EMERGING DATA ISSUES IN APPLIED FOOD DEMAND ANALYSIS

Proceedings of a Workshop Held by the S216, Food Demand
and Consumption Behavior Regional Commitee

October, 1993

David Eastwood and Benjamin Senauer, Editors,
Department of Agricultural Economics, The University of Tennessee
and
Department of Agricultural and Applied Economics, The University of Minnesota

TABLE OF CONTENTS

AUTHOR AFFILIATIONS

Margaret Andrews is an Economist in the Food and Nutrition Service, U. S. Department of Agriculture.

Peter Basiotis is an Economist/Chief Diet Appraisal Research Branch in the Human Nutrition Information Service, U. S. Department of Agriculture.

Oral Capps, Jr. is a Professor in the Department of Agricultural Economics, Texas A and M University.

Wen Chern is a Professor in the Department of Agricultural Economics and Rural Sociology, Ohio State University.

David Eastwood is a Professor in the Department of Agricultural Economics and Rural Sociology, University of Tennessee.

Mary Hama is an Economist in the Food Consumption Research Branch, Human Nutrition Information Service, U. S. Department of Agriculture.

Helen Jensen ia an Associate Professor in the Department of Economics, Iowa State University.

Vickie McCracken is an Associate Professor in the Department of Agricultural Economics, Washington State University.

David Price is a Professor in the Department of Agricultural Economics, Washington State University.

Dorothy Price is a Professor in the Department of Human Development, Washington State University.

David Smallwood is an Economist in the Economic Research Service, U. S. Department of Agriculture.

EDITORS' NOTE

This Tennessee Experiment Station Bulletin is the edited collection of seven papers presented by members of the Changing Patterns of Food Consumption (S216 Regional Committee) at a 1993 Workshop held by the Regional Committee. They focus on a variety of emerging issues associated with data sets used in applied demand analysis. These pertain to topics that are not discussed in the extant literature but are quite germane to the extension of empirical models of food consumption.

POOLED TIME-SERIES AND CROSS-SECTION DATA FROM THE CONSUMER EXPENDITURE SURVEY

Wen S. Chern and Ben Senauer[1]

Introduction

The objectives of this paper are to describe the Consumer Expenditure Survey (CES) and to identify some uses of this continually expanding survey. The CES is one of the most comprehensive household survey data bases available in the United States. The CES is conducted by the Census Bureau for the Bureau of Labor Statistics (BLS) in the U.S. Department of Labor. The primary purpose of the CES is to provide a data base to support and maintain the Consumer Price Index (CPI). CES data were collected approximately every 10 years with surveys in 1888-1891, 1901, 1917-1919, 1934-1936, 1941-1942, 1950, 1960-61, and 1972-73. Beginning in 1980, the survey has been conducted annually to provide more current data. These annual surveys were originally referred to as the Continuing Consumer Expenditure Surveys (CCES) but are now typically just called the BLS Consumer Expenditure Survey.

The CES provides a rich source of data for research, especially for demand analysis. By pooling the yearly data since 1980, a longitudinal data set with both cross-sectional and time-series variation can be created. Expenditures for a wide variety of categories are gathered. However, neither data on prices nor quantities (so that expenditures can be divided by quantities to obtain implicit prices) are collected. The expenditure categories can be matched with CPI data for the same item by time period, and possibly geographic area, which allows prices to be introduced into the demand analysis.

Description of Data Bases: 1980-1990

The CES collects data from a national probability sample of consumer

---

[1]Wen S. Chern is a professor in the Department of Agricultural Economics and Rural Sociology, The Ohio State University and Ben Senauer is a professor in the Department of Agricultural and Applied Economics, University of Minnesota. The authors wish to thank Rafael Cortez for his helpful input into this paper, particularly to the section on Creating Prices for Individual Households, which draws heavily on his work. Zho Jun also provided valuable input, especially with the graphs.

units (typically households) that represents the total noninstitutional population and a portion of the institutional population of the United States. The CES actually consists of two separate surveys: an Interview Survey and a Diary Survey. The former gathers data on large purchases from a panel of approximately 5,000 households rotated on a quarterly basis. The latter collectes data from a household for five consecutive quarters and then it is dropped from the sample and replaced by a new consumer unit. The Diary Survey is collected from another independent sample also of approximately 5,000 households. This survey covers items purchased frequently, typically on a daily or weekly basis. The data are collected for two one-week periods, and the sample is spread over the entire year. At the beginning of the period, the interviewer visits the household, records household characteristics information, and leaves a daily expense record to record purchases. At the end of one week, the interviewer comes to the household again, picks up and reviews the completed diary, leaves another diary for the following week, and collects additional socioeconomic data for the household.

