No. 270

ANALYZING A DATA BASE FOR ECONOMIC MODELING

by

Rueben Buse and Aaron C. Johnson, Jr.

Department of Agricultural Economics
University of Wisconsin-Madison

ANALYZING A DATA BASE FOR ECONOMIC MODELING

Many problems are encountered in using cross-sectional household surveys for estimating consumption and consumption functions. This paper discusses a number of these problems. It stresses the need for the researcher to carefully inspect the data, using such simple procedures as cross-tabulations, plots, and other description statistics, prior to estimation. The paper also argues that often the researcher must (or should) modify, redefine, and reorganize theoretical models to maximize model compatibility with available data. Goldberger says it best:

> "...(ask) not what the sample can do for us but what we can do for the sample." (p. 82).

The advent of cross-section food consumption and expenditure surveys in the 1950s was, with the benefit of hindsight, a mixed blessing. Prior to that time, food demand analysis in the U.S. relied primarily on aggregate time-series data. Such data sets are restrictive, not only because of a small number of observations, but also, and more important, because of the limited amount of information on the socio-demographic variables that theory says, through the "taste and preference" variable, are important determinants of consumption behavior. Cross-section data sets, on the other hand, seem to offer much. Typically, the unit of observation is the household, a large number of observations are available that permit more efficient estimates of the parameters (Prais and Houthakker), and a wealth of information is usually provided on socio-demographic characteristics of the household.

The earliest household consumer surveys in the U.S. were designed and implemented for the single purpose of providing expenditure weights for computing and updating various price indices. The lack of computer hardware and software at the time resulted in little other use being made of these data sets. However, changes have occurred and researchers are now making extensive use of these surveys for research. Indeed, some researchers believe these data sets are fundamentally useful for policy analysis because they yield accurate estimates of long-run elasticities (George and King; Intrilligator).

However, this broadening of the data base for research has not come without a serious cost to the uncritical. Efficiently utilizing large data bases requires sophisticated data manipulation and organizational techniques, sometimes taxing the capacity of the software. More important for research quality, the researcher is a passive participant in data production--the data are imply "there." Because many hidden theoretical assumptions and empirical definitions underlie all published data, the unwary researcher uses the data at considerable peril.

And it is easy to be uncritical. With vast amounts of data readily available in machine-readable format, it is tempting to accept the data as is and proceed uncritically to analysis. This contrasts sharply to the golden years of time-series data sets when simple time plots and scatter diagrams quickly revealed potential data problems. Nowadays, with 5,000 observations on 40 variables, quick-and-dirty data checking is not possible and there lurks the danger that the time-honored procedure of checking data prior to analysis is no longer followed. The consequence may be nonsense research results.

The econometric literature recognizes two fundamental sources of error: "errors-in-data" and "errors-in-equations." Since much of this literature focuses on the latter, we need not consider it here. Rather, this paper discusses potential problems with large data sets, viewing them from the "errors-in-data" perspective. Examples are drawn from the authors' experience working with the 1972-73 BLS Consumer Expenditure Survey (CED/CES), the 1972-73

BLS Consumer Expenditure Diary Survey (BSL/CED), and the 1977-78 USDA Nationwide Food Consumption Survey (USDA/NFCS). These data sets are used only to demonstrate that problems are not unique to a particular survey or data set; it must be clearly stated that this paper does not criticize the BLS or USDA data. Rather, the data are used simply to emphasis the need for caution and for a liberal application of skepticism when using any large data base. It is ultimately the researcher's responsibility to become thoroughly familiar with the data before using it for estimating models of consumer behavior.

Three fundamental types of error are typically present in published data: conceptual errors, operational errors, and measurement errors.[1] A conceptual error occurs when the data being used do not measure the concepts of the theory guiding the research. For example, if the theory is based on the concept of an individual consumer, but the data used to estimate regression coefficients reflect household behavior, a conceptual error is encountered. Operational error refers to the rules used to transform the concepts of theory to empirically observable phenomenon. Finally, measurement error covers the range of errors from sampling error to improper coding of data. This framework is used here to illustrate some of the problems encountered in the household surveys mentioned above.

CONCEPTUAL/OPERATIONAL ERRORS

Since the interest of this paper is household survey data, it would be well to use "consumer unit" to illustrate the nature of conceptual and operational errors. Following are the definitions of "household" used by two major organizations publishing household data often used by researchers.

1972-73 BLS Consumer Survey: The household, the basic reporting unit for the survey, was defined as (BLS, (1977) p. 94):

(1) a group of two or more persons, usually living together, who pool their income and draw from a common fund for their major items of expense; or

(2) a person living alone or sharing a household with others, or living as a roomer in a private home, lodging house, or hotel, but who was financially independent--that is, income and expenditures were not pooled with other residents. Never-married children living with parents or away at school were always considered members of the consumer unit.

1980-81 BLS Consumer Expenditure Survey: The basic reporting unit for the survey was defined as (BLS, (1985), p. 131):

(1) all members of a particular household who are related by blood, marriage, adoption, or other legal arrangements such as a foster child; or

(2) a person living alone or sharing a household with others or living as a roomer in a private home or lodging house or in permanent living quarters in a hotel or motel, but who is financially independent; or

(3) two or more persons living together who pool their income to make joint expenditure decisions.

Financial independence was determined by the three major expense categories: housing, food, and other living expenses. To be considered financially independent, at least two of the three major expense categories had to be provided by the respondent.

A careful reading of these reveals that while the same concept, "consumer unit," is being measured, the operational rules used to identify (describe) a consumer unit differs between the two surveys. Moreover, the 1980-81 survey sampled only urban households; the 1977-73 survey sampled both urban and rural households. It follows that the two surveys are not directly comparable--they provide statistical pictures of a different 'reality'. This must be recognized, especially if, as is sometimes done, the two surveys are combined for research purposes.

