. The asymptotic distribution of pseudo-$R^2$s*

An alternative, much faster, method for estimating the distribution of a pseudo-$R^2$ is to estimate its *asymptotic distribution*, i.e., the expected distribution as the number of events tends to infinity. Hu et al. (2006) derive analytically the asymptotic distribution of the Maddala pseudo-$R^2$ in the multinomial logit model (a discrete choice model that is similar to the CL model). Here, we adapt their analysis to derive the asymptotic distribution of the McFadden pseudo-$R^2$, our own rescaled McFadden pseudo-$R^2$ specified above, and the Maddala pseudo-$R^2$ for the CL model (for an outline proof, see Appendix B).

**Proposition 2.** *Assume that the independent variables $x_{ij}$, $j = 1, 2, \ldots, N$, $i = 1, 2, \ldots, n_j$ are independent and identically distributed random $m$-vectors with finite second moment (i.e., $E[x_{ij}^2]$ finite). Let*

$$\begin{aligned} H_1 &= E[\ln n_j], \\ H_2 &= -E\left[\sum_{i=1}^{n_j} y_{ij} \ln p_{ij}\right], \end{aligned} \tag{11}$$

*and let*

$$\begin{aligned} \Sigma &= \begin{pmatrix} Var(n_j) & \eta \\ \eta & \epsilon \end{pmatrix}, & g_1 &= \frac{1}{\ln\lambda}\begin{pmatrix} \frac{H_2}{\lambda\ln\lambda} \\ 1 \end{pmatrix}, \\ g_2 &= \begin{pmatrix} \frac{1}{\lambda} \\ 1 \end{pmatrix}, & g_3 &= \frac{2e^{2H_2}}{\lambda^2}\begin{pmatrix} \frac{1}{\lambda} \\ 1 \end{pmatrix}, \end{aligned} \tag{12}$$

*where*

$$\begin{aligned} \lambda &= E[n_j], \\ \eta &= E\left[n_j \sum_{i=1}^{n_j} y_{ij} \ln p_{ij}\right] + \lambda H_2, \\ \epsilon &= E\left[\sum_{i=1}^{n_j} y_{ij} (\ln p_{ij})^2\right] - H_2^2. \end{aligned} \tag{13}$$

*Then, as $N \to \infty$,*

$$\begin{aligned} &\sqrt{N}\left[R_M^2 - (1 - H_2/H_1)\right] \to_d N(0, \sigma_1^2), \\ &\sqrt{N}\left[\tilde{R}_M^2 - (H_1 - H_2)\right] \to_d N(0, \sigma_2^2), \\ &\sqrt{N}\left[R_D^2 - (1 - e^{2(H_2 - H_1)}\right] \to_d N(0, \sigma_3^2), \end{aligned} \tag{14}$$

*where $\sigma_i^2 = g_i^\top \Sigma g_i$ for $i = 1, 2, 3$.*

The above proposition gives the asymptotic distribution of the pseudo-$R^2$s. Hence, to obtain the estimates of the variance of point estimates of these pseudo-$R^2$s, we can replace the unknown quantities with consistent estimators. So, denote by $\bar{n}$, $\tilde{n}$, and $s^2(n)$ the arithmetic mean, geometric mean, and sample variance of the number of alternatives, respectively, i.e.,

$$
\begin{aligned}
\bar{n} &= (1/N)\sum_{j=1}^{N} n_j, \\
\tilde{n} &= \left(\prod_{j=1}^{N} n_j\right)^{1/N}, \\
s^2(n) &= \frac{1}{N-1}\sum_{j=1}^{N}(n_j - \bar{n})^2.
\end{aligned}
\tag{15}
$$

Then, let

$$
\begin{aligned}
\hat{\Sigma} &= \begin{pmatrix} s^2(n) & \hat{\eta} \\ \hat{\eta} & \hat{\epsilon} \end{pmatrix}, & \hat{g}_1 &= \frac{1}{\ln \bar{n}} \begin{pmatrix} \frac{\hat{H}_2}{\bar{n}\ln\bar{n}} \\ 1 \end{pmatrix}, \\
\hat{g}_2 &= \begin{pmatrix} \frac{1}{\bar{n}} \\ 1 \end{pmatrix}, & \hat{g}_3 &= \frac{2e^{2\hat{H}_2}}{\bar{n}^2} \begin{pmatrix} \frac{1}{\bar{n}} \\ 1 \end{pmatrix},
\end{aligned}
\tag{16}
$$

where

$$
\begin{aligned}
\hat{\eta} &= (1/N)\sum_{j=1}^{N} n_j \sum_{i=1}^{n_j} y_{ij}\ln p_{ij} + \bar{n}\hat{H}_2, \\
\hat{\epsilon} &= (1/N)\sum_{j=1}^{N} n_j \sum_{i=1}^{n_j} y_{ij}(\ln p_{ij})^2 - \hat{H}_2^2.
\end{aligned}
\tag{17}
$$

