Received: 28 July 2020 | Revised: 14 January 2022 | Accepted: 17 July 2022

DOI: 10.1002/jae.2921

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

Regression with an imputed dependent variable

Thomas F. Crossley[1] | **Peter Levell**[2] | **Stavros Poupakis**[3]

[1]European University Institute, Institute for Fiscal Studies and ESCoE, Italy

[2]Institute for Fiscal Studies, UK

[3]University College London, UK

Correspondence
Peter Levell, Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, UK.
Email: peter_l@ifs.org.uk

Summary

Researchers are often interested in the relationship between two variables, with no single data set containing both. A common strategy is to use proxies for the dependent variable that are common to two surveys to impute the dependent variable into the data set containing the independent variable. We show that commonly employed regression or matching-based imputation procedures lead to inconsistent estimates. We offer a consistent and easily implemented two-step estimator, "rescaled regression prediction." We derive the correct asymptotic standard errors for this estimator and demonstrate its relationship to alternative approaches. We illustrate with empirical examples using data from the US Consumer Expenditure Survey (CE) and the Panel Study of Income Dynamics (PSID).

KEYWORDS

consumption, imputation, measurement error

1 | INTRODUCTION

Empirical researchers 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 which contain limited information on income or wealth. Income or wealth, and particularly changes in income and wealth, is 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 (or variables). 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 "regression prediction."

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 prediction error, or Berkson measurement error, that the regression prediction procedure introduces into the dependent variable leads to inconsistent estimates of the regression coefficients of interest. While classical measurement errors, which are orthogonal to the true dependent variable, do not cause bias, Berkson measurement errors, which are orthogonal to the imputed dependent variable, do cause an attenuation bias. We then show that under mild assumptions, the asymptotic attenuation factor is equal to one minus the population R^2 on the first stage regression of the variable to be imputed on the proxy or proxies. This leads us to propose a "rescaled-regression-prediction" (hereafter **RRP**) estimator. We demonstrate that this

easily implemented procedure provides consistent estimation of the second stage regression parameters, and those parameters are identified even when the set of explanatory variables is larger than the set of proxies. We then derive the correct standard errors for our two-step estimator.

When multiple proxies are available for a single dependent variable, we demonstrate an overidentification test is possible and that data combination by full nonlinear GMM (**nlGMM**, Arellano & Meghir, 1992) may be asymptotically more efficient than **RRP**. However, Monte Carlo experiments show that **RRP** frequently dominates **nlGMM** in finite samples.

We further show that, in the case a single proxy variable is used, the **RRP** estimator is numerically identical to the estimator from a procedure developed by Blundell et al. (2004, 2008) (hereafter **BPP**), also for imputing consumption, in which the first stage involves, in contrast to regression prediction, regressing the proxy on the variable to be imputed, and then inverting. However, the procedure proposed by **BPP** does not naturally extend to the case of multiple proxies. Moreover, we show that the usual OLS standard errors from a regression of an imputed dependent variable, derived from the **BPP** procedure, are too small (a point that the literature also seems to have overlooked). The equivalence of **RRP** and the **BPP** procedure means that the correct asymptotic standards errors for the two-step **RRP** estimator can be used for the **BPP** procedure as well.

Both regression prediction and the **BPP** procedure are currently used to impute dependent variables. However, the choice of imputation methods used in different papers appears to be ad hoc. In a well-known paper, Skinner (1987) proposed using the US Consumer Expenditure Survey (CE) and a regression prediction procedure to impute a consumption measure into the Panel Study of Income Dynamics (PSID). **BPP** does not give explicit reasons for favouring their imputation method over regression prediction, and on occasion, the same authors have switched from using the **BPP** procedure to regression prediction in later papers. Examples of papers using versions of regression prediction include Mulligan (1999), Browning et al. (2003), Meyer and Sullivan (2003), Attanasio and Pistaferri (2014), Charles et al. (2014), Arrondel et al. (2015), Fisher et al. (2016) and Kaplan et al. (2020). Examples of papers using the **BPP** procedure include Schulhofer-Wohl (2011), Guvenen and Smith (2014) and Attanasio et al. (2015). Some studies have observed that, empirically, imputation by regression prediction seems to lead to biased estimates in specific contexts. However, they neither offer an explanation nor realise that the problem is a general one. For example, Charles et al. (2014) note that the intergenerational elasticity of consumption spending is lower when regression prediction is used to impute consumption to the PSID than when true consumption data is used. Palumbo (1999) also obtains lower estimates of risk aversion when using regression prediction to impute consumption than when using a version of the **BPP** procedure. We account for these findings by formally setting out the nature of the biases associated with regression prediction, and demonstrating that it specifically leads to an attenuation bias.

In the next section, we lay out our basic framework, and derive the main results. We also relate our results to the prior literature, including regression prediction, the **BPP** procedure and **nlGMM**, and also contrast our estimator with the 2-sample IV approach proposed by Klevmarken (1982) and Angrist and Krueger (1992) (which Lusardi, 1996), applies to combine CE consumption data with PSID income data). Section 3 takes up the question of efficiency and inference. Monte Carlo evidence on finite sample performance of our estimator is presented in an Appendix. Section 4 provides two empirical examples using the CE and PSID. Section 5 concludes.

2 | CONSISTENCY AND IDENTIFICATION

2.1 | Set-up and assumptions

Consider the following linear regression model

$$y = X\beta + \epsilon. \tag{1}$$

where β is the $K \times 1$ parameter vector of interest. To make things concrete, the $n \times 1$ vector y could be consumption (or nondurable consumption), and the $n \times K$ matrix X would include income or wealth and other determinants of consumption. To keep the notation compact, variables have been de-meaned so there is no constant, but the addition of constants (and non-zero means) is not important for the analysis that follows.

We assume that for any random sample $i = 1, \ldots, n$ of $\{y_i, X_i\}_{i=1}^n$ from the population the following hold:

A1 $E(X_i'X_i) = \Sigma_{XX}$ is finite and non-singular, and $E\left(X_i'\epsilon_i\right) = 0$.

This means that given such a sample, an unbiased and consistent estimate of β can be obtained by OLS on Equation (1).

Suppose however that we have no such data on $\{y_i, X_i\}$. In this case, there are conditions under which we can consistently estimate β given a proxy for y, denoted Z, and samples where y and Z (but not X), and Z and X (but not y) are observed.

Let subscripts 1 and 2 denote whether variables correspond to sample 1 or sample 2; from here forward, the absence of a sample subscript indicates a population quantity. Using this notation, we would have a sample of data on $\{y_{1i}, Z_{1i}\}$ for $i = 1, \ldots, n_1$ and a second sample of data on $\{X_{2j}, Z_{2j}\}$ $j = 1, \ldots, n_2$. Z_m is an $n_m \times L$ matrix of proxies ($l = 1, \ldots, L$) for y from sample m. In our consumption example, Z is often food spending. Food spending is captured in many general purpose surveys, and is thought to be well-measured.

To derive asymptotic results for different estimators of β that impute y using Z, we make the following additional assumptions:

A2 $\{y_{1i}, Z_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}, Z_{2j}\}_{j=1}^{n_2}$ are i.i.d random samples from the same population, with finite second moments and which are independent.

A3 $E\left(Z_{1i}'Z_{1i}\right) = \Sigma_{ZZ}$. Σ_{ZZ} is non-singular. $E\left(X_{2j}'X_{2j}\right) = \Sigma_{XX}$ (which from A1 is also non-singular) and $E(y_{1i}^2) = E(y_{2j}^2) = \sigma_{yy} > 0$.

In the derivations below we do not need to impose that $E\left(Z_{2j}'Z_{2j}\right) = \Sigma_{ZZ}$ or that $E\left(X_{1i}'X_{1i}\right) = \Sigma_{XX}$, as it is never ambiguous which samples are being used to calculate these objects. These equalities are however guaranteed by Assumptions A2 and A3.

Assumptions A2 and A3 guarantee the existence of linear projections of Z onto y and of y onto Z

$$Z_{1i} = y_{1i}\gamma' + u_{1i} \tag{2}$$

where γ is $L \times 1$ and u_{1i} is $1 \times L$ and

$$y_{1i} = Z_{1i}\zeta + \xi_{1i}. \tag{3}$$

The residuals u_{1i} and ξ_{1i} satisfy the conditions for Equations (2) and (3) to be linear projections (i.e., $E(y_{1i}u_{1i}) = 0$ and $E(Z_{1i}'\xi_{1i}) = 0$). However, note that this is completely general in that we are not making any structural assumptions about the joint distributions of y_{1i} and Z_{1i}; the orthogonality of y and Z variables with the error terms u and ξ arise by construction. In addition, no homoscedasticity assumptions are placed on u or ξ.

We also assume that:

A4 $\lim_{n_2 \to \infty} \frac{n_2}{n_1} = \alpha$ for some $\alpha > 0$.

This ensures that, as n_1 tends to infinity, n_2 does as well.