The USDA/NFCS Survey, 1977-78: The Nationwide Food Consumption Survey collected data from 14,930 households of one or more members. The sample was statistically selected from the population of all private households in the 48 coterminous states, stratified by region, urbanization, and geographic or demographic similarities. About six percent, or 900 sample households, were excluded as non-housekeeping--no member ate 10 or more meals from the household food supply during the 7 days preceding the interview (USDA (1983) p. 302-305).

No explicit definition of a consumer unit is available for this survey. By inference, the definition appears to key on eating out of a common household food supply. How a household in this survey compares to a household in the two BLS surveys discussed above is not at all clear; indeed, they are in all likelihood not comparable. Consequently, the BLS and USDA data should not be merged for analysis, and research results obtained from using the two sets should not be compared without a detailed understanding of their similarities or differences.

The danger of doing so is suggested by the following.

Making indiscriminate comparisons across data sets, as is often done, is hazardous to healthy research. For example, the BLS/CED provides expenditure data on 10 food groups that have the same descriptive titles as the food groups used in the 1977-78 USDA/NFCS (see Table 6). A close examination of the two surveys reveals that their data are simply not comparable. They differ substantially for a number of reasons.

First, the USDA data represent the money value of food and beverages consumed rather than purchased, where consumption includes home produced food and food received as a gift or pay. Second, the USDA classified a consumer unit as ineligible and excluded it from the survey if at least one member did not consume at least 10 or more meals from the household food supplies during the 7 day reference period. The BLS, on the other hand, imposed no minimum on the number of meals eaten from household food supplies. As a result, there are more 1-person households in the BLS survey than in the USDA survey. Third, the USDA excludes sales tax on the value of food purchased for home consumption; the BLS data includes the tax. Fourth, the value of food produced at home was not included in the BLS data; it is included in the USDA data. Fifth, income was recorded by BLS/CED for the 12-month period prior to the date of the interview. The effect is a rolling definition of a year. In contrast, the USDA recorded income for the calendar year preceding the data collection period.

In addition to these differences in operational rules, the two surveys differ substantially in the manner by which the individual expenditures were aggregated into the food groups. To get a clearer picture of the aggregation procedures, individual expenditures for the two surveys were re-aggregated into comparable expenditure groups (Buse and Glaze). The results are presented in Table 1.

[Insert Table 1]

None of the major food groups is comparable: exactly the same subset of foods is used in the aggregation. One example is illustrated by dairy products. The BLS aggregation includes only 88 out of 125 items in the USDA expenditure category by the same name. Of the remaining 37 items, BLS used 14 in the fats and oils group, 6 in the beverage group, and 17 in miscellaneous. Similar patterns are present for the other food groups. The lesson of Table 1 is that even at quite high levels of aggregation, there are serious problems with direct comparisons of reported values. It is clear that consumption models estimated using the reported expenditure groups are likely to differ substantially between the two data sets.

More could, and perhaps should, be said about the importance of determining the concepts measured and the operational rules used to transform these concepts into empirically observable phenomenon before proceeding to model estimation. However, this discussion is sufficient to make the point that careful checking is required and that the researcher is responsible for doing it prior to analysis.

MEASUREMENT ERROR

Measurement error, the third type of data error of concern to the researcher, is broadly defined to include errors that arise from sampling, using a faulty questionnaire, improper coding, errors in data entry, and so on. These types of errors are referred to collectively by the sampling statistician as sampling and nonsampling errors. Additional errors can be introduced by the researcher applying transformations to the raw data, such as aggregating expenditures.

Data Inconsistencies

A household expenditure survey, by its very nature, involves handling hundreds and sometimes thousands of pieces of data. And those skilled in data handling are all too aware of the many errors that can creep in. Agencies responsible for surveys and for producing the data diligently try to locate and reconcile coding errors, transcription errors, an internal inconsistencies.

They do a commendable job. But errors of consequence to the researcher may escape the error checking routines used by the data producer. This can arise in the following way. Data are collected for a specific objective, or use (weights for computing the CPI), and the data checking routines are based on this objective. To use a data set for research might require quite a different type of error check. Hence, the data may be error free for one use, yet error filled for another. Improved communications between data producers and researchers, along with accumulated research experience, can mitigate but never eliminate this type of problem.

A large number of tests looking for potential data problems for consumption analysis were performed on the three data sets discussed above. Some errors were obvious mistakes in translating the data into machine readable format. Some involved data interpretation or coding errors. In many cases, the problems were easily resolved by applying simple logic. In others, resolution was impossible without reference to the original survey instrument. Because government regulations do not permit access to the original schedules, diaries, or copies of the basic data, these problems cannot be resolved by the researchers. Consequently, the researcher must decide whether to delete certain observations from the data set prior to analysis. Tables 2, 3, and 4 summarize some of our findings.

Selected data problems in the BLS expenditure diary surveys are shown in Table 2. The overall error rate is small--639 diaries out of 23,186, or less than three percent, contained detectable errors. Although the error rate is small, the errors present can prove troublesome. They may be outliers in a statistical analysis, such as zero dates, zero expenditures for food stamps, miscoded start dates, and negative earnings. Alphabetical data in numerical fields often result in expensive computer run failures, a problem most researchers prefer to avoid.

[Insert Table 2]

A similar summary of data problems with the 1972-73 BLS Consumer Expenditure Survey (BLS/CES) is presented in Table 3. Again, the error rate is not high, but there remain potential problems for the uncritical researcher. Two categories are shown in the table: data errors, which usually can be corrected directly by the researcher, and data warnings, which cannot be resolved by the researcher due to lack of sufficient information. In some cases, an inconsistency created both a correctable error and a warning because there was insufficient information available to correct all the affected fields. Data error #2 is an example. In 30 cases, one or more of the household members (FM's) over one year of age were listed as being in the household (CU) zero weeks in the past year. The code for "weeks in the CU" was changed from 0 to 99 (unknown) to eliminate the zero, but the calculated average family size could not be changed. A data warning for the CU was put into the data base to flag that observation as a potential problem.