Here

$$
\hat{H}_2 = -(1/N)\sum_{j=1}^{N}\sum_{i=1}^{n_j} y_{ij}\ln p_{ij}. \tag{18}
$$

Then estimates of the variance of the McFadden, rescaled McFadden, and Maddala pseudo-$R^2$s are given by

$$
\begin{aligned}
s^2(R_M^2) &= \frac{1}{N}\left(\hat{g}_1^\top \hat{\Sigma}\hat{g}_1\right), \\
s^2(\tilde{R}_M^2) &= \frac{1}{N}\left(\hat{g}_2^\top \hat{\Sigma}\hat{g}_2\right), \\
s^2(R_D^2) &= \frac{1}{N}\left(\hat{g}_3^\top \hat{\Sigma}\hat{g}_3\right),
\end{aligned}
\tag{19}
$$

respectively.

### *3.3. An empirical comparison of the asymptotic and bootstrap methods*

We now verify the two methods by estimating variances of the three pseudo-$R^2$s described above. It has been shown that the bootstrap method overestimates standard errors in large samples for standard logistic regression, relative to the asymptotic distribution (Teebagy & Chatterjee, 1989). Here, we compare values estimated from the bootstrap and asymptotic distribution methods on the same real data, which are described in section 2. This time odds from two different markets (exchange and bookmaker) are used. We again fit CL models with just the log of odds-implied probability as the single explanatory variable; the coefficient of this variable is $\beta$. From the results presented in Table 2, it is clear that both methods produce reasonably similar estimates, with the differences not being significant according to $F$ tests for difference of variances. Since the asymptotic method is very fast to calculate, we would therefore recommend that pseudo-$R^2$s are always reported with a measure of dispersion, so that comparing two non-nested models can always be carried out with a significance test.

## 4. Pseudo-$R^2$ as a predictor of the economic value of a discrete choice model

### *4.1. Probabilistic forecasting of competitive events*

We now turn to a specific application of discrete choice models: probabilistic forecasting of CEs for identifying and measuring the degree of inefficiency in betting markets. CEs, such as horseraces, usually have an associated market for trading on the outcome. From the prices available in the market (the odds), it is possible to obtain 'public' forecasts of the probabilities of each outcome. If these probabilistic forecasts are inaccurate, then the market is inefficient. Skilled forecasters are able to exploit this inefficiency for profit (e.g., Benter, 1994). We argue that, in this context, the primary interpretation of pseudo-$R^2$s, as improvement from the null to fitted model, is misleading. This is because we can easily specify a 'public' model that has a single variable, which is the log

Table 2: A comparison of the asymptotic and bootstrap methods for estimating the distributions of the McFadden, rescaled McFadden, and Maddala pseudo-$R^2$s, with $F$-tests for difference of variances.

| Market type | Exchange | Bookmaker |
|---|---|---|
| Number of events | 6064 | |
| Total number of competitors | 62224 | |
| Mean number of competitors | 10.3 | |
| $\ln L(0)$ | -13686.0 | |
| $\hat{\beta}$ | 1.06 | 1.21 |
| $\ln L(\hat{\beta})$ | -11137.0 | -11195.0 |
| Asymptotic | | |
| $R^2_M$ | 0.186 | 0.182 |
| $S.E.(R^2_M)$ | 0.00446 | 0.00448 |
| $\tilde{R}^2_M$ | 0.420 | 0.411 |
| $S.E.(\tilde{R}^2_M)$ | 0.0104 | 0.0105 |
| $R^2_D$ | 0.569 | 0.560 |
| $S.E.(R^2_D)$ | 0.0090 | 0.0092 |
| Bootstrap | | |
| $R^2_M$ | 0.186 | 0.182 |
| $S.E.(R^2_M)$ | 0.00455 | 0.00455 |
| $\tilde{R}^2_M$ | 0.421 | 0.411 |
| $S.E.(\tilde{R}^2_M)$ | 0.0103 | 0.0103 |
| $R^2_D$ | 0.569 | 0.560 |
| $S.E.(R^2_D)$ | 0.0089 | 0.0090 |
| $F_{6063,6063}(R^2_M)$ | 1.041 | 1.033 |
| $F_{6063,6063}(\tilde{R}^2_M)$ | 1.028 | 1.035 |
| $F_{6063,6063}(R^2_D)$ | 1.031 | 1.037 |

of odds-implied probabilities (the model we used in sections 2 and 3), and this model supersedes the null model that has no variables at all. Rather, pseudo-

$R^2$ in this context should measure the improvement of the relevant model over the public, where the model contains both the log of odds-implied probabilities together with variables that it is believed are not fully discounted by the public. In this section we define such a measure, which we call *relative* pseudo-$R^2$, and link it directly to the model's 'edge', i.e., the economic significance of the inefficiency identified by the model.

