Consumer Research with Big Data: Applications from the Food Demand Survey (FooDS)

Jayson L. Lusk

July 15, 2016

Abstract: In three separate studies based on data from Food Demand Survey (FooDS), which has been conducted monthly for over three years, this paper explores heterogeneity in preference across consumers in traditional demand systems, heterogeneity in preferences over time in choice experiments, and the tail of the distribution for a particular food consumption pattern. Results show elasticities of demand for food at home and away from vary widely across different groups of consumers defined by *a priori* cluster analysis utilizing demographic and attitudinal variables. Results from a choice experiment are found to depend on when the experiment was conducted and on the market-prices prevailing at the time of the survey. Given the large sample of consumers observed over time, I am able to demographically characterize a small portion of the population, vegetarians, using traditional logit models and a machine learning method, classification trees.

Keywords: CART, choice experiment, demand system, food at home, food away from home, vegetarianism

* This project was supported by the Willard Sparks Endowed Chair at Oklahoma State University, the Oklahoma Agricultural Experiment Station, and the Agriculture and Food Research Initiative Competitive Grant no. 2015-67023-23134 from the USDA National Institute of Food and Agriculture.

Researchers now have access to tens of thousands if not millions of observations arising from online search and shopping behaviors, retail scanner data, and panel surveys, opening the door to deeper insights about consumer preferences than has previously been possible.

A number of recent studies have begun to explore "big data" in relation to modeling economic phenomenon and consumer choice (e.g., Bajari, et al., 2015; Belloni, Chernozhukov, and Hansen, 2014; Varian, 2014). Such papers have discussed new "tricks" and techniques being applied to large datasets, many of which fall under the umbrella of machine learning.

This paper is less focused on new techniques (though one such machine learning method is used in the third application) and instead explores how these large data sets might yield new insights even with traditional methods. As discussed by Einav and Levin (2014), research in top economics journals has become increasingly empirical, partly in response to growing data quality and availability. Hamermesh (2013), for example, calculated that in 2011 more than 72% of papers in the top three economics journals were empirical in nature, a figure up from only 48% in 1963. Over this time period, empirical work in top journals has shifted from using "ready-made" data (e.g., government survey data) and has increasingly relied on author-created datasets, including surveys and experiments. These emerging data sets are timelier than traditional government surveys and they often capture previously unmeasured variables tailor made to answer questions of interest. The publishing trends discussed by Hamermesh (2013) in terms of the types of papers and number of co-authors seems to suggest that research in top economic journals is beginning to look more similar to the real-world, problem-solving type of research that has been the mainstay of agricultural economics, and it suggests opportunities for food and agricultural economists who are well trained in empirical methods and analyses.

This paper explores several questions from a new author-generated data set, the Food Demand Survey (FooDS). FooDS is an online survey that has been repeatedly delivered to over 1,000 consumers every month for over three years yielding more than 36,000 observations as of this writing. With this large survey, I study (i) preferences for food at home and away from as revealed by structural demand systems, (ii) preferences for meat products over time as revealed by repeated choice experiments, and (iii) the characteristics of a group of consumers representing a small share of the overall population (vegetarians). A common theme of the analyses presented here is heterogeneity – across people, over time, and in the tails of the distribution.

Accounting for respondent heterogeneity has been a theme of many previous papers in the food and agricultural economics literature. What I add here is the exploration of temporal or cross-sectional heterogeneity in models and approaches that have previously been less amenable to such questions. For example, analyses of consumer demand using demand systems (e.g., see Marsh and Piggott, 2011; Unnevehr et al., 2010) has typically relied on aggregate time-series data where analysts look for structural shifts in demand by the representative consumer over time. While such structural model can include demographic demand shifters, as I discuss below, it is often done in a restrictive manner using national aggregates.

Other demand models have relied on the random-utility approach, where consumer choice is explained by product attributes. While modern analyses working within the random utility framework frequently model heterogeneity across consumers, the underlying data are often cross-sectional, providing little insight into how preferences might vary over time. A common method to generate such data is the so-called choice experiment; however, as I will show the "experiment" is not conducted in a vacuum, and the market prices prevailing at the

time of the survey significantly impact consumers' choices. Thus, changes in market conditions in conjunction with other temporal factors lead to variation in willingness-to-pay over time.

Finally, most consumer demand analysis tends to focus on the average or representative consumer. However, sometimes interest is in the people in the tail of distribution, as they may be responsible for setting new trends or for a disproportionate share of a firm's profitability. A simple random sample will only yield a small number of people in the tails of a distribution, so a (very) large sample is needed if one wishes to characterize the demographics or attitudes of these "extreme" people. The application in this study focuses on vegetarians, a group of consumers who have heretofore been difficult to characterize given that they represent a small share of the overall population.

While each of the three studies explored in this paper make contributions to the literature in their own right, the analyses fit into a broader literature that tackles consumer demand issues using big data. For example, Zhen et al. (2014) used quarterly Nielsen scanner data from over 37,000 households to estimate demands for 23 food categories, and they explored variation in demand elasticities across households with different demographics, finding the burden of a soda tax varied by income. Similarly, Allender and Richards (2010) used household scanner data to project the *ex ante* welfare impacts of impending California laws on egg consumers and explored how the impacts varied by household demographics. Other studies using large scanner data sets have explored how prices paid vary across consumers of different income levels and by package size, location of purchase, etc., even after controlling for quality differences by utilizing product bar codes (Broda et al., 2009; Griffith et al., 2009). Examples extend beyond scanner data applications. For example, Taylor and Villas-Boas (2016) made use of the relatively new National Household Food Acquisition and Purchase Survey (FoodAPS) from the USDA

Economic Research Service to study over 50,000 food acquisition choices made by 4,826 households, finding preferences for different types of food outlets varying by household income levels and participation in food assistance programs. These studies often involve multiple datasets, merging survey and scanner data with data on products' nutritional contents, store location, prices and wages reported by the Bureau of Labor Statistics, and more.

The next section of the paper describes FooDS and the survey sample. Then, three separate studies are discussed in turn, each focused on different questions from FooDS. Within each study, motivation, methods, and results are discussed. The last section summarizes and concludes.

The Food Demand Survey (FooDS)

FooDS is an online survey that has been conducted monthly since May 2013. The survey is delivered on the 10th of the month unless the 10th falls on a weekend, in which case delivery is moved to the following Monday. Each month over 1,000 completed surveys are obtained, a sample size that produces an approximate 3% sampling error with 95% confidence for a dichotomous choice question. All completed observations are typically acquired within three days of delivery.

Because of concerns about cost and attrition, FooDS was not constructed as a true panel. That is, a different sample of people is interviewed each month (aside from a handful of observations that can be judged to have come from the same IP address in multiple months). While the participants are not the same each month, the questions are identical over time, aside from a few ad hoc questions added at the end of each survey. Survey programing and hosting is done by Oklahoma State University researchers using the software Qualtrics. The survey is

delivered to a sample each month generated by Survey Sampling International (SSI), an organization that uses a diversity of methods to recruit and maintain a large opt-in panel of online respondents. Participants are compensated for completions in a variety of ways depending on how they were recruited including cash, charitable donations, points redeemable for goods, and frequent flyer miles. In monthly reports releasing the results, the data are weighted to assure correspondence with the US population in terms of gender, age, education, and region of residence. Unless otherwise noted, the weights are not used in this paper.

The three studies in this paper rely on 32,683 surveys collected in the 32 months from June 2013 to January 2016. The first month's data, May 2013, is not utilized because afterward we changed the way the alternatives in the choice experiment questions were randomized. The end date, January 2016, was chosen because that is when the analyses discussed here began. It should be noted that the third study also makes use of more recent FooDS data (5,175 observations from February 2016 to June 2016) as a holdout sample to test the prediction performance of competing models.

Responses to a variety of questions are tracked each month including awareness and concern for a host of food issues, food values, food challenges, and changes in expected food prices, purchases and spending. The particular questions used in this paper are discussed in each of the sections that follow.

Study 1. Heterogeneity in Preferences for Food at Home and Away from Home

Motivation

A common model used to analyze consumer demand is the almost ideal demand system (AIDS) introduced by Deaton and Muelbauer (1980). Despite the popularity of the AIDS model and

other related demand systems, there has emerged a lack of clarity on how model consumer heterogeneity. A common approach is demographic scaling where demographic variables (often aggregate measures like the share of women in the workforce) are included as linear shifters in the expenditure share equations. A downside to this approach is that demographics enter in a restrictive fashion as the non-linear price index parameters are assumed independent of demographics. More problematic, however, is an often unappreciated result derived by Alston, Chalfant, and Piggott (2001) who showed that inclusion of demographics in this manner causes the estimates to depend on the units of measurement (e.g., elasticities will change depending on whether prices are measured per kilogram or pound). This finding has led researchers interested in demand shifters such as, for example, food safety recalls (Piggott and Marsh, 2004) to utilize demand systems like the generalized AIDS (GAIDS) model. However, to avoid the scale invariance problem in the GAIDS model, demand shifters enter the model in an even more restrictive fashion: demographics only affect so-called pre-committed quantities and do not affect price sensitivity or income effects directly. While there are other possible solutions to this problem, here I utilize an approach that takes advantage of a large data set like FooDS. The application relates to demand for food at home vs. food away from home, a topic that has received increasing attention as consumers have been spending more on food away from home in recent decades and because of concerns about the relative healthfulness of food away from home (Stewart, 2011).