[Insert Table 3]

In 127 cases, the household head (FM-1) was listed as married, but there was no adult female listed among the other household members. In most cases, the marital status of FM-1 was changed from 1 (married) to 2 (not married). This decision was arbitrary, but at least the observation was made internally consistent with the other data on the record. The alternative is to drop the observation from the analysis. But this, too, is hazardous. Arbitrarily deleting observations with minor problems, such as a miscoded marital status, can quickly result in losing the representativeness of the sample data.

Data error #7 is more critical. It indicates that total expenditures did not include auto registration fees. A researcher using this data to estimate a model of complete consumer demand would have a hidden adding-up problem or biased coefficients if he were estimating a subset of expenditures.

Problems discovered in the 1977-78 USDA/NFCS survey data are summarized in Table 4. The model to which the data were to be fitted included food-price indices based upon the interview date. A histogram showed that some dates preceded and other dates followed the date of the survey--some were as much as 6 months out of line. Another example: the data base had two different measures of household income: one for the previous pay period and one for the previous calendar year. In more than 500 cases, these income measures were either widely inconsistent or had impossible values, such as $50,000 of welfare income or $35,000 of unemployment benefits. In many cases a detailed examination revealed that annual income had been recorded instead of income for the previous pay period. As a consequence, the income variable was either 12 or 26 times overstated.

[Insert Table 4]

Very large weekly expenditures on particular food items are particularly troublesome. Some examples: 13.5 pounds of tea, 21 pounds of coffee, 18 pounds of lard, 30 pounds of baby food, 103 pounds of non-dairy creamers and toppings, 17 pounds of icings and artificial sweeteners, 300 pounds of fresh fruit purchased in the winter, 500 pounds of pudding, spending more than 100 percent of weekly income on fresh whole milk.

Finally, economic models implicitly assume positive household income, yet more than 1,600 households in the USDA survey reported negative annual income. Probably the income data came from the respondents' tax returns, but unfortunately there is no way of determining that this is the case. Thus the choice: Are these atypical observations to delete from the data set or bona fide cases that should be included in the analysis? The point is that such things can be found in large data sets. The researcher must first find them and then decide whether to delete them.

Data Density

The analyst must answer a fundamental question: What level of expenditure detail (aggregation) provides the best results for the given research objective? Analyzing very detailed expenditure data is an alternative, but data handling and computing costs are high. This suggests recasting the question: What is the maximum level of aggregation that will not obscure real differences in the demand parameters? For one study, aggregating to total meat expenditure may be unsatisfactory because it hides the substitution effect induced by differences in price ratios among pork, beef, and poultry. For another study, the distinction between chuck roast and round steak may not be worth the added cost of estimating that distinction. Aggregate national estimates of the demand for fluid milk may be satisfactory for national dairy policy analysis but unsatisfactory for a dairy marketing board managing a regional or subregional dairy promotional campaign.

Regardless of the level of aggregation, the researcher must face the issue of data density. Typically, the lower the level of aggregation, the larger the proportion of households that will have zero values for the dependent variable. This is not an inconsequential consideration because many estimation procedures require positive values for the dependent variable. If the level of aggregation results in a large number of zero values for the dependent variable, the dependent variable is said to be truncated. An immediate and non-trival consequence is that the OLS estimates will be biased. Recent theoretical developments (Tobin; Heckman; and others) have provided methods for handling the truncation problem in single equation models. These estimation procedures are more complex than OLS procedures. Moreover, there is some evidence that the level of truncation has implications for choosing the most appropriate estimation method (Cox and Ziemer, 1985, p. 11). Thus, data density is a prime consideration in any decision regarding the level of aggregation and the estimator to use.

The higher the level of aggregation, the more likely the household is to have consumed or purchased that item. Few household buy buttermilk, but most use some dairy products. Thus, at some minimum level of aggregation, the data will show a household reporting positive expenditure or consumption quantities. However, below this level of aggregation, some households will show zero quantities, and the lower the level of aggregation, the greater the proportion of zero values for the dependent variable. Zero values of 5, 30, or 60 percent or more of the values of the dependent variable seriously affect the estimation method and the policy implications of the results.

To demonstrate the seriousness of the problem in most cross-sectional data sets, the data density of various expenditure aggregates in the three data sets is presented in the following tables. For this summary, data density is defined as the proportion of observations containing non-zero values for a particular expenditure item.

[Insert Table 5]

The BLS/CES data were aggregated into 13 major expenditure categories each of which covers a broad range of expenditures. Four categories--Alcohol, Tobacco, Education, Other--have a density of less than 66 percent (Table 5). For the other categories, 85 percent or more of the households reported expenditures. This relatively high density resulted from collecting the data over 12 months by four quarterly interviews. This length of time increases the chances that a household will purchase some item within each category.

The 1972-73 BLS/CED survey, conducted over a two-week period, reports expenditures for nine major categories. Most of the emphasis and detail in that survey was on food expenditures and frequently purchased non-food items, such as household supplies, toothpaste, alcohol, utilities, and the like. Five of the nine major expenditure groups have a density of less than 66 percent (Table 6). Within food expenditures, "sugar and sweets" and "fats and oils" show densities of less than 66 percent. In Food-away-from-Home, all three of the subgroups show low densities. Note that, for comparable expenditure categories, the density in the BLS/CED is less than in the BLS/CES, probably reflecting the shorter reporting period.

[Insert Table 6]

For comparison, the density of food expenditures for the USDA/HFCS are shown in the last column of Table 6. For the first five sub-items under food-at-home, the density for the BLS/CED data is much lower (about 10 percentage points on average) than for the USDA/HFCS data. The remaining four categories differ by about two percentage points. This comparison may not be particularly interesting because, as shown above in Table 1, the categories differ substantially by the individual items making up the categories. Nevertheless it emphasizes the point that the data may differ substantially from one survey to the next, making direct comparisons questionable.

