Developing a Model-Based Approach to Identifying and Correcting Misclassified Data

Clayton Knappenberger

March 16, 2018

CE Surveys Program Report Series

Table of Contents

Introduction

Classification error occurs when a survey response is reported (or is recorded) in an incorrect category. For example, a respondent might be prompted about computer expenditures, but then recall and report a tablet expenditure. This is a common, but rarely studied, problem in survey data. Using the Consumer Expenditure Surveys' definitions, classification error is a form of measurement error as the response provided is different from the true value of the measurement (Gonzalez et al., 2009). Correlation of classification error with other explanatory variables of interest can cause bias in regression estimates or when calculating population estimates based on survey weights.

Survey practitioners often have limited tools at their disposal for identifying and correcting cases of classification error. Typically researchers who studied classification error have relied on subsequent re-interviews or on administrative data to estimate the presence and extent of classification error. The first approach was adopted by Feng and Hu (2013) in their analysis of the impact of labor force status misclassification on the unemployment rate. Food and Nutrition Services administrative records were used in Bollinger and David's examination of misclassification in the 1984 *Survey of Income and Program Participation* (1997). These and similar studies assumed auxiliary data sources report the "true" classification of one or more variables of interest. This approach is appropriate post-processing and where such data are available, but is less useful to survey producers who may wish to identify and mitigate classification errors during survey processing.

This report presents a novel approach to the problem of classification error in the Bureau of Labor Statistics' Consumer Expenditure Quarterly Interview Survey (CEQ) data. Previously, Consumer Expenditure Surveys staff relied only on outlier detection methods to identify and correct respondent-reported misclassified expenditures. These methods assume that two different item categories have different cost distributions, but in many cases this is not a reasonable assumption. I use respondent

provided text descriptions of purchases to train a predictive model whose output is a predicted item category for each expenditure. This prediction is then compared to the reported category to identify likely cases of respondent-reported item misclassification.

In this report, I explain this new process called Item Code Estimation (ICE) that is currently being adopted within the Consumer Expenditure Surveys to improve upon existing data review procedures. In addition, I estimate a lower bound on classification error for a single expenditure category and generate new expenditure estimates for specific items within that expenditure category after correcting for identified misclassification. ICE is currently being used in production on a single expenditure category and can be expanded for use across CEQ expenditure categories.

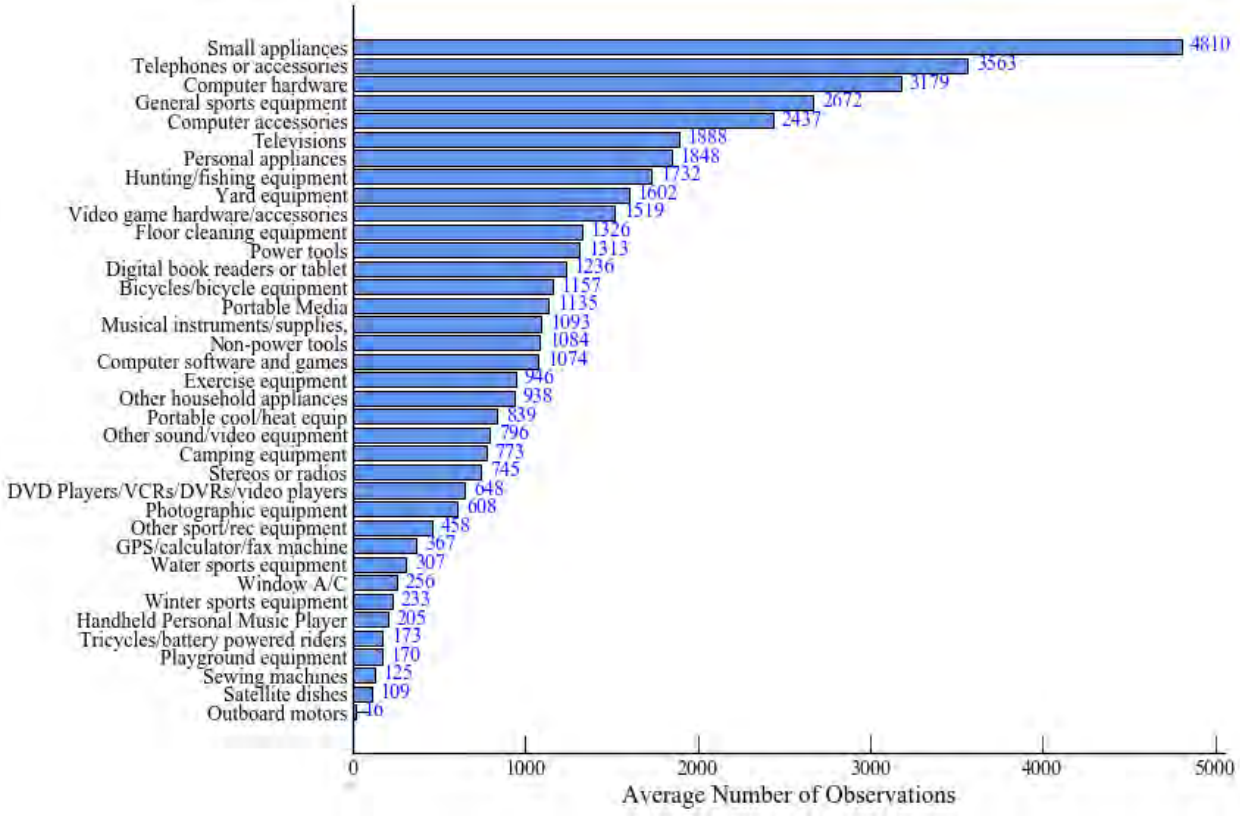
Analytics Base Table

Analytics Base Table (ABT) is a term used in prediction settings to refer to the single file containing all the information used for training and testing a predictive model. Typically in an ABT each row represents a distinct observation. The source files for my ABTs are the CEQ production microdata files that are based on collection quarter. Using the 11 quarters of data from 2013Q2 – 2015Q4 allows me to construct four rolling 8 quarter ABTs for the 2015 collection year. The first seven quarters in each ABT are designated the training dataset and the remaining quarter is set aside as the test dataset. This rolling process is done to simulate the quarterly nature of CE expenditure processing.

