Journal of Behavioral and Experimental Economics 73 (2018) 86–92

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

Assessing the forecast performance of models of choice

Dale O. Stahl

Department of Economics, University of Texas, Austin, TX, United States

ARTICLE INFO

Keywords:
Forecast performance
Over-fitting
Cross-validation
Lottery choice

J.E.L. Classifications:
C52
C53
C91
D81

ABSTRACT

We often want to predict human behavior. It is well-known that the model that fits in-sample data best is not necessarily the model that *forecasts* (i.e. predicts out-of-sample) best, but we lack guidance on how to select a model for the purpose of forecasting. We illustrate the general issues and methods with the case of Rank-Dependent Expected Utility versus Expected Utility, using laboratory data and simulations. We find that poor forecasting performance is a likely outcome for typical laboratory sample sizes due to over-fitting. Finally we derive a decision-theory-based rule for selecting the best model for forecasting depending on the sample size.

1. Introduction

The desire to understand and predict human behavior motivates theorizing and building models of choice behavior, especially for situations in which the consequences of the choices are uncertain. Von Neumann and Morgenstern (1953) axiomatized Expected Utility (EU) theory, which has become the mainstream model in economics. From its inception, the EU model has faced heavy criticism outside and inside economics and has been subjected to laboratory testing. However, with rare exception (e.g. Wilcox, 2008 and Murphy and ten Brincke, 2017), the testing has focused on showing how well EU and alternative models fit the data rather than assessing the model's ability to forecast (i.e. predict out-of-sample). In contrast, the focus of this paper is on assessing the forecast performance of alternative models of choice under uncertainty.

We will illustrate the general issues and methods with the case of Rank-Dependent Expected Utility (RDEU) versus EU.[1] Since the RDEU model nests EU, RDEU can fit any data it is confronted with at least as well as EU. However, a better fit does not imply a better forecast, especially given the small sample sizes provided by laboratory experiments. On a small sample, RDEU may fit significantly better than EU because the extra parameter gives it the ability to fit the noise in the sample (called “over-fitting”), which leads to biased parameter estimates; hence, the RDEU forecast could be worse than the EU forecast.

To convince the reader that over-fitting is a real danger, we will demonstrate the problem using the data from Hey and Orme (1994; hereafter HO), which is one of the first papers to confront a variety of decision theories with experimental data from a large number of well-designed choice tasks. The standard statistical method to assess over-fitting is the split-sample method of cross-validation[2]: the data is divided into an “estimation” subset and a “holdout” subset. One estimates the model parameters on the estimation subset, and then tests whether this fitted model is the data generating process (DGP) for the holdout subset. We find that the answer is negative, and we further show that RDEU forecasts worse than EU.

There are two possible explanations for this finding: (i) RDEU over-fit the data, and/or (ii) the behavior of the humans was not governed by a single DGP throughout the experiment (i.e. the behavioral process was not stationary). The second question that arises is whether the poor forecast performance result for the HO data is statistically significant or an artifact of this particular data.

Both of these questions can best be addressed using simulation methods. First, in a simulation, the data generation process can be held fixed for both the estimation and the holdout pseudo-data, so non-stationary behavior is ruled out as a possible explanation of poor forecast performance. Second, a simulation can generate a good approximation of the properties of any statistic for a given sample size, so the question of statistical significance can be answered without relying inappropriately on asymptotic theory (e.g. Van der Vaart, 1998). Third, a simulation can determine how large a sample should be for the over-fitting danger to be negligible. Last, but not least, the simulations can be used to find an optimal decision rule for which model to use when

E-mail address: stahl@eco.utexas.edu.

[1] Bruhim et al. (2010) and Fehr-Duda and Epper (2012) explore the econometric issues when the population has heterogeneous RDEU preferences.

[2] There is a voluminous literature on cross-validation: e.g. Stone (1974), and Cawley and Talbot (2010).

https://doi.org/10.1016/j.socec.2018.02.006
Received 5 September 2017; Received in revised form 2 February 2018; Accepted 12 February 2018
Available online 14 February 2018

forecasting based on the size of the in-sample data.