The Quarterly Interview Survey covers all major consumer expenditure categories: food, housing, apparel, transportation, health care, entertainment, and other expenses. Food purchases are broken down only into at home and away from home categories. The Diary Survey provides much more detailed information on food purchases. Food away from home expenditures, however, are only disaggregated into meal occasions: breakfast and brunch, lunch, dinner, and snacks and nonalcoholic beverages, plus various alcoholic beverage purchases away from home.

Table 1 shows detailed information on annual sample sizes of the CES Interview Survey during 1980-1990. The survey has been able to achieve a relatively high response rate of more than 85 percent. If all data from 1980 to 1990 are used, the total sample has 245,767 households, a huge data base, by any standard. The estimation task could be burdensome, especially, for a large demand system.

Table 2 presents sample sizes of the CES Diary Survey during 1980-1990.

Since the actual sample sizes from the public use tapes sometimes differ slightly from the sample sizes indicated in the BLS documentations, they are also presented for comparison. Again, the response rates are relatively high, ranging from 83.7 percent in 1986 to 91.1 percent in 1985. Over this 11 year period, the survey has accumulated a large total sample of respondents. When all data during this period are used, the sample contains 126,594 households. There are many households with missing or unreasonable data on income before taxes. During the entire period, there were about 15.26% of surveyed households with missing income data or negative income. Table 2 also shows annual sample sizes with households having a positive income. The large sample size of this continuously expanded CES survey data offers tremendous flexibility and different options for creating various data bases for undertaking different demand analyses. The next section explores some of these options and discuss some econometric issues related to the uses of the CES data.

Data Organization and Modeling Issues

Aggregations

With large numbers of surveyed households and commodity items, aggregation is usually necessary in order to use the CES data for a particular research objective in demand analysis. Aggregations take place across goods as well as households. With respect to the number of goods and services surveyed, the CES interview survey contained several hundred expenditure items. These expenditure data can be aggregated into different groups. In most of the BLS bulletins, there are 14 aggregate expenditure groups such as food, housing, and transportation are defined. Many empirical studies are based on this classification of goods and services. In the Dairy survey, there are approximately 95 categories for at-home food and alcohol expenditures (see Table 3). From this list, one can aggregate these items into broader food groups. Typical aggregations are into either six or nineteen food categories. The six are cereal and bakery goods, meats, dairy products, fruits and vegetables, others, and food away from home. The nineteen are cereal, bakery products, beef, pork, other meats, poultry, seafood, eggs, milk, other dairy products, fresh fruits, fresh vegetables, processed fruits, processed vegetables, sweets, beverages, fats and oils, other foods, and food away from home.

In aggregating good and service items, one important consideration is to maintain consistency with the classification used in the CPI data. For example, for the six or nineteen food categories, the expenditure data can be matched exactly to the CPI data. However, for other detailed food items, the corresponding CPI may not be available. As an illustration, the CES has separate categories of rice and pasta-corn meal-other cereal products (Table 3), but the CPI is available only for the combined group of rice, pasta and cornmeal.

Regarding aggregations over households, one can make good use of the rich demographic data available in the survey. These demographic variables such as region, race, ethnicity, or education can be used to separate the whole sample into several subsamples (see, for example, Fan). Demand estimation can be conducted for each subsample. Furthermore, average expenditure can be computed for various demographic groups, thus creating time-series data from these continuing annual surveys. This latter approach will be described in more detail later.

