



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

On the Evaluation of Probability Forecasts: An Application to Qualitative Choice Models

Senarath Dharmasena
Department of Agricultural Economics
Texas A&M University
sdharmasena@tamu.edu

David A. Bessler
Department of Agricultural Economics
Texas A&M University
d-bessler@tamu.edu

Oral Capps, Jr.
Department of Agricultural Economics
Texas A&M University
ocapps@tamu.edu

***Selected Paper prepared for presentation at the 2016 Agricultural & Applied Economics
Association Annual Meeting, Boston, Massachusetts, July 31-August 2***

Copyright 2016 by Senarath Dharmasena, David A. Bessler and Oral Capps, Jr., All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies

On the Evaluation of Probability Forecasts: An Application to Qualitative Choice Models

Abstract

Using data from Nielsen HomeScan scanner panel for calendar year 2003, we develop binary choice models to focus on the decision made by a sample of U.S. households to purchase various non-alcoholic beverages. We evaluate the probabilities generated through those qualitative choice models using an array of techniques such as expectation-prediction success tables; receiver operating characteristics (ROC) curve, Kullback-Leibler Information criteria; calibration; resolution (sorting); the Brier score; and the Yates partition of the Brier score.

In using expectation-prediction success tables, we paid attention to *sensitivity* and *specificity*. Use of a naïve 0.50 cut-off to classify probabilities resulted in the over or under estimation of *sensitivity* and *specificity* values compared to the use of the market penetration value. Area under the ROC curve is suggested as an alternative to the use of 0.5 cut-off as well as cut-off at market penetration level to classify probabilities, because this method treats a wide range of cut-off probabilities to come up with a coherent measure in classifying probabilities. The area under the ROC was highest for coffee for with-in-sample probabilities while it was highest for fruit juice model for out-of-sample probabilities. Kullback-Leibler Information Criteria which selects the model with the highest log-likelihood function value observed at out-of-sample observations (OSLLF) to evaluate probabilities show “closeness” or deviation of model generated probabilities to the true data generating probability overall, although this method does not offer classification of probabilities for events that occurred versus that did not. Again, with respect to OSLLF value, probabilities associated with fruit juice model outperform all other beverages. Forecast probabilities with respect to most of the beverage purchases were well calibrated. All resolution graphs were almost flat against a 45-degree perfect resolution graph, indicative of poor sorting power of choice models. The Brier score was lowest for fruit juices and the highest for low-fat milk. According to the calculated Brier score, probability forecasts for fruit juices outperformed other non-alcoholic beverages.

Although the Brier score gave an overall indication of the ability of a model to forecast accurately, the components of the Yates decomposition of the Brier score provided a clearer and broader indication of the ability of the model to forecast.

With-in-sample probabilities generated through logit model for coffee outperforms probabilities generated for other beverages based on area under the ROC curve, covariance between probabilities and outcome index and slope of covariance. Out-of-sample probabilities generated through logit model for fruit juice performs better than any other beverage category based on area under the ROC curve, Brier Score, and OSLLF value.

In the event where researchers are confronted with alternative models that issue probability forecasts, the accuracy of probability forecasts in determining the best model can be measured through myriad of metrics. Even though traditional measures such as expectation-prediction success tables, calibration and log-likelihood approaches are still used, ROC charts, resolution, the Brier score and the Yates partition of the Brier score to evaluate probabilities generated through alternative models are highly recommended.

JEL Classification: C25, C52, D12

Keywords: Qualitative choice, probability forecasts, forecast evaluation, calibration, resolution, Brier score, Yates partition, receiver operating characteristics curve, log-likelihood

Background

Discrete choice models are widely used in economic modeling when the dependent variable corresponds to discrete outcomes. They are used to model choices of decision-makers. Decision-makers can be individual persons, households, or firms. Choice alternatives available might represent competing products, different actions, such as to buy or not to buy a product, or any other option or items over which choices must be made (Train, 2003).

Among a wide range of discrete choice models available to model different situations¹, dichotomous probit and logit models are important to model choices where the dependent variable is set up as a zero-one (0-1) variable. Once appropriately modeled, discrete choice models determine the probability of the choice decision. Importantly, these probit and logit models are used to identify statistically significant factors that are related to the choice decision.

In addition, with an appropriate decision rule, these models provide predictions of various choices. A key question relates to the accuracy of these predictions. Such accuracy can be measured using traditional metrics such as expectation-prediction success tables, where the percentage of correct (incorrect) predictions are calculated in comparison to the total number of predictions based on a predetermined probability cut-off level for success (or not success). The expectation-prediction table is limited in its ability to correctly classify and evaluate probabilities in the absence of accurate predetermined cut-off levels. Receiver Operating Characteristics (ROC) curves (Hsieh and Turnbull, 1996; Reiser and Faraggi, 1997) offer somewhat relief for this predetermined cut-off probability levels in classifying probabilities by calculating and plotting probability outcomes based on a wide range of cut-off probabilities. Alternatively, a log-likelihood function approach which selects models closest to the true data generating process based on Kullback-Leibler Information Criteria has been used to assess performance of models generating probabilities (Stone, 1977; Shao, 1993; Norwood, Lusk and Brorsen, 2004). On the other hand, other techniques such as calibration, calibration graphs, resolution, resolution graphs, and optimum scoring rules such as the Brier Score (BS) and the Yates partition of the Brier Score (Yates, 1982) can be used to measure accuracy of predictions.

¹ Wide range of discrete choice models are available such as probit, logit, mixed logit, ordered probit, generalized extreme value, nested logit, multinomial probit, multinomial logit, etc.

The expectation/prediction success table perhaps is the standard method to evaluate the predictive performance of qualitative choice models (Stock and Watson, 2007). It is noteworthy that to our knowledge, Zellner *et al.* (1991), and Bessler and Ruffley (2004) were the only studies so far that have used optimal scoring rules (such as the Brier score and the Yates-partition of the Brier score) to evaluate probability forecasts from econometric models. According to best of our knowledge, our work is the first such attempt to evaluate within-sample and out-of-sample probability forecasts developed from qualitative choice models (probit and logit) using optimal scoring rules such as the Brier score and the Yates-partition of the Brier score for a wide variety of non-alcoholic beverage purchase decisions in the United States. We develop binary probit and logit models to focus on the decision made by a sample of U.S. households to purchase various non-alcoholic beverages. Then we evaluate the probabilities generated through those models both within-sample and out-of-sample using an array of probability evaluation techniques.

The specific categories of non-alcoholic beverages considered in these analyses are: isotonics (sports drinks); regular soft drinks; diet (low-calorie) soft drinks; high-fat milk (whole milk and 2% milk); low-fat milk (1% milk and skim milk); fruit drinks; fruit juices; bottled water; coffee; and tea. The general objective of the study is to consider and apply methods to evaluate probabilities emanating from discrete choice models of non-alcoholic beverage purchase decisions in the United States. Specific objectives of the study are, to evaluate within-sample and out-of-sample probabilities generated through the respective models using the following metrics: (1) expectation-prediction success tables; (2) Receiver Operating Characteristics (ROC) curves; (3) Log-Likelihood function approach; (4) probability calibration and calibration graphs (Dawid, 1986); and (5) probability resolution and resolution graphs (Dawid, 1986); (6) mean probability score (the Brier Score); and (7) the Yates partition of the Brier Score. (Yates, 2008 and Yates, 2010)

Data

Nielsen Homescan scanner data for 2003 for at-home purchases of non-alcoholic beverages was used in this analysis. Monthly household purchases of these beverages

(expenditure and quantity information) are captured over the period January 2003 through December 2003 (we have a total of 7642 households). These data are demographically representative from various cities and rural markets of the 48 contiguous states of the United States. Each household was provided with a scanner machine in which it could scan and record all items purchased in different retail trade locations throughout a given time period. Panelists recorded the expenditure and quantity of non-alcoholic beverages purchased in that household for that time period.

For each household in the sample, the expenditure and quantity data are summed over 12 months to generate an annual value for each non-alcoholic beverages considered in this study. The quantity data are standardized in terms of gallons per household per year and expenditure data are expressed in terms of dollars per year. It should be noted that some households may have not purchased some beverage products, hence an observation corresponding to zero for quantity purchased.

Since all households did not buy all beverages, a weighted average price was generated taking the ratio of sum of total expenditures of dry good, frozen and dairy beverages to the sum of quantities of dry, dairy and frozen beverages. This weighted average price was used as a proxy for the price of each non-alcoholic beverage considered in this study. Demographic categories used in this analysis are as follows: age of household head, employment status of household head, education status of household head; region; race; Hispanic household; age and presence of children; gender of household head; poverty status of household.

Methodology² and Empirical Results

To evaluate forecast probabilities, we generated two samples of observations and estimated the model using one sample and reserved the data from the second sample to perform out-of-sample forecast evaluation analysis. We divided the sample of 7642 observations in half to generate two random samples of data, each with 3821 household level observations using SAS 9.2 Enterprise Minor data mining software (SAS Institute, 2015). We called these samples, Sample A and B.

² It should be noted that we have generated within-sample and out-of-sample probabilities using both probit and logit models for all ten non-alcoholic beverage purchases. However, for brevity we present the results associated with only the logit model.

Initially, Sample A was used to fit probit and logit models to model the decision to buy a non-alcoholic beverage. Subsequently, within-sample forecast probabilities were generated. Next, we ran the estimated coefficients from Sample A model, through data from Sample B to generate out-of-sample forecast probabilities. To evaluate these within- and out-of-sample forecast probabilities, we consider the following performance methods; expectation-prediction success tables, ROC charts, Log-Likelihood function estimates, calibration and calibration graphs, resolution and resolution graphs, the Brier score and the Yates partition of the Brier score.

Expectation-Prediction Success Tables and ROC Curves

Expectation-prediction success table is a two-way table that shows the relationship between the expected outcome and predicted outcome. Expected outcome is known beforehand; such as the decision to buy or not-to-buy a beverage expressed using an index (or latent) variable. The predicted outcome is generated through the model given the information available at hand (exogenous variables) and it is a probability value when dealing with dichotomous choice models.

A two-by-two contingency table based on expected outcome and predicted probabilities of purchase decision³ provide the number of $y=1$ values correctly and incorrectly predicted, and the number of $y=0$ values correctly and incorrectly predicted. For classification purposes, conventionally, the 0.5 cut-off probability value was used. Therefore, we predict $y=1$, if the estimated probability of $y=1$ exceeds 0.5. In other words, if predicted probability is greater than 0.5, that observation is said to be associated with an event that occurred. The other side of the scenario is where if an event actually did not occur, the predicted probability for that event is below 0.5. As a result, the cut-off level 0.5 classifies the predicted probabilities for events that occurred versus that did not occur.

Consider the following classification table that shows correct and incorrect predictions of an event. One can use 0.5 level of cut-off. There are two events in the example, 0 for not

³ Y being the purchase or non-purchase decision

observing the behavior and 1 for observing the behavior. Let a , b , c , and d be number of occurrences for an event.

		Actual	
		0	1
	Predicted		
	0	a	B
	1	c	D

For a predetermined cut-off probability value, the fraction of $y = 1$ observations that are correctly predicted is termed “*sensitivity*” and is depicted as $d/(b + d)$. The fraction of $y = 0$ observations that are correctly predicted is termed “*specificity*” and is denoted by $a/(a + c)$.

Within-sample and out-of-sample forecast probabilities were generated for probit and logit models. Forecast probabilities are evaluated using a conventional 0.5 cut-off value and a cut-off value generated using the frequency of purchase of a given non-alcoholic beverage (or market penetration level). We center attention to sensitivity and specificity for each beverage for each cut-off value (see Table 1).

For example, isotonics have a low market penetration (about 0.22). We observe an under-estimated sensitivity and overestimated specificity if the 0.5 cut-off is used. However, when the market penetration is used as the cut-off level to classify probabilities, we observe a better sorting of probabilities associated with events that occurred versus events that did not occur. All other beverage categories have high market penetrations compared to isotonics (as high as 0.94 for fruit juices), hence over estimated sensitivity and under estimated specificity are observed if naïve 0.5 cut-off is used to classify probabilities. If respective market penetrations are used to classify probabilities, we observe a better sorting of probabilities. We do not observe a large difference between sensitivity and specificity values generated for within-sample and out-of-sample probabilities.

If cut-off probability 0.5 is used as the convention for classifying (or sorting) probabilities our results show that the sorting for events that occurred versus that did not occur is poor. This is too naïve such that a predicted probability value that is close to 0.5 (say 0.51) or 1 (say 0.99) is associated with an event actually occurred and a predicted probability values that is close to 0 (say 0.01) or 0.5 (say 0.49) is associated with an event that did not occur. According to preceding argument, choice of probability 0.5 as cut-off value is appropriate for an event that

has realized relative frequency value close to 0.5 (event occurs only 50% of the time).

Therefore, choice of cut-off value that is close to the realized relative frequency (or market penetration) value to correctly classify predicted probabilities would be a better way to classify probabilities.

Receiver Operating Characteristics (ROC) curve offers an alternative method to classify probabilities based on sensitivity and specificity derived for a range of cut-off probabilities (alternatively wide range of threshold probabilities), hence the error of over or under classifying probabilities based on a naïve 0.5 cut-off probability is eliminated. In evaluating probabilities using ROC curves, we use the area underneath the ROC curve as a measure of accuracy of probabilities being classified. That is to say, the larger the area under ROC curve, the better the classification of probabilities of events that occurred and did not occur. A ROC curve is a plot of sensitivity against specificity (which would results in the downward sloping graph) or in some instances plot of sensitivity against 1-specificity (which would results in upward sloping graph). In decision making, the model with the largest area underneath the ROC curve is chosen, which is referred to as the generalized ROC criterion by Reiser and Faraggi, (1997). Recent advances in ROC curves and associated statistical tests for area under the ROC curve were developed by Hsieh and Turnbull, (1996), Blume (2002), Reiser and Faraggi (1997), and Vekataraman and Begg, (1996). More updated treatment of ROC curves, see Norwood, Lusk and Brorsen (2004).

According to the results reported in Table 2, area under the ROC curve is generally high for all probability forecasts associated with decision to purchase non-alcoholic beverages. However the highest with-in-sample ROC area is associated with probabilities associated with coffee purchases, which is 0.76. The lowest was reported with purchases associated with low-fat milk, bottled water and tea, which is 0.63. Also as shown in Table 2, probabilities associated with out-of-sample forecasts are sorted relatively well as well, with highest being 0.74 associated with fruit juice purchases and lowest being 0.63, again associated with bottled water and tea. Figure 1 shows the ROC charts generated for with-in-sample probabilities from logit model for different types of non-alcoholic beverage purchases. Figure 2 shows ROC charts associated with probabilities that are generated through out-of-sample forecasts. These figures show and offer confirmation to ROC area under the curve depicted in Table 2.

Kullback-Leibler Information Criteria Approach:

This approach is based on an information criterion called Kullback-Leibler Information Criteria which selects the models closest to the true data generating process, developed by Kullback and Leibler, 1951), and as implemented by Stone (1977), Shao (1993), Norwood *et.al.*, (2004), Royston (2006), and Desmarais and Harden (2013). Kullback-Leibler Information Criteria calculates distance between two probability distributions, one being the true distribution and the other being the model generated probability distribution. One would want to minimize this distance to make sure that the data follow the true generating process. This criterion selects the model with the highest log-likelihood function value observed at out-of-sample observations (Norwood *et.al.*, 2004), hence referred to as out-of-sample log-likelihood function (OSLLF)⁴ by Norwood *et.al.*, (2004). Again, unlike the aforementioned expectation-prediction success table method used to classify probabilities based on predetermined cut-off probability value, Kullback-Leibler Information Criteria method does not require specification of a threshold probability cut-off. For variables that take zero and one values (dichotomous discrete choice models, such as probit or logit), the OSLLF is calculated as follows.

$$OSLLF = \sum_{t=1}^T (1 - G_t) \ln[1 - P_t] + \sum_{t=1}^T G_t \ln[P_t] \quad (1)$$

where G is the 0,1 dichotomous variable and P is the predicted probability associated with G . Table 3 shows the average OSLLF values (averaged across 3819 forecasts) for logit model generated probabilities for ten nonalcoholic beverages considered in the study. The highest OSLLF is calculated for fruit juices, which is -0.2086. The lowest is -0.6182 for low-fat milk. That is to say, the logit model generated probabilities from fruit juice model represent the true (observed) 0,1, probabilities more accurately than they do for the logit model generated probabilities associated with low-fat milk. Although no threshold (or cut-off) probability value is

⁴ These dichotomous models are estimated by maximizing a log-likelihood function. The likelihood function will be higher than its expected value, if with-in-sample observations are used, because, some of the observations are used to estimate parameters (Akaike, 1972; Sawa, 1978). Therefore, in this study, we center attention only to out-of-sample observations to generate log-likelihood function values to compare logit model generated probabilities across models.

used to find the accuracy of probabilities generated, this method does not look deeply into probabilities in such a manner to accurately classify probabilities (which will be shown in methods below).

Probability Calibration and Calibration Graphs

Calibration is a metric of goodness of performance. It is the correspondence between the issued probability for an event *ex-ante* and its long-run realized relative frequency *ex-post*. The calibration criterion is similar to the relative frequency definition of probability. However, calibration does not require a background of repeated trials under identical conditions (Dawid, 1982 and Kling & Bessler, 1989). More formally, for a model to be well-calibrated, for all those events where an x percent probability was assessed, the frequency of occurrence must be x percent for all x (Bunn, 1984). In graphical terms, a well-calibrated qualitative choice model should plot along a 45-degree line with issued probability on x -axis and realized long-run relative frequency on the y -axis. This plot is called “*calibration graph*” or “*calibration function*”. The closer the calibration function is to the 45-degree line, the better the probabilities issued from the qualitative choice model. On the other hand, a model can be consistently overconfident if it issues high probabilities for events that actually do not occur resulting in a calibration curve below the 45-degree line. Also a model can be consistently issuing lower probabilities for events that actually have higher relative frequencies after the fact showing underconfidence, resulting in a calibration curve that is above the 45-degree line.

According to Dawid (1984, page 281) and Bunn (1984 page 150), a continuous random variable (X_n) with a continuous distribution function (F_n), the random fractiles generated (U_n) are distributed uniform $U[0,1]$, i.e. $U_n = F_n(X_n)$. This result is obtained through the *probability integral transform* method explained in Rosenblatt (1952). In other words, when the outcome of the variable (X_n) becomes known, we can define (U_n) as $U_n = F_n(X_n)$, which is the fractile of the distribution function that was actually realized. Since (U_n) is uniformly distributed, it takes the values between 0 and 1. For a perfectly calibrated forecaster, the probability for a particular value U^* would be $P(U \leq U^*) = U^*$ (implying that U should have a uniform probability density function in the ideal situation of perfect calibration (Bunn, 1984)).

