USING Q-Q PLOTS TO ASSESS COMPARABILITY BETWEEN NSAF AND CPS

Simon Pratt and Tamara Black, The Urban Institute
Simon Pratt, UI, 2100 M Street NW, Washington, DC 20037

Key Words: Quantiles, Q-Q Plots, Scatterplot Matrix, Outliers.

1. PURPOSE

The purpose of this paper is to assess the comparability between income data from the National Survey of America's Families (NSAF) and Current Population Surveys (CPS) in 1997 and 1999. Income variables are examined because of the importance of income in the NSAF, given that it is a survey which studies welfare issues. Also, dollar values lend themselves to being sorted and grouped into percentiles.

If the two surveys are found to be comparable, then it may be possible in some way to draw conclusions about NSAF data that could not otherwise be drawn. For example, the NSAF focuses on low-income households (where low-income is defined as less than 200% of the Federal Poverty Level), and so the sample sizes for high-income households are small relative to the CPS. Attention will be on the low-income populations which are roughly the same size in the CPS and the NSAF.

Perhaps more importantly, if the two surveys are not found to be comparable, then it is necessary to consider the reason for the difference, and make adjustments to the NSAF if necessary. Some areas of divergence have been discovered as a result of this study, and will be mentioned later.

2. SURVEY BACKGROUND

The two surveys compared in this paper are described briefly to give some background to the analysis that follows.

NSAF

The NSAF is a survey of the economic, health, and social characteristics of children, adults under the age of 65, and their families. Two rounds of interviews were conducted in 1997 and 1999, yielding information on over 40,000 households and 100,000 persons in each round. Westat conducted the data collection for the Urban Institute and Child Trends.

Large representative samples of households were taken in the nation as a whole with oversampling in 13 targeted states and the balance of the nation. The 13 states were Alabama, California, Colorado, Florida, Massachusetts, Michigan, Minnesota, Mississippi, New Jersey, New York, Texas, Washington, and Wisconsin.

These 13 states account for over half of the U.S. population and have a broad array of government programs, fiscal capacity, and child well-being. The sample results from the first round provide a wide range of characteristics for each of the targeted study areas and for the country as a whole, in the period just before the era of the New Federalism (when major changes in U.S. federal and state policies occurred). The sample results from the second round provide information on the characteristics of the targeted study areas and for the country as a whole after many of the changes of New Federalism had been implemented. Collectively, they form a sound baseline from which many of the changes brought about by the New Federalism can be measured, assessed, and tracked.

The NSAF sample is representative of the civilian, noninstitutionalized population under age 65. The first round of data was obtained from February to November 1997, and the second round of data was collected from February to October 1999. The NSAF sample had two parts: a main sample of an RDD survey of households with telephones; and a second (area probability) sample of households without telephones. Telephone households were subsampled, with the subsampling rates depending on the presence of children in the household and their response to a single household income-screening question. All households screened with children and classified as low-income were given a full interview, while higher-income households with children and all households without children (but with someone under 65) were subsampled before in-depth interviewing. Households with only adults age 65 and over were screened out of the survey.

CPS

The Current Population Survey (CPS) is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years.

The CPS is the primary source of information on the labor force characteristics of the U.S. population. In the same way as the NSAF, the sample is selected to represent the civilian, noninstitutionalized population. The employment status of each member of the household 15 years of age and older is obtained, though published data focus on those ages 16 and over.

Estimates of a variety of demographic characteristics including age, sex, race, marital status, and educational attainment are available. Indicators such as employment, unemployment, earnings, hours of work, occupation, industry, and class of worker are also available. The CPS contains monthly supplements which ask in-depth questions in areas such as school enrollment, income, previous work experience, health, employee benefits, and work schedules. In the March Supplement, in-depth questions are asked that relate to income and work experience, so it is this supplement that is most relevant to this particular study.

Coverage Differences

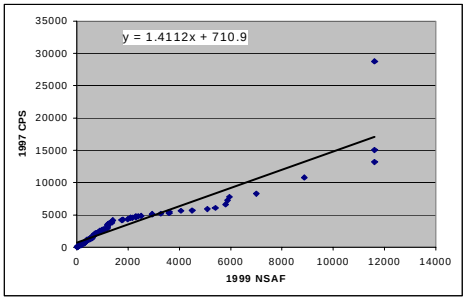
There are important differences between the NSAF and the CPS. Most important is that, in the NSAF, people aged 65 and over are only sampled if they live in a household in which people under the age of 65 also live. Thus, there are no households in which there are only people over the age of 64. There is no similar restriction in the CPS so, in order to make the two surveys comparable, it is necessary to remove such households from the CPS. Also, the CPS contains information about the employment status of those aged 15 years and over, whereas the NSAF starts at 18. However, the impact of those aged 15-17 on family income is very small, and thus had no impact on the results.