Recall that our dataset consists of $N$ events, where each event $j$ is between $n_j \geq 2$ competitors; for each event, there is one winner, given by $y_{ij} = 1$, with $y_{ij} = 0$ otherwise. 'Decimal odds' are denoted by $D_{ij} > 1$, with $d_{ij} = 1/D_{ij}$ denoting market prices. The decimal odds represent the potential return to a bettor from a bet on competitor $i$ in race $j$, with a winning bet of \$1 returning $D_{ij}$ if the bet wins. If the bettor assigns winning probabilities $p_{ij}$, the expected profit from a \$1 bet is $p_{ij}D_{ij} - 1$. Denote the winning probability that the bettor assigns to the eventual winner by $p_j$, its market price $d_j$, its decimal odds $D_j$, and the winning probability as implied by the odds $q_j = d_{ij}/(1 + B_j)$, where the 'over-round' $B_j = \sum_{i=1}^{n_j} d_{ij} - 1$ represents the market's transaction costs. Then expected profit from a bet on this competitor, or 'edge', is given by

$$W_j = \frac{p_j}{q_j(1 + B_j)} - 1. \tag{20}$$

Now, we define the relative McFadden and Maddala pseudo-$R^2$s by

$$\bar{R}^2_M = 1 - \frac{\ln L(p)}{\ln L(q)}, \tag{21}$$

$$\bar{R}^2_D = 1 - \exp\{-(2/N)\,[\ln L(p) - \ln L(q)]\},$$

where the log-likelihood of the bettor's and the public models are given by

$$\ln L(p) = \sum_{j=1}^{N} \sum_{i=1}^{n_j} y_{ij} \ln p_{ij}, \tag{22}$$

$$\ln L(d) = \sum_{j=1}^{N} \sum_{i=1}^{n_j} y_{ij} \ln q_{ij},$$

respectively. Defined in this way, the relative pseudo-$R^2$s measure the degree of improvement over the public odds for the winning competitor, which we now show is directly related to the bettor's realized edge (profit per \$1 bet).

Substituting (20) into (21), we can write the relative McFadden pseudo-$R^2$ as

$$\bar{R}_M^2 = \frac{\ln GM(1+W_j) + \ln GM(1+B_j)}{\ln GM(D_j) + \ln GM(1+B_j)}, \tag{23}$$

where $GM(x_j) = \left(\prod_{j=1}^N x_j\right)^{1/N}$ denotes geometric mean. As in section 2, this has a dependence on the data: in this case, the average over-round and average odds of the winner. So, we define the relative rescaled McFadden pseudo-$R^2$ by

$$\bar{\bar{R}}_M^2 = (\ln GM(D_j) + \ln GM(1+B_j))\bar{R}_M^2 = \ln GM(p_j/q_j). \tag{24}$$

Then this measure has a log-proportional relationship with average edge. Rearranging,

$$GM(p_j/q_j) = e^{\bar{\bar{R}}_M^2}. \tag{25}$$

Similar relationships for the Maddala pseudo-$R^2$ are

$$\bar{R}_D^2 = 1 - \frac{1}{GM(1+W_j)^2 GM(1+B_j)^2}, \tag{26}$$

$$GM(p_j/q_j) = \frac{1}{\sqrt{1-\bar{R}_D^2}}.$$

These relative pseudo-$R^2$s are a novel tool for evaluating and comparing probabilistic forecasts of competitive events, since they allow us to quantify the relative gain of the out-of-sample probability forecasts of a model over those of the relevant benchmark, which is the public model and not the irrelevant null model.

### *4.2. Relative pseudo-$R^2$s in betting markets*

To illustrate the usefulness of relative pseudo-$R^2$s, we conduct analysis using real betting market data. We train a number of CL models on a dataset consisting of 18040 horse races from the years 2007-2009, and the results are presented in Table 3. Each model includes a transformation of the public odds (LN_FIXED_ODDS_PROB), which itself adds some predictive power over and above the raw odds-implied probability, along with a successively increasing set of independent variables based on fundamental information pertaining to the horse's winning chances, such as its lengths beaten in past races, speed in past races, ability of the jockey/trainer and weight carried. With model coefficients $\beta_k$ fixed from the in-sample data, we then evaluate the probability forecasts derived from each model on a holdout set of 6309 races from 2010, and calculate relative pseudo-$R^2$s over this data (Table 4). To illustrate how relative pseudo-$R^2$s derived from a model are related to its economic importance, we simulate a Kelly betting strategy (Kelly, 1956) on the outcomes of each race in the holdout set using the actual market prices available. A proportion of capital equal to $\max((p_{ij}D_{ij}-1)/(D_{ij}-1),0)$ is bet on each horse, which is the amount that maximizes the log of expected returns from the bet (it is assumed that the bookmaker will take bets of any size).