I use seven quarters of training data for two reasons. First, seven quarters allows me to capture monthly seasonal changes in expenditures. Second, many item categories have relatively few observations in a single collection quarter. Even bringing in this much data does not always resolve this problem. In Figure 1 below, I have graphed the average number of expenditures for the four ICE ABTs in 2015. The category with the fewest expenditures, 530 – outboard motors, has an average of only 16

observations in the ICE source data. However, the category with the next fewest expenditures, 670 – satellite dishes/receivers/accessories, has on average over 100 observations in the ICE source data.

Figure 1: Item Category Counts for 2015 CEQ

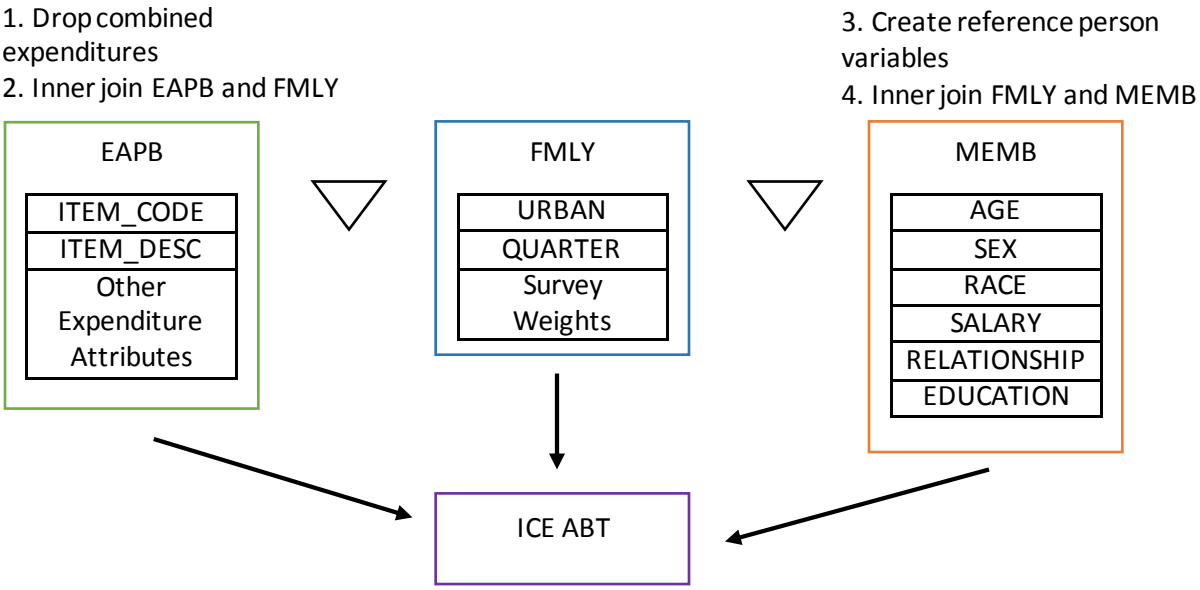


For each quarter the ICE ABT is formed by merging three different CEQ files. The components of each ICE ABT are: (1) the data collected in Section 6 Part B of the CEQ interview: Appliances, household equipment, and other selected items (EAPB) which contains expenditure level information including the item code, the text description of the item, and other variables that can help classify the purchase; (2) the Family file (FMLY) which offers household[1] demographics, the survey weights, and the quarter and

[1] Technically, CE collects expenditure information for consumer units rather than households. A consumer unit is defined as: 1) all members of a particular housing unit who are related by blood, marriage, adoption, or some other legal arrangement, such as foster children; 2) a person living alone or sharing a household with others, or living as a roomer in a private home, lodging house, or in permanent living quarters in a hotel or motel, but who is financially independent; or 3) two or more unrelated persons living together who pool their income to make joint expenditure decisions. In most cases the consumer unit is equivalent to the household. For this reason, and to aid the reader in interpreting these results, I refer to households in place of consumer units throughout this article.

year of the interview; and (3) the Member file (MEMB) which provides member level demographic and income variables that I later recode into household variables.

Figure 2: Creation of the ICE ABT



The basic steps to create the ICE ABT are outlined above in Figure 2. First, I drop combined expenditures[2] from the EAPB file as their descriptions would add unnecessary noise to the ICE model and because combined items are more likely to have truncated descriptions. Second I transform the EAPB variables into either binary or standard normal variables (i.e. Z-Scores transformation). Third, I perform an inner join of the EAPB and FMLY files. Fourth I perform variable transformations on the selected member-level variables found in the MEMB file with the goal of transforming them into

[2] A combined expenditure is any expenditure where the respondent reported multiple items in a single expenditure. For example: the respondent might report spending $100 on "a toaster and a blender." This is a combined expenditure because the respondent included two items (toaster and blender) within a single expenditure.

standardized household-level demographic and income variables. These transformations are outlined in Table 1.

Table 1: Variable Transformations

Variable Name	Description	Table	Calculations and Transformations
GFTCMIN	Gift or not gift	EAPB	Binary transformation[3]
MIN_MO	Month of purchase	EAPB	Binary transformation
PURCH	Whether the item was purchased or rented	EAPB	if MINPURX == 'B' then PURCH = 0; else PURCH = 1;
DK_PURCH	Binary indicating that the respondent did not know or refused to say whether purchased or rented	EAPB	if MINPURX == 'F' then DK_PURCH = 1; else DK_PURCH = 0;
MINPURX	Standardized expenditure on item	EAPB	1. if MINPURX in ('B', 'F') then MINPURX = 0; else MINPURX = MINPURX; 2. Standardized[4]
DK_RENT	Binary indicating that the respondent did not know or refused to say whether purchased or rented	EAPB	if MINRENTX == 'F' then DK_RENT = 1; else DK_RENT = 0;
MINRENTX	Standardized rental expenditure on item	EAPB	1. If MINRENTX in ('B', 'F') then MINRENTX = 0; else MINRENTX = MINRENTX; 2. Standardized
INSTLSCR	If the expenditure includes installation costs	EAPB	Binary transformation
MINTAX	Whether sales tax was applied to expenditure	EAPB	Binary transformation
URBAN	Whether the household is categorized as living in an urban area or not	FMLY	Binary transformation
AGE_REF	Age of the reference person	MEMB	1. Select AGE as AGE_REF From MEMB Where CU_CODE = 1 2. Standardized
SEX_REF	Sex of the reference person	MEMB	1. Select SEX as SEX_REF From MEMB Where CU_CODE = 1 2. Binary transformation
SALARYX	Total reported salary of all household members	MEMB	1. If SALARYX in ('A', 'B', 'F', 'G') then SALARYX = 0; else SALARYX = SALARYX; 2. Sum of all member salaries

