Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 3, pp. 851~859

A Comparison of NLSY and CPS Data[1)]

Yoonae Jo[2)]

Abstract

The family income distributions of NLSY97 and CPS youth data are compared by using the generalized beta distribution of the second kind. The null hypothesis that the two data sets represent the same underlying population is rejected. The ML estimation suggests that NLSY97 data are oversampled in an income group of $11,308 or less, by about 15.7% compared to CPS data.

Keywords : CPS, Generalized beta of the second kind, Income distribution, NLSY

1. Introduction

The sampling weight of NLSY97 (National Longitudinal Survey of Youth 1997) data is based on the respondent's characteristics such as gender, ethnicity, year of birth, sample type and location.[3)] However it does not account for family income, which is one of the most important variables in explaining the behavioral aspects of the young cohort of NLSY97. Therefore the estimated distributions of the behavioral variables based on the current sampling weights may not be representative of the underlying population.

In this paper, the family income distributions of NLSY97 and 1997 CPS (Current Population Survey) March Annual Demographic File are compared to see if the two data sets represent the same population. The NLSY97 data are collected for youth of age 12 to 16. For comparison, the 1997 CPS personal data are restricted to the same age groups so that the two can represent the same population in principle. <Table 1> shows the frequency tables of family income of the NLSY97

---

1) This research was supported by Sangji University Research Fund, 2005.

2) Assistant Professor, Dept. of Pubic Administration, Sangji University, Wonju, 220-702, Korea
E-mail : yoonaejo@sangji.ac.kr

3) There are two sets of NLSY data; NLSY79 and NLSY97.

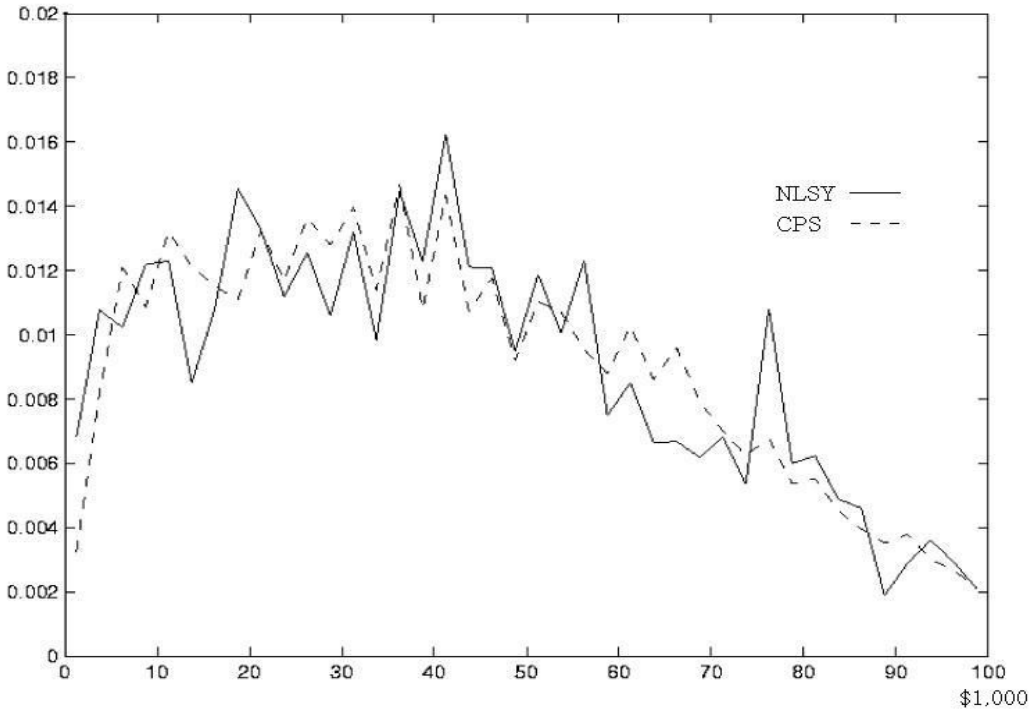
and CPS data. The frequencies are given for a total of 41 income groups with the corresponding survey weight. They are also graphically illustrated in <Figure 1>.