The density falls rapidly by disaggregating. For example, 99.7 of the CUs in the BLS/CES data reported housing expenditure (Table 5). Disaggregating this yields:

Sub-Item	% Non-Zero
shelter	55
operations	98
utilities	less than 90
telephone	less than 90

Disaggregating food expenditures yields a similar pattern. About 90 percent, of the completed BLS diaries reported meat, fish, poultry, or egg expenditures (Table 6). The density falls to between 66 percent and 72 percent by disaggregating the meat category into "beef and veal" and "pork-except canned" (Table 7). Further disaggregation of "beef and veal" reduces the percent of households reporting non-zero expenditures to 46.9 for ground beef, 16.2 for chuck roast, 10.7 for round steak, 16.9 for other beef, and 3.5 for veal. Similar changes in data density are shown for dairy products.

[Insert Table 7]

If the research objective is estimating a complete demand system, data density can quickly limit the available number of observations. Even when the density of individual expenditures is quite high, the number of households reporting non-zero values across a pre-specified list of expenditures falls off rapidly as the number of expenditure equations in the model increases. In Table 8, respondents in the three surveys are grouped according to the number of non-zero expenditures reported. The table shows that the researcher quickly faces a density problem, particularly in the BLS/CED data. Eliminating all households reporting two or less food expenditure items reduces the sample size by 2,170 or almost 10 percent. On the other hand, estimating a complete food demand system that includes at least five different food items would exclude at least 19 percent of the BLS/CED households, but less than one percent of the USDA observations. The USDA/NFCS was limited to foods. This is an example of how a specialized data base may, in some sense, be more useful to the researcher than a more generalized data base.

[Insert Table 8]

A higher level of aggregation in the survey data reduces the severity of the density problem but does not eliminate it. Only four percent of the 1972-73 BLS/CES households report expenditures on six or fewer of 13 major categories. The problem is more complex than indicated by a simple tabulation. The more usual complete demand model requires simultaneous expenditures for a specific set of goods, a much more restrictive requirement on the data. For example, if the researcher were to use the BLS/CED data to estimate a model that included five expenditure categories (food, housing, clothing, medical care, transportation and recreation), 14.5 percent of the sample observations would have to be deleted (Table 9). There is little assurance in the literature that any model resulting from such subsets represent anything more than the subset of sample observations.

[Insert Table 9]

In summary, large data sets, such as those discussed here, with many observations on a range of socio-economic variables are exciting to the applied researcher. And the rush to use them in models is understandable. However, the summary results discussed above should temper the enthusiasm--all may not be well on the data tape. At a minimum, the researcher must first prepare tabulations, such as those above, to determine the limits imposed by the data before "SASing" them.

Outliers

Another problem often encountered with household survey data is the presence of unusually large or small values of the dependent variable. The problems with and the implications of outliers in estimating regression equations are well known and need not be discussed here. It is sufficient to say that an outlier can seriously affect statistical results, particularly OLS estimates. Thus, an awareness of their prevalence in the data and an evaluation of their likely effect of required. Often outliers in the data indicate special circumstances warranting further investigation as to the appropriateness of the model, its functional form, or the need for additional explanatory variables.

This section summarizes some of our findings regarding outliers in the three data sets under discussion. For this summary, an outlier is defined as a value that is greater than five standard deviations from the mean.[2]

The mean, standard deviation, and largest value for each of the 13 major expenditure categories in the BLS/CED are tabulated in Table 10. Average annual total food expenditures per CU (household) was $1,774, with a standard deviation of $1,262. Forty-eight CU's reported expenditures of $8,084 or larger (over $150/week), more than five standard deviations above the mean. The largest was $46,433 (28 standard deviations above the mean) by a household with a reported annual income of $13,295. Other expenditure categories exhibit similar, very large expenditures. Since non-disclosure rules prohibit examining the original documents, the researcher faces a dilemma. Including unexplainable extreme observations may substantially distort the parameter estimates, increase estimated variances, or both. Excluding them introduces several possible biases into the results. Whichever decision is taken, the researcher must think carefully about the impact of the decision on the analysis.

[Insert Table 10]

Tables 11 and 12 illustrate the same phenomenon for the 1972-74 BLS/CED and the 1977-78 USDA/HFCS. It is clear that extreme values are easily encountered for almost any expenditure at any level of aggregation. The largest reported values in a category can range from 25 to 100 standard deviations above the mean expenditure of all CU's reporting non-zero values.

Generally, a household does not report an extreme value for all expenditure categories; rather, the value is usually found in only one subcategory. For example, in the BLS/CED data, one household reported a weekly food expenditure of $4,099, of which $4,000 was for meals away from home. Another household reported spending $1,069 for total food, at which $1,030 was for beef and veal purchases. Finally, one household spent $216 for food at home, of which $175 was for fresh whole milk. Similar examples exist in the USDA/NFCS (Table 12).

[Insert Tables 11, 12]

Often outliers are deleted on the argument that they represent miscoding; transcription errors; or unusual households with large families, large incomes, medical problems requiring special needs, and so on. However, examination of each extreme value in the context of the socio-economic characteristics of the observation does not support this proposition. The only exception seems to be in the Diary data, where households reporting 1-week expenditures (in contrast to the usual 2-weeks) have a slightly higher probability of reporting extreme values. Otherwise, there is little evidence to indicate that those values are aberrations that can simply be excluded from an analysis. The question remains, should those large values be "adjusted" because they appear "inconsistent" with other data? There is no unequivocal answer. If extreme values are modified, the question of when to stop arises. If a value that is 30 standard deviations above the mean, and clearly seems to be miscoded, is adjusted, why not adjust those that are 25, 10, 5, ..., standard deviations from the mean?[3]

Adjusting for income level by using expenditure proportions as the dependent variable might reduce the number of households reporting extreme values. The theoretical literature describes a number of budget share models and, since budget shares are dimensionless, some argue that budget shares are more suitable for comparison. However, Tables 13 and 14 show that using budget shares as the dependent variable can aggravate the problem. Food, housing, house furnishings and equipment, clothing, and transportation expenditures have fewer extreme values (Table 13). Yet, for the remaining eight expenditure categories, the number of CU's with expenditure proportions more than five standard deviations above the mean increases. The pattern is similar in the BLS/CED. Using budget shares increases the number of CU's more than five standard deviations above the mean (Table 14). Only meat, fish and poultry, and miscellaneous expenditures have fewer large values.

