Journal of Official Statistics
Vol. 1, No. 1, 1985, pp. 5–19

The Utility of the Cube Root of Income

Joseph E. Schwartz[1]

Abstract: Social scientists, especially economists, have long thought that the distribution of income is roughly lognormal. This belief has justified, at least implicitly, using the standard deviation of the logarithm as a measure of income inequality and using the logarithm of income as a dependent variable in multivariate analyses. This paper examines the full family of power transformations with several years of American income data and finds that the cube root – a transformation intermediate between no transformation and the log transformation – most closely approximates a normal distribution. The cube root of income exhibits additional statistical properties that make it perhaps the most suitable transformation for multivariate analyses of income.

1. Introduction

This paper is about income and the way we think about income. At its core lies the assumption that the best way to study income may be to study a transformation of it. This is obviously not an original assumption since, after all, economists usually analyze the logarithm of income rather than income. There is also a second assumption: that the use of alternative transformations has some effect on one's analyses and a *substantial* effect on the interpretation of these analyses. Because of this, it is important to choose an appropriate transformation before applying the usual multivariate methods to income data. In the following, some of the principal criteria for selecting a transformation are outlined and subsequently applied in order to find the best transformation for the analysis of income data.

[1] Columbia University, Department of Sociology and Center for the Social Sciences and Stockholm University, Institute for Social Research, 1981 Level of Living Project. The author is indebted to Paul Holland, Christopher Jencks, James Mirrlees, Lee Rainwater, Harrison White, and two anonymous reviewers for their comments on earlier versions of this paper.

There are four principal criteria for selecting a transformation. The first two criteria are derived from the basic assumptions of analysis-of-variance and regression models; that the residuals/errors of the model are independent and identically distributed with a normal (Gaussian) distribution having a mean of zero and a variance of σ^2, i.e. $N(0, \sigma^2)$. In analysis-of-variance this implies within-group normality – that the dependent (endogenous) variable within each group, defined by the independent (exogenous) variable(s), has a normal distribution – and homogeneity of variances – that the within-group variances are all equal (to σ^2). Thus, two reasons for transforming a variable are:

i. to *increase the normality of the within-group distributions* of the dependent variable; and
ii. to reduce the heterogeneity of the within-group variances or to *increase the homogeneity of variances.*

The third criterion for transforming a variable is:

iii. to *increase the linearity and/or additivity* of the relationship between the dependent and independent variables.

Non-linearity and non-additivity can frequently be treated by incorporating higher-order polynomials or multiplicative interaction terms for the independent variables into one's model. However, an alternative is to seek a transformation of one or more of the variables, perhaps especially the dependent variable, such that the resulting relationships are linear and additive. Frequently such a transformation exists and the resulting model is more parsimonious than those with polynomials or interactions. By transforming a variable, we alter the *functional form* of the relationship that is being estimated between it and the other variables.

The final criterion for transforming a variable is not statistical, but substantive. In general, there is no reason why the form in which data are collected should dictate the form in which they are analyzed. As Blalock (1982) has emphasized, theory may suggest or specify which transformation is appropriate, usually by implying a specific functional form for the relationship between two or more variables. For example, since an income elasticity is estimated by regressing the logarithm of the demand for a good on the logarithm of income, those who think that elasticities are theoretically the most appropriate tool (perhaps because they are unitless) for describing demand curves will transform income into its logarithm: the coefficient from any other regression would not be an elasticity. The theoretical reason for estimating an elasticity is likely to be an implicit assumption that income elasticities are approximately constant at different levels of income. If so, this amounts to an assumption of linearity and an attempt should be made to test its validity. Thus, we may also transform a variable for:

iv. *theoretical/conceptual reasons.*

Both statistical criteria and theory should guide the choice of functional forms. However, since this is not a paper on the translation of theory into mathematical equations, little more will be said about this criterion for choosing a transformation except to suggest that when the three statistical criteria strongly support a particular transformation, the researcher should seriously search for possible substantive interpretations of the transformed variable; the data may be trying to tell her something.

Traditionally, income has either been transformed to its logarithm or not been transformed at all: the transformation which leaves a variable untransformed is called the "identity" transformation. These are two special cases of the one-parameter family of power transformations:

$$f(x) = (1/p)x^p \quad \text{for} \quad p \neq 0, \text{ and}$$

$$f(x) = \ln(x) \quad \text{for} \quad p=0.^2$$

It is generally accepted that the log transformation ($p=0$) behaves better than the identity transformation ($p=1$) with respect to the three above-mentioned statistical criteria. This paper considers the whole family of power transformations and determines that, according to each of the criteria, a different (intermediate) transformation – the cube root ($p = 1/3$) – is markedly superior to both of the traditional transformations. For this reason, we shall conclude that the cube root of income is a more appropriate dependent variable for multivariate analyses.

The data for these analyses come from several sources. Sections 2 and 3 use published tables (Table 176 of the Handbook of Labour Statistics, U.S. Department of Labor (1972))

[2] Those who are unfamiliar with the practice of treating the logarithmic transformation as the power transformation when p equals zero might prefer to define this family of transformations as

$$f(x:p) = \int x^{p-1} dx.$$

These transformations are usually only applied to non-negative incomes though a subset of them, including the cube root are also appropriate for zero incomes and even negative incomes.

of the U.S. Current Population Survey's (CPS) annual family income distributions for seven years, disaggregated by race and education. For each year there are five levels of education for two races, resulting in seventy income distributions. Section 3 also uses longitudinal microdata from the Michigan Panel Study of Income Dynamics (PSID), while Section 4 relies on cross-sectional data from several sources including the PSID and the 1970 U.S. Census.

