Institute for Fiscal Studies

Regression with an Imputed Dependent Variable

IFS Working Paper W19/16

Thomas F. Crossley
Peter Levell
Stavros Poupakis

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Regression with an Imputed Dependent Variable

Thomas F. Crossley

European University Institute, University of Essex, Institute for Fiscal Studies and ESCoE

Peter Levell

Institute for Fiscal Studies and University College London

Stavros Poupakis

University College London

June, 2019

Abstract

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 an easily-implemented correction and correct asymptotic standard errors. We illustrate these with Monte Carlo experiments and empirical examples using data from the US Consumer Expenditure Survey (CE) and the Panel Study of Income Dynamics (PSID).

Keywords: Imputation; Measurement error; Consumption.

JEL codes: C81, C13, E21

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). 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.

1 Introduction

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 which 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).[1] In this paper we consider the consequences of estimating a regression with an imputed dependent variable, and how those

[1] For panel data on consumption, an alternative approach is to invert the inter-temporal budget constraint and calculate spending as income minus saving where the latter is often approximated by changes in wealth. This was initially suggested by Ziliak (1998) for the PSID, but has more recently been adopted for administrative (tax) data on income and wealth (Browning et al., 2003). While attractive this procedure has several drawbacks. First it identifies only total household spending, and in many applications the distinctions between consumption spending, nondurable consumption and household investment spending are important. Second, in the case of our motivating example, this procedure results in income or wealth being on both the right and left-hand side of the equation so that any measurement error in income or wealth can cause quite serious problems (Browning et al., 2014). Baker et al. (2018) show that even with administrative data on income and wealth there can be significant measurement error in implied spending.

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 R^2 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 **RRP**) procedure. We then show that with a single proxy, the **RRP** 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 **RRP** 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.

In the next section we layout our basic framework, and derive the main results. We also relate our results to the prior literature, including Lusardi (1996), who combines CE consumption data with PSID income data using the 2-sample IV approach proposed by Klevmarken (1982) and Angrist and Krueger (1992). We clarify the relationship between that approach and the imputation procedures we study.

Section 3 takes up the question of inference. We show that the usual OLS standard errors from a regression of an imputed dependent variable (derived from the **RRP** or **BPP** procedures) are too small, and provide an estimator of the correct asymptotic standards errors of the regression coefficient of interest. Section 4 illustrates our main points with a Monte Carlo study, and Section 5 provides two empirical examples using the CE and PSID. Section 6 concludes.

2 Set-up And Main Results

Consider estimating the regression

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

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. Assume that the usual regression assumptions hold, so that the vector β could be consistently estimated by Ordinary Least Squares (OLS) if we had complete data. In particular,

Assumption A1. For any representative sample j (of size n_j):

a. $y_j = X_j\beta + \epsilon_j$

b. $plim\left(\frac{1}{n_j}X_j'X_j\right) = \Sigma_{XX}$, which is of full rank (K).

c. $plim\left(\frac{1}{n_j}X_j'\epsilon_j\right) = 0$

d. $plim\left(\frac{1}{n_j}\epsilon_j'\epsilon_j\right) = \sigma_\epsilon^2 > 0$

However, we have no data that allows us to calculate the empirical analogue $(\frac{1}{n_j}X_j'y_j)$ of the population covariances Σ_{Xy}. Subscripts $j = 1, 2, ...$ index the data set (or sample); absence of a subscript indicates a population quantity. We do have a sample of size n_1 of data on (y_1, Z_1) and a second sample of size n_2 of data on (X_2, Z_2). Z_j is a $L \times n_j$ matrix of proxies $(l = 1, ..., L)$ for y; if we have only a single proxy (a vector) we denote it by z; one of a set of multiple proxies is denoted by z_l. Both data sets are independent, random samples from the population of interest. 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.

We posit a relationship between our proxies and the dependent variable of interest. In our motivating example with total nondurable consumption as our quantity of interest and food consumption as a single proxy, this relationship is an Engel Curve in the form of $z = y\gamma + u$. This implies a reduced form relationship between z and X, such as

$$z = X\beta\gamma + \epsilon\gamma + u. \tag{2}$$

With more than one proxy we have a set of relationships between the proxies and y

$$Z = y\gamma + u. \tag{3}$$

where γ is $1 \times L$ and u is $n \times L$. This in turn implies a set of reduced form relationships:

$$Z = X\beta\gamma + \epsilon\gamma + u. \tag{4}$$

Note that Equation (4) makes clear that Z *must* depend on ϵ: Z has some information about y that is not contained in X. This is why we refer to Z as proxy. Given Z with these properties, one can *impute y using Z*. For clarity of exposition, we begin with the cross-sectional case and abstract from additional covariates in the Engel curve. Our assumptions regarding the proxy or proxies are collected in A2.