Therefore, the cumulative density function for U , which is $F_u(U)$ will in this case describe a straight line, on a graphical representation where U is on the horizontal axis and $F_u(U)$ on the vertical axis. Furthermore, the straight line is $F_u(U) = U$. This graphical representation gives us a perfect calibration function for a continuous random variable. For a more realistic situation of imperfect calibration, the calibration function is generated as follows. Let us suppose that a set of n values of U are available from the realized sequence and they are arranged in the ascending order U_1, U_2, \dots, U_n . To estimate $F_u(U)$ from above data, we can use the following relationship,

$$F_u(U) = \frac{j}{n} \text{ for } j = 1, 2, \dots, n \quad (2)$$

In our analysis of purchase decisions of non-alcoholic beverages, we have a discontinuous random variable to begin with, i.e. purchase or do not purchase, *0,1 type* dichotomous random variable-. When the random variable under consideration is discontinuous, the generation of the calibration function takes a slightly different path. Suppose the dichotomous random variable is Y and the associated cumulative distribution function is $F_Y(Y)$. When the outcome of the variable becomes known, we can define V as $V = F_Y(Y)$. The realized fractile in this case is V . According to David and Johnson (1950), such a realized fractile from a discontinuous random variable is not uniformly distributed; rather they give rise to different moments. In our work on purchase decisions of non-alcoholic beverages, the realized fractile is the probability of purchase of a given non-alcoholic beverage by a household. When the realized fractile is not uniformly distributed, a calibration function with the realized fractile on the horizontal axis and the cumulative probability density function of the realized fractile on the vertical axis cannot be generated. Therefore, we take the following approach.

First, the realized fractiles (in this case probability) are arranged in ascending order and discretized, so that they form desired number of discrete class intervals with a desired class width. For such class intervals, we need to find out the relative frequency of occurrence of the event after the event occurred. One can plot the calibration function for a discrete random variable, where, the probability of occurrence is on the horizontal axis and the realized relative frequency on the vertical axis.

A statistical test for calibration (Dawid, 1984) can be made by testing the observed fractiles from a discrete random variable (V_n in this case) from the sequence of probability forecasts, that is, probabilities of purchase of a given non-alcoholic beverage. If we have J non-overlapping probability subintervals⁵ that exhaust the unit interval, then we can calculate a goodness-of-fit statistic χ^2 as follows;

$$\chi^2 = \sum_{j=1}^J \frac{(a_j - n\pi_j)^2}{n\pi_j} \quad (3)$$

In equation (3), a_j is the actual number of observed fractiles in the interval j (in our study the actual number of observed fractiles is the number of households that did purchase a given non-alcoholic beverage), π_j is the length of probability interval j (or the midpoint of probability class as stated in Seillier and Dawid, 1993) where $(0 \leq \pi \leq 1)$ (Kling and Bessler, 1989 and Seillier and Dawid, 1993), n is the frequency or the total number of households that are found under each probability class (or n such probability forecasts) and $n * \pi$ gives us the expected number of fractiles under each probability class interval. In establishing the test statistic in equation 2, the expected number of fractiles, i.e. $n\pi_j$ is compared against the actual number of observed fractiles, i.e. a_j . The number calculated in equation (2) is compared against the *chi*-squared distribution with $J - 1$ degrees of freedom. Seillier and Dawid (1993) recently have shown under very weak conditions (not requiring independence) on the

⁵ Historically, the number of sub intervals that has to be included in calculating a chi-square test has always been a debate amongst researchers. One of main reasons for this being the influence on the power of the test by the number of sub intervals that one chooses in calculating the *chi-squared* test. Seiller and Dawid (1993) use 11 sub intervals and their justification for that is rather simple, thus, “all forecasts were given to one decimal place, thus dividing the unit interval into 11 ranges”, However, Mann and Wald (1942) and Williams (1950) suggest a formula to come

up with an optimum number of sub intervals as follows: if number of sub intervals in denoted by J , $J = 4 * \sqrt[5]{\frac{2(N-1)^2}{c^2}}$ where N is the total number of observations, and c is the probability of the critical region under the null hypothesis assuming a standard normal distribution,

i.e. $c = \int_c^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} ds$. Furthermore, a similar formula to Mann and Wald (1942) is arrived at by Schorr (1974) using an alternative

distance norm (see Schorr (1974) page 358 for Mann and Wald (1942) distance norm and page 359 for Schorr (1974) distance norm). Nevertheless, Hamdan (1963) states that Mann and Wald (1942) procedure gives too many class intervals and that reduces the power of the *chi*-square test. Therefore, Hadman (1963) argues that optimum number of class intervals that one can take is about 10 to 20 to maintain a high power of the test. In our analysis of testing for calibration of probabilities generated through qualitative choice models using the *chi*-square test, we use 11 equally distributed class intervals (uniformly distributed class intervals) within the unit interval. Our results are robust for class intervals less than 11 and as high as 22. Therefore, we maintain 11 uniformly distributed class intervals for our analysis.

distributions underlying the forecasts and under the null hypothesis of calibration, this test statistic is distributed *chi*-squared asymptotically (Kling and Bessler, 1989).

Calculation of the goodness-of-fit test statistic takes a slightly different path in Seillier and Dawid (1993) compared to for example Kling and Bessler (1989). We used the Seillier and Dawid (1993) approach to evaluate probabilities generated using probit and logit models for calibration. For each probability class interval, Seillier and Dawid (1993) calculated a test statistic which has properties of the asymptotic standard normal distribution irrespective of the properties of the joint distribution associated with observed and expected fractiles. It is called a *Z* statistic and is calculated as follows⁶;

$$Z_j = \frac{(a_j - n\pi_j)}{\sqrt{n\pi_j}} \text{ where } j = 1, 2, \dots, n \quad (4)$$

Define the observed relative frequency of the probabilities as ρ_j where $\rho_j = \frac{a_j}{n_j}$. For probabilities generated through qualitative choice models to be regarded as to be “empirically valid” as stated in Seillier and Dawid (1993) or well calibrated, the discrepancy between ρ_j and π_j must be tend to zero at least as sample size increases. In other words, if the observed relative frequency and expected probabilities were extended to infinity, one might demand that for the forecasts to be valid, they have to be perfectly calibrated in the limit: $\rho_j - \pi_j \rightarrow 0$ as $n \rightarrow \infty$ (Seillier and Dawid, 1993). Therefore, the Z_j statistic calculated in equation (4) is a normalized measure of discrepancy which attempts to capture deviation from perfect

⁶ In calculating the *Z* statistic, Seillier and Dawid, 1993 bring in a small correction for grouping called Sheppard’s correction (see Hald, 2001, Sheppard’s second moment correction for grouping) through a weight variable introduced to the denominator of equation 3 in the text. The weight variable *w* is calculated using the *n*, the number of forecast probabilities and π_j , the width of the probability class interval. Hence the weight, $W = n_j \pi_j (1 - \pi_j)$. According to Seillier and Dawid, 1993, the equation for

Z statistic is as follows; $Z_j = \frac{(a_j - n\pi_j)}{W_j^{1/2}}$. However, according to Ferguson (1941) and Davies and Burner (1943), use of

Sheppard’s correction may introduce a downward bias for the moments of grouped data especially if the underlying distribution of the random variable does not taper off at extreme points. In other words for Sheppard’s correction to work, the underlying distribution for the random variable concerned must taper-off to zero at extreme points. In our analysis of probabilities generated through qualitative choice models for purchase decisions of selected non-alcoholic beverages, we observe distributions that do not taper off to zero at extreme points. Therefore, we do not use the Sheppard’s correction to adjust for grouping of data in calculating the above *Z* statistic.

calibration (note that we say probabilities are perfectly calibrated if there is no discrepancy between observed relative frequency and expected probabilities). This property constitutes our null hypothesis. Our null hypothesis states that probabilities are well (perfectly) calibrated. Any statistically significant deviation from perfect calibration gives rise to imperfect calibration or over or under-calibrated scenarios.

According to Seillier and Dawid (1993), distribution of the Z_j statistic is standard normal regardless of the joint distribution between expected probabilities and observed events and under such an independence structure; we could simply examine such a test statistic. If the test statistic is too far out in the tail of the standard normal distribution, we can regard this case as evidence against perfect calibration. Under the same independence structure, we could form a “portemanteau” test statistic $X^2 = \sum_{j=1}^J Z_j^2$, which has an asymptotic *chi-squared* distribution. This

calculated number is compared with table *chi-squared* distribution values with $J - 1$ degrees of freedom. If we fail to reject the null hypothesis, our model generated probability forecasts are said to be well calibrated.

We have analyzed probabilities generated through probit and logit models (both within-sample and out-of-sample scenarios) for calibration using graphical and mathematical/statistical approaches. Graphical analysis on calibration is focused on over or under-calibration (or over and under-confident probabilities issued by the model respectively) looking at the deviation of the calibration plot away from a 45-degree perfect calibration line. Statistical analysis is performed focusing on the statistical significance of the calculated X^2 statistic which is distributed *chi-squared* with degrees of freedom $J - 1$. Notice that we have used 11 probability classes in calculating this statistic, hence the degrees of freedom for the *chi-squared* test is 10. The critical, $\alpha = 0.05$, level *chi-squared* value to test the null hypothesis is 18.31. Our null hypothesis is “*issued probabilities are well calibrated*”.

For brevity, we show only the calibration graphs for probabilities generated through the logit model (for both with-in sample and out-of sample scenarios) (see Figure 3). Calculated *chi-squared* test statistics for calibration of probabilities generated through logit model are depicted in Table 4.

For isotonics, model issued probabilities are consistently over calibrated for probabilities up to about 0.6 and beyond that, probabilities are under calibrated for the with-in sample scenario. For out-of-sample probabilities we observe consistently over-calibrated probabilities. According to the calculated *chi-squared* statistic we observe poorly calibrated probabilities for both scenarios for isotonics.

Within-sample forecast probabilities are consistently over-confident while out-of-sample forecast probabilities show mixed results for regular soft drinks. Significance of the calculated *chi-squared* statistic testifies well calibrated probabilities. For diet soft drinks, within-sample probabilities are slightly over-confident and out-of-sample probabilities show very small under-confidence around probability 0.30 and a very small over-confidence above probability 0.40. The calculated *chi-squared* statistic is indicative of poor calibration of probabilities, overall.

For both high-fat milk and low-fat milk, within-sample forecast probabilities are slightly over-calibrated. Out-of-sample forecast probabilities show mixed results, where they are slightly under-calibrated for low probabilities and over-calibrated for high probabilities. According to the calculated *chi-squared* value probabilities are well calibrated for both types of milk. For fruit drinks and fruit juices, within-sample model generated probabilities show a slight over-calibration. Out-of-sample probabilities show mixed results, where they are under-calibrated for low probabilities and over-calibrated for higher probabilities. Calculated *chi-squared* statistic show well calibrated probabilities, overall.

For bottled water, calibration graphs generated for within-sample forecast probabilities show slight under calibration for low probabilities and consistent small over calibration for higher probabilities. According to the *chi-squared* test, forecast probabilities generated for within-sample data are well calibrated. However, probabilities generated out-of sample are not well-calibrated.

For coffee, calibration graphs generated for within-sample forecast probabilities show mixed results indicating a slight under-confidence for probabilities below forecast probability 0.4 and a small over-confidence in forecast probabilities above 0.4 probability level. According to the *chi-squared* test, within-sample forecast probabilities are well calibrated. Calibration

curves for out-of-sample forecast probabilities show consistent under-confident forecast probabilities. According to the *chi-squared* statistic, these probabilities are not well calibrated. For tea, within-sample generated forecast probabilities show consistent over-confidence, while out-of-sample generated forecast probabilities show some under-confidence for forecast probabilities below 0.50 and over-confidence for forecast probabilities above 0.50. According to *chi-squared* test, forecast probabilities are well-calibrated for within-sample as well as out-of-sample data for tea.

Probability Resolution and Resolution Graphs (Covariance Graphs)

Resolution is a metric of goodness of sorting power of a forecasting model. In our work, this issue corresponds to the ability of the model to sort probabilities into two classes, such as probabilities associated with events that occurred versus probabilities associated with events that did not occur. Say for example our model is designed to generate probabilities associated with an event that occurs (probability of purchase of a given non-alcoholic beverage). We would like to see high probabilities associated with the events that occurred (in our study high probabilities should be associated with all those events where a purchase of a given non-alcoholic beverage occurred) and low probabilities associated with all those events that did not occur (in our study low probabilities must be associated with all those events where a purchase of a given non-alcoholic beverage did not occur). Furthermore, for a perfect sorting model, we would like to see probability very close to 1 associated with all those events that occur and probability very close to 0 associated with all those events that do not occur. In other words, according to Yates (1982), events that are assigned probabilities close to 1 occur frequently, whereas those assigned probabilities near 0 occur rarely.

This information can be used to plot a *resolution graph (covariance graph)* where probabilities are plotted in y-axis and outcome index is on x-axis (outcome index is a zero (0) one (1) type index where zero is associated with an event that did not occur and one is associated with an event that did occur)⁷. Our method first will plot a resolution graph and

⁷ However, it is imperative to understand that well calibration does not necessarily mean good resolution or sorting power (Dawid, 1986). Dawid (1986) further states that it is unreasonable in general to expect for perfect sorting, because, perfect sorting is equivalent to absolutely correct or absolutely incorrect categorical forecasting.

then regress forecast probabilities on an outcome index to see the statistical validity of the resolution graph.

In our resolution regression (also called covariance regression in the extant literature), we would like to see intercept terms that are statistically not different from zero and slope coefficients that are statistically not different from one. This finding will correspond with perfect resolution. Any deviation of slope from one and intercept from zero would be characterized by poorly resolved probabilities. In explaining the goodness of sorting of probabilities, we concentrate on the mean values of those forecast probabilities associated with outcome index zero and one. Dispersion (variance) of those forecast probabilities also are used in the analysis.

Figure 2 shows resolution graphs for out-of-sample probability forecasts for each beverage category⁸. According to these graphs, for isotonics, outcome index zero⁹ is modestly associated with low probabilities, even though it shows a large dispersion. The mean forecast probability associated with zero outcome index is about 0.20, which is generally speaking low enough to say that we observe a good sorting behavior for forecast probabilities that are associated with zero outcome index. However, mean of the forecast probability that is associated with outcome index 1 is about 0.29. We would like it to be high (close to one) if we were to observe good probability sorting behavior.

For all other non-alcoholic beverages considered in this study, we observe high probabilities associated with outcome indexes both zero and one. It should be noted that we expected to have high probabilities associated with outcome index one (event where a purchase of a beverage occurred). To support that contention we observe relatively high mean probability associated with outcome index one. This observation is a positive result where the model sorts forecast probabilities associated with events that have outcome index one more correctly. However, the models do not sort forecast probabilities associated with outcome index of zero well for all non-alcoholic beverages but isotonics. Overall, resolution graphs for all

⁸ Resolution graphs for within-sample forecast probabilities were found to be very similar to those plotted using out-of-sample probabilities, hence not reported in the paper. They are available from authors upon request.

⁹ Outcome index zero is associated with events that did not occur (did not purchase a given non-alcoholic beverage)

non-alcoholic beverages considered are upward sloping; nevertheless they are relatively flat compared to the 45 degree perfect sorting line.

Resolution regression results are depicted in the Table 5. The resolution graphs are supported by the covariance regressions. Intercept coefficients of covariance regressions show the mean probability value associated with events with outcome index zero and they are statistically significant at the 5% level for all beverages considered. Also, the calculated slope coefficients are significantly different from one indicating poor sorting of probabilities. Overall, this result would reject the null hypothesis of perfect sorting of probabilities¹⁰.

The Brier Score and the Yates Partition of the Brier Score

The following discussion on the Brier score (BS) and the Yates partition of the Brier score follows from Brier (1950), Yates (1982), Yates and Curley (1985) and Yates (1988).

Let f represent the probabilistic forecast for an event that the forecaster is trying to predict (in our analysis, probabilities are generated using qualitative choice models). Let d represent the outcome index where, $d = 1$ if the event occurs and $d = 0$ if the event does not occur. As shown in equation (5), the probability score (PS) is formally defined as the squared difference between f and d .

$$PS(f, d) = (f - d)^2 \quad (5)$$

The PS is bounded $0 \leq PS \leq 1$. Over N occasions, indexed by $i = 1, \dots, N$, the mean of the PS (or \overline{PS} or the Brier score) is given by

$$\overline{PS}(f, d) = \frac{1}{N} \sum_{i=1}^N (f_i - d_i)^2 \quad (6)$$

Sanders (1963) and Murphy (1972a, 1972b, 1973) have decomposed the Brier score into various components including measures of calibration and resolution. However, Yates (1982), Yates and Curley (1985), and Yates (1988) further decomposed the Brier score into its variance and covariance components allowing for additional analysis. His formulation called “*covariance decomposition*” is given as follows.

¹⁰ Covariance regressions have very low R-squared values (as low as 0.05 for fruit juices and as high as 0.16 for coffee). This poor fit also is indicative of poor sorting of probabilities generated through discrete choice models.

$$\overline{PS}(f, d) = \text{Var}(d) + \text{MinVar}(f) + \text{Scat}(f) + \text{Bias}^2 - 2 * \text{Cov}(f, d) \quad (7)$$

The various components of \overline{PS} on the right hand side of equation (7) have following definitions and interpretations. $\text{Var}(d)$ represents the variance of the outcome index and defined as:

$$\text{Var}(d) = \bar{d}(1 - \bar{d}) \quad (8)$$

where

$$\bar{d} = \frac{1}{N} \sum_{i=1}^N d_i \quad (9)$$

Equation (8) shows the relative frequency or the “base rate” with which the target event occurs, where the target event for our analysis would be the decision to *buy* a given non-alcoholic beverage. This decision is completely out of control of the forecaster (in our analysis the forecaster is the qualitative choice model), hence the $\text{Var}(d)$ is not determined through our model. The remaining terms reflect the factors that are under the model’s control. Thus we want to minimize, $\text{Scat}(f)$ and Bias^2 , while maximizing $\text{Cov}(f, d)$ for an allowable minimum variance ($\text{MinVar}(f)$) to obtain the lowest \overline{PS} . It should be noted that our objective is to minimize the \overline{PS} in evaluating probabilities, because the lower the Brier score, the higher the ability of the model to correctly classify probabilities.

Bias is defined as follows.

$$\text{Bias} = (\bar{f} - \bar{d}) \quad (10)$$

where

$$\bar{f} = \frac{1}{N} \sum_{i=1}^N f_i \quad (11)$$

In the equation (10), \bar{f} is the mean of the probabilities generated from the model. Bias reflects the overall miscalibration of the forecast. The square of the bias, which is what actually appears in the covariance decomposition (equation 7), reflects the calibration error regardless of the direction (+ or -) of the error.

The $\text{Cov}(f, d)$ term is defined as follows.

$$\text{Cov}(f, d) = [\text{slope}][\text{Var}(d)] \quad (12)$$

The slope is defined as the difference between the means of conditional probability of events that actually occurred and conditional probability of events that actually did not occur.

Algebraically the slope is defined as follows.

$$\text{Slope} = (\bar{f}_1 - \bar{f}_0) \quad (13)$$

where

$$\bar{f}_1 = \frac{1}{N_1} \sum_{j=1}^{N_1} f_{1j} \quad (14)$$

$$\bar{f}_0 = \frac{1}{N_0} \sum_{j=1}^{N_0} f_{0j} \quad (15)$$

Here \bar{f}_1 represents the conditional mean probability forecast for event under consideration over the N_1 occurrences for which the event actually occurs; \bar{f}_0 represents the conditional mean probability for event under consideration over the N_0 occurrences that the event does not occur, with $N = N_1 + N_0$. The maximum value that Slope can have is 1, which occurs when the model always reports $f = 1$ when the target event is going to occur and $f = 0$ when it is not. Furthermore, Slope is the gradient of the regression line when probabilities generated through the model are regressed on outcome indexes. For a perfect forecast, all the probabilities associated with events that do not occur must have probabilities equal to zero and all probabilities associated with events that did occur must have probabilities equal to one, resulting in a slope equal to one. Therefore, it makes sense for Slope to contribute to mean probability score negatively. In other words, steeper the Slope, the more appropriate the classification of probabilities for events that occurred and that did not occur (high probabilities for event that occurred and lower probabilities for events that did not occur, the smaller the Brier score the better).

