The Effects of Power Transformations on Consumer Expenditure Survey Data

Taylor J. Wilson

July 20, 2018

Consumer Expenditure Surveys Program Report Series

U.S. BUREAU OF LABOR STATISTICS • bls.gov | @BLS_gov

I. Introduction

The Consumer Expenditure Surveys (CE) are the only Federal surveys that cover the complete household profile of expenditures, income, and demographics. The CE data are a powerful analytical tool utilized by many researchers, academics, and policy makers. However, due to the complex nature of the data, there are nuances that are often overlooked when performing analysis using them. For example, the surveys produce continuous expenditure and income series that have a variety of data distributions. Both expenditure and income data usually have a fixed domain beginning at zero and can be extremely high.[1] This implies that it is unlikely for outliers to exist in the left tail of the data distribution, and as a result, expenditure and income data will typically be right skewed. Asymmetric or skewed distributions can be addressed by transforming the data. Implementing a power transformation on an expenditure or income variable may result in a data distribution that is easier to handle in a regression framework (e.g. normalizing residuals, reducing heteroskedasticity). Additionally, the distribution may aid in satisfying some of the underlying assumptions of parametric statistical tests that require normality as a basic assumption. Specifically, with CE data, it is important to understand how these transformations affect the data and what type of transformation should be implemented to achieve the desired effects. This paper examines the CE summary variables for income and total expenditures, and the effects of both logarithmic and, more generally, power transformations on the distributions of these variables. Additionally, this paper explores some of the practical applications of power transformations in economic analyses using CE data.

II. An Overview of Power Transformations

Power transformations are useful tools for altering a set of data to a desired distributional shape. They have been explored in length in the literature by Box and Cox (1964), Andrews (1971), Atkinson (1973), Hinkley (1975), and Taylor (1985). These authors note that is possible to apply a transformation to every positive data point in a data set by some constant parameter, λ, such that skewness is as close to zero as possible. Azzalini (1985) calls this a "shape parameter." Because skewness is defined as a measure of asymmetry in a distribution, skewness of zero is an important feature of data that are normally distributed. After performing this transformation, the data will be distributed in such a way

[1] With the notable exception of the healthcare category in which negative expenditures are common and represent the reimbursement of expenditures. Regarding income, certain sources, such as rental and self-employment income, can take negative values when losses occur.

that the normality assumption is satisfied.[2] These types of transformations are generalized by Box and Cox (1964) in the following way:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \ln(y) & (\lambda = 0) \end{cases} \tag{1}$$

Notably, for very small values of λ, both cases of equation (1) are approximately equal. That is[3],

$$\lim_{\lambda \to 0} \frac{y^{\lambda} - 1}{\lambda} = \ln(y) \tag{2}$$

When applying these transformations to a dataset, it is worth noting that in a linear context, case one simplifies nicely. Equation (3) below is the equation for a line or the systematic non-random component of a regression model.

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x} \tag{3}$$

Substituting case one from equation (1) yields the following,

$$\frac{y^{\lambda} - 1}{\lambda} = \beta_0 + \beta_1 \mathrm{x} \tag{4}$$

Performing the necessary algebra, equation (4) becomes the following,

$$y^{\lambda} = (\lambda\beta_0 + 1) + \lambda\beta_1 \mathrm{x} = \widehat{\beta_0} + \widehat{\beta_1} x \tag{5}$$

Because lambda is a constant parameter it can be absorbed into new parameters represented in equation (5). It is these new absorbed coefficients that will be estimated in an OLS regression model when the dependent variable is transformed by raising every observation to the selected shape parameter.

---

[2] Conversely, it should be noted that it is possible to select a parameter to alter the shape to something non-normal, should that be of interest to the researcher.