The key assumptions that we make to allow consistent estimation of β are:

A5 $E\left(Z_{1i}'y_{1i}\right) = E\left(Z_{2j}'y_{2j}\right) = \Sigma_{Zy}$ which has at least one non-zero entry.

A6 $E\left(X_{2j}'u_{2j}\right) = 0$. Assumption A5 ensures that the proxies Z_1 have information about y (that the slope of the linear projections in Equations (2) and (3) are not zero). Assumption A6 will be discussed further below.

Assumptions A1–A3 and A5 allow us to define the population R^2 from a regression of y on Z, $\phi_{y,Z} \equiv \Sigma_{yZ}\Sigma_{ZZ}^{-1}\Sigma_{Zy}/\left(\sigma_{\epsilon\epsilon} + \beta'\Sigma_{XX}\beta\right)$, and to guarantee that $0 < \phi_{y,Z} \leq 1$. Assumption A3 is necessary to ensure that this quantity is defined. Assumption A5 ensures that it is strictly positive and thus that its reciprocal is also defined.

To compute variances for different estimators allowing for general forms of heteroscedasticity, we make the following further assumptions:

A7 $E(Z_{1i}'\xi_{1i}\xi_{1i}'Z_{1i}) = \Omega_{Z\xi}$ which is finite and positive semi-definite.

A8 $E(X_{2j}'\delta_{2j}\delta_{2j}'X_{2j}') = \Omega_{X\delta}$ which is finite and positive semi-definite. δ_2 are residuals from a regression of $Z_2/\phi_{y,Z}$ on X_2.

A9 $\{y_{1i}, Z_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}, Z_{2j}\}_{j=1}^{n_2}$ have finite fourth moments.

Finally, in what follows, we also make use of the following definitions and notation:

D1 $E\left(X_{2j}'y_{2j}\right) = \Sigma_{Xy}$ and $E\left(X_{2j}'Z_{2j}\right) = \Sigma_{XZ}$.

D2 We define for instance $\Sigma_{Zy}' = \Sigma_{yZ}$.

D3 $E\left(u_{1i}'u_{1i}\right) = \Sigma_{uu}$ which under A2 is finite and positive semi-definite and $E\left(\epsilon_{2j}^2\right) = \sigma_{\epsilon\epsilon} \geq 0$. With a single proxy, $E\left(u_{1i}^2\right) = \sigma_{uu} \geq 0$.

D4 R_{y_1,Z_1}^2 is the sample analogue of $\phi_{y,Z}$ (taken from sample 1).

D5 $E(y_{1i}) = \mu_y$.

2.2 | Rescaled regression prediction (RRP)

Consider regressing y_1 on Z_1 in the first sample using the resulting coefficients to predict (impute) y in the second, and then regressing this regression prediction, $\hat{y}_2^{RP}$, on X_2. Denote the resulting vector of coefficient estimates by $\hat{\beta}^{RP}$. We first show that $\hat{\beta}^{RP}$ is not, in general, a consistent estimator of β. We then show that under a mild assumption the asymptotic bias can be characterised, and this leads immediately to the consistent two-step estimator we propose.

Proposition 1. *Given assumptions A1–A5, Regression of $\hat{y}_2^{RP}$ on X yields inconsistent estimates of β unless (i) X is contained within the span of Z or (ii) y is contained within the span of Z or (iii) $\Sigma_{Xy} = \Sigma_{XZ} = 0$.*

Proof.

$$plim\left(\hat{\beta}^{RP}\right) = plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'Z_2}{n_2}\left(\frac{Z_1'Z_1}{n_1}\right)^{-1}\frac{Z_1'y_1}{n_1}\right\}$$

$$= plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'Z_2}{n_2}\left(\frac{Z_1'Z_1}{n_1}\right)^{-1}\frac{Z_1'(Z_1\zeta+\xi_1)}{n_1}\right\}$$

$$= plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'Z_2\zeta}{n_2}\right\} = plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'(y_2-\xi_2)}{n_2}\right\}$$

where ξ_2 is the difference between Z_2 and the (unobserved) value of y_2

$$= \beta - plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'\xi_2}{n_2}\right\}$$

$$= \beta - plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{(X_2'y_2 - X_2'Z_2(Z_2'Z_2)^{-1}Z_2'y_2)}{n_2}\right\}.$$

Given Assumption A5, the second term will be zero if and only if (i) $\Sigma_{Xy} = \Sigma_{XZ} = 0$ or (ii) $\exists$ some finite, non-zero $L \times K$ matrix ϕ s.t $X = Z\phi$ or (iii) $\exists$ some finite, non-zero L -vector λ s.t $y = Z\lambda$. □

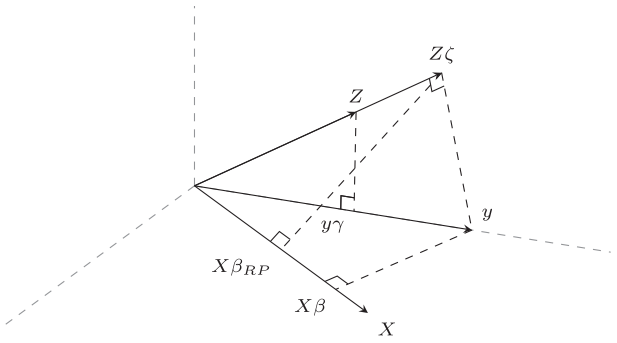
If (i) holds both parts of the bias term are zero but note that this would imply that $\beta = 0$ so that the estimator is consistent at only one point in the parameter space. If (ii) or (iii) hold, the two parts of the bias term are equal (and so cancel). Note though that (iii) implies that the first stage R^2 is one, and if (ii) holds there is no need for data combination. Thus, the regression prediction procedure only consistently estimates β in extreme cases.

The source of the problem is that regression prediction results in a prediction, $\hat{y}_2^{RP}$, that differs from y_2 by a prediction error or Berkson measurement error, ξ_2 that is uncorrelated with Z_2 but not uncorrelated with y_2 and, in general, not uncorrelated with X_2. As is well known, classical measurement errors in an independent variable causes bias in linear regression, but classical measurement errors in the dependent variable do not.[1] This is because classical measurement errors in y are by assumption (and in contrast to Berkson errors) uncorrelated with y and X. It is also widely recognised that Berkson errors in an independent variable do not cause bias in a linear regression (Berkson, 1950; Wansbeek & Meijer, 2000). What is less frequently recognised is that Berkson errors in a dependent variable do cause bias.

Figure 1 gives a geometric intuition for the problem for a case with a single proxy and independent variable. The solid lines represent the vectors y, Z and X. The dashed lines illustrate orthogonal projections. The orthogonal projection of y onto X (which would be obtained by regression with complete data) is labelled $X\beta$. The regression prediction procedure first projects the y onto Z, giving $\hat{y} = Z\zeta$, and then projects this vector onto X giving $X\beta^{RP}$. Note that $X\beta^{RP} \neq X\beta$.

[1]Nonclassical measurement in a dependent variable often causes bias, but not always. For example, Pischke (1995) shows that mean reverting—and hence nonclassical—measurement errors in earnings reports affect only the transitory part of earnings. If schooling affects only the permanent part of earnings, this explains with such mean-reverting measurement error does not attenuate estimates of the returns to schooling.

FIGURE 1 Regression prediction imputation procedure as projections



The same problem arises when we add residuals to the regression prediction to mimic the variance of the missing variable. The true value of (unobserved) y_{2j} can be decomposed into its projection onto Z and an orthogonal error

$$y_{2j} = \hat{y}_{2j}^{RP} + \hat{\xi}_{2j}. \tag{4}$$

Consider then drawing a random residual from the first stage regression (ξ_{1j}) to create a stochastic imputation

$$\hat{y}_{2j}^{RP} = \hat{y}_{2j}^{RP} + \hat{\xi}_{1j} = y_{2j} - \hat{\xi}_{2j} + \hat{\xi}_{1j}. \tag{5}$$

Then $\hat{y}_2^{RP}$ differs from y_2 by the error $\hat{\xi}_1 - \hat{\xi}_2$ which is by construction orthogonal to Z_1, but not y_2 or X_2. Note that, because it is randomly drawn from a separate random sample, $\hat{\xi}_1$ is orthogonal to y_2. The problem with the composite error $\hat{\xi}_1 - \hat{\xi}_2$ lies in the prediction error $\hat{\xi}_2$.

The proof of Proposition 1 notes that, given Assumptions A1-A5, $E(X_2'\xi_2) = 0$ can hold only in extreme cases. The same is not true of the alternative projection error, u_2, associated with (2).

The next proposition shows that the further assumption A6 ($E(X_2'u_2) = 0$) implies a bias in $\hat{\beta}^{RP}$ that takes a simple form. The proofs of this and all subsequent propositions are collected in the Online Supplemental Appendix (Crossley et al., 2021).

