Learning about Respondents' Characteristics Using Standard Exploratory Data Analysis (EDA) Tools November 2016

MoonJung Cho
Cho.Moon@bls.gov *

Larry Lang†

Key Words: Classification and regression tree, Multidimensional scaling, Survey nonresponse, Spectral clustering, U.S. International Price Program

1. Introduction

With aid of computing power, EDA has made remarkable advancements especially in visualization, clustering, and dimension reduction. We examined the characteristics of survey non-respondents using standard EDA tools such as classical multidimensional scaling (MDS), spectral clustering, and topological networks. In addition, we applied classification and regression tree methods to identify the important variables which could better assess and interpret nonresponse rates. We applied these tools to analyze non-respondents' characteristics at the initiation stage in the U.S. International Price Program Survey.

2. Illustrative Example: International Price Program

2.1 Sampling

The International Price Program (IPP) of the Bureau of Labor Statistics (BLS) produces two of the major price series for the United States: the import price indices and the export price indices. The IPP, as the primary source of data on price change in the foreign trade sector of the U.S. economy, publishes index estimates of price change for internationally traded goods. The target universe of the import and export price indices consists of all goods and services sold by U.S. residents to foreign buyers (exports) and purchased from abroad by U.S. residents (imports).

In particular, the IPP selects establishments (primary sampling units) within each broad product category (stratum), and then detailed product groups (secondary sampling units) within a selected establishment. The first and the second stage samplings are done

*U.S. Bureau of Labor Statistics, Office of Survey Methods Research, 2 Massachusetts Avenue NE, Washington, DC 20212

†U.S. Bureau of Labor Statistics, Office of Price and Living Conditions, 2 Massachusetts Avenue NE, Washington, DC 20212

in the national office. At the final stage of sampling, a field economist visits a selected establishment and initiates actual items.

We considered nonresponses based on a respondents' refusal at the initiation stage. Specifically, we identified the important explanatory factors which would better assess and interpret the nonresponse rate; and we studied the relationship between the nonresponse rate and establishment characteristics.

2.2 Data and Variable Descriptions

The IPP divides the import and export merchandise universes into two halves referred to as panels. Samples for one import half and one export half are selected each year and sent to the field offices for collection. Import Panel A consists of Food and Beverages, Crude Materials, Vehicles, and Miscellaneous Goods; while Import Panel B consists of Machinery, and Minerals and Chemicals. We applied exploratory data analysis tools to IPP M37 import data which were sampled from Panel B and initiated in 2011.

Table 1: Variable Descriptions

Name	Type (# of levels)	Description
Cert	binary	Certainty
Crank	ordinal (7)	Consistency rank
Dollar	continuous	Trading dollar value
Int	nominal (64)	Interviewers
List	binary	List aided
Qreq	discrete	Number of quotes requested
Region	nominal (9)	Field offices
Size	binary	Establishment size
Type	binary	Establishment type
Visit	binary	Visit by interviewer
Y	binary	Outcome (dependent var)
W	continuous	Weight

Cert Cert=1 if an establishment was selected with certainty.
(i.e., selection probability=1).

Crank Consistency rank ranges from 1-7, and is based on the number
of months and quarters an establishment traded a particular product category.
It is also based on how frequently an establishment traded.
Establishments with consistency rank 5-7 are called as consistent or frequent traders,
and establishments with consistency rank 1-4 as inconsistent or infrequent traders.

	The IPP selects approximately 99% of the establishments from the consistent traders and 1% from the inconsistent traders.
Dollar	Amount of trading in dollar value by an establishment.
Int	Interviewer ID.
List	Indicates whether a checklist exists for the classification group which requires the field economist to collect specific item characteristics.
Qreq	Number of quotes requested (Qreq) to an establishment is based on frequency of trade, size of establishment in stratum, and number of product category traded in stratum.
Region	Region includes eight regional offices and National office. Their description with codes are: (1) Boston; (2) New York; (3) Philadelphia; (4) Atlanta; (5) Chicago; (6) Dallas; (7) Kansas City; (8) San Francisco; (9) National Office.
Size	Size=1 if an establishment is classified as a large establishment.
Type	Type=1 if an establishment is a regular-type establishment. Non-regular-type establishments belong to the Foreign Trade Zone (FTZ).
Visit	Visit=1 if "Personal Visit Conducted"' is selected 'Yes'. That is if a field economist conducted an in-person initiation interview with an authorized establishment official.
Y	Outcome variable, Y, is 1 if an establishment responded for any quote.
W	Weight is standardized value of inverse of first stage selection probability.