2. The Shape of the Income Distribution: Transforming to Promote Normality

Those who study income know that its distribution is positively skewed. Beyond this, different economists have claimed that the distribution of income conforms to one or another family of distributional forms including Pareto (primarily for the upper tail of the income distribution), lognormal (occasionally called Gibrat), and displaced lognormal[3]. As their name suggests, the latter two types imply that the income distribution is a transformation of a normal (Gaussian) distribution. The family of displaced log transformations, $f(x) = \ln(x + \text{constant})$, is one generalization of the log transformation. However, the family of power transformations is an alternative generalization that incorporates both the log and identity transformations, the two forms of income that are used most commonly in multivariate analyses. Power transformations are also effective at altering the skewness of a distribution; the lower the power, the less positive (or more negative) the skewness of the transformed variable becomes. For these reasons, we shall determine which power transformation of income has the most normal distribution.

[3] Gibrat's (1931) analyses were probably the first to suggest that income has a lognormal or displaced lognormal distribution. Metcalf (1972) provides a useful summary of the relevant literature on the distribution of personal income. (Also, see chapter 6 of Pen (1971) or Bronfenbrenner (1971).)

In order to find the most normal power transformation, one needs a measure of deviation from normality. Since the seventy (within race-by-education-by-year) CPS income distributions are already categorized, the χ^2-statistic is appropriate. For each of several powers, we have transformed each of the seventy income distributions and measured its deviation from the best-fitting normal distribution[4]. This yields a measure of the non-normality of each transformation of each income distribution.

The logic of inference does not allow one to prove or even demonstrate the validity of the null hypothesis that a particular power transformation of income is normally distributed. Inference only enables us to test the *statistical significance* of observed deviations from the null model. Furthermore, this statistical test is a function of two parameters: a) the magnitude of the deviations of the observed probability distribution from the predicted; and b) the sample size. It is clear that there exists a sample size for which any power transformation of an income distribution will differ significantly from normality. It is equally clear that there also exists a (much smaller) sample size for which the deviations of a range of transformations, including the log and identity, from normality would not be significant. But neither of these facts should distract from the primary concern of *comparing* the magnitude of the deviations of alternative power transformations of the observed distributions from a normal distribution. Therefore, the χ^2-statistics have been standardized to a constant sample size of 1 000. Fig. 1 shows how the

[4] A description of the algorithms that have been used for estimating the mean and standard deviation of that Gaussian distribution which is most similar to an empirically observed categorized distribution is available upon request. It is possible to minimize either the log likelihood ratio or Pearson goodness-of-fit χ^2-statistic. While the reported results are based on minimizing the latter, the difference between the alternative analyses are minor.