3. Q-Q BACKGROUND

The f^{th} quantile, $q(f)$, of a set of data is a value along the measurement scale of the data with the property that approximately a fraction f of the data are less that or equal to f (Cleveland, 1993). A quantile-quantile plot, or Q-Q plot, is the graphical representation of the magnitude of one set of quantiles plotted against that of another. Thus, it is a good visual means for understanding patterns across two sets of univariate numerical data.

When comparing sets of data, it is common to consider measures of central tendency such as the median. If greater distributional detail is required, then finer gradations like quartiles or, in this case, percentiles, may be examined. As an example, the Q-Q plot for 1999 NSAF vs 1997 CPS for non-AFDC public assistance is presented below:

Figure 1. Q-Q Plot of Non-AFDC Public Assistance: 1999 NSAF vs. 1997 CPS

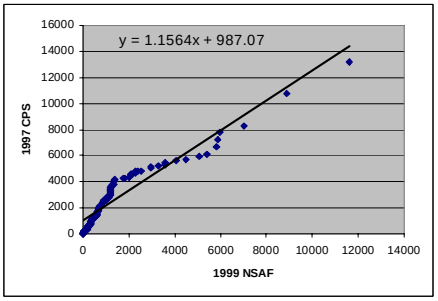


The Q-Q plots we are using are created by plotting all 100 percentiles. A trendline is then fitted to the plotted points using ordinary least squares regression. Working with quantiles is, in effect, the same as working with two sets of data that have been ordered from the smallest value to the largest value. Thus, the trendline will always have a positive slope. A Q-Q plot can sometimes identify an outlier. In Figure 1, the data point in the top-right corner of the Q-Q plot is an outlier. Outliers may be explained by transcription error, in which case the original source data should be corrected. However, if the data value appears to be correct, then there is a dilemma: should the value by kept given that it is a genuine part of the data, or should it be ignored on the basis that it appears to be anomalous and its inclusion is affecting the underlying trend of the data?

In the case above, we know what is causing the outlier. The final three quantiles for the NSAF share the same value (this is evident from the Q-Q plot because all three points lie on the same vertical line). They have the same value because the NSAF data have been topcoded (the procedure whereby the data values are not allowed to exceed a certain level, in order to minimize the risk of disclosure and to correct for what are most likely data errors). The CPS data have not

been topcoded (or they have, but at a much higher value so these data points are not affected). Thus, the outliers are a result of topcoding. They are removed so as not to affect the underlying trend of the other values.

Figure 2. Q-Q Plot of Non-AFDC Public Assistance: 1999 NSAF vs. 1997 CPS (with outliers removed)



In Figure 2, the final two observations from the previous figure have been removed, and it is interesting to note the change in the results. The x-coefficient from Figure 1 (with outliers included) is 1.41, and it falls to 1.16 in the figure above. This is a substantial difference, and demonstrates the importance of correcting for outliers.

Several features are common to all Q-Q plots. If two distributions are taken such that $F(x) = G((x- \mu/\sigma)$ (so that both are from the location-and-scale family of distributions) for all x, and specifically:

$$F(X_F - \mu_F / \sigma_F) = G(X_G - \mu_G / \sigma_G)$$

it can be shown that: [1]

$$F^{-1}(p) = \mu + \sigma G^{-1}(p) \quad \text{for } 0<p<1. \qquad (1)$$

$$\text{Where: } \mu = \mu_G - (\sigma_G / \sigma_F)\, \mu_F$$
$$\sigma = \sigma_G/\sigma_F$$

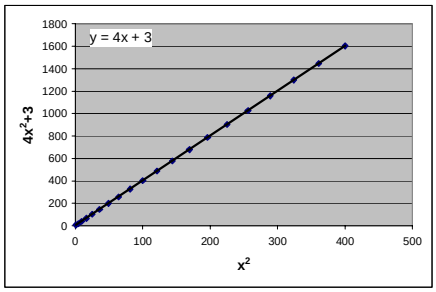
This means that the Q-Q plot takes the form of a straight line (with slope σ and intercept μ) provided that both F and G are from the same family of distributions. This can be demonstrated in simple terms by considering twenty observations from two polynomial distributions, $F(x) = x^2$ and $G(x) = 4x^2+3$.