<Table 1> Family income distributions of CPS and NLSY97 data

Income ($1,000)	CPS (%)	NLSY (%)	Income ($1,000)	CPS (%)	NLSY (%)
0.0-2.5	0.8101	1.7032	50.0-52.5	2.7564	2.9704
2.5-5.0	2.0358	2.6914	52.5-55.0	2.6775	2.5112
5.0-7.5	3.0247	2.5577	55.0-57.5	2.3830	3.0809
7.5-10.0	2.7144	3.0460	57.5-60.0	2.1988	1.8718
10.0-12.5	3.3035	3.0750	60.0-62.5	2.5671	2.1275
12.5-15.0	3.0300	2.1275	62.5-65.0	2.1515	1.6625
15.0-17.5	2.8774	2.7030	65.0-67.5	2.3987	1.6683
17.5-20.0	2.7722	3.6389	67.5-70.0	1.9779	1.5462
20.0-22.5	3.3140	3.3134	70.0-72.5	1.7517	1.7090
22.5-25.0	2.9248	2.7960	72.5-75.0	1.5623	1.3370
25.0-27.5	3.4087	3.1390	75.0-77.5	1.6938	2.7030
27.5-30.0	3.1983	2.6449	77.5-80.0	1.3414	1.4997
30.0-32.5	3.4929	3.3017	80.0-82.5	1.3782	1.5579
32.5-35.0	2.8459	2.4589	82.5-85.0	1.1310	1.2207
35.0-37.5	3.6665	3.6156	85.0-87.5	0.9890	1.1510
37.5-40.0	2.6986	3.0692	87.5-90.0	0.8785	0.4708
40.0-42.5	3.5823	4.0574	90.0-92.5	0.9469	0.7208
42.5-45.0	2.6828	3.0285	92.5-95.0	0.7575	0.9010
45.0-47.5	2.9406	3.0169	95.0-97.5	0.6628	0.7324
47.5-50.0	2.2988	2.3717	97.5-100.0	0.5418	0.5232
			100.0-∞	9.6318	9.6785
			Mean	54.6071	52.8924

Note: Mean is computed from the raw data, not from the grouped data.

Strictly speaking, the frequencies of the two data sets given in the table and figure are not the same. However it does not necessarily mean that the two data sets represent different populations. Even though being drawn from the same population, the observed distributions of the two data sets may, and are most likely to, differ from each other because of the randomness of sampling. Comparing the numbers of the frequencies alone does not give a clear-cut answer to the question on their underlying population distributions. Instead of relying on the subjective conjecture based on the frequency numbers, a more direct statistical test is conducted in this paper. The population distributions of the two data sets are estimated from the observed distributions by using popular income distribution models. And the null hypothesis that the two underlying distributions are the same is statistically tested with their estimation.



<Figure 1> Family income distributions of CPS and NLSY97 data

This paper consists of 4 sections. The next section briefly surveys the estimation method of income distribution. Section 3 provides the estimation results of the family income distribution of NLSY97 and CPS data. Conclusion will follow in the final section.

2. ML estimation of income distribution

Various forms of distribution functions have been suggested either to capture the regularity characteristics observed in an empirical distribution of income, or just to better fit to empirical data. The parameters of a distribution can be estimated by either a maximum likelihood (ML) method, a minimum χ^2 method, or maximum product spacing (MPS) estimation when parameters are on the boundary of parameter space. In this paper, ML estimation is adopted.

Suppose there are T intervals $\{[y_{i-1}, y_i)\}_{i=1}^{T}$, such that $0 = y_0 < y_1 < \cdots < y_T = \infty$. Then the log-likelihood is given by,

$$logL(\theta) = log(N!) + \sum_{i=1}^{T}(n_i log p_i(\theta) - log(n_i!)) \tag{2.1}$$

where, n_i is the number of observations in the i-th interval, $N = \sum_{i=1}^{T} n_i$. And $p_i(\theta) = F(y_i;\theta) - F(y_{i-1};\theta)$ where $F(.)$ is the cumulative distribution function with distribution parameters θ.

The criteria of judging 'goodness of fit' include the log-likelihood, χ^2, sum of squared errors (SSE), and sum of absolute errors (SAE). The χ^2, SSE, and SAE are given by estimating,

$$\chi^2 = N\sum_{i=1}^{T}\frac{(\frac{n_i}{N} - p_i(\hat{\theta}))^2}{p_i(\hat{\theta})} \tag{2.2}$$

$$SSE = \sum_{i=1}^{T}(\frac{n_i}{N} - p_i(\hat{\theta}))^2 \tag{2.3}$$

$$SAE = \sum_{i=1}^{T}|\frac{n_i}{N} - p_i(\hat{\theta})| \tag{2.4}$$

respectively.