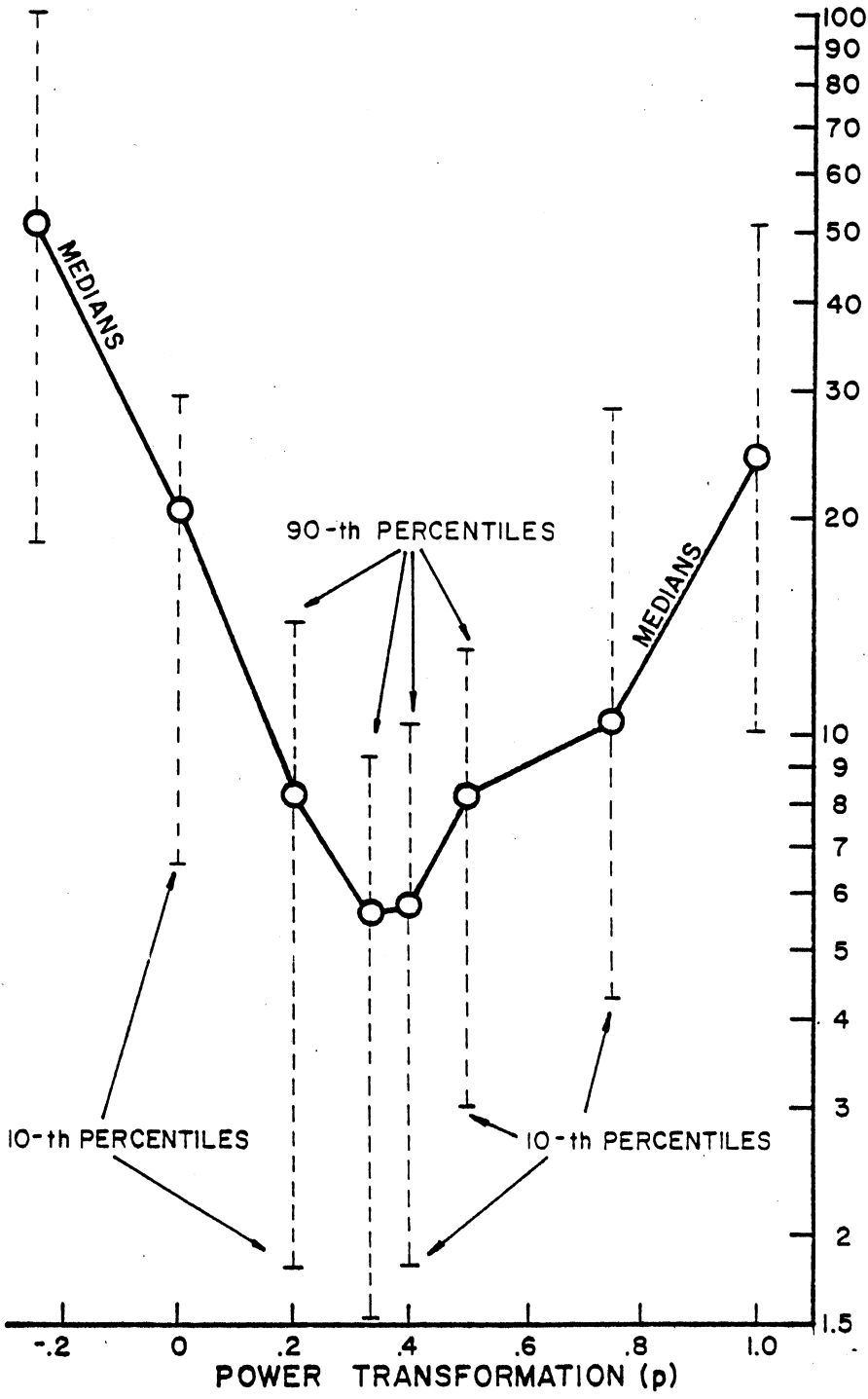

Fig. 1: Weighted Distributions of Standardized χ^2-statistics of Deviations from Normality for Assorted Power Transformations of Income-Medians, Tenth Percentiles, and Ninetieth Percentiles

distributions of these standardized measures of non-normality vary by power transformation. It indicates, for example, that of the seventy log ($p=0$) distributions, fifty percent have a standardized χ^2-statistic less than 21, ten percent have scores less than 7, while another ten percent have scores greater than 30.

It has been argued that one of the reasons for transforming income data to the log is that the transformed distribution is more normal. Fig. 1 shows that while this is indeed true (sixty-three percent of the distributions are more normal in the log than when they are untransformed), the cube root of income is more normal than either the log or identity transformations. In fact, when these three transformations are compared, seven and three percent of the income distributions are most normal under the logarithmic and identity transformations respectively, while the remaining ninety percent are most normal in the cube root. The evidence overwhelmingly supports the conclusion that the power transformation which deviates least from normality, across seventy separate CPS income distributions, is approximately the cube root.

3. Comparing Income Distributions: Transforming to Promote Homogeneous Variances

When comparing two income distributions (e.g., of blacks and whites) it is common to compute either the ratio of the two medians or the absolute difference between the medians. The purpose of comparing two medians (or means) is to summarize the difference between the distributions and not simply the difference between a single point (albeit, the center) of each distribution. However, unless one distribution equals the other plus a constant – implying that they have comparable amounts of spread – the absolute difference between their medians will differ from the absolute difference at other percentiles of the two distributions. Similarly, unless one distribution is a multiple of the other (in the logs they would differ by a constant and have equal variances), the ratio of their two medians (related to the difference in the logs) will differ from the ratio taken at another percentile. If the income distributions are not multiples of each other, we must question both the *significance* of the fact that the ratio of the two medians is .65 and the *validity* of a statement such as "Black income is about 65 percent of white income."

To say that two distributions differ only by a constant is equivalent, statistically, to saying that the second and higher moments of the two distributions are equal. But if the distributions are roughly normal, then it is the equality (homogeneity) of their variances which is crucial, since a normal distribution is completely determined by its first two moments. This is also the assumption that underlies conventional significance testing in multivariate data analysis. The most common and important type of heteroscedasticity is the presence of a significant relationship between the variance and mean of the different distributions. Once again, power transformations are often effective in reducing this type of heterogeneity.

Tukey (1970) describes a method by which one can use the bivariate regression of the log of the inter-quartile range on the log of the median to approximately determine which power transformation will minimize the monotonic relationship between them. From the unstandardized regression,

$$\ln(\text{IQR}) = a + b \ln(\hat{x}),$$

where IQR is the inter-quartile range and $\hat{x}$ is the median,
the appropriate power (p) transformation is given by,

$$p = 1.00 - b.$$

It is readily observed that if there is initially no monotonic relationship between the median

and inter-quartile range, then b will equal zero and the power will equal unity, the identity transformation. Similarly, if distributions tend to be multiples of one another, then b will be unity and the power will equal zero indicating the log transformation.