3. Application on Binary Variables

We examined the following binary predictor variables in the data. The distributions of the

Table 2: Binary Predictor Variable

Code	Name	Description
1	**C**ert	Certainty
2	**L**ist	List aided
3	**S**ize	Establishment size
4	**T**ype	Establishment type
5	**V**isit	Visit by interviewer

binary predictor variables showed in Figure 1 that there were more non-certainty, smaller-size, regular-type establishments than certainty, larger-size, non-regular-type (FTZ) ones. They also showed that more interviews were conducted through personal visits and without any available checklist for the classification group. Furthermore, we grouped establishments according to response pattern: establishments which responded to all or some of requested

items and establishments which responded to none:

$$Y_i = \begin{cases} 1 & \text{if “all” or ”some”} \\ 0 & \text{if “none”} \end{cases}$$

3.1 Association analysis using multidimensional scaling

Using the multidimensional scaling (MDS) method, we examined a pairwise relationship among binary variables including both binary predictor variables and response variable. With information of pairwise relationship, the MDS finds coordinates and visualizes relationships in a low-dimensional space, while preserving the proximity relationships. In addition, the relationship doesn't have to be an Euclidean distance matrix and could be similarity or other measures. With χ^2 independence test, we tested the independence of each pair of variables; obtained χ^2 test statistics and p values on each combination of pairs. p values near zero cast doubts on the assumption of independence.

In keeping with Martinez and Cho (2015), Figure 2 presents results from multidimensional scaling. It displays distances between binary predictor variables and response variable in a three dimensional space. The smaller the p value that a pairwise independence test produced, the smaller the distance of the pair was measured and visualized. We observed that Size (3), Type (4), and Visit (5) variables are closer to the response variable (Y), while Cert (1) and List (2) are farther away from Y. The Figure 3 displays the color matrix using χ^2 independence test statistics. The brightness of the color indicates a stronger association. Since one was strongly associated with oneself, the diagonal elements displayed the brightest colors. Visit (5) showed a strong association with the response variable (6).

Literature has shown that clustering can have a substantial effect on the distribution of the standard Pearson χ^2 test statistic. Rao and Scott (1981) proposed a correction to χ^2 which required the knowledge of design effect for individual cells in the goodness of fit problem. Our analysis was conducted at the establishment level. Establishments are a primary sampling units (PSUs) and they are considered to be independent from each other within a stratum. It would be interesting to investigate further how to apply the standard Pearson χ^2 tests on the data from complex sample surveys.

3.2 Correlation analysis using topological network

We computed a pairwise correlation coefficient between each pair of binary variables. Graph command used a correlation coefficient matrix as an adjacency matrix, and plotted a graph from an adjacency matrix. With threshold > 0.1, only cases with correlation coefficient values greater than 0.1 were connected. Figure 4 showed that the response variable (6) had direct connections with Size (3), Type (4), or Visit (5), but it did not have a direct connection to Cert (1) and List (2).

With threshold > 0.2, Figure 4 showed that the response variable (6) had direct connections with Visit (5) but it was disconnected from all other variables.

3.3 Clustering analysis using pattern

We applied k-means clustering to our data with five binary variables to see whether there was any clustering. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. We didn't observe much separation through this method.

We divided our data into 32 patterns which were all possible patterns of five binary variables. For example, each of **C**ert,**L**ist, **S**ize, **T**ype, and **V**isit variables takes either 1 or 0. All possible combination of $\boxed{\mathbf{C}}\ \boxed{\mathbf{L}}\ \boxed{\mathbf{S}}\ \boxed{\mathbf{T}}\ \boxed{\mathbf{V}}$ could be mapped to 1 to 32 (decimal number) as shown below:

$$\begin{array}{lcl} \boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0} & \Longrightarrow & 1 \\ \boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{1} & \Longrightarrow & 2 \\ & \vdots & \\ \boxed{1}\boxed{1}\boxed{1}\boxed{1}\boxed{1} & \Longrightarrow & 32 \end{array}$$

Meanwhile the pattern labels, 1 through 32 (decimal number) could be mapped back to binary combination.

Figure 6 showed how many observations belonged to each pattern. For each pattern, the figure also showed the number of respondents and non-respondents. The fourth pattern has the most observations. We can map the decimal number 4 back to the binary combination which is $\boxed{0}\ \boxed{0}\ \boxed{0}\ \boxed{1}\ \boxed{1}$. It represents that non-certainty, no-list-aided, no-large-size, regular-type, and interviewer-visited case. Not surprisingly, these are the most common cases according to the Figure 1.

25th pattern, $\boxed{1}\ \boxed{1}\ \boxed{0}\ \boxed{0}\ \boxed{0}$, didn't have any observations. This is certainty, list-aided, no-large-size, no-regular-type, and no-interviewer-visited case. Note the odd matching of certainty and no-large-size establishment as we usually expect certainty establishments to be large-size ones.

We examined the patterns with high rates of nonresponse, e.g., the third, seventh, and 23rd patterns:

$$\begin{array}{lcl} \boxed{0}\boxed{0}\boxed{0}\boxed{1}\boxed{0} & \Longrightarrow & 3 \\ \boxed{0}\boxed{0}\boxed{1}\boxed{1}\boxed{0} & \Longrightarrow & 7 \\ \boxed{1}\boxed{0}\boxed{1}\boxed{1}\boxed{0} & \Longrightarrow & 23 \end{array}$$

We noted that the most common of these three patterns are no-list-aided (i.e., the second digit was 0), regular-type (i.e., the fourth digit was 1), no-interviewer-visited (i.e., the last digit was 0) case. Considering no-list-aided interviews occurred more often and regular-type establishments were more common, it followed that we observed no-list-aided interviews and regular-type establishments. It was remarkable that these high nonresponse patterns were all common in interviews without a personal visit when majority of interviews were conducted by personal visits.