The first column in Table 4 is based on the $q_{ij}$ directly, without any model. Hence, the relative pseudo-$R^2$s are 0 and no betting opportunities are found. Each of the other models are effectively being compared with this one. The relative pseudo-$R^2$s here are small, so we present them as percentages, but they translate into increasingly large expected and actual profits, which are determined from the $p_{ij}$ and from the identity of the winning horse, respectively (odds-implied number of wins are determined assuming win probabilities are $q_{ij}$). Note that while the expected profits increase consistently as the number of predictor variables is increased, actual profits are subject to the natural variance in the data, and therefore do not necessarily increase monotonically. However, because relative pseudo-$R^2$s increase consistently with expected profits and have a natural interpretation of 0 being no expected profits, it is clear that the relative pseudo-$R^2$s can be used as an indicator of the economic value of the model.

In this section we have demonstrated that direct relationships can be derived between forecast probabilities estimated by CL models containing multiple predictor variables, and relative pseudo-$R^2$ measures. These relationships are crucial because they represent the economic value of forecasting models and can be used to assess the economic significance of any market efficiency that the models unearth. They also contribute to an understanding of the context in which pseudo-$R^2$s should be reported. In a broader range of problems, the

Table 3: Conditional logit models with additional predictor variables, training set of 184048 competitors in 18040 events, from the years 2007-2009 ($\ln L(0)$ = -40904, * indicates significance at the 1% level).

| variable/coefficient | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\beta}_5$ | $\hat{\beta}_6$ | $\hat{\beta}_7$ |
|---|---|---|---|---|---|---|---|---|
| LN_FIXED_ODDS_PROB | 0.99* | 0.98* | 0.96* | 0.96* | 0.93* | 0.92* | 0.91* | 0.91* |
|  | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| LENGTHS_BEATEN |  | 0.01 | 0.01 | 0.01 | 0.02* | 0.02* | 0.02* | 0.03* |
|  |  | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| SPEED_RATING |  |  | 0.84* | 0.86* | 0.90* | 0.96* | 1.00* | 1.04* |
|  |  |  | (0.15) | (0.15) | (0.15) | (0.15) | (0.15) | (0.15) |
| PREFERENCES |  |  |  | 0.42* | 0.44* | 0.46* | 0.47* | 0.48* |
|  |  |  |  | (0.10) | (0.10) | (0.10) | (0.10) | (0.10) |
| JOCKEY |  |  |  |  | 0.92* | 0.78* | 0.84* | 0.93* |
|  |  |  |  |  | (0.20) | (0.21) | (0.21) | (0.21) |
| TRAINER |  |  |  |  |  | 0.71* | 0.77* | 0.80* |
|  |  |  |  |  |  | (0.19) | (0.19) | (0.19) |
| DRAW_BIAS |  |  |  |  |  |  | 2.88* | 2.88* |
|  |  |  |  |  |  |  | (0.37) | (0.37) |
| WEIGHT_CARRIED |  |  |  |  |  |  |  | -0.01* |
|  |  |  |  |  |  |  |  | (0.00) |
| $\ln L(\hat{\beta})$ | -34230 | -34230 | -34215 | -34205 | -34195 | -34187 | -34157 | -34152 |
| $R_M^2$ | 0.1632 | 0.1632 | 0.1635 | 0.1638 | 0.1640 | 0.1642 | 0.1649 | 0.1651 |
| $\tilde{R}_M^2$ | 0.3700 | 0.3700 | 0.3708 | 0.3714 | 0.3719 | 0.3723 | 0.3740 | 0.3743 |
| $R_D^2$ | 0.5228 | 0.5229 | 0.5237 | 0.5242 | 0.5247 | 0.5251 | 0.5267 | 0.5269 |