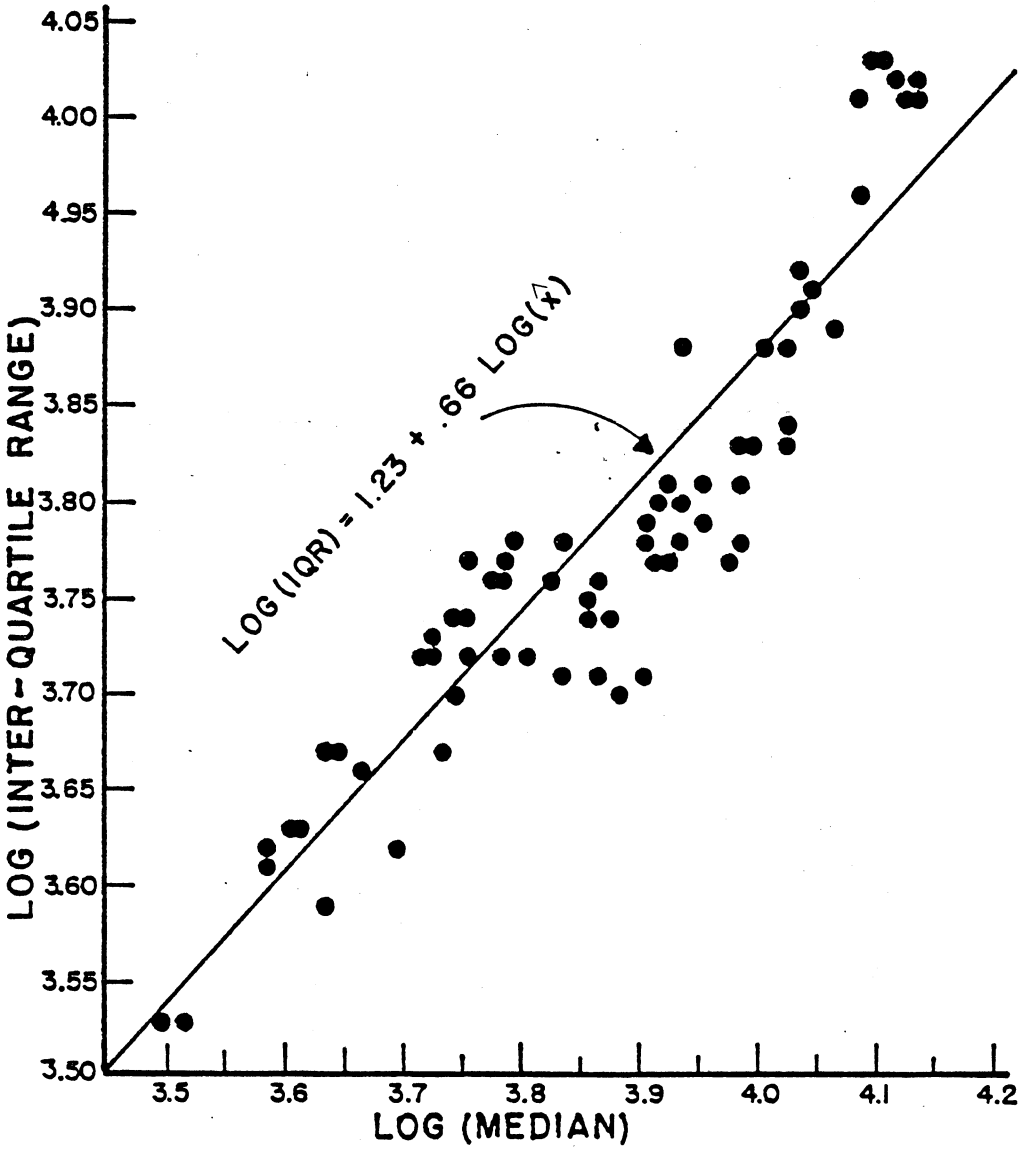

Fig. 2. Plot of the Log-log Relationship Between Medians and Inter-quartile Ranges of Income Distributions (Constant Dollars)

*Table 1. Unstandardized coefficients from regression of ln(IQR) on ln($\hat{x}$) for ten income distributions, for each year**

Year	1963	1964	1966	1967	1968	1969	1970
Regression Coefficient	.589	.642	.576	.598	.656	.702	.668
Correlation	.94	.96	.93	.93	.88	.90	.91

* Average of regression coefficients = .633

Using this method, it is possible to examine the heterogeneity of variances of the seventy CPS income distributions and determine which power transformation will reduce it. Table 1 shows the unstandardized coefficient which results from regressing ln(IQR) on ln($\hat{x}$) for the ten distributions for each year, one equation for each of the seven years. Because each equation includes data from only one year, the results in Table 1 are unaffected by any changes from current to constant dollars. The plot of all seventy *constant* dollar income distributions is shown in Fig. 2. This plot has a slope of .66 (r=.92)[5]. Thus, there is a strong positive relationship between the median and IQR of income distributions, but this relationship is *not* simply multiplicative. Applying Tukey's method, we find that the best power transformation for obtaining homogeneous variances should be around .35, approximately the cube root. This indicates that distributions of the cube root of income (rather than raw income or log income) tend to differ from each other by an additive constant and, therefore, that this constant is probably a better one-parameter description of the difference between two income distributions than either the difference or ratio of medians.

Figures 3, 4, and 5 illustrate several of the above-described features of income distributions. They show six cumulative probability distributions: for blacks and whites in each of three education groups for 1967. The percentage (horizontal) axis is scaled in standard deviations so that the graph will be a straight line *if* the distribution is Gaussian. This type of "probability plot" is discussed in Wilk and Gnanadesikan (1968): if the distribution is in fact Gaussian, then the y-intercept of the line – the expected value of the ordinate at the 50th percentile – is a good estimate of the mean (and median), and its slope is an estimate of the standard deviation. The only difference between the three graphs is in the scale of the income (vertical) axis, which employs a linear, logarithmic, and cube root scale respectively. The distributions in Fig. 3 (4) curve upward (downward) demonstrating that the income (log income) distribution clearly deviates from normality by being skewed to the right (left). When the same income distributions are plotted with a cube root scale, the plots are approximately linear, indicating that the distribution of the cube root of income is approximately Gaussian. Similarly, an exami-

[5] When analyzing distributions from different years, one faces the problem that differences between distributions of *current* income are the result of changes in the cost of living as well as differences in real income. If, as a first approximation, changes in the cost of living reflect a constant percentage change in the costs of all goods and services, then these changes should have a multiplicative effect on the income distribution. Under these circumstances, differences among current income distributions from different years will be more multiplicative than differences between constant income distributions and, therefore, the slope of the log-log plot (analogous to Fig. 2) should be somewhat nearer to 1.00. This is indeed the case; the plot of the seventy current income distributions has a slope of .71. Since this deviates more from the within-year regressions of Table 1, I conclude that income distributions should be converted to constant dollars in order not to confound the differences in real income with the multiplicative effect of inflation.

