Statistical Addendum to:

"A Concave Log-Like Transformation Allowing Non-Positive Values"

Economics Letters, forthcoming

Martin Ravallion[1]

Department of Economics, Georgetown University
Washington DC., 20057, U.S.A.
mr1185@georgetown.edu

This addendum provides examples illustrating points raised in the main paper.

Artificial data: Figure 1 plots four transformations applied to numbers in the interval (-2, 8). (These values would be typical of the low-middle incomes in rich countries in units of $1,000 per person per year, or incomes in developing countries in units of $'s per person per day.) We see the expected convexity of the IHS transformation for negative values, and that the curvature is reversed using the h-transformation (as shown in the main text). It is also evident from Figure 1 that $h(y)$ approaches $\ln y$ quite quickly in the region $y > 0$. Figure 1 also gives the started log with c set at the minimum value (-2) plus 0.1% of the mean. (The graph looks very similar if one uses, say, 1% of the mean.) This deviates much more from $\ln y$ in the range of the data.

The h-transformation in Figure 1 is for $\alpha = 1$. Figure 2 also gives the transformation in equation (2) of the paper for $\alpha = 0.5$ and $\alpha = 1.5$. We see that $h(y;\alpha)$ approaches $\ln y$ quickly for these values of α. For $\alpha = 1.5$, the hyperbolic sine transformation ($\sinh(y) \equiv 0.5(e^{y} - e^{-y})$) yields a deep curvature, with large negative values at the extremes.

As an intuitive summary statistic for inequality, Figure 2 also gives:

$$H(\alpha) \equiv h(\bar{y};\alpha) - \frac{1}{n}\sum_{i=1}^{n} h(y_i;\alpha) \text{ (where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{)} \qquad \text{(A1)}$$

This will be recognized as a modified MLD in which $\ln y$ is replaced by $h(y;\alpha)$. So it is of interest to compare $H(\alpha)$ to MLD, which is 0.28 for these data. (Using the started log it is 0.35.)

[1] For useful comments or discussions on this paper the author thanks Francois Bourguignon, Christopher Camenares, Dan Cao, Denis Cogneau, Shaohua Chen, Garance Genicot and Franco Peracchi.

We see that $H(\alpha)$ is substantially higher. This suggests that how one deals with negative net wealth or income can make a big difference to measures of inequality using log or log-like transformations. Values of $\alpha < 1$ are probably a sensible choice if one is measuring inequality using the h-transformation.

For negative values, none of these transformations except $g(y)$ provide satisfactory approximations to the proportionate changes, $\Delta y/|y_{-1}|$, namely the discrete version of $dy/|y|$. This is clear from Figure 4 which plots to changes from one observation to the next. Indeed, all except the IHS indicate decreasing differentials, while the proportionate changes are increasing for $y < 0$. The changes in $h(y;0.5)$ might be considered to provide a reasonable approximation at the lower end but then deviate substantially from the proportionate changes in the interval (-1, 1). Of course, it is expected that the discrete change in $g(y)$ is virtually indistinguishable from the proportionate changes given that $dg(y) = dy/|y|$.

Income data: This example uses real data on incomes, namely the European Union's SILC survey for Belgium in 2008. There are only 21 non-positive values in a survey with 6,000 households. The density function for the raw data on income per person in Figure 4 shows that there are long tails in both directions.

Figure 5 gives $h(y;\alpha)$ $(\alpha = 0.5,1,1.5)$ for these data. Three points are notable. First, unlike the first example, the values of $H(\alpha)$ in Figure 5 are lower than MLD. Even though there is a small proportion of negative values, the IHS transformation is less concave in the (0,1) interval and (as noted) it is convex in the negative region. This difference reflects the high density of data in the (0,1) interval, where the log transformation becomes highly concave (bending down sharply toward negative infinity).

Second, for values of $\alpha \geq 1$, the hyperbolic sine transformation yields very large negative numbers at the negative extremes; intuitively, higher values of α make the function bend down more sharply in the region $y < 0$. This makes the inequality index (and variance) explosive unless α is set at a sufficiently low value. (This feature is shared by the ordinary log transformation with small values of *y* less than unity.) For example, for the raw data, $H(1) = 98.84$, but on dropping the two lowest (most negative) incomes it drops to 0.074! The

measure is more stable for $\alpha = 0.5$. Re-scaling and/or trimming the data or setting a lower value of α appears to be justified.