For the overall relative frequency, taking the logarithm would make it easier to distinguish among patterns which have less than 200 observations as shown in Figure 7.

4. Which Weight to Use?

When the IPP computed the weight as detailed in the BLS Handbook of Methods (1997), the main components were trading dollar value and selection probability. Bobbitt et al. (2005) described some features of the weighting procedure used for the IPP. For the current discussion, the important element is sample selection probability at the first stage. We wanted to have weight (w_i) be close to the inverse of the first-stage sample selection probability (π_i). The sum of the sample selection probability of a stratum was not 1, but the number of sample units to be selected in the stratum. We chose to normalize this value. Specifically, for each stratum,

1. Compute the inverse of sample selection probability, $w_i^* = \pi_i^{-1}$
2. Sum individual weights of a stratum, $w^* = \sum_i w_i^*$
3. Finally, obtain new weights by standardizing previous weights $w_i = w_i^*/w^*$

Weighting made a considerable difference when we compared distributions between respondents and non-respondents.

5. Comparing Distributions of Continuous Variable

We compared distributions of trading dollar value of respondents and non-respondents using the two-sample Kolmogorov-Smirnov test and kernel smoothing function.

The two-sample Kolmogorov-Smirnov test evaluated the difference between the cumulative distribution functions (cdfs) of the distributions of the trading dollar values from non-respondents and respondents. The two-sided test uses the maximum absolute difference between the cdfs of the distributions of the trading dollar values. Suppose $F_1(x)$ is the cdf of the trading dollar values from respondents, $F_2(x)$ is the cdf of the trading dollar values from non-respondents, $F_{1m_1}(x)$ and $F_{2m_2}(x)$ are corresponding empirical cdfs,

m_1 and m_2 are number of respondents and non-respondents respectively. Then the test statistic is:

$$D_{m_1 m_2} = \left(\frac{m_1 m_2}{m_1 + m_2} \right)^{1/2} sup \ |F_{1m_1}(x) - F_{2m_2}(x)|.$$

In our discussion, $m1 = 2364, m2 = 1075, D_{m_1 m_2} = 0.0984$, p value $1.0811 * 10^{-6}$, and the test had rejected the null hypothesis at the 5% significance level.

In kernel smoothing function, the estimate was based on a normal kernel function, and was evaluated at equally-spaced points that covered the range of trading dollar value. It estimated the density at 100 points for *log* (trading dollar value), and vertical axis showed a probability density estimate for trading dollar value. The default bandwidths were the optimal for normal densities: bandwidths for respondent and non-respondents of *log* (**unweighted** trading dollar value) were 0.2154 and 0.2967 respectively; bandwidths for respondent and non-respondents of *log* (**weighted** trading dollar value) were 0.2200 and 0.2875 respectively.

As shown in Figure 7, for *log* (unweighted trading dollar value), respondents seemed to have a much larger trading dollar value than non-respondents. However, for *log* (weighted trading dollar value), the distribution of respondents seemed less variable than the distribution of non-respondents as shown in Figure 8. In fact, weighted trading dollar values from both respondents and non-respondents had distributions much less variable than the ones in unweighted case.

6. Importance Scoring using Classification and Regression Tree Methods

Classification and regression trees are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. Among various tree methods, we chose to use GUIDE (Loh, 2002): GUIDE stands for generalized, unbiased, interaction detection and estimation. Specifically, GUIDE has following advantages: selection unbiasedness; fast computation speed; missing value treatment.

6.1 Importance Scoring

GUIDE has a facility to rank variables in order of their importance. In addition, it provides a threshold for distinguishing the important variables from the unimportant ones. It treats any variables with score 1 or less as unimportant ones. Table 3 showed that GUIDE ranked Visit and Interviewer variables as top two important variables. Although importance scores could rank order the variables, they did not explain how the variables influence the predictions. Single-tree classification and regression models can provide their model interpretability, and it is the biggest advantage of using single-tree models.