Table 4: Relative pseudo-$R^2$s and betting simulation results, holdout set of 60815 competitors in 6309 events, from the year 2010 ($\ln L(0)$ = -13931, initial capital = \$1000)

| model | public | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| $\ln L$ | -11695 | -11667 | -11667 | -11656 | -11655 | -11654 | -11652 | -11650 | -11646 |
| $R^2_M$ | 0.1606 | 0.1625 | 0.1626 | 0.1633 | 0.1634 | 0.1635 | 0.1636 | 0.1637 | 0.1640 |
| $\tilde{R}^2_M$ | 0.3545 | 0.3589 | 0.3589 | 0.3606 | 0.3609 | 0.3610 | 0.3613 | 0.3616 | 0.3622 |
| $R^2_D$ | 0.5079 | 0.5122 | 0.5122 | 0.5138 | 0.5141 | 0.5142 | 0.5145 | 0.5148 | 0.5154 |
| $\bar{\tilde{R}}^2_M$ (%) | 0.00 | 0.44 | 0.44 | 0.61 | 0.64 | 0.65 | 0.68 | 0.70 | 0.77 |
| $\bar{R}^2_D$ (%) | 0.00 | 0.87 | 0.88 | 1.20 | 1.26 | 1.29 | 1.34 | 1.40 | 1.52 |
| $GM(p_j/q_j)$ | 1.0000 | 1.0044 | 1.0044 | 1.0061 | 1.0064 | 1.0065 | 1.0068 | 1.0071 | 1.0077 |
| Total bet (\$) | 0.00 | 2265.01 | 2423.60 | 10567.38 | 13920.45 | 20156.71 | 20598.82 | 43981.29 | 47380.08 |
| no. events bet on | 0 | 97 | 113 | 540 | 803 | 1148 | 1328 | 2193 | 2294 |
| no. bets | 0 | 127 | 143 | 593 | 861 | 1235 | 1430 | 2476 | 2606 |
| Odds-implied no. wins | 0 | 46.7 | 52.7 | 205.4 | 276.0 | 367.6 | 407.9 | 599.1 | 630.0 |
| $E$(no. wins) | 0 | 50.5 | 57.2 | 233.4 | 314.7 | 421.9 | 469.5 | 703.2 | 740.6 |
| no. wins | 0 | 54 | 61 | 224 | 318 | 412 | 469 | 704 | 727 |
| $E$(Profit) (\$) | 0.00 | 101.61 | 106.91 | 451.19 | 623.08 | 996.00 | 1082.48 | 3210.24 | 3542.52 |
| Profit (\$) | 0.00 | 271.45 | 294.81 | 1260.52 | 1405.56 | 1454.94 | 598.49 | 2114.31 | 3482.54 |

.

context-specific public model might be replaced by other types of ‘base’ model that are not simply the $1/n$ null model. For example, an organization, wishing to maximize their subjective expected utility from their decision making, might already be in possession of hard data on some possible courses of action, such as their current practices. In this case, it might be more appropriate to compare their forecasts with their existing data, in which case it is the relative expected gain that is important; in a choice model context this could be measured by relative pseudo-$R^2$.

## 5. Conclusion

In this paper, we have described and evaluated properties of pseudo-$R^2$s as a measure of accuracy of probabilistic forecasts from discrete choice models. These are a class of models primarily employed to predict choices made by individuals, and thus have applications in marketing, public choice (e.g., healthcare provision), econometrics, operations research, and other areas. While $R^2$ in ordinary least squares linear regression is a widely-used and well-justified measure, the same is typically not true of pseudo-$R^2$s, but these are nevertheless invaluable measures of forecast accuracy, particularly in the context of competitive events. We have shown both theoretically and empirically that at least one of the definitions of pseudo-$R^2$ (McFadden’s definition) is not robust to variations in the number of alternatives in each event. We have therefore suggested a rescaling to correct for the resulting bias. This has important implications for the comparability of pseudo-$R^2$ measures across models, particularly non-nested models or models tested on different datasets, and comparability is a key desirable property of any forecast evaluator. We have also described two methods for estimating the variance of pseudo-$R^2$s so that their values can be statistically compared: the bootstrap and asymptotic methods. A comparison of the two methods on real-world data demonstrates that the estimates that they produce are reasonably close, so we would recommend that the faster asymptotic method is used. Finally, in a specific application to the forecasting of competitive events, we have derived simple relationships between relative pseudo-$R^2$ measures and the expected profit to a trader from betting on competitive events. This relationship is crucial because choice modelling is often employed in the context of competitive events in order to assess market efficiency, an important concern with implications for the health of markets as allocative mechanisms.

Our findings contribute to an understanding of the use of pseudo-$R^2$s, and our results suggest that the common practice of simply reporting these without a justification or without standard errors when comparing them across models is not advisable. Moreover, the methods described in this paper (rescaled McFadden pseudo-$R^2$, bootstrap and asymptotic methods for standard errors, and relative pseudo-$R^2$) are crucial when the pseudo-$R^2$ itself is the value of interest in hypothesis testing; for instance, in comparing the out-of-sample predictive power of discrete choice models, or evaluating the efficiency of speculative financial markets. Accurate probabilistic forecasting is a desirable goal in any context, but it is particularly important to be able to assess that accuracy in a consistent and intuitive manner. This paper helps achieve this consistency by providing a more rigorous understanding of the value of pseudo-$R^2$s in evaluating probabilistic forecasts from discrete choice models.