Proposition 2. *Given assumptions A1–A6,*

$$plim\left(\hat{\beta}^{RP}\right) = \beta\phi_{y,Z}. \tag{6}$$

Proof. See Online Supplemental Appendix. □

Thus, with A6, $\hat{\beta}^{RP}$ is attenuated, and the degree of attenuation is one minus the first stage population R^2 (1-$\phi_{y,Z}$). It is important to note that we are working with de-meaned versions of the variables: More generally, $R^2_{y_1,Z_1}$ is the *centred* sample R^2, $\phi_{y,Z}$ is the *centred* population R^2 and the result holds without demeaning the data.

As noted above, many authors have followed Skinner (1987) in regressing total consumption expenditure (y_1) on food expenditure Z_1 (and possible other proxies) in the Consumer Expenditure Survey (CE) and using the resulting coefficients to predict $\hat{y}_2^{RP}$ in the PSID (and then regressing $\hat{y}_2^{RP}$ on X_2). With a single spending category as the proxy, the first stage linear projection here resembles an "inverse" Engel curve. R^2s for food Engel curves are typically between 50 and 70%, implying attenuation of between 30 and 50% in this literature.

In A6, note that u_2 is not observed so when $L = 1$ this condition is not empirically verifiable (though see below for cases when $L > 1$). It states that X should not affect Z independently of Y.

Condition A6 is an exclusion restriction, analogous to the exclusion restriction imposed in instrumental variable (IV) procedures (discussed further below). It states that Z is partially uncorrelated with X, given y. The reason it may often be possible to find proxies for which $E(X_{2i}'u_{2i}) = 0$, whereas $E(X_{2i}'\xi_{2i}) = 0$ can only be satisfied in very special cases follows from the fact that y and ξ are correlated by construction while y and u are by construction uncorrelated.

As the attenuation in the regression prediction procedure is an estimable quantity, the bias can be corrected. One can rescale $\hat{y}_2^{RP}$ by the estimated first stage (centred) $R^2_{y_1,Z_1}$, or, equivalently, rescale $\hat{\beta}^{RP}$ by the estimated first stage (centred) $R^2_{y_1,Z_1}$. We refer to the resulting estimator as "Rescaled Regression Prediction" (**RRP**), with the rescaled impute of y_2 denoted $\hat{y}_2^{RRP}$ and the resulting estimate of β denoted $\hat{\beta}^{RRP}$. The consistency of $\hat{\beta}^{RRP}$ is a consequence of Proposition 2.

Corollary 1.

$$plim\left(\hat{\beta}^{RRP}\right) = plim\left(\frac{\hat{\beta}^{RP}}{R^2_{y_1,Z_1}}\right) = \beta.$$

Note that we do not need to assume that $L \geq K$. Identification requires only that we have at least one proxy for the dependent variable, regardless of dimensions of X.

It is also worth noting that the problem we highlight with regression prediction extends to several closely-related imputation procedures. In particular, Lillard et al. (1986) and David et al. (1986) note that commonly employed hot-deck imputation procedures can be interpreted as regression prediction plus an added residual. Such procedures draw a matched observation, y_{1i}, of the missing variable y_{2j}, from a cell in the donor data set. Cells are defined by categorical variables derived from Z. This replacement of the missing y_{2j} with y_{1i} from a donor observation matched on Z can be viewed as a prediction using the coefficients from a saturated first stage regression on those categorical indicators for Z, plus a residual from the first stage regression, and so is an example of the regression prediction procedure (with added residuals), and our results apply. The **RRP** estimator is easily extended to allow for covariates and to panel data. This is discussed in the Online Supplemental Appendix (Crossley et al., 2021).

2.3 | Comparison with other estimation strategies

It is useful to compare **RRP** to two other (consistent) estimation strategies. First, Blundell et al. ((2004), (2008)), again using the CE and PSID, first regress Z_1 (food spending) on y_1 then predict $\hat{y}_2^{BPP} = Z_2\frac{1}{\hat{\gamma}}$ (the **BPP** procedure). That is, they estimate an Engel curve and then invert it to predict consumption. Second, one could not impute y_2 at the observation level at all, but to recover the parameter of interest (β) from a combination of moments taken from the two surveys. This two-sample nonlinear GMM estimator (**nlGMM**) was first suggested (for a different application) by Arellano and Meghir (1992).[2] Note that the **BPP** procedure does not extend naturally to multiple proxies (because of the need to invert). However, full **nlGMM**, like our consistent two-step estimator (**RRP**), does extend to the case of multiple proxies.

Before taking up the case of multiple proxies, we note the relationship between these three estimators in the case of a single proxy. Denote the estimates of β from the **BPP** procedure and two-sample **nlGMM** by $\hat{\beta}^{BPP}$ and $\hat{\beta}^{nlGMM}$ respectively.

Proposition 3. *If and only if there is a single proxy Z (a vector)* $\hat{\beta}^{RRP}$, $\hat{\beta}^{BPP}$ *and* $\hat{\beta}^{nlGMM}$ *are numerically identical.*

Proof. See Online Supplemental Appendix. □

This implies consistency of the **BPP** procedure. As the **BPP** procedure is identical to **RRP** (or **nlGMM**) when there is one proxy, but does not extend to multiple proxies, it can be seen as a special case of **RRP** (or **nlGMM**).

Turning to the case of multiple proxies, assumptions A1–A6 imply the following $KL + L$ moment conditions

$$G(\gamma,\beta) = \begin{array}{l} E\left[X'_{2j,k}\left(Z_{2j,l} - X_{2j,k}\beta_k\gamma_l\right)\right] = 0\ \forall k,l \\ E\left[y'_{1i}\left(Z_{1i,l} - y_j\gamma_l\right)\right] = 0\ \forall l \end{array} \tag{7}$$

where $Z_{1j,l}$ is for example the lth column of Z_{1j}. Following Arellano and Meghir (1992), these moments can be used to consistently estimate (γ,β) using **nlGMM**.

By contrast, **RRP** exploits the restricted set of $K + L$ moments.

$$\begin{array}{l} E\left[Z_{2j}\zeta/\phi - X_{2j}\beta\right] = 0 \\ E\left[Z'_{1i}(y_{1i} - Z_{1i}\zeta)\right] = 0. \end{array} \tag{8}$$

[2]Here, in the just-identified case of a single proxy, one could regress Z_1 on y_1 to get $\hat{\gamma}$, then regress Z_2 on X_2 to get $\widehat{\beta\gamma}$, and take the ratio of the two to estimate β. Thus, in this case, we would not need to use the generalised method of moments. For simplicity, we also refer to this estimator as β^{nlGMM}.

We could also exploit a larger set of moments if we adopted **RRP** separately for each proxy l, and then combined the results into a single estimate of β using GMM. In particular, we could exploit the $KL + L$ moments

$$\begin{aligned} E\left[X_{2j,k}(Z_{2j,l}\zeta_l/\phi_l - X_{2j,k}\beta_k)\right] &= 0 \ \forall k, l \\ E\left[Z_{1i,l}(y_{1i} - Z_{1i,l}\zeta_l)\right] &= 0 \ \forall l. \end{aligned} \tag{9}$$

It turns out that this "proxy-by-proxy" **RRP** estimator is numerically equivalent results to the **nlGMM** estimator using the moments in (7).

Proposition 4. *Proxy-by-proxy **RRP** using the moment conditions in (9) is numerically equivalent to **nlGMM** using the moments in (7).*

Proof. See Online Supplemental Appendix. □

To summarise, with a single proxy, all of these approaches yield identical estimates. An attraction of **RRP** and **nlGMM** over the **BPP** procedure is that they extend naturally to multiple proxies. Relative to **nlGMM** with multiple proxies, **RRP** has the advantage that it is easier and quicker to estimate. Applied researchers may be more comfortable with a simple, two-step regression-based approach. Further, we show below that it often performs better than **nlGMM** in finite samples.

2.4 | Testing overidentification restrictions

As we noted above, A6 is not testable when $L = 1$. However, when $L > 1$, that is, when we have multiple proxies, it is possible to test the overidentifying restrictions imposed by A6 in an analogous way to similar tests carried out for IV estimators. Intuitively, we can test to see whether estimates of β obtained from different proxies are consistent with one another.

For **nlGMM**, the overidentifying restrictions can be tested using Hansen's J-statistic (Hansen, 1982).

It is also possible to test the overidentifying restrictions for the **RRP** estimator. Although this could be done in a number of ways, our suggestion is a "double length" artificial regression (Davidson & MacKinnon, 1988), which can be made heteroscedasticity robust by use of a sandwich variance estimator (Wooldridge, 1991). Consider the first "reduced form" moments from (7). These moments are the source of the overidentification (when $L > 1$) and they can be tested after **RRP** as follows.