Covariance between the probabilities generated through the model and outcome index $Cov(f, d)$ is the heart of the forecasting problem (Yates, 1988). It reflects the ability of the model to make distinctions between individual occasions in which the event occurs or does not occur. In other words, it represents how responsive the forecast is to information related to the

event. Our objective with respect to minimum variance is that the model needs to maximize the value associated with the $Cov(f, d)$ to achieve a lower Brier score.

Scatter is defined as the mean of the weighted variances of probabilities associated with events that occurred and that did not occur. The algebraic representation of scatter is depicted in equation (16) below.

$$\text{Scat}(f) = \frac{1}{N} [N_1 \text{Var}(f_1) + N_0 \text{Var}(f_0)] \quad (16)$$

where

$$\text{Var}(f_1) = \frac{1}{N_1} \sum_{i=1}^{N_1} (f_{1i} - \bar{f}_1)^2 \quad (17)$$

and

$$\text{Var}(f_0) = \frac{1}{N_0} \sum_{i=1}^{N_0} (f_{0i} - \bar{f}_0)^2 \quad (18)$$

$\text{Var}(f_1)$ is the conditional variance of the probabilities generated from the model associated with the events on those N_1 occasions when the event actually occurred and $\text{Var}(f_0)$ is the conditional variance of the probabilities generated from the model associated with the events on those N_0 occasions when the event actually did not occur. $\text{Var}(f_1)$ and $\text{Var}(f_0)$ measure variability in model generated probabilities which is unrelated to whether or not the target event occurs. Scatter can be interpreted as an index of overall noise contained in model generated probabilities. It is expected that the Scatter will be minimized to achieve a lower mean probability score.

$\text{MinVar}(f)$ is defined as follows.

$$\text{MinVar}(f) = \text{Var}(f) - \text{Scat}(f) \quad (19)$$

where $\text{Var}(f)$ is the variance of the entire collection of probabilities generated for the target event. Minimum variance can also be shown as follows.

$$\text{MinVar}(f) = (\bar{f}_1 - \bar{f}_0)^2 [\bar{d}(1 - \bar{d})] \quad (20)$$

which contains the elements of the covariance of judgments and outcome indexes (Yates, 1988). To give more perspective to the relationship between minimum variance and overall

variance of the probabilities generated through the models, we can rearrange equation (19) as follows.

$$\text{Var}(f) = \text{MinVar}(f) + \text{Scat}(f) \quad (21)$$

Minimum variance can also be defined as the variance of probabilities on top of scatter that contributes toward the overall variance, i.e. $\text{Var}(f)$.

Since $\text{Var}(f)$ contributes to the Brier score positively, one would want to minimize it. That is to say, in the equation (21), we have to minimize the components in the right hand side, i.e. $\text{MinVar}(f)$ and $\text{Scat}(f)$. It would make sense to minimize $\text{Scat}(f)$ of probabilities as lower the $\text{Scat}(f)$ the tighter the distribution of probabilities around conditional means of probabilities for events that actually occurred and events that did not occur the better the model's ability to sort probabilities for events that occurred versus events that did not occur. However, it would not make sense to minimize the $\text{MinVar}(f)$ in trying to minimize the overall variance of the probabilities generated. This is clear when one looks at the equation (20). $\text{MinVar}(f)$ is a function of Slope and variance of index variable, where the latter is not determined through the model that we used to generate probabilities. The only manipulatable component is the Slope, which is a function of conditional probabilities. What is desired is to have a maximum slope of one at the extreme in minimizing the Brier score. However, in trying to minimize the $\text{Var}(f)$, if one minimizes the $\text{MinVar}(f)$, it will eliminate the Slope, which is not desirable. Therefore, we need to have some Slope, hence some $\text{MinVar}(f)$ in the model, in minimizing $\text{Var}(f)$ and trying to achieve the minimum Brier score. Therefore, $\text{MinVar}(f)$ essentially reflects the *maximum allowable model variability* (or amount of model variability that must be tolerated) which is required to minimize the $\text{Var}(f)$, hence the Brier score.

Since $\text{Cov}(f, d)$ and $\text{MinVar}(f)$ are both functions of Slope, $(\bar{f}_1 - \bar{f}_0)$ and Variance of outcome index, $(\bar{d}(1 - \bar{d}))$, we can establish a relationship between $\text{Cov}(f, d)$ and $\text{MinVar}(f)$ as follows. Equation (20) can be rearranged to represent the Slope as follows;

$$(\bar{f}_1 - \bar{f}_0) = \sqrt{\frac{\text{MinVar}(f)}{\text{Var}(d)}} \quad (22)$$

Substituting (22) into (12) and after simplification we arrive at the following relationship that combines Covariance of forecast probabilities, Minimum Variance and Variance of outcome index as follows.

$$\text{Cov}(f, d) = \sqrt{\text{MinVar}(f) * \text{Var}(d)} \quad (23)$$

According to equation (23), variance of outcome index and Minimum Variance are positively related to the covariance of forecast probabilities and outcome index. It is an obvious fact that variance of the outcome index, $\text{Var}(d)$ is beyond the control of the forecasting model and only determined externally by the actual observations. Therefore, the only model generated variable that affect the $\text{Cov}(f, d)$ is $\text{MinVar}(f)$. We can conclude that higher the Slope, the higher the $\text{MinVar}(f)$, the higher the $\text{Cov}(f, d)$. In other words, high $\text{MinVar}(f)$ is associated with high $\text{Cov}(f, d)$. This result has leverage in explaining the forecasting model's sorting power (resolution) and $\text{Cov}(f, d)$. We also can conclude that, high resolution is associated with high $\text{Cov}(f, d)$.

It is important to note that, although the Brier score gives an overall indication of the ability of the model to forecast (the lower the Brier score, the better the forecast), the components of the covariance decomposition of the Brier score provides a clearer indication of the ability of the model to forecast as well as to sort probabilities.

Tables 6 and 7 show the Brier score and covariance decomposition of the Brier score for forecast probabilities generated using logit model. We have generated both within-sample and out-of-sample forecasts and evaluated them using the Brier score and the Yates partition of the Brier score.

The Brier Score

Fruit juices show the lowest Brier score (0.06 for within-sample estimates and it is 0.05 for out-of-sample estimates). The Brier score associated with low-fat milk is 0.22 and 0.21 for within-sample and out-of-sample forecasts respectively. Notice that the out-of-sample Brier score is lower than the within-sample value. One may sometimes erroneously conclude that out-of-sample forecasts are better, because they are associated with a low Brier score value. However, one must remember that the Brier score value can be decomposed into its covariance parts, which would provide a better explanation to the realized Brier score. Other non-alcoholic

beverages have varying values of the Brier score depending on the forecast probabilities and outcome index values observed for each observation.

Even though the Brier score provides a simple yet rigorous number to compare forecast probabilities generated through alternative models, it does not tell anything about the calibration or resolution property of forecast probabilities. However, it is a good measure independent of cut-off probability values in sorting probabilities which were used in expectation-prediction success tables.

Variance of the Outcome Index (*DVar*)

Variance of the outcome index is a measure that cannot be controlled through the model under consideration. It is determined through the behavior of the agent (purchasing behavior in our study). Market penetration value for a given non-alcoholic beverage or the number of individuals that actually purchased a non-alcoholic beverage has a direct leverage on the variance of the outcome index. Figure 3 shows the plot of market penetration value against the variance of the outcome index. According that, highest variance value of the outcome index (0.25) could be observed for the market penetration value 0.50. Any other market penetration value is associated with the variance value less than 0.25. In our study, fruit juices have the highest market penetration value, which is 0.93. It is associated with the variance of outcome index 0.0651, which is the lowest variance of the outcome index reported. Highest variance of the outcome index is 0.23 which is reported for low-fat-milk and it is associated with a market penetration value 0.63 (close to 0.50). Therefore, the market penetration value which is outside the control of the forecasting model has a direct impact on the value of the calculated variance of the outcome index (*DVar*). Since variance of the outcome index is a component of the covariance decomposition of the Brier score, it has a direct influence on the calculated Brier score. Therefore, a highly inflated Brier score value may be a result of a contribution coming from a large variance of the outcome index. In our study, the highest Brier score value is reported for low-fat milk and it also has the highest variance value of the outcome index exhibiting the large contribution of the variance of the outcome index toward the Brier score.

Minimum Variance and Scatter

Unlike the variance of the outcome index, variance of the forecast probabilities i.e. $Var(f)$ is something that the forecast model has control of. We would like to have small $Var(f)$ to be associated with a good probability forecast, hence lower scatter. It was made clear earlier that the Minimum Variance is the variability that is tolerated to have a positive slope of the covariance graph while minimizing scatter, then in turn minimizing $Var(f)$.

The highest Scatter is associated with coffee within-sample forecasts, which is 0.027. Coffee out-of-sample forecasts show slightly high Scatter (0.0283) compared to that of within-sample forecasts, indicating more spread of the forecast probabilities around their mean values. We observe the lowest Scatter with forecast probabilities associated with fruit juices within-sample estimates, which is recorded at 0.0029.

Minimum Variance has a direct relationship with the Slope (defined as $\bar{f}_1 - \bar{f}_0$) where higher slope is associated with high Minimum Variance. The highest Slope is observed with respect to forecast probabilities associated with coffee, which is 0.16, hence largest Minimum Variance 0.0051 for within-sample forecasts. For out-of-sample forecasts we observe a low Minimum Variance, 0.0040, hence lower Slope (0.14) compared to within-sample forecasts. All other non-alcoholic beverages showed very small Minimum Variance values, hence very small slope indicating more flat covariance graphs.

Bias

Bias is the ability of the model to match mean forecasts to relative frequencies. The model has to minimize the Bias in evaluating forecast probabilities. It is clear from Tables 4 and 5 that the Bias associated with the covariance decomposition is very small (almost negligible) compared to other part of the covariance decomposition. It must be emphasized that for all non-alcoholic beverages considered, the Bias associated with out-of-sample forecasts are relatively larger than those of within-sample forecasts. This is indicative of presence of more mis-calibration with respect to out-of-sample generated probability forecasts compared to those generated within-sample. Overall, we should emphasize that the contribution of the Bias toward the Brier score is very small in our analysis.

Covariance of Forecast Probabilities and Outcome Index (2cov)

Covariance of forecast probabilities and outcome index is the most important part of the forecasting property of a model. Covariance enters negatively to the Yates partition of the probability score; hence in order to get a low Brier score, we need to maximize the value associated with covariance.

Highest covariance value is associated with coffee within-sample forecasts. The covariance value obtained from out-of-sample forecast probabilities is slightly lower than that of within-sample counterpart, indicating better forecasts obtained from within-sample forecasts compared to out-of-sample forecasts. Notice that if one considered the covariance of forecast probabilities and outcome index to comment on the forecasting ability of a model, probability forecasts associated with coffee outperforms forecasts for other beverages. However, coffee has a higher Brier score compared to other beverages. On the other hand, fruit juices not only have the lowest Brier score but also the lowest covariance of forecast probabilities and outcome index. Even though the low Brier score indicates better forecasting ability, low covariance of forecast probabilities and outcome index is an indication of poor forecasting performance. We also find relatively higher covariance values associated with fruit drinks, diet soft drinks and bottled water for both within-sample and out-of-sample forecasts even though they were not necessarily associated with low Brier scores.

Therefore, the use of just the Brier score to comment on the goodness of the probability forecasts can be misleading because the results may be different if one had partitioned the Brier score into its covariance components. The use of this decomposition introduces more accuracy to forecast evaluation and therefore improved decision making.

According to equation (22), we observe the relationship with $Cov(f, d)$ and $MinVar(f)$ for forecast probabilities generated for decision to purchase non-alcoholic beverages. Coffee has the highest $Cov(f, d)$ and highest $MinVar(f)$. It also has the highest calculated Slope (0.16). On the other hand, fruit juices have the lowest $Cov(f, d)$ and lowest $MinVar(f)$ hence the lowest Slope (0.04). Therefore we can conclude that, in terms of the Yate's partition of the Brier score, models do an excellent job in generating probability forecasts with respect to coffee and do a very poor job in generating probability forecasts for fruit juices. Probability

forecasts generated for other non-alcoholic beverages lie somewhere in-between the probability forecasts generated for coffee and fruit juices. Despite the fact that Yate's partition of the Brier score does an exceptional job in evaluating probability forecasts, we are not in position to test the numbers statistically, because sampling distributions of these decompositions are yet to be derived.

Summary of Key Findings

In using expectation-prediction success tables and a desired cut-off probability level to correctly classify probabilities, we paid attention to *sensitivity* and *specificity* values. Use of naïve 0.50 cut-off value to classify probabilities resulted in over- or under-estimated *sensitivity* and *specificity* values for all models compared to the use of market penetration value as cut-off probabilities.

Receiver Operating Characteristics (ROC) charts show the evaluation of probabilities over wide range on cut-off probabilities (including 0.05) in determining the model that provide the best probability forecasts. The model with the highest calculated area under the ROC chart provides evidence for the best model in terms of generated probabilities for events that occurred vis-à-vis that did not. For with-in-sample probabilities, logit model associated with coffee shows the highest area under the ROC curve, followed by fruit juices. The lowest area under the ROC curve for this case is associated the probabilities generated for low-fat milk and tea. The highest area under the ROC curve for out-of-sample probabilities is associated with logit model generated probabilities for fruit juices followed by coffee.

Next we used Kullback-Leibler Information Criteria which selects the model with the highest log-likelihood function value observed at out-of-sample observations (OSLLF) to evaluate probabilities (and the best model). According to OSLLF approach, we tested for "closeness" or deviation of model generated probabilities to the true data generating probability distribution using calculated log-likelihood function value for each model. According to this criterion, logit model associated with fruit juices gave rise to probabilities that moved more closely with true data generating process (true probability distribution) than did probabilities generated from other logit models.

Next we used calibration (graphs and chi-squared statistical test) to evaluate probabilities. According to the calculated chi-squared statistics of forecast probabilities, probabilities generated for most of beverage purchases were well calibrated both with-in- and out-of-sample. However, calibration graphs show varying degree of over and under calibrated probabilities for all models

Next, we used resolution graphs and covariance regressions to evaluate forecast probabilities. All resolution graphs were almost flat against a 45-degree perfect resolution graph. This result is indicative of poor sorting of forecast probabilities all models. According to covariance regressions, we found that for all non-alcoholic beverages, the intercept coefficient was statistically different from zero and the slope parameter was statistically different from one, indicating weak sorting power of probabilities generated through choice models.

Finally, we investigated the forecast probabilities generated through choice models using the Brier score and the Yates partition of the Brier score. We expected to have a low Brier score for well issued forecast probabilities. In the Yates partition of the Brier score, we expected to have a smaller scatter and a bias with a minimally allowed minimum variance. More importantly, we expected to have a high Covariance associated with forecast probabilities and outcome index. The Brier score was lowest for fruit juices and the highest for low-fat milk. According to the calculated Brier score, probability forecasts for fruit juices outperformed other non-alcoholic beverages.

Although the Brier score gave an overall indication of the ability of a model to forecast accurately, the components of the covariance decomposition of the Brier score provided a clearer and broader indication of the ability of the model to forecast. Highest variance of the outcome index was associated with low-fat milk and also low-fat milk had the highest Brier score. This inflated Brier score was primarily due to the large variance of the outcome index which has a direct relationship with the market penetration value for a given beverage. Bias was almost negligible for all forecast probabilities associated with all non-alcoholic beverages. Scatter and minimum variance directly contributed to the variance of the forecast probabilities. The lowest scatter was associated with fruit juices for all scenarios, hence the lowest spread of forecast probabilities. The highest scatter was associated with coffee, hence the largest spread

of forecast probabilities. The highest minimum variance was recorded with coffee; consequently the highest slope of the resolution graph. Highest covariance of outcome index and forecast probabilities were observed for coffee. Therefore, in terms of the Yates partition of the Brier score, coffee outperforms all other beverages in issuing forecast probabilities.

Conclusions

The choice of cut-off probability level in classifying probabilities was important for all non-alcoholic beverages. The market penetration probability level as a cut-off probability value to correctly classify probabilities outperformed the naïve 0.50 cut-off probability level. Therefore, it is recommended to use market penetration as the appropriate cut-off to classify probabilities. This recommendation is consistent with the works by Park and Capps (1997) and Briggeman (2002).

Area under the ROC curve is suggested as an alternative to the use of 0.5 cut-off as well as cut-off at market penetration level to classify probabilities, because this method treats a wide range of cut-off probabilities to come up with a coherent measure, i.e. area under the ROC curve, to offer evidence for better classification of model-generated probabilities. The Kullback-Leibler Information criterion measured through OSLLF value offers important information in terms of closeness of model-generated probabilities to the true (observed) probabilities, although this measure does not offer a classification of probabilities for events occurred versus that did not.

Most calibration graphs with respect to purchase decision of non-alcoholic beverages revealed that almost always there was a certain degree of over- and under-calibration with respect to probabilities generated. However, forecast probabilities were well calibrated. Resolution regression analysis revealed that forecast probabilities generated for the decision to purchase all non-alcoholic beverages were not well resolved (or sorted). However, all resolution graphs were upward sloping, indicating some degree of sorting power in choice models. Yates decomposition of the Brier score offers rich set of measures to speak to the goodness of probabilities compared to other measures, such as bias, scatter, minimum variance, variance of outcome index and covariance between outcome index and associated probability in correctly

classifying probabilities. With-in-sample probabilities generated through logit model for coffee outperforms probabilities generated for other beverage products based on having highest area under the ROC curve, highest covariance between probabilities and outcome index (part of Yates partition of Brier Score) and highest slope of covariance graph in classifying probabilities. On the other hand, out-of-sample probabilities generated through logit model for fruit juice performs better than any other beverage category based on having highest area under the ROC curve, lowest Brier Score, and highest OSLLF value.

In the event where researchers are confronted with alternative models that issue probability forecasts, the accuracy of probability forecasts in determining the best model can be measured through myriad of metrics. Even though traditional measures such as expectation-prediction success tables, calibration and log-likelihood approaches are still used, ROC charts, resolution, the Brier score and the Yates partition of the Brier score to evaluate probabilities generated through alternative models are highly recommended.