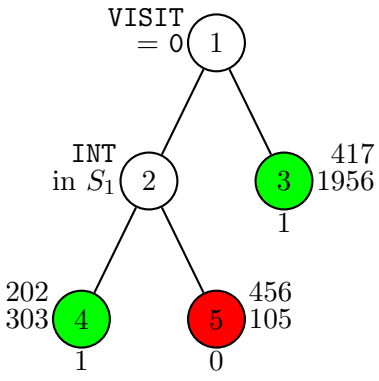
Table 3: Important Variable Rank

Rank	Score	Variable
1	97.3	Visit
2	52.7	Interviewer
3	18.3	Region
4	12.4	Size
5	8.2	Qreq
6	6.3	Dollar
7	1.4	Section
8	1.2	Certainty
9	1.0	List
10	0.3	Type
11	0.0	Crank

6.2 Classification Tree Methods

GUIDE classification tree method is accomplished by carrying out the following steps: (1) select the most significant X variable through χ^2 independence test to split a node, (2) find the split point or split set for X to minimize the Gini index, (3) recursively repeat steps (1) and (2) until too few observations in each node, (4) and use the CART method to prune the tree to minimize cross-validation (CV) estimate of misclassification cost. There are several ways to control the tree size. One way is by setting a minimum sample size per node: the larger minimum sample sample size, the smaller tree size one gets. As the minimum sample size per node gets smaller, tree size gets bigger and interpretability rapidly diminishes.

Classification Tree (a minimum sample size per node greater than 100)

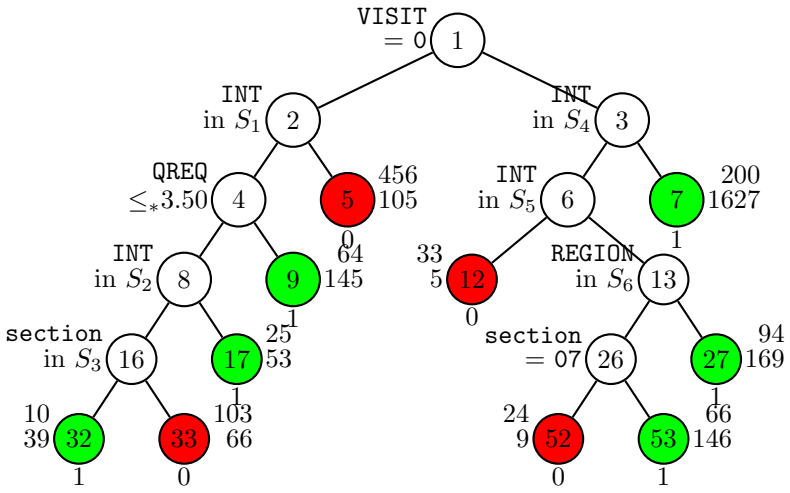


At each node, an observation goes left if and only if the condition is satisfied.

At node 1, an observation went to node 3 if an interview was conducted by a personal visit. GUIDE predicted establishments of node 3 as respondents. At node 1, an observation went to node 2 if an interview was not conducted by a personal visit.

It showed that an interview by a personal visit was an important factor to get responses from establishments. In addition, interviewers of S_1 were still able to have many establishments responded even when interviews were conducted without personal visits.

Classification Tree (minimum sample size per node greater than 10)



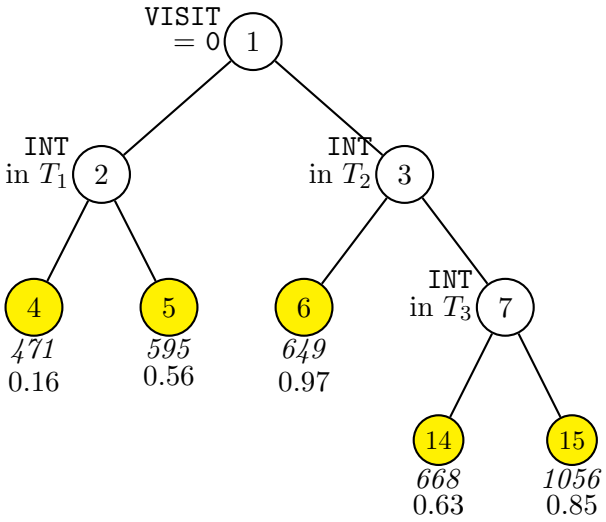
As minimum sample size per node gets smaller, the tree split deeper. Note that GUIDE predicted establishments of node 12 as non-respondents: the majority of establishments interviewed by interviewers of S_5 did not respond even though interviews were conducted by personal visits.

In addition, the original dataset had distinct labels for 64 interviewers. Application of the GUIDE procedure led to data-driven groupings of these interviewers into sets that here are labeled S_1 through S_6 shown in both graphs above. Note that these sets are partially overlapping. In particular, $S_3 \in S_2 \in S_1$, $S_5 \in S_4$, and $S_6 = S_4 \setminus S_5$.

6.3 Regression Tree Methods

Regression tree methods could tell more about the data visually. GUIDE carries out the following steps recursively at each node: (1) fit a model to the training data, (2) cross-tabulate the signs of the residuals with each predictor variable to find the one with the most significant chi-square statistic, and (3) search for the best split on the selected variable, using the appropriate loss function. After a large tree is constructed, it is pruned with the cross-validation method of CART.

Piecewise Linear Least-Squares Regression Tree



Sample sizes (*in italics*) and means of `Y` are printed below the nodes. Also, a separate application of regression procedures in GUIDE led to a different grouping of interviewers into sets labeled T_1 through T_3. The regression tree also showed that an interview by a personal visit was an important factor to get responses from establishments. Furthermore, response rates varied considerably depending on interviewers in both personally visited and not-visited cases.