Methods

The demand system estimation is based on three questions asked about weekly expenditures on food at home, weekly expenditures on food away from home, and total annual expenditures (i.e.,

income which is divided by 52 to produce a weekly measure).[1] See Lusk (2013) for exact question wording. These data are merged with month- and region-specific price data from the Bureau of Labor Statistics (BLS), namely the consumer price index (CPI) for food price at home, the CPI food price away from home, and the CPI less food. Variation in these prices over time and across region is what identifies consumer preferences. These data characterize a three-good demand system for food at home, food away from home, and all non-food goods (the expenditures of which are calculated as the difference between income and total food spending). Given that we utilize total income rather than group-expenditures, we avoid some of the problems associated with expenditure-conditional demand systems (LaFrance, 1991).

The issue at hand is how to incorporate respondent heterogeneity into demand systems like the AIDS model. I utilize an approach somewhat akin to Deaton (1985), who suggested creating pseudo-panels consisting of averages of demographically-identical cohorts over time. Unlike Deaton (1985), however, my interest is not in including cohort-fixed effects to an aggregate model using cohort averages, but rather in estimating completely separate models for different groups of people based on all the underlying data in each group.

To identify groups of consumers who may share similar preferences, cluster analysis was utilized prior to the demand system estimation. In particular, groups were formed based on their similarity/differences in fourteen different variables (household size, gender, age, race,

[1] FooDS data indicate mean expenditures of $96.23, $53.54, and $149.77 per week for food at home, food away from home, and total food; figures which amount to 11.8%, 5.6%, and 17.4% of total income. The at-home estimate is somewhat higher than that suggested by the BLS Consumer Expenditure Survey (CES), which implies for 2014 mean expenditures of $76.37, $53.60, and $129.90 per week for food at home, food away from home, and total food, representing 5.9%, 4.2%, and 10.1% of average weekly income (note: the CES totals do not include an additional $8.92 week spent on alcohol; also the CES asks expenditures using an open-ended question whereas FooDS provides ranges for respondents – the present analysis assigns people the midpoint of the range). By contrast, total food spending from FooDS is lower than the $165.13/week amount measured by FooDAPS (Taylor and Villas-Boas, 2016). As a final point of comparison, 2015 data from the Bureau Economic Analysis (BEA) Personal Consumption Expenditures (PCE) (see BEA Table 2.4.5U) implies that total aggregate spending on food at home and away from home are 7.3% and 5.4% (totaling 12.7%) of total aggregate spending.

education, income, marital status, total food spending, and responses to questions that ask about the most/least important food challenges such as "finding affordable foods that fit my budget" and "finding convenient, quick-to-make alternatives"). Proc FASTCLUS in SAS was used to identify clusters and to assign individual observations to each cluster. The procedure creates clusters by minimizing the Euclidean distances (the sum of squared differences) between multiple variables (the approach is also called k-means clustering). An observation is assigned to the group (or cluster) that is closest in distance. The procedure is sensitive to units of measurement and outliers, so data used in the cluster analysis are standardized (to have mean zero and standard deviation of one) prior to clustering. The total sample size is 32,683 observations, and I created 50 clusters that have an average size of about 654 respondents (maximum and minimum cluster sizes are 932 and 452 respondents).[2]

An AIDS model was estimated for each cluster *c*, where the expenditure share for individual *i* and good *j* ($w_{i,j} = x_{i,j}p_{i,j}/X_i$) is specified as:

$$(1) \qquad w_{i,j} = \alpha_{c,j} + \sum_{k=1}^{3} \gamma_{c,jk} \ln(p_{i,k}) + \beta_{c,j} \ln\left(\frac{X_i}{P_{c,i}}\right)$$

where X_i is total expenditure (in this case income) by individual *i*, $p_{i,j}$ is the price of the j^{th} good type faced by individual *i*, and $P_{c,i}$ is price index for cluster *c* and individual *i* defined by:

$$(2) \qquad \ln(P_{c,i}) = \alpha_{c,0} + \sum_{k=1}^{3} \alpha_{c,k} \ln(p_{i,k}) + 0.5 \sum_{l=1}^{3} \sum_{g=1}^{3} \gamma_{c,lg} \ln(p_{i,l}) \ln(p_{i,g}).$$

Homogeneity, adding-up, and symmetry are imposed for each cluster, *c*:

$\sum_{j=1}^{3} \gamma_{c,jk} = 0$, $\sum_{j=1}^{3} \alpha_{c,j} = 1$, $\sum_{j=1}^{3} \delta_{c,j} = 0$, and $\gamma_{c,jk} = \gamma_{c,kj}$. This process produces sets of parameter estimates for each of the 50 clusters, which are used to calculate cluster-specific

[2] The number of clusters was chosen somewhat arbitrarily; however, I sought to create enough clusters to allow for an exploration of heterogeneity while ensuring each cluster contained enough observations that sampling error was not too large, and 50 clusters seemed a good compromise. I have re-conducted the analysis here with 10, 20, and 60 clusters, and the overall pattern of results is broadly similar in terms of the means and distributions of elasticities that emerge.

elasticities, as well as the compensating variation resulting from the price change that occurred from July 2013 to June 2015 (over this time, the prices of food at home, away from home, and non-food increased approximately 3.6%, 5.1%, and 1.8%).

Results

The mean own-price elasticities of demand for food at home, away from home, and non-food across the 50 clusters are -1.673, -1.372, and -1.073; by contrast, when a single model was fit to the aggregate data, the own price elasticities were -1.773, -1.449, and -1.056, respectively. The estimated elasticities for food away from home are somewhat more elastic than reported in the previous literature based on time-series data (e.g., see review in Okrent and Alston, 2012). The income elasticities for food at home and away from home averaged 0.031 and 0.153 across the 50 disaggregate models, but were 0.227 and 0.428 in the aggregate pooled model.

While the results suggest the tendency for some bias in the elasticities stemming from the aggregate model relative to the disaggregate model means, more interesting is the heterogeneity in elasticities that arises from the 50 disaggregate models. Figure 1 shows the distribution of elasticities for food at home and away from home. Several clusters are highly elastic, with elasticities less than –3. Eleven clusters (or 22% of the sample) have an own-price elasticity of demand for food at home of between -0.01 and -1. There are only 2 clusters (representing 4% of clusters) which violate curvature conditions with elasticities greater than zero.

Figure 2 shows a wide dispersion of cross price elasticities of demand for food at home and away from home, with the two goods being substitutes for about a two-thirds of the sample and compliments for the other third of the sample. Figure 3 shows the compensating variation resulting from the price changes that occurred from July 2013 to June 2015. The mean welfare

effect was -$3.96/week but the amount varied a low of -$1.82/week to -$6.97/week across the 50 clusters.

The elasticities vary widely across cluster characteristics. For example, figure 4 shows how the cross price elasticity of demand for food at home with respect to a change in the price of food away from home varies with mean group spending on food at home. While there does not appear much of linear relationship between these two variables, it does not follow that the data are useless or uninformative. For example, grocers may be most interested in retaining those consumer clusters that spend a relatively large proportion of their income on food at home and who are most responsive to a change in the price of food away from home (i.e., those clusters in the upper-right hand portion of the graph).

There are some significant correlations between elasticities and mean cluster demographics. For example, running linear regressions (N=50 each) that include the explanatory variables income, proportion white, proportion on SNAP, and proportion with children under 12 in the household reveals that higher income clusters have smaller income elasticities of demand for food at home, more inelastic demand for food at home, and larger compensating variation losses from the price changes than do clusters with lower incomes. Clusters that have more SNAP participants have higher income elasticities of demand for food at home (going from a cluster with no SNAP participants to one with 100% SNAP participants is predicted to increase the income elasticity by 1.12). Clusters with more households who have children at home have lower income elasticities of demand for food at home, more inelastic own-price elasticities of demand for food at home, and larger welfare losses from the 2013-2015 price changes than clusters that have fewer households with children at home.

