2020

36th IARIW General Conference

Paper Prepared for the 36th IARIW General Conference, Oslo, Norway, August 24-28, 2020

Regression with an Imputed Dependent Variable

Thomas Crossley

Peter Levell

Stavros Poupakis

In empirical research we are often interested in the relationship between two variables, but no available data set contains both variables. For example, a key question in fiscal policy and macroeconomics is the effect of income or wealth (or changes in income or wealth) on consumption. Traditionally, consumption has been measured in dedicated household budget surveys that contain limited information on income or wealth. Income or wealth, and particularly changes in income and wealth, are measured in panel surveys with limited information on consumption.

A common strategy to overcome such problems is to use proxies for the dependent variable that are common to both surveys and impute that dependent variable into the data set containing the independent variable. In the first stage the dependent variable is regressed on the proxies in the donor data set. In the second stage, the coefficients, and possibly residuals, from the donor data set are combined with observations on the proxies in the main data set to generate an imputed value of the missing dependent variable in the main data set. Hereafter we refer to this as the RP procedure (for ``regression prediction"). The addition of residuals to the regression prediction seeks to give the imputed variable a stochastic component and mimic the dispersion of the missing variable, and we refer to this as the RP+ procedure. For example, in a well-known paper, Skinner (1987) proposed using the U.S Consumer Expenditure Survey (CE) and the RP procedure to impute a consumption measure into the Panel Study of Income Dynamics (PSID).

In this paper we consider the consequences of estimating a regression with an imputed dependent variable, and how those consequences depend on the imputation procedure adopted.

We show that the RP procedure leads to an inconsistent estimate of the regression coefficient of interest, as does the RP+ procedure. We show that under reasonable assumptions the asymptotic attenuation factor is equal to the population R2 on the first stage regression of the variable to be imputed on the proxy or proxies. This leads us to suggest a ``rescaled-regression-prediction" (hereafter RPP) procedure. We then show that with a single proxy, the RP procedure is numerically identical to a procedure developed by Blundell et al (2004, 2008) (hereafter BPP

after the authors), also for imputing consumption, in which the first stage involves, in contrast to RP, regressing the proxy on the variable to be imputed, and then inverting.

The issue we point to is much more general than our motivating application. In particular, widely used ``hot deck" imputation procedures are in many cases equivalent to the RP+ procedure. An important implication of our analysis for data providers is that the preferred method of imputation may depend on the intended application. While the RPP and BPP procedures allow for consistent estimation of a regression coefficient, they are less attractive if the object of interest is the (unconditional) variance of the missing variable.

We complicate our theoretical analysis with a Monte Carlo study and two applications to data from the CEX and PSID.