FIT OR HIT IN CHOICE MODELS

KHALED BOUGHANMI, RAJEEV KOHLI, AND KAMEL JEDIDI

Abstract. The predictive validity of a choice model is often assessed by its hit rate. We examine and illustrate conditions under which a choice model with a higher likelihood value may obtain a lower hit rate. We also show that the solution obtained by maximizing a likelihood function can be different from the solution obtained by maximizing the hit rate. The analysis and results suggest that the hit rate should not be overly emphasized when the objective is testing a theory and/or statistical inference. But if the aim is prediction, then the expected hit rate can be maximized.

1. Introduction

Marketing researchers often assess the predictive validity of a discrete choice model by its hit rate. A good model is expected to have a higher likelihood value[1] and a higher hit rate in estimation and holdout data. We examine if this is an appropriate expectation.

A choice model with a higher likelihood value can be shown to guarantee a better *lower bound* on the hit rate. But this does not guarantee that its actual hit rate is also higher. Let p_k denote the choice probability of the alternative selected from the kth choice set. Then the difference between the actual hit rate and the lower bound depends on the variance of $\sqrt{p_k}$. We obtain conditions under which a model can obtain a higher maximum likelihood value but a lower hit rate than a competing model.

The present result implies that it is better not to mix likelihood maximization and hit rate performance. If the objective is statistical inference based on sample data, then it is appropriate to compare models using maximum likelihood and related criteria like BIC.

Date: August 17, 2016.

[1] Or a higher value on a related measure, like AIC and BIC.

If the objective is prediction, competing models should be estimated using hit rate as the criterion. For example, the latter criterion may be appropriate when an online retailer wishes to recommend products to consumers.

2. Likelihood maximization and prediction

The following analysis is relevant for any probabilistic choice model. The data used for estimating the model consist of the choices from each of n choice sets C_k, $k = 1, \dots, n$. We focus on maximum likelihood estimation of the model parameters.

Let p_k denote the choice probability for the alternative selected from C_k. For example, in a multinomial logit model,

$$p_k = \frac{e^{v_k}}{\sum_{j \in C_k} e^{v_j}}, \ k = 1, \dots, n,$$

where v_j is the deterministic utility of alternative $j \in C_k$. The value of v_j can be a function of covariates. Let $l = p_1 p_2 \dots p_n$ denote the likelihood function, and $\hat{p}_k$ the maximum likelihood estimate of p_k, for each $k = 1, \dots, n$. Let $\hat{l} = \hat{p}_1 \dots \hat{p}_n$ denote the maximum value of the likelihood function.

The predictive validity of a choice model is commonly assessed by its hit rate. One method of prediction assumes that an alternative is chosen from a choice set if it has the highest deterministic utility. The problem with this method is that it ignores the very uncertainty in choices that are intended to be captured by estimating a random utility model. Another method predicts that an alternative is chosen if it has the highest estimated choice probability in a choice set. Its limitation is that it does not distinguish between predictions in which an alternative is almost certainly chosen or is barely chosen (that is, when one choice probability is close to one, and another just exceeds $1/s$, where s is the number of alternatives in a choice set). A more suitable approach, used for example by Gilbride and Allenby (2006), employs the parameter estimates to obtain the choice probabilities for all alternatives in a choice set, and then uses these to simulate the alternative chosen from the choice set. A "hit" is recorded if a simulated choice matches the alternative actually chosen in a choice set. Since this occurs with probability $\hat{p}_k$ in choice set C_k, the expected value of a hit for choice set C_k is

$$\hat{p}_k \cdot 1 + (1 - \hat{p}_k) \cdot 0 = \hat{p}_k.$$

The expected hit rate across the n choice sets, $C_1, \ldots, C_n$, is

$$\hat{h} = \frac{1}{n} \sum_{k=1}^{n} \hat{p}_k. \tag{1}$$

That is, the (expected) hit rate is equal to the arithmetic mean of the predicted choice probabilities for the alternatives that are chosen from the n choice sets. For brevity, we refer to the expected hit rate as simply the hit rate in the rest of the paper.

Let

$$\hat{g} = (\hat{p}_1 \hat{p}_2 \ldots \hat{p}_n)^{1/n} \tag{2}$$

denote the geometric mean of the probabilities. Then $\hat{g} = (\hat{l})^{1/n}$. Since the arithmetic mean of a set of numbers is no smaller than their geometric mean, $\hat{h} \geq \hat{g}$: the expected hit rate obtained by maximizing the likelihood function is at least as large as the nth root of the maximum likelihood value.