Study 2. Meat Demand in Repeated Choice Experiments

Motivation

Stated preference methods in general, and choice experiments in particular, have become a popular method for estimating consumer preferences in the environment, food, and transportation literatures (Louviere, Hensher, and Swait, 2000). These studies generate willingness-to-pay (WTP) values that are used to inform policy makers via cost-benefit analysis and agribusinesses making decisions about pricing and new product introduction. Aside from a few isolated examples, these studies are almost universally conducted using data from survey delivered to a cross-section of respondents at a single point in time. This raises the question about the temporal stability of WTP, which relates to the robustness and generalizability of stated preference studies. While it is often presumed that choices in a contingent valuation or choice experiment study are unaffected by outside-survey influences, respondents do not enter a vacuum when they take the survey. Changes in general economic conditions like wages and unemployment, occurrence of news stories about the topic in question, or prices of substitutes for the good being studied in the "real world" would all rationally be expected to influence consumers' WTP, and yet a single "snap shot", cross-sectional study cannot identify such effects. This study utilizes data from an identical choice experiment conducted over 32 consecutive months to study the temporal stability in WTP for meat products.

Methods

In study 1, consumers' preferences were identified by variation in prices over time; in study 2, preferences are identified by variation in prices at a point in time via a choice experiment. The choice experiment used in this study is a simple "branded" choice experiment, where the only

attributes are the food type and price. In each choice question, respondents were shown images of eight different uncooked food items (beef steak, ground beef, pork chop, deli ham, chicken breast, chicken wing, rice and beans, and tomato pasta), and were asked which item they would choose (subjects could also choose a "none of these" option) when shopping for a meal for their family. Each respondent answered nine choice questions that were identical to each other except for the prices assigned to each option. The prices were varied across options such that they were uncorrelated with each other across all the choice questions. Given that each respondent answered 9 choice questions, the dataset consists of 32,683*9 = 294,147 choice observations. More details about the choice experiment are available in Lusk (2013) and Lusk and Tonsor (2016).

To explore whether the choice experiment results were affected by outside influences, prevailing market prices for each of the eight products used in the choice experiment were obtained from the BLS. The BLS price data were merged with the choice experiment data by month and region. There is wide variation in BLS prices; for example, the BLS price of ground beef varied from a low of $2.936/lb (in July 2013 in Midwest region) to a high of $4.674/lb (in June 2016 in West region).

For sake of convenience and parsimony, consumer preferences were estimated via the multinomial logit (MNL) model. Consumer *i* responding at time *t* is assumed to derive the following utility from choice option *j* in choice task *c*: $U_{ijtc} = V_{ijtc} + \varepsilon_{ijtc}$, where ε_{ijtc} is a stochastic term assumed to be known to the individual but not the analyst. The systematic portion of the utility function is defined as:

(3) $$V_{ijtc} = \theta_{ijt} + \gamma_{ijt} p_{jc},$$

where p_{jc} is the price of alternative *j* in choice task *c*, γ_{ijt} is the marginal (dis)utility of a price change, and θ_{ijt} is an alternative specific constant indicating the utility of option *j* relative to the utility of the "no purchase" option which is normalized to zero for identification purposes. WTP for food type *j* relative to "none" is: $-\theta_{ijt}/\gamma_{ijt}$. WTP for food type *j* relative to food type *k* is: $-(\theta_{ijt} - \theta_{ikt})/\gamma_{ijt}$. For sake of simplicity and exposition, the foregoing analysis reports only two values, WTP for ground beef relative to "none" and WTP of chicken breast relative to ground beef.

As indicated by the subscripts in (3), the preference parameters are allowed to vary by individual and time. In particular, let $\theta_{ijt} = \delta_0 + \sum_{k=1}^{10} \delta_{jk} z_{ik} + \sum_{g=1}^{6} w_{jg} d_g + \pi_j BLS_{ijt} + \rho_{jt}$, where z_{ik} are demographic variables representing region of residence, household size, presence of children in the household, gender, education, income, and race that affect choice through the parameters δ_{jk}, d_g are dummy variables indicating the day of week (e.g., Monday, Tuesday) the survey was taken, w_{jg} are coefficients associated with alternative-specific day-of-week effects, BLS_{ijt} is the BLS price for good *j* reported by BLS in month *t* for consumer *i* (the variation across consumers within a month arises from region of residence), π_j is a coefficient showing how preferences for food type *j* in the experiment are affected by prices for the same type of food outside the experiment, and ρ_{jt} are alternative- and time-specific monthly fixed effects. The coefficient related to price responsiveness, γ_{ijt}, is similarly specified as a function of demographics, day of week, time, and BLS prices. The result is a model that consists of 441 coefficients.[3]

[3] Full estimation results are available from the author upon request. The total number of coefficients is as follows. There are 8 alternative specific constants (the ninth "none" option is normalized to zero) and one price effect = 9 coefficients; 10 demographic variables * (8 alternative specific constants + 1 price effect) = 90 coefficients; 31 time periods (one time period is normalized to zero for identification) * (8 alternative specific constants + 1 price effect) = 279 coefficients; 6 day of week effects (one day is normalized to zero for identification) * (8 alternative specific

Assuming the ε_{ijtc} follow a Type I extreme value distribution and are independently and identically across individuals, choices, and alternatives, then the probability of individual *i* choosing option *j* in choice set *c* in month *t*:

(4) Prob(*i* chooses *j*) = $\frac{e^{V_{ijct}}}{\sum_{k=1}^{9} e^{V_{ikct}}}$.

Results

Likelihood ratio tests reject the null of no demographic effects (χ^2=21688 with 90df, p-value <0.001), no BLS price effects (χ^2=162 with 9df, p-value <0.001), no day-of-week effects (χ^2=205 with 54df, p-value <0.001), and no month effects (χ^2=1425 with 279df, p-value <0.001). In short, there are structural changes in the preference parameters over time, day of week, demographics, and preference parameters depend on prevailing market prices at the time of the experiment.

Figure 6 reports estimated WTP for ground beef as compared to "none" and in comparison to chicken breast, holding BLS prices and demographics constant at mean levels. For sake of comparison, the nationwide BLS price for ground beef is also plotted in the figure. Results suggest temporal variability in WTP. From April to September 2015, WTP for ground beef vs. none is significantly higher than it was prior to December 2013. Interestingly, WTP for ground beef seems to follow a pattern that is similar to that exhibited by the BLS market price for ground beef despite the fact that the plotted WTP values hold BLS values constant at mean BLS price levels. The correlation between WTP for ground beef vs. "none" and the BLS price of ground beef shown by the data points in figure 6 is 0.70. The figure also shows that WTP for chicken breast vs. ground beef changes over time. At the same time ground beef WTP was

---

constants + 1 price effect) = 54 coefficients; 1 alternative-specific BLS price * (8 alternative specific constants + 1 price effect) = 9 coefficients. The sum is 9+90+279+54+9=441.

increasing relative to "none", so too was chicken breast, and in fact relative WTP for chicken breast over ground beef was higher in late 2015 than earlier in the time period.

Figure 7 shows variation in WTP for ground beef and chicken breast by the day of the week the survey was conducted holding constant demographics, BLS prices, and month. Surveys conducted on Monday, Tuesday, and Wednesday tend to produce similar WTP values; however, there is a slight decline in WTP for ground beef vs. none and WTP for chicken breast vs. ground beef toward the end of the week. WTP confidence intervals are wider for Saturday and Sunday because relatively fewer observations were collected on these days.

Figure 8 illustrates the change in WTP as BLS prices change holding constant demographics, month, and day-of-week. WTP for ground beef vs. none falls by about $0.20 over the range of BLS prices show in figure 8; conversely, WTP for chicken breast vs. ground beef increases about $0.20 over the range of BLS prices plotted. The confidence intervals overlap for all the price ranges, suggesting that BLS prices do not have a statistically significant effect on these WTP values. However, as previously noted, the null hypothesis of no BLS price effects is strongly rejected by a likelihood ratio test. One of the reasons for the apparent lack of impact on these WTPs is that BLS prices have competing effects on the coefficients in the numerator and denominator of the WTP calculation. Increasing the BLS price simultaneously reduces price sensitivity as measured by the experiment (increasing WTP) and decreases the alternative-specific constant or non-pecuniary portion of utility (decreasing WTP). For example, a $1 increase in the BLS price of ground beef outside the choice experiment causes a statistically significant 0.013 increase in the marginal utility of a price change (the denominator in the WTP formula) in the choice experiment, say from -0.513 to -0.500, while also decreasing the alternative-specific constant (the numerator in the WTP formula) by a statistically significant -

0.104 utility points, say from 2.604 to 2.500. In this example, WTP would fall from 2.604/0.513 = $5.076 to 2.500/0.500 = $5.000 from a $1 price change in the BLS price of ground beef.

Study 3. Who are the Vegetarians?