[Insert Tables 13 & 14]

Tables 13 and 14 also illustrate another problem. In almost all expenditure categories there is at least one household reporting spending all (100 percent) of its weekly or annual expenditures on that item. This is particularly surprising for the survey data since it reports annual expenditures. Furthermore, in the survey there is at least one CU in each expenditure category reporting spending more than 50 percent of its total annual expenditures on one category. There are also 11 households spending more than 83 percent of total annual expenditures on food, and one household reporting 100 percent of its annual expenditures on food. Similar dubious, reported expenditures can be found in the other data sets.

Non-Response

Using cross-sectional data is often made difficult by the absence of values for one or more of the socio-demographic variables. There is no best approach for dealing with non-response, but there are several alternative solutions. One is to drop those observations with missing data. As illustrated earlier, this can quickly decrease the degrees of freedom. Furthermore, the literature shows there is a decrease in the efficiency of the estimates from the reduced sample. One alternative for quantitative variables is to replace the missing value with the sample mean of the complete observations. If there is a pattern in the missing observations, a second method capitalizes on that pattern to fill in the variable having missing values. This approach finds a set of variables (instruments) that are highly correlated with the missing values. The variable with missing values is regressed on the set of instrumental variables and the calculated values are used to fill in the missing values. The choice of the proper instruments can affect the results. If the instruments are appropriate, the resulting estimates, using the observations with computed values, are consistent estimates. With many missing observations, this procedure must be used with caution since it can produce heteroskedastic error variances. If more than one right hand side variable has missing values, the problem becomes more complex. Finally, the degrees of freedom in the model must be adjusted by the number of observations that were filled in. A summary of these techniques can be found in Donner.

If values are missing for a qualitative variable such as race, occupation, or marital status, the problem can be handled by creating a separate dummy for that group and including it as a separate variable in the analysis. This is more complex than it first appears because the researcher must decide a priori if the variable is to be a slope shifter, an intercept shifter, or both.

SUMMARY

There are many potential problems in the cross-sectional consumer expenditure data bases. The researcher must give at least as much attention, thought, and diligence to the data base as to the model being estimated. One wonders how many times hypotheses have not been rejected or rejected because of the influence of an outlier, or because the researcher either ignored or was ignorant of other problems with the data. Our empirical models definitely lag behind our data since they have not been adjusted to reflect the current state of the art in data collection. They are highly oversimplified explanations of how the household operates. In contrast, the data in the various consumer expenditure surveys reflect the full spectrum of consumer behavior. Fitting such models to survey data is akin to putting round pegs in square holes. They can be viewed as fitting--if the square hole is large enough. But the fit is not very satisfactory. The researcher must carefully match the statistical model to the data, not vice versa. Data inspection, reorganization, and cleaning must be major areas of concern. After working with cross-sectional data sets for over more than 10 years, we are very sensitive to the need for carefully inspecting the data through cross-tabulations, plots, liberal use of the descriptive statistics that computer software can easily produce, and checking results against other published descriptive statistics. The process requires several steps: (1) carefully checking the internal consistency of the data items within each observation; (2) becoming familiar with the outer limits of the data and what it implies for the model and resulting estimates; and (3) modifying, redefining, and reorganizing the data and the model to maximize compatibility.

Understanding consumer demand will be successful if it builds upon past and current work in a coherent way. This means retesting old models with new data and comparing results across data sets, models, and methods. Theoretical models are simply hypothesized relationships that require testing and retesting before their validity can be provisionally accepted. They can only be tested if the assumptions of the models and those implicit in the data used to test them coincide. We need good descriptive analysis of household expenditures to use as input into our theories.

This is not to say that theory has little to contribute. On the contrary, theory must be used. But theoretical models must be recast as new research results are obtained. This demands diligence, keen powers of observation, and the patience to test and retest the results on new data sets. The time is ripe for systematically developing models useful for explaining household-to-household variation in demand or expenditures. We have the technology and detailed data bases. And more are coming on stream. It is this that gives rise to optimism. Complex microeconomic models that first explain and then simulate household demand behavior with some degree of detail are within the realm of feasibility, if researchers avoid the data traps and use efficiently the available data. The wealth of detail these data sets contain on expenditures and on the characteristics of the consuming unit are fertile fields for quantum jumps in our knowledge--provided we are willing to carefully and industriously forge ahead.

> "...The data are imperfect not by design, but because that is all there is. Empirical economists have over generations adopted the attitude that having bad data is better than having no data at all, that their task is to learn as much as is possible about how the world works from the unquestionably lousy data at hand. While it is useful to alert users to their various imperfections and pitfalls, the available economic statistics are our main window on economic behavior. In spite of the scratches and the persistent fogging, we cannot stop peering through it and trying to understand what is happening to us and to our environment, nor should we. The problematic quality of economic data presents a continuing challenge to econometricians. It should not cause us to despair, but we should not forget it either." (Griliches, p. 199).

NOTES

[1] See Jacobs, p. 15ff or Judge, et.al., p. 509-516.

[2] Tchebysheffs theorem says that at least $1-1/(K^2)$ of the total observations should lie within k standard deviations of their mean. Thus, in the BLS diary no more than 1/9, or approximately 2,500 observations, should be more than three standard deviations. Similarly, in the Survey no more than 2,200 should lie beyond three standard deviations from the mean and no more than 800 observations five or more standard deviations.

[3] Recent unpublished and preliminary work with Engel functions for dairy and for meats by the authors indicates that eliminating all observations with expenditure values more that five standard deviations above the mean has little effect on the OLS parameter estimate, but it does not reduce their variation.

References

Buse, Rueben C. "Data Problems in the BLS/CES PU-2 Diary Tape: The Wisconsin 1972-73 CES Diary Tape," Ag. Econ. Report #164, Department of Agricultural Economics, University of Wisconsin-Madison, July 1979.