7. Discussion

In household interview surveys, Groves and Couper (1998) put greater emphasis on the interaction between interviewer and householder during the survey. Our study of the IPP (which is an establishment interview survey) pointed in the same direction. This leads to working closely with field offices in studying nonresponses. By gathering more factors on dynamics in the field, we may be able to understand the nature of nonresponse to a greater degree.

8. Acknowledgment

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics. The authors thank Jeff Blaha and Jim Himelein for helpful comments on the International Price Program; and Wei-Yin Loh for helpful suggestions on classification and regression tree methods. We are especially grateful to John Eltinge for reading the entire manuscript and making helpful suggestions.

REFERENCES

BOBBITT, P. A., CHO, M. J., and EDDY, R. M. (2005), "Comparing Weighting Methods in the International Price Program," *Proceedings of the American Statistical Association*, Government Statistics Section [CD-ROM], 1006-1014.

BOBBITT, P. A., PABEN, S. P., CHO, M. J., HIMELEIN, J. A., CHEN, T-C., and ERNST, L. R. (2007), "Application of the Bootstrap Method in the International Price Program," *Proceedings of the American Statistical Association*, Survey Research Methods Section [CD-ROM], 2910-2917.

BUREAU OF LABOR STATISTICS (1997), *BLS Handbook of Methods*, International Price Indexes, Chapter 15 of BLS Handbook of Methods, U.S. Department of Labor,
Available at http://www.bls.gov/opub/hom/home.htm (accessed September 2016).

CHO, M. J., CHEN, T-C, BOBBITT, P. A., HIMELEIN, J. A., PABEN, S. P., ERNST, L. R., and ELTINGE, J. L.(2007), "Comparison of Simulation Methods Using Historical Data in the U.S. International Price Program," *Proceedings of the American Statistical Association*, Third International Conference on Establishment Surveys [CD-ROM], 248-255.

CHO, M. J. and ELTINGE, J. L.(2009). "Identification of Functional Forms and Predictor Variables in Generalized Variance Functions for Price Indices," *Proceedings of the American Statistical Association*, Third International Conference on Establishment Surveys [CD-ROM], 1393-1407.

GROVES, R. M. and COUPER, M. P. (1998), *Nonresponse in Household Interview Surveys*, John Wiley and Sons, New York.

LOH, W.-Y. (2002), "Regression trees with unbiased variable selection and interaction detection," *Statistica Sinica*, vol. 12, 361-386.

Loh, W.-Y. (2009), "Improving the precision of classification trees," *Annals of Applied Statistics*," vol. 3, 1710-1737.

LOH, W.-Y. (2014), "Fifty Years of Classification and Regression Trees," *International Statistical Review*, 0, 0, 120.

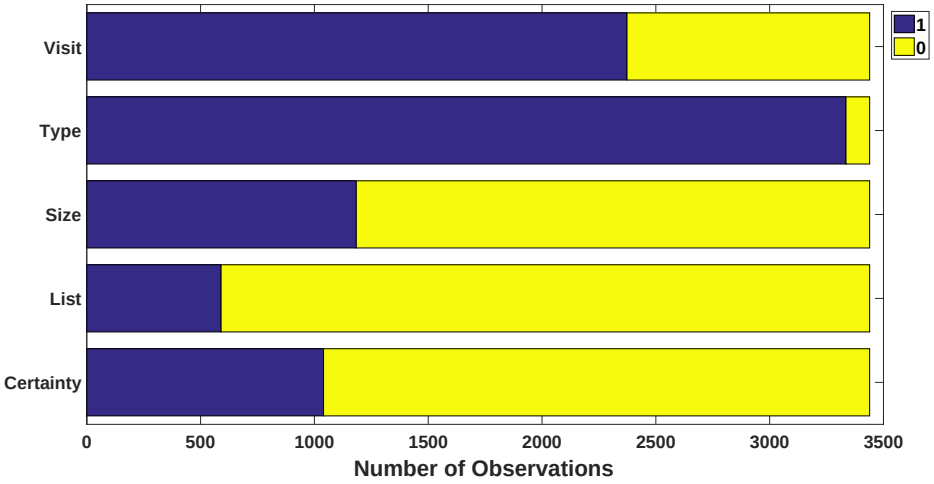
MARTINEZ, W. L. and CHO, M. J. (2015), *Statistics In Matlab A Primer*, CRC Press, New York.

MARTINEZ, W. L. and MARTINEZ, A. R. (2011), *Exploratory Data Analysis With Matlab*, Second Edition, CRC Press, New York.

POWERS, R., ELTINGE, J. L., and CHO, M. J. (2006), "Evaluation of the Detectability and Inferential Impact of Nonresponse Bias in Establishment Surveys," *Proceedings of the American Statistical Association*, Survey Research Methods Section [CD-ROM], 3577-3583.