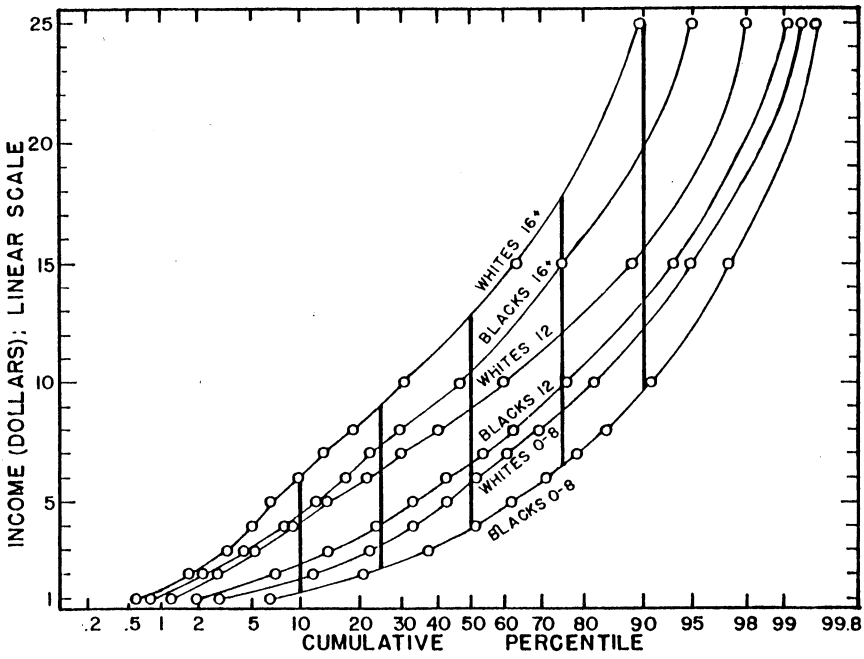

Fig. 3. Probability Plot of 1967 Distributions of Income, by Race and Education

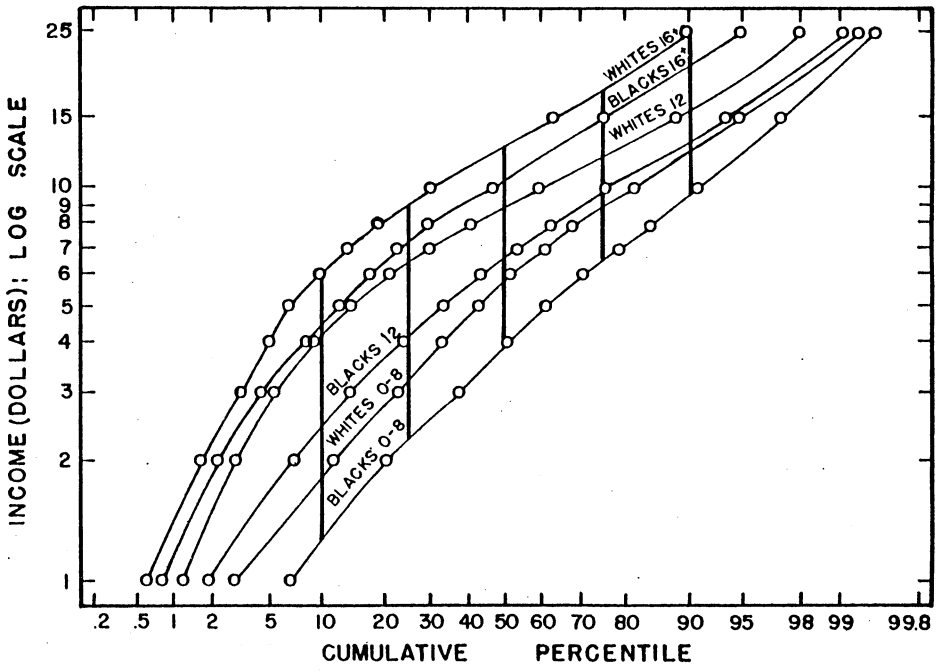

Fig. 4. Probability Plot of 1967 Distributions of ln(income), by Race and Education

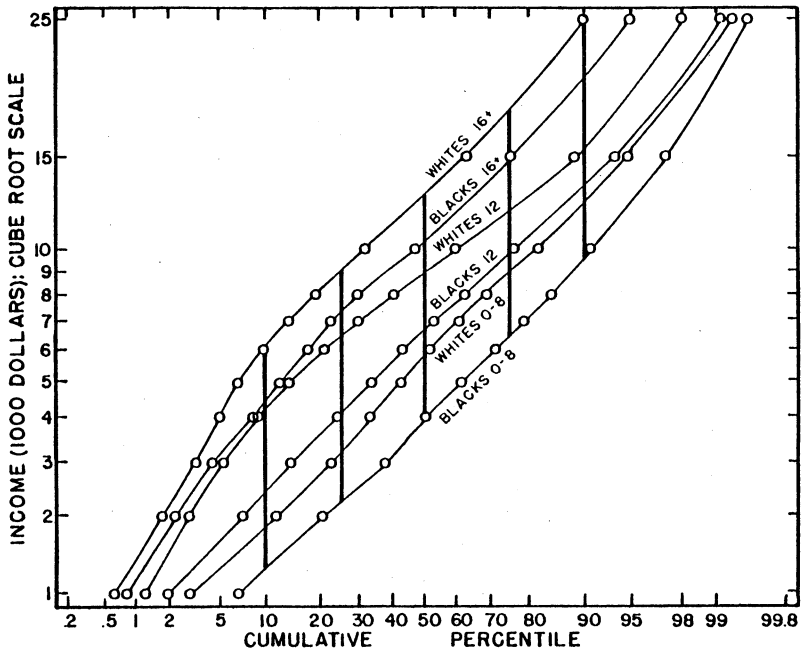

Fig. 5. Probability Plot of 1967 Distributions of Income $^{1}/_{3}$, by Race and Education