Let M_0 and M_1 denote two choice models estimated using the same data. Let $\hat{l}_0$ and $\hat{l}_1$ be the maximum values of the likelihood function for M_0 and M_1. Let $\hat{g}_0 = (\hat{l}_0)^{1/n}$ and $\hat{g}_1 = (\hat{l}_1)^{1/n}$ be the geometric means, and $\hat{h}_0$ and $\hat{h}_1$ the arithmetic means for M_0 and M_1, respectively. Thus, $\hat{l}_0 \leq \hat{l}_1$ implies $\hat{g}_0 \leq \hat{g}_1$. Since the arithmetic mean of any n numbers is no smaller than their geometric mean, $\hat{g}_0 \leq \hat{h}_0$ and $\hat{g}_1 \leq \hat{h}_1$. Thus, there are three possible orderings of $\hat{g}_0, \hat{h}_0, \hat{g}_1$ and $\hat{h}_1$:

$$(1)\ \hat{g}_0 \leq \hat{h}_1 \leq \hat{g}_0 \leq \hat{h}_2$$

$$(2)\ \hat{g}_0 \leq \hat{g}_1 \leq \hat{h}_0 \leq \hat{h}_1$$

$$(3)\ \hat{g}_0 \leq \hat{g}_1 \leq \hat{h}_1 \leq \hat{h}_0$$

Case (1) is consistent with the expectation that a model with a lower likelihood value also has a lower hit rate ($\hat{g}_0 < \hat{g}_1$ and $\hat{h}_0 < \hat{h}_1$). The difference between cases 2 and 3 is a matter of degree. Case 2 allows the two models to have different likelihood values but the same hit rates ($\hat{g}_0 \leq \hat{g}_1$ and $\hat{h}_0 = \hat{h}_1$). Case 3 is more extreme, allowing one model to have a lower likelihood value but a higher hit rate ($\hat{g}_0 < \hat{g}_1$ and $\hat{h}_1 < \hat{h}_0$). Below, we examine the conditions under which each of these cases can occur.

Let $\hat{p}_k$ and $\hat{p}_k^{'} = (1+\epsilon_k)\hat{p}_k$ denote the maximum likelihood estimates of the choice probabilities obtained using models M_0 and M_1, where ϵ_k is a suitable positive or negative number. The maximum likelihood values for the two models are

$$\hat{l}_0 = \prod_{k=1}^{n} \hat{p}_k \ \text{ and } \ \hat{l}_1 = \prod_{k=1}^{n} \hat{p}_k^{'} = \prod_{k=1}^{n} (1+\epsilon_k)\hat{p}_k. \tag{3}$$

Let

$$\hat{g}_0 = \hat{l}_0^{1/n} \ \text{ and } \ \hat{g}_1 = \hat{l}_1^{1/n} \tag{4}$$

denote the geometric means of the probabilities $\hat{p}_k$ and $\hat{p}_k^{'}$, $k = 1, \ldots, n$. The corresponding arithmetic means, which are equal to the expected hit rates for models M_0 and M_1, are

$$\hat{h}_0 = \frac{1}{n}\sum_{k=1}^{n} \hat{p}_k \ \text{ and } \ \hat{h}_1 = \frac{1}{n}\sum_{k=1}^{n}(1+\epsilon_k)\hat{p}_k = \hat{h}_0 + \frac{1}{n}\sum_{k=1}^{n}\hat{p}_k\epsilon_k. \tag{5}$$

Observe that $\hat{h}_1 \leq h_0$ is equivalent to

$$\sum_{k=1}^{n} \hat{p}_k\epsilon_k \leq 0. \tag{6}$$

Similarly, $\hat{g}_1 \geq \hat{g}_0$ is equivalent to

$$\frac{\hat{g}_1}{\hat{g}_0} = \prod_{k=1}^{n}(1+\epsilon_k)^{1/n} \geq 1,$$

which can be rewritten as

$$\sum_{k=1}^{n} \log(1+\epsilon_k) \geq \log(1) = 0. \tag{7}$$

Since $\epsilon_k > \log(1+\epsilon_k)$, this condition implies that $\hat{g}_1 \geq \hat{g}_0$ if

$$\sum_{k=1}^{n} \epsilon_k \geq \sum_{k=1}^{n} \log(1+\epsilon_k) \geq 0. \tag{8}$$

Thus, a necessary condition for model M_1 to have a higher likelihood value but a lower hit rate than model M_0 (i.e., $\hat{g}_1 \geq \hat{g}_0$ and $\hat{h}_1 \leq h_0$) is

$$\sum_{k=1}^{n} \hat{p}_k \epsilon_k \leq 0 \quad \text{and} \quad \sum_{k=1}^{n} \epsilon_k \geq 0. \tag{9}$$

The condition in equation (9) is also sufficient if each ϵ_k value is small, because in this case $\log(1+\epsilon_k) \approx \epsilon_k$.