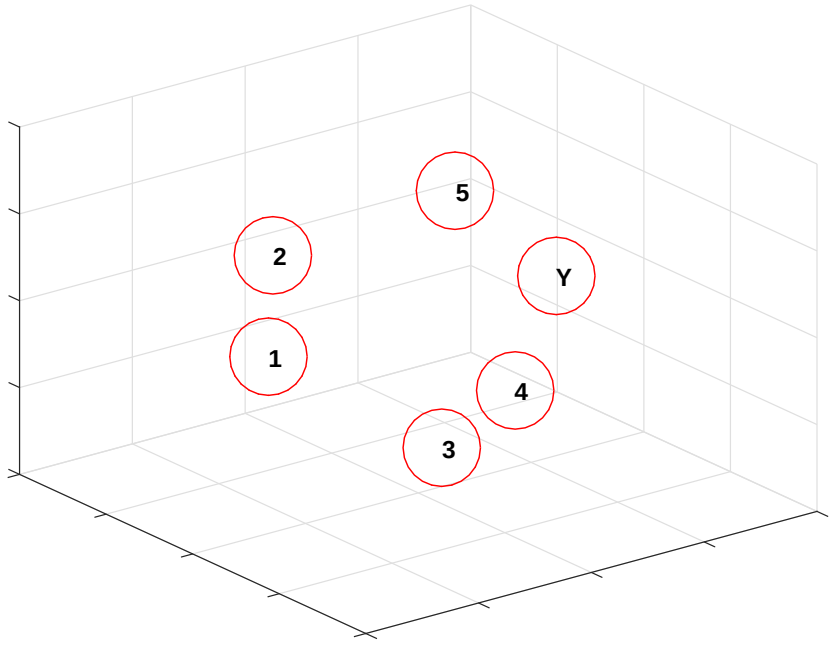
RAO, J. N. K. and SCOTT, A. J. (1981). "The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence of two-way table," *J. Amer. Statist. Assoc.*, 76, 221-230.

Figure 1: Distribution of Each Binary Variable (M37)



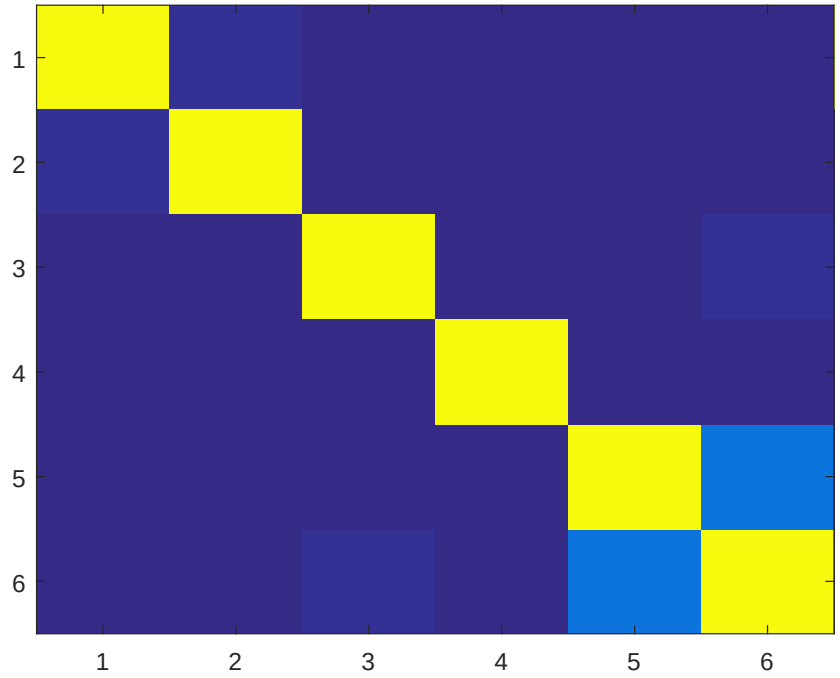
Notes: Description of binary variables can be found in Table 2. Note that there were more non-certainty, smaller-size, regular-type establishments than certainty, larger-size, non-regular-type (FTZ) ones. They also showed that more interviews were conducted through personal visits and without any available checklist for the classification group.