Income distribution functions are nicely summarized in a distribution tree by McDonald (1984), and McDonald and Xu (1995). And they can be broadly categorized by the number of distribution parameters. In this paper, the generalized beta of the second kind, Burr type 3, and lognormal distributions are estimated. Each of three is reportedly the better distribution among distributions with parameters 4, 3, and 2 respectively.

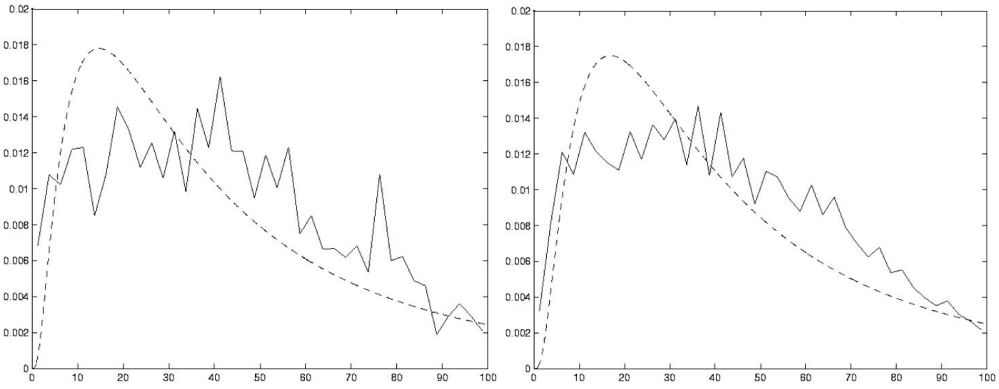
The lognormal distribution may be among the oldest distribution estimated for income distribution. It would be because its computation can be done less costly and more feasibly, especially in early days. The Burr type 3 distribution is better known as an income distribution model by Dagum (1977). In general, it gives a better fit of U.S. income than any other 3 parameter distributions including the generalized gamma, Singh and Maddala (1976) (Burr type 12), beta of the first kind and the second kind. The generalized beta of the second kind nests the Burr type 3, and therefore gives better fitting than the other two. The four parameter distribution by Majumder and Chakravarty (1990) is just a reparemeterization of the generalized beta of the second kind, as pointed by McDonald and Mantrala (1995).

3. Estimation results

<Table 2> summarizes the ML estimation results of a lognormal distribution model for both NLSY97 and CPS data. The lognormal distribution is restrictive in that it only attempts to approximate the mean and variance within its long-tail shape. The ML estimation is not efficient, which is shown in the large values of the χ^2, SSE and SAE, and the lower value of the log-likelihood. <Figure 2> shows the resulting fitting. It is easy to find that it is rather loose. The table also reports the estimated population mean of family income.

<Table 2> ML estimation of family income distributions: lognormal

	CPS	NLSY97
μ	3.6369	3.6112
σ	0.9007	0.9612
SSE	0.0039	0.0060
SAE	0.2742	0.3368
χ^2	2516.7169	3978.2229
$logL$	-1184.3657	-1650.4282
Mean	56.9680	58.7388



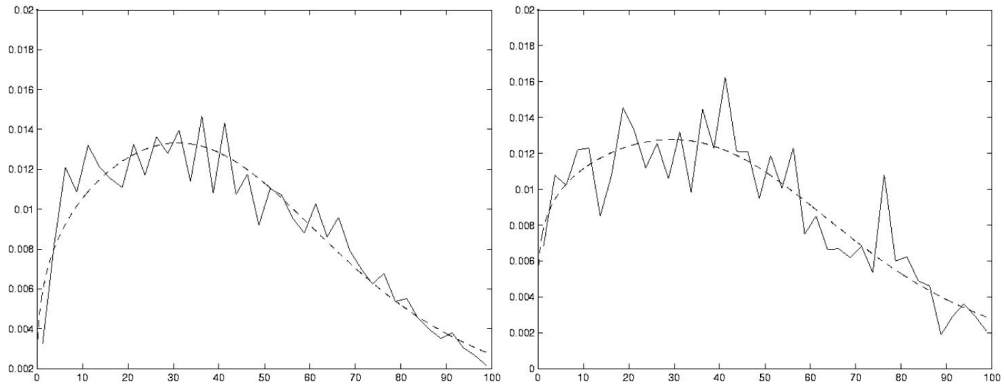
(a) CPS (b) NLSY

<Figure 2> Observed and estimated family income distributions: lognormal

<Table 3> summarizes the estimation results of the Burr type 3 distribution model. The ML estimation becomes quite efficient compared to the lognormal model. All fitting statistics are much better than before. And the resulting fitting in <Figure 3> is very tight in both NLSY97 and CPS data.