Buse, Rueben C., and John A. Glaze. "Differences in Household Food Expenditures; 1972-74 to 1977-78," Research Report to SEA/USDA, Washington, D.C., 1984.

Cox, Thomas L., R.F. Ziemer, and Jean-Paul Chavas. "Household Demand for Fresh Potatoes: A Disaggregated Cross-Sectioned Analysis," Western Journal of Ag. Econ, (1984) 9:1 pp. 41-57.

Cox, Thomas L., and Rod F. Ziemer. "An Empirical Comparison of Alternative Tobit Estimate," Staff Paper #234, Department of Agricultural Economic, University of Wisconsin-Madison, March, 1985.

Daberkow, Stan. "Regression Analysis with Complex Survey Data: A Comparison of Estimation Techniques." American Agricultural Economics Association Meeting. Cornell University, Ithaca, New York, August 1985.

Donner, Allan. "The Relative Effectiveness of Procedures Commonly Used in Multiple Regression Analysis for Dealing with Missing Values." The American Statistician, Vol. 78, No. 383 (1983), p. 535-543.

George, P.S., and G.A. King. "Consumer Demand for Food Commodities in the U.S. with Projections for 1980," University of California, Giannini Foundation Monograph, No. 36, March 1971.

Goldberger, Arthur S. Topics in Regression Analysis, London: MacMillan Co., 1968.

Griliches, Zvi. "Data and Econometricions -- The Uneasy Alliance." American Economic Review. Vol. 75 (1985), No. 2, pp. 196-200.

Heckman, James J. "Sample Selection Bias as a Specification Error," Econometrica, 47(1979):153-161.

Intriligator, Michael D. Econometric Models, Techniques and Applications, Englewood Cliffs, NJ, Prentice Hall, 1978.

Jacobs, Herbert. "Using Published Data, Errors and Remedies." Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-042. Beverly Hills, SAGE Publs. (1984).

Judge, George G., Griffiths, William E., Hill, R. Carter, and Lee, Tsoung-Chao. The Theory and Practice of Econometrics. New York, Wiley (1980).

Prais, F.J., and H.S. Houthakker. The Analysis of Family Budgets, Cambridge, Massachusetts, The Cambridge University Press, 1955.

Tobin, James. "Estimation of Relationships for Limited Dependent Variables," Econometrica, 26(1958):74-36.

U.S. Bureau of Labor Statistics. "Consumer Expenditure Survey Series: Interview Survey, 1972-73." USDL/BLS Report No. 455-4, Washington, D.C. 1977.

U.S. Bureau of Labor Statistics. "Consumer Expenditure Survey: Interview Survey, 1980-81." USDL/BLS Bulletin 2225, Washington, D.C. 1985.

U.S.D.A. "Food Consumption: Households in the United States, Seasons and Year 1977-78." NFCS 1977-78 Report No. H-6, Washington, D.C. 1983.

TABLE 1: Comparison of Expenditures Included in Major Food Groups 1972-74 BLS/CED and 1977-78 USDA/NFCS

		BLS/CED Food Groups									
USDA/HFCS		Cereal	Meat, Fish				Sugar	Fats			
	No. of	and	Poultry &	Dairy			and	and			Alcohol
Food Groups	Items	Bakery	Eggs	Products	Fruit	Vegetables	Sweets	Oils	Beverages	Miscellaneous	At-Home
Cereal & Bakery	693	660				2				31	
Meat, Fish, Poultry & Eggs	1,230		1,170							60	
Dairy Products	125			88				14	6	17	
Fresh Fruits	146				146						
Fresh Vegetables	312				1	299				12	
Processed Fruits & Vegetables	454				192	231				31	
Sugar & Sweets	126	1					94			31	
Fats & Oils	67			4				63			
Beverages	84				17		3		49	3	12
Miscellaneous	594									594	
TOTAL	3,831	661	1,170	92	356	532	97	77	55	779	12

TABLE 2: Number of Households Exhibiting Selected Data Problems in the

1972-73 Bureau of Labor Statistics Consumer Expenditure Diaries

Data Problem	Diary Year	
	1972-73	1973-74
1. Inconsistent Interview Dates	20	51
2. Missing Interview Dates	38	29
3. Family Member Detail Inconsistent with Reported Aggregates	77	108
4. Incomplete Food Stamp Data	17	232
5. Weeks Worked Missing	28	--
6. Alphanumeric Data in a Numeric Field	11	25
7. Income Detail Does Not Sum to Total Income	--	3
Total	191	448
Total Households	11,065	12,121
percent with errors	1.7	3.7

Source: Buse (1979)

TABLE 3: Number of Households Exhibiting Selected Data Problems in the 1972-73 Bureau of Labor Statistics Consumer Expenditures Survey

Data Problem Description	Number of CU's
A. Data Errors	
1. Marital Status Inconsistent with Reported Data on Family Member - 2	127
2. Weeks Family Member Was in the Household Not Reported	30
3. Average Family Size Incorrectly Calculated	33
4. Incorrect Sign on Earnings	20
5. Earnings of "Other" Was Set to Zero Because No Other Family Member Was Listed	20
6. The Earnings of "Other" has Wrong Sign	2
7. State and Local Auto Registration Fees Not Included in Calculated Total Expenditures	a/
B. Data Warnings	
1. Calculated Family Size Inconsistent with Details on Family Member	27
2. Exchange Value of Food Stamps Was Less Than Cost	12
3. Exchange Value of Food Stamps Was Not Reported	2
4. Cost of Food Stamps Not Reported	57
5. Earnings Detail Inconsistent with Reported Total Earnings	6
Total	306
Number of Observations	19,975
Percent of CU's with Data Problem	1.6 percent

a/ Count is not included since there were more than 6,000 CU's with this error in their total expenditure record. It was a programming oversight in producing the original data tapes.