Assumption A2. For any representative sample j (of size n_j):

a. $Z = y\gamma + u$, with $\gamma_l \neq 0 \; \forall \; l$

b. $plim \left(\frac{1}{n_j} Z_j' Z_j \right) = \Sigma_{ZZ}$, which is of full rank (L).

c. $plim \left(\frac{1}{n_j} X' u \right) = 0$

d. $plim \left(\frac{1}{n_j} y' u \right) = 0$

e. $plim \left(\frac{1}{n_j} u' u \right) = \Sigma_{uu}$, with diagonal elements strictly positive.

Assumption (A2d) would fail, for example, if there was measurement error in y. Below, we take up all of: additional covariates in the Engel curve, panel data, and measurement error in y.

A final assumption we make in our analysis is that the ratio of the sizes of our two samples approaches a positive constant as n_1 tends to infinity.

Assumption A3. $\lim_{n_1 \to \infty} \frac{n_1}{n_2} = \alpha$ for some $\alpha > 0$.

2.1 Alternative Imputation Strategies

Skinner (1987) suggested regressing y_1 on Z_1 in the CE and using the resulting coefficients to predict $\hat{y}_2$ in the PSID (and then regressing $\hat{y}_2$ on X). Note that with a single spending category as the proxy, the first stage here is an "inverse" Engel curve. This **RP** procedure was advocated by Browning et al. (2003) and recent applications include Attanasio and Pistaferri (2014), Arrondel et al. (2015) and Kaplan et al. (2016).[2] Alternatively, Blundell et al. (2004, 2008), again using the CE and PSID, first regress z_1 (food spending) on y_1 then predict $\hat{y}_2 = z_2 \frac{1}{\hat{\gamma}}$. That is, they estimate an Engel curve and then invert it to predict consumption. This is the **BPP** procedure and it has also recently been employed by Attanasio et al. (2015). Finally, an alternative is to not impute consumption at the household level at all, but to recover the parameter of interest (β) from a combination of moments taken from the two surveys. This was first suggested (for a different application) by Arellano and Meghir (1992) (hereafter **AM**). Here, (again with a single proxy) one could regress z_1 on y_1 to get $\hat{\gamma}$, then regress z_2 on X_2 to get $\widehat{\beta\gamma}$ in Equation (4), and take ratio of the two to estimate β.

We first consider the **RP** procedure (with possibly multiple proxies). Regression of $\hat{y}_2^{RP}$ on X does not give a consistent estimate of β.

[2] Kaplan et al. (2016) regress county-level consumption spending on local house prices in the US. Since data on total nondurable consumption is not available at county level, they use county-level data on a subset of nondurable expenditures (grocery spending) from the Kilts-Nielsen Retail Scanner Dataset (KNRS) as their dependent variable, and then scale up their coefficients using household-level data on the relationship between grocery and total spending from the CE Survey. This is analogous to our set-up in a case where the regression of interest is $y_c = X_c\beta + \epsilon_c$ (where the subscript c denotes county), the first stage regression is $z_h = \gamma y_h + u_h$ (where h denotes a household), and where the researcher takes the additional step of projecting z_h onto a set of county dummies to obtain z_c (and then proxying y_c using an estimate of $\frac{1}{\gamma} z_c$). Kaplan et al. (2016) additional include controls in the first stage regression (such as age) that are not used to impute y_c in their main regression. This needn't cause a problem so long as the county-level averages of these variables are conditionally uncorrelated with X_c.

Proposition 1. *Assuming that both samples are random samples of the same population, and (A1a), (A1b), (A1c), (A2a), (A2c), (A2d) and (A3) hold,*

$$plim\left(\hat{\beta}^{RP}\right) = \beta\phi_{y,Z} \tag{5}$$

where $\phi_{y,Z}$ is the population R^2 from a regression of y on Z $(0 \ < \ \phi_{y,Z} \ < \ 1)$.