## Appendix A. A description of UK betting markets

The two major types of betting market in the UK are bookmakers and exchanges, together accounting for 94% of horserace betting turnover (over £5.7 billion) in the year to March 2010 (Gambling Commission, 2010). In bookmaker markets, bettors place bets at fixed odds set by the bookmaker. The bettor must accept the odds currently offered by the bookmaker or the unknown starting price (the odds available at the close of the market). Bookmakers have the highest operating costs (e.g., maintaining an estate of betting offices) so their margins are typically higher than exchange markets. Bets can be placed at racetracks (on-course market) as well at betting offices around the UK or online (off-course market). Prices are distinct in the on-course and off-course markets until 10 minutes before the race starts, at which point the two markets converge.

A betting exchange is an online platform that allows bettors to back horses to win or to lay them to lose. Bets are only matched when two or more bets of the appropriate stake and odds are made, with the exchange automatically pairing backers and layers to settle bets. Exchanges typically have lower margins than bookmakers. For more information on exchanges, see Smith & Vaughan Williams (2008).

## Appendix B. Proofs

**Proof of Proposition 1.** For the McFadden pseudo-$R^2$,

$$\begin{aligned} R_M^2 &= 1 - \frac{\sum_{j=1}^N \ln\left[f(n_j)/n_j\right]}{\sum_{j=1}^N \ln(1/n_j)} \\ &= \frac{\sum_{j=1}^N \ln f(n_j)}{\sum_{j=1}^N \ln n_j} \\ &= \frac{\ln \prod_{j=1}^N f(n_j)}{\ln \prod_{j=1}^N n_j} \\ &= \frac{\ln \tilde{f}}{\ln \tilde{n}}. \end{aligned} \tag{B.1}$$

For the Maddala pseudo-$R^2$,

$$\begin{aligned} R_D^2 &= 1 - \exp\left\{-(2/N)\left[\sum_{j=1}^N \ln\{f(n_j)/n_j\} - \sum_{j=1}^N \ln(1/n_j)\right]\right\} \\ &= 1 - \exp\left\{-(2/N)\ln \prod_{j=1}^N f(n_j)\right\} \\ &= 1 - \left[\prod_{j=1}^N f(n_j)\right]^{-2/N} \\ &= 1 - 1/\tilde{f}^2. \end{aligned} \tag{B.2}$$

**Lemma.** *Assume that the independent variables $x_{ij}$, $j = 1, 2, \ldots, N$, $i = 1, 2, \ldots, n_j$, are independent and identically distributed random $m$-vectors with* *finite second moment (i.e., $E(x_{ij}^2)$ finite). Let*

$$\begin{aligned} H_1 &= E[\ln n_j], \\ H_2 &= -E\left[\sum_{i=1}^{n_j} y_{ij} \ln p_{ij}\right]. \end{aligned} \tag{B.3}$$

*Then, as $N \to \infty$, $R_M^2 \to_p 1 - H_2/H_1$, $\tilde{R}_M^2 \to_p H_1 - H_2$, and $R_D^2 \to_p 1 -$* $\exp\left[2(H_2 - H_1)\right]$.

*Proof.*

$$\begin{aligned} R_M^2 &= 1 - \frac{(1/N)\ln L(\hat{\beta})}{(1/N)\ln L(0)} \\ &= 1 - \frac{(1/N)\ln L(\beta) - (1/N)\left[\ln L(\beta) - \ln L(\hat{\beta})\right]}{(1/N)\sum_{j=1}^{N}\ln(1/n_j)} \\ &= 1 - \frac{(1/N)\left[\ln L(\beta) - \ln L(\hat{\beta})\right]}{(1/N)\sum_{j=1}^{N}\ln n_j} \\ &\quad - \frac{(1/N)\sum_{j=1}^{N}\sum_{i=1}^{n_j} y_{ij}\ln p_{ij}}{(1/N)\sum_{j=1}^{N}\ln n_j}. \end{aligned} \tag{B.4}$$

Similarly,

$$\begin{aligned} \tilde{R}_M^2 = (1/N)\sum_{j=1}^{N}\ln n_j + (1/N)\sum_{j=1}^{N}\sum_{i=1}^{n_j} y_{ij}\ln p_{ij} \\ - (1/N)\left[\ln L(\beta) - \ln L(\hat{\beta}\right]. \end{aligned} \tag{B.5}$$