- Construct the residual $v_{2j,l} = (Z_{2j,l} - X_{2j}\beta\gamma_l) \ \forall l$. Note that the vector of γ coefficients is not estimated directly in the **RRP** procedure, but they can be recovered from the projection of y_1 on Z_1 using the fact that $\gamma = (y_1'y_1)^{-1}Z_1'Z_1\zeta$.
- Replicate the second sample L times and stack.
- Generate $v = v_l$ for replicate l, and a set of replicate indicator variables $d = d_1, \ \dots \ , d_L$.
- Regress v on the full set of interactions between d and x in the stacked data.
- Test that all coefficients are zero.

Davidson and MacKinnon (1988) show that tests based on double-length regressions are asymptotically valid. In the Online Supplemental Appendix (Crossley et al., 2021), we show using Monte Carlo simulations that this test also performs well, and is appropriately sized, in finite samples. The same is true for its heteroscedastic version when the errors are heteroscedastic. However, it is worth noting that in the case of weak proxies, neither version is well sized.

2.5 | Estimating other moments of y

It is useful also to think about other moments, as these imputation procedures have been used to study dispersion as well as regression coefficients. For example, Blundell et al. (2008), Attanasio and Pistaferri (2014) and Fisher et al. (2016) study consumption inequality. There are a number of reasons why one might wish to impute y from some other sample to calculate means, variances and covariances rather than calculating them directly in the initial sample. For example, one might want to calculate variances among subsets of the population that cannot be defined in the first sample but can be defined in the second. In addition, we might be interested in the growth of y among particular individuals along with variances and covariances for these growth rates from panel data, but may only directly observe y in cross-sectional data.

For instance, Blundell et al. (2008) calculate variances of growth rates in total consumption expenditures for households in the PSID, and covariances of these growth rates with the growth in incomes, where total consumption is imputed to the PSID from the cross-sectional CE.

We continue with the case of a single proxy to allow comparison of regression prediction to **BPP** and **RRP** and consider the case of a single X variable for ease of exposition (though the results extend to cases where $K > 1$). **nlGMM** recovers β directly and does not generate unit level estimates of y. The imputes $\hat{y}_2^{RP}$ and $\hat{y}_2^{RRP}$ are numerically different,

$$\hat{y}_2^{RP} = Z_2(Z_1'Z_1)^{-1}Z_1'y_1, \tag{10}$$

$$\hat{y}_2^{RRP} = Z_2(Z_1'Z_1)^{-1}Z_1'y_1/R^2_{y_1,Z_1}. \tag{11}$$

Algebra analogous to the proof of Proposition 3 shows that $\hat{y}_2^{RRP}$ and $\hat{y}_2^{BPP}$ are numerically identical for the case when all variables have been de-meaned. They will differ by an additive constant in the event a non-zero intercept shift is present in Equation (2).

Proposition 5. *Denote sample moments based on $\hat{y}_2^{RP}$ by s_{yy}^{RP} and S_{yX}^{RP}; and analogously for $\hat{y}_2^{RRP}$. Then, given assumptions A1–A6,*

(i) $plim\left(s_{yy}^{RP}\right) = \sigma_{yy} \times \phi_{y,z}$
(ii) $plim\left(S_{yX}^{RP}\right) = \Sigma_{yX} \times \phi_{y,z}$
(iii) $plim\left(s_{yy}^{RRP}\right) = \sigma_{yy}/\phi_{y,z}$
(iv) $plim\left(S_{yX}^{RRP}\right) = \Sigma_{yX}$

where again $\phi_{y,Z}$ is the population R^2 from the first stage regression and, for instance, $s_{yy}^{RP} = \frac{1}{n_2}\hat{y}_2^{RP\prime}\hat{y}_2^{RP}$ and $S_{yX}^{RP} = \frac{1}{n_2}\hat{y}_2^{RP\prime}X_2$.

Proof. See Online Supplemental Appendix. □

Proposition 5 states that the sample variance of $\hat{y}_2^{RP}$ underestimates the population variance of y, while S_{yX}^{RP} is not a consistent estimator of Σ_{yX}. This gives an additional intuition for the inconsistency of $\hat{\beta}^{RP}$ as an estimator of β as with a scalar X the OLS estimate of β is just S_{yX}^{RP}/s_{yy}^{RP}. Moreover, adding a residual to $\hat{y}_2^{RP}$ does not correct this. Similarly, the **RRP** estimator is consistent for β because it is consistent for Σ_{yX}.

Simple algebra also establishes that, when $L = 1$

$$s_{yy}^{RRP} = s_{yy}^{BPP} \tag{12}$$

and

$$S_{yX}^{RRP} = S_{yX}^{BPP}. \tag{13}$$

This follows from the numerical equivalence of the de-meaned values of $\hat{y}_2^{RRP}$ and $\hat{y}_2^{BPP}$. Thus, $plim\ s_{yy}^{BPP} = plim\ s_{yy}^{RRP} > \sigma_{yy} > plim\ s_{yy}^{RP}$. Turning again to our motivating consumption example, Attanasio and Pistaferri (2014) show that trends in s_{yy}^{BPP} and s_{yy} (where y is observed) are similar, but that there is a level difference. The similarity in trends suggests that the first stage $R^2_{y_1,Z_1}$ is roughly constant across years in their data. We confirm this in our empirical example below.

For completeness, we can also consider means. Had we not de-meaned the data, then it is straightforward to show that the average of $\hat{y}_2^{RP}$ gives a consistent (and unbiased) estimate of the population mean of y. However, if **RRP** is implemented by rescaling $\hat{y}_2^{RP}$ (rather than rescaling β^{RP}), it then immediately follows that the mean of this rescaled prediction of y is not a consistent estimator of the mean of y. One implication is that a statistical agency aiming to add an imputed $\hat{y}$ to a data release could not add a single variable that would be appropriate both for use as a dependent variable and for estimating quantities that depend on the first moment of y (poverty rates, for example).

2.6 | Related literature

In this paper, we study the use of proxies to predict a dependent variable.[3] Regression prediction of a dependent variable induces a prediction or Berkson measurement error. Berkson measurement errors in a dependent variable cause bias in a linear regression, and this seems to be much less noted than the innocuous cases of Berkson measurement error in an independent variable, or classical measurement error in a dependent variable.[4] Two exceptions are Hyslop and Imbens (2001) and Hoderlein and Winter (2010). Hyslop and Imbens (2001) show attenuation bias in a regression of $\hat{y}$ on X where $\hat{y}$ is an optimal linear prediction generated by a survey respondent (not the econometrician). Relative to the imputation problem we study, key differences include the fact that it is the survey respondent doing the prediction, and the assumption that the respondent's information set includes Z, β and $E(X)$. They also assume (in our notation) that $Z = y + u$; ($\gamma = 1$). Hoderlein and Winter (2010) study a similar problem, but in a nonparametric setting. Again, in their model it is the survey respondent, rather than the econometrician, doing the predicting. They illustrate their results using self-reported data on consumption expenditure.

Dumont et al. (2005) study corrected standard errors in a regression with a "generated regressand." Their work is motivated by the two-stage procedure for mandated-wage regression proposed by Feenstra and Hanson (1999). In this paper, domestic prices are first regressed on some structural determinants (trade and technology variables). The estimated contributions of these variables to price changes are then in turn regressed on factor shares to identify the changes in factor prices "mandated" by changes in product prices.

In this context, the first stage is

$$y_i = Z_i\zeta + \xi_i \tag{14}$$

(where y_i are in this case product prices and Z_i are some structural determinants of those prices) and the second stage is not (1) but rather L separate equations:

$$Z_{i,l}\zeta_l = X_i\beta_l + \epsilon_{i,l} \tag{15}$$

(where X_i are factor shares). Here, $Z_{i,l}\zeta_l$ is not observed and so is replaced by the first stage estimate $Z_{i,l}\hat{\zeta}_l$. Of course $\hat{\zeta}$ differs from ζ by an estimation error $(Z'Z)^{-1}Z'\hat{\xi}$, but, given the set-up, the stochastic element $\hat{\xi}$ is orthogonal to Z, and so also X, and thus causes problems for inference but not inconsistency. Although the motivation and second-stage regressand are different, this procedure is analogous to the **BPP** procedure (using y_i as a proxy for $Z_{i,l}\zeta_l$), so the Berkson measurement error problem does not arise.

We also contribute to the literature on data combination; see the excellent survey by Ridder and Moffitt (2007). One approach to data combination problems is statistical matching. This can be used to create synthetic observations which each have values of y, X and Z. As first noted in Sims (1972), the use of this approach to identify features of the joint distribution $f(y, X, Z)$ makes the strong and untestable assumption that, conditional on Z, y is independent of X. An alternative is to assume that, conditional on X, y is independent of Z. This leads to the 2-sample IV (2SIV) and 2-sample 2SLS approaches. 2SIV was first proposed by Klevmarken (1982) and popularised by Angrist and Krueger (1992). Inoue and Solon (2010) show that 2SIV is not in general efficient because it does not take account of the fact that Z_1 and Z_2 will be different in finite samples. They suggest the 2-sample two-stage least squares estimator is therefore preferred. Pacini and Windmeijer (2016) provide robust inference for that estimator. 2SIV has been applied to the combination of CE consumption data and PSID income data by Lusardi (1996).