[3] The process of binary transformation involves taking a categorical variable and turning into $(k-1)$ binary variables where k is the number of categories.

[4] Standardization is performed on continuous variables so that they have equivalent scales (standard normal).

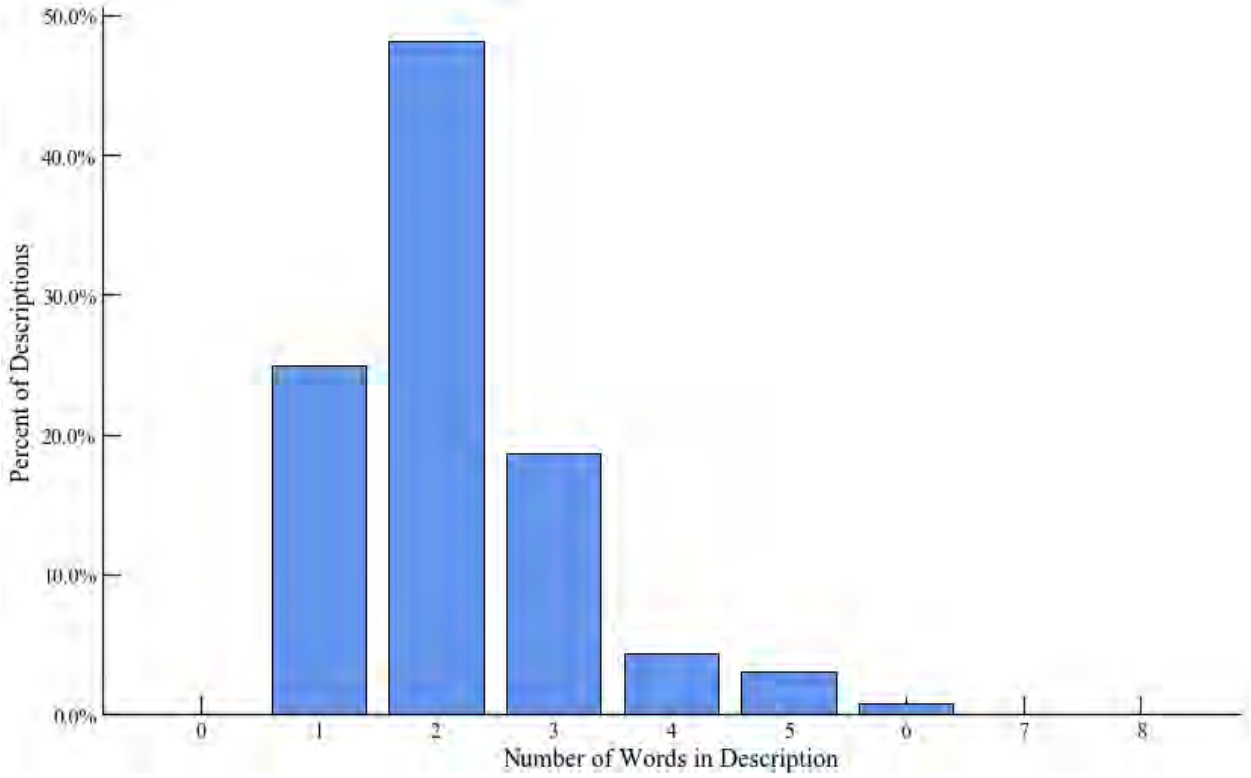
Variable Name	Description	Table	Calculations and Transformations
REF_RACE	Race of the reference person	MEMB	1. Select MEMBRACE as REF_RACE From MEMB Where CU_CODE = 1 2. Binary transformation
HIGH_EDU	Highest level of education attained by any household member	MEMB	1. Select max(EDUCA) as HIGH_EDU From MEMB Group by NEWID; 2. Standardized
FAM_SIZE	Number of household members	MEMB	1. Select count(NEWID) as FAM_SIZE From MEMB Group by NEWID; 2. Standardized

The Text Descriptions

In addition to the above described variable transformations, the text item descriptions must be cleaned and transformed to be used as independent variables for predicting item codes. During the CEQ interview as respondents are prompted to report expenditures within a given item category, the interviewer is able to record a brief text description of the item the respondent is reporting. If the interviewer fails to record a description, the instrument fills in a default description based on the item category it was reported in. For example, suppose the interviewer asks the respondents if they purchased anything in the item category "small kitchen appliances". The respondent may inadvertently report the purchase of a vacuum cleaner. If the interviewer writes a description it could be something like "vacuum cleaner," but if there is no written input, the instrument fills in the default description of "small kitchen appliances."

Typically these descriptions provided by the interviewer are very short, the average length being only two words while the maximum description length found in the source data is eight words. As can be seen in Figure 3, the majority of item descriptions contains three or fewer words. For my analysis shorter descriptions are often preferable as they require less cleaning, and likely contain information more pertinent to identifying the expenditure. Longer descriptions are more likely to contain ancillary words that may make predicting the correct item codes more difficult.