Motivation

It is typically difficult or impossible to characterize small segments of the population using conventional surveys and polls. Survey researchers often seek to obtain a completed sample of between 1,000 and 1,500 respondents, which yields a sampling error of around 2.5 to 3%. Reducing the sample error further requires exponentially larger numbers of respondents, which is costly relative to the increased accuracy attained. For example, achieving a sampling error of 1% would require more than 9,600 completed observations. Although a sample size of 1,000 can provide a reasonably accurate estimate of the statistic of interest (e.g., mean WTP in the population), it is difficult to make meaningful statements about the characteristics of small sub-samples. For example, according to the Census Bureau, people who claim Native American as their only race make up about 0.8% of the population. Thus, a typical 1,000 person survey would likely only have about eight people in the sample that are Native American; hardly a sufficient sample size to say much about the WTP, income, education, or health status of this group.

In the context of food demand, a small population group that has been gaining more attention in recent years is vegetarians and vegans. Concerns about animal welfare, environmental impacts of livestock production, and health impacts from eating red meat have been in the news, and research has increasingly been taking up such issues (e.g., Pan, 2012; Golub et al., 2013; West et al., 2014). Previous estimates suggest that between 5% and 6% of

the US population considered themselves vegetarians between 1999 and 2012, and 2% considered themselves vegan in 2012 (Newport, 2012). Thus, a typical survey would yield a mere 50 vegetarians to study. As a result, little is known about trends in vegetarianism or about the characteristics of people who are vegetarian (see Lusk and Norwood, 2009 and Norwood and Lusk, 2016 for a couple papers touching on the economics of vegetarianism).

Methods

FooDS has been tracking vegetarian status monthly for over three years. The question simply asks, "Are you a vegetarian or a vegan?" and the response categories are "Yes" or "No."[4] The main interest of this study is in characterizing vegetarian/vegan (hereafter vegetarian) status – i.e., whether vegetarian status can be predicted by socio-economic and demographic characteristics. To address this question, I first use a conventional modeling approach – a logit model – and compare its performance to a machine-learning method popular in analysis of big data: classification and regression trees (CART).

A classification tree is used to predict the outcome of a dichotomous (or multichotomous) variable by splitting the sample into multiple sub-samples based on a series of predictor variables (see Varian (2014) for additional discussion of the method). The basic idea is to "grow" a tree that splits the sample based on the outcome of one of the predictor variables and then create "branches" that further subdivide the sample based on the levels of additional predictor variables. Splits and branches are determined by minimizing the variability in the outcome variable created by the split. This procedure typically produces a large tree that over-fits the data, and as such the

[4] As we note in Norwood and Lusk (2016), a large fraction of the people who say "Yes" also choose a meat option in the choice experiment. This is not necessarily an inconsistency as the choice experiment asks about choices when shopping for the family, whereas this question on vegetarianism asks about individual behavior.

tree is “pruned” by eliminating branches that do not add predictive power in cross-validation exercises.

For this analysis, I used the PROC HPSPLIT in SAS to construct a classification tree to predict vegetarian status. Predictor variables used to create tree include age, education, income, gender, SNAP participation status, race, obesity status, status as primary shopper, presence of children in the household, marital status, household size, and political ideology (measure on a five point scale, where 1=very liberal; 5=very conservative). The initial tree was grown by using an entropy measure to evaluate possible candidate splits in the data. The tree was then pruned by using cross-validation and a measure of cost-complexity that aims to trade off complexity (number of leaves in the tree) and prediction performance. Although the number of leaves that minimized the prediction error in this application was more than 30, for sake of parsimony and exposition, this paper utilizes a tree based on only seven leaves, a choice consistent with the “one minus standard error rule” commonly applied in CART studies. For sake of comparison, a binary logit model was also estimated on the data where the dependent variable was vegetarian status and independent variables were the same as those used in the classification tree analysis.

The logit model and the classification tree were constructed using the same dataset as in study 1 and 2 including 32,683 observations from June 2013 to January 2016. To compare the prediction performance of the logit and classification tree, vegetarian status is also predicted in a hold-out sample comprised of more recent data from February 2016 to June 2016 consisting of 5,175 total observations.

Results

Figure 9 shows the percent of respondents in FooDS indicating they were vegetarian or vegan over time. The percent of the population self-declaring vegetarian or vegan status ranged from a low of 3.3% in March 2014 to high of 7.9% in April 2016; the average over the entire time period was 4.9%. If anything there is a slight upward trend toward the end of the sample period.

While there are only about 30 to 80 vegans each month in the sample, when the data are pooled across months, the FooDS data set contains what is perhaps the largest collection of identified vegetarians and vegans to date. In the portion of the dataset used to estimate the logit model and construct the classification three (from June 2013 to January 2016), there were 1,767 self-identified vegetarians. In the hold-out data set (from February 2016 to June 2016), there are 344 vegetarians (out of a total sample size of 5,175).

The logit and CART models performed similarly in-sample. The percent of correct predictions was 94.6% and 94.7% for the logit and CART models, respectively. The receiver operating characteristic (ROC) was higher for the logit (0.72) than the CART (0.60), where a value of 0.5 indicates no predictive power and a value of 1 represents perfect prediction. However, the CART analysis performed better in-sample in terms of sensitivity (i.e., correctly classifying vegetarians as vegetarian) at 6.1% than the logit, which had a sensitivity measure of only 0.3%.

Of more interest are the out-of-sample predictions. The percent correct predictions were very similar for the two models; the logit was 93.3% and the CART was 93.4%. Whereas the logit had a superior out-of-sample log likelihood function value (-1164.97) than the classification tree (-1193.01), the classification tree, again, did a better job identifying true positives. In fact, the logit model had a sensitivity measure of 0%; the model did not correctly any of the vegetarians as vegetarian in out-of-sample prediction. By contrast, the classification tree

correctly classified 7% of the vegetarians as vegetarian. The two approaches were virtually identical in terms of specificity (i.e., correctly classifying non-vegetarians as non-vegetarian) at 99.9% and 99.6% for the logit and CART models.

Looking at the coefficients from the logit model, the following statistically significant results emerge: vegetarians are more likely to be higher income, younger, non-white, more highly educated, liberal, non-obese women on SNAP in small households. The largest effect sizes relate to age (the odds of being vegetarian are 2.85 times higher for individuals under 25 years of age relative to those 65 and older), income (the odds of being vegetarian are 2.56 times higher for individuals with income $160,000/year or higher relative to those with incomes less than $20,000/year), ideology (odds of being vegetarian are 3.91 times higher for very liberal vs. very conservative individuals), and SNAP participation (the odds of being vegetarian are 3.53 times higher for SNAP participants than non-participants).

Rather than making predictions based on *ceteris paribus* regression coefficients, the classification tree categorizes individuals based cut-offs from multiple predictor variables. Figure 10 shows the seven-leaf classification tree. The first branch is based on political ideology. All individuals with an ideology score greater than one are predicted to be non-vegetarian; by contrast, those individual with a score of one (i.e., those that are "very liberal) ultimately have branches predicted to be vegetarian. Working down the tree, there are three "leaves" or sub-categories of respondents (colored in light grey) where the majority of individual are vegetarian. In short, the classification tree predicts vegetarians are very liberal, on SNAP, with incomes higher than $60,000, with children, who are either 35 and older or younger than 35 with a household size less than three. There is an approximate two-third chance that individuals meeting these criteria are vegetarian according to the classification tree.

Conclusion

Big data has allowed progress in empirical economics by giving answers to new questions and by providing new ways to address old questions. Aggregate disappearance data from the US Department of Agriculture have long allowed food and agricultural economists to estimate aggregate, representative demands for agricultural commodities. But these data lack information on cross-sectional heterogeneity, and they are annual or quarterly measures released with significant lags. Datasets like those associated with the National Health and Nutrition Examination Survey (NHANES) provide insight about demographics, health, and nutritional quality of foods consumed but NHANES is conducted only periodically and lacks information on key economic variables like prices. Over the past two decades, scanner datasets both from household panels and retail outlets have emerged to provide timely, detailed information about consumers' purchase behaviors. While household scanner panels contain some information about demographics, often missing is ongoing collection of information about health, attitudes, beliefs, SNAP participation, etc. Scanner data also present challenges in causal identification due to unmeasured variables (e.g., unobserved quality attributes) that can bias demand estimates. In recent decades, surveys and experiments have emerged that allow insights into consumer preferences in a more controlled environment, but often at the expense of smaller sample sizes and lack of temporal variability.

No single data source is likely to be a panacea. Future insights are likely to be garnered by combining multiple datasets both from government and private sources, and in fact these big, combined "unstructured" datasets are already being routinely used in consumer research. Newer data related to online search and shopping, movements tracked by smart phones and watches,

and even data on consumers' eye movements and brain waves are beginning to inform consumer research, and they present new challenges in terms of memory and storage, aggregation, sampling, and statistical analysis. These new sources of information are also being joined with new cross-sectional, time series surveys that focus on the economics of food consumption, including the USDA Economic Research Service's FoodAPS survey and my Food Demand Survey (FooDS).