Our simulation exercise demonstrates that the poor forecast performance found using the HO data does not vanish when the DGP is fixed, and that such poor forecast performance should be expected given the typical laboratory sample sizes. Our simulations also indicate that for accurate estimation and forecasting of the RDEU model, we should have 200 or more binary lottery tasks - otherwise it would be better to use the EU model. Finally, we show that a decision-theory based conditional rule about which model to use for forecasting can improve forecast performance, but still one should have at least 100 binary lottery tasks in the estimation data.

The paper is organized as follows. Section 2 specifies the RDEU models. Section 3 describes the HO experiment and measures the forecast performance on that data. Section 4 describes the simulation exercise and presents the findings. Section 5 addresses the question of how to choose a model for forecasting. Section 5 concludes with a discussion.

2. The rank-dependent expected utility model

A convenient encompassing model is Rank-Dependent Expected Utility[3] (RDEU) (Quiggin, 1982, 1993), which nests EU. RDEU allows subjects to modify the rank-ordered cumulative distribution function of lotteries as follows. Let $Y \equiv \{y_0, y_1, \ldots, y_n\}$ denote the set of potential outcomes of a lottery, where the outcomes are listed in rank order from worst to best. Given rank-ordered cumulative distribution for a lottery on Y, let F_j denote the cumulative probability up to and including y_j. It is assumed that the subject transforms F_j by applying an increasing function $H(F_j)$ with $H(0) = 0$ and $H(1) = 1$. From this transformation, the individual derives modified probabilities of each outcome:

$$h_0 = H(F_0),\ \ h_1 = H(F_1) - H(F_0),\ \ldots \text{ and } h_n = 1 - H(F_{n-1}). \tag{1}$$

Common parametric specifications of the transformation functions are

$$H(F_j) \equiv (F_j)^{\beta}/[(F_j)^{\beta} + (1 - F_j)^{\beta}], \tag{2a}$$

$$H(F_j) \equiv (F_j)^{\beta}/[(F_j)^{\beta} + (1 - F_j)^{\beta}]^{1/\beta}, \tag{2b}$$

$$H(F_j) \equiv (\alpha\, F_j)^{\beta}/[(\alpha\, F_j)^{\beta} + (1 - F_j)^{\beta}],\ \alpha > 0, \tag{2c}$$

$$H(F_j) \equiv \exp\{-\beta\,[-ln(F_j)]^{\alpha}\},\ \alpha > 0, \tag{2d}$$

where $\beta > 0$. Arguing from symmetry that H(0.5) should equal 0.5, Quiggin (1982) recommended Eq. (2a). Tversky and Kahneman (1992) suggested Eq. (2b) because it allows the interior fixed point to differ from 0.5. Lattimore et al. (1992) suggested Eq. (2c) which allows a greater range on the shape and fixed point. Prelec (1998) provides an axiomatic foundation for an alternative two-parameter transformation (2d). For ease of reference, RDEU0 will refer to the EU model (i.e. $\beta = 1$ and $\alpha = 1$); RDEU1 will refer to the model with Eq. (2a), RDEU2 to the model with Eq. (2b), RDEU3 to the model with Eq. (2c), and RDEU4 to the model with Eq. (2d).

Given value function $v(y_j)$ for potential outcome y_j, the *rank-dependent expected utility* is

$$U(F) \equiv \sum_j v(y_j)h_j(F). \tag{3}$$

To confront the RDEU model with binary choice data (F^A vs. F^B), we assume a logistic choice function:

$$\mathrm{Prob}(F^A) = exp\{\gamma U(F^A)\}/[exp\{\gamma U(F^A)\} + exp\{\gamma U(F^B)\}] \tag{4}$$

where $\gamma \geq 0$ is the precision parameter. Without loss of generality, we can assign a value of 0 to the worst outcome and a value of 1 to the best outcome.[4] Accordingly, we specify $v_0 \equiv v(y_0) = 0$ and $v_n \equiv v(y_n) = 1$. This leaves n-1 free utility parameters: $v_j \equiv v(y_j)$ for j = 1,…, n-1, with the monotonicity constraint that $v_j \geq .\ v_{j\text{-}1}$ for j = 1,…, n. Hence, the empirical RDEU0 model entails n parameters: (γ, $\underline{v}$), the RDEU1 and RDEU2 models entail n + 1 parameters: (γ, $\underline{v}$, β), and the RDEU3 and RDEU4 models entail n + 2 parameters (γ, $\underline{v}$, β, α). It is obvious that RDEU3 nests RDEU1 (when $\alpha = 1$), and RDEU1 and RDEU2 nest RDEU0 (when $\alpha = 1$ and $\beta = 1$).