REFERENCES

- Alviola, P. 2009. "Essays on Choice and Demand Analysis of Organic and Conventional Milk in the United States." Unpublished PhD Dissertation, Texas A&M University.
- Bessler, D.A., and R. Ruffley. 2004. "Prequential Analysis of Stock Market Returns." *Applied Economics* 36:399-412.
- Blume, J.D. 2002, "Estimation and Covariate Adjustment of Smooth ROC Curves." Center for Statistical Sciences, Brown University
- Brier, G.W. January 1950. "Varification of Forecasts Expressed in terms of Probability." *Monthly Weather Review* 78(1):1-3.
- Briggeman, B.C., 2002. "Consumption of Fresh and Processed Pork in the At-Home and Away-from-Home Markets, *Unpublished Master's Thesis*, Texas A&M University, May 2002.
- Covey, T. 1999. "Banker's Forecast of Farmland Value." Paper presented at NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management. Chicago, IL.
- David, F.N., and N.L. Johnson. 1950. "The Probability Integral Transformation When the Variable is Discontinuous." *Biometrika* 37(1/2):42-49.
- Davies, G.R., and N. Bruner. March 1943. "A Second Moment Correction for Grouping." *Journal of American Statistical Aassociation* 38(221):63-68.
- Dawid, A.P. 1984. "Present Position and Potential Developements: Some Personal Views Statistical Theory The Prequential Approach." *Journal of Royal Statistical Society* 147(2):278-292.
- . 1986. *Probability Forecasting: Encyclopedia of Statistical Sciences*, Vol 7 Wiley, New York.
- . 1982. "The Well Calibrated Bayesian." *Journal of American Statistical Aassociation* 77(379):605-610.
- Desmarais, B.A. and J. J. Harden, 2013, "Testing for Zero Inflation in Count Models: Bias Correction for the Vuong Test". *The Stata Journal*, 13(4): 810-835
- Epstein, E.S., and A.H. Murphy. April 1965. "A Note on the Attributes of Probabilistic Predictions and the Probability Score." *Journal of Applied Meteorology* 4:297-299.
- Ferguson, G.A. 1941. "The Application of Sheppard's Correction for Grouping." *Psychometrika* 6(1):21-27.
- Hald, A. 2001. "On the History of the Correction for Grouping, 1873-1922." *Scandinavian Journal of Statistics* 28:417-428.
- Hamdan, M.A. 1963. "The Number and Width of Classes in the Chi-Square Test." *Journal of American Statistical Aassociation* 58(303):678-689.
- Hsieh, F., and B.W. Turnbull. 1996. "Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve." *Annals of Statistics*, 24(1):25-40
- Kling, J.L., and D.A. Bessler. 1989. "Calibration-based Predictive Distributions: An Application of Prequential Analysis to Interest Rates, Money, Prices and Output." *Journal of Business* 62(4):477-499.
- Lahiri, K., and J.G. Wang 2005. Evaluating Probability Forecasts: Calibration Isn't Everything, Working Paper, Department of Economics, Univeristy of Albany, State Univeristy of New York.
- Mann, H.B., and A. Wald. September 1942. "On the Choice of the Number of Class Intervals in

- the Application of the Chi Square Test." *The Annals of Mathematical Statistics* 13(3):306-317.
- Murphy, A.H. January 1973. "A Sample Skill Score for Probability Forecasts." *Monthly Weather Review* 102:48-55.
- . March 1972a. "Scalar and Vector Partitions of the Probability Score: Part I. Two-State Situation." *Journal of Applied Meteorology* 11:273-282.
- . 1972b. "Scalar and Vector Partitions of the Probability Score: Part II. N-State Situation." *Journal of Applied Meteorology* 11:1183-1192.
- Murphy, A.H., and R.L. Winkler. 1977. "Reliability of Subjective Probability Forecasts of Precipitation and Temperature." *Applied Statistician* 26(1):41-47.
- Norwood, F. B., J.L. Lusk, and B. W. Brorsen. 2004. "Model Selection for Discrete Dependent Variables: Better Statistics for Better Steaks." *Journal of Agricultural and Resource Economics*, 29(3):404-419
- Parks, J. L., and Capps, O. Jr. 1997. "Demand for Prepared Meals by US Households." *American Journal of Agricultural Economics* 79: 814-824.
- Reiser, B., and D. Faraggi. 1997. "Confidence Intervals for the Generalized ROC criterion." *Biometrics*, 53:644-652
- Royston, P., 2006. "Explained Variation for Survival Models". *The Stata Journal*, 6(1): 83-96
- Sanders, F. April 1963. "On Subjective Probability Forecasting." *Journal of Applied Meteorology* 2(2):191-201.
- Schorr, B. 1974. "On the Choice of the Class Intervals in the Application of the Chi-Square Test." *Mathematische Operationsforschung und Statistik* 5(4/5):357-377.
- Seillier-Moiseiwitsch, F., and A.P. Dawid. 1993. "On Testing the Validity of Sequential Probability Forecasts." *Journal of American Statistical Association* 88(421):355-359.
- Shao, J. 1993. "Linear Model Selection by Cross-Validation." *Journal of American Statistical Association*, 88(422):486-494
- Stock, J.H., and M.W. Watson. 2007. *Introduction to Econometrics*. 2 ed: Pearson Education Inc.
- Stone, M. 1977. "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion." *Journal of Royal Statistical Society, Series B (Methodological)* 39(1):44-47
- Train, K.E. 2003. *Discrete Choice Methods with Simulation*. 1 ed: Cambridge University Press.
- Venkatraman, E.S., and C.B. Begg, 1996, "A Distribution-Free Procedure for Comparing Receiver Operating Characteristics Curves from a Paired Experiment." *Biometrika*, 83(4):835-848.
- Williams, C.A., Jr. March 1950. "On the Choice of the Number and Width of classes for the Chi-Square Test of Goodness of Fit." *Journal of American Statistical Association* 45(249):77-86.
- Yates, J.F. 1988. "Analyzing the Accuracy of Probability Judgements for Multiple Events: An Extension of the Covariance Decomposition." *Organizational Behavior and Human Performance* 41:281-299.
- . 1982. "External Correspondence: Decompositions of the Maen Probability Score." *Organizational Behavior and Human Performance* 30:132-156.
- Yates, J.F., and S.P. Curley. 1985. "Conditional Distribution Analyses of Probabilistic Forecasts." *Journal of Forecasting* 4:61-73.

Zellner, A., C. Hong, and C. Min. 1991. "Forecasting Turning Points in International Output Growth Rates Using Bayesian Exponentially Weighted Autoregression, Time-varying Parameter, and Pooling Techniques." *Journal of Econometrics* 49:275-304.

Table 1: Results from Expectation-Prediction Success Table for Logit Model Generated Probabilities

Beverage	With-in-Sample				Out-of-Sample			
	0.5 Cut-off		Market Penetration Cut-off		0.5 Cut-off		Market Penetration Cut-off	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Iostoncis	0.09	0.98	0.58	0.70	0.06	0.98	0.63	0.62
Regular Soft Drinks	1.00	0.01	0.66	0.64	1.00	0.00	0.68	0.64
Diet Soft Drinks	0.90	0.25	0.72	0.51	0.90	0.19	0.73	0.46
High-Fat Milk	1.00	0.01	0.58	0.66	1.00	0.00	0.51	0.71
Low-Fat Milk	0.89	0.27	0.66	0.55	0.87	0.27	0.55	0.67
Fruit Drinks	0.97	0.10	0.61	0.68	0.99	0.03	0.58	0.67
Fruit Juices	1.00	0.00	0.67	0.69	1.00	0.00	0.60	0.66
Bottled Water	0.97	0.12	0.66	0.56	0.98	0.07	0.66	0.54
Coffee	0.94	0.25	0.69	0.68	0.93	0.27	0.71	0.60
Tea	0.97	0.07	0.62	0.60	0.98	0.06	0.58	0.60

Table 2: Results from area under the ROC curve for probabilities generated by logit model -

Beverage	With-in-Sample	Out-of-Sample
Iostoncis	0.69	0.67
Regular Soft Drinks	0.68	0.69
Diet Soft Drinks	0.65	0.64
High-Fat Milk	0.67	0.67
Low-Fat Milk	0.63	0.66
Fruit Drinks	0.65	0.68
Fruit Juices	0.72	0.74
Bottled Water	0.63	0.63
Coffee	0.76	0.73
Tea	0.63	0.63

Table 3: Logit model generated probability evaluation using out-of-sample log-likelihood function value (OSLLF)

Beverage	Average OSLLF value
Iostoncis	-0.4698
Regular Soft Drinks	-0.2801
Diet Soft Drinks	-0.6182
High-Fat Milk	-0.4322
Low-Fat Milk	-0.6189
Fruit Drinks	-0.5070
Fruit Juices	-0.2086
Bottled Water	-0.5848
Coffee	-0.5268
Tea	-0.5623

Table 4: Chi-squared test Statistics for calibration for logit model generated probabilities

Beverage	With-in Sample	Out-of Sample
Iostoncis	45.42	86.76
Regular Soft Drinks	11.34	13.36
Diet Soft Drinks	18.35	21.21
High-Fat Milk	11.19	8.29
Low-Fat Milk	16.63	4.50
Fruit Drinks	12.62	9.68
Fruit Juices	9.12	7.22
Bottled Water	14.89	20.06
Coffee	15.09	23.50
Tea	13.09	12.37

Note: Critical value for chi-squared statistic with degrees of freedom 10 is 18.31 at alpha=0.05

Table 5: Covariance Regression of Forecast Probabilities and Outcome Indexes for logit model

Beverage		Within-sample	Out-of-sample
Isotonics	Intercept	0.2028 (0.0021)	0.2086 (0.0021)
	Slope	0.0845 (0.0045)	0.0655 (0.0048)
Regular Soft Drinks	Intercept	0.8391 (0.0038)	0.8639 (0.0032)
	Slope	0.0692 (0.0040)	0.0425 (0.0035)
Diet Soft Drinks	Intercept	0.6063 (0.0033)	0.6140 (0.0033)
	Slope	0.0714 (0.0041)	0.0542 (0.0040)
High-Fat Milk	Intercept	0.7694 (0.0034)	0.7785 (0.0034)
	Slope	0.0582 (0.0037)	0.0444 (0.0037)
Low-Fat Milk	Intercept	0.5737 (0.0030)	0.5695 (0.0031)
	Slope	0.0602 (0.0038)	0.0628 (0.0038)
Fruit Drinks	Intercept	0.6799 (0.0041)	0.7060 (0.0034)
	Slope	0.0938 (0.0048)	0.0620 (0.0039)
Fruit Juices	Intercept	0.8820 (0.0035)	0.8904 (0.0036)
	Slope	0.0522 (0.0036)	0.0446 (0.0037)
Bottled Water	Intercept	0.6570 (0.0034)	0.6715 (0.0030)
	Slope	0.0680 (0.0041)	0.0519 (0.0036)
Coffee	Intercept	0.6162 (0.0052)	0.6300 (0.0052)
	Slope	0.1629 (0.0060)	0.1410 (0.0061)
Tea	Intercept	0.6834 (0.0031)	0.6940 (0.0029)
	Slope	0.0518 (0.0036)	0.0407 (0.0034)

Note: all coefficients are significant at p-value 0.001 level

Table 6: The Brier Score and the Yates Partition of the Brier Score: Logit Within-Sample

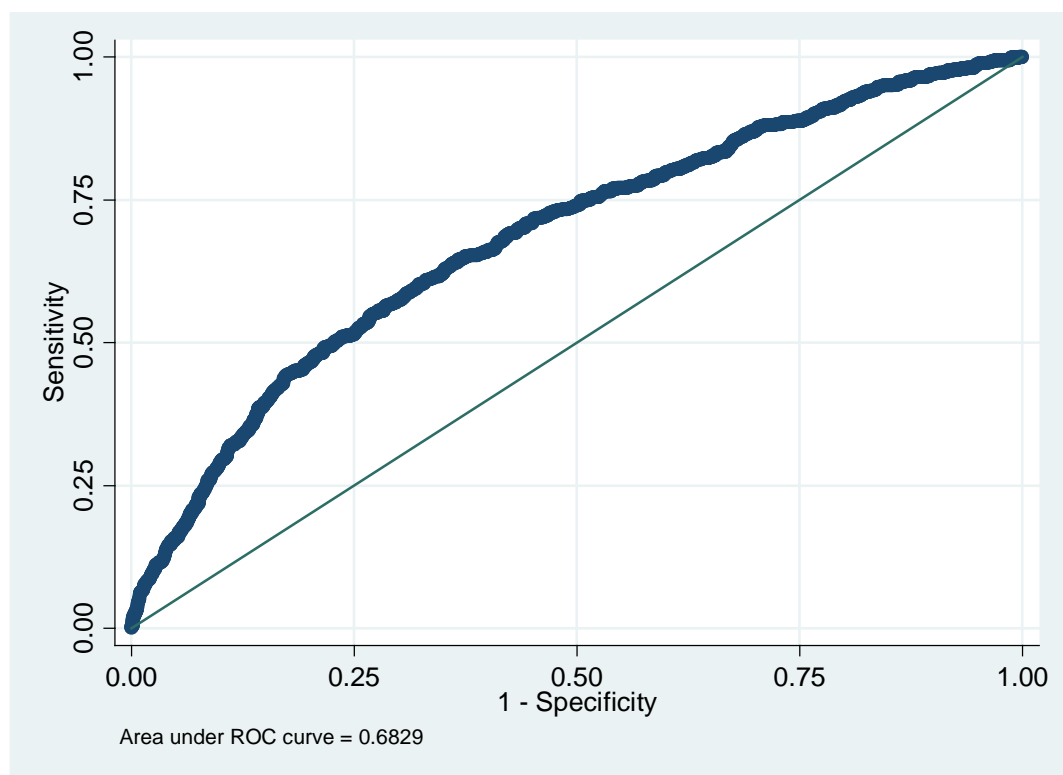
	Isotonics	Regular Soft Drinks	Diet Soft Drinks	High Fat Milk	Low Fat Milk	Fruit Drinks	Fruit Juices	Bottled Water	Coffee	Tea
Brier Score	0.1578	0.0824	0.2102	0.1407	0.2235	0.1701	0.0613	0.1938	0.1631	0.1910
Dvar	0.1724	0.0887	0.2266	0.1495	0.2378	0.1874	0.0646	0.2080	0.1942	0.2013
Min Var	0.0012	0.0004	0.0012	0.0005	0.0009	0.0017	0.0002	0.0010	0.0052	0.0005
Scatter	0.0133	0.0055	0.0148	0.0081	0.0134	0.0162	0.0033	0.0132	0.0270	0.0100
Bias	9.0E-16	0.0E+00	1.0E-16	0.0E+00	0.0E+00	1.0E-16	4.0E-16	4.0E-16	1.0E-16	0.0E+00
2Cov	0.0291	0.0123	0.0324	0.0174	0.0286	0.0351	0.0067	0.0283	0.0633	0.0208

Table 7: The Brier Score and the Yates Partition of the Brier Score: Logit Out-of-Sample

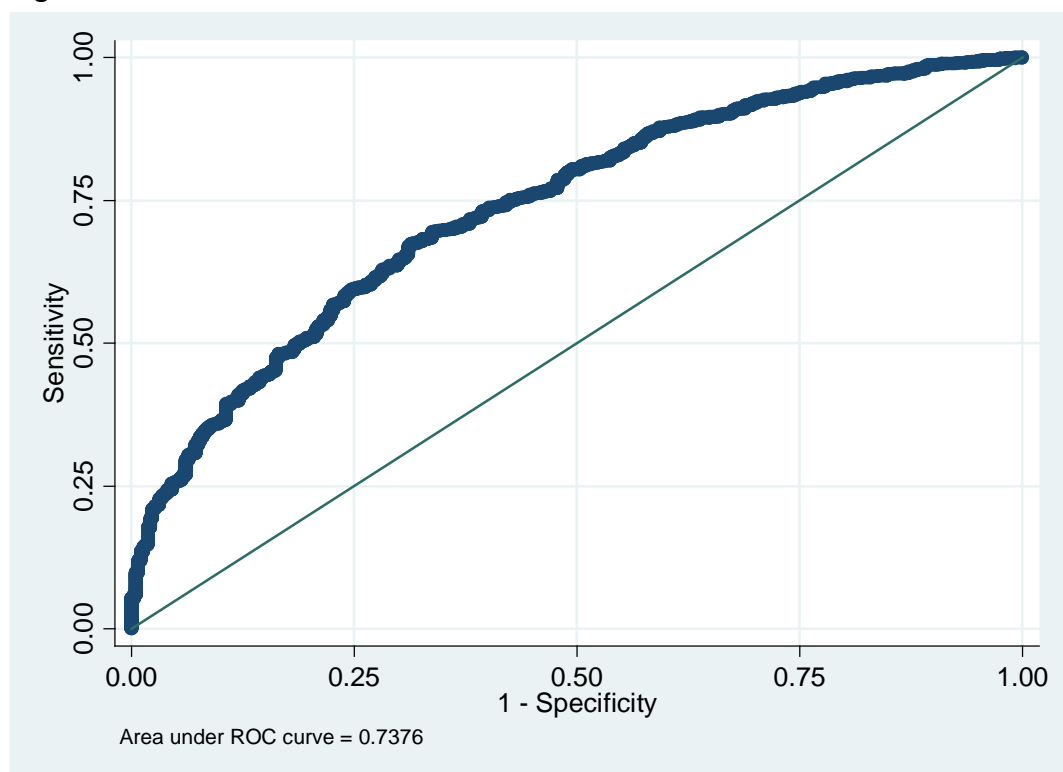
	Isotonics	Regular Soft Drinks	Diet Soft Drinks	High Fat Milk	Low Fat Milk	Fruit Drinks	Fruit Juices	Bottled Water	Coffee	Tea
Brier Score	0.1542	0.0785	0.2164	0.1377	0.2164	0.1684	0.0558	0.1982	0.1778	0.1905
Dvar	0.1599	0.0815	0.2262	0.1426	0.2307	0.1796	0.0577	0.2092	0.2020	0.1974
Min Var	0.0007	0.0001	0.0007	0.0003	0.0009	0.0007	0.0001	0.0006	0.0040	0.0003
Scatter	0.0141	0.0037	0.0140	0.0075	0.0129	0.0103	0.0030	0.0101	0.0286	0.0087
Bias	4.8E-04	6.2E-05	2.5E-05	1.4E-04	8.6E-04	1.4E-04	3.8E-05	3.4E-05	1.3E-04	3.1E-05
2Cov	0.0209	0.0069	0.0245	0.0129	0.0290	0.0223	0.0051	0.0217	0.0570	0.0161

Figure 1: Receiver Operating Characteristics Curves (ROC) for logit model generated probabilities (with-in-sample)