This paper used FooDS data to elucidate how the emergence of a new dataset allows answers to questions that have been previously out of reach, and provides new ways to address old questions. I showed, for example, that there is significant cross-sectional heterogeneity in traditional demand systems. Although this is not a novel insight, I present a new way of capturing such preference heterogeneity while preserving the theoretical properties of the demand system. This is accomplished by exploiting the size and complexity of the FooDS data set; using spending, demographic, and attitudinal variables to create numerous clusters of consumers for which separate demand systems are estimated.

This research also suggested some caution in utilizing "snap shot", cross-sectional surveys for making policy and marketing recommendations. Using data from what is perhaps the largest and longest-running choice experiment to date, I show that willingness-to-pay (WTP) values exhibit significant temporal variability, some of which is explained by changes in market prices outside the survey. The results show that choice experiments do not take place in a vacuum, and even if prices of goods within the experiment are uncorrelated, marginal utilities of price changes are likely to be affected by prices outside the experiment.

Finally, while there has been interest in characterizing small segments of the population, such as vegetarians, the analyses was impossible until a large-scale survey came along that

yielded a sufficient sample size of vegetarians to draw meaningful analyses. Traditional logit models provided insight into the characteristics of vegetarians, and they yield good overall prediction performance; however, this traditional approach performed poorly at detecting true positives (i.e., correctly classifying vegetarians as vegetarian). This might be a more general problem when trying to predict outcomes which only pertain to a small share of the population. I find that a new machine learning method, a classification tree, performs better than the logit model in term of detecting true positives while providing additional insights into the characteristic of vegetarians.

This paper presented only a small, and somewhat eclectic, set of examples of how one new large-scale dataset can be used to garner insights about consumer demand for food. Many exciting questions lay in the future. For example, might classification and regression trees better predict consumer choice than demand systems or random utility models? How do consumers' social networks affect food consumption and the flow of information about food? Which types of consumers are most affected by media stories, weather, and food safety events? What are the distributional impacts of state and federal food policies across consumers with different socio-economic and demographic characteristics? Food and agricultural economics are well equipped to answer such questions given the right kinds of data. The challenge for the future is in creating the datasets we will need to answer these and other emerging questions.

References

Allender, W.J. and T.J. Richards. 2010. "Consumer Impact of Animal Welfare Regulation in the California Poultry Industry. *Journal of Agricultural and Resource Economics* 35(3):424-442.

Alston, J.M., J.A. Chalfant, and N.E. Piggott. (2001). "Incorporating Demand Shifters in the Almost Ideal Demand System." *Economics Letters* 70(1):73-78.

Bajari, P., D. Nekipelov, S.P. Ryan, S.P. and M. Yang. 2015. "Machine Learning Methods for Demand Estimation." *American Economic Review* 105(5):481-485.

Belloni, A., V. Chernozhukov, and C. Hansen (2014), "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives*, 28(2): 29-50.

Broada, C., E. Leibtag, and D.E. Weinstein. "The Role of Prices in Measuring the Poor's Living Standards." *Journal of Economic Perspectives* 23(2):77-37.

Deaton, A. 1985. "Panel Data from Time Series of Cross-Sections." *Journal of Econometrics* 30(1-2):109-126.

Deaton, A. and J. Muellbauer. (1980). "An Almost Ideal Demand System." *American Economic Review* 70(3):312-326.

Einav, L. and J. Levin. 2014. "Economics in the Age of Big Data." *Science* 346(6210): 1243089.

Griffith, R. E. Leibtag, A. Leicester, and A. Nevo. 2009. "Consumer Shopping Behavior: How Much Do Consumers Save?" *Journal of Economic Perspectives* 23(2):99-120.

Golub, A.A., B.B. Henderson, T.W. Hertel, P.J. Gerber, S.K. Rose, and B. Sohngen. 2013. "Global Climate Policy Impacts on Livestock, Land Use, Livelihoods, and Food Security." *Proceedings of the National Academy of Sciences* 110(52): 20894-20899.

Hamermesh, D.S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature* 51(2013):162-172

LaFrance, J.T. 1991. "When is Expenditure 'Exogenous' in Separable Demand Models?" *Western Journal of Agricultural Economics* 16:49–62

Louviere, J.J., D.A. Hensher, J.D. Swait, J.D. 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge: Cambridge University Press.

Lusk, J.L. 2013. "Food Demand Survey: Technical Information on Survey Questions and Methods." Department of Agricultural Economics, Oklahoma State University. May 22, 2013. Available online at: http://agecon.okstate.edu/files/Survey%20Info%20(pdf4556).pdf

Lusk, J.L. and F.B. Norwood. 2009. "Some Economic Benefits and Costs of Vegetarianism." *Agricultural and Resource Economics Review* 38(2):109-124.

Lusk, J.L. and G.T. Tonsor. 2016. "How Meat Demand Elasticities Vary with Price, Income, and Product Category." *Applied Economic Perspectives and Policy*. forthcoming.

Newport, F. 2012. "In U.S., 5% Consider Themselves Vegetarians." Gallup.com. July 26, 2012. http://www.gallup.com/poll/156215/consider-themselves-vegetarians.aspx

Norwood, F.B. and J.L. Lusk. 2016. "Some Vegetarians Spend Less Money on Food, Others Don't." *Ecological Economics*. forthcoming.

Okrent, A.M. and J.M. Alston. 2012. "The Demand for Disaggregated Food Away-From-Home and Food-at-Home Products in the United States." USDA Economic Research Service, Research Report Number 139, August.

Pan, A., Q. Sun, A.M. Bernstein, M.D. Schulze, J.E. Manson, M.J. Stampfer, W.C. Willett, and F.B. Hu. 2012. "Red Meat Consumption and Mortality: Results from 2 Prospective Cohort Studies." *Archives of Internal Medicine* 172(7):555-563.

Piggott, N.E., and T.L. Marsh. 2004. "Does Food Safety Information Impact US Meat Demand?" *American Journal of Agricultural Economics* 86(1):154-174.

Stewart, H. 2011. "Food Away from Home." in J.L. Lusk, J. Roosen, and J.F. Shogren (eds). *The Oxford Handbook of the Economics of Food Consumption and Policy*, 647-666.

Taylor, R. and S.B. Villas-Boas. 2016. "Food Store Choices of Poor Households: A Discrete Choice Analysis of the National Household Food Acquisition and Purchase Survey (FoodAPS)." *American Journal of Agricultural Economics* 98(2):513-532.

Unnevehr, L., J. Eales, H. Jensen, J. Lusk, J. McCluskey, and J. Kinsey. (2010). "Food and Consumer Economics." *American Journal of Agricultural Economics* 92(2):506-521.

Varian, H.R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28(2):3-28.

West, P.C., J.A. Gerber, P.M. Engstrom, N.D. Mueller, K.A. Brauman, K.M. Carlson, E.S. Cassidy, M. Johnston, G.K. MacDonald, D.K. Ray, and S. Siebert. 2014. "Leverage points for improving global food security and the environment." *Science*, 345(6194): 325-328.

Zhen, C., E.A. Finkelstein, J.M. Nonnemaker, S.A. Karns, and J.E. Todd. 2014. "Predicting the Effects of Sugar-Sweetened Beverage Taxes on Food and Beverage Demand in a Large Demand System." *American Journal of Agricultural Economics* 96(1):1-25.