Figure 2: Distance by p-value from χ^2 Independence Test



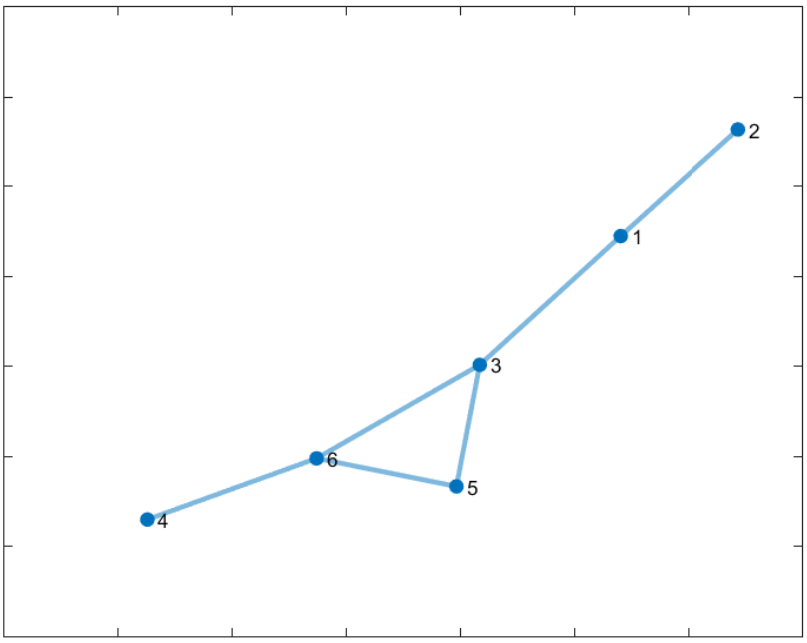
Notes: The labels 1 *thorough* 5 *are the five variables described in Table 2, and* Y *represents the outcome variable. Note that Size (3), Type (4), and Visit (5) variables are closer to the response variable (Y), while Cert (1) and List (2) are farther away from* Y*.*

Figure 3: Association by Stat from χ^2 Independence Test



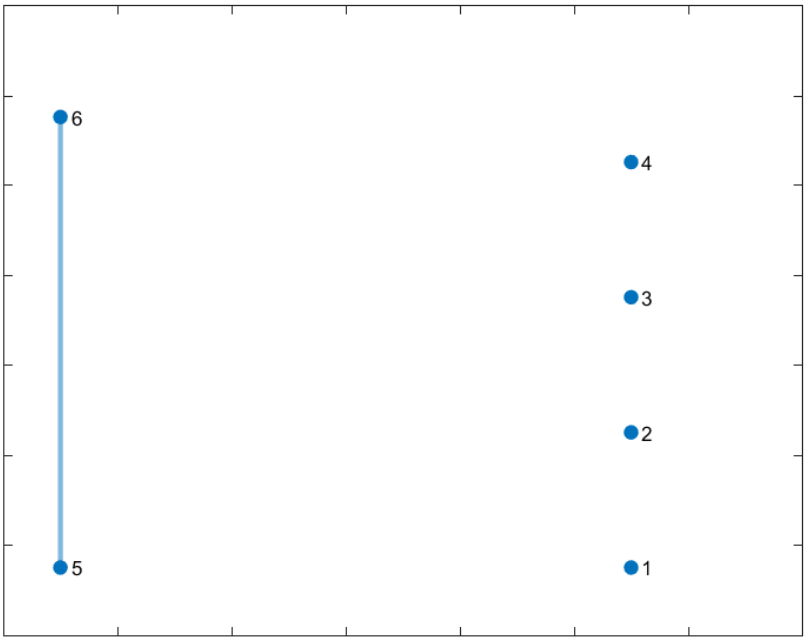
Notes: The labels 1 *thorough* 5 *are the five variables described in Table 2, and* 6 *represents the outcome variable. Note that Visit (5) showed a strong association with the response variable (6).*

Figure 4: Correlation among Binary Variables (threshold > 0.1)



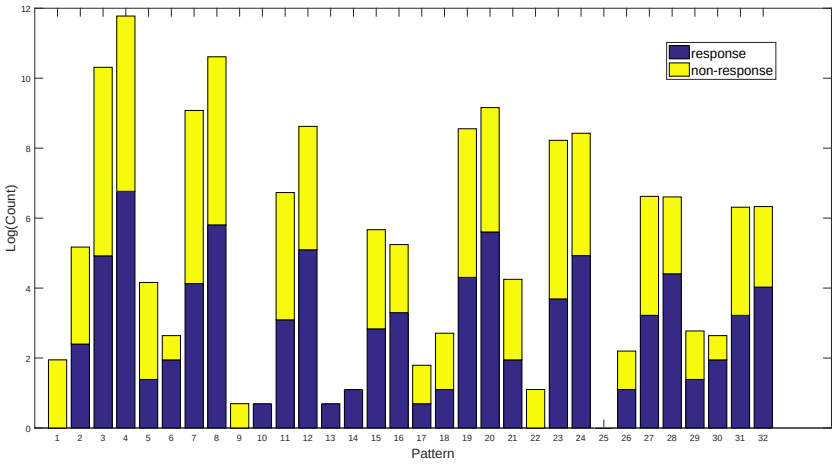
Notes: The labels 1 *thorough* 5 *are the five variables described in Table 2, and* 6 *represents the outcome variable. Note that the response variable (6) had direct connections with Size (3), Type (4), or Visit (5), but it did not have a direct connection to Cert (1) and List (2).*

Figure 5: Correlation among Binary Variables (threshold > 0.2)