Since $\hat{p}_k \geq 0$, the condition $\hat{p}_k \epsilon_k \geq 0$ is equivalent to $\epsilon_k \geq 0$. Model M_1 always has a higher likelihood value and a higher hit rate than model M_0 if $\epsilon_k > 0$ for all $k = 1, \ldots, n$. Otherwise, equation (9) implies that M_1 can have a higher likelihood value but a lower hit rate than M_0. Situations favoring this outcome occur if $\epsilon_k < 0$ when p_k is large, and $\epsilon_k \geq 0$ when p_k is small; that is, M_1 predicts a lower choice probability than M_0 (i.e., p_k is large and $\epsilon_k < 0$) in cases where M_0 predicts a high choice probability for the selected alternative; and M_1 predicts a choice probability no smaller than M_0 (i.e., p_k is small and $\epsilon_k \geq 0$) in cases where M_0 predicts a low choice probability for the selected alternative. The following example illustrates how this can occur. We generated $n = 1,000$ binary outcomes, 200 of which corresponded to purchases, and 800 to non-purchases, of a product by consumers. A single independent variable, x, was used to predict the buy/no buy outcome in a binary logit model. The predictor variable had a value of $x = 4$ for the 200 purchase observations, $x = -5$ for 700 of the no-purchase observations, and $x = 10.5$ for the remaining 100 no-purchase observations. That is, there was no purchase if the value of

x was much smaller or much larger than $x = 4$. Model M_0 was the logistic regression

$$\hat{p}_k = \frac{1}{1 + e^{\hat{\beta}_1 x_k}},$$

in which the purchase probability $\hat{p}_k$ was a function of x. Model M_1 was the logistic regression

$$\hat{p}_k = \frac{1}{1 + e^{\hat{\beta}_0}},$$

in which the purchase probability was a function of an intercept term, but not of x.[2] The maximum likelihood estimates were $\hat{\beta}_0 = -0.25$ and $\hat{\beta}_1 = 1.38$ for models M_0 and M_1. Table 1 shows the log-likelihood values and the hit rates for the two models.

Table 1. Log-likelihood and hit rate for M_0 and M_1

Model	Log-likelihood	Hit rate
M_0	-508.46	0.70
M_1	-500.40	0.68

Model M_1 has the higher log likelihood value but the lower hit rate. Since each model has $m = 1$ parameter, the BIC values, $-2 \ln \hat{l}_i + m \ln n$, are 1007.71 for model M_1 and 1023.83 for model M_0. The BIC criterion favors the selection of model M_1 over M_0. The hit rate favors the selection of M_0 over M_1.

2.1. Lower bound for hit rate.

As noted, the nth root of the likelihood function is a lower bound for the expected hit rate of a model. Without loss of generality, suppose $\hat{p}_1 \leq \cdots \leq \hat{p}_n$, where $\hat{p}_k$ is the predicted choice probability of the alternative chosen from set C_k, for all $k = 1, \ldots, n$. Then a result by Aldaz (2012) implies that

$$\hat{h} \geq \hat{g} + \frac{1}{n-1} \sum_{k=1}^{n} \left(\sqrt{\hat{p}_k} - s \right)^2, \tag{10}$$

[2] These are not the best fitting models for the data. For example, a model with both the intercept and the covariate can fit the data better than either model. M_0 and M_1 were chosen only to illustrate a situation in which a model with higher maximum likelihood does not necessarily result in a higher hit rate.

where

$$s = \frac{1}{n} \sum_{k=1}^{n} \sqrt{\hat{p}_k}. \tag{11}$$

That is, the hit rate is no less than the geometric mean plus the variance of the square roots of the choice probabilities. Observe that $\frac{1}{n-1} \sum_{k=1}^{n} \left(\sqrt{\hat{p}_k} - s\right)^2 = 0$ only if all $\hat{p}_k$ values are equal. In this case, the arithmetic and geometric means are also equal; maximizing the likelihood function is equivalent to maximizing the expected hit rate. As the variance of $\sqrt{\hat{p}_k}$ increases, so does the minimum difference in the value of $\hat{h} - \hat{g}$.

3. Empirical Illustrations

In the previous section, we showed that it is possible for a model to have a higher maximum likelihood value but the same or lower hit rate than another model estimated using the same data. In this section, we illustrate the results with three empirical applications. The first application compares a latent-class logistic regression with a latent-class probabilistic disjunctive model for binary (acceptable/unacceptable) data. The second application compares a nested logit model with a multinomial logit model. The third application compares nine different models estimated using a hierarchical Bayesian approach. In the first two applications, one model is assessed to be better based on maximum likelihood and BIC values, but shows no improvement on the hit rate, compared to a competing model. In the third application, there is a tradeoff between the log marginal density and the hit rate, and the model with the highest log marginal density has the lowest hit rate.