Matching CPI Data

Since the CES survey is now conducted continuously, it would be less justifiable to use only one year’s survey as a cross-section data basis without considering the effects of prices. Fortunately, the CPI data provide a readily accessible source of prices that can be matched to the expenditure categories of the CES. The current base period for the CPI is 1982-84 = 100. A price index reflects the variation in price relative to the base period. Most researchers would use the CPI-U which relates to all urban consumers. The CPI series are available monthly for a number of major cities, for urban areas of various sizes, and as a national city average for the United States. CPI data are also published for the four major regions: Northeast, North

Central, South, and West. The CPI can, therefore, be matched to pooled times-series and cross-section CES expenditure categories by time period and geographic location. An example would be matching the beef expenditure data with the beef CPI series by month within each year and by major region. If using the CPI indexes, dummy variables for the regions would need to be added to the demand equations to account for regional differences in the price levels in the base period, since 1982-84 equals 100 in every region.

Alternatively, one can use the quarterly data of the cost of living index (CLI) published by the American Chamber of Commerce Researchers Association (ACCRA). The ACCRA data base contains the CLI covering most of the Metropolitan Statistical Areas (MSAs) of the Primary Metropolitan Statistical Areas (PMSAs). However, the data base includes merely a composite index and six indexes for grocery items, housing, utilities, transportation, health care, and miscellaneous goods and services. Therefore, these price data can only be matched with expenditure data from the Interview Survey. One approach of creating prices for individual households will be detailed later.

Nonpurchasers

In the analysis of disaggregated (detailed) expenditure categories, a substantial number of households may not report any purchases during the survey period. The result is that the dependent variable in demand analysis is zero for a significant portion of the observations. More generally in econometrics it is referred to as a limited dependent variable problem. Nonpurchasers raise both conceptual and econometric issues. Initially, due to a lack of awareness of the problem, ordinary least squares (OLS) was simply used to estimate demand functions. Then there was a period in which Tobit analysis was applied without giving any thought to the underlying issues.

There are several reasons why a household may report no purchases, say for butter. First, people seek variety in their diets, and households maintain inventories of many food products, so purchases are periodic. If the survey period was extended to cover purchases over a month, or two months, many more households would report expenditures for butter. Second, there may be households which still reported no purchases, but which would purchase some butter if its price was lower or their incomes higher; they are potential purchasers. Third, there may be some households which would never purchase any butter regardless, perhaps because of concerns about its high fat and cholesterol content; they are true nonpurchasers.

Tobit is really only an appropriate technique in the second case, because it assumes the non-purchasers are not just a frequency issue, yet they would purchase the product at some level of the independent variables. In the first case, if all households purchase the product over some longer period and those observed purchasing it during the survey are random, OLS may be used to estimate demand equations with the observations for purchasing households. The issue can be viewed as a possible sample selection bias problem. Is the sample of purchasers random or not? If not, the Heckman procedure represents an alternative to tobit analysis. Since the probit stage of the Heckman approach needs not contain the same explanatory variables as the regression analysis, it is more flexible than tobit.

The underlying problem remains, though. When a household reports not purchasing any butter, the researcher does not know the reason. The next set of questions one would like answered for nonpurchasers are do they consume the product, when did they last purchase it, or do they ever purchase it, and if not, why not? Another possibility would be to lengthen the survey period, say from two consecutive weeks for the Diary CES to a month. The problem with these suggestions is the increase in respondent burden, which is already considerable. Alternatively, by aggregating expenditure categories or households, researchers can average out the nonpurchases.

POOLING ANNUAL SURVEY DATA

Creating Prices for Individual Households

This section outlines more specific uses of CPI data with the CES to derive prices, composite prices, and implicit quantities. Define the following variables

CPI_t^i: Consumer Price Index for good i at time t,

p_t^i: current price for good i at time t,

p_0^i: current price for good i at base time period 0,

RP_t^i: real price of good i at time period t, and

CPI_t: Consumer Price Index for all goods at period t.

Since $CPI_t^i = (p_t^i / p_0^i) \times 100$, solving for the current price for good i at time t, one obtains:

$$p_t^i = p_0^i \cdot \frac{(CPI_t^i)}{100}$$

The real price of good i at time period t is estimated by:

$$RP_t^i = \frac{p_t^i}{CPI_t} \cdot 100$$

For an aggregate good, S_t^i denotes the composite price for item i at time period t. Item i contains j=1,..,n goods. Then:

$$S_0^i = \sum_{j=1}^{n} p_0^j \cdot W_0^j$$

where:

S_0^i denotes the composite price for item i at base period 0,

W_0^j denotes the share proportion for good j at base period 0, and

p_0^j is the current price for good j at time period 0.