Finally, letting $f(x) = (1/2)\ln(1-x)$,

$$\begin{aligned} f(R_D^2) &= (1/N)\ln L(0) - (1/N)\ln L(\hat{\beta}) \\ &= (1/N)\sum_{j=1}^{N}\ln(1/n_j) - (1/N)\ln L(\beta) \\ &\quad + (1/N)\left[\ln L(\beta) - \ln L(\hat{\beta}\right] \\ &= -(1/N)\sum_{j=1}^{N}\ln n_j - (1/N)\sum_{j=1}^{N}\sum_{i=1}^{n_j} y_{ij}\ln p_{ij} \\ &\quad + (1/N)\left[\ln L(\beta) - \ln L(\hat{\beta})\right]. \end{aligned} \tag{B.6}$$

By the law of large numbers, as $N \to \infty$,

$$(1/N) \sum_{j=1}^{N} \ln n_j \to_p H_1, \tag{B.7}$$

$$-(1/N) \sum_{j=1}^{N} \sum_{i=1}^{n_j} y_{ij} \ln p_{ij} \to_p H_2.$$

Moreover, Hu et al. (2006) show that, as $N \to \infty$,

$$(1/N) \left[ \ln L(\beta) - \ln L(\hat{\beta}) \right] \to_p 0. \tag{B.8}$$

Hence, as $N \to \infty$, $R^2_M \to_p 1 - H_2/H_1$, $\tilde{R}^2_M \to_p H_1 - H_2$, and $f(R^2_D) \to_p H_2 - H_1$. So, by the continuous mapping theorem, as $N \to \infty$, $R^2_D \to_p f^{-1}(H_2 - H_1)$, i.e., $R^2_D \to_p 1 - \exp[2(H_2 - H_1)]$. □

**Proof of Proposition 2.** Define $Z_j = (n_j, W_j)$, where

$$W_j = \sum_{i=1}^{n_j} y_{ij} \ln p_{ij}. \tag{B.9}$$

Then the $Z_j$ form an independent and identically distributed random sequence with

$$\mu = E[Z_j] = (\bar{n}, -H_2) \tag{B.10}$$

and $Cov(Z_j) = \Sigma$. To see this, note that it follows immediately that

$$\begin{aligned} Cov(n_j, n_j) &= Var(n_j), \\ Cov(n_j, W_j) &= \eta, \end{aligned} \tag{B.11}$$

and

$$\begin{aligned} Cov(W_j, W_j) &= E[W_j^2] - E[W_j]^2 \\ &= E\left[ \left( \sum_{i=1}^{n_j} y_{ij} \ln p_{ij} \right)^2 \right] - H_2^2 \\ &= E\left[ \sum_{i=1}^{n_j} y_{ij} (\ln p_{ij})^2 \right] - H_2^2. \end{aligned} \tag{B.12}$$

By the central limit theorem (in two dimensions),

$$\sqrt{N}(\bar{Z} - \mu) \to N(0, \Sigma). \tag{B.13}$$

Let

$$\begin{aligned} \phi_1(x_1, x_2) &= 1 + \frac{x_2}{\ln x_1}, \\ \phi_2(x_1, x_2) &= \ln x_1 + x_2, \\ \phi_3(x_1, x_2) &= 1 - \exp\left[2(-\ln x_1 - x_2)\right]. \end{aligned} \tag{B.14}$$

Applying the delta method with $\phi_i$, $i = 1, 2, 3$, to (B.13) gives

$$\sqrt{N}\left[\phi_i(\bar{Z}) - \phi_i(\mu)\right] \to_d N\left(0, \nabla\phi_i(\mu)^\top \Sigma \nabla\phi_i(\mu)\right). \tag{B.15}$$

From the Lemma, and since

$$\begin{aligned} \nabla\phi_1(x_1, x_2) &= \begin{pmatrix} -\frac{x_2}{x_1(\ln x_1)^2} \\ \frac{1}{\ln x_1} \end{pmatrix}, \\ \nabla\phi_2(x_1, x_2) &= \begin{pmatrix} \frac{1}{x_1} \\ 1 \end{pmatrix}, \\ \nabla\phi_3(x_1, x_2) &= \begin{pmatrix} \frac{2}{x_1} e^{2(-\ln x_1 - x_2)} \\ 2e^{2(-\ln x_1 - x_2)} \end{pmatrix}, \end{aligned} \tag{B.16}$$

the results in (14) follow.

## References

Abe, M. (1999). A generalized additive model for discrete-choice data. *Journal of Business and Economic Statistics, 17*(2), 271-284.

Armstrong, J. S. (2001). Evaluating forecasting methods. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 443-472). Norwell, MA: Kluwer Academic Press.

Benter, W. (1994). Computer based horse race handicapping and wagering systems: a report. In D. B. Hausch, V. S. Y. Lo, & W. T. Ziemba (Eds.), *Efficiency of Racetrack Betting Markets* (pp. 183-198). London: Academic Press.