[1] Hoaglin, Mostellar, and Tukey (1996) P.432

x	F(x)	G(x)
1	1	7
2	4	19
3	9	39
4	16	67
5	25	103
..	…	….

Given that the two distributions are both quadratic in form, the Q-Q plot will return a straight line:

Figure 3. Example of Q-Q Plot for Two Polynomial Distributions



Of course, if F and G are from different families of distributions, then the Q-Q plot will return a curve. For example, if one distribution was quadratic, and the other cubic, the Q-Q plot would return a curve concave to the vertical axis (if this measures the cubic function), and if one distribution was quadratic and the other quartic, then this curve would be more concave, and so on.

There are two other important properties that can be gleaned from (1). If the scale of each distribution is the same ($\sigma_F = \sigma_G$), then the Q-Q plot will have a slope of one and a non-zero intercept. If the scale of each distribution is the same, and the location is also the same ($\mu_F = \mu_G$), then the straight line on the Q-Q plot will pass through the origin and have a slope of one. In effect, the two distributions are identical (F = G).

In this paper, we focus on the Q-Q plots for child support income for families under 200% of the Federal Poverty Level. Child Support is characteristic of most of the income variables that we compared and also has some additional analytical features of interest. The reason for choosing the universe of families under 200% poverty is to maintain accordance with NSAF's focus on low-income families, and also because child support is a non-need-based variable. We calculate the quantiles for the CPS and the NSAF in both 1997 and 1999. This produces six Q-Q plots: two are across

survey but within year (1997 CPS vs 1997 NSAF and 1999 CPS vs 1999 NSAF); two are across year but within survey (1997 CPS vs 1999 CPS and 1997 NSAF vs 1999 NSAF); and two are across both survey and year (1997 CPS vs 1999 NSAF and 1999 CPS vs 1997 NSAF). These Q-Q plots are arranged into a scatterplot matrix. The scatterplot matrix for child support is presented at the end of this paper. A given label constitutes the x-axis for every graph above it, and the y-axis for every graph to the right of that label. Below each label there is, in parentheses, the percentage of families receiving child support. This is the recipiency rate. The Q-Q plots are based only on the noted percentage of respondents reporting recipiency.

4. FINDINGS

Of all of the plots in a given scatterplot matrix, we have particular interest in those that compare across surveys within a year. That is, we focus on comparisons of 1997 NSAF to 1997 CPS and comparisons of 1999 NSAF to 1999 CPS. There are many differences between the survey methods of the NSAF and the CPS. For example, the NSAF is largely a random-digit–dial (RDD) survey, whereas the CPS starts out as an in-person survey. This results in the CPS having a much higher response rate. Both are national, but the NSAF focuses on 13 states, whereas the CPS focuses on all states more or less equally. The exact wording of the child support question is slightly different. Nevertheless, essentially each survey is asking the same question of members of the same population during the same year, so it would be very encouraging if the responses from each survey were similar. This would result in a Q-Q plot where y = x and a value of R^2 close to 1.

The other plots of main interest are those that compare across years within a given survey. We are working with income variables, so we might expect the income values to increase over time (between 1997 and 1999 there was modest inflation), resulting in a trendline with an x-coefficient of less than 1 (note that in the scatterplot matrix, when comparing within a survey and across years, the 1997 survey is always on the y-axis and the 1999 survey is always on the x-axis.) The R^2 for these comparisons will be close to 1. The plots that compare across surveys and across years simultaneously are of less analytical value for our present purposes.

On the scatterplot matrix, the plots comparing 1997 NSAF to 1997 CPS and 1999 NSAF to 1999 CPS prove consistent with our expectations for a plot of two similar distributions. That is, the x-coefficient of the 1997 survey comparisons is 1.0027 and for the 1999 comparison it is 0.9726. Additionally, the R^2 values on both of these plots are close to 1. In 1997, a y-intercept of 145.72 seems high, but compared to a scale of nearly 15000, it is relatively small. Similarly, the y-intercept of 25.784 for 1999 is more than acceptable.

In the comparisons across years within a given survey, our expectations are also met. The x-coefficient for both NSAF and CPS comparisons is less than one, and the values of the y-intercepts are small. Once again the values for R^2 are very high. Overall, the scatterplots for child support demonstrate that, on this variable, respondents on all four distributions are similar, both across surveys and across years. Hypothesis testing will be conducted to determine whether there are any significant differences.