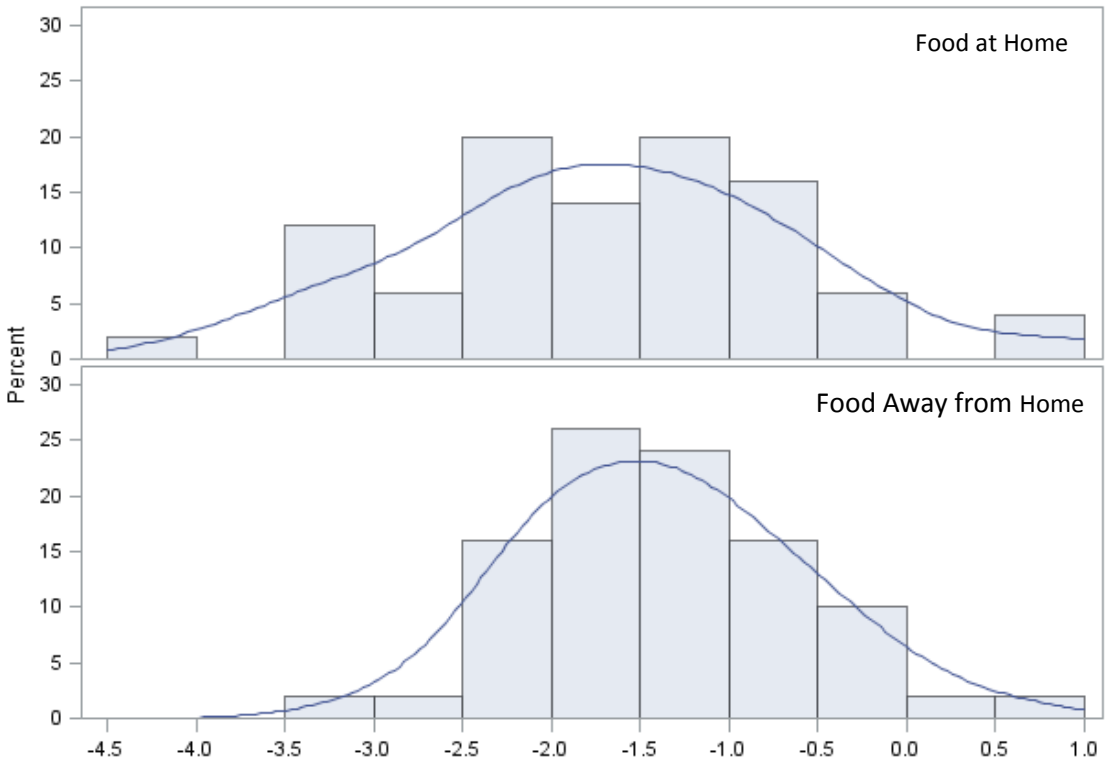

Figure 1. Distribution of Own Price Elasticities of Demand

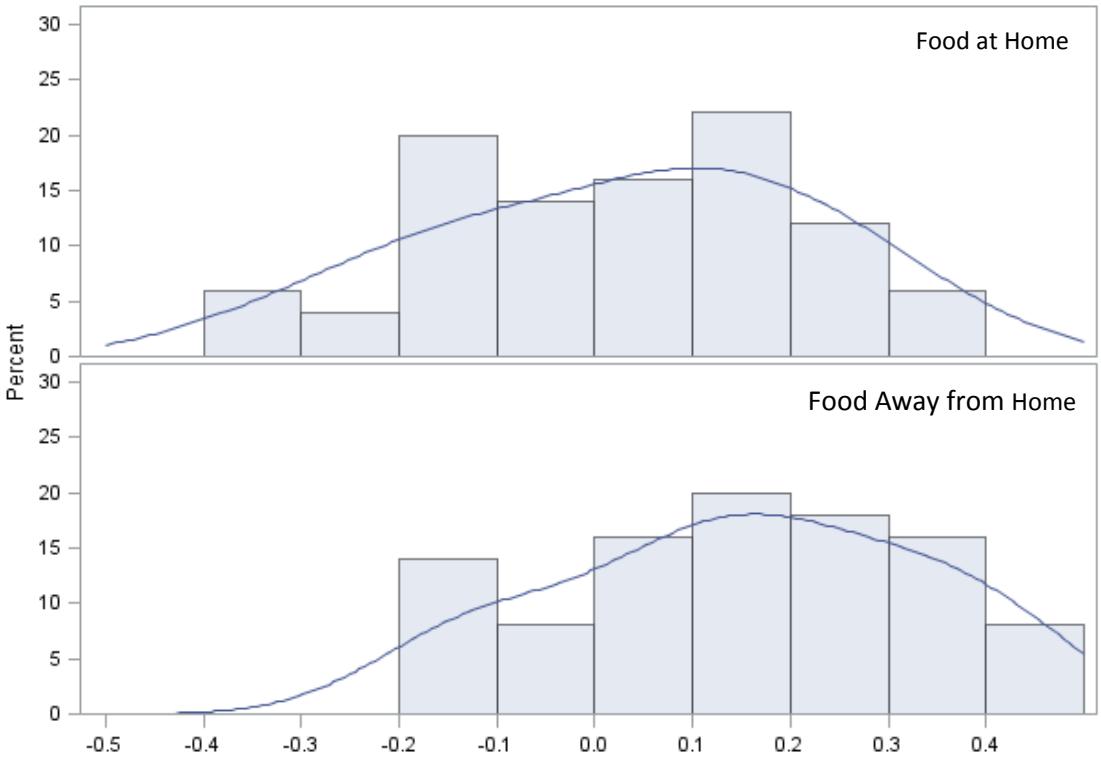

Figure 2. Distribution of Income Elasticities

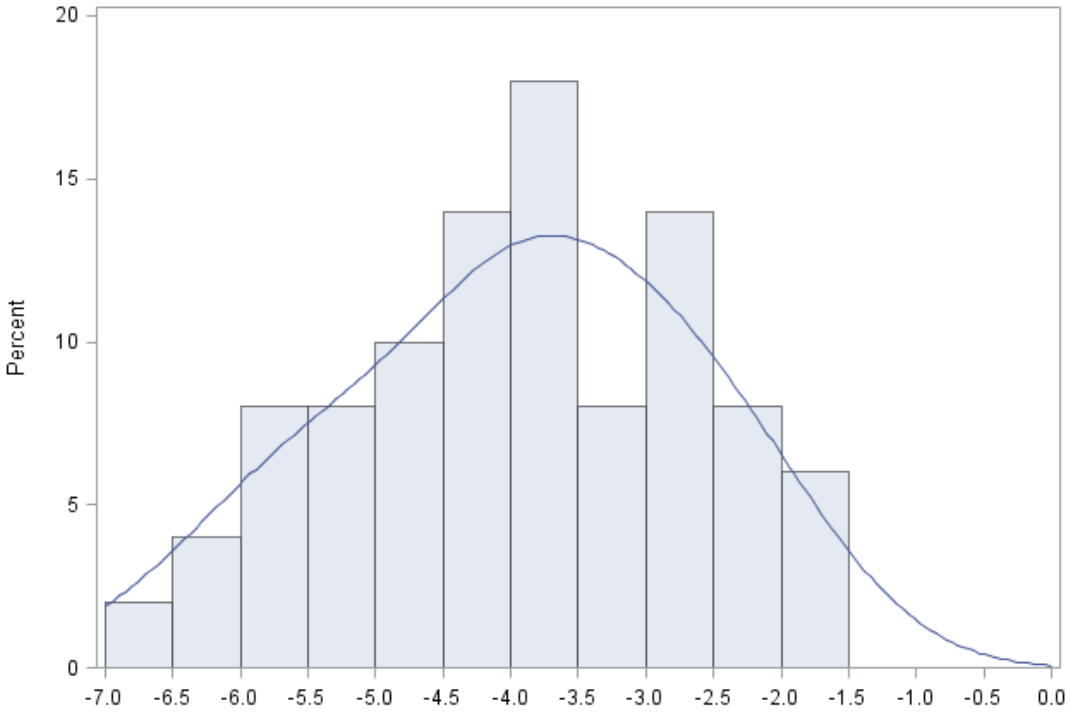

Figure 3. Distribution of Compensating Variation Resulting Price Change from July 2013 to July 2015 ($/week)

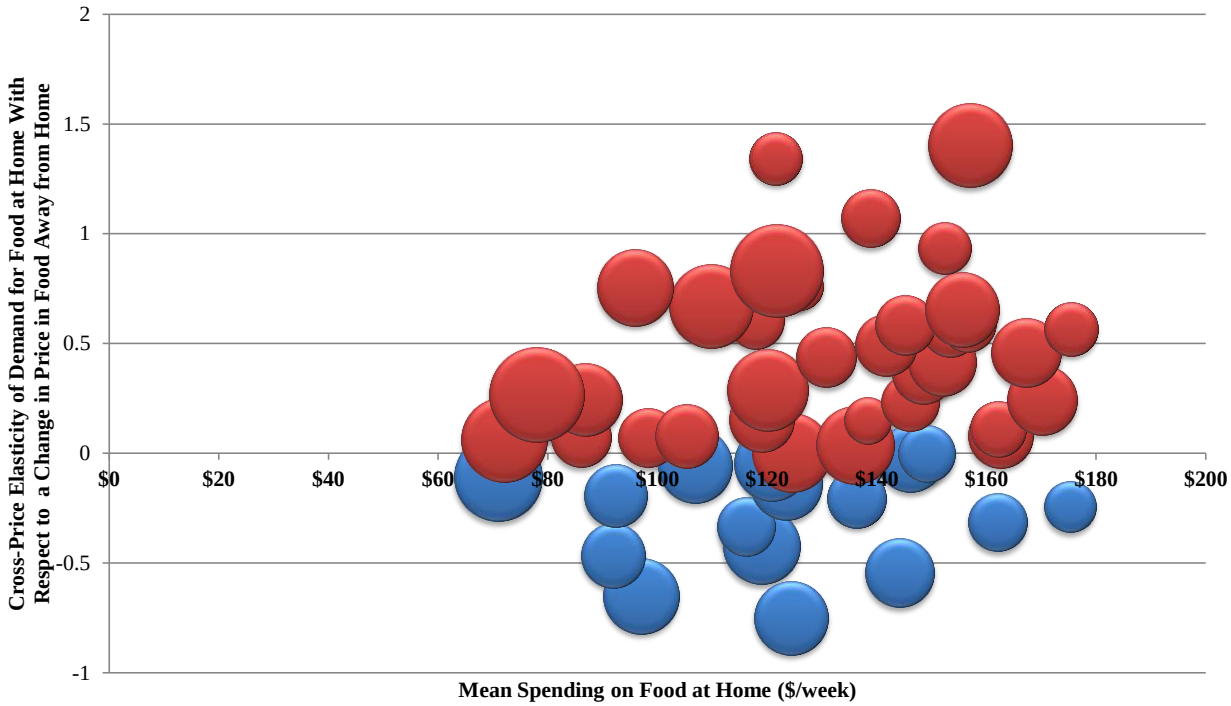

Figure 4. Relationship between mean spending on food at home and cross price elasticity of food at home with respect to a change in price of food away from home (note: size of circle represents number of consumers; red (blue) circles indicate substitute (complement) relationship between food at home and food away from home)