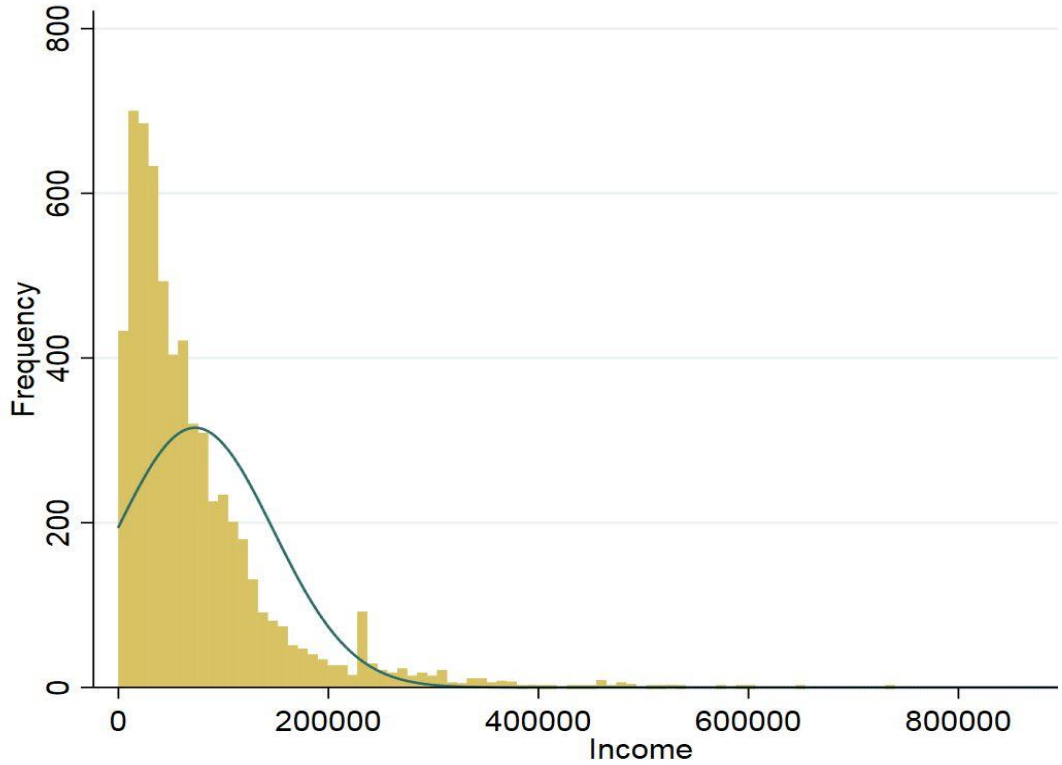
[3] Directly applying the limit to this function produces an indeterminate form. Applying L'Hopital's Rule, the first case of equation (1) takes the form $\mathrm{f(y)} = y^{\lambda}\ln(\mathrm{y})$; the limit of which as λ approaches 0 is $\ln(y)$.

III. Selecting Lambda

Determining whether a transformation is necessary usually occurs by examining the data distributions at the beginning of a project. Examining a frequency histogram of the variables of interest is a reasonable place to start. In the CE, income is one example of a variable that exhibits right-skewness. Given that the transformations often take the form of a logarithm or a root, it is an implicit requirement that the values of y need to be strictly positive to perform the transformation. This is easy to accomplish with most expenditure and income data, which are rarely negative save for some special cases like health insurance reimbursements or income losses. However, when losses occur, the affected variables require alternative methods.

As noted, income data sometimes do contain some negative values that reflect the impact of business losses. However, these are usually a very small percentage of the distribution—0.08 percent of the values reported for all consumer units in the Interview Survey in 2017 quarter 1. For clarity of presentation, these are ignored for computing the results of this paper. Figure I below shows the income frequency distribution of the remaining Interview reports from the first quarter of 2017.

Figure I. Distribution of Income – Interview Survey, First Quarter 2017



Source: Consumer Expenditure Survey, Interview Public Use Microdata 2016

The skew of this distribution can be visualized from the above histogram. The sample skewness can also be mathematically represented by S in equation (6) below.[4]

$$S = \frac{n\sqrt{n-1}}{n-2} \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{3/2}} \tag{6}$$

Applying equation (6) to the above distribution reveals a skewness estimate of approximately 2.62. The higher the value of skewness the more right asymmetric the distribution will appear in the histogram. For a sufficiently large sample size, the difference between $(n-1)$ and $(n-2)$ is small so it is common to substitute n for these values leading to a simplified constant ($\frac{n\sqrt{n}}{n} = \sqrt{n}$). After making this substitution, the statistic becomes the 'population skewness' as opposed to the 'sample skewness'.

It is the goal of these power transformations to select a lambda where S is minimized. Tukey (1977) suggests that an easy way to explore the data and determine an approximately optimized transformation is to plot the skew against the chosen lambda parameter for a given data distribution. It should be noted that this is largely a pedagogical exercise because methods exist, using statistical software, to find the optimal shape parameter. Although, it may be useful to do this anyway to better understand the data and to check that the statistical program is producing the desired result. Figure II and the associated table shows the results of applying common transformations, from which one can determine those that achieve a reasonably normal distribution.[5,6]

