*Building a Lorenz Curve from Grouped Income Data*

Figure 1 illustrates the steps though which Lorenz interpolation approximates a Lorenz curve. The vertical dashed lines represent cumulative population shares, which can be computed from the grouped income data. The income boundaries associated with these lines are used to determine several slopes along the Lorenz curve, which are represented by the red line segments.

[Insert Figure 1 About Here]

Lorenz interpolation approximates a Lorenz curve by estimating several cubic functions in a sequence. These functions are constrained to pass through points on the Lorenz curve and to have slopes at x-coordinates that are determined by the income boundaries. The left plot of Figure 1 shows the first function estimated by this method. This function is constrained to pass through (0, 0), the point where the Lorenz curve begins. The slope of the function is constrained to equal at , at , and at , where is the lower bound of the first bin, and are the upper bounds of the first and second bins, and is the cumulative population share plotted on the x axis. Once this function, , is estimated, the y-coordinate of the point on the Lorenz curve associated with the lower bound of the second bin is estimated as . Then, the method estimates a function for the second bin that is constrained to go through (, ) and to have slopes at , at , and at (see the center plot in Figure 1).[[1]](#endnote-1) These steps are repeated for the remaining bins except for the top bin.[[2]](#endnote-2)

The absence of an upper bound makes defining a cubic function for the top bin of the grouped income data particularly challenging. Cubic functions fit to points along the Lorenz curve potentially underestimate the variance of incomes at the top of the income distribution (Kakwani 1976:489). To remedy this issue, Lorenz interpolation uses a slope constraint to reduce the upward trajectory of the cubic function applied to this bin. First, a quadratic function is defined for the top bin. The slope of this function is then incrementally reduced at the bin midpoint until the function is no longer convex, at which point the method selects the last cubic function that does not violate this convexity check. This results in a gradually increasing cubic function that captures that variance at the upper tail of the income distribution. [[3]](#endnote-3)

Having estimated a Lorenz curve, the final step is to create a sample of exact incomes based on this curve. A computationally efficient way to approximate a sample from the Lorenz curve is to plot equidistant points along the estimated curve and multiply the slopes of the line segments connecting these points by the income distribution mean. This generates samples from the underlying income distribution, which can be weighted using the frequencies provided in the grouped income data. This weighted sample is plugged into an income inequality formula to produce an income inequality estimate.

1. The coefficients of each cubic function defined by these constraints can be calculated by solving the following system of equations.

   = (3 )

   Where , , , and are the coefficients of the cubic function applied to each bin and is the function for the approximated Lorenz curve of the preceding bin. The first row of the system ensures that the function passes through the x and y-coordinates of the bin’s lower bound. The next three rows represent the slope constraints: the expressions in the left matrix show the first derivative of the cubic function with the function input set to the bin lower bound, the bin upper bound, and the upper bound of the following bin. [↑](#endnote-ref-1)
2. The cubic function defined for the final closed bin uses three of the four constraints used for the other closed bins. In place of a third slope constraint the cubic function is constrained to pass through (1, 1), the point associated with the upper bound of the top bin. [↑](#endnote-ref-2)
3. This convexity check is also applied to the other segments of the cubic spline function that Lorenz interpolation estimates using the closed bins of grouped income data. Specifically, any cubic function that has a negative second derivative at either of its boundaries is adjusted so that it is as close to the original cubic function as possible without violating this convexity rule. [↑](#endnote-ref-3)