Therefore, the current composite price for item i at time period t is computed as follows:

$$S_t^i = S_0^i \cdot \frac{(CPI_t^i)}{100}$$

The implicit quantities of goods may also be derived. E_t^i is the expenditure on item i at time period t. S_t^i denotes the current composite price of item i at time period t. Therefore, the implicit quantity for item i at time period t is

$$Q_t^i = \frac{E_t^i}{S_t^i}$$

Case Studies

Falconi for his Ph.D. dissertation at the University of Minnesota (1991) pooled data from the 1980-87 Consumer Expenditure Surveys and matched the CPI data by month and region. He analyzed the effects of aggregation over consumers when estimating an Almost Ideal Demand System for a six food demand system. His analysis incorporated demographic variables into the demand model. In addition to the model estimated with household data, he aggregated households by month, region, adult equivalent household size, and income levels in various combinations to compare the demand estimates.

Cortez, in his dissertation at the University of Minnesota (1994), pooled the CES data for 1980-90 and matched them with the CPI by month. He used nonparametric techniques applied to the revealed preference axioms to test for structural change in consumer preferences for 19 food expenditure categories for specific socioeconomic groups. Previous nonparametric studies have primarily relied on time-series data and, therefore, have been able to test for taste change for the aggregate population only (presumably the representative consumer if aggregation assumptions are valid). Although the CES is not a true panel because the households sampled change from year to year, the annual surveys can be pooled to analyze population subgroup behavior over time. The household observations were aggregated into socioeconomic population subgroups by income level, age, and education. For example, one subgroup was households with incomes less than twice the poverty level, whose female head is less than 45 years old and has only a high school education or less. Substantial differences in preference trends between population groups were found for many of the food commodities.

CREATING TIME SERIES FROM SURVEY DATA

Despite the efforts to match CPI data with the CES expenditure data and the attainment of satisfactory results in several case studies cited earlier, the matching remains imperfect. The CPI data were not computed by households and thus they do not reflect price differences within a region, nor do they account for the differences in the composition of goods in each group among households. One way to provide exact matches between CES and CPI data is by creating a time-series from survey data. In the Diary Survey, there are about 800 households available in each month, so average expenditures can be computed from these many households month by month. By using the weights provided in the public use tapes, the weighted averages reflect the population more closely. Based on the entire sample from 1980 to 1990, the monthly time-series has 132 observations. Alternatively, the aggregation (averaging) can be done on a quarterly basis. The quarterly time-series would have 44 observations during the 1980-90 period. Unfortunately, the quarterly CPI data are not readily available, although they can be created from monthly data using simple averaging.

For the Interview Survey, it would be more complex to create monthly data because the household expenditure data for broadly defined expenditure groups are not identified by month; they are only identified by quarter. In order to create monthly averages, one has to use the Detailed Expenditure File (MTAB) to compute averages for original expenditure items (numbered in hundreds) and then follow the BLS procedure to aggregate them to broadly defined groups such as food and transportation.

The advantages and disadvantages of creating time-series data from the

CES are discussed in the Conclusions. Figures 1-5 show the trends of weekly expenditures by month obtained from the Diary Survey for beef, pork, poultry, fresh milk, and other dairy products. Figures 6-7 show the expenditure trends of food at home and away from home derived from the Diary and Interview Surveys, respectively.

Case Studies

Lee of The Ohio State University estimated a 19 food categories demand system, using the monthly time-series data created from the CES Diary Survey from 1980 to 1986. He estimated several versions of the LA/AIDS model for these 19 foods in order to compare static, dynamic, and autoregressive specifications. In addition, he also conducted nonparametric tests and clustering analysis to group these 19 foods into six groups. He then estimated a two-stage model and compared the alternative functional forms based on the LA/AIDS, translog, and Lewbel’s general system. One difficulty was encountered in estimation. The price of beef did not change much during 1980-1986. Consequently, the estimated beef price elasticities were not very stable. After the time-series were updated to 1990, considerable variations in beef price after 1986 were noted. Lee’s model is currently being updated at The Ohio State University.