3.1. **Battery conjoint study.**

Data. Jedidi and Kohli (2005) reported a conjoint study using acceptable/unacceptable data for household batteries from 175 consumers. Thirty-two battery concepts were generated using an orthogonal, main-effects design. The following attributes were used in generating the concepts: (1) incremental price (0%, 25%, and 50% higher than the current price), (2) built-in charge meter (yes/no), (3) environmental safety (yes/no), (4) rapid recharge (yes/no), (5) battery life (standard, 50% longer life), and (6) brand name (A, B, C, D). Each respondent saw each battery concept in random order and reported if he or she would consider purchasing the battery if it were available in the market. On average, respondents were willing to consider 82% of the battery concepts.

Estimation results. We used the data to estimate two latent-class models of battery consideration by consumers. The first was a disjunctive model (M_0), in which an alternative was acceptable if at least one of its attribute levels was acceptable. The second was a latent-class logistic regression (M_1). The full estimation results are available in Jedidi and Kohli (2005). Here we report the log-likelihood values and the expected hit rates for a three-segment latent class model (the three-segment solution was selected because it obtained the best fitting logistic regression). Each model was estimated using a randomly selected subset with 90% of the observations. The remaining observations were used for holdout predictions. The procedure was replicated fifty times. Table 2 shows the average log likelihood values and hit rates across the replications.[3]

Table 2. Model performance statistics for latent class disjunctive and logistic regression models for battery data

Latent class model	# of par.	Log Likelihood	Average expected hit rate in sample	 holdout
Logistic regression	23	-1454.1	0.86	0.86
Disjunctive	35	-1482.6	0.86	0.86

The logistic regression has a higher likelihood value, and performs significantly better than the disjunctive model based on the BIC criterion (3106.70 vs. 3267.27). Despite superior fit, it has the same in-sample and out of sample hit rates (86%) as the disjunctive model. Thus, we conclude that the logistic regression is the better model because it has a higher likelihood value, fewer parameters, and the same hit rate as the disjunctive model.

[3] The expected hit rates in Table 2 differ from the hit rates reported by Jedidi and Kohli (2005), because the latter predicted choices using a deterministic, maximum utility rule.

3.2. **King salmon fishing in Alaska.**

Data. The Alaska Department of Fish and Game (ADFG) sponsored a study to assess the choice of recreational fishing destinations by state residents. We analyzed the data for 440 respondents who made 1327 trips (an average of over three trips per person), to catch King salmon during the summer season. Since ADGF closed some sites each week, the number of sites open during a week ranged between three and twenty. Much of the population lived in a few locations in a geographically contained area. Fifteen of the twenty sites were accessible to most residents by car. The other five were accessible only by air. Carson et al. (2009) describe the data collection and the calculation of individual travel costs to fishing sites, which we used in the following analysis.

Estimation results. We used the site choice data to estimate a multinomial logit model (M_0) and a nested logit model (M_1). The best-fitting nested logit model grouped the fifteen sites accessible by car in one nest, and the five sites accessible by air in the other nest. The utility of a site varied by week, and was a function of the following four covariates: (1) the quality rating of a fishing site for King salmon during a week (1=poor, 8=excellent), (2) the site crowd rating during a week, (3) cost of traveling to a site during a week, and (4) cabin ownership at a site. The first two covariates reflected the weekly variability in site attractiveness, and the latter two heterogeneity among respondents.

Table 3 shows the log-likelihood values for the models, obtained using the full information maximum likelihood method (Proc NLP in SAS).[4] It also shows the average in-sample and holdout (expected) hit rates. The latter were obtained by estimating each model using a randomly selected sub-sample with 90% observations and predicting the choice probabilities for the remaining 10% observations. The holdout percentages reported in Table 3 are the average (expected) hit rates over 100 replications.

[4] The parameter estimates of the two models are available from the authors upon request.

Table 3. Model performance statistics for multinomial logit and nested logit model for King salmon fishing in Alaska

Model	# of par.	Log Likelihood	Average hit rate in sample	 holdout
Multinomial logit	23	-2497.12	0.27	0.27
Nested logit (fly, drive)	25	-2486.96	0.27	0.27

The nested logit model obtains significantly better fit than the multinomial logit model ($\chi^2 = 20.32$; $p < 0.001$). However, both models have the same in-sample and out-of-sample hit rates. Thus, the improvement in the maximum likelihood value suggests that the nested logit model is the better model. But if we also consider the hit rate, we would conclude that it is no better than the simpler multinomial logit model, which also has two less parameters.