TABLE 4: Number of Households Exhibiting Selected Data Problems in the 1977-78 U.S. Department of Agriculture Nationwide Food Consumption Survey

Problem Description	Number
1. Alphabetic Codes in Data Fields	3
2. Interview Dates Out of Limit	280
3. Inconsistent Income Values	569
4. Income or Earnings Incorrectly Coded	88
5. Very Large Expenditures	41
6. One Expenditures More Than 80 percent of Weekly Income	40
7. Total Food Expenditures More Than 2.5 Time Income	64
8. Income Negative or Unreported	1,684
Total Number of Observations	14,930
Number of Observations with Data Problems	11.3 percent*

* Overstates error rate since some observations contained more than 1 error.

TABLE 5: Number of CU's Reporting Non-Zero Expenditure by Major Expenditure Category, 1972-73 BLS/CES

Expenditure Category	Households Reporting Non-Zero Expenditures	
	Number	Percent
1. Total Food	19,924	99.7
2. Alcohol	12,773	63.9
3. Tobacco	11,286	56.5
4. Housing	19,910	99.7
5. House Furnishings and Equipment	17,705	88.6
6. Clothing and Material	19,734	98.8
7. Transportation	18,818	94.2
8. Medical Care	19,237	96.3
9. Personal Care	16,866	84.4
10. Recreation	18,161	90.9
11. Reading	16,835	84.3
12. Education	4,920	24.6
13. Other Expenditures	13,016	65.2
Total No. of Observations	19,975	100.0

TABLE 6: Density of Major Expenditure in the 1972-73 BLS Diaries: Number of Households Reporting Non-Zero Expenditure by Type of Expenditure; 1972-73 BLS/CED and 1977-78 USDA/NFCS

Expenditure Category	Percent of Sample Reporting Non-Zero Expenditures	
	BLS/CED	USDA/HFCS
1. Total Food	95.9	100.0
A. Food At Home	94.2	99.8
cereal and bakery products	90.7	99.2
meat, fish and poultry	89.6	98.8
dairy	90.3	98.7
fruit	81.0	92.7
vegetables	83.7	96.2
sugar and sweets	64.4	66.8
fats and oils	64.7	67.2
non-alcoholic beverages	80.2	82.8
miscellaneous foods	80.9	83.6
B. Food Away From Home	78.0	80.1
meals	67.4	
snacks	49.7	
beverages	38.4	
2. Alcoholic Beverages	42.0	
3. Tobacco and Smoking Supplies	52.0	
4. Personal Care	70.7	
5. Non-prescription Drugs and Medicines	41.8	
6. Housekeeping Supplies	82.6	
7. Utilities and Fuels	41.0	
8. Automobile Fuel and Lubricants	73.6	
9. Miscellaneous	60.0	

TABLE 7: Density of Selected Subaggregates of Household Food Expenditures in the 1972-73 BLS Diary

Expenditure	CU's Reporting Non-Zero Expenditure	
	Number	Percent
1. Meat, Fish and Poultry	20,771	89.6
Beef and Veal	16,182	71.9
Ground Beef	10,869	46.9
Chuck Roast	9,767	16.2
Round Steak	2,492	10.7
Other Beef	3,913	16.9
Veal	823	3.5
Pork-Except Canned	15,331	66.1
Bacon	8,483	36.6
Chops	6,014	25.9
Sausage	5,882	25.4
Roasts	1,302	5.6
Other	3,909	16.9
2. Dairy Products	20,943	90.3
Fresh Milk and Cream	20,576	88.7
Processed Milk	5,181	22.3
Cheese	14,032	60.5
Yogurt	1,172	5.1
Ice Cream	8,603	37.1
Butter	5,232	22.6
Total No. of CU's	23,186	

TABLE 8: Distribution of Observation According to the Number of Major Categories of Expenditures Reported*, 1972-73 BLS Diary, Survey; and 1977-78 USDA/NFCS

Number of Non-Zero Expenditures Reported Observ.	BLS/CES Survey		BLS/CED Diary		USDA/NFCS	
	Number	Percent	Number	Percent	Number	Percent
0	0	–	812	3.5	7	–
1 or fewer	3	–	1,215	5.2	31	–
2 or fewer	18	.1	2,170	9.4	48	–
3 or fewer	60	.3	3,703	16.0	67	–
4 or fewer	170	.9	6,092	26.3	100	–
5 or fewer	426	2.1	9,631	41.5	178	–
6 or fewer	810	4.1	14,242	61.4	353	–
7 or fewer	1,542	7.7	18,858	81.3	870	.1
8 or fewer	2,672	13.4	22,061	95.1	2,268	15.2
9 or fewer	4,754	23.8	23,186	100.0	6,203	41.5
10 or fewer	8,431	42.2	--	--	14,903	100.0
11 or fewer	13,407	67.1	--	--		
12 or fewer	18,203	91.1	--	--		
13 or fewer	19,975	100.0	--	--		

* See Tables 5 and 6 for a definition of the major expenditure categories.

TABLE 9: Number of Households Reporting Specific Combinations of Expenditures: 1972-73 BLS Survey

Expenditure Set*	Households Reporting the Expenditures	
	Number	Percent
I	18,977	95.0
I, II	17,080	85.5
I to III	15,860	79.4
I to IV	12,855	64.4
I to V	4,761	23.8
All 6 Categories	1,772	8.9
Total	19,975	

* Expenditure Sets are defined as follows:

I = Food, Housing, Clothing, Medical Care

II = Transportation, Recreation

III = House Furnishings and Equipment

IV = Reading, Personal Care

V = Alcohol, Tobacco, Other

VI = Education

TABLE 10: Descriptive Statistics on Average Annual Expenditures of Households Reporting Non-Zero Expenditures, 1972-73 BLS/CES

	Expenditure	Annual Expenditures		Largest	Number of
		Mean	Standard Deviation	Value	Outliers*
1.	Food	$ 1,774	$ 1,262	$ 17,449	48
2.	Alcohol	131	219	6,000	78
3.	Tobacco	227	166	2,028	31
4.	Housing	2,129	1,591	46,433	71
5.	House Furnishings & Equipment	442	648	12,907	97
6.	Clothing & Materials	627	660	11,463	85
7.	Transportation	1,912	2,109	39,690	56
8.	Medical Care	493	578	31,239	67
9.	Personal Care	120	127	2,356	50
10.	Recreation	339	450	21,596	64
11.	Reading	58	75	1,581	101
12.	Education	444	885	11,469	40
13.	Miscellaneous	126	393	26,417	68
	Total Expenditure	$ 804	$ 5,201	$ 99,716	38
	Total Annual Income	$11,443	$15,127	$1,002,000	46

* Reported expenditures more than 5 standard deviations above the mean.