Figure 3: Percent Distribution of Description Lengths



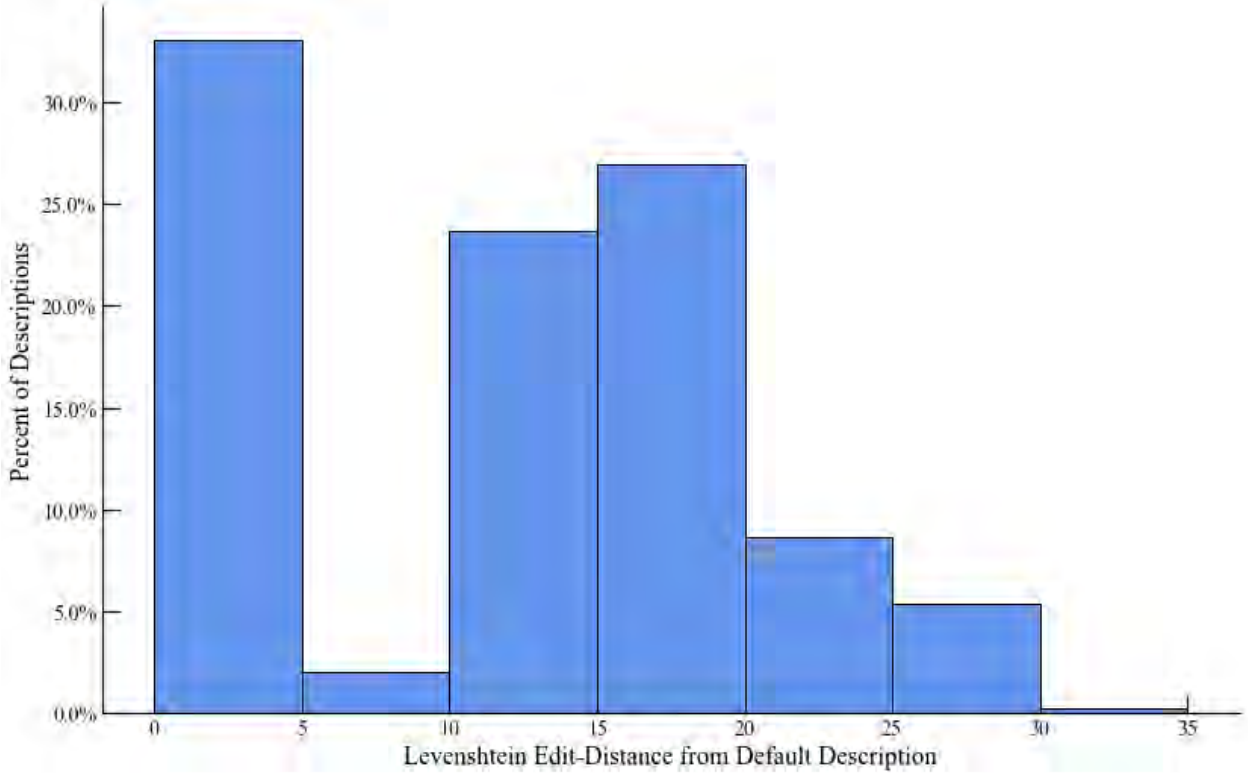
As noted, if the interviewer does not input a description, a default description is automatically added to the record based on the item category selected. Overall, approximately 33 percent of the descriptions provided use the default description, although this varies from 14 to 89 percent across item categories. The roughly 67 percent of descriptions that did not contain the default description are non-trivially different from the default description. This suggests that when the interviewer recorded an item description, it contained more specific information beyond what would have been provided by the default.

The Levenshtein Edit-Distance is a distance function used in natural language processing and is calculated as the number of character operations (substitutions, additions, and deletions) that would need to be changed to transform one string into another. The Levenshtein Edit-Distance, $L(x, y)$, satisfies the definition of a distance function as:

1. $L(x, y) > 0$ for all strings x and y
2. $L(x, y) = 0 \Leftrightarrow x = y$
3. $L(x, y) = L(y, x)$
4. $L(x, y) \leq L(x, y) + L(y, z)$

In Figure 4, I show the distribution of the Levenshtein Edit-Distances of the provided description from the default description. Over one third of the descriptions have a distance of less than 5 which indicates that in these cases either the default description or a description that was only slightly different from the default was used. For the entire distribution the average distance between the provided description and the default description is 11 characters. However, for the records that did not use the default description the average distance looks to be around 18 characters. Given that the average number of words in a description is 2, this signifies that when interviewers input a description that was different from the default description, it was markedly different from the default.

Figure 4: Percent Distribution of Levenshtein Edit-Distances



Despite the average description having only 2 words (Figure 3), the sheer number of items results in over 9,000 unique words. Text cleaning is therefore necessary to reduce the number of unique words. This process can range from the very simple to the extremely complex. I use a relatively simple five step process aimed at distilling the item descriptions into only information about the item purchased – potentially leaving out things like how many were purchased and who it was for.

1. Convert all letters to lowercase
2. Remove punctuation
3. Convert contractions into two words
4. Remove stop words[5]
5. Remove numbers and number words

The text cleaning process could be made more sophisticated if desired. For example, I made no effort to correct misspellings or to enforce a standard form for words like television/t.v./tv. I also did not attempt to address the issues presented by having different prefixes and suffixes attached to similar words. For example: "digital converter box", "digital conversion box", and "digital converting box" each refer to the same item and the model's prediction accuracy could be improved by replacing the three different forms of "convert" with a single example.[6] Despite the relative simplicity of this approach the total number of unique words present is reduced from over 9,000 to around 4,000 unique words in each of the training data sets.

With the cleaned text data, I must next convert it into a format that can be used in a predictive model. The approach that I employ with ICE is called the "bag of words" wherein each unique word that appears is represented in the model as an independent variable. The most naïve form the "bag of words" can take is to encode each word variable as a binary 1 or 0 if the word is present in a given item

[5] Many software packages that allow for natural language processing include standard lists of stop words. Examples of stop words depend on the context, but commonly include words like *a*, *and*, *the*, and *it*. For this analysis I used Python's Natural Language Toolkit's (NLTK) stop word list.
[6] The process of removing word prefixes and suffixes is called stemming and NLTK provides access to many different stemming algorithms, one of the most popular being Porter's.

description or not. The next logical step would be to let each word variable represent the number of times a word appears in a given item description. I take this one additional step by utilizing what are known as term frequency – inverse document frequency weights (tf-idf).