3.3. **Choice among price tariffs.**

Data. Schlereth (2013) examined the relation between Internet use and the structure of two-part tariffs in the usage plans available to consumers. The models simultaneously predicted purchase, plan (tariff) choice and usage, and were estimated using data from an online discrete choice experiment. Each of 206 student subjects evaluated 21 different pairs of tariff plans. Each plan has a fixed monthly fee, which ranged between 11 and 32 Euro, and a usage fee, which ranged between 0.30 and 1.20 Euro per hour. A subject could choose one, or reject both, plans in a pair.

Estimation results. Schlereth (2013) compared ten different models predicting the choices made by the respondents. Model M_1 was a standard multinomial logit model. Models M_2 to M_5 used different formulations of consumer utility. Models M_6 to M_9 used different formulations of a consumer's willingness to pay. All models, except M_2 and M_6, allowed for usage uncertainty. We exclude model M_{10} from the discussion below because it alone used a two-step estimation procedure; all other models were estimated in one step, using a hierarchical Bayesian procedure. The estimation sample consisted of nineteen randomly chosen pairs of plans for each respondent. The other two pairs were held out for validation.

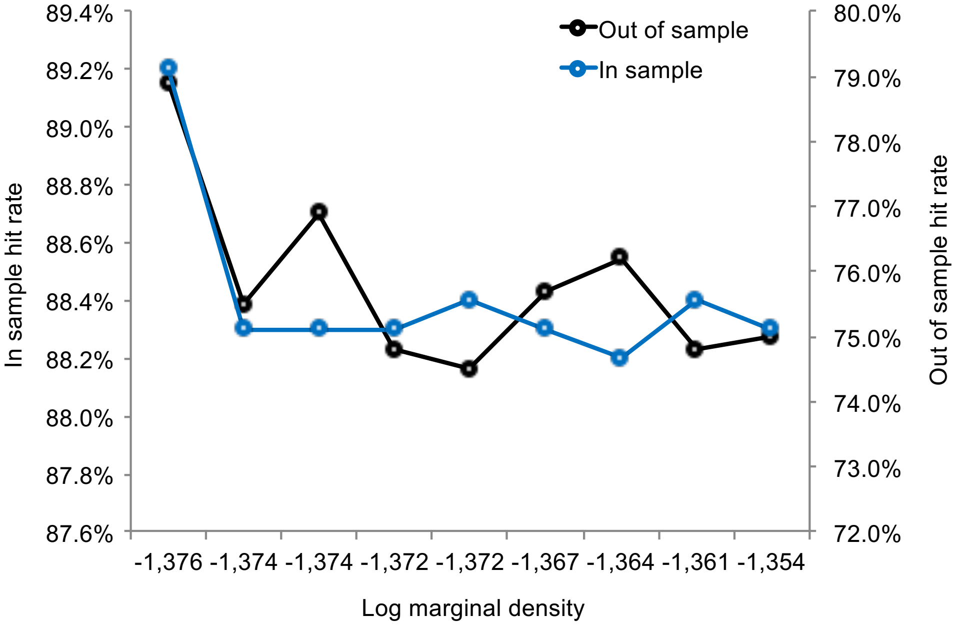


Figure 1. Log marginal density vs. in-sample and out of sample hit rates for the nine models estimated by Schlereth (2013).

Figure 1 plots the in-sample and out-of-sample hit rates against the log marginal density (LMD) for each of the nine models.[5] The general pattern is that models with lower values of the log marginal densities have higher hit rates, both in and out of sample. The out of sample hit rates have greater variability (74.5% to 78.9%) than the in-sample hit rates (88.2% to 89.2%), but both have the same pattern of a negative relation between the log marginal density and the hit rate. Based on the likelihood marginal density values (which penalize for over-parametrization), the best model is M7 (LMD=-1354.4, in-sample

[5] The data for the figure were obtained from Table 4 in Schlereth (2013, p. 13).

h=88.3%, out-sample h=75.0%); based on the hit rates, the best model is M1 (LMD=-1376.2, h=89.2%, out-sample h=78.9%).

4. Fit or hit: maximizing likelihood or hit rate

The preceding analysis shows that a model with a higher maximum likelihood need not have a higher hit rate. It suggests that if the objective is to estimate a best fitting model using sample data (and draw inferences about population parameters), then we should compare the maximum likelihood values (and related measures) of competing models. However, there can be situations in which predicting choices is the main modeling objective. For example, an online retailer may be interested in making product recommendations to consumers. In this case, the objective may be to find parameter values that maximize the expected hit rate instead of the likelihood value. How different can the hit rates and likelihood values be if the parameters of the same model are estimated to maximize a hit rate instead of a likelihood function? And how different can the parameters estimates be from their maximum likelihood values?