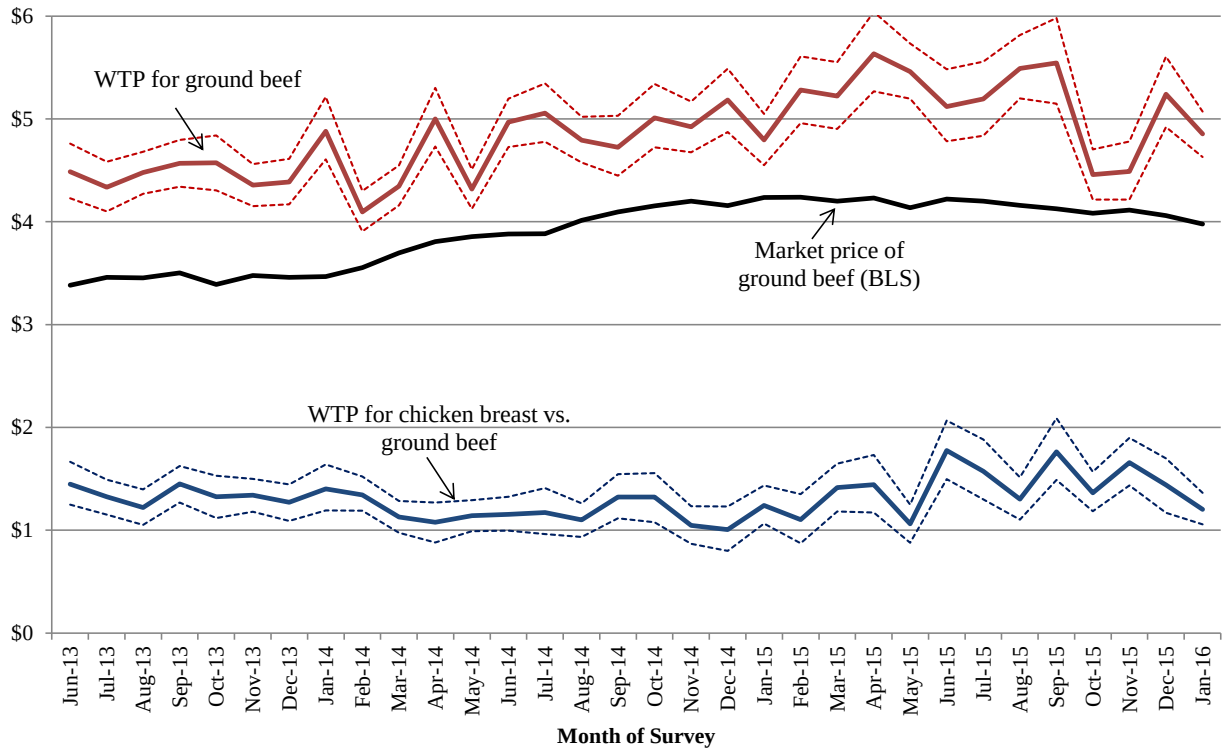

Figure 6. Variation in Willingness-to-Pay (WTP) from Repeated Choice Experiments Conducted in Different Months (note: dashed lines are 95% confidence intervals)

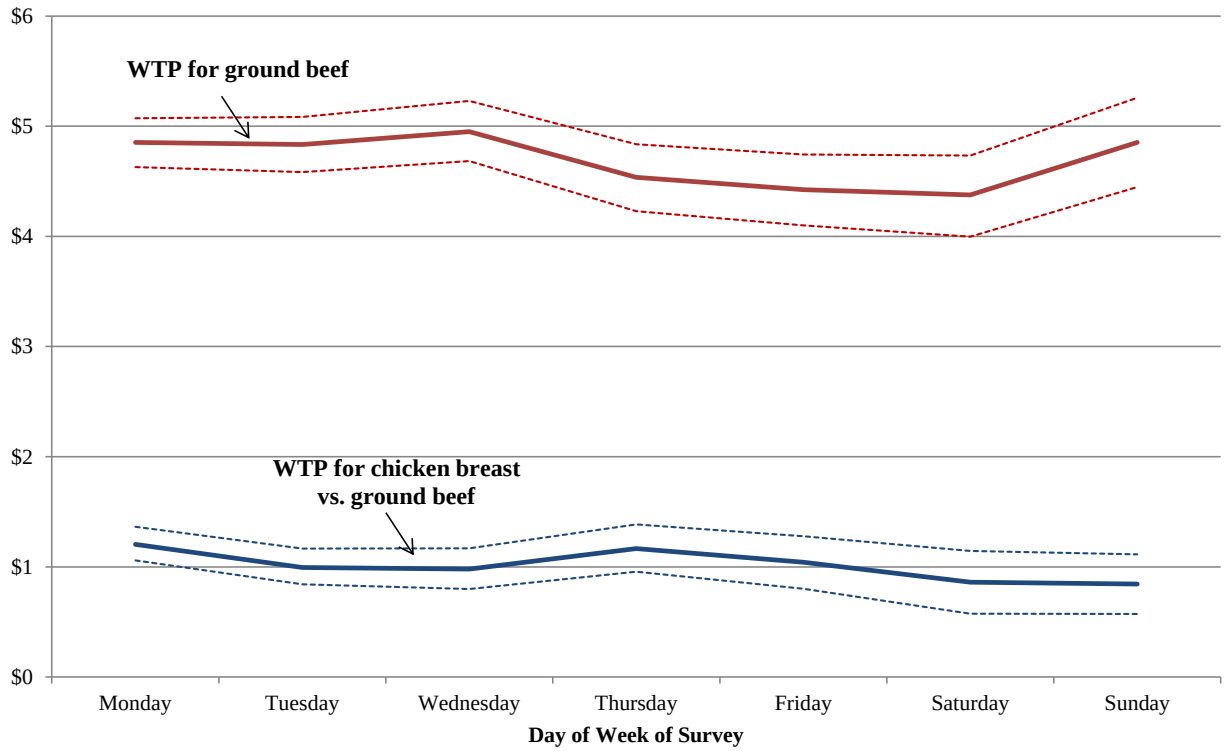

Figure 7. Variation in Willingness-to-Pay (WTP) from Repeated Choice Experiments Conducted on Different Days of the Week (note: dashed lines are 95% confidence intervals)