To see the contrast between 2-sample two-stage least squares and our approach, start from (1), and consider the linear projection of X_2 on Z_2:

$$X_{2j} = Z_{2j}\theta + \nu_{2j}. \tag{16}$$

[3]Wooldridge (2002) contains an excellent overview of the use of proxies for independent variables and Lubotsky and Wittenberg (2006) and Bollinger and Minier (2015) are recent papers on the optimal use of multiple proxies for an independent variable. Ridder and Moffitt (2007) gives a broad survey of the literature on data combination.

[4]Berkson measurement error in an independent variable is also a problem in nonlinear models. See for example Blundell et al. (2019). The distinction between a "measurement error" in the sense of a classical measurement error and a "forecast error," or Berkson measurment error has also been important in the literature on GDP revisions (Mankiw & Shapiro, 1986).

Note that $E(Z'_{2j}\nu_{2j}) = 0$ by definition. Then replace assumptions with A5 and A6 with parallel assumptions:

A10 $E\left(Z'_{1i}X_{1i}\right)$ and $E\left(Z'_{2j}X_{2j}\right)$ both have rank K.

A11 $E\left(Z'_{1i}\epsilon_{1i}\right) = 0$.

alongside the order condition, that

A12 $L \geq K$.

Under these assumptions, the 2-sample-2SLS estimator is

$$\hat{\beta}^{2S2SLS} = (\hat{X}_1{}'\hat{X}_1)^{-1}\hat{X}_1{}'y_1 \tag{17}$$

where $\hat{X}_1 = Z_1(Z'_2Z_2)^{-1}Z'_2X_2$ is consistent for β.

This approach is typically taken where Z is a grouping variable or variables (e.g., birth cohort, occupation, birth cohort × education).

Again, A11 states that, conditional on X, y is uncorrelated with Z. In contrast, our A6 states that conditional on y, Z is (partially) uncorrelated with X. In fact, a useful proxy must have information about y over and above the information in X. To see this, considering the combination of Equations (1) and (2):

$$Z_{1i} = X_{1i}\beta\gamma' + \epsilon_{1i}\gamma' + u_{1i}. \tag{18}$$

Given A5, then A11 can only hold in the knife-edge case, $E\left(\gamma\epsilon'_{1i}\epsilon_{1i}\right) = -E(u'_{1i}\epsilon_{1i})$ (note also that A5 implies that $\gamma \neq 0$). This means that a variable may be a plausible instrument or a plausible proxy, or neither; but not both.[5]

With 2-sample 2SLS, we use Z to impute$X$ (and as the resulting prediction or Berkson error is in an independent variable, this two-stage procedure does not cause inconsistency). An additional virtue of this procedure is that we can relax the exogeneity requirement in A1, for example, allowing for classical measurement error in X.

On the other hand, there is no equivalent of the order condition A12 in the proxy framework set out in Section 2.1. That is, there is no need to ensure that there are at least as many proxies as explanatory variables for **RRP** and **nlGMM**, even though we require that there are at least as many instruments as endogenous variables in X to employ two-sample IV.

Cross and Manski (2002) consider a quite general data combination problem. They consider identification of the "long" regression $E(y_i|X_i, Z_i)$ when only the joint distributions $f(y, Z)$ and $f(X, Z)$ are known. Of course, if $E(y_i|X_i, Z_i)$ can be recovered, then the missing "short" regression, $E(y_i|X_i)$, follows from the application of the law of total expectation (using $f(X, Z)$). Cross and Manski (2002) consider both partial identification when nothing else is known, and identification under exclusion restrictions. The exclusion restrictions they study are those proposed by Sims (1972) and Klevmarken (1982). Thus, again, the most important difference in our approach is that we consider the alternative exclusion restriction A6.

3 | INFERENCE AND PRECISION

3.1 | Intuition from the one proxy, homoscedastic case

If we strengthened the assumptions listed in Section 2 to include homoscedasticity ($E\left(\epsilon_i^2|X_i\right) = \psi_{\epsilon\epsilon} > 0$) and conditional independence of the error term ($E(\epsilon_i|X_i) = 0$), then direct estimation of (1) on complete data would result in an asymptotic variance for $\hat{\beta}$ of $(\Sigma_{XX})^{-1}\psi_{\epsilon\epsilon}$. When we impute $\hat{y}$ from one data set to another, there are two losses of precision resulting from (i) imputation and (ii) the combination of two different samples of the underlying population. Moreover, applying the usual OLS standard error formula, the regression of $\hat{y}$ on X results in standard errors that are too small. We use the one-proxy (and single X variable), and homoscedastic case to illustrate these points, and then give a correct formula for the asymptotic standard errors with possibly multiple proxies and heteroscedastic errors.

[5]A similar point is made with respect to proxy and IV approaches to an "omitted variable" (a missing independent variable) in Wooldridge (2002).

With a single proxy, $\hat{\beta}^{nlGMM}$, $\hat{\beta}^{RRP}$ and $\hat{\beta}^{BPP}$ are numerically identical, so we derive the asymptotic variance from the **nlGMM** approach. The first stage (2) and reduced form (18) give two moments

$$E\left(y_{1i}'(Z_{1i} - y_{1i}\gamma')\right) = E\left(y_{1i}'u_{1i}\right) = 0,$$
$$E\left(X_{2j}'(Z_{2j} - X_{2j}\beta\gamma')\right) = E\left(X_{2j}'(\epsilon_{2j}\gamma' + u_{2j})\right) = 0$$

which identify the parameters γ and β.

It is informative to first consider implementing $\hat{\beta}^{nlGMM}$ (or equivalently $\hat{\beta}^{BPP}$ or $\hat{\beta}^{RRP}$) on a single sample, containing all of y, Z, X (of course, a researcher would have no reason to do this, but it delivers a useful intuition). In this one-sample case, given the further assumptions $E\left(u_i^2|y_i\right) = \psi_{uu}$ and $E(u_i|y_i) = 0$, the asymptotic variance-covariance matrix of the moments is

$$F = \begin{bmatrix} \psi_{uu}\sigma_{yy} & \psi_{uu}\Sigma_{XX}\beta \\ \psi_{uu}\Sigma_{XX}\beta & \left(\gamma^2\psi_{\epsilon\epsilon} + \psi_{uu}\right)\Sigma_{XX} \end{bmatrix} \tag{19}$$

where the off-diagonal terms are not zero because the moments come from the same random sample. The asymptotic variance covariance matrix of (β, γ) is $\left(G'F^{-1}G\right)^{-1}$ where G is the gradient of the moments with respect to the parameters. The asymptotic variance of $\hat{\gamma}$ is of course $\sigma_{yy}^{-1}\psi_{uu}$. The asymptotic variance of $\hat{\beta}$ is

$$Asymp\,Var(\hat{\beta}) = \frac{(\Sigma_{XX})^{-1}\psi_{\epsilon\epsilon}}{\phi_{y,Z}}. \tag{20}$$

Thus the loss of asymptotic precision due to imputation (relative to the direct estimation of (1)), is inversely related to the first stage population R^2 ($\phi_{y,Z}$). Note the similarity of this precision loss to the precision loss in the case of linear IV estimation (relative to OLS), which is related to a first stage R^2 in the same way (Shea, 1997).

Turning now to the realistic two-sample case, the asymptotic variance-covariance matrix of the moments becomes

$$F = \begin{bmatrix} \alpha\psi_{uu}\sigma_{yy} & 0 \\ 0 & \left(\gamma^2\psi_{\epsilon\epsilon} + \psi_{uu}\right)\Sigma_{XX} \end{bmatrix}$$

where note that the off-diagonal terms are now zero because the moments come from independent random samples. The asymptotic variance covariance matrix of (β, γ) is again $\left(G'F^{-1}G\right)^{-1}$ where G is the gradient of the moments with respect to the parameters. The asymptotic variance of $\hat{\gamma}$ is the same as before (though now multiplied by the term α). $\alpha\left(\sigma_{yy}\right)^{-1}\psi_{uu}$. The asymptotic variance of $\hat{\beta}$ is

$$\begin{aligned} Asymp\,Var(\hat{\beta}) &= (\Sigma_{XX})^{-1}\left(\psi_{\epsilon\epsilon} + \gamma^{-2}\psi_{uu}\right) + \alpha\Sigma_{yy}^{-1}\beta^2\gamma^{-2}\psi_{uu} \\ &= (\Sigma_{XX})^{-1}\psi_{\epsilon\epsilon} + \gamma^{-2}(\Sigma_{XX})^{-1}\psi_{uu} + \alpha\beta^2\gamma^{-2}\sigma_{yy}^{-1}\psi_{uu}. \end{aligned}$$

This can be written as

$$Asymp\,Var(\hat{\beta}) = \frac{(\Sigma_{XX})^{-1}\psi_{\epsilon\epsilon}}{\phi_{y,Z}} + (1+\alpha)\beta^2\left(\frac{1-\phi_{y,Z}}{\phi_{y,Z}}\right). \tag{21}$$

The second term inside the brackets represents the loss of asymptotic precision, due to the use of two different samples. Precision is greater in (20) because the covariances between moments in Equation (19) have a stabilising influence on the estimates $\hat{\beta}$. These covariance terms are zero in the two sample case.