We examined these questions in the context of an empirical application. The results show that changing the objective function can lead to very different results. The hit rate increased from 58.39% to 80% when the objective function maximized hit rate instead of the likelihood value. Simultaneously, the log likelihood value decreased from a maximum of -174.22 to -4470.00. The estimated choice probabilities were all close to zero or one, and the parameter estimates had very large positive or negative values, when the objective function maximized the hit rate. Less extreme values of the choice probabilities and parameter estimates were obtained when the objective function maximized the likelihood function. The results suggest that there can be a significant tradeoff between likelihood maximization and hit rate maximization. Likelihood maximization is more appropriate when the purpose is to test hypotheses and use sample data to estimate the population parameters. Hit rate maximization may be more appropriate when the main objective is prediction.

Data. We analyzed data on transportation choices by 210 non-business travelers between Sydney, Canberra and Melbourne. The data have been previously analyzed by Louviere, Hensher and Swait (2000) and Hensher and Greene (2002). Each individual chose one of four travel alternatives, plane, car, bus and train. The alternatives were described in terms of (1) in-vehicle time (in minutes), (2) in-vehicle cost (in dollars) for all stages of a journey, and (3) waiting time (in minutes) at a terminal for a plane, bus or car. Each respondent also provided information on (4) household income (in $'000s) and (5) party size (≥ 1), which refers to the number of individuals traveling together. These five variables were used as covariates in a multinomial logit model.

Estimation results. We used a nonlinear optimization procedure in SAS (Proc NLP) to obtain the following maximum likelihood solution.

$$\text{Maximum log likelihood: } \ln(\hat{l}_1) = -174.22. \text{ Hit rate: } \hat{h}_1 = 0.5839.$$

Figure 2 shows hit rates and geometric means for other solutions that are in the neighborhood of the maximum likelihood solution. These solutions have lower likelihood values (geometric means), but higher hit rates. For example, there is a solution for which

$$\text{Log likelihood: } \ln(\hat{l}_2) = -205.41. \text{ Hit rate: } \hat{h}_2 = 0.662.$$

The highest hit rate shown in Figure 2 is associated with the rightmost solution, for which the log-likelihood value and the hit rate obtain the following values:

$$\text{Log likelihood: } \ln(\hat{l}_3) = -305.46. \text{ Hit rate: } h_3 = 0.6948.$$

To further examine the tradeoff between the hit rate and the likelihood value, we used the iterative optimization procedure to maximize the hit rate $(\hat{p}_1 + \cdots + \hat{p}_n)/n$, where $n = 210$. The procedure was started with the maximum likelihood solution and obtained the following solution after sixteen iterations:

$$\text{Log likelihood: } \ln(\hat{l}_4) = -4470.00. \text{ Maximum hit rate: } h_4 = 0.7999.$$

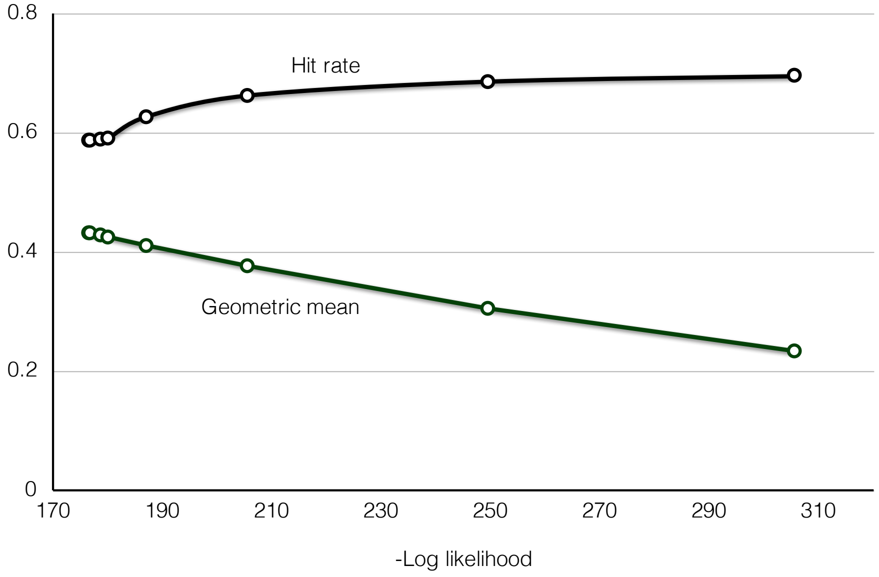
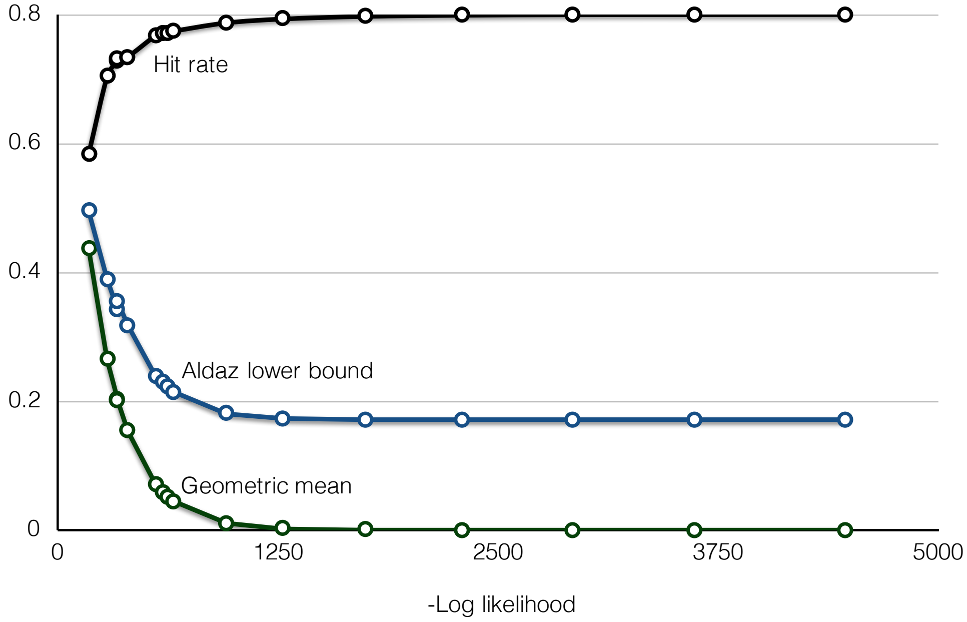