Third, for these data, the started log (with $c = y^{\min} + \varepsilon$ where ε is 0.1% of the mean) becomes quite un-log-like. While the function remains concave, it is close to linear, so the inequality index falls to nearly zero. (If one trims the data of the two largest negative values then one obtains an inequality measure of 0.012; trimming the three largest it rises to 0.030.) These data illustrate that the use of the started log to deal with non-positive values can greatly attenuate measured inequality by implicitly reducing the function's built-in inequality aversion. Clearly this is deceptive.

Net-wealth data: Suppose one is interested in the relationship between the wealth distributions of different age groups. The household survey in use is not longitudinal and did not ask about the wealth of non-resident parents or children. Instead, one estimates the elasticity of the wealth quantiles for young families with respect to those of middle-aged families.

Let $y_j(p)$ denote the wealth quantile function (the inverse of the distribution function) for age group *j*; in other words, $y_j(p)$ is the net wealth level of the *p*'th household ranked by wealth. For this illustrative purpose, the quantile was calculated for 99 ("percentile") values of *p* for various age groups from the 2013 Federal Reserve's Survey of Consumer Finances (SCF).[2] Figure 6 plots the resulting quantile functions for two groups according to the age of the household head, namely 18-24 years (interpretable as the "Millennials" once they have left home) and 40-44 (roughly their parents). Using the latest available data at the time of writing, one finds that 34% of households headed by someone 18-24 years of age in 2013 had negative net wealth; in the age group of 40-44 (say) it is 14%. So if one estimates the elasticity using the log transformation, the regression is heavily censored with the risk of a large bias.

Using the 65 observations with positive values for $\ln y_{24}(p_i)$ (in obvious notation), the OLS regression coefficient of $\ln y_{24}(p_i)$ on $\ln y_{44}(p_i)$ is 1.220 (s.e.=0.133).[3] This can be compared with a regression of $h[y_{24}(p_i)]$ on $h[y_{44}(p_i)]$.[4] Using only positive observations

[2] I used the convenient DQYDJ net worth calculator based on the SCF. Top coding does not allow an estimate of the top percentile; this is another source of bias, but not the one of interest here.
[3] Only White standard errors are reported in this paper.
[4] On these data the hyperbolic sine function produced numeric overflows at the extremes of both tails of the distribution of net wealth in $'s. To avoid this, net wealth was scaled to units of $10,000.

(equivalent to using the IHS), the regression coefficient is 0.715 (s.e.=0.025). If one regresses $h[y_{24}(p_i)]$ on $h[y_{44}(p_i)]$ using all observations then the coefficient falls markedly to 0.230 (s.e.=0.011). The p_1 value is a clear outlier; if one excludes that observation then the regression coefficient is 0.486 (s.e.=0.081). So it is clear that switching to the h-transformation and including all real values in the data yields an appreciably lower estimate of this elasticity.

These calculations are for $\alpha = 1$. As noted, α can also be treated as an estimable parameter; in this case the regression specification is:[5]

$$h[y_{24}(p_i), \alpha_{24}] = \pi_0 + \pi_1 h[y_{44}(p_i), \alpha_{44}] + \varepsilon_i \quad (i=1,\ldots,98) \qquad \text{(A2)}$$

The properties of the $h(.)$ transformation (including continuity in α) suggest that Nonlinear Least Squares (NLS) is an appropriate estimator for the parameters of (2). [6] The NLS estimates are $\hat{\alpha}_{24} = 0.037$ (s.e.=0.010), $\hat{\alpha}_{44} = 0.024$ (s.e.=0.005), $\hat{\pi}_0 = -0.606$ (s.e.=0.282) and $\hat{\pi}_1 = 0.328$ (s.e.=0.062).[7] The estimates of α are both quite low (i.e., the implied transformation has a low slope at zero), though still significantly different from zero. They are also quite similar, suggesting a stable parameter value across age groups.[8] The estimated elasticity of 0.33 is higher than when α is set to unity but it remains appreciably lower than is found for the censored regression using logs.