Finally, the usual OLS standard errors from a regression of an imputed dependent variable (derived from **RRP** or the **BPP** procedure) are incorrect, but can easily be corrected. The OLS standard errors (as produced by standard software packages) are

$$\hat{V}^{OLS}(\hat{\beta}^{RRP}) = (X_2X_2)^{-1}\left(\hat{y}_2 - X_2\hat{\beta}\right)'\left(\hat{y}_2 - X_2\hat{\beta}\right) = (X_2X_2)^{-1}\left(\hat{y}_2'\hat{y}_2 - \hat{y}_2'X_2(X_2'X_2)^{-1}X_2'\hat{y}_2\right)$$
$$= (X_2'X_2)^{-1}\left[y_1'y_1(Z_1'y_1)^{-1}Z_2'Z_2(Z_1'y_1)^{-1}y_1'y_1 - y_1'y_1(Z_1'y_1)^{-1}Z_2'X_2(X_2'X_2)^{-1}X_2'Z_2(Z_1'y_1)^{-1}y_1'y_1\right].$$

With some algebra, it is straightforward to show that

$$plim\left(\hat{V}^{OLS}(\hat{\beta}^{RRP})\right) = \frac{(\Sigma_{XX})^{-1}\psi_{\epsilon\epsilon}}{\phi_{y,Z}} + \beta^2\left(\frac{1-\phi_{y,Z}}{\phi_{y,Z}}\right)$$
$$= Asym\ Var(\hat{\beta}^{RRP}) - \alpha\beta^2\left(\frac{1-\phi_{y,Z}}{\phi_{y,Z}}\right). \tag{22}$$

So, the usual OLS standard errors are too small, by $\alpha\beta^2\left(\frac{1-\phi_{y,Z}}{\phi_{y,Z}}\right)$. Given assumption A9, the OLS standard errors can be corrected using available consistent estimates of α, β and $\phi_{y,Z}$, $\frac{n_2}{n_1}$, $\hat{\beta}$ and $R^2_{y_1,Z_1}$ (or simply by using the **nlGMM** standard errors).

3.2 | Asymptotic standard errors—general case

If there is more than one proxy, $\hat{\beta}^{RRP} \neq \hat{\beta}^{nlGMM}$. Here, we derive the asymptotic variance of $\hat{\beta}^{RRP}$ and relate it to uncorrected, "naive" estimates of the standard errors one would obtain from the second stage **RRP** regression (of $\hat{y}$ on X). Our formula allows for possibly heteroscedastic errors and can, for example, straightforwardly be extended to provide cluster-robust standard errors. Our approach closely follows that of Pacini and Windmeijer (2016) who provide robust standard errors for two sample 2SLS.

Proposition 6. *Given assumptions A1–A8, $\hat{\beta}^{RRP}$ has asymptotic variance*

$$Asymp\ Var(\hat{\beta}^{RRP}) = \Sigma_{XX}^{-1}\left[\Omega_{X\delta} + \alpha\frac{\Sigma_{XZ}}{\phi_{y,Z}}\Sigma_{ZZ}^{-1}\Omega_{Z\xi}\Sigma_{ZZ}^{-1}\frac{\Sigma_{ZX}}{\phi_{y,Z}}\right]\Sigma_{XX}^{-1}.$$

Proof. See Online Supplemental Appendix. □

Given the further assumption A9, this formula can be estimated using sample analogues of Σ_{XX}, $\phi_{y,z}$, Σ_{XZ}, α, $\Omega_{Z\xi}$ and $\Omega_{X\delta}$. It can also be written as

$$Asymp\ Var(\hat{\beta}^{RRP}) = V^{OLS}(\hat{\beta}^{RRP}) + \alpha\Sigma_{XX}^{-1}\frac{\Sigma_{XZ}}{\phi_{y,Z}}V_{OLS}(\hat{\gamma})\frac{\Sigma_{XZ}}{\phi_{y,Z}}\Sigma_{XX}^{-1}$$

where $V^{OLS}(\hat{\beta}^{RRP})$ are once again the (asymptotic) "naive" variances estimated using the second stage residuals. This expression can be used to adjust robust estimates for the variances of coefficients in the first and second stage regressions that are provided by Stata and other software packages. A Stata package that estimates the **RRP** estimator and provides the correct standard errors is available from the authors at https://github.com/spoupakis/rrp. The results in Proposition 6 can straightforwardly be extended to situations where we impute the dependent variable into panel data and where we use instrumental variables for X.

3.3 | Efficiency of alternative estimators

Standard results show that $\hat{\beta}^{nlGMM}$ is asymptotically efficient when estimated via GMM using the optimal weight matrix.

We showed in Proposition 4 that it is possible to write the nonlinear GMM moment conditions expressed in (7) as a set of **RRP** equations (each using a single proxy). This implies that that $\hat{\beta}^{nlGMM}$ can be thought of as an optimally weighted combination of individual **RRP** estimators. We also showed that $\hat{\beta}^{RRP}$ with multiple proxies can be expressed as the solution to a (restricted) set of $K + L$ moment conditions. In particular, the first KL "reduced form" moments in (7) are replaced by just K moments, with each using a (rescaled) linear combination of all L proxies.

The analogy to 2SLS, where a linear combination of instruments replaces the instrument-by-instrument moments of full GMM, is obvious. As with 2SLS, there is a potential loss of asymptotic efficiency. However, as with 2SLS, **RRP** is easier to implement than full GMM (it can be implemented with two regression), and it may be more efficient in finite samples (because it avoids estimation of the GMM weighting matrix).

While **RRP** is not in general asymptotically efficient, we can show that it has desirable asymptotic properties. In particular, **RRP** minimises the variance of imputation errors

Proposition 7. *Given assumptions A1–A6, **RRP** minimises the variance of imputation errors among the class of consistent, two-sample, estimators for β.*

Proof. See Online Supplemental Appendix. □

The Online Supplemental Appendix (Crossley et al., 2021) provides Monte Carlo evidence on the finite sample performance of alternative estimators under various assumptions. We find that **RRP** outperforms (or performs as well as) **nlGMM** in finite samples under different cases of varying strength, precision, and correlation of the proxies, under both homoscedastic and heteroscedastic error terms.

4 | EMPIRICAL ILLUSTRATIONS

In this section, we illustrate our results with two empirical examples using the PSID (Panel Study of Income Dynamics, 2019) and the CE Interview Survey.

4.1 | Housing wealth effects

We begin with an exercise similar to that of Skinner (1989) (making use of the imputation procedure set out in Skinner, 1987). This is to estimate the elasticity of consumption spending with respect to changes in housing wealth by regressing nondurable consumption spending on demographics, lags and leads of total family income and house values. We do this using the 2005–2013 waves of the PSID when a more-or-less complete measure of nondurable expenditures is available. Following the approach taken by Skinner (1989) for an earlier period, when spending data were only available for a subset of goods in the PSID, we also impute nondurable consumption spending from the CE into the PSID. This allows us to compare results from different imputation procedures with the complete data case (using the PSID's own consumption measures). In this respect, our exercise is similar to that used in Attanasio and Pistaferri (2014), who assess the accuracy of the imputed consumption measures they use in the early years of the PSID with those available in the PSID in later years.

Our measure of nondurable consumption is the sum of spending on food at home, food away from home, utilities (including gas and electricity), gasoline, car insurance, car repairs, clothing, vacations and entertainment. For proxies, we use the log sum of total food spending (whether at home or away from home) and log utility spending. Our demographics controls are the size of the household, age, age squared, the log earnings of the household head (set to zero for those with zero earnings) and a dummy for having zero earnings. We annualise consumption measures and then take logs in both surveys.

Our sample selection choices in the PSID are chosen to mirror those used in Skinner (1989). In particular, we take a sample of homeowners, who are observed in all waves from 2005 to 2013, who do not move, are not observed with zero incomes and who are not observed renting over the sample period. To prevent our results being driven by extreme values, we exclude those with incomes or house values in the top and bottom 1% of the PSID sample.