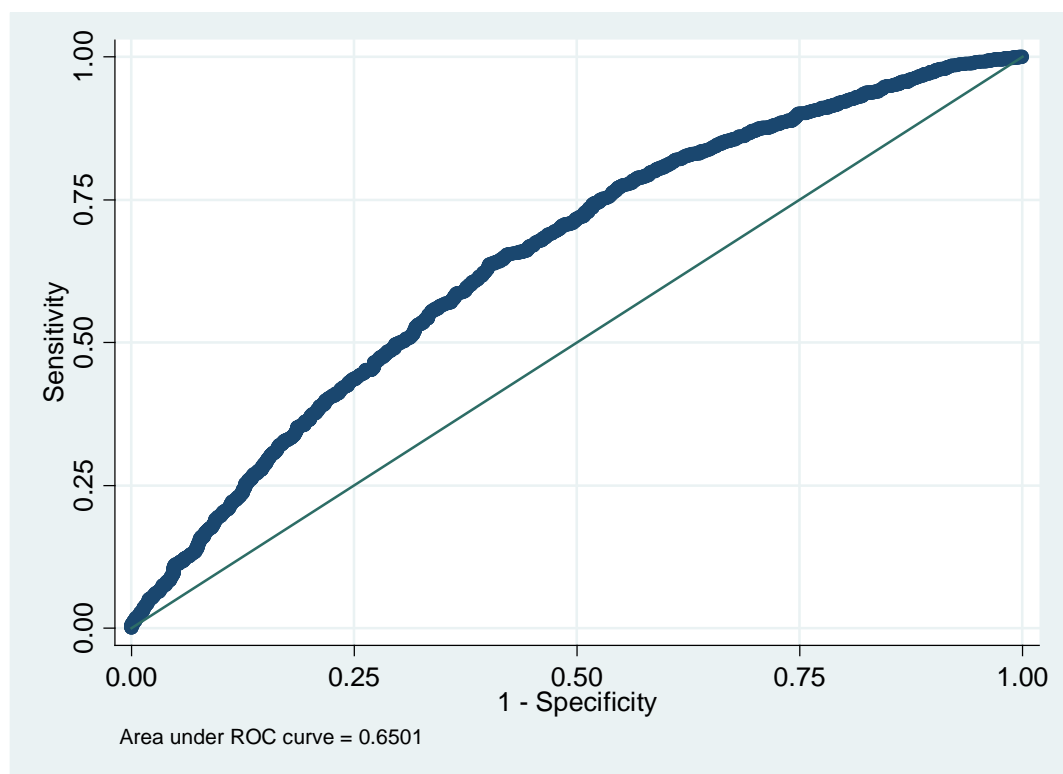
Isotonics



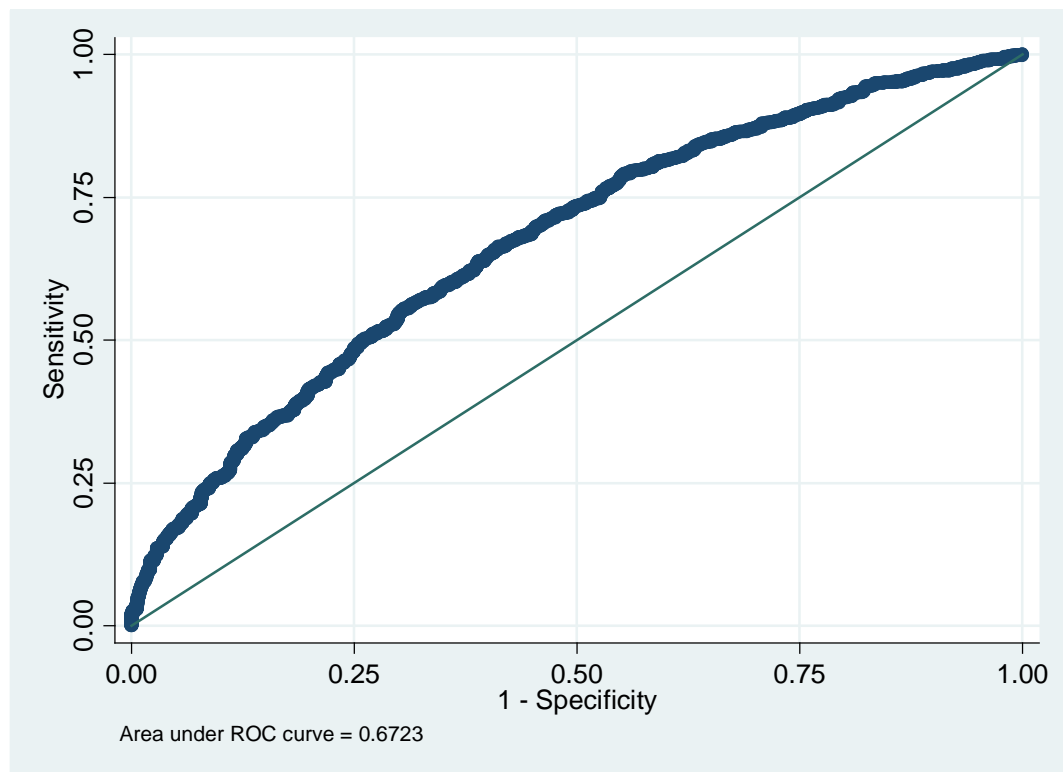
Regular Soft Drinks



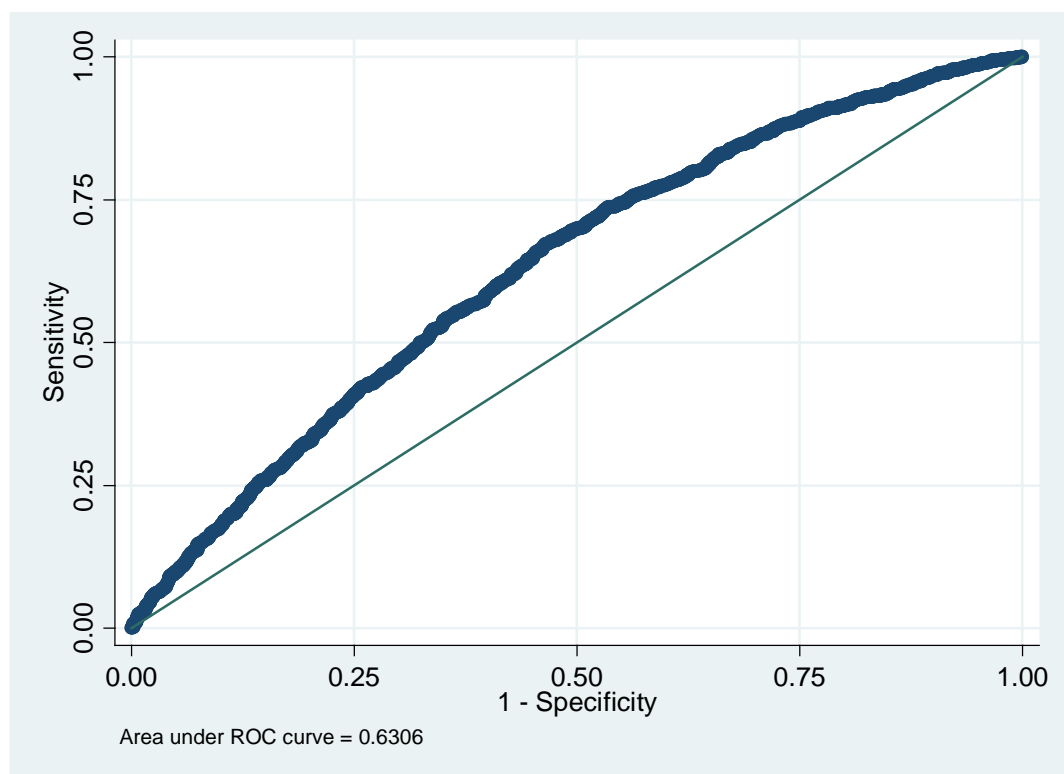
Diet Soft Drinks



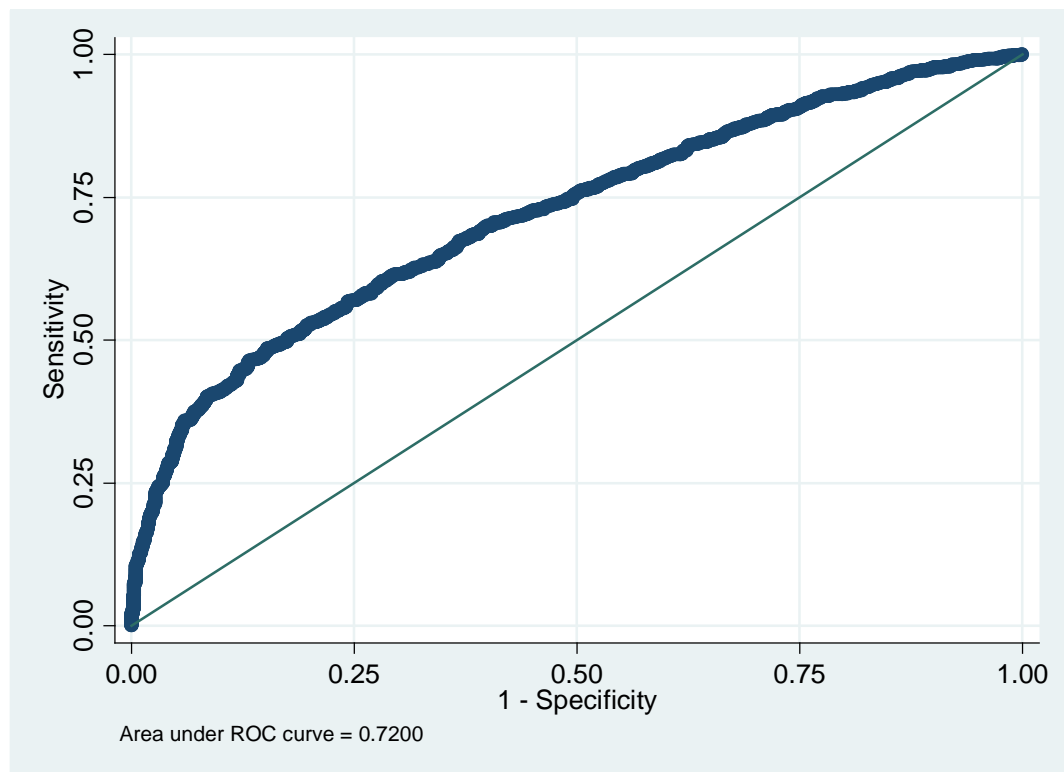
High-Fat Milk



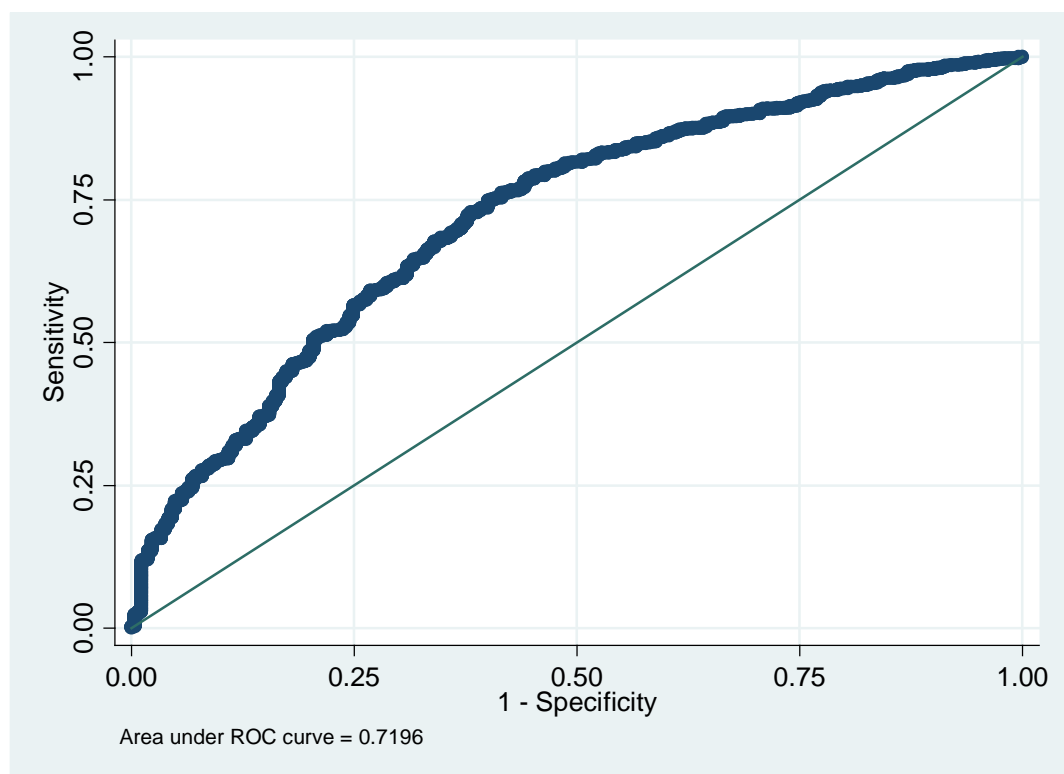
Low-Fat Milk



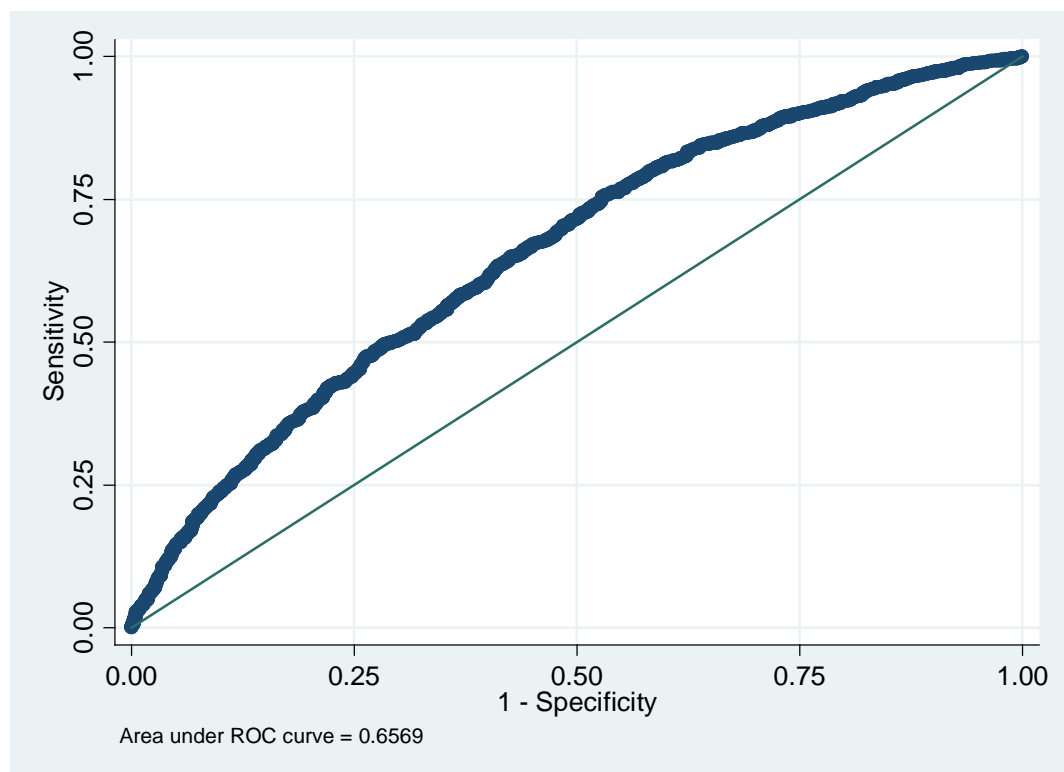
Fruit Drinks



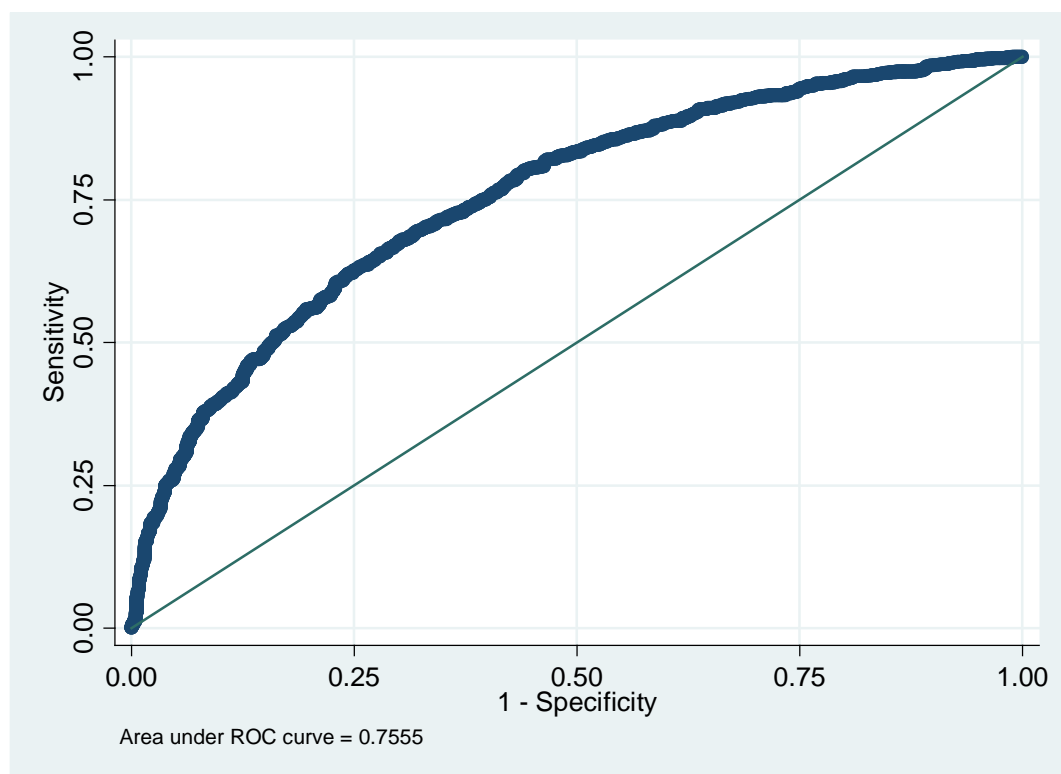
Fruit Juices



Bottled Water



Coffee



Tea

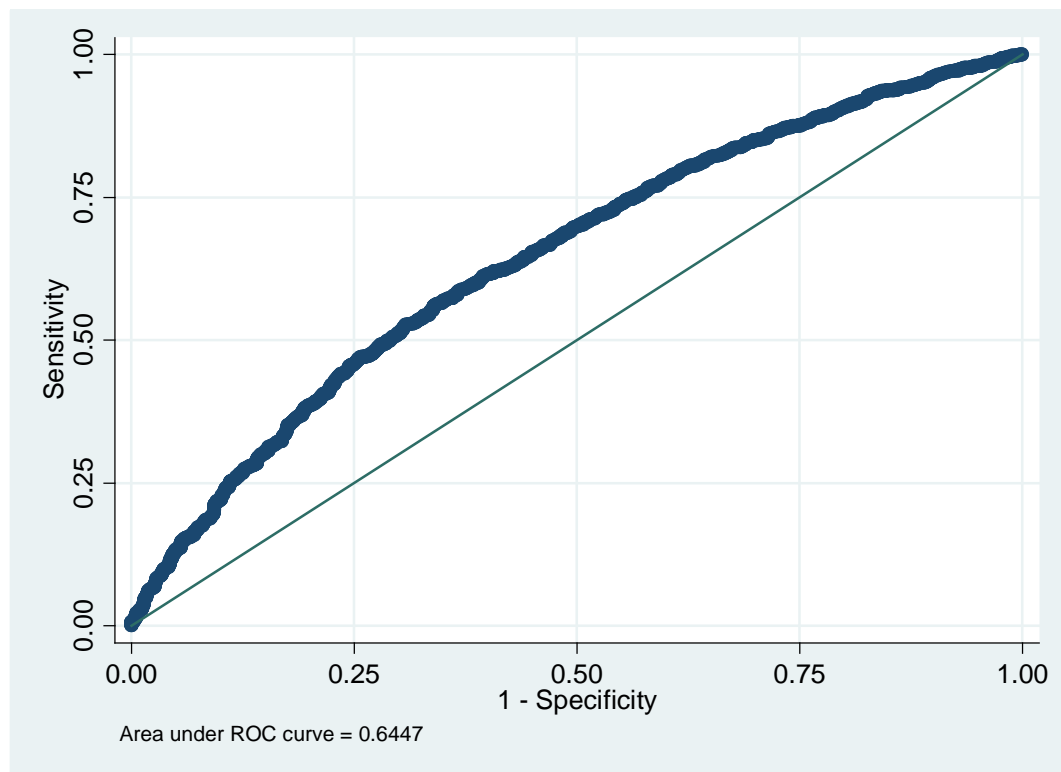
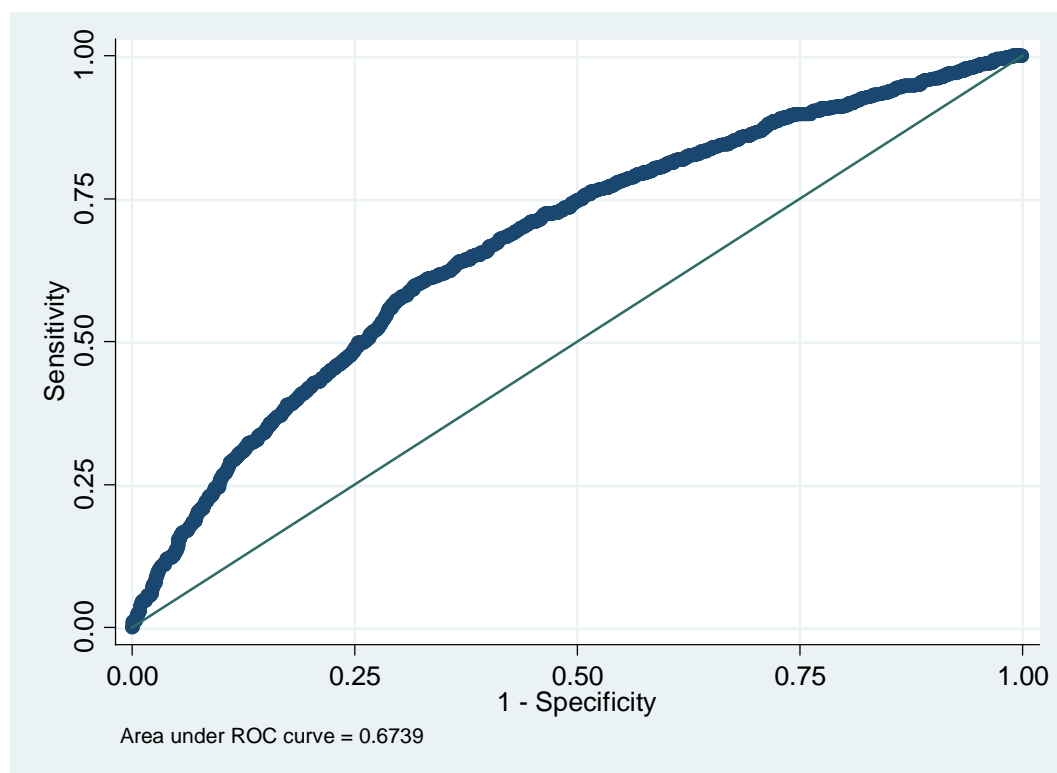
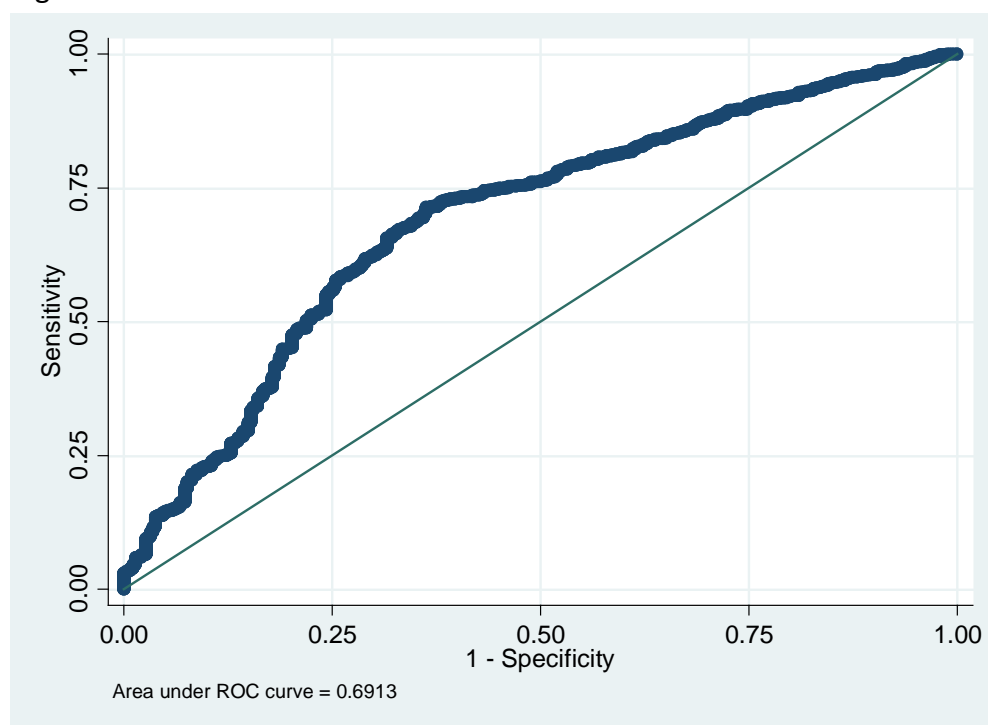


Figure 2: Receiver Operating Characteristics Curves (ROC) for logit model generated probabilities (out-of-sample)