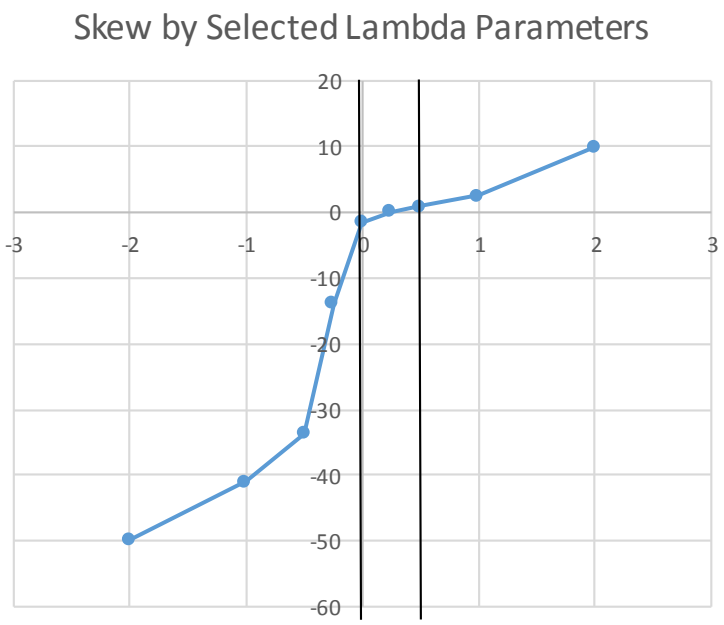
[4] The formula presented here is how most statistical software compute this statistic. See, http://support.sas.com/documentation/cdl/en/procstat/66703/HTML/default/viewer.htm#procstat_univariate_details03.htm and https://support.office.com/en-us/article/skew-function-bdf49d86-b1ef-4804-a046-28eaea69c9fa The derivation of this formula, which details the link between formula (6) and the formula presented in the SAS documentation, is found here, http://www.macroption.com/skewness-formula/ For more information about skewness see Doane and Seward (2011).

[5] Common transformations meaning powers that imply a named functional form like square roots, quarter roots, reciprocals, squares, etc. This is sometimes referred to as the Tukey ladder of powers.

[6] In this case of the CE income variable, depending on whether the full specification from equation (1) or simply the left-hand side of equation (5), the plot will either resemble a cubic spline or a parabola respectively. Figure II uses the full specification.

Figure II. Table and Graph of Skewness of Income Distribution for Selected Lambda Parameters

Lambda Parameter (λ) – X Axis	Skewness of the Income Dist. – Y Axis
-2	-49.810
-1	-40.899
-0.5	-33.628
-0.25	-13.768
0 (LN Transform)	-1.428
0.25	0.096
0.5	0.930
1	2.620
2	9.966



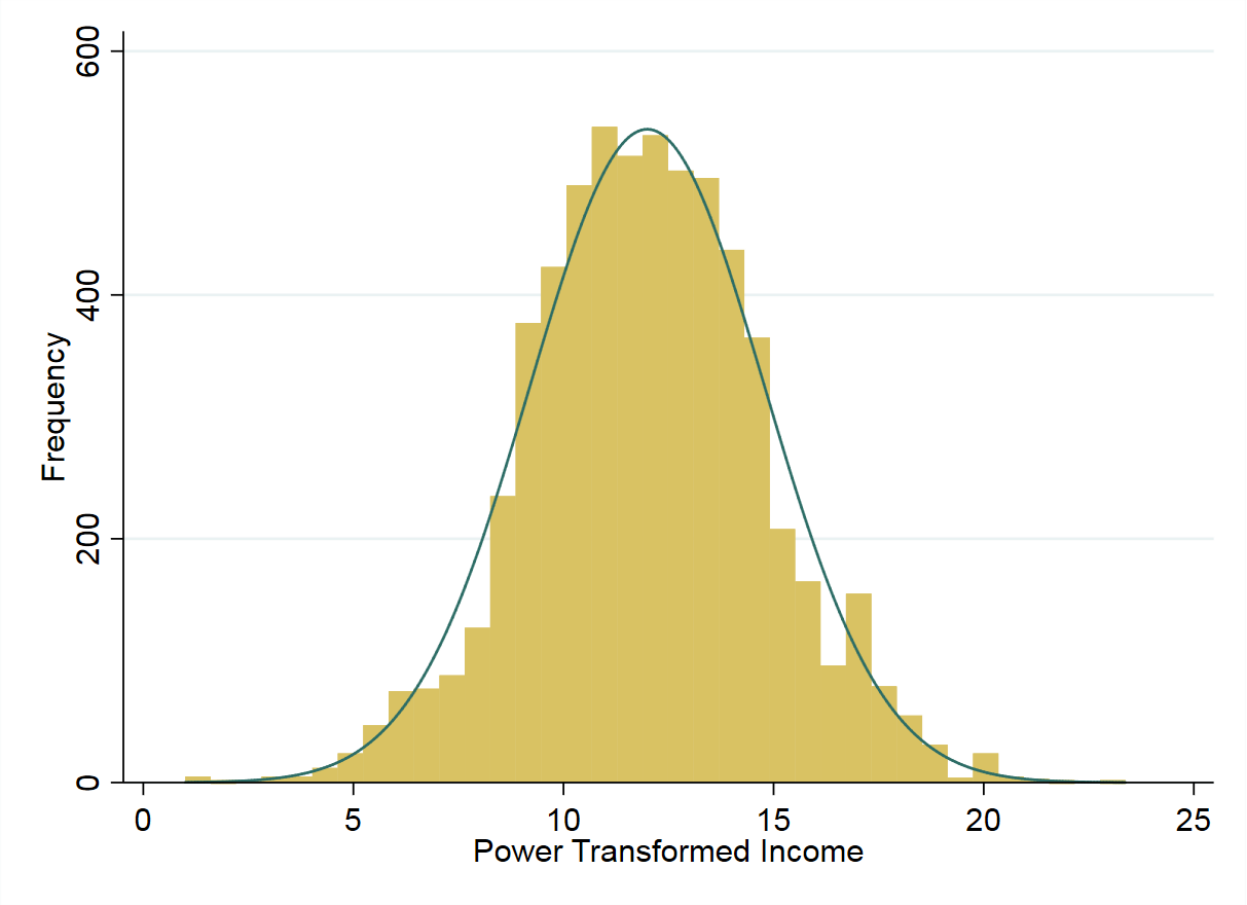
Source: Consumer Expenditure Survey, Interview Public Use Microdata 2016