In the CE Interview Survey, we take a sample of homeowners. The CE Interview Survey aims to interview households over four quarters, asking retrospective consumption questions over the previous three months in each interview. We take only those individuals who were observed in all four interviews, and whose final interview was held a year coinciding with the biennial PSID survey waves from 2005–2013. We then average spending over each of the previous four quarters they were observed and keep only one observation per household. By averaging over multiple waves, we reduce measurement error in consumption and get consumption values which are more in the spirit of the questions households are asked in the PSID (households in the PSID are asked about their spending over the previous year, or "usual" spending in an average week or month). We run our imputation regressions separately in each year, which would for example allow for

TABLE 1 Imputing nondurable consumption spending using CES

	(1) 2005	(2) 2007	(3) 2009	(4) 2011	(5) 2013	(6) All years
log Food	0.568**	0.550**	0.545**	0.554**	0.560**	0.555**
	(0.012)	(0.008)	(0.008)	(0.008)	(0.009)	(0.004)
log Utilities	0.387**	0.387**	0.415**	0.395**	0.396**	0.396**
	(0.015)	(0.010)	(0.011)	(0.011)	(0.011)	(0.006)
Partial R^2	0.720	0.746	0.744	0.751	0.747	0.743
N	1590	2896	2759	2668	2470	9899

Note: Standard errors in parentheses. Additional controls for age, age squared, family size, log of head's earnings (set to zero if earnings are zero), a dummy for head's earnings being greater than zero, and (in the pooled regression) year dummies. The partial R^2 reported here is obtained by regressing our dependent variables on our proxies after partialling out the effects of other covariates in an initial regression.

*$p < 0.05$.

**$p < 0.01$.

the fact that changes in relative prices might change the relationship between food and total spending from one period to the next.[6]

Table 1 shows the results from our first stage imputation regressions. We note that the relationships between the proxy variables and total nondurable consumption and the fit of the imputation regressions remain very stable across the different survey years. Column (6) shows results pooling across all years (2005, 2007, 2009, 2011 and 2013).

Table 2 shows the results from regressions of consumption spending on income variables and house values in the PSID. The first column shows results using the consumption measure available in the PSID. This is the complete data case. The second column shows results using the regression prediction procedure employed by Skinner, and the third column shows results using **RRP**. The fourth column shows results estimated using **nlGMM**.

The complete data results from the PSID suggests that each 10% increase in house values is associated with a 1.14% increase in consumption spending. When we impute consumption using unscaled regression prediction, we underestimate the effects of housing wealth on consumption (with the estimated effect falling to 0.87%). Using **RRP** (inflating 0.87 by inverse of the partial R^2 in Table 1), we obtain a value of 1.16% which is very similar to that obtained using the complete data in the PSID. Column (4) shows results using **nlGMM**. This implies that a 10% increase in home values results in a 1.29% increase in spending, which is again closer to the result using complete data than regression prediction. Since we have more than one proxy, we can also test the exclusion restriction A6 by calculating Hansen's J-statistic for **nlGMM** and using our double-length artificial regression for **RRP**. Both tests fail to reject the overidentification restrictions.

These results illustrate the theoretical results of Section 2. The similarity between the results using **RRP** and **nlGMM** and the results using complete data confirms that A6 holds in these data, and moreover, that our demographic covariates adequately control for any sample differences between the PSID and the CE Interview Survey.

4.2 | Consumption inequality

As a second exercise, we examine the evolution of consumption inequality using actual and imputed nondurable consumption measures. This is in the spirit of the longer run analysis of consumption and inequality carried out in Attanasio and Pistaferri (2014).

To do this, we impute consumption measures for all households in the PSID (including non-homeowners) and plot the standard deviation over time for imputed consumption from the regression prediction procedure and from **RRP**. We then compare this with the standard deviation of nondurable consumption spending as measured in the PSID. To prevent this measure being unduly influenced by extreme values, we trim the top and bottom 1% of consumption spending in the PSID. The results are shown in Figure 2.

[6]Our approach differs from the approach used in Skinner (1989) in two key respects. First, Skinner (1989) imputes the absolute level of consumption using the absolute levels of food and utility spending before taking logs of the imputed values in the PSID, while we use the log of nondurable consumption, food and utility spending throughout. To avoid the need to remove observations with no spending on food away from home, we combine food at home and food away from into a food spending variable. Second, we use a measure of nondurable consumption that is narrower than that used in Skinner (who takes the sum of all spending, less mortgage interest, furniture and automobiles and including imputed spending on owner-occupied housing). This allows us to compare the results we obtain with our imputed spending measures with those in the PSID.

TABLE 2 Empirical Example: Log nondurable consumption on house values

	(1) PSID	(2) CE (RP)	(3) CE (RRP)	(4) CE + PSID (nlGMM)
log Income$_{t-3}$	0.047**	0.035*	0.047*	0.048*
	(0.017)	(0.016)	(0.021)	(0.021)
log Income$_{t-2}$	0.065**	0.043*	0.057*	0.045**
	(0.016)	(0.014)	(0.019)	(0.018)
log Income$_{t-1}$	0.040**	0.023	0.031	0.024
	(0.015)	(0.014)	(0.019)	(0.018)
log Income$_{t}$	0.109**	0.075**	0.101**	0.123**
	(0.022)	(0.020)	(0.027)	(0.026)
log Income$_{t+1}$	0.105**	0.073**	0.098**	0.098**
	(0.016)	(0.015)	(0.020)	(0.019)
log House value	0.114**	0.087**	0.116**	0.129**
	(0.016)	(0.016)	(0.021)	(0.020)
DLR F-stat			0.85	
J-statistic				9.32
χ^2 test (p value)			0.62	0.16
N	5407	5407	5407	5407

Note: Standard errors in parentheses. Standard errors are clustered at the individual level. Additional controls for age, age squared, family size, log of head's earnings (set to zero if earnings are zero), a dummy for head's earnings being greater than zero, and year dummies. Column (1) shows results using the measures of nondurable consumption contained in the PSID as the dependent variable. Column (2) uses the unscaled regression prediction procedure to impute consumption spending into the PSID from the CE survey. Column (3) shows results when nondurable consumption is imputed to the PSID from the CE using re-scaled regression prediction (RRP). Column (4) shows results when coefficients are estimated using nonlinear GMM.

*$p < 0.05$.

**$p < 0.01$.

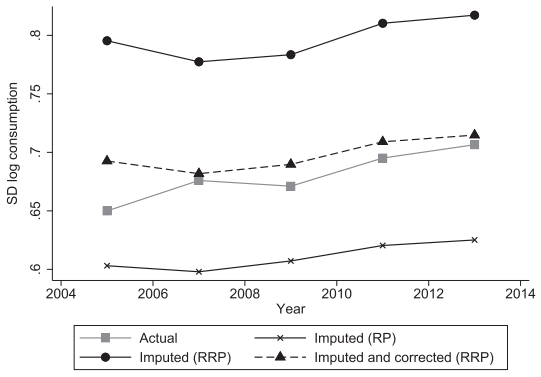


FIGURE 2 Standard deviation of log consumption. *Note*: Authors' calculations using the PSID. Lines show the standard deviation of log nondurable spending in the PSID ("Actual"), the standard deviation of imputed log consumption using regression prediction ("Imputed (RP)"), the standard deviation of imputed log consumption using re-scaled regression prediction ("Imputed (RRP)"), and the standard deviation of log consumption using rescaled regression prediction corrected using the relationship in Proposition 5 ("Imputed and corrected (RRP)")

The standard deviation of consumption spending shows similar trends in all three series. The fact that imputed and observed consumption move in similar ways over time is consistent with the findings of Attanasio and Pistaferri (2014) who use the latter as a check for the former in their analysis. The link between movements in the regression prediction and **RRP** imputed measures reflects the stability of the first stage R^2 over time.

We also note that the regression prediction measure tends to understate the *level* of consumption inequality, while **RRP** tends to overstate it. This was shown analytically in Section 2. This example reinforces the point that while **RRP** does not lead to biased estimates of regression coefficients, it does lead to biased estimates of the unconditional population mean and variance. When we apply the correction implied by Proposition 5 to the RRP estimate of the standard deviation, we obtain roughly the correct standard deviation. Once again, this suggests that the key assumption A6 is appropriate in this application.

5 | SUMMARY AND CONCLUSION

Although using regression prediction to impute the dependent variable in a regression model induces measurement errors "on the left," it is not necessarily innocuous. We have shown that the resulting Berkson errors in the dependent variable result in inconsistent estimates of the regression slope. This procedure has been much used to impute consumption to data sets with income or wealth, following a suggestion by Skinner (1987). We propose an alternative two-step estimator (**RRP**) which overcomes this inconsistency by rescaling by the first stage (imputation) R^2. Even then, we have shown that the usual OLS standard errors are not correct, but they can be corrected with estimable quantities.

Our results have use beyond the applications we demonstrate. For example, suppose a researcher has data with which to estimate a regression, but suspects that the dependent variables is measured with (possibly non-classical) error. The researcher also has a validation sample including the same dependent variable and a "gold-standard" measure of the dependent variable (for example from administrative data). This maps naturally into our set-up, treating the original measure of the dependent variable as the proxy, Z, and the "gold-standard" measure as y. If the assumptions of our framework hold, the **RRP** estimator is a consistent estimator of the regression parameters of interest.