nation of the differences between any pair of distributions in Fig. 3 (4), as suggested by the solid vertical lines at the 10th, 25th, 50th, 75th, and 90th percentiles, reveals that the difference between the two medians (50th percentile) generally over- (under-) estimates the difference between the distributions at the 10th and 25th percentiles, while it under- (over-) estimates their differences at the 75th and 90th percentiles. In contrast, Fig. 5 shows that the difference between the medians of two cube root income distributions is generally about the same as (and therefore a good summary of) the difference at other percentiles. This property of the cube root of income follows not only from the fact that the graphs are approximately linear, but more importantly from the result that the slopes of the graphs (estimates of the standard deviation) are approximately equal, a reflection of the general homogeneity of variances.

3.1. Annual Fluctuations in Income: Heterogeneity of Variances, Revisited

The possibility of heterogeneous variances also occurs at the individual level because each individual actually has a distribution of annual incomes. However, since cross-sectional surveys contain only a single observation per respondent, they cannot provide information about individual distributions and the problem of heterogeneous variances remains latent. With panel data, on the other hand, we can study individual distributions and observe the problem of heterogeneous variances.

A brief example will illustrate the issue. Consider a hypothetical doctor with annual incomes over three years of $79 507, $94 196,

and \$66 430 and a carpenter with corresponding incomes of \$31 256, \$18 610, and \$24 389[6]. Their average incomes are \$80 045 and \$24 752 respectively with standard deviations of \$13 887 and \$6 331. How should we compare the incomes of the doctor and the carpenter? Clearly, the doctor earns more, but how much more? To simply conclude that the doctor earns \$55 293 or \$55 118 more than the carpenter (the differences in observed means and medians) is inadequate because the lower portions of the two distributions are much closer together than the higher portions. On the other hand, it is also inaccurate to conclude that the carpenter only earns 30.9, 30.6, or 30.7 percent as much as the doctor (the ratio of means, geometric means, and medians) since the *ratio* between the lower portions of the carpenter's and doctor's distributions is greater than that between the upper portions. The difficulties in comparing the incomes of a doctor and a carpenter are caused by the heterogeneous variances; doctors have a higher income variance, while carpenters have a higher variance of ln(income). In this case, since by design the variances of the cube root of their incomes are equal, the difference in the means of the cube roots of their incomes completely summarizes the difference between the two individual distributions.

Note that due to the relative behavior of different power transformations, this summary implies, mathematically, that the lower portions of the two income distributions will converge while the lower portions of the log distributions will diverge. From one perspective, the cube root summary is simply a more parsimonious (and perhaps precise) way of describing the differences between the two hypothetical distributions.

This discussion can be translated into an analysis of variance framework. Here, individual i's annual income for year t is expressed as: $Y_{it} = Y_i + e_{it}$, where the e_{it} are assumed to be independent and normally distributed $N(0, \sigma_i^2)$. Conceptually, this separates the stable component (Y_i) from the transitory component (e_{it}) of annual income. As the example of the doctor and the carpenter shows, it is difficult to compare different individuals' income distributions if the variances of the transitory component (σ_i^2) are heterogeneous[7]. Heterogeneous σ_i^2 also imply that the "reliability" of annual income as an estimate of the stable component of income varies across individuals or, equivalently, that individuals experience different amounts of uncertainty/risk in the determination of their income.

Comparisons between individuals are facilitated both conceptually and statistically if the σ_i^2 are homogeneous. Most importantly, the amount of observed annual fluctuation, s_i^2, should be independent of the stable component of income, Y_i. The actual relationship between the s_i and Y_i can be examined empirically.

Appropriate data for this purpose are contained in the Panel Study of Income Dynamics conducted by the Institute for Social Research (1972) at the University of Michigan, Ann Arbor, USA. The present analysis was restricted to male, non-student heads-of-households aged 25–65 in 1972 who reported positive annual earnings (measured in constant 1967 dollars) for each of the five previous years.

[6] These numbers are completely artificial except for the general difference between a typical doctor and carpenter. Specifically, they were rigidly created to be 1, 0, and –1 standard deviations from the means of two normal cube root distributions; $N(43, 2.5)$ for the doctor and $N(29, 2.5)$ for the carpenter.

[7] Permanent attributes of the individual cannot, by definition, explain or predict the transitory component of income. However, non-permanent attributes can interact with permanent attributes to affect the transitory component. Though not discussed in this paper, the same issue of heteroscedasticity arises for the recent and more sophisticated structural equation models (e.g., with autocorrelated/lagged transitory components) of income dynamics.

The average and standard deviations of the heads' annual earnings were computed for each of the approximately 2 000 cases satisfying the above criteria.