This method suggests that a quarter root transformation is best applied to the CE income variable achieves the best approximation to a normal distribution. The lambda parameter can be applied continuously as opposed to the above method which selects cutoff points that have function names like "square root" or "quarter root." Because lambda is a continuous parameter, there must exist a lambda somewhere between 0.25 and the log transform for which the skewness is zero. In order to find this optimal transformation, the most common application in statistical programming is to compute maximum likelihood estimates of the shape parameters.[7,8] Doing so shows that the parameter that produces the zero skew result is approximately 0.229.[9] Considering that expenditure and income data are not perfectly smooth, it is not possible to guarantee perfect normality. However this parameter will create a distribution as statistically close to normal as possible whereby deviations in the parameter, within a few significant digits, will always move the skewness away from zero. If the data are less coarse or if more significant digits are used in the transformation, the skew can move closer to zero. Eventually,

[7] The SAS programming language manual details some of the process here, support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_transreg_sect015.htm
[8] The Stata programming language manual has more of the details here, https://www.stata.com/manuals13/rboxcox.pdf
[9] Computed with Stata.

the benefit of adding more significant digits will be outweighed by the computation time. Figure III below shows the optimally transformed income distribution with three significant digits.

Figure III. Distribution of Income under an Optimal Parameter - Interview Survey, First Quarter 2017



Source: Consumer Expenditure Survey, Interview Public Use Microdata 2016

IV. Interpreting Transformed Data

The value of a normal distribution comes from its impact on the analyses that are done with this newly generated normal variable. Many statistical tests are robust to the normality assumption.[10] In other cases, the assumption of normality is not about the individual variables themselves but rather that the sampling distribution of the means is normally distributed. The real value of performing data transformations, particularly with CE data, is correcting heteroskedastic relationships that emerge as a result of the underlying data distributions. One common type of transformation used in analyses is the

[10] See Khan and Rayner (2003)

transformation by the natural logarithm.[11] A log transformation is convenient because in a regression context it allows for a direct interpretation of estimated coefficients. Consider equation (7), a log-transformed systemic non-random component of a regression model.

$$\ln(\mathrm{y}) = \beta_0 + \beta_1 \ln(\mathrm{x}) \tag{7}$$

By differentiating this expression with respect to x, the result is expressed in equation (8).

$$\frac{\mathrm{dy}}{\mathrm{dx}}\frac{1}{y} = \beta_1 \frac{1}{x} \tag{8}$$

The resulting expression, implies that the estimated coefficient β_1 is equivalent to a percent change in x with respect to a percent change in y. That is,

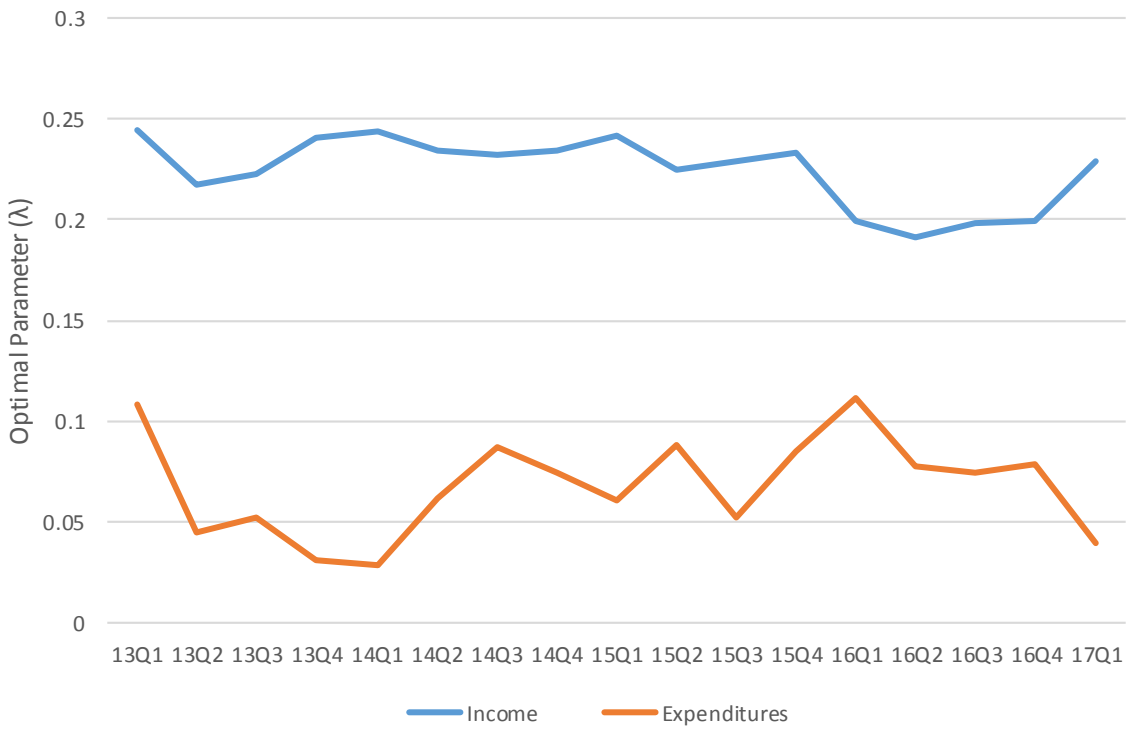
$$\beta_1 = \frac{\mathrm{dy}}{\mathrm{dx}}\frac{x}{y} \tag{9}$$