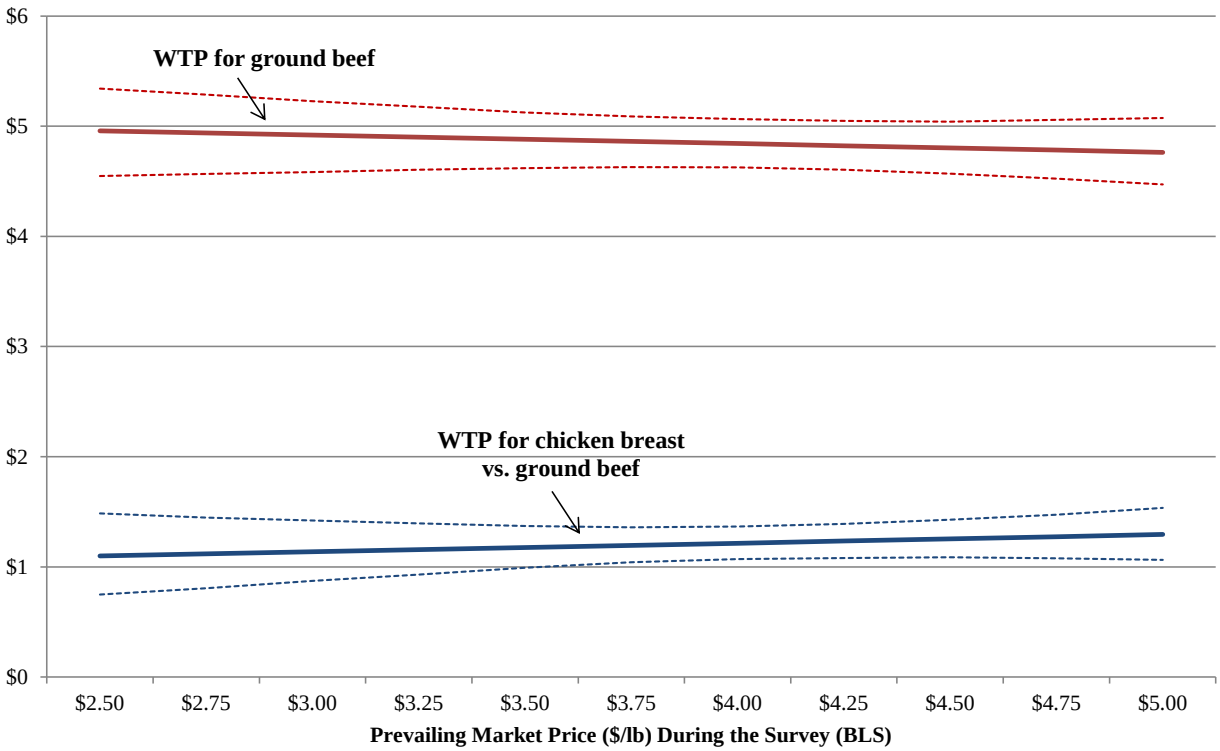

Figure 8. Variation in Willingness-to-Pay (WTP) from Repeated Choice Experiments Conducted at Times with Different Prevailing Market Prices for Ground Beef (note: dashed lines are 95% confidence intervals)

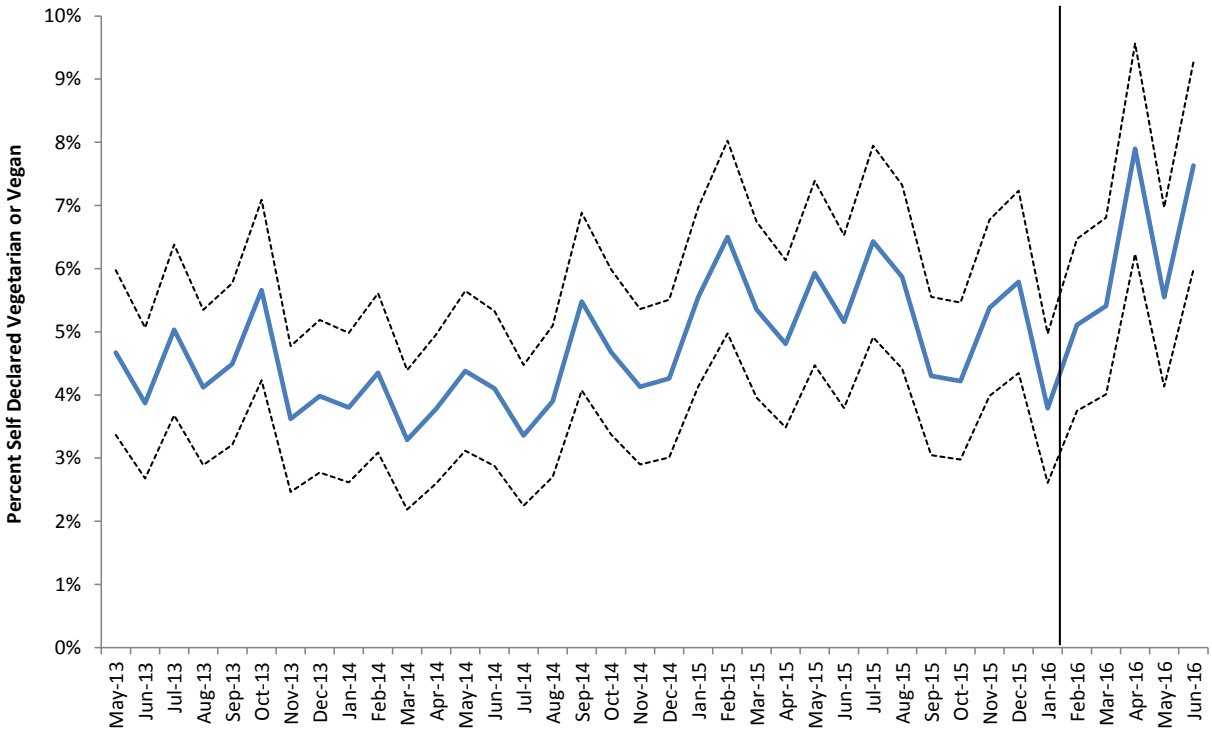

Figure 9. Estimated Percent of US Population Self-Identified as Vegan or Vegetarian from Food Demand Survey (FooDS) (note: estimates are weighted to match the demographics of the US population in terms of age, education, gender, and region of residence; dashed lines are 95% confidence intervals; data prior to the solid vertical line are used for estimation whereas data after the vertical line are used for validation)

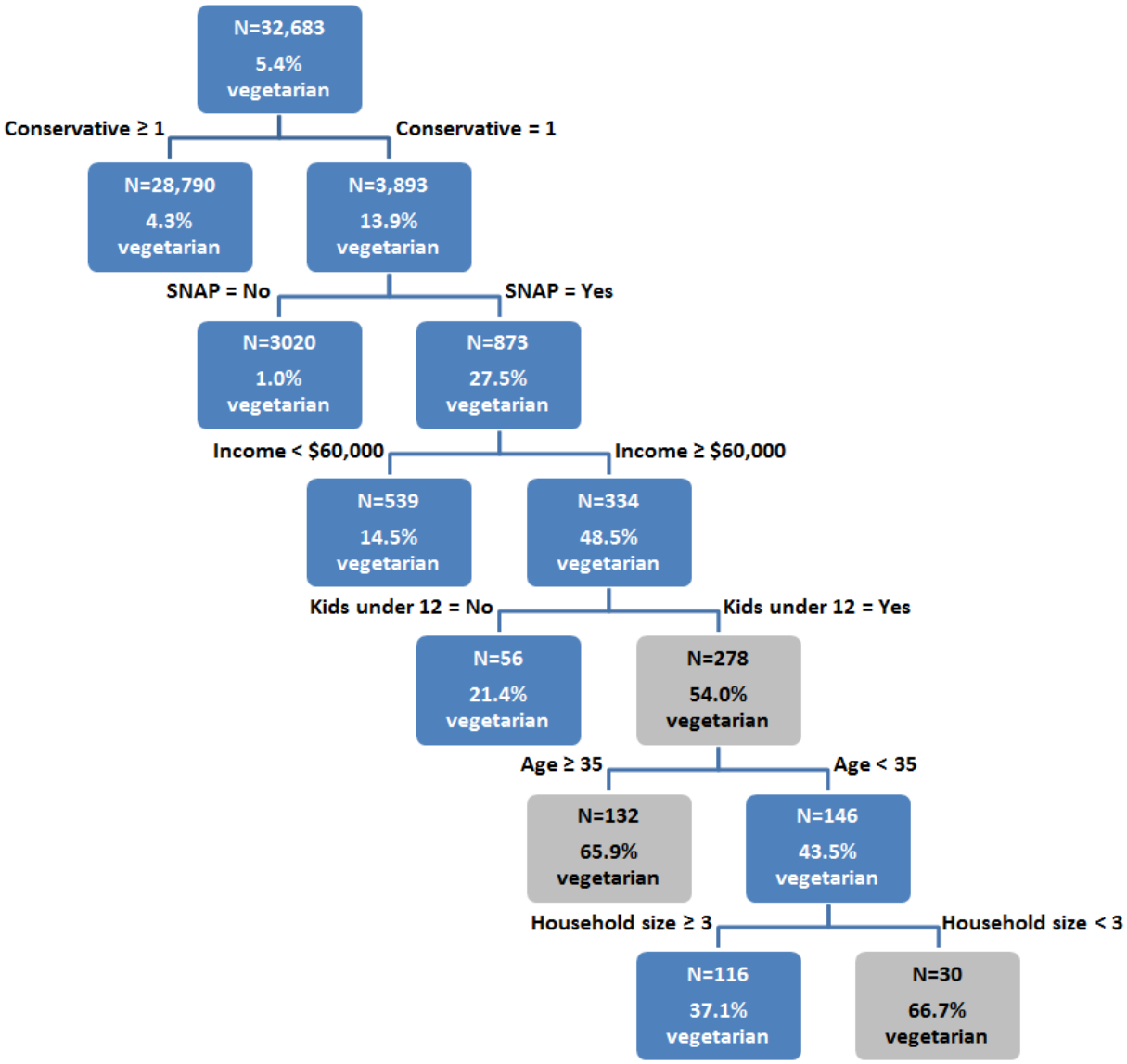

Figure 10. Classification Tree for Vegetarian or Vegan Status