FIGURE 2. Hit rates and geometric means for solutions close to the maximum likelihood solution.

This hit rate is virtually identical to the maximum hit rate of 0.80, which is obtained when the probabilities have 0-1 values (that is, when the likelihood function has a value of zero, and the log likelihood function has a value of minus infinity).

Figure 3 shows the hit rate, the geometric mean and the Aldaz lower bound, $\hat{g}+\mathrm{Var}(\sqrt{\hat{p}})$, as a function of the minus log likelihood value for the sixteen iterations of the optimization algorithm. The leftmost solution shown in the figure is the maximum likelihood solution. The rightmost solution maximizes the hit rate. The maximum likelihood solution has the highest geometric mean, $\hat{g} = \exp(-174.22/210) = 0.4367$ and the lowest hit rate $\hat{h} = 0.5839$. The estimated choice probabilities for this solution have an entropy value of 0.401. The solution with the highest hit rate, $\hat{h} = 0.7999$, has the lowest geometric mean, $\hat{g} = \exp(-4470/210) = 5.698 \times 10^{-10}$. The estimated choice probabilities for this solution approach zero or one values (entropy=0.997) and correctly predict 168 of the 210 choices (the remaining 42 choices are wrongly predicted). The Aldaz lower bound increases with

Figure 3. Hit rate, geometric mean and the Aldaz lower bound as a function of minus log likelihood.



the value of the likelihood function. It is substantially larger than the geometric mean when the latter is close to zero, because the variance of the square root of the choice probabilities is large. Its value increases with the likelihood value, and is the largest for the maximum likelihood solution.

Table 4 shows the parameter estimates obtained in each of the sixteen iterations used to maximize the hit rate. As the hit rate increases in each successive iteration, the value of the likelihood function decreases (see Figure 3), and all parameter values become much larger. The large values of the parameter estimates push the choice probabilities towards zero or one values. Notably, the parameter estimates maximizing the likelihood function and the hit rate have almost perfect correlation (0.996).

Validation. To assess for the validity of the preceding results, we re-estimated the multinomial logit model described above using a random sub-sample with 90% of the data (189

Table 4. Parameter estimates for the multinomial logit model at successive iterations of the algorithm optimizing the hit rate.