Breslow, N. E., & Day, N. E. (1994). *Statistical methods in cancer research. Volume I - the analysis of case-control studies.* Lyon: International Agency for Research on Cancer.

Cheng, S., & Stough, R. S. (2006). Location decisions of Japanese new manufacturing plants in China: a discrete-choice analysis. *The Annals of Regional Science, 40*(2), 369-387.

Cowgill, B., Wolfers, J., & Zitzewitz, E. (2009). Using prediction markets to track information flows: evidence from Google. *Mimeo.*

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis.* New York: Wiley.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26.

Franck, E., Verbeek, E., & Nüesch, S. (2010). Prediction accuracy of different market structures - bookmakers versus a betting exchange. *International Journal of Forecasting, 26*, 448-459.

Gambling Commission. (2010). Industry Statistics 2009/10. http://www.gamblingcommission.gov.uk/pdf/Gambing Industry Statistics 2009 2010 WEB - January 2011.pdf (last accessed: 15 February 2011).

Greene, W. H. (2012). *Econometric Analysis.* Upper Saddle River, NJ: Prentice Hall.

Hu, B., Shao, J., & Palta, M. (2006). Pseudo-$R^2$ in logistic regression model. *Statistica Sinica, 16*, 847-860.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*, 679-688.

Johnstone, D. J. (2011). Economic interpretation of probabilities estimated by maximum likelihood or score. *Management Science, 57*(2), 308-314.

Judge, G. G., Garrett, G., & Griffiths, W. E. (1985). *The Theory and Practice of Econometrics*. New York: Wiley.

Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal, 35*(4), 917-926.

Kvålseth, T. O. (1985). Cautionary note about $R^2$. *The American Statistician, 39*(4), 279-285.

Lessmann, S., Sung, M., & Johnson, J. E. V. (2009). Identifying winners of competitive events: a SVM-based classification model for horserace prediction. *European Journal of Operational Research, 196*(2), 569-577.

Lessmann, S., Sung, M., Johnson, J. E. V., & Ma, T. (2012). A new methodology for generating and combining statistical forecasting models to enhance competitive event prediction. *European Journal of Operational Research, 218*(1), 163-174.

Lin, K. Y., & Sibdari, S. Y. (2009). Dynamic price competition with discrete customer choices. *European Journal of Operational Research, 197*(3), 969-980.

Liu, Y-H. (2011). Incorporating scatter search and threshold accepting in finding maximum likelihood estimates for the multionomial probit model. *European Journal of Operational Research, 211*(1), 130-138.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105-142). New York: Academic Press.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17-24.

Nagelkerke, N. J. D. (1991). A note on general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692.

Ohtani, K. (2000). Bootstrapping $R^2$ and adjusted $R^2$ in regression analysis. *Economic Modelling, 17*(4), 473-483.

Plott, C. R., & Chen, K-Y. (2002). Information aggregation mechanisms: concept, design and implementation for a sales forecasting problem. *CalTech, Division of the Humanities and Social Sciences, Working Paper 1131.*

Press, S. J., & Zellner, A. (1978). Posterior distribution for the multiple correlation coefficient with fixed regressors. *Journal of Econometrics, 8*(3), 307-321.

Schnytzer, A., Lamers, M., & Makropoulou, V. (2010). The impact of insider trading on forecasting in a bookmakers' horse betting market. *International Journal of Forecasting, 26*(2), 537-542.

Smith, M. A., & Vaughan Williams, L. (2008). Betting exchanges: a technological revolution in sports betting. In D. B. Hausch, & W. T. Ziemba (Eds.), *Handbook of Sports and Lottery Markets* (pp. 403-418). Amsterdam: North Holland.

Smith, M. A., & Vaughan Williams, L. (2010). Forecasting horse race outcomes: new evidence on odds bias in UK betting markets. *International Journal of Forecasting, 26*(3), 543-550.

Sung, M., Johnson, J. E. V., & Peirson, J. (2012). Discovering a profitable trading strategy in an apparently efficient market: exploiting the actions of less informed traders in speculative markets. *Journal of Business Finance & Accounting, 39*(7-8), 1131-1159.

Teebagy, N., & Chatterjee, S. (1989). Inference in a binary response model with applications to data analysis. *Decision Sciences, 20*(2), 393-403.

Veall, M. R., & Zimmerman, K. F. (1996). Pseudo-$R^2$ measures for some common limited dependent variable models. *Journal of Economic Surveys, 10*(3), 241-259.

Wolfers, J., & Zitzewitz, E. (2006). Five open questions about prediction markets. In R. Hahn, & P. Tetlock (Eds.), *Information Markets: A New Way of Making Decisions in the Public and Private Sectors* (pp. 13-36). Washington DC: AEI-Brookings Press.