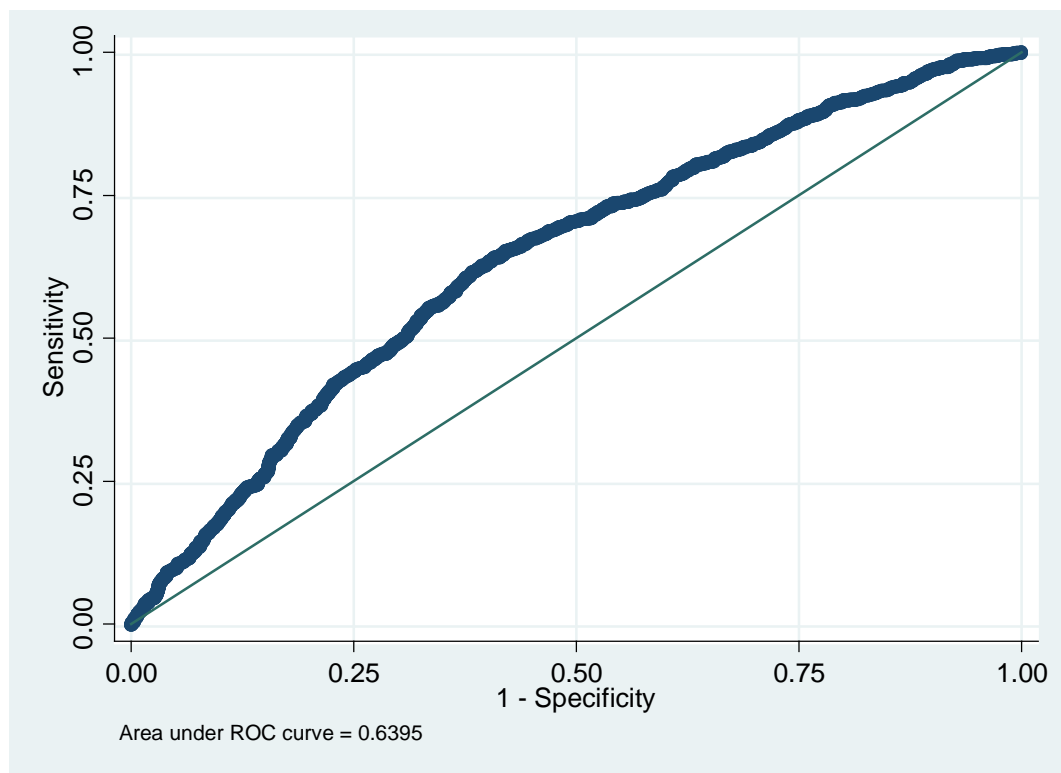
Isotonics



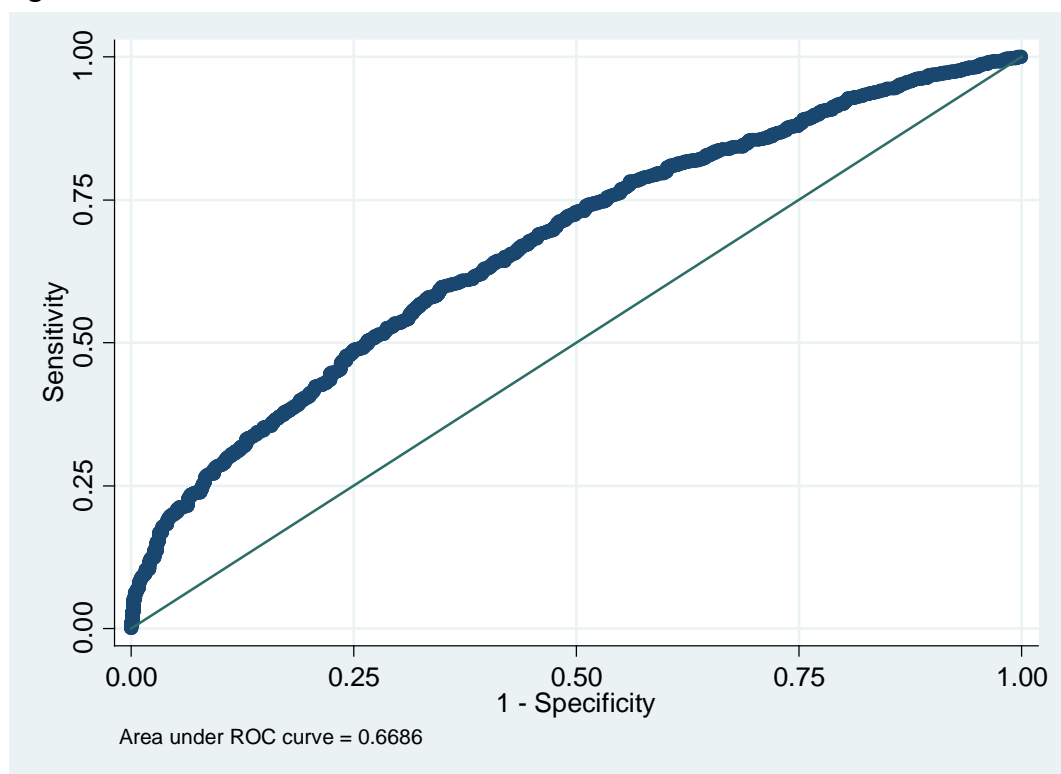
Regular Soft Drinks



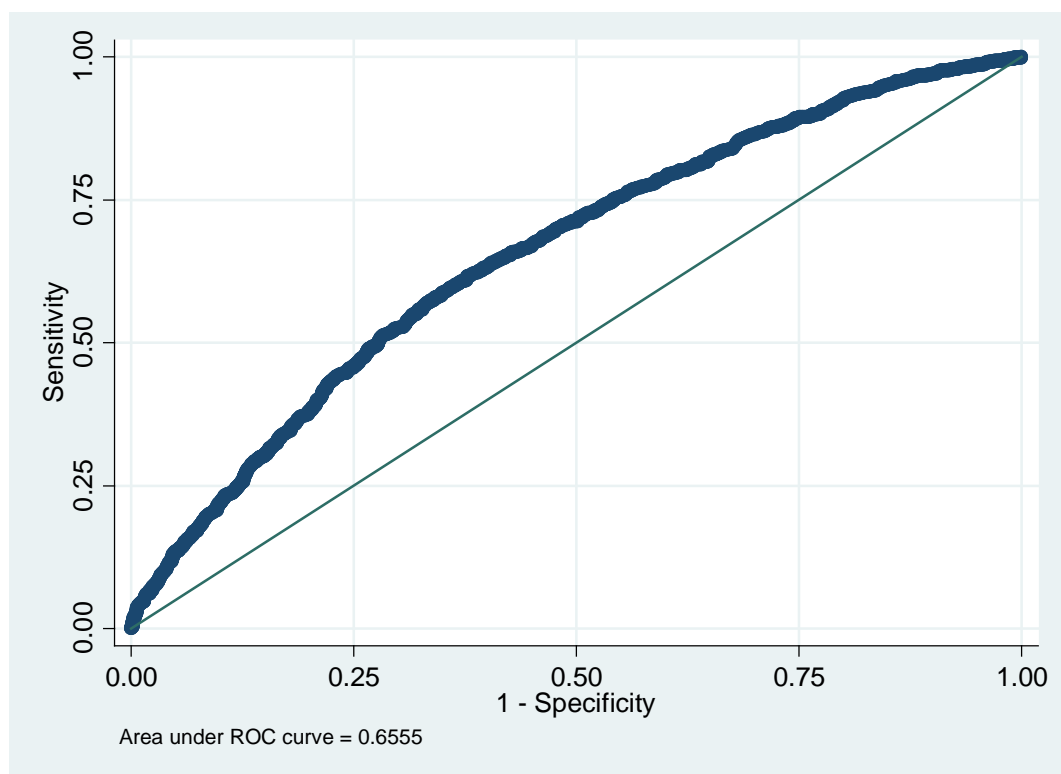
Diet Soft Drinks



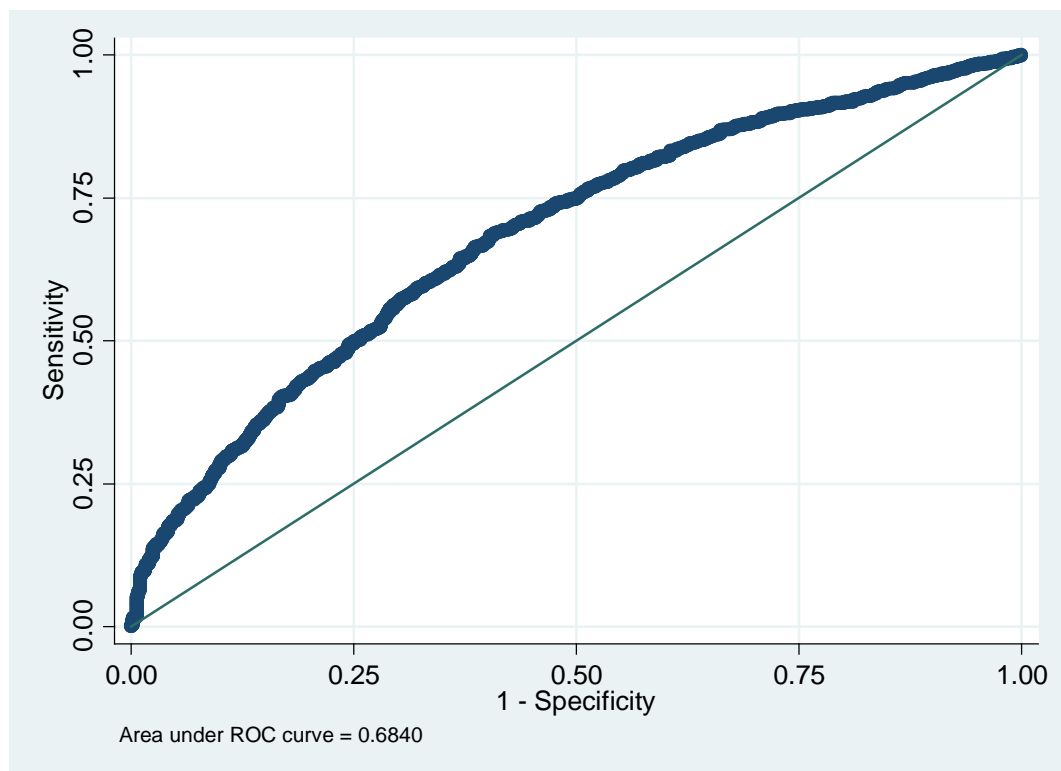
High-Fat Milk



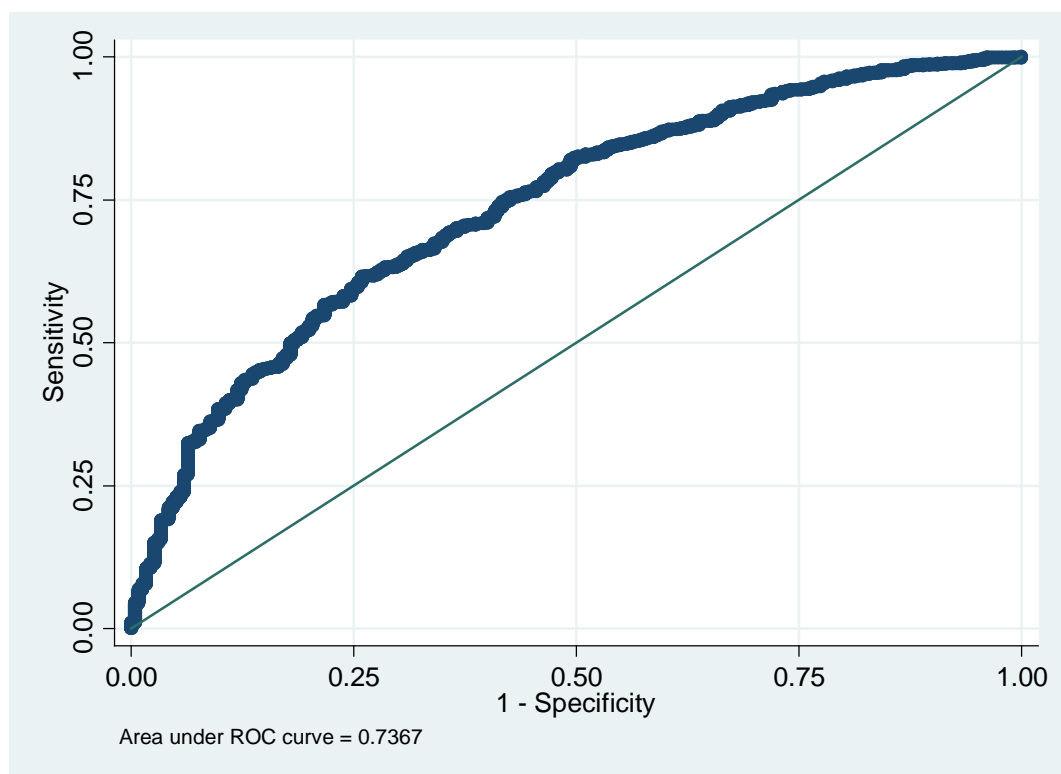
Low-Fat Milk



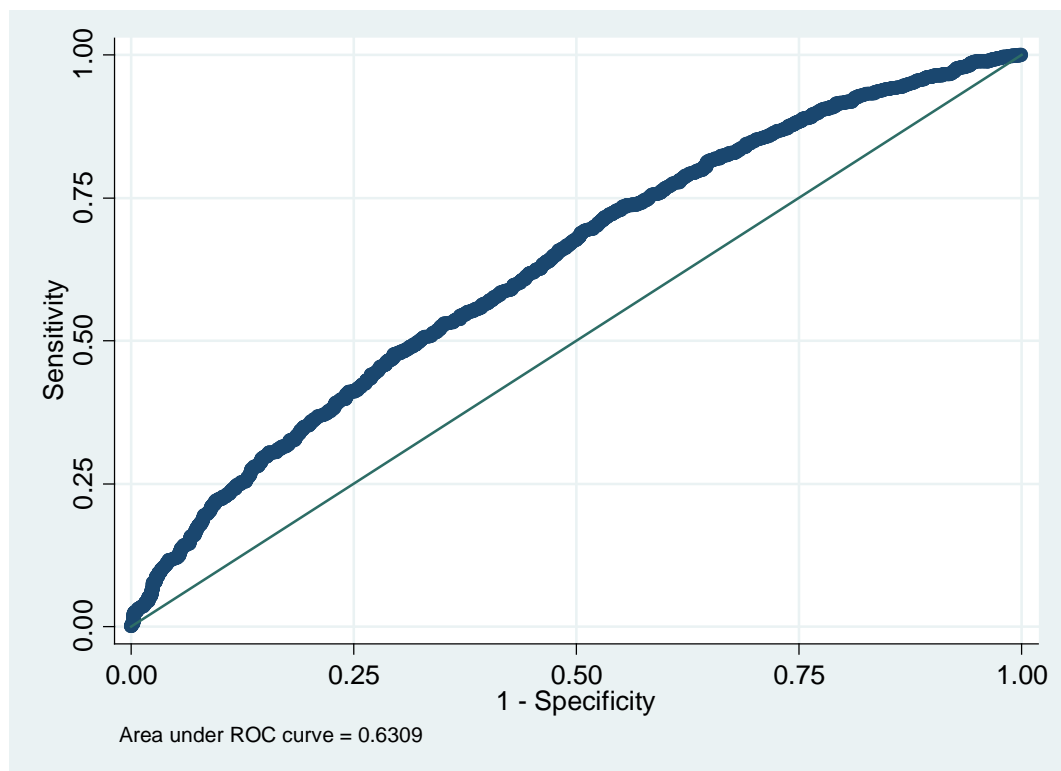
Fruit Drinks



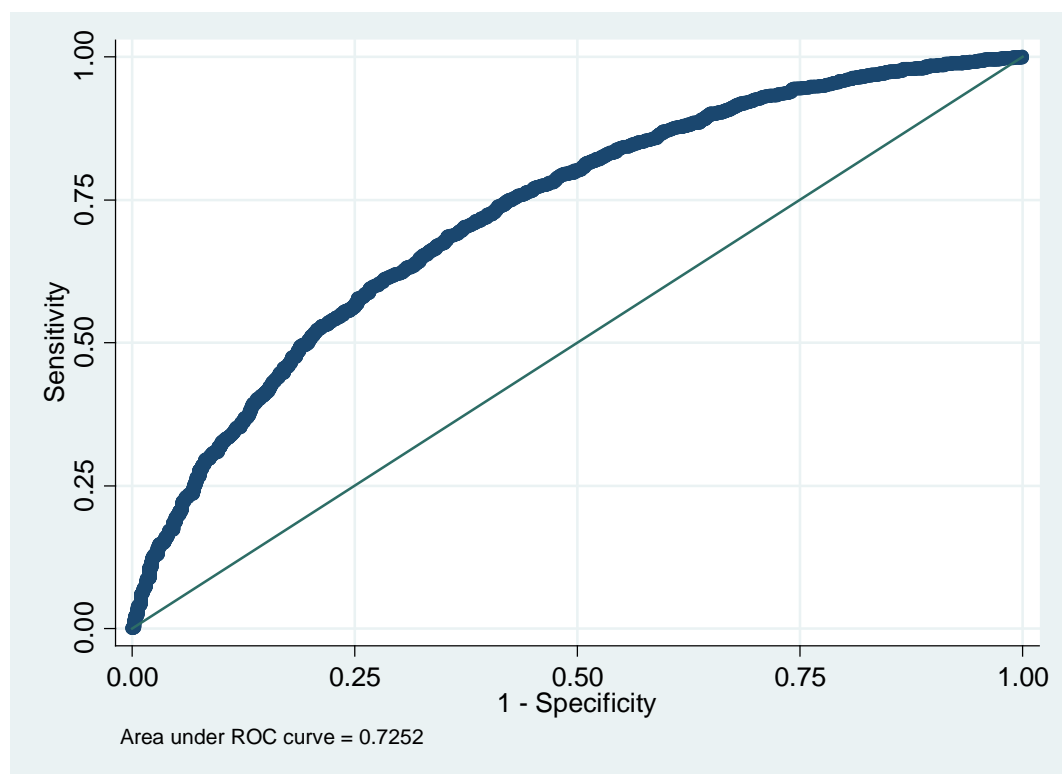
Fruit Juice



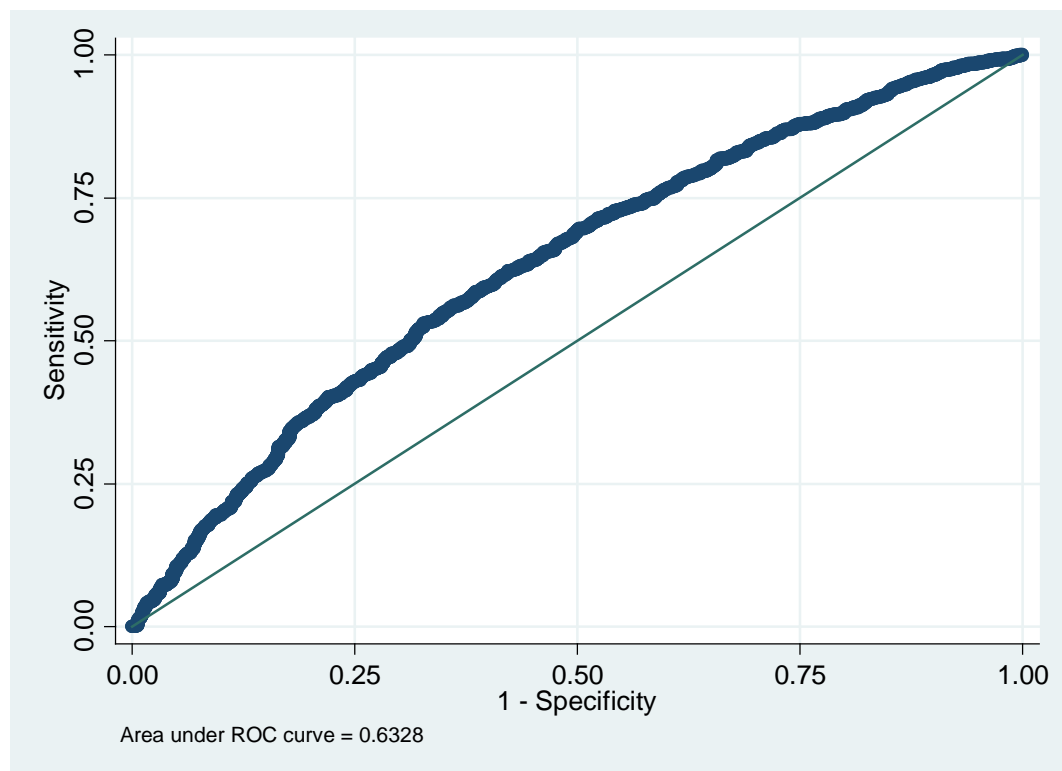
Bottled Water



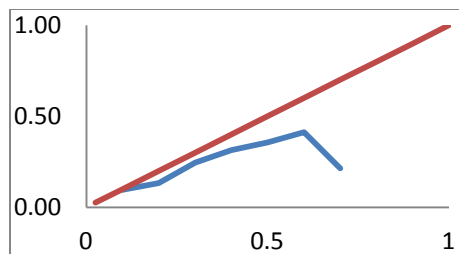
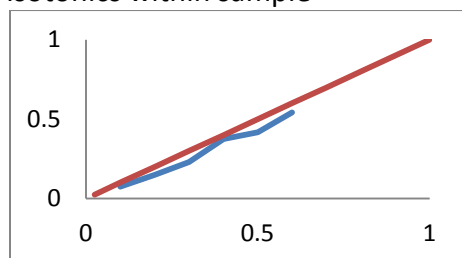
Coffee



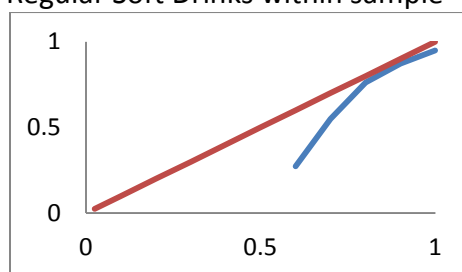
Tea



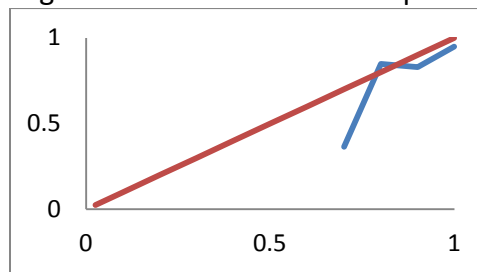
Isotonics within sample



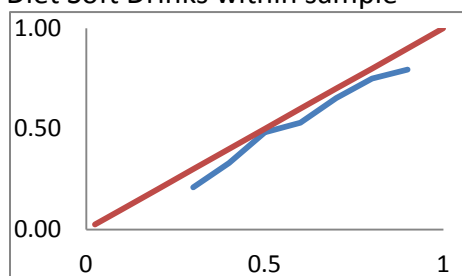
Regular Soft Drinks within sample



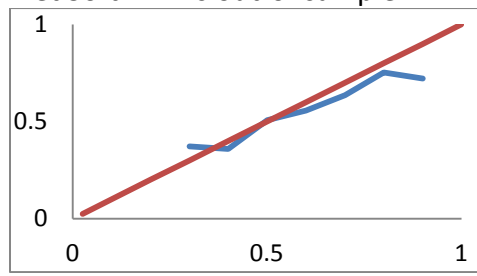
Regular Soft Drinks out-of sample



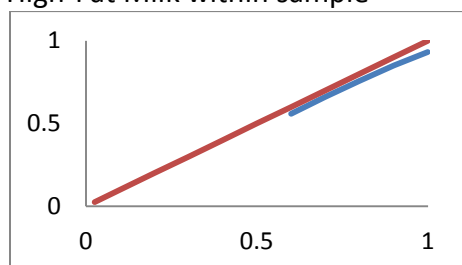
Diet Soft Drinks within sample



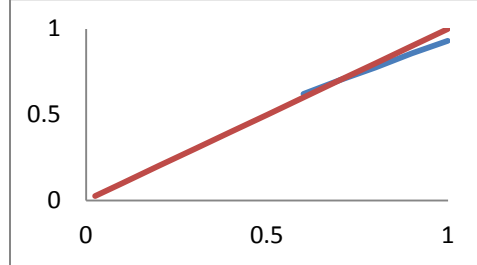
Diet Soft Drinks out-of sample



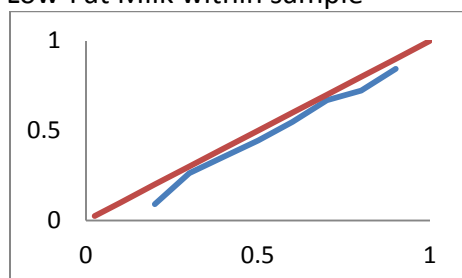
High-Fat Milk within sample



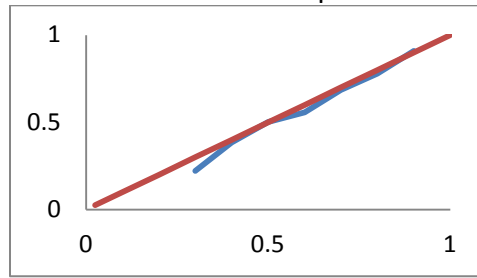
High-fat Milk out-of sample



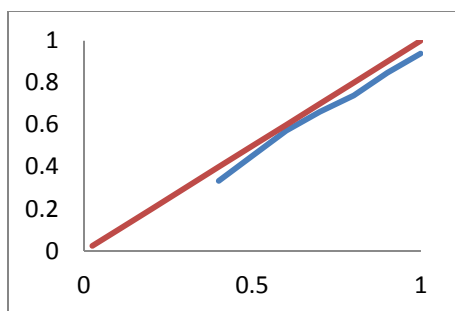