Efforts are also underway to create monthly time-series expenditure data by income groups, and by race at The Ohio State University. These data series will be used to examine the differences in the demand for tobacco and alcoholic beverages across different income and racial groups.

CONCLUSIONS

This paper discussed the CES, its Interview and Diary Surveys, sample sizes, and contents, and its uses. These surveys have accumulated two large samples of household expenditures which can be organized in various ways for consumer demand analysis. The paper presents two general approaches of using these vast data bases--one by pooling household-level data over time and the other by creating time-series, particularly, monthly series from household data. There are advantages and disadvantages of these two approaches.

Pooling household data over time, one can use the household-level data directly, including rich demographic information about the households. For most expenditure categories, there are corresponding CPI data available nationally and by region. Therefore, household-level expenditures can be matched with national or regional CPIs for conducting demand analysis. Pooling can also be done for different demographic groups using such variables as race and income. The disadvantages of this pooling method are the following. First, the matching with CPI data is not perfect because for all households in a region, there is only one CPI. With limited price variations, the price-demand relationships may not be successfully estimated. Second, there are many nonpurchases of goods and services at the household level. How to deal with zero observations in a complete consumer demand model remains an unresolved econometric problem.

Creating time-series data from household-level data escapes the problem of nonpurchases. Econometrically, the limited dependent variable problem is avoided and the estimation issue is related to those of a time-series model. Another advantage is the perfect matching with CPI data. Since the time-series data are considered national observations, expenditure and CPI data are perfectly consistent. The estimated price-demand relationships should be more reliable with this set of data. On the other hand, this approach will necessarily sacrifice some of the richness of demographic information available in the CES surveys. This is because for continuous demographic variables such as household size and age, one can only use the average among households in a month or quarter. The categorical variables such as education, it is only possible to create a continuous variable such as percentage of household heads with college education. These demographic variables tend to have reduced explanatory power in a time-series model. Despite this loss of demographic information, we have learned from our experience that for several nonlinear demand systems such as the translog, AIDS, LES, QES, and the Lewbel’s model, one typically cannot incorporate more than three demographic variables. It would be impossible to use most, if not all, demographic variables available in the CES in estimating most nonlinear demand systems.

In conclusion, the CES provides rich household level expenditure data which can be used to meet a particular research objective in a demand analysis. Much more can be explored in the uses of this data source.

REFERENCES

American Chamber of Commerce Researchers Association, ACCRA Cost of Living Index, 1990.

Cortez, Rafael, "Taste Changes in the Demand for Food by Demographic Groups in the United States: A Nonparametric Empirical Analysis", unpublished Ph. D. Dissertation, University of Minnesota, 1994.

Falconi, Cesar, "Estimation of An Almost Ideal Demand System for U.S. Food with Household and Aggregate Data", unpublished Ph.D. Dissertation, University of Minnesota, 1991.

Fan, Xiaojing, "Ethnic Differences in Preference Structure and Budget Allocation Patterns", unpublished Ph.D. Dissertation, The Ohio State University, 1993.

Lee, Hwang-Jaw, "Nonparametric and parametric Analysis of Food Demand in the United States", unpublished Ph.D. Dissertation, The Ohio State University, 1990.

Table 1. Sample Sizes of CES Interview Survey[a]

Year	Designated Sample	Eligible Sample[b]	Interviewed Sample[c]	Response Rate (%)
1980-81	58,898	51,126	42,830	83.8
1982-83	64,219	53,859	45,971	85.3
1984	33,658	28,027	23,977	85.5
1985	37,842	31,211	26,625	85.3
1986	30,582	25,214	21,466	85.1
1987	31,156	27,377	23,536	86.0
1988	29,009	23,881	20,507	85.9
1989	28,826	23,628	20,338	85.7
1990	29,064	23,929	20,517	85.7
Total sample	343,254	288,252	245,767	85.3

[a]Note that these are not all independent observations because most households stayed in the survey for 5 quarters and each quarterly interview is considered as one observation in Interview Survey.