<Table 3> ML estimation of family income distributions: Burr type 3

	CPS	NLSY97
a	3.3514	3.5371
b	70.0719	74.4848
p	0.3827	0.3310
SSE	0.0003	0.0007
SAE	0.0878	0.1259
χ^2	275.2585	558.9367
$logL$	-294.7251	-422.2922
Mean	51.9785	51.1907



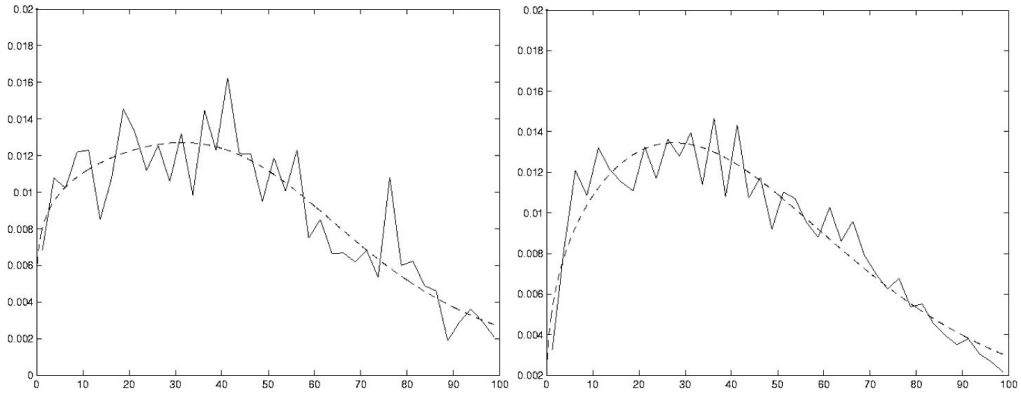
(a) CPS (b) NLSY

<Figure 3> Observed and estimated family income distributions: Burr type 3

<Table 4> summarizes the estimation results of the generalized beta distribution of the second kind. The ML estimation is the most efficient of the three models in all fitting statistics. The resulting fitting is shown in <Figure 4> and it is very tight. Although the generalized beta of the second kind is slightly better than the Burr type 3, the estimation results of the two distributions are very comparable. And it is also consistent with the results of the other studies.

<Table 4> ML estimation of family income distributions: generalized beta of the second kind

	CPS	NLSY97
a	2.3224	4.2004
b	95.5273	68.4288
p	0.5823	0.2749
q	2.2823	0.7281
SSE	0.0003	0.0007
SAE	0.0898	0.1250
χ^2	267.8493	558.8202
$logL$	-289.3554	-420.3178
Mean	50.7563	51.9234