Low-Fat Milk within sample



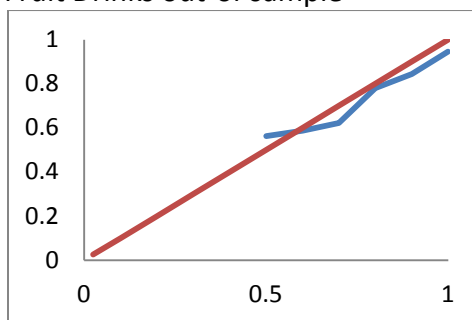
Low-fat Milk out-of sample



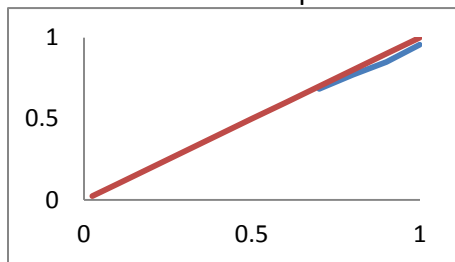
Note: Horizontal axis is the probability and the vertical axis is the realized relative frequency. Red line represents the perfect calibration line and blue line represents the model generated calibration line



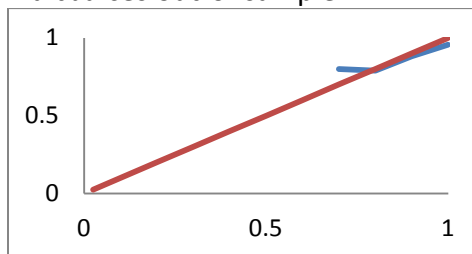
Fruit Drinks out-of sample



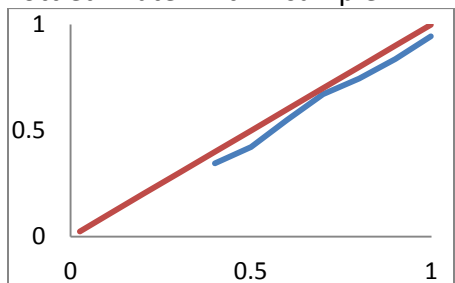
Fruit Juices within sample



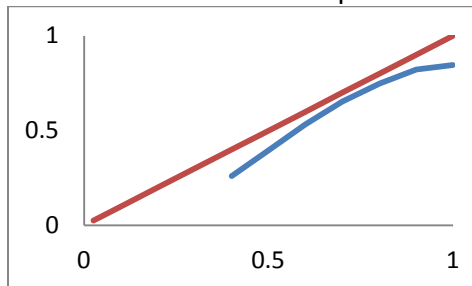
Fruit Juices out-of sample



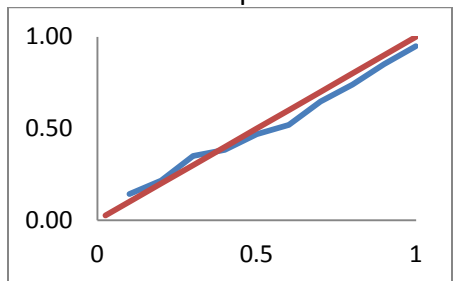
Bottled Water within sample



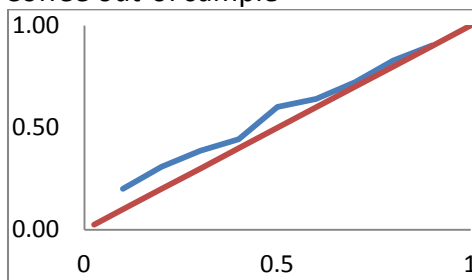
Bottled Water out-of sample



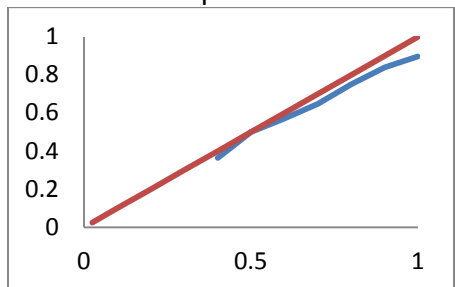
Coffee within sample



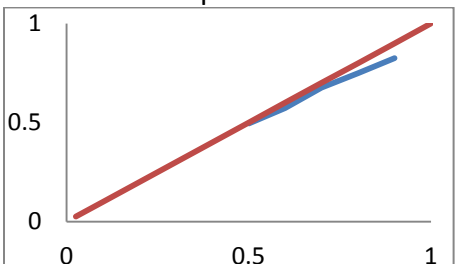
Coffee out-of sample



Tea within sample



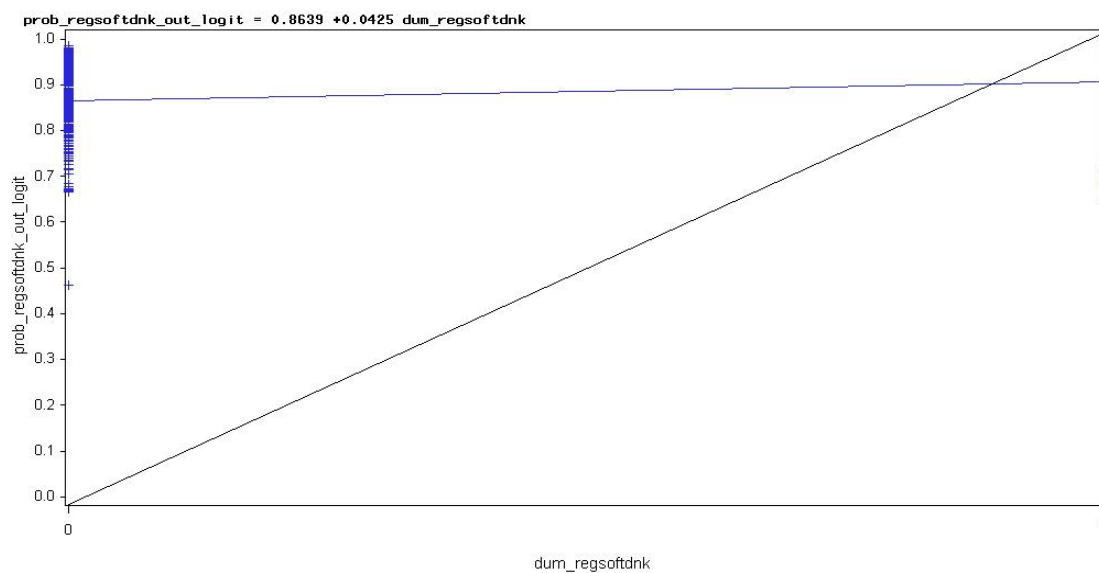
Tea out-of sample



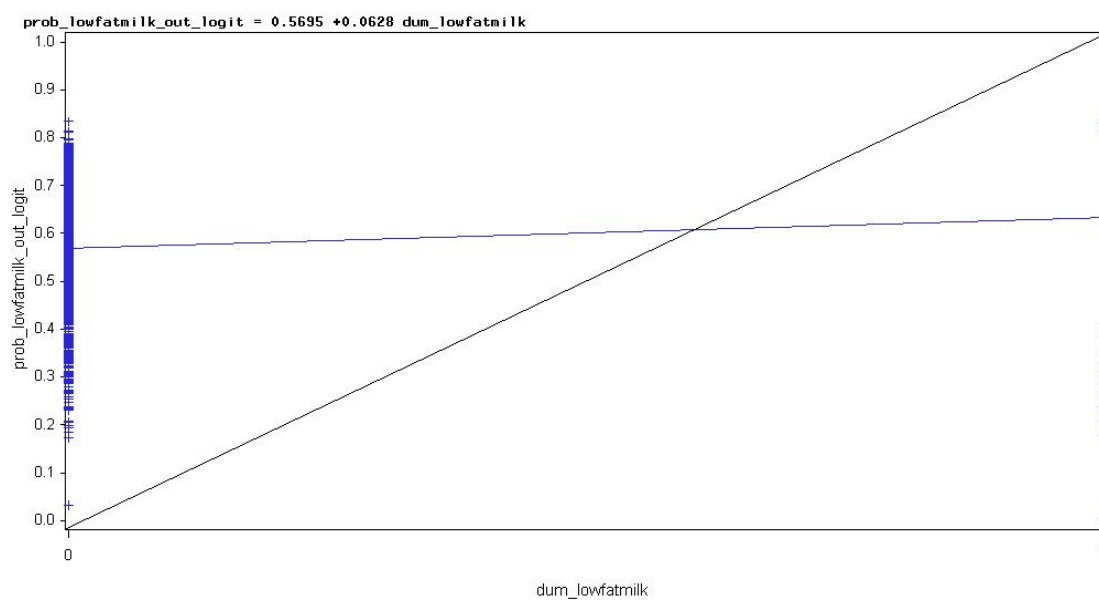
Note: Horizontal axis is the probability and the vertical axis is the realized relative frequency. Red line represents the perfect calibration line and blue line represents the model generated calibration line