Next, to specify the likelihood function for our data, let $\underline{x}_i \equiv \{x_{i1}, \ldots, x_{iT}\}$ denote the choices of subject i for T lottery pairs indexed by $t \in \{1, \ldots T\}$, where $x_{it} = 1$ if lottery A was chosen, and 0 otherwise. Then the probability of the T observed choices of subject i is the product of the probability of each choice given by Eq. (4).[5] For notational convenience, let $\theta_i \equiv (\gamma_i, \underline{v}_i, \beta_i, \alpha_i)$. Then, in log-likelihood terms:

$$LL(x_i, \theta_i) \equiv \sum_{t=1}^{T} [x_{it}\ln(\mathrm{Prob}[F^A(\theta_i)]) + (1 - x_{it})\ln([1 - \mathrm{Prob}[F^A(\theta_i)]])]. \tag{5}$$

Then, we define the total log-likelihood of the data as

$$LL(\mathbf{x}, \theta) \equiv \sum_i LL(\underline{x}_i, \theta_i) \tag{6}$$

where $\boldsymbol{\theta} \equiv \{\theta_i, i = 1, \ldots, N\}$, and $\mathbf{x} \equiv \{\underline{x}_i, i = 1, \ldots, N\}$.

3. The Hey-Orme experiment and performance tests

3.1. The experiment

Hey and Orme (1994; hereafter HO) is one of the first papers to confront a variety of decision theories with experimental data from a large number (100) of choice tasks.[6] Each task was a choice between two lotteries with three prizes drawn from the set {£0, £10, £20, £30}.[7] A crucial design factor was the ratio of (i) the difference between the probability of the high outcome for lottery A and the probability of the high outcome for lottery B to (ii) the difference between the probability of the low outcome for lottery A and the probability of the low outcome for lottery B. It is insightful to represent this choice paradigm in a Machina (2008) triangle, as shown in Fig. 1.

The ratio for the A-B pair is the slope of the dotted line connecting A and B, which is greater than 1. The ratio for the A' -B' pair (dashed line) is clearly less than 1. According to EU indifference curves are parallel straight lines with positive slope in this triangle, and the indifference curves of a risk neutral subject would have slope equal to 1. A wide range of ratios was used in order to identify indifference curves and to test the implications of EU (as well as alternative theories).

[3] This model is the same as the Cumulative Prospect (Tversky and Kahneman, 1992) model restricted to non-negative monetary outcomes.

[4] Since we estimate one precision parameter for all choice tasks, this scale specification is not simply the assumption of affine invariance; it is also an assumption about the magnitude of "noise" implicit in the logistic function relative to the payoffs. Wilcox (2008) argues for a re-scaling for each choice task. While we agree that re-scaling may be needed for diverse choice tasks, we feel that in the context of the HO tasks, since all four payoffs were encountered many times in succession, a re-scaling for the entire set is more appropriate. To test our intuition, we estimated the Wilcox-type EU model for the HO data (which he used), and we found it fit slightly worse than a EU model without rescaling for each task. This different finding may be due to our using only the first 100 tasks of HO and estimating individual parameters rather than a random coefficient specification.

[5] As pointed out by Harrison and Swarthout (2014), this specification implicitly assumes the "compound independence axiom". Since we view EU and RDEU as behavioral models, we are comfortable with this implicit assumption.

[6] These 100 tasks were presented to the same subjects again one week later. We do not consider that data here because the test that the same model parameters that best fit the first 100 choices are the same as those that best fit the second 100 choices fails. Possible explanations for this finding are (i) that learning took place between the sessions, (ii) preferences changed due to a change in external (and unobserved) circumstances, and (iii) the subjects did not have stable preferences. Therefore, we focus our attention on the first 100 choice tasks.

[7] £ is the British pound.