Our analysis demonstrates that the preferred method of imputation may depend on the intended application. For example, it matters if the imputed variable will be a dependent variable or an independent variable, or whether the parameter of interest is a regression slope, or something else. This poses a challenge to data providers who may wish to include imputed variables in a standardised data set for multiple users.

Imputation of a dependent variable from a complementary data set is a potentially useful part of the applied econometrician's toolkit, but it must be done with care.

ACKNOWLEDGEMENTS

We thank Rob Alessie, Richard Blundell, Chris Bollinger, Mick Couper, Abhimanyu Gupta, Jörn-Steffen Pischke, Joachim Winter and participants in a workshop New Perspectives on Consumption Measures (STICERD LSE, 2016), for useful comments. Financial support from the ESRC through a grant to Essex University for "Understanding Household Finance through Better Measurement" (reference ES/N006534/1) is gratefully acknowledged. Crossley and Levell acknowledge support from the ESRC through the ESRC-funded Centre for Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies (grant reference ES/M010147/1). Crossley also acknowledges support through the Research Centre on Micro-Social Change (MiSoC) at the University of Essex, (reference ES/L009153/1) and Levell also acknowledges support through the grant "Advancing Microdata Models and Methods" (grant reference ES/P008909/1). The collection of PSID data used in this study was partly supported by the National Institutes of Health under grant number R01 HD069609 and R01 AG040213, and the National Science Foundation under award numbers SES 1157698 and 1623684.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results.

DATA AVAILABILITY STATEMENT

All data used to generate the results in this paper are available at http://qed.econ.queensu.ca/jae/datasets/crossley001/.

REFERENCES

Angrist, J. D., & Krueger, A. B. (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, *87*(418), 328–336.

Arellano, M., & Meghir, C. (1992). Female labour supply and on-the-job search: An empirical model estimated using complementary data sets. *The Review of Economic Studies*, *59*(3), 537–559.

Arrondel, L., Lamarche, P., & Savignac, F. (2015). Wealth effects on consumption across the wealth distribution: Empirical evidence. (*Working Paper 1817*): ECB Working Paper Series.

Attanasio, O., Hurst, E., & Pistaferri, L. (2015). The evolution of income, consumption, and leisure inequality in the United States, 1980–2010. In: Carroll, C. D., Crossley, T. F., & Sabelhaus, J. (Eds.), *Improving the measurement of consumer expenditures* (pp. 100–140). National Bureau of Economic Research, University of Chicago Press.

Attanasio, O., & Pistaferri, L. (2014). Consumption inequality over the last half century: Some evidence using the new PSID consumption measure. *American Economic Review: Papers & Proceedings*, *104*(5), 122–126.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, *45*(250), 164–180.

Blundell, R., Horowitz, J., & Parey, M. (2019). Estimation of a nonseparable heterogenous demand function with shape restrictions and berkson errors. (*Technical report*): Cemmap.

Blundell, R., Pistaferri, L., & Preston, I. (2004). Imputing consumption in the PSID using food demand estimates from the CEX. (*IFS Working Paper WP04/27*): The Institute for Fiscal Studies.

Blundell, R., Pistaferri, L., & Preston, I. (2008). Consumption inequality and partial insurance. *American Economic Review*, *98*(5), 1887–1921.

Bollinger, C. R., & Minier, J. (2015). On the robustness of coefficient estimates to the inclusion of proxy variables. *Journal of Econometric Methods*, *4*(1), 101–122.

Browning, M., Crossley, T. F., & Weber, G. (2003). Asking consumption questions in general purpose surveys. *Economic Journal*, *113*(491), F540–F567.

Charles, K. K., Danziger, S., Li, G., & Schoeni, R. (2014). The Intergenerational Correlation of Consumption Expenditures. *American Economic Review: Papers & Proceedings*, *104*(5), 136–140.

Cross, P. J., & Manski, C. F. (2002). Regressions, short and long. *Econometrica*, *70*(1), 357–368.

Crossley, T. F., Levell, P., & Stavros, P. (2021). Online supplemental appendix to regression with an imputed dependent variable. *Journal of Applied Econometrics*.

David, M., Little, R. J. A., Samuhel, M. E., & Triest, R. K. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association*, *81*(393), 29–41.

Davidson, R., & MacKinnon, J. G. (1988). Double length artificial regressions. *Oxford Bulletin of Economics and Statistics*, *50*(2), 203–217. https://ideas.repec.org/a/bla/obuest/v50y1988i2p203-17.html

Dumont, M., Rayp, G., Thas, O., & Willemé, P. (2005). Correcting standard errors in two-stage estimation procedures with generated regressands. *Oxford Bulletin of Economics and Statistics*, *67*(3), 421–433.

Feenstra, R. C., & Hanson, G. H. (1999). The impact of outsourcing and high-technology capital on wages: Estimates for the United States, 1979-1990. *The Quarterly Journal of Economics*, *114*(3), 907–940.

Fisher, J., Johnson, D., Latner, J. P., Smeeding, T., & Thompson, J. (2016). Inequality and mobility using income, consumption, and wealth for the same individuals. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *2*(6), 44–58.

Guvenen, F., & Smith, A. A. (2014). Inferring labor income risk and partial insurance from economic choices. *Econometrica*, *82*(6), 2085–2129.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*(4), 1029–1054. http://www.jstor.org/stable/1912775

Hoderlein, S., & Winter, J. (2010). Structural measurement errors in nonseparable models. *Journal of Econometrics*, *157*(2), 432–440.

Hyslop, D. R., & Imbens, G. W. (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics*, *19*(4), 475–481.

Inoue, A., & Solon, G. (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, *92*(3), 557–561.

Kaplan, G., Mitman, K., & Violante, G. L. (2020). Non-durable consumption and housing net worth in the great recession: Evidence from easily accessible data. *Journal of Public Economics*, *189*, 104176.

Klevmarken, A. (1982). Missing variables and two-stage least-squares estimation from more than one data set. (*Working Paper Series 62*): Research Institute of Industrial Economics.

Lillard, L., Smith, J. P., & Welch, F. (1986). What do we really know about wages: The importance of non-reporting and census imputation. *Journal of Political Economy*, *94*(3), 489–506.

Lubotsky, D., & Wittenberg, M. (2006). Interpretation of regressions with multiple proxies. *The Review of Economics and Statistics*, *88*(3), 549–562.

Lusardi, A. (1996). Permanent income, current income, and consumption: Evidence from two panel data sets. *Journal of Business & Economic Statistics*, *14*(1), 81–90.

Mankiw, G. N., & Shapiro, M. D. (1986). News or noise: An analysis of GNP revisions. *Survey of Current Business*, *66*, 20–25.

Meyer, B., & Sullivan, J. X. (2003). Measuring the well-being of the poor using income and consumption. *Journal of Human Resources*, *38*(4), 1180–1220.

Mulligan, C. (1999). Galton versus the human capital approach to inheritance. *Journal of Political Economy*, *107*(S6), 184–224.

Pacini, D., & Windmeijer, F. (2016). Robust inference for the two-sample 2SLS estimator. *Economics Letters*, *146*(C), 50–54.

Palumbo, M. (1999). Uncertain medical expenses and precautionary saving near the end of the life cycle. *The Review of Economic Studies*, *66*(2), 395–421.

Panel Study of Income Dynamics (2019). Public Use Dataset. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Pischke, J.-S. (1995). Measurement error and earnings dynamics: Some estimates from the PSID validation study. *Journal of Business & Economic Statistics*, *13*(3), 305–314.

Ridder, G., & Moffitt, R. (2007). The econometrics of data combination. *Handbook of Econometrics*, *6*, 5469–5547.

Schulhofer-Wohl, S. (2011). Heterogeneity and tests of risk sharing. *Journal of Political Economy*, *119*(5), 925–958.

Shea, J. (1997). Instrument relevance in multivariate linear models: A simple measure. *The Review of Economics and Statistics*, *79*(2), 348–352.

Sims, C. A. (1972). Comments and rejoinder (on Okner (1972)). *Annals of Economic and Social Measurement*, *1*(3), 343–345.

Skinner, J. (1987). A superior measure of consumption from the panel study of income dynamics. *Economics Letters*, *23*(2), 213–216.

Skinner, J. (1989). Housing wealth and aggregate saving. *Regional Science and Urban Economics*, *19*(2), 305–324.

Wansbeek, T., & Meijer, E. (2000). *Measurement Error and Latent Variables in Econometrics*. Amsterdam: Elsevier.

Wooldridge, J. M. (1991). On the application of robust, regression-based diagnostics to models of conditional means and conditional variances. *Journal of Econometrics*, *47*(1), 5–46.
Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of the article.

How to cite this article: Crossley, T. F., Levell, P., & Poupakis, S. (2022). Regression with an imputed dependent variable. *Journal of Applied Econometrics*. 1–18. https://doi.org/10.1002/jae.2921