Figure 4: Resolution Graphs for Probabilities and Outcome Index: Logit Model Out-of-Sample

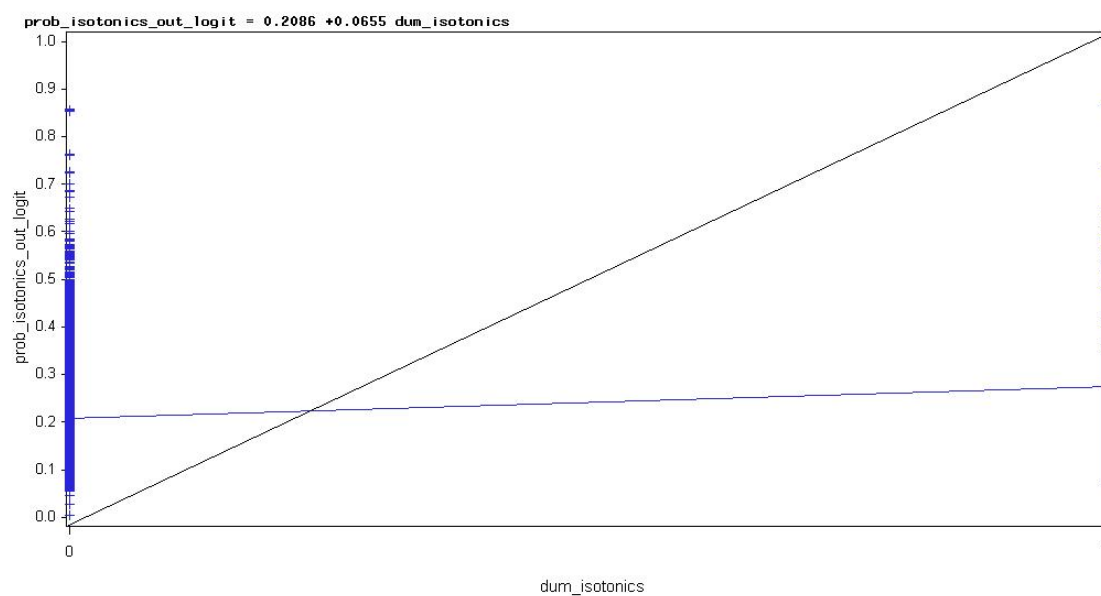
Resolution Graph:Logit Out-of-Sample Regular Soft Drinks



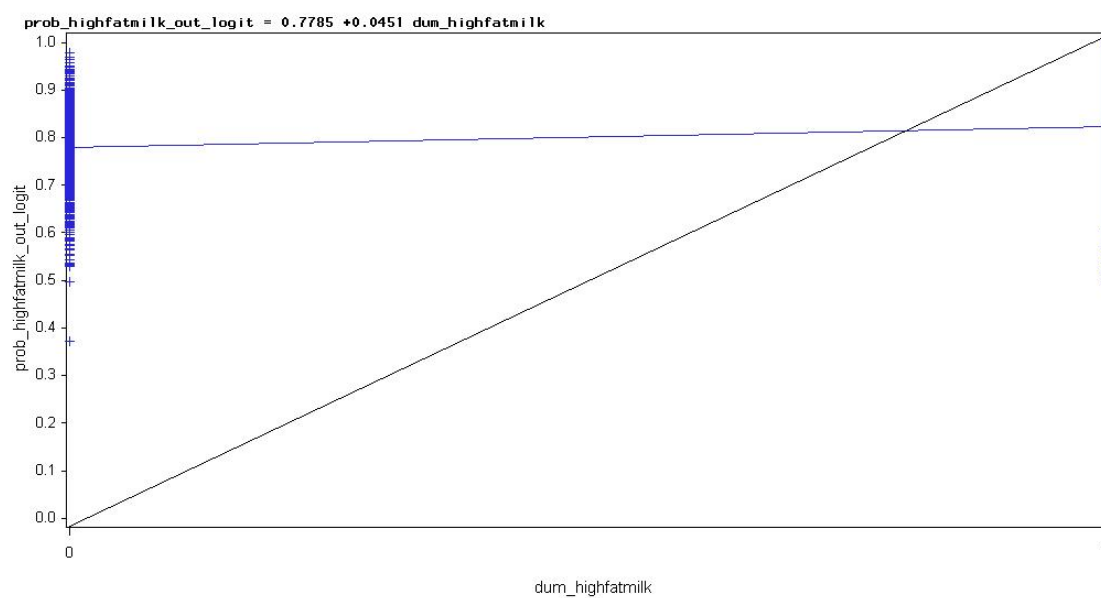
Resolution Graph:Logit Out-of-Sample Low Fat Milk



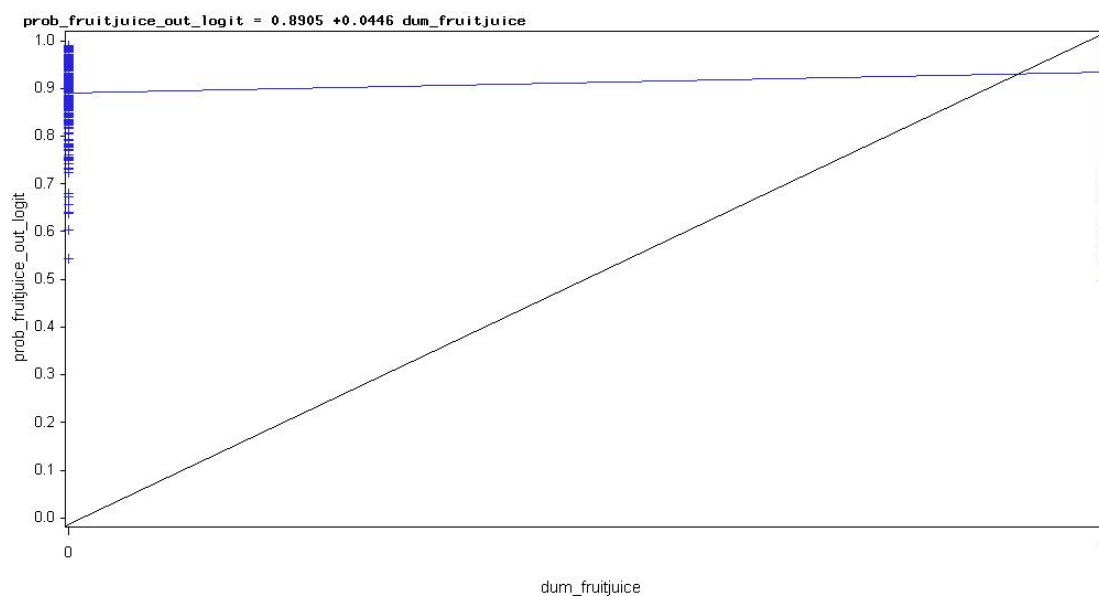
Resolution Graph:Logit Out-of-Sample Isotonics



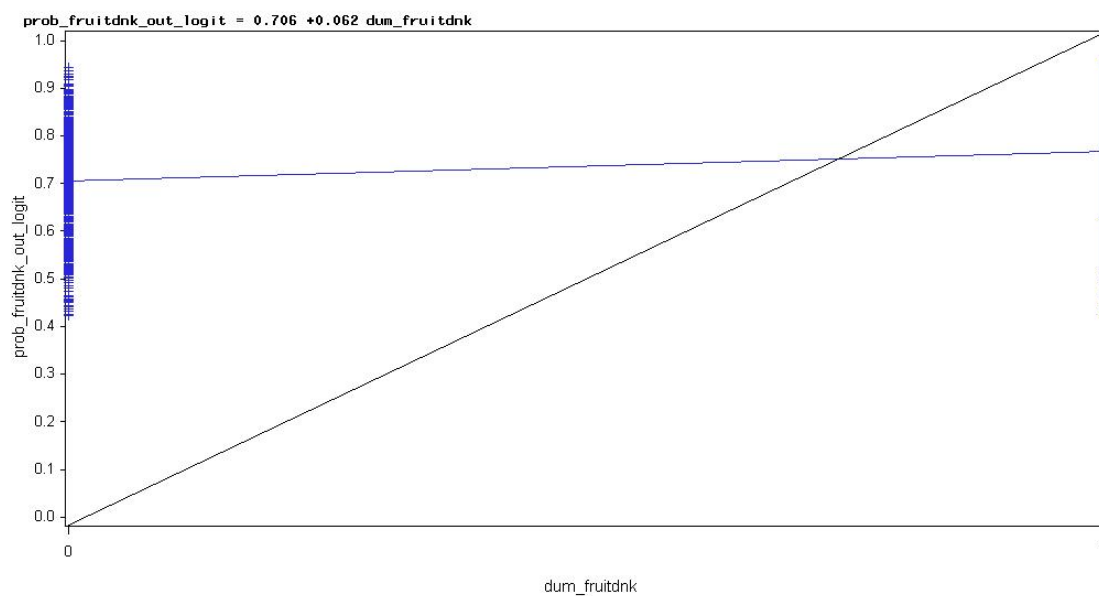
Resolution Graph:Logit Out-of-Sample High Fat Milk



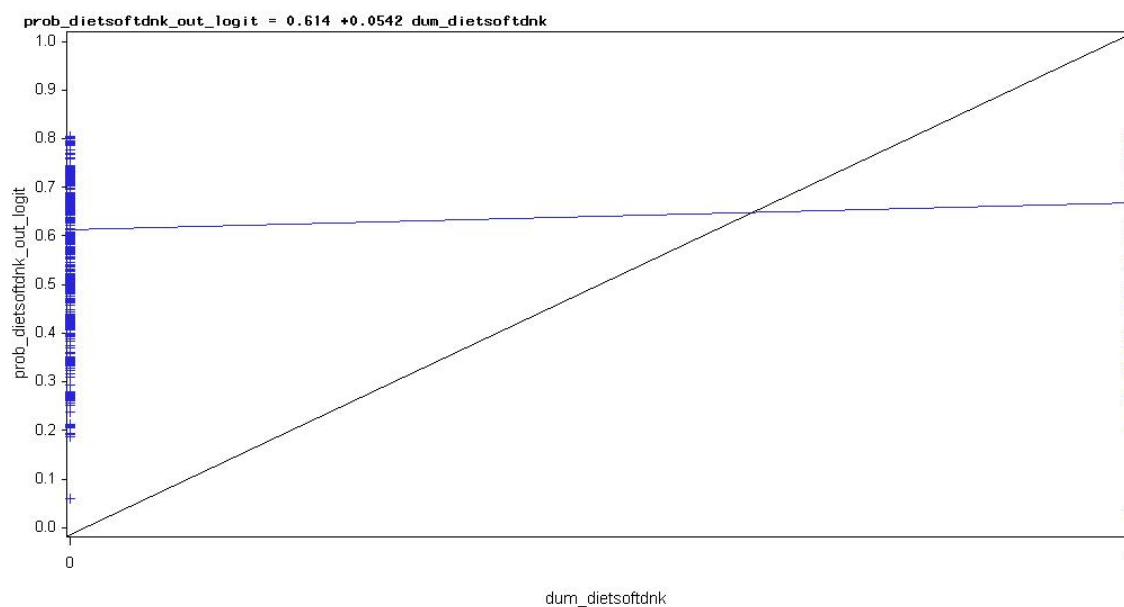
Resolution Graph:Logit Out-of-Sample Fruit Juices



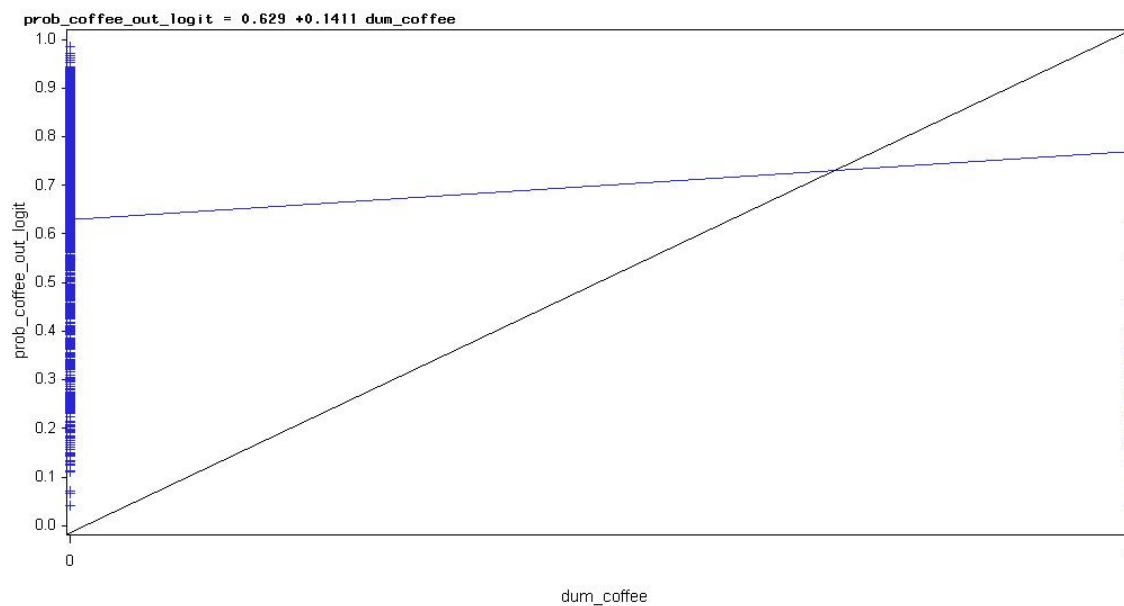
Resolution Graph:Logit Out-of-Sample Fruit Drinks



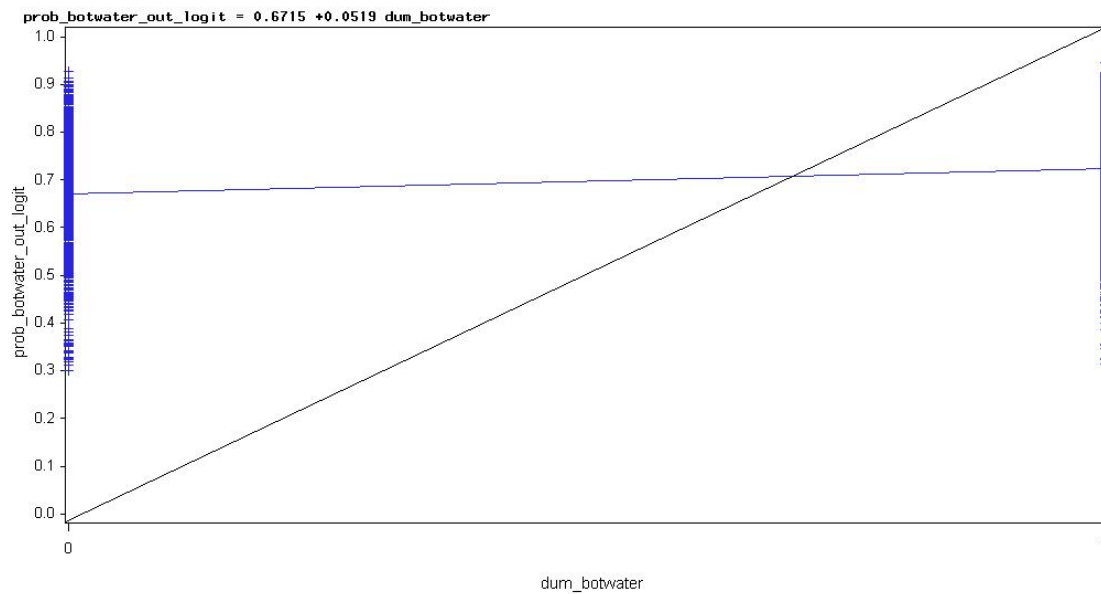
Resolution Graph:Logit Out-of-Sample Diet Soft Drinks



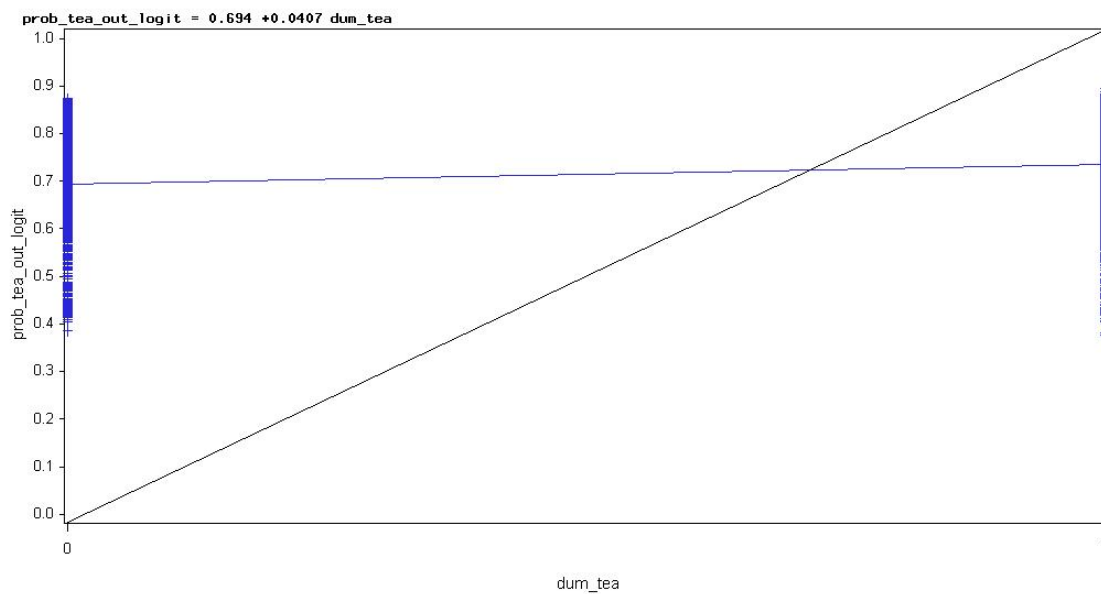
Resolution Graph:Logit Out-of-Sample Coffee



Resolution Graph:Logit Out-of-Sample Bottled Water



Resolution Graph:Logit Out-of-Sample Tea



Note: Horizontal axis refer to the zero (0), one (1) outcome index. Vertical axis is the model generated probability

Figure 3: Market Penetration Versus Variance of Outcome Index