TABLE 11: Descriptive Statistics on Average Bi-Weekly Expenditure of Households Reporting Non-Zero Expenditures, 1972-73 BLS/CED

		Bi-Weekly Expenditures			
	Expenditure	Mean	Standard Deviation	Largest Value	Number of Outliers*
1.	Total Food	$ 69.04	$ 58	$ 4,099.93	45
	Food at Home	51.26	40	1,018.52	63
	Cereal and bakery	6.30	5	118.63	56
	Meat, fish, poultry, & eggs	21.40	24	1,022.28	86
	Dairy products	7.32	6	207.13	66
	Fruit	3.98	5	285.20	34
	Vegetables	4.45	4	76.33	64
	Sugar and sweets	2.26	3	66.70	70
	Fats and oils	2.05	2	56.00	64
	Non-alcoholic beverages	4.40	4	117.35	48
	Miscellaneous	4.86	5	231.54	50
	Food Away From Home	22.96	40	4,000.00	30
	Meals	21.94	41	4,000.00	25
	Snacks	4.63	5	88.78	55
	Beverages	2.14	5	355.65	25
2.	Alcoholic Beverages	10.76	15	434.76	49
3.	Tobacco and Smoking	8.41	7	122.87	38
4.	Personal Care	8.20	11	731.50	51
5.	Non-prescription Medicines	5.74	14	518.97	73
6.	Housekeeping Supplies	6.40	8	267.40	87
7.	Utilities and Fuels	32.21	32	988.80	32
8.	Gasoline, oil and coolants	19.01	20	1,245.78	65
9.	Miscellaneous	8.29	19	640.24	81
	Total Expenditures	125.18	92	4,109.52	39
	Total Annual Income	$11,658.00	$13,823	$6,650,000.00	49

* Reported expenditures more than 5 standard deviations above the mean.

TABLE 12: Descriptive Statistics on Average Bi-Weekly Expenditures of Households Reporting Non-Zero Expenditures, 1977-78 USDA/NFCS

	Weekly Expenditures			
Expenditures	Mean	Standard Deviation	Largest Value	Number of Outliers[1/]
1. Total Food	$ 55.39	$ 36	$ 501.42	32
2. Food-At-Home	40.13	25	243.06	35
Cereals	4.74	4	41.37	42
Meat, fish, poultry and eggs	14.84	11	120.96	45
Dairy	5.52	4	58.97	42
Fruit	3.22	3	28.62	51
Vegetables	3.84	3	33.51	46
Sweets	1.27	2	61.61	49
Fats and oils	1.49	1	13.84	42
Beverages	3.53	4	32.14	50
Miscellaneous	2.95	3	54.32	54
3. Food-Away-From-Home	20.08	22	461.00	47
4. Last Months Income[2/]	13,480.00	12,046	270,000.00	56
5. Last Years Income	14,140.00	11,300	226,800.00	41

[1/] Reported expenditures more than 5 standard deviations above the mean.

[2/] On an annual basis.

TABLE 13: Descriptive Statistics of Selected Expenditure Categories as a Percent of Total Expenditures, 1972-73 BLS Survey

Expenditure	Mean Proportion*	Standard Deviation	Largest Value	Number of CU's Greater than 5 S.D.**
1. Total Food	23.7	11.9	100	11
2. Alcohol	1.0	2.2	47	152
3. Tobacco	1.9	2.8	56	85
4. Housing	29.2	14.3	99	0
5. House Furnishings	2.9	5.5	63	72
6. Clothing	7.1	5.1	83	45
7. Transportation	19.2	14.7	100	1
8. Medical Care	6.6	6.6	100	90
9. Personal Care	1.3	1.6	56	59
10. Recreation	3.4	3.8	87	70
11. Reading	.6	.9	19	106
12. Education	.9	3.2	85	194
13. Miscellaneous	.9	2.6	64	174

* Mean of those CU's reporting non-zero expenditures.

** S.D. = Standard Deviation.

TABLE 14: Descriptive Statistics of Specific Expenditure Categories as a Percent of Total Expenditures, 1972-73 BLS Survey

		Expenditure Proportions			
	Expenditure	Mean Proportion*	Standard Deviation	Largest Value	Number of CU's Greater than 5 S.D.**
1.	Total Food	57.2	18.6	100	0
	Food At Home	43.8	20.5	100	0
	Cereal and bakery	5.7	4.8	100	88
	Meat, fish & poultry	17.4	11.6	100	43
	Dairy Products	6.6	5.7	100	78
	Fruits	3.6	3.7	100	78
	Vegetables	3.9	3.2	67	82
	Sugar & sweets	2.0	2.4	60	78
	Fats & oils	1.8	1.9	59	71
	Non-alcoholic bev.	3.8	3.4	100	68
	Miscellaneous	4.1	4.0	100	85
	Food Away From Home	17.5	15.4	100	0
	Meals	16.2	14.9	100	0
	Snacks	4.0	4.5	100	104
	Beverages	1.8	3.2	86	52
2.	Alcoholic Beverages	7.8	8.8	100	52
3.	Tobacco & Smoking	7.8	8.3	100	74
4.	Personal Care	6.5	7.4	100	88
5.	Non-prescription Medicines	4.3	8.2	100	87
6.	Housekeeping Supplies	5.1	5.0	100	98
7.	Utilities & Fuel	22.2	15.7	100	0
8.	Gasoline, oil & coolants	15.2	11.5	100	74
9.	Miscellaneous	5.6	8.1	100	112

* Mean of those CU's reporting non-zero expenditures.

** S.D. = Standard Deviation.