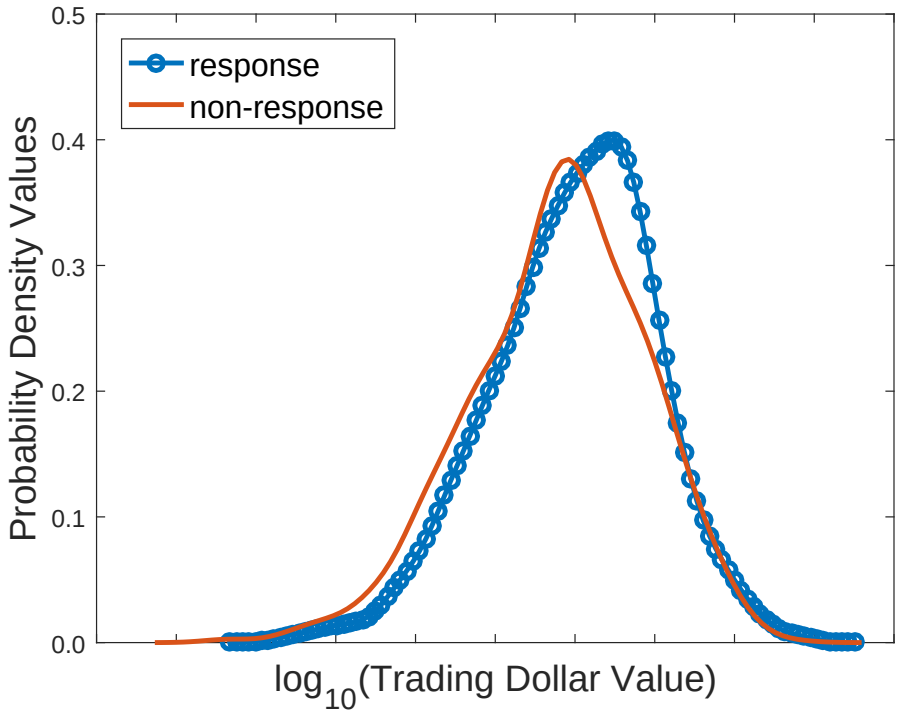
Notes: The labels 1 *thorough* 5 *are the five variables described in Table 2, and* 6 *represents the outcome variable. With threshold* > 0.2*, the response variable (6) had direct connections with Visit (5) but it was disconnected from all other binary predictor variables.*

Figure 6: Log of Frequency of Patterns: Binary Predictor Variables



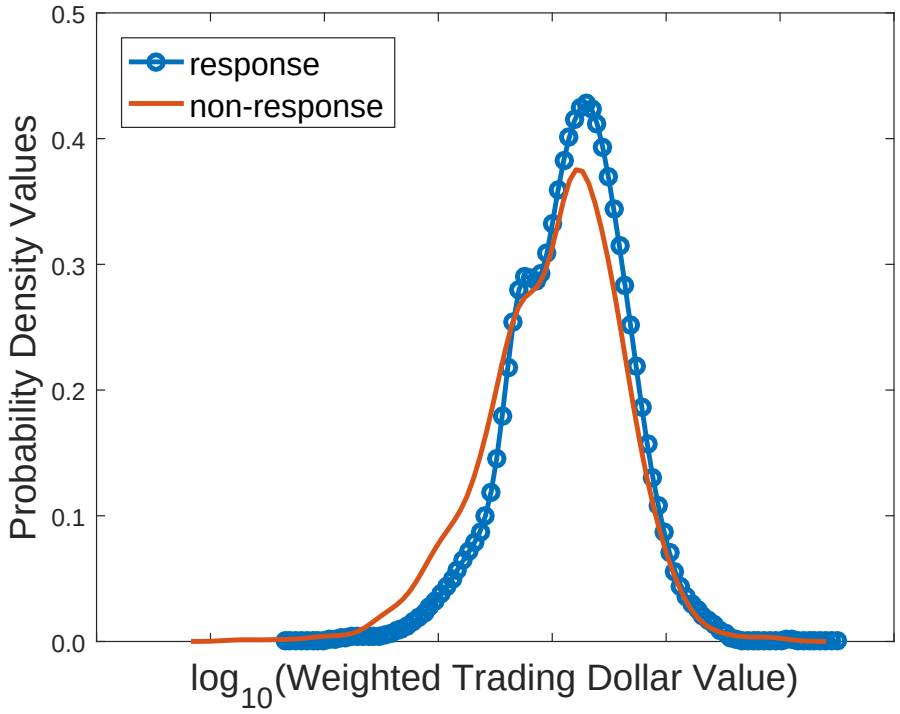
Notes: Formation and labeling of patterns can be found in Section 3.3.

Figure 7: Estimated Densities for Respondents and Non-Respondents



Notes: For $\log_{10}$(unweighted trading dollar value), respondents seemed to have a much larger trading dollar value than non-respondents.

Figure 8: Estimated Densities for Respondents and Non-Respondents



Notes: For log_{10}(weighted trading dollar value), the distribution of respondents seemed less variable than the distribution of non-respondents. In fact, weighted trading dollar values from both respondents and non-respondents had distributions much less variable than the ones in unweighted case.