The value, and interpretation, of the R^2 needs to be mentioned. R^2, in this case, should <u>not</u> be thought of as a measure of the extent to which the independent variable is explaining the variation in the dependent variable. The NSAF does not explain the variation in the CPS. Rather, R^2 is a measure of comparability.

Another area of concern is the value attached to the R^2. The values are always very high. However, this should not be surprising. The two data sets have been ordered from lowest to highest, and then plotted, with a trendline fitted to those data points. It is inevitable that the value for R^2 will be very high. The important question is what value of R^2, in this context, is good, and what value is poor?

In order to address this question, we observed the behavior of random numbers. It was mentioned earlier that two sets of data from the same family of distributions, when plotted on a Q-Q plot, will produce a straight line. Conversely, two sets of data from different families of distributions will produce a curve. If a straight line is then fitted to the curve, one would expect the value of R^2 to be quite low. Thus, this would be a good test of a poor value for R^2.

We took two ordered sets of 100 random numbers, one from the uniform distribution, and one from the normal distribution, and plotted them on a Q-Q plot. A trendline is fitted to the signature ‘S’ shape of the curve, yielding a value of R^2=0.9. This tells us that a value of R^2=0.9 is poor given that the data sets have been ordered. Another test is to consider two sets of observations from the same distribution. We took two ordered sets of 100 random numbers both from the

uniform distribution. This time, R^2=0.98. However, this value is still lower than we might expect from our Q-Q plots. This is because the sample size of 100 is much smaller than the sample size of several thousand that we are presented with by the surveys. The next test, therefore, was to take two ordered sets of 1000 uniform random numbers. This time, R^2=0.9988. This gives us some idea of what might be a 'good' value for R^2.

It is for these two reasons – the different interpretation of R^2, and the high value attached to all values of R^2 – that lead us to change the notation to Q^2 on the Q-Q plots. Despite the fact that it is calculated in the same way, R^2 behaves quite differently in the context of quantile regression, and so it is helpful to give it a different label as a reminder that it is conceptually different to the R^2 in ordinary least squares regression.

Figure 4. Two ordered sets of 100 random numbers from the Normal and the Uniform Distributions

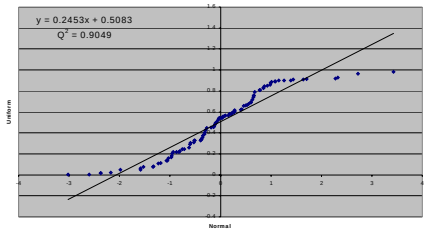
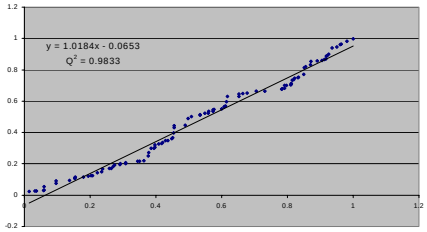


Figure 5. Two ordered sets of 100 uniform random numbers



Another area of concern is the recipiency rates, or the percentage of families receiving child support, especially in 1999. The values for 1999 show a large discrepancy between the NSAF and the CPS. The value for the NSAF is 13.4% whereas the value for the CPS is just 6.9%. Why are the NSAF and CPS rates so different? One possible explanation is that the NSAF has a series of questions prior to the question on child support which are given to any household in which one of the biological parents lives outside that household. This may serve as a trigger or reminder to the respondent that they receive child support.

Even if this is the explanation, however, it is still necessary to know if this difference is affecting the outcome of the Q-Q plot predictions. We ran a significance test on the difference between the two percentages, and found the difference to be significant. At this point, it is necessary to ask if there is a systematic bias. That is: are a disproportionate amount of the extra respondents in the NSAF clustered in either the lower- or upper-end of the distribution? If so, then we would expect this to significantly alter the value of the x-coefficient. There is no sign of clustering in the Q-Q plots themselves. Similarly, a breakdown by race indicated that the additional NSAF respondents are not concentrated in any particular racial category. Thus, we can tentatively conclude that the discrepancy in recipiency rates is not undermining the analytical value of the Q-Q plots, though it may still be a matter of survey interest.

5. NEXT STEPS

Further work needs to be conducted on the value of Q^2 that is acceptable for Q-Q plots. This includes both experimental and theoretical work.