The tf-idf[7] is defined in the following manner:

$$tfidf = tf * \log\left(\frac{|D|}{1 + |\{d: t \in d\}|}\right)$$

Where $|D|$ is the cardinality of the document space and $|\{d: t \in d\}|$ is the number of documents where the term t appears. This weighting scheme therefore increases as a word appears more frequently in a given description, but decreases as a word appears more frequently across the collection of descriptions. Conceptually if a word appears many times in an item description, but appears infrequently across all item descriptions, that word is more likely to convey useful information about an item than a word that appears frequently in an item description and frequently across all item descriptions.

The Predictive Model

Below are the steps I took to build a predictive model for identifying which records were misclassified. Before starting that discussion, I will discuss my methodological approach. Predictive analysis uses many of the same tools as descriptive analysis, but emphasizes prediction accuracy. In the analysis that follows I estimate models that incorporate and select from several thousand variables. Additionally I use regularization techniques that I know will result in biased coefficients. For these reasons, individual coefficients are not examined or even reported. Instead the predictive model's performance is measured entirely on its ability to correctly classify test data.

[7] There are alternate definitions for the inverse document frequency weighting. This formulation is what is used in the Python machine learning package Scikit-Learn.

Logistic regression is often used in solving binary classification problems, but can be extended to solve multiclass classification problems either with a one-vs-rest technique (OVR) or through multinomial logistic regression. The OVR approach fits one binomial regression to each of the N class labels. For each observation, the predicted class label is the class label whose logistic regression had the highest estimated probability. Multinomial logistic regression on the other hand estimates the joint probability by estimating parameters that maximize the log-likelihood function. A disadvantage to using multinomial logistic regression is that it assumes the independence of irrelevant alternatives (IIA). This assumption posits that the presence of an irrelevant option does not affect the relative probability of other options. This is not a safe assumption to make here as I expect at least a portion of the misclassification to be the result of two item categories lying in close proximity to each other in the instrument. Thus I use the OVR method as it does not rely on an IIA assumption.

As described above, I take the "bag of words" approach to text analysis which allows each word reported in the cleaned text to be represented as an independent variable in the model. The predictive model therefore incorporates all of the text variables in addition to the expenditure and demographics variables identified in table 1. For m expenditure and demographics variables, and w word variables, the predictive model takes the form:

$$\hat{y} = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \sum_{t=1}^{w} \beta_t x_t$$

With over 4,000 unique words in the cleaned text there is cause for concerns of overfitting in the model. Overfitting occurs when a model is overly complex to the point of describing random noise in the training data such that it fails to generalize when the model is used on test data. Three steps have been taken which should mitigate the concern of overfitting. First, my training dataset is quite large. For a typical quarter, there are approximately 40,000 observations in the training data – with around 4,000

predictors[8] this leads to an average 9.35:1 ratio of observations to predictors. Various rules of thumb suggest that ratios between 10:1 and 15:1 mitigate the impact of overfitting. The next steps taken to reduce overfitting focus on reducing the number of predictors.

Regularization is a popular approach to reducing overfitting. Regularization adds an additional term to the optimization problem that penalizes overly complex models by shrinking the regression coefficients toward zero. Specifically I use an L1 penalty (LASSO) which is the sum of the absolute value of coefficients $\lambda \sum_{i=1}^{p} |\beta_i|$ where λ is a regularization parameter setting how harsh the penalty is (Hastie *et al.,* 2009). This λ is determined using 5-fold cross validation on the training dataset. Regularization is made more effective by my use of the OVR method since for each class label (with its own binary logistic regression) there will only be a subset of words that add predictive power to the model.

Results and Misclassification Estimates

Overall, the process described above is able to correctly predict approximately 90 percent of the reported item codes. However, just because the model predicted the item code that was reported does not guarantee that the item itself was correctly classified by the respondent and interviewer. A clear example of this would be if the respondent reports an item incorrectly and the interviewer uses the default description for that item code. In this case, no predictive model would be able to tell the difference between a correctly classified and a misclassified record using the item description. The remaining 10 percent of records for which the model predicted an item code that did not match what was reported are considered to be potentially misclassified. Assuming that the model has generated useful rules for classifying items based on their descriptions, the fact that the model disagrees with what was coded by the interviewer is evidence for misclassification. Table 2 below shows a few examples of

[8] Almost all of the predictors used in the logistic regression come from the different words that appear in the training data's item descriptions, however they also include those variables identified in Table 1. Thus my predictors include the text descriptions, Consumer Unit characteristics, and variables directly relating to the expenditure like purchase amount.

what the ICE process might output as potentially misclassified expenditures. Upon manual review around 50 percent of the potentially misclassified records are found to be misclassified in the four quarters analyzed.

Table 2: Examples of ICE Output

Reported UCC	Text Description	Predicted UCC	Changed
610230: photographic equipment	cell phone	320232: telephones and accessories	Yes
690119: computer software	tablet	690118: digital book readers	Yes
600210: general sport/exercise equipment	recumbent bike	600310: bicycles	No

Table 3 below summarizes some of the main results of performing the ICE process on the four collection quarters covering the 2015 collection year. The sample size N refers to the number of records in each production quarter that ICE was run on. The classification accuracy is simply the percentage of records whose predicted item category was equal to the item category reported. Hence 1 minus the classification accuracy returns the percentage of records in the production quarter that were considered potentially misclassified. Those candidates were then manually reviewed for misclassification and marked as either misclassified or not – yielding the change rate as the percentage of records that were manually reviewed and edited. From this I calculate an implied misclassification rate for the quarter as 1 minus the classification accuracy times the change rate. The misclassification rate is therefore the percentage of reported expenditures that were misclassified when reported.

$$misclassification\ rate = (1 - accuracy) * change\ rate$$

Performing this calculation on each of the quarters analyzed I find that the average misclassification rate is 5.85 percent. This represents a lower bound on the true rate of misclassification in the CEQ data. The analysis assumes that only the records identified as potentially misclassified were candidates for manual review. Using the example of misclassified records that were given the default descriptions, I demonstrated the potential error in relying on this assumption. While this estimate of the

misclassification rate could no doubt be improved, it represents the best effort to estimate this error source thus far.