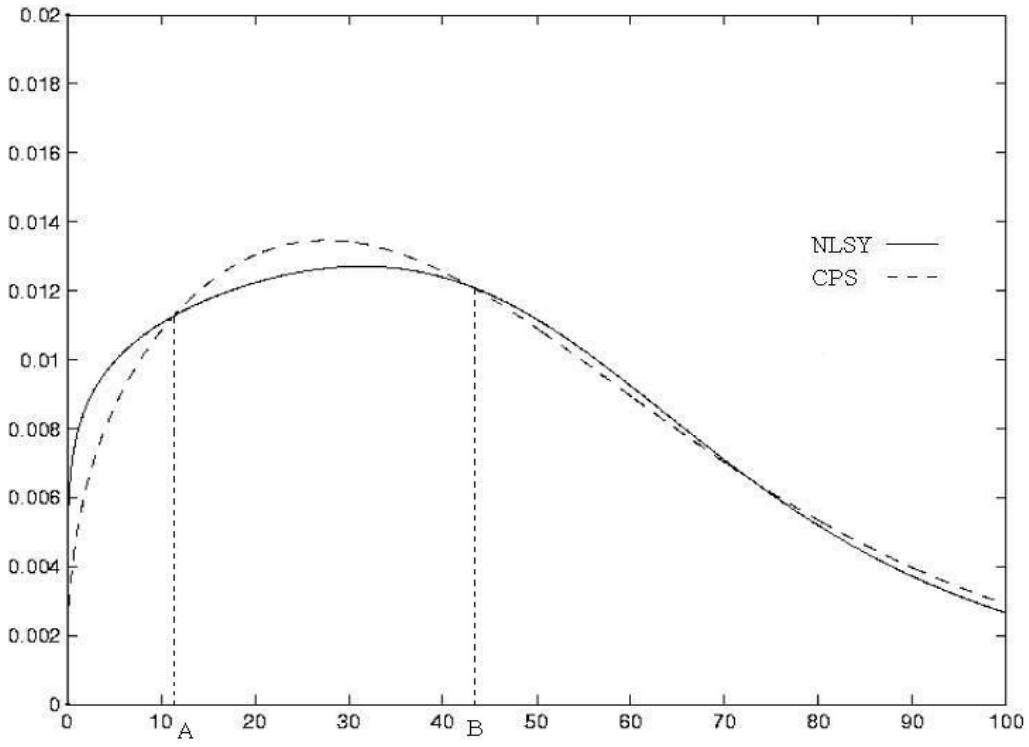
(a) CPS (b) NLSY

<Figure 4> Observed and estimated family income distributions: generalized beta of the second kind

Assuming that the two distributions can be represented by the same distribution model, the null hypothesis that the two income distributions are the same is equivalent to that the parameters of the two distributions coincide. The null hypothesis is tested by the Wald test, using the generalized beta of the second kind estimation. The estimated Wald test statistic is 17255.4166 and its p -value is virtually zero, which clearly rejects the null hypothesis.

<Figure 5> compares the estimated family income distributions of NLSY97 and CPS data by the generalized beta of the second kind distribution. The probability density functions intersect at A and B. They are estimated as $11,308 and $43,041 respectively, and define three income groups. <Table 5> reports the estimated frequencies for the three income groups. They are 9.53%, 40.85% and 49.62% for

the CPS data, and 11.03%, 39.12%, and 49.85% for the NLSY97. In other words, NLSY97 data are relatively oversampled by 15.70% and 0.47% for the low and high income groups, and are undersampled by 4.23% for the middle income group, compared to CPS data.



<Figure 5> Estimated family income distributions: generalized beta of the second kind

<Table 5> Estimated family income frequencies

income interval (%)	estimated frequency (%)		absolute oversampling (%)	relative oversampling (%)
	CPS	NLSY		
0-11,308	9.5334	11.0303	1.4969	15.7013
11,308-43,041	40.8488	39.1208	-1.7280	-4.2302
43,041-∞	49.6178	49.8489	0.2311	0.4658

4. Concluding remarks

The null hypothesis that the NLSY97 and CPS represent the same population is rejected. The oversampling of the low income group and the undersampling of the middle may account for this. The NLSY97 cohort is designed to oversample Hispanic and black people, who are relatively poor in population. Although a weight variable is expected to adjust for this oversampling, this paper shows that it still may not be enough. Researchers should take extra care in generalizing their statistical analysis of NLSY97 data, considering this possible sampling bias. And it would be fruitable to reconstruct the NLSY97 survey weight so that it can more appropriately capture the family income distribution of the U.S. youth population.

References

1. Dagum, C. (1977). A new model of personal income distribution: Specification and estimation, *Economie Appliquée*, 30(3), 413-437.
2. Majumder, A., & Chakravarty, S. R. (1990). Distribution of personal income: Development of a new model and its application to U.S. income data, *Journal of Applied Econometrics*, 5(2), 189-196.
3. McDonald, J. B. (1984). Some generalized functions for the size distribution of income, *Econometrica*, 52(3), 647-663.
4. McDonald, J. B., & Mantrala, A. (1995). The distribution of personal income: Revisited, *Journal of Applied Econometrics*, 10(2), 201-204.
5. McDonald, J. B., & Xu, Y. J. (1995). A generalization of the beta distribution with applications, *Journal of Econometrics*, 66(1), 133-152.
6. Singh, S. K., & Maddala, G. S. (1976). A function for size distribution of income, *Econometrica*, 44(5), 963-970.

[received date : May. 2006, accepted date : Jul. 2006]