[b]Designated sample less Types B or C nonresponses, representing the housing units that are vacant, nonexistent, or ineligible for interview.

[c]Eligible sample less Type A nonresponses, including housing units which the interviewers were unable to contact or those refused to participate in the survey.

Table 2. Sample Sizes of CES Diary Survey[a]

	From BLS Documentation				From Tape Processing		
Year	Designated Sample	Eligible Sample[b]	Inter- viewed Sample[c]	Response Rate	Readable Sample	Sample with Data	Sample with Positive Income Data
1980	d	d	d	d	10,433	10,423	8,810
1981	d	d	d	d	10,547	10,547	8,695
1982	29,105[e]	23,987[e]	21,721[e]	90.5	10,927	10,925	9,224
1983					10,792	10,791	9,169
1984	16,721	13,637	12,144	89.1	11,873	11,873	9,925
1985	16,602	13,491	12,286	91.1	11,619	11,618	9,797
1986	18,650	15,312	12,817	83.7	12,817	12,815	10,957
1987	19,065	15,436	13,098	84.8	13,098	13,095	11,227
1988	16,599	13,327	11,413	85.6	11,413	11,413	9,727
1989	16,670	13,378	11,470	85.7	11,470	11,444	9,787
1990	16,934	13,705	11,735	85.6	11,651	11,650	9,952
Total Sample					126,640	126,594	107,270

[a]Note that these are not all independent observations because most households completed two 1-week periods. Each one-week period is considered as one observation in the Diary Survey.

[b]Designated sample less Types B or C nonresponses, representing the housing units that are vacant, nonexistent, or ineligible for interview.

[c]Eligible sample less Type A nonresponses, including housing units which the interviewers were unable to contact or those refused to participate in the survey.

[d]Not available.

[e]For 1982 and 1983 combined.

Table 3. CES - Diary Survey - Food Expenditure Categories

Flour

Prepared flour mixes
Cereal
Rice
Pasta cornmeal other cereal products
White bread
Bread other than white
Fresh biscuits, rolls, muffins
Cakes and cupcakes
Cookies
Crackers
Bread and cracker products
Doughnuts, sweetrolls, coffeecakes
Frozen & refrig. bakery prod.
Fresh pies, tarts, turnovers
Ground beef exclude canned
Chuck roast
Round roast
Other roast
Round steak
Sirloin steak
Other steak
Other beef (exclude canned)
Bacon
Pork chops
Ham (exclude canned)
Other pork
Pork sausage
Canned ham
Frankfurters
Bologna, liverwurst, salami
Other lunchmeat
Lamb and organ meats
Mutton, goat, game
Fresh whole chicken
Fresh or frozen chicken parts
Other poultry
Canned fish and seafood
Fresh and frozen shellfish
Fresh and frozen fish
Eggs
Fresh whole milk
Other fresh milk and cream
Butter
Cheese
Ice cream and related products
Other dairy products
Apples
Bananas
Oranges
Other fresh fruits
Potatoes

Lettuce
Tomatoes
Other fresh vegetables
Frozen orange juice
Frozen fruit, oth. fruit juice
Fresh/canned/bottled fruit juice
Canned and dried fruit
Frozen vegetables
Canned beans
Canned corn
Other processed vegetables
Candy and chewing gum
Sugar
Artificial sweeteners
Other sweets
Margarine
Oth fats/oils/salad dressings
Nondairy cream substitutes
Peanut butter
Cola drinks
Other carbonated drinks
Roasted coffee
Instant/freeze dried coffee
Noncarb. fruit flavored drinks
Tea
Other noncarb. beverages
Soup
Frozen meals
Froz/prep. food oth than meals
Potato chips and other snacks
Nuts
Salt/other seasonings & spices
Olives, pickles, relishes
Sauces and gravies
Other condiments
Prepared salads/desserts
Baby food
Misc. prepared foods
Lunch
Dinner
Snacks and nonalcoholic bev.
Breakfast and brunch
Board (includes at school)
Catered affairs
Beer and ale at home
Whiskey at home
Other alcoholic bev. at home
Wine at home
Beer and ale away from home
Wine away from home