Table 3: ICE Model Results

Quarter	N	Accuracy (%)	Change (%)	Misclassified (%)
2015Q1	4,460	89.00%	54.30%	5.99%
2015Q2	5,292	89.00%	49.20%	5.41%
2015Q3	5,652	88.00%	54.30%	6.52%
2015Q4	5,333	89.00%	49.20%	5.41%

The amount of misclassification differs across items, identified by Universal Classification Codes (UCCs). It would be impractical to provide tables showing how many records from each of the 38 UCCs mapped from EAPB were misclassified and which UCCs they were misclassified to. Nevertheless to give a sense of the findings, I present a few of the worst offenders. These misclassification prone UCCs are worth keeping in mind for future development work on the questionnaire and on the rules mapping expenditures to UCCs.

Table 4: Top Misclassified UCCs

Reported	Corrected	Count	Total Reports
600902: Other Sport Equipment	600210: General Sport/Exercise Equipment	56	182
690120: Computer Accessories	690111: Computers and Computer Hardware	48	1,152
320902: Hand Tools	320410: Lawn and Garden Equipment	47	532
690245: Other Household Appliances (Owned)	320522: Portable Heating/Cooling Equipment	36	352
310316: Radios/Speakers/Sound Systems	310333: Accessories and Other Sound Equipment	33	368
320420: Power Tools	320410: Lawn and Garden Equipment	28	649
310333: Accessories and Other Sound Equipment	310316: Radios/Speakers/Sound Systems	26	402
600210: General Sport/Exercise Equipment	600310: Bicycles	23	1,780
600310: Bicycles	600210: General Sport/Exercise Equipment	23	653
320902: Hand Tools	320420: Power Tools	21	532

Examining the top sources of UCC misclassification, none of them are particularly surprising from a survey perspective. These are examples of cases where two or more UCCs capture similar items. Notice

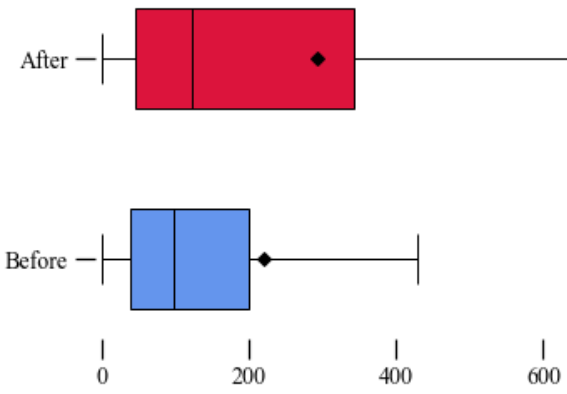
that the 5th and 7th ranked misclassifications are merely inverses of each other and that the 1st, 8th, and 9th ranked misclassifications are also closely related to each other. This lends credence to the idea that misclassification is occurring primarily where respondents and interviewers are likely to be confused by the difference between two or more item categories.

There are some key limitations that hinder the ability to estimate the impact of item misclassification on UCC expenditure estimates. First, all of the data analyzed had already gone through the entire production process and it was impossible to recreate this process for my analysis. The outlier detection and expenditure imputation processes in particular have a detrimental impact on my ability to estimate the classification error. By changing an expenditure's classification, I necessarily changed two cost distributions. Some expenditures may have been outliers in addition to being misclassified. Not being able to run the standard production outlier reviews means that I do not catch and correct these outliers. Alternatively, misclassified expenditures could have been incorrectly classified and edited as outliers in production when a correct classification would have placed the expenditure more in line with the rest of the distribution. Expenditure imputation results would also have been different had misclassified records been correctly classified beforehand. Second, because I did not perform this analysis on every section, I cannot say that there are no records in other sections that would be correctly classified in EAPB. This limitation is likely stronger for UCCs that have greater cross-section misclassification.

In spite of these two limitations, I attempt to estimate the impact of item misclassification on UCC expenditure estimates for three of the UCCs identified in Table 3. Figures 5, 6, and 7 below show that misclassification can have a relatively large impact on estimated means for specific UCCs. In dollar terms, UCC 600902 (Other Sport Equipment) shows a classification error effect on the mean of $73. It corresponds to a roughly 33 percent underestimate of the UCC mean. Similarly UCCs 690245 (Other Household Equipment) and 310333 (Accessories and Other Sound Equipment) both overestimated the

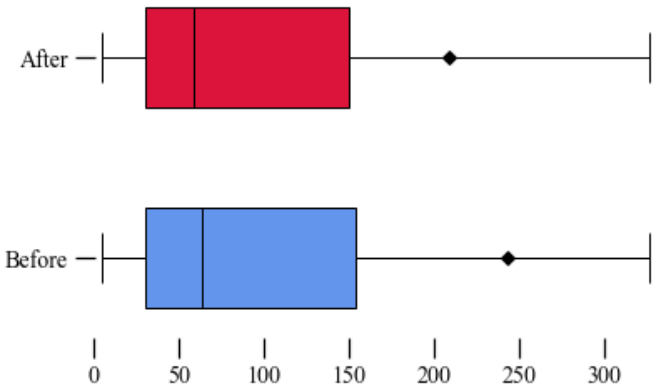
expenditure means by roughly 14 percent and 18 percent respectively. These three had some of the largest misclassification errors and are presented for that reason.