This is convenient for many analyses, especially those interested in computing elasticities, a common use for CE data. While convenient, it may lead to an erroneous outcome if the untransformed equation takes optimal parameters on y and x that are not the natural logarithm.

Those variables that become normal upon transformation by the natural logarithm are called log-normal. However, it is unlikely that income variables will be log-normal because of their typically large right skew. Therefore, a log transformation is not the most efficient transformation for these variables. Nevertheless, some variables are closer to a log-normal distribution than others. The closer a variable is to being log-normal the less biased a transformation by the natural logarithm will be when applied as transformation for convenience. Therefore, it is useful to know how close the variable under consideration is to a log-normal distribution. Battistin, et al. (2009) show that expenditure data are closer to a log-normal distribution than income from a variety of sources including the CE. This result still holds given the most recent consumer expenditure data. Figure IV below shows the optimal transformation parameters for total expenditures and income by quarter. The closer the value of the optimal parameter is to zero, the more log-normal the distribution will be in that given year and quarter.

[11] Azzalini and Dalla Valle (1996) discuss log transformations as the device which often 'cures non-normality.'

Figure IV. Optimal Transformation Parameter for Income and Expenditures by Quarter, 2013-2017



Source: Consumer Expenditure Survey, Interview Public Use Microdata 2013-2016

Empirically, the reason income consistently requires a larger shape parameter for transformation to a normal distribution is due to the presence of relatively larger outliers in the data. This typically causes the skew for income distributions to be greater than for expenditure distributions. The effects of performing a log transformation for convenience will thus have less impact on expenditure variables than on income variables. There is a tradeoff between optimizing the transformation and preserving the interpretability of the output. Since log transformations allow beta coefficients from a regression model to be interpreted easily, it may be worth sacrificing some of the optimality for the ease of interpretation. The differences between the estimated coefficients for lambdas sufficiently close to zero and a natural log are likely to be negligible, depending on the variable examined. This is ultimately the decision of the researcher. For example, in figure IV, there is more variance observed in the expenditure parameters than the income parameters, but both are relatively stable, typically not varying by more than 0.05 between any two quarters. In the context of the ladder of powers, this difference would be unlikely to change the selected common functional transformation.

There is an additional layer of complexity with respect to interpretability that is introduced as a result of using a transformation that is not a log. Computing elasticities using log transforms is straightforward via the procedure detailed in equations (7) through (9). Consider equation (10) below which uses a different transformation parameter for both y and x. In this case, β_1 is estimated for the researcher by a statistical program, and the lambdas on y and x are 0.5 and 0.33, respectively.

$$\sqrt{y} = \beta_0 + \beta_1 \sqrt[3]{x} \tag{10}$$

Assuming the end goal is still to compute the x elasticity of y, then a few additional steps are required to achieve this using the optimal transformations in place of the natural logarithm.

$$\frac{\mathrm{dy}}{\mathrm{dx}} \frac{1}{2\sqrt{y}} = \beta_1 \frac{1}{3\sqrt[3]{x^2}} \tag{11}$$

Rearranging equation (11) to get elasticity represented on the left hand side in terms of β_1.

$$\frac{\mathrm{dy}}{\mathrm{dx}} \frac{x}{y} = \beta_1 \left(\frac{\sqrt{y}}{\sqrt[3]{x^2}} \right) \left(\frac{\frac{1}{3}}{\frac{1}{2}} \right) \left(\frac{x}{y} \right) \tag{12}$$