Proof.

$$\begin{aligned}
plim\left(\hat{\beta}^{RP}\right) &= plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'Z_2}{n_2}\left(\frac{Z_1'Z_1}{n_1}\right)^{-1}\frac{Z_1'y_1}{n_1}\right\} \\
&= plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'Z_2}{n_2}\left(\frac{Z_1'Z_1}{n_1}\right)^{-1}\frac{Z_1'y_1}{n_1}\frac{1}{R^2_{y_1,Z_1}}R^2_{y_1,Z_1}\right\} \\
&= plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'Z_2}{n_2}\left(\frac{Z_1'Z_1}{n_1}\right)^{-1}\frac{Z_1'y_1}{n_1}\left[\frac{y_1'Z_1}{n_1}\left(\frac{Z_1'Z_1}{n_1}\right)^{-1}\frac{Z_1'y_1}{n_1}\right]^{-1}\frac{y_1'y_1}{n_1}R^2_{y_1,Z_1}\right\} \\
&= \beta\gamma\Sigma_{ZZ}^{-1}\gamma'\Sigma_{yy}\left(\Sigma_{yy}\gamma\Sigma_{zz}^{-1}\gamma'\Sigma_{yy}\right)^{-1}\Sigma_{yy}\phi_{y,Z} \\
&= \beta\gamma\Sigma_{ZZ}^{-1}\gamma'\left(\gamma\Sigma_{ZZ}^{-1}\gamma'\right)^{-1}\phi_{y,Z} = \beta\phi_{y,Z}
\end{aligned}$$

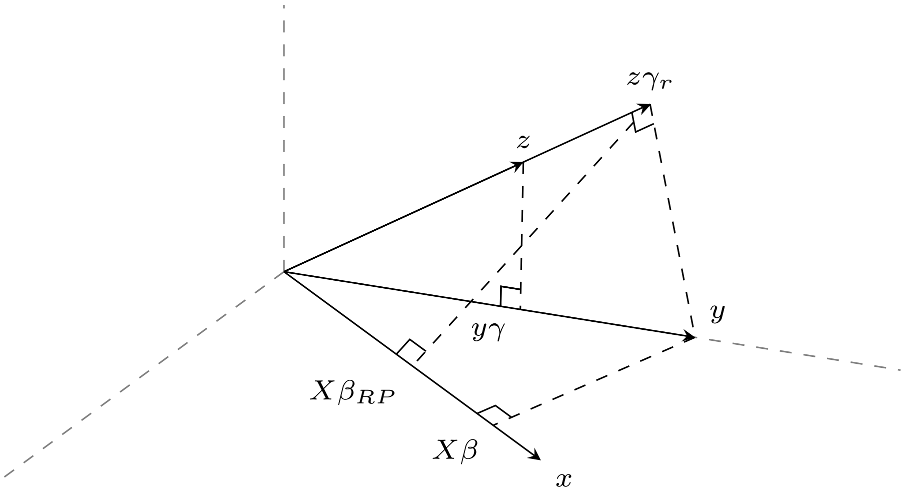
where $\Sigma_{ZZ} = plim\left(\frac{Z_1'Z_1}{n_1}\right)$, the scalar $\Sigma_{yy} = plim\left(\frac{y_1'y_1}{n_1}\right)$, and $R^2_{y_1,Z_1}$ is the sample R^2. □

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 *centered* sample R^2, $\phi_{y,Z}$ is the *centered* population R^2 and the result holds without demeaning the data.

Figure 1 gives a geometric intuition for the problem. The solid vectors y, z and X represent data. 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 **RP** procedure first projects the y onto z, giving $\hat{y} = z\gamma$, and then projects this vector onto X giving $X\beta^{RP}$. Note that $X\beta^{RP} < X\beta$.

The source of the problem is that regression prediction results in a $\hat{y}$ that differs from y by a prediction error or Berkson measurement error, that is uncorrelated with z but not

Figure 1: RP Imputation Procedure as Projections



uncorrelated with y and, in general, not uncorrelated with X. As is well known, classical measurement in an independent variable causes bias in linear regression, but classical measurement errors in the dependent variable does not. 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 recognized that Berkson errors in an independent variable does not cause bias in a linear regression (Berkson, 1950; Wansbeek and Meijer, 2000). What is less frequently recognized is that Berkson errors in a dependent variable do cause bias.

The same problem arises with the **RP+** procedure. The true value of (unobserved) y_2 can be decomposed into its projection onto Z and an orthogonal error

$$y_2 = \hat{y} + \hat{v}_2. \tag{6}$$

Consider then drawing a random residual from the first stage regression to create a stochastic imputation

$$\hat{\hat{y}}_2 = \hat{y} + \hat{v}_1 = y_2 - \hat{v}_2 + \hat{v}_1. \tag{7}$$

Then $\hat{\hat{y}}_2$ differs from y_2 by the error $\hat{v}_1 - \hat{v}_2$ which is by construction orthogonal to Z_1, but not y_2 or X_2.[3]

Again the degree of attenuation (and hence bias) in the **RP** procedure depends on the first stage population R^2 ($\phi_{y,