Figure 5: Misclassification Impact on UCC 600902 – Other Sport Equipment



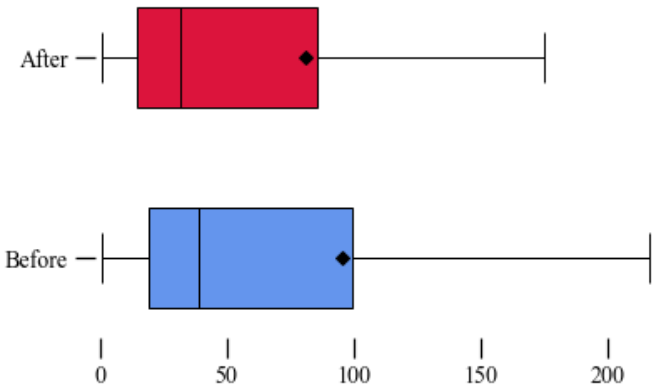
	Before ICE	After ICE
Count	182	92
Mean	$220	$293
Std.	$360	$435
Min	$1	$1
25%	$40	$47
50%	$99	$122
75%	$200	$343
Max	$2,136	$2,136

Figure 6: Misclassification Impact on UCC 690245 – Other Household Appliances (Owned)



	Before ICE	After ICE
Count	352	265
Mean	$243.13	$208.79
Std.	$1,246.28	$1,095.29
Min	$5.00	$5.00
25%	$30.75	$30.00
50%	$64.00	$59.00
75%	$154.00	$150.00
Max	$17,200.00	$17,200.00

Figure 7: Misclassification Impact on UCC 310333 – Accessories and Other Sound Equipment

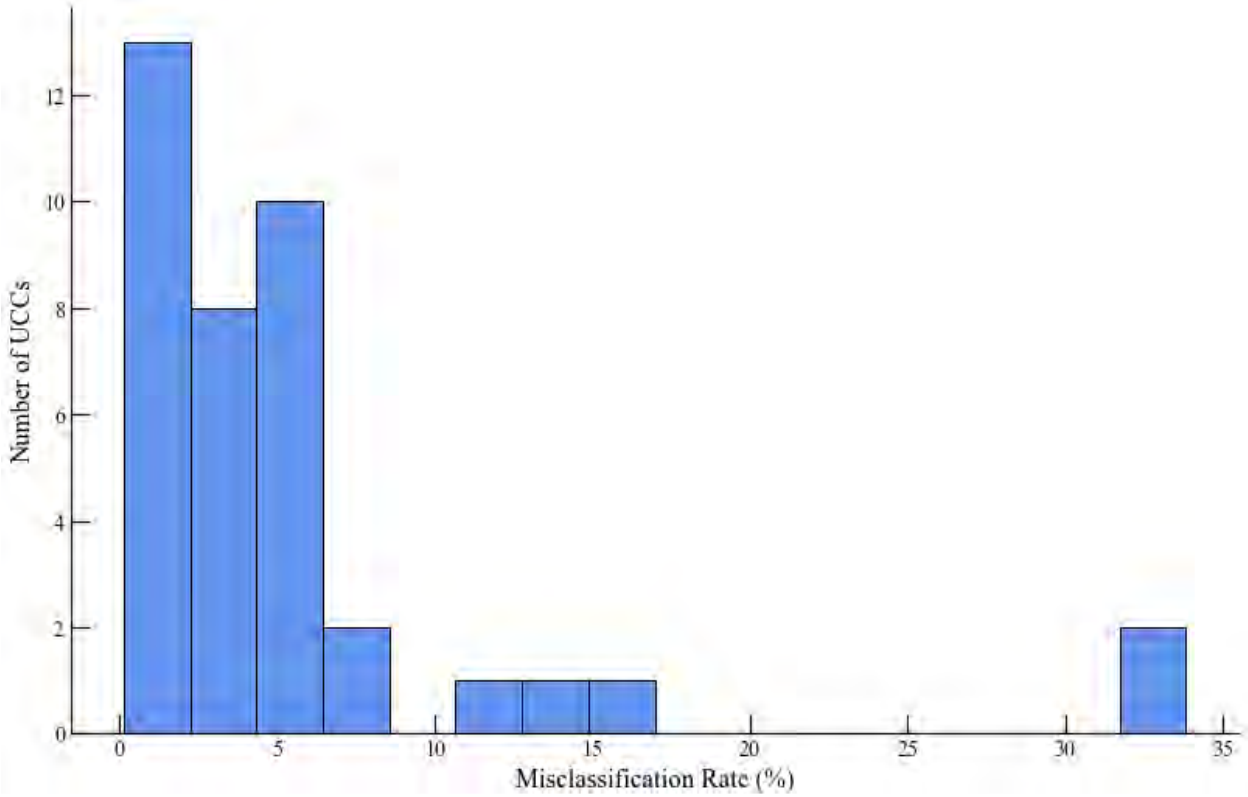


	Before ICE	After ICE
Count	402	406
Mean	$95.79	$80.88
Std.	$177.69	$154.56
Min	$1.00	$1.00
25%	$19.00	$15.00
50%	$39.00	$32.00
75%	$99.75	$85.75
Max	$1,500.00	$1,500.00

Some UCCs showed virtually no change in expenditure means as a result of ICE despite a high level of misclassification. These include UCCs like 690119 (Computer Software) of which around 4 percent of the records being mapped to this UCC were determined to be misclassified. Despite this relatively high incidence of misclassification, the mean moved only $3.72 from $102.62 to $98.90. This relatively small impact from classification error stems from the fact that many of the misclassified records in this UCC are video games that have been incorrectly reported with computer games. These two items often have similar prices (and are often the same games) so even a large number of misclassifications would not dramatically change the estimated mean.