This equation can be generalized for any power combination of powers, m and n, on x and y respectively,

$$\frac{\mathrm{dy}}{\mathrm{dx}} \frac{x}{y} = \beta_1 \left(\frac{x^{m-1}}{y^{n-1}} \right) \left(\frac{m}{n} \right) \left(\frac{x}{y} \right) \tag{13}$$

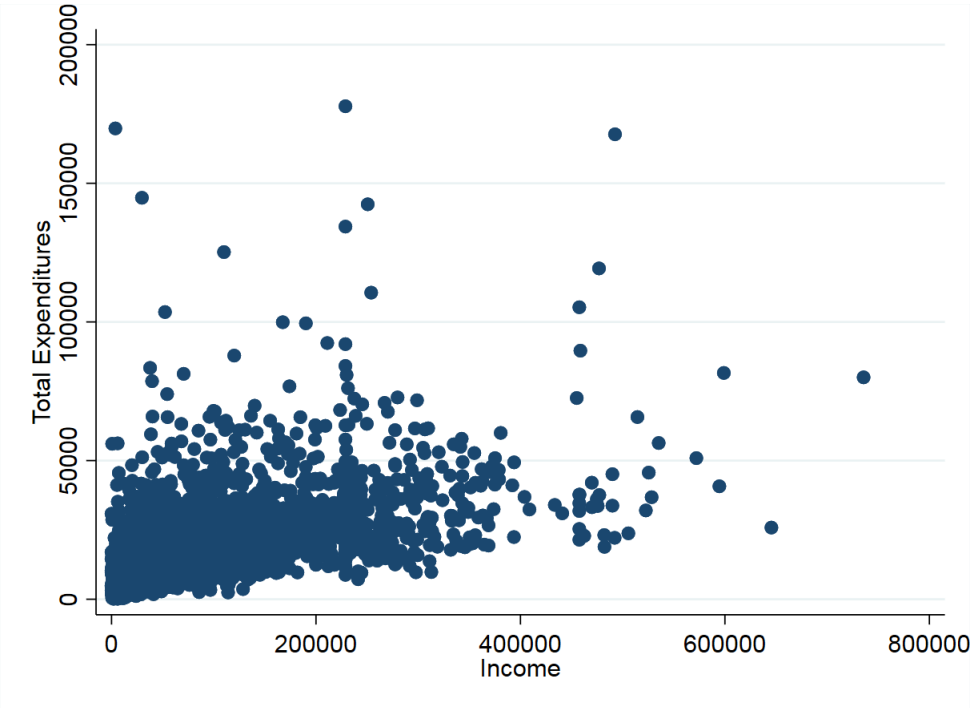
From equation (13), it is clear that the elasticity can be derived from the estimated β_1 by multiplying it by the correction terms in parentheses. Elasticities are evaluated at a particular point, so for most relationships, the predicted value of y at the average value of x is appropriate. However, if interested in the elasticity at other points, equation (13) can take any combination of values.

V. Example from Consumer Expenditure Data

In this section, it is shown how the estimated coefficients are altered in a regression context depending on what transformations are selected. The processes detailed in the previous section are

explored here in a single year, bivariate context for demonstration purposes with CE income and expenditure variables. The primary objective is to demonstrate how the coefficients change, depending on which transformation is used. First, the relationship is explored with no transformation; then a log transformation is introduced, and finally the optimal transformation is used to describe the relationship. To fully research the income-expenditure relationship, it is important to introduce appropriate control variables and methods for working with imputed data.[12] The value of implementing a transformation in the first place is often to address heteroskedastic relationships. That is, one of the assumptions of the linear model is that the conditional mean of the unobserved errors is zero. A heteroskedastic relationship will prevent the researcher from reasonably making this assumption. For these types of relationships, Figure V shows the characteristic 'fan shape' of the heteroskedastic bivariate relationship between income and expenditure.

Figure V. Untransformed Income-Expenditure Relationship, 2017 Quarter 1[13]