The method for investigating the homogeneity of individual variances is the same as above except that this time the data are disaggregated into individuals' income distributions rather than into race-by-education-by-year distributions. Analogous to the earlier analysis, the logarithm of the standard deviation of annual fluctuations was regressed on the logarithm of the mean. The resulting equation from the Panel Study is:

$$\ln(s) = .4592 + .6696 \ln(\bar{Y}) \quad (r = .7048),$$

indicating that $\bar{Y}$ and s are far from independent. The correlation of .70, despite being attenuated because the observed $\bar{Y}$ and s are unreliable estimates (since they depend on only five observations per individual) of each individual's underlying distribution of annual income, is quite high for analyses at the individual level. The power transformation of annual income for which an individual's average would not be related to the size of his annual fluctuations about this average is once again the cube root (i.e., 1.0 – .67 = .33). It is interesting to observe that applying this same analysis to the logarithm of income, one observes a substantial *negative* correlation between individuals' mean log income and the standard deviation of the annual fluctuations about the mean log; poorer people have higher proportional fluctuations around the stable component of their income than richer people.

While the focus of the discussion has been on the desirability of separating the stable and transitory components of income, social scientists are also interested in the behavioral, social, and psychological effects of different amounts of fluctuation. In economics this is a problem in uncertainty or risk theory. Those studying this problem presumably want a measure of this fluctuation or risk which is independent of the average level of income. Our analysis suggests that the standard deviation of the transitory component of the cube root of income would be suitable.

The variances of the stable and transitory components of income can be estimated by applying the standard analysis-of-variance model to the cube root of income. The results are summarized in Table 2. The analysis allows for between-year changes in average real cube root income (across all individuals). While the variance attributable to differences between years is significant, it accounts for a

Table 2. Summary table for analysis of variance of the cube root of annual income over five years for male heads in PSID

Source of Variation	D.F.	SS	MS	E(MS)
Persons	1 982	128 970	65.069	$5\sigma_p^2 + \sigma^2$
Years	4	184	46.110	
Persons x Years	7 928	20 226	2.551	σ^2
Years + (P x Y)	7 932	20 411	2.573	(σ^2 assuming no between-year variance)
Persons + (P x Y)	9 910	149 196	15.055	(average within-year variance)
Total	9 914	149 380	15.067	

trivial amount of the total variance. Allowing for differences between years, the persons-by-years mean-square (2.55) is an estimate of the variance of the annual fluctuation about each individual's average cube root income. Using the expectation of the mean-square for persons, the variance of the stable component of cube root earnings is estimated to be:

$$\sigma^2 = (MS_p - MS_{PxY}) / 5 = 12.50.$$

Since the best estimate of the total variance for a given year is 15.06 (the mean-square for persons-within-years), the stable component of cube root income accounts for 83 percent of the total variance in a given year, while the transitory component accounts for the remaining 17 percent. Thus, the maximum percentage of the annual variance that one could possibly "explain" (R^2), using only permanent attributes of the individual, is 83 percent.

Thus far we have tried to justify the decision to transform income data for the purpose of increasing both the normality of the within-group distributions and the homogeneity of their variances at both the individual and group levels. The substantive justification for transforming has been that comparisons of differences among distributions become more meaningful and interpretable. The statistical justification has been that the common multivariate statistical methods, such as analysis of variance and regression, *assume* that the disaggregated distributions of data are normal and have homogeneous variances. (The maximum-likelihood techniques which are being increasingly used in econometrics are especially sensitive to departures from these assumptions.) For income, the cube root transformation satisfies these assumptions considerably better than either the logarithmic or identity transformations.

4. The Functional Form of the Relationship Between Income and Its Determinants

The third statistical reason for transforming one's data is to promote linearity and/or additivity in the relationship between the endogenous and exogenous variables. The most parsimonious models are those in which the endogenous variable is an additive (without interactions), linear function of the exogenous variables. The two statistical criteria for comparing alternative functional forms are their relative parsimony and some measure of their relative goodness-of-fit to the data. Once again, there is evidence that the cube root of income makes a better dependent variable, in this respect, than other power transformations of income. Since the evidence has been published elsewhere, the results will only be summarized here.

Although they did not argue specifically for the cube root transformation, the best evidence appears in Heckman and Polachek (1974). In their article they use maximum-likelihood techniques developed by Box and Cox (1964) and Box and Tidwell (1962) "to determine the empirical functional relationship between earnings and schooling" for three separate sets of data. Unfortunately, they conclude that, "the natural logarithm of earnings is statistically preferable to any other simple dependent variable (p. 350)." They apparently assume that only the log and identity transformations are "simple." In the body of their paper they present four graphs showing how the log-likelihood statistic (a measure of fit) varies for different power transformations (of income) ranging between the identity ($p = 1$) and the reciprocal ($p = -1$). While their graphs do show that using the logarithm of earnings as the dependent variable results in a better fitting model than using untransformed earnings, they also show that

the cube root of earnings yields a better fitting model than the logarithm. (According to their graphs, the power transformations at which the four plots show the maximum goodness-of-fit are .40, .22, .43, and .33. The cube root is a good summary of this range, especially since the graphs are relatively flat around their maxima.) In fact, three of their graphs show that the improvement of the cube root over the log is greater than the improvement of the log over the untransformed earnings variable.

The analyses of Schwartz and Williams (1979) on the functional form of the relationship between earnings, education, and race are not as sophisticated as those of Heckman and Polachek. Schwartz and Williams compare OLS regression equations predicting earnings, the natural logarithm of earnings, and the cube root of earnings for each of three surveys. They also conclude that the cube root of earnings makes the best dependent variable. A large number of additional comparisons among equations predicting these three earnings variables appear in Tables A2.2 through A2.12 of Jencks et al. (1979) and they generally support this conclusion. Schwartz and Williams also discuss the effects of the three transformations on the resulting regression coefficients and the likely impact of these effects on the substantive conclusions that are drawn regarding black/white differences in the returns to education and work experience.