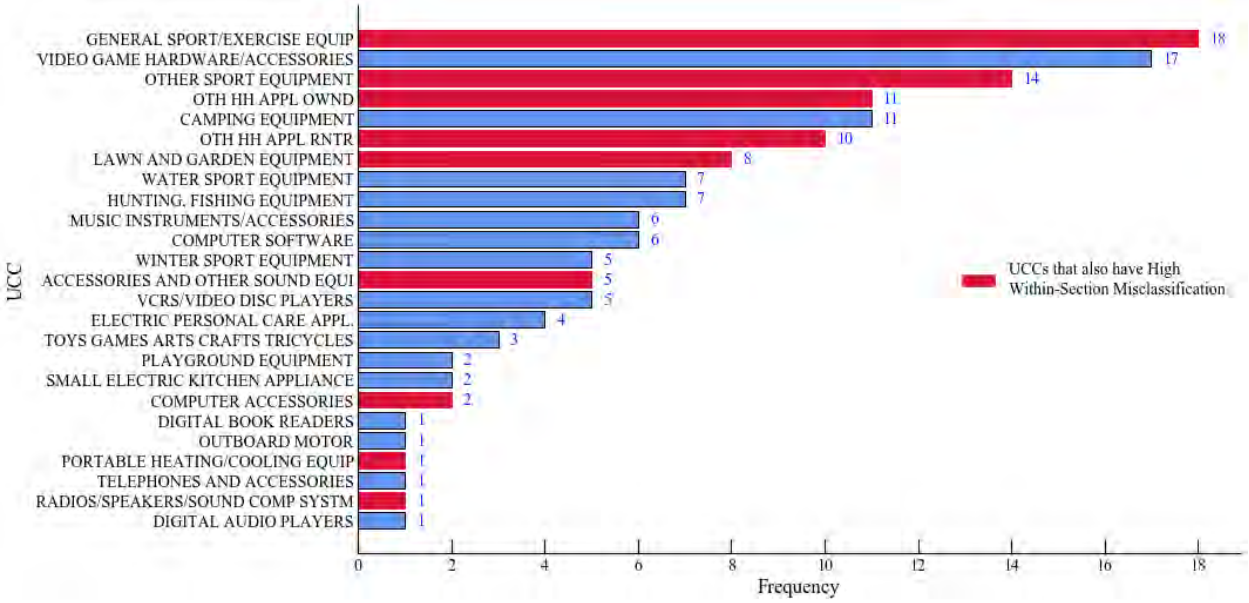
Overall, most UCCs had relatively small misclassification rates. Figure 8 below shows that about 2/3rd of the UCCs analyzed had misclassification rates less than 5 percent and that only 5 out of 38 had misclassification rates greater than 10 percent. The impact of classification error depends greatly on the aggregation of data analyzed. At higher levels of aggregation, UCCs that might have high rates of cross-misclassification are combined and so the classification error disappears.

Figure 8: Frequency Distribution of UCC Misclassification Rates



Complicating matters, the video games that were incorrectly reported in UCC 690119 (Computer Software) should not have even been collected in EAPB. Video games are supposed to be collected later in Section 17 of the CEQ interview: subscriptions and entertainment expenses. This means that beyond the already identified problem of items being misclassified within a section, ICE has helped uncover cases where items are misclassified across sections. In the four quarters analyzed, 149 records were found to belong in a different section. Together these 149 represented $30,705 in unweighted expenditures reported in the wrong categories. At this stage it is not possible to determine how greatly this misclassification impacted UCC estimates, but it is possible to point out a few UCCs that had higher levels of cross-section misclassification. Figure 9 below highlights those UCCs that suffered from most cross-section misclassification.

Figure 9: Count Frequencies of UCCs with Cross-Section Misclassification



Three UCCs in particular stand out in Figure 4: 600210 (General Sport/Exercise Equipment), 310232 (Video Game Hardware and Accessories), and 600902 (Other Sport Equipment). It is particularly troubling to see that how many of these UCCs in Figure 9 were also featured prominently in Table 3 as having high levels of within-section misclassification. These UCCs which also appeared in Table 3 are highlighted in red. The misclassifications appearing in 310232 tended to be of the same kind as those in 690119 (Computer Software): video games incorrectly reported with the video game hardware. Unfortunately the kinds of records being incorrectly reported in 600210 and 600902 do not seem to follow any kind of pattern. Misclassifications reported here included rafts, clothing, medical supplies, furniture, and major appliances.

Conclusions and Future Research

The findings represent a first effort at measuring the impact of classification error on Consumer Expenditure Surveys estimates. Despite the limitations outlined above, I was able to estimate that on average 5.85 percent of the reported items in EAPB for the four quarters of 2015Q1 – 2015Q4 are misclassified. In addition, I was able to show that at least 3 UCCs had fairly large classification errors on

the order of a 14 to 33 percent change in the estimated means. This analysis is also helpful in pointing to specific UCCs that may need to be clarified or redefined so that respondents are better able to correctly classify their expenditures.

Future work in this area should focus on expanding this analysis to additional sections and items. ICE has already been applied successfully to Section 17, subscriptions and entertainment expenses, and to Section 15, non-health insurance, data in production – it merely remains to extend this analysis to those sections and their corresponding UCCs. The ICE process itself could also be improved. The roughly 90 percent accuracy already achieved is certainly impressive, but improvements of even a few percentage points would reduce the number of records that need to be manually reviewed by 100 or more. There is also the potential to modify ICE so that it automatically reclassifies records. This would require significant research but has the potential to reduce classification error without adding another burdensome manual review process.

References

Bird, Steven, Edward Loper and Ewan Klein, "Natural Language Processing with Python". O'Reilly Media Inc. 2009.

Bollinger, Christopher R. and Martin H. David. "Modeling Discrete Choice with Response Error: Food Stamp Participation." Journal of the American Statistical Association Vol. 92, No. 439, Sept 1997: 827 – 835.

Feng, Shuaizhang and Yingyao Hu. "Misclassification Errors and the Underestimation of the US Unemployment Rate." American Economic Review Vol. 103, No. 2, April 2013: 1054 - 1070.

Gonzalez, Jeffrey, Catherine Hackett, Nhien To, and Lucilla Tan. "Definition of Data Quality for the Consumer Expenditure Survey: A Proposal." October 22, 2009. https://www.bls.gov/cex/research_papers/pdf/ovrvwdataqualityrpt.pdf.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Second Ed. Springer Series in Statistics. 2009.

Pedregosa *et al.* "Scikit-learn: Machine Learning in Python." JMLR Vol. 12, 2011: 2825 – 2830.