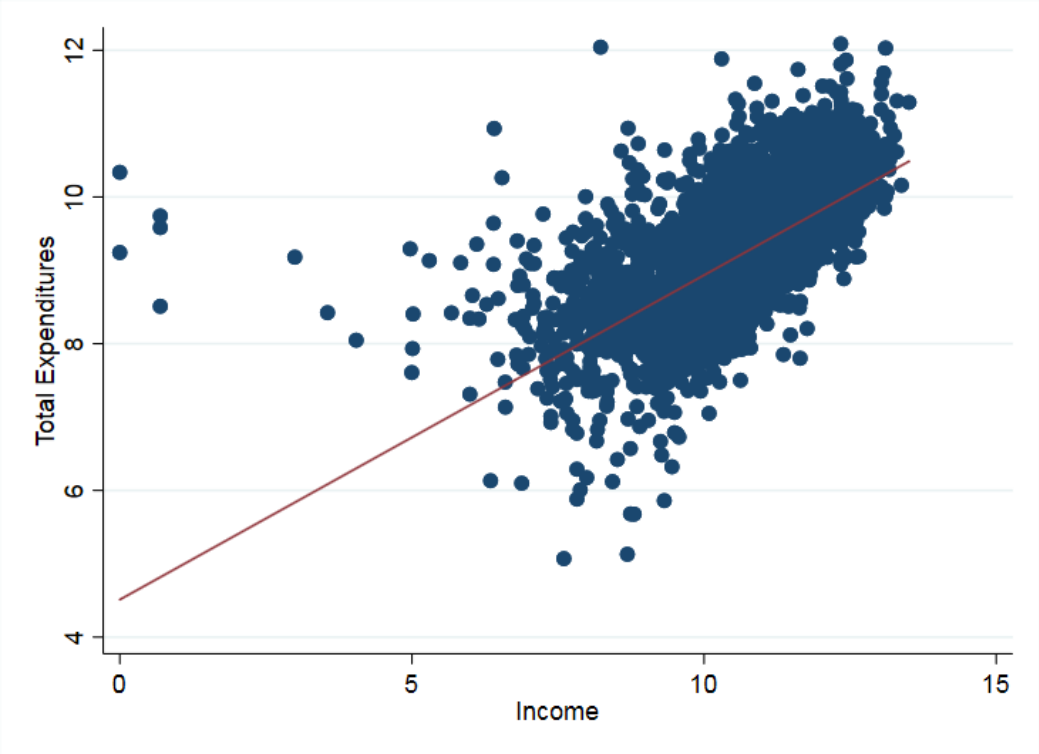
Source: Consumer Expenditure Survey, Interview Public Use Microdata 2016

[12] FINCBTXM (imputed income before taxes) is used as the measure for income. To produce true income elasticity equations, control variables should be introduced and the multiply imputed versions of the variable (i.e., FINCBTX1, FINCBTX2, FINCBTX3, FINCBTX4, and FINCBTX5) should be analyzed separately in order to accurately estimate the coefficient for meaningful economic interpretation. For more information, see the User's Guide to Income Imputation in the CE found here, https://www.bls.gov/cex/csxguide.pdf

[13] One quarter is used here for demonstration purposes so that the relationship can be reasonably visualized on a graph. (n=6,199)

After performing a log transformation to both income and expenditures, the resulting relationship between the variables improves. In a regression context, the bivariate coefficient on income can be directly interpreted as the income elasticity of expenditures. However, using the non-optimal transformation can still preserve some of the outliers in the relationship. The outliers biasing the regression line are low-income, high-expenditure households, which the log transformation does a poor job of addressing. Figure VI below shows the scatter plot of the log-transformed variables.

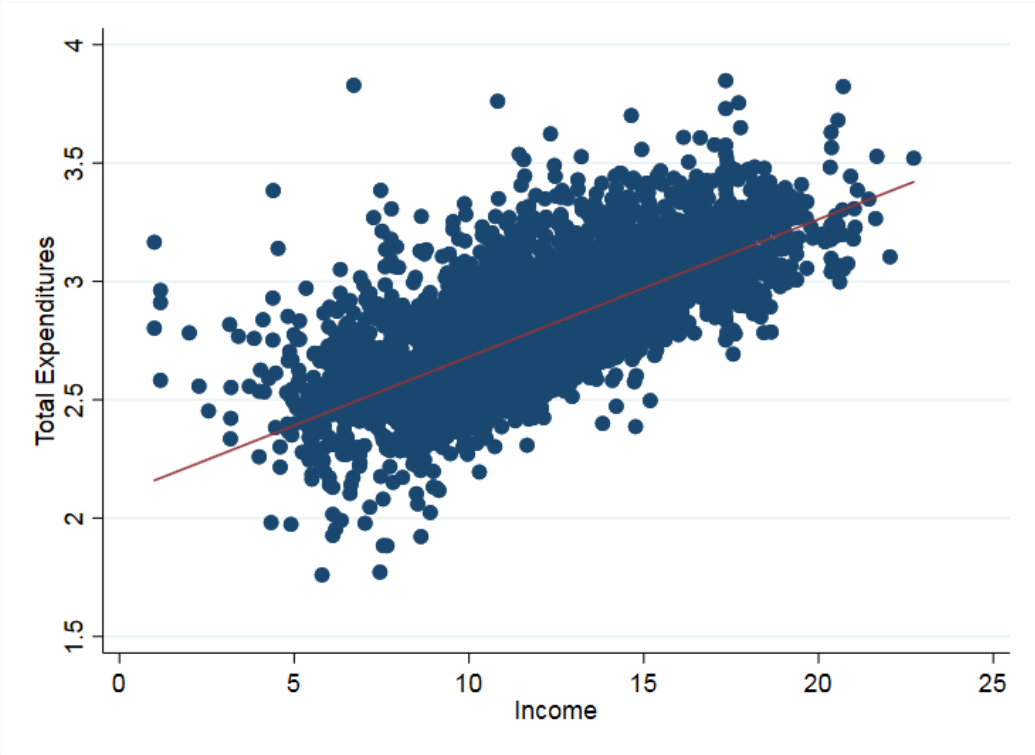
Figure VI. Log Transformed Income-Expenditure Relationship, 2017 Quarter 1



Source: Consumer Expenditure Survey, Interview Public Use Microdata 2016; Red line indicates the predicted log of total expenditures as a function of the log of income.

