There is a meaningful coding error
In Deaton and Lubotsky (2003)

Michael Ash* Dean Robinson†

14 October 2013

In “On weights and coding errors: odd coincidence or dress rehearsal?” an Econbrowser blog post on October 9, 2013, Angus Deaton critiques our 2009 paper, “Inequality, race, and mortality in U.S. cities: A political and econometric review of Deaton and Lubotsky.” Contradicting a claim in our paper, Deaton’s post states “there was no coding error” in his co-authored “Mortality, inequality and race in American cities and states” (Deaton and Lubotsky, 2003). Deaton and Lubotsky (DL) also previously denied this coding error in their 2009 response to our review (Deaton and Lubotsky, 2009). DL’s 2003 paper, our review, and DL’s response were all published in Social Science & Medicine after peer review.

Deaton is wrong. There was a coding error in the regression weighting in DL 2003. The DL results depend in a meaningful way on the error and the choice of weighting. The hypothesis under consideration is that inequality is bad for health, examined using the correlation between inequality and mortality across U.S. cities. DL found a positive association between inequality and mortality controlling for income. However, DL found that the association disappears when the racial composition of cities is additionally taken into account. Our review examined both DL’s econometric analysis and DL’s interpretation of the relationships between inequality, racial composition, and mortality. As our paper reported, the estimated relationship between mortality and inequality taking into account racial composition is highly sensitive to the relative importance of small and large cities—the issue of regression weighting. DL’s published results, which lead DL to reject the inequality-mortality hypothesis, reflect a coding error in implementing regression weights. Weights corrected to the specification in DL’s text produce a substantially different result that is consistent with the inequality-mortality hypothesis. Our review identified the error, explored sensitivity to weighting, and recommended further research.

This note describes the coding error in DL 2003 and summarizes our econometric replication. The issues Deaton raises regarding Ash’s working paper with Herndon and Pollin

*Department of Economics and Center for Public Policy and Administration, University of Massachusetts Amherst, mash@econs.umass.edu.

†Department of Political Science and Center for Public Policy and Administration, University of Massachusetts Amherst. Robinson was listed as author on 17 October 2013 and several pronouns were changed to reflect the change. We thank Robert Pollin for valuable comments.
replicating Reinhart and Rogoff “Growth in a Time of Debt,” are, obviously, an entirely separate matter.

In describing their regression weighting, DL wrote, “Each regression is weighted by the square root of the population at risk in each state or MSA” (p. 1142) and the table notes for each regression table stated, “All regressions are weighted by the square root of the relevant population.” “Weighted by the square root of population” means that each observation or row is multiplied through by the square root of that observation’s population.\footnote{Weighting a regression by the square root of population is a standard approach when the data for a regression are estimated averages for the units. In DL, cities are the units of observation, and the variables are averages at the city level: the average mortality rate, the average income, the inequality level, and the percent black. The logic of square-root weighting is that the variance of sampling error for the outcome variable is proportional to the size (population) of the unit, or \( E(u_j^2) \equiv \sigma_j^2 \propto 1/n_j \).} A regression weighted by the square root of population is expressed as:

\[
y_j\sqrt{n_j} = \beta_0\sqrt{n_j} + \beta_1x_{1j}\sqrt{n_j} + \beta_2x_{2j}\sqrt{n_j} + u_j\sqrt{n_j}
\]

where \( j \) indexes observations, \( y \) and the \( x_k \)’s are variables, \( u \) is an error term, the \( \beta \)'s are regression coefficients, and \( n_j \) is the population of observation \( j \).

Following standard professional practice, DL shared all code and data for the study with us. The relevant code appears in the Stata script file \texttt{table2.do} (provided by DL). DL define the weight variable on line 7:

\texttt{gen weight = sqrt(popul);}

and use it in all subsequent regressions including the key regression specification on line 29:

\texttt{regress aa_mrate mnlypq ginipq_a pct_b [aw=weight] if racesex==1;}

The Stata code provided by DL produces exactly the results published in DL (2003), but the DL code does not weight the regressions by the square root of population as DL claim. The Stata manual Technical Note on weighted regression in the section “regress–linear regression” states:

A popular request on the help line is to describe the effect of specifying \texttt{[aweight=exp]} with \texttt{regress} in terms of transformation of the dependent and independent variables. The mechanical answer is that typing

\texttt{. regress y x1 x2 [aweight=n]}

is equivalent to fitting the model

\[
y_j\sqrt{n_j} = \beta_0\sqrt{n_j} + \beta_1x_{1j}\sqrt{n_j} + \beta_2x_{2j}\sqrt{n_j} + u_j\sqrt{n_j}
\]

(Stata 11, p. 1530)
That is, in Stata, weighting by the square root of population requires specifying population (not the square root of population) as the weight term in the `regress` command. The correct code to implement the weighting specified by DL is `gen weight = popul`, rather than the term used by DL, `gen weight = sqrt(popul)`.

Because DL’s code first took square root and then included the already square-rooted term in the `regress` command, DL’s code did not weight by the square root of population but rather by the square root of the square root, or the fourth root. Because of this coding error, DL’s code actually estimated,

\[ y_j \sqrt{\sqrt{n_j}} = \beta_0 \sqrt{\sqrt{n_j}} + \beta_1 x_{1j} \sqrt{\sqrt{n_j}} + \beta_2 x_{2j} \sqrt{\sqrt{n_j}} + u_j \sqrt{\sqrt{n_j}}. \]

DL’s results depended in a meaningful way on the choice and coding of weights. Both DL’s original paper and our replication explored alternative weighting. DL wrote:

We have also repeated Table 2 without using population weights. This is not our preferred strategy—the MSAs are of very different size, and the regressions should be thought of as on an individual basis not an MSA basis—but is nevertheless a useful specification text. (1146)

Table 3 of our paper shows the key results using a full range of four alternative weights: (1) unweighted, (2) the fourth-root weights that DL actually used, (3) the square-root weights that DL intended, and (4) full population weights. As we reported, “The results differ substantially for the coefficient on the key variable based on the weighting scheme applied.” Some of these weights substantially and significantly decrease support for the inequality-mortality hypothesis with a \( t \)-statistic as low as \(-1.92\), and some of the weights substantially and significantly increase support for the inequality-mortality hypothesis with a \( t \)-statistic as high as \(+8.57\). If we focus exclusively on DL’s coding error, DL’s incorrect fourth-root weights generate a \( t \)-statistic of \(-0.6\); DL’s intended square-root weights generate a \( t \)-statistic of \(+1.64\). Our paper reports and discusses the full range of results. Our discussion of which weights are correct was detailed and appropriately provisional. We conclude the discussion of the DL coding error in regression weighting:

We take the extraordinary sensitivity of the gini-mortality results to the weighting system as an indication of fundamental misspecification or likely heterogeneity...Sensitivity to population weighting...suggests directions for further empirical research, but it is hard to consider the case closed.

In conclusion, DL 2003 does contain a coding error in the regression weighting. Correcting the error and varying the regression weights changes the basic results of DL with respect to inequality and mortality in a meaningful way. Our paper identified a coding error and demonstrated the sensitivity of results to reasonable alternative weightings, and we presented and discussed all of the results. We continue to encourage further research and replication on the inequality-mortality hypothesis.
References