5. The Substantive Implications of Transforming Income: A Brief Example

Previous sections have described reasons why one should want to transform a variable and some methods for selecting an appropriate transformation. Despite several statements to the contrary, the reader may have reached the conclusion that the issue of whether or not to transform a variable is based on narrow statistical criteria and is of little practical or substantive importance. This final section provides a brief example of how the transformation one uses can and does affect the conclusions that will be reached.

Consider an obvious policy-related question: "Did the difference between black and white incomes decrease between 1963 and 1970, during the height of the civil rights movement and the war on poverty?" Table 3 shows the median family incomes (aggregated across education) for blacks and whites for each of these years. The average change per year in the median white income is $595 while the average change for blacks is $474. The absolute difference between black and white median incomes was therefore *increasing* at the average rate of $121 per year. If this is the "right" way to think about income, one must conclude that blacks will never have incomes equal to whites as long as this pattern continues.

Table 3. Median Income of Blacks and Whites, by Year, in Dollars

Year	Blacks	Whites
1963	3 465	6 548
1964	3 839	6 858
1966	4 628	7 722
1967	5 232	8 471
1968	5 684	9 179
1969	6 340	10 089
1970	6 692	10 545

What happens when the data are transformed into logarithms? Over this eight-year period, white median and black median log incomes increased at the rate of .0715 and .0967 log-dollars per year respectively. Translated into ratios, these imply rates of increase of 7.4 and 10.2 percent per year. Now one sees that the difference between black and white median log income was decreasing an average

of .0252 log-dollars, or 2.6 percent, per year. According to this model one would predict that blacks and whites will have the same median income by 1987, a prediction that no longer seems likely to be born out.

When the data are transformed to the cube root, the same type of analysis can be performed. The median cube root of income increased an average of .482 cube-root-dollars per year for whites and .546 cube-root-dollars per year for blacks. The difference between blacks and whites (3.085 cube-root-dollars in 1970) was decreasing at the average rate of .064 cube-root-dollars per year. At this rate, black median income would equal white median income in the year 2017. This simple example illustrates that the transformation one uses to analyze data can substantially affect one's projections as to how quickly (if at all) the black and white income distributions will converge.

6. Summary

Statistical theory tells us that there are three primary reasons for transforming a variable before subjecting it to multivariate analyses: (1) to make the within-group distributions more normal; (2) to reduce the heterogeneity of within-group variances; and (3) to alter the functional form of its relationship with other variables in order to increase additivity, linearity, and/or goodness-of-fit. While the logarithm of income, the form most frequently analyzed by economists, is preferable from each of these perspectives to the untransformed income variable, the present analyses demonstrate that across a fairly broad range of American data covering the period 1963–1975, the optimal power transformation is very close to the cube root. Future work should investigate whether these results can be generalized to other countries and other time periods.

As mentioned in the introduction, there is no logical basis for assuming that a single transformation will optimally satisfy each of the statistical criteria for transforming a variable. However, having found this to be the case, it is important to consider what the substantive significance of the cube root of income might be. Analysis of independent psychometric data (to be reported in another paper) suggests that the cube root of income is linearly related to the American public's general conception of the utility of income, making it theoretically, as well as statistically, a desirable variable.

7. References

Blalock, H. (1982): Conceptualization and Measurement in the Social Sciences. Beverly Hills, CA: Sage.

Box, G. and Cox, D. (1964): An Analysis of Transformations. Journal of the Royal Statistical Society, Series B, 26 (2), pp. 211–252.

Box, G. and Tidwell, P. (1962): Transformation of the Independent Variables. Technometrics, 4 (Nov.), pp. 531–550.

Bronfenbrenner, M. (1971): Income Distribution Theory. London: Aldine.

Gibrat, R. (1957), (1931): On Economic Inequalities, International Economic Papers. Vol. 7, pp. 53–70. This article was first published in 1931 in Les Inégalités Economiques, Paris, Sirey, Chapters V–VII, pp. 62–90. The 1957 version is a translation from French.

Heckman, J. and Polachek, S. (1974): Empirical Evidence on the Functional Form of the Earnings-Schooling Relationship. Journal of the American Statistical Association, (June), pp. 350–354.

Institute for Social Research (1972): A Panel Study of Income Dynamics: Study Design, Procedures, Available Data, 1968–1972 Interviewing Years. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Jencks, C.S., Bartlett, S., Corcoran, M., Crouse, J., Eaglesfield, D., Jackson, G.,

McClelland, K., Mueser, P., Olneck, M., Schwartz, J., Ward, S., and Williams, J. (1979): Who Gets Ahead? A Study of the Determinants of Economic Success in America. New York: Basic Books.

Metcalf, C. (1972): An Econometric Model of the Income Distribution. Chicago: Markham.

Pen, J. (1971): Income Distribution. Great Britain: Penguin.

Schwartz, J. and Williams, J. (1979): The Effects of Race on Earnings. In C.S. Jencks et al., Who Gets Ahead?, Ch. 7.

Tukey, J.W. (1970): Exploratory Data Analysis. Limited preliminary edition, 3 Vols., Addison-Wesley, Reading, Mass.

U.S. Department of Labor, Bureau of Labor Statistics (1972): Handbook of Labor Statistics.

Wilk, M.B. and Gnanadesikan, R. (1968): Probability Plotting Methods for the Analysis of Data. Biometrika, 55 (1), pp. 1–17.

Received June 1984
Revised September 1984