Also important are hypothesis tests for each Q-Q plot. Tests for x-coefficient = 1, and y-intercept = 0 will give a more definite answer to the comparability of the two surveys considered in this study. However, hypothesis tests for Q-Q plots cannot be conducted in the usual manner due to the existence of serial correlation between the error terms, caused by the ordering of the two sets of data. Standard errors can be calculated, however, for the CPS by using the alpha and beta coefficients, and in the NSAF using the replicate weights.

6. REFERENCES

Hoaglin, Dave, Frederick Mostellar and John Tukey. *Exploring Data Tables, Trends, and Shapes.*

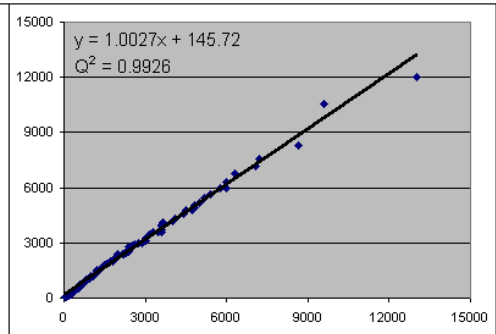
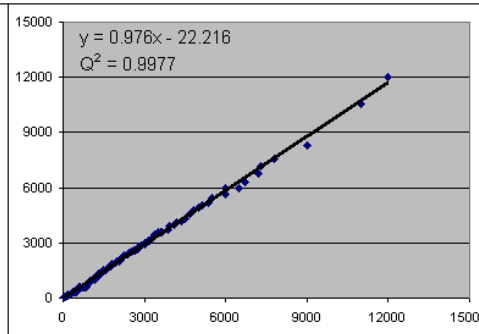
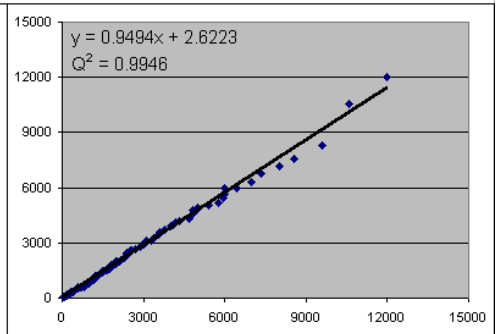
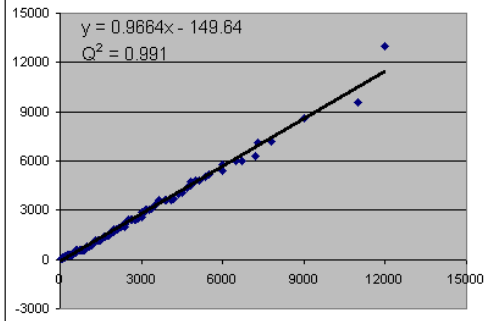
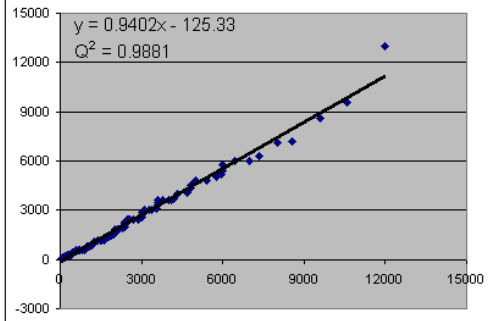
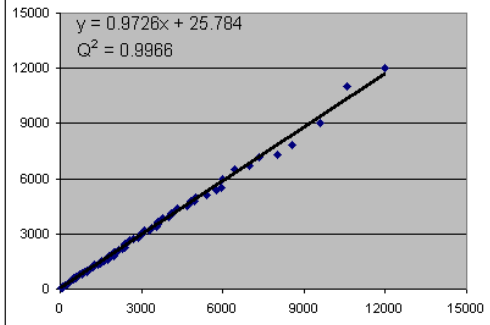
Wang, Kevin, David Cantor and Nancy Vaden-Kiernan (2001). *1999 NSAF Questionnaire: Methodology Report No. 1.*

Wheaton, Laura and Linda Giannarelli. "Underreporting of Means-Tested Transfer Programs in the March CPS."

Wilk, M.B. and R. Gnanadesikan. "Probability Plotting Methods for the Analysis of Data."

William Cleveland. (1993) *Visualizing Data.*

Figure 7. Child Support Receipt Among Families Below 200% Federal Poverty Level

1997 CPS (% Families Receiving Child Support =7.3%)			
	1997 NSAF (% Families Receiving Child Support =12.4%)		
		1999 CPS (% Families Receiving Child Support = 6.9%)	
			1999 NSAF (% Families Receiving Child Support =13.4%)