The resulting elasticity computed from the log relationship is 0.44—a one percent change in income leads to a 0.44 percent change in total expenditures. Though the relationship is less heteroskedastic, the line is still biased by outliers in the left tail of the distribution. Perhaps the convenience of the log relationship is worth sacrificing for a more accurate estimate of the income elasticity. Consider the optimally transformed version of this relationship, where expenditures are raised to the power of 0.11 and income is raised to the power of 0.23. Figure VII shows that the data are even more homoscedastic and less biased by outliers than both the untransformed and the log transformation versions.

Figure VII. Optimally Transformed Income-Expenditure Relationship, 2017 Quarter 1



Source: Consumer Expenditure Survey, Interview Public Use Microdata 2016; Red line indicates the predicted transformed total expenditures as a function of transformed income.

Applying formula (13) on the average values of income and total expenditures, the resulting elasticity computed from the optimally transformed relationship is 0.56—for a consumer unit with average income and predicted expenditures, a one percent change in income leads to an estimated 0.56 percent change in total expenditures. The optimally transformed relationship produces a higher elasticity because of how the outliers on the left hand side of the relationship are treated. By mitigating the effect of the outliers through the transformation, as opposed to simply dropping them, we maintain more statistical power and preserve the sample size. Given that the outliers are no longer 'flattening' the regression line, the optimally transformed relationship shows a more statistically representative relationship of the core sample.

VI. Conclusion

Overall, the use of transformations can allow for a more accurate interpretation of relationships between CE variables. Given the skew of income and expenditure distributions, power transformations

provide a quick and efficient way to accurately assess relationships between these variables. It should be noted that power transformations are not the only way to correct the problems described in this paper. Other types of transformations may do a better job of correction, while preserving other attributes about which a researcher may be concerned (e.g. negative values in health care expenditures and business losses in income). However, for linearizing data and achieving constant variance across the domain, the power transformation technique addresses those concerns and reliably produces estimates that better describe the underlying relationships, specifically compared to the logarithmic case (unless the log happens to be the optimal transformation).

VII. Works Cited

1. Andrews, D. F. "A Note on the Selection of Data Transformations." *Biometrika*, vol. 58, no. 2, 1971, pp. 249–254. *JSTOR*, JSTOR, www.jstor.org/stable/2334514.
2. Atkinson, A. C. "Testing Transformations to Normality." *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 35, no. 3, 1973, pp. 473–479. *JSTOR*, JSTOR, www.jstor.org/stable/2985112.
3. Azzalini, A. "A Class of Distributions Which Includes the Normal Ones." *Scandinavian Journal of Statistics*, vol. 12, no. 2, 1985, pp. 171–178. *JSTOR*, JSTOR, www.jstor.org/stable/4615982.
4. Azzalini, A. and A. Dalla Valle "The Multivariate Skew-Normal Distribution." *Biometrika*, vol. 83, no. 4, 1996, pp. 715–126. *JSTOR*, JSTOR, https://www.jstor.org/stable/2337278.
5. Battistin, Erich, et al. "Why Is Consumption More Log Normal than Income? Gibrat's Law Revisited." *Journal of Political Economy*, vol. 117, no. 6, 2009, pp. 1140 1154. *JSTOR*, JSTOR, www.jstor.org/stable/10.1086/648995.
6. Box, G. E. P., and D. R. Cox. "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, no. 2, 1964, pp. 211 252. *JSTOR*, JSTOR, www.jstor.org/stable/2984418.
7. Doane, David and Lori Seward. "Measuring Skewness: A Forgotten Statistic?" *Journal of Statistics Education,* vol. 19, no. 2, 2011 http://ww2.amstat.org/publications/jse/v19n2/doane.pdf
8. Hinkley, David V. "On Power Transformations to Symmetry." *Biometrika*, vol. 62, no. 1, 1975, pp. 101–111. *JSTOR*, JSTOR, www.jstor.org/stable/2334491.

9. Khan, Azmeri, and Glen D. Rayner. “Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem.” *International Scholarly Research Notices*, Hindawi, 2003, www.hindawi.com/journals/ads/2003/657201/abs/.

10. Taylor, Jeremy M. G. “Power Transformations to Symmetry.” *Biometrika*, vol. 72, no. 1, 1985, pp. 145–152. *JSTOR*, JSTOR, www.jstor.org/stable/2336344.

11. Tukey, J. W. (1977). Exploratory Data Analysis. Reading, Massachusetts: Addison Wesley.