							Household	Household	Party
Iteration	Intercept	Intercept	Intercept	In-vehicle	In-vehicle	Waiting	income	income	size
	(air)	(train)	(bus)	cost	time	time	(air)	(train)	(air)
0	5.847	5.613	3.626	-0.009	-0.004	-0.102	0.014	-0.049	-0.973
1	10.271	8.171	6.359	-0.024	-0.007	-0.193	0.020	-0.049	-1.479
2	14.486	12.082	9.103	-0.030	-0.010	-0.280	0.033	-0.075	-2.193
3	21.425	17.599	12.817	-0.050	-0.014	-0.392	0.031	-0.111	-3.147
4	32.162	25.827	18.074	-0.087	-0.023	-0.550	-0.038	-0.159	-4.110
5	49.514	39.236	25.370	-0.154	-0.029	-0.740	-0.080	-0.271	-5.946
6	67.610	53.594	34.571	-0.221	-0.040	-1.007	-0.103	-0.368	-8.008
7	90.462	72.719	46.533	-0.309	-0.056	-1.358	-0.131	-0.508	-10.683
8	115.344	94.592	59.925	-0.398	-0.078	-1.787	-0.159	-0.659	-14.224
9	115.339	94.598	59.925	-0.398	-0.082	-1.820	-0.163	-0.645	-14.235
10	144.497	120.387	75.150	-0.506	-0.106	-2.305	-0.179	-0.836	-18.510
11	169.144	141.856	88.055	-0.599	-0.127	-2.719	-0.184	-0.988	-22.399
12	194.329	163.197	101.151	-0.694	-0.147	-3.125	-0.203	-1.140	-25.946
13	217.102	182.457	113.032	-0.780	-0.165	-3.491	-0.219	-1.278	-29.121
14	237.390	199.547	123.608	-0.858	-0.181	-3.815	-0.233	-1.401	-31.909
15	255.361	214.605	132.981	-0.928	-0.195	-4.100	-0.245	-1.510	-34.345
16	271.432	227.991	141.385	-0.990	-0.207	-4.354	-0.254	-1.606	-36.493

Note: The parameters in iteration 0 maximize the likelihood function; those in iteration 16 maximize the hit rate.

choice sets), then used the parameter estimates to compute the likelihood value and predict the hit rate for the remaining 10% of the data (21 choice sets). We separately maximized the two objective functions, likelihood value and hit rate. We also examined a solution that was close to the maximum likelihood solution but had a substantially higher hit rate. The reason for examining this solution was to assess if there were solutions in the vicinity of the maximum likelihood solution that provided substantially higher hit rates. We repeated the procedure 100 times. Table 5 shows the results.

Table 5. In-sample and out-of-sample likelihood values and hit rates

	In sample		Out of sample	
	Log likelihood	Hit rate	Log likelihood	Hit rate
Maximum likelihood solution	-155.9548	0.5856	-18.9014	0.5737
Solution near maximum likelihood	-182.8306	0.6658	-22.4868	0.6547
Maximum hit rate solution	-6701.4500	0.7997	-779.4163	0.7996

The first row of Table 5 shows the average values of the in-sample and out-of-sample log likelihood values and hit rates across the 100 replications for the maximum likelihood solutions. The second row shows these averages for a solution close to the maximum likelihood solution (corresponding to the solution in the full sample with log likelihood $\ln(\hat{l}_2) = -205.41$ and hit rate $\hat{h}_2 = 0.662$). The third row shows the averages for the solutions that maximize the hit rate. The results are consistent with those obtained using the full-sample estimates. When maximizing hit rate, both the in-sample and the out-of-sample hit rates increased as the likelihood values decreased. In each solution, the out-of-sample hit rates were close to the in-sample hit rates.

5. Conclusion

The present analysis and results suggest that the use of hit rate as a measure of predictive validity should not be overly emphasized when the objective is testing a theory and/or statistical inference. Instead, it may be better to use predicted log likelihood for validating such models. However, if the aim is prediction, then it is better to explicitly maximize the expected hit rate. Mixing likelihood maximization and hit rate maximization can lead to either the rejection of better statistical models, or to the choice of a suboptimal predictive model.

References

[1] Aldaz, J. M. (2012) "Sharp bounds for the difference between the arithmetic and geometric means," *Archiv der Mathematik*, 99 (4), 393–399.

[2] Carson, R.T., W. M. Hanemann and T. C. Wegge (2009), "A nested logit model of recreational fishing demand in Alaska," *Marine Resource Economics*, 24, 101–129.

[3] Gilbride, T.J. and G.M. Allenby (2006), "Estimating heterogeneous EBA and economic screening rule choice models," *Marketing Science*, 25 (5), 494–509.

[4] Hensher, D.A. and W.H. Greene (2002), "Specification and estimation of the nested logit model: alternative normalisations," *Transportation Research Part B: Methodological*, 36 (1), 1–17.

[5] Jedidi, K. and Kohli, R. (2005), "Probabilistic subset-conjunctive models for heterogeneous consumers," *Journal of Marketing Research*, 42 (4), 483–494.

[6] Louviere, J.J., D.A. Hensher and J. Swait (2000), *Stated Choice Methods: Analysis and Applications in Marketing: Transportation and Environmental Valuation*, Cambridge: Cambridge University Press.

[7] Schlereth, C. (2013), "A Comparison of Nonlinear Pricing Preference Models for Digital Services," *Thirty Fourth International Conference on Information Systems*, Milan.