[5] The Eviews code for this regression specification is: (log((c(1)*y24 + (((c(1)*y24)^2) + 1)^0.5)))*(1-d24) + (0.5*(exp(c(1)*y24)) - 1/(exp(c(1)*y24)))*d24 - log(2*c(1)) = c(2) +c(3)*(log((c(4)*y44 + (((c(4)*y44)^2) +1)^0.5)))*(1-d44) + (0.5*(exp(c(4)*y44))-1/(exp(c(4)*y44)))*d44-log(2*c(4))

[6] On the assumptions for consistency of the NLS estimator see Wooldridge (2002, Chapter 12).

[7] On again dropping the first observation one obtains $\hat{\pi}_1 = 0.290$ (0.071).

[8] One cannot reject the null that $\hat{\alpha}_{24} = \hat{\alpha}_{44}$ (t=1.37). On imposing this restriction one obtains $\hat{\pi}_1 = 0.254$ (0.025).

Figure 1: Alternative transformations of a hypothetical income or net wealth distribution

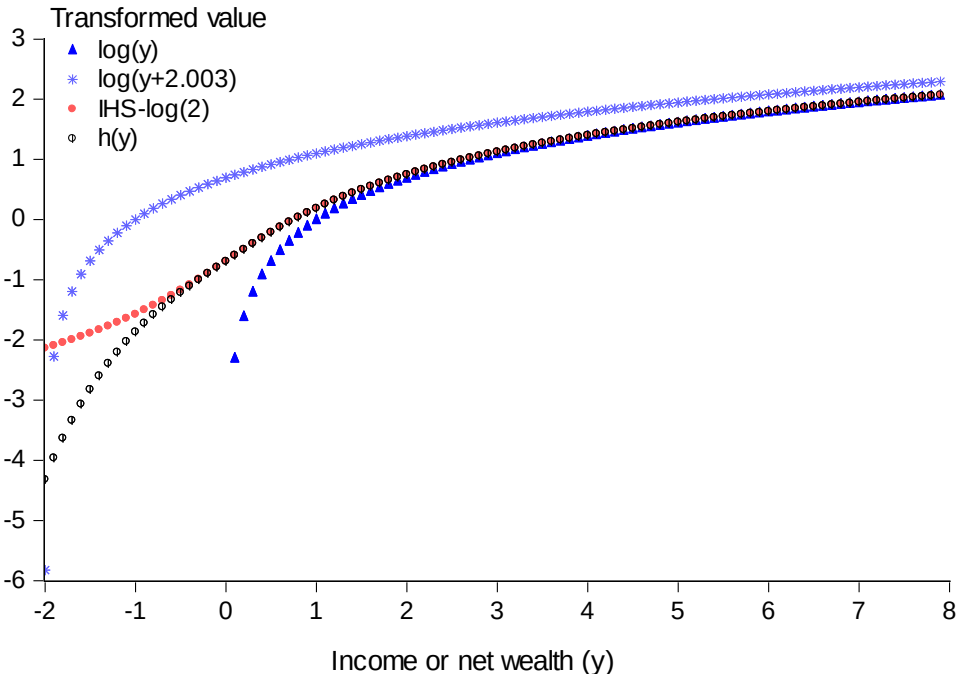


Figure 2: h-transformations for different parameter values

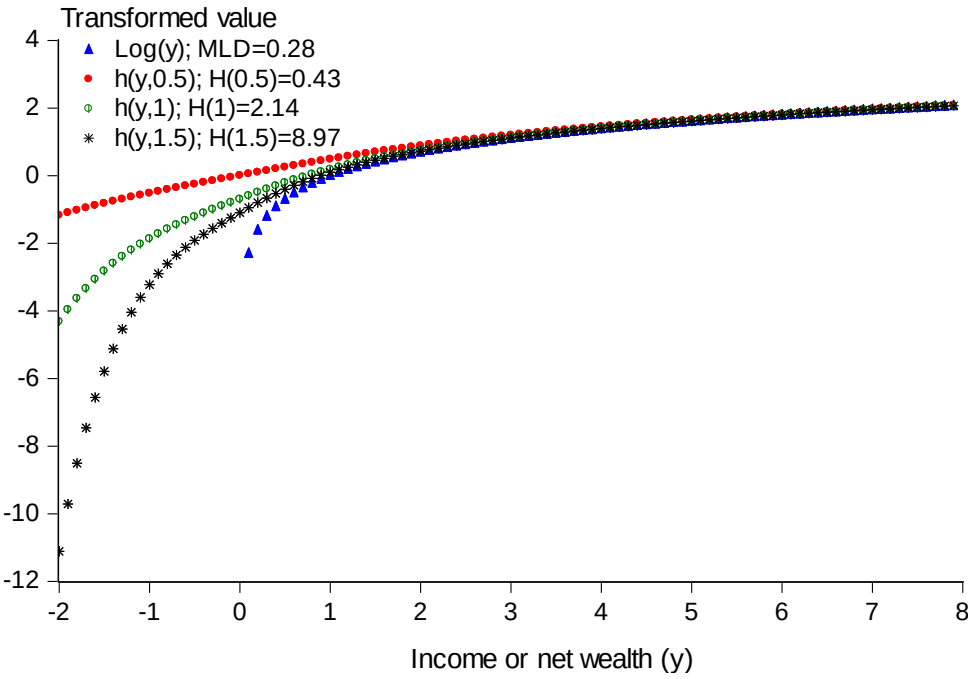
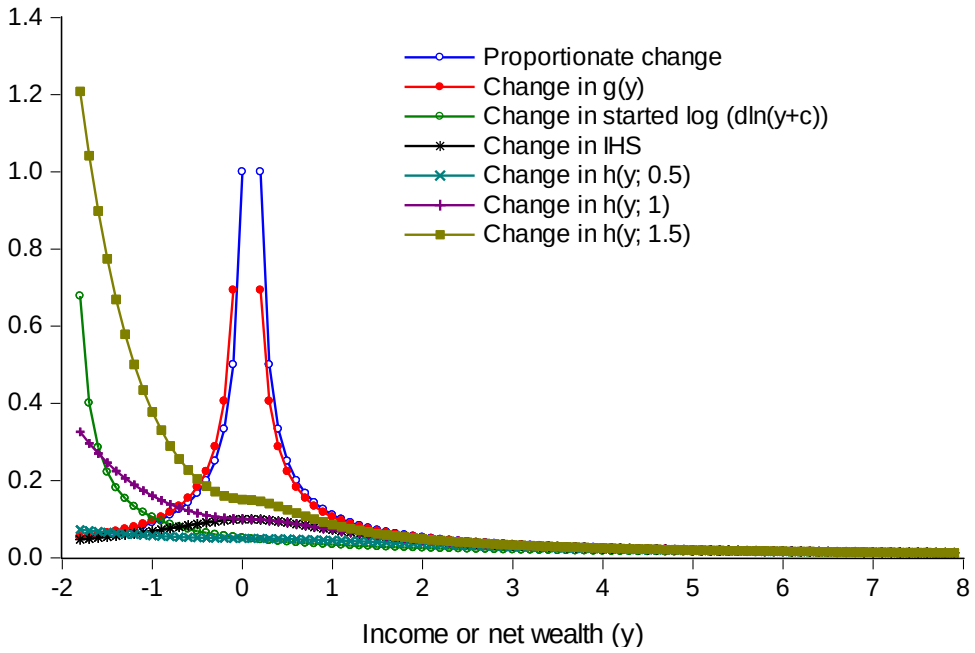


Figure 3: Changes in various transformation from one observation to the next



Note: The proportionate change is defined as $dy/|y|$ for all $y \neq 0$.

Figure 4: Kernel density function for household income per person in Belgium 2008

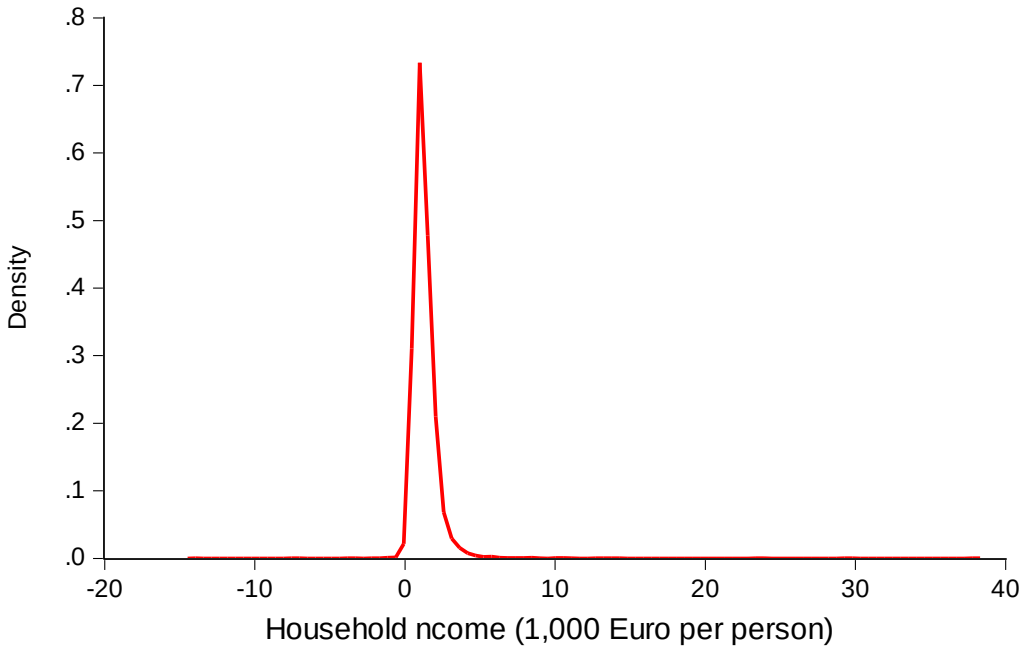
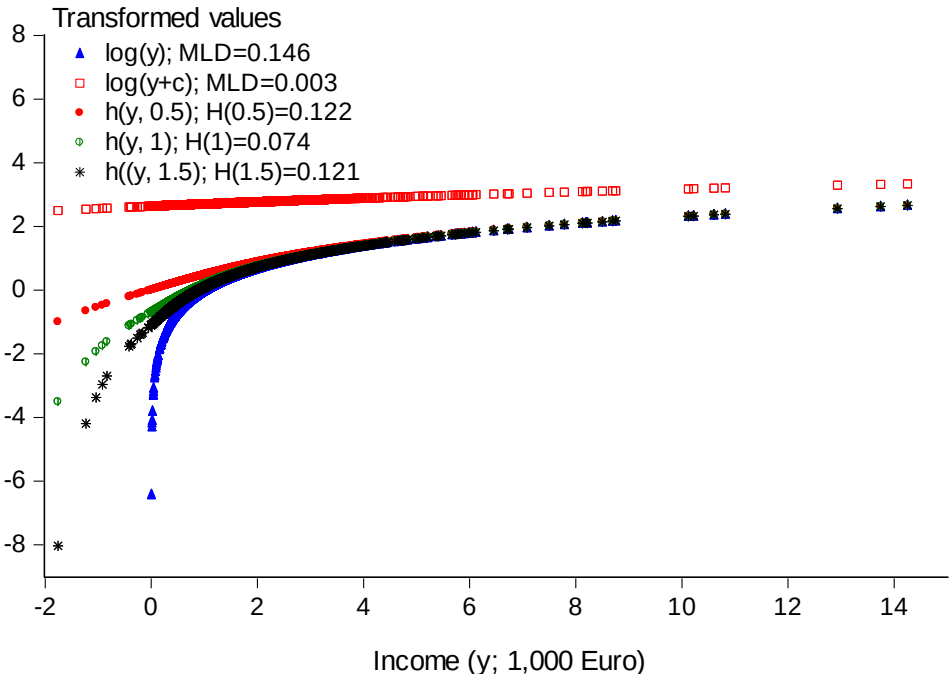


Figure 5: Transformed values of the Belgian household income distribution



Note: Data from the European Union Statistics on Income and Living Conditions (EU-SILC). Household income per person in 1,000 Euros. The data are trimmed at the top and bottom to make the figure more readable (though this only affected 0.2% of the 6026 data points). The values of MLD and H(0.5) are based on the full data set. The values of H(1) and H(1.5) are based on n=6024, after dropping the two largest negative incomes. On also doing so for MLD one obtains 0.149 using normal logs and 0.012 using started log.

Figure 6: Quantile functions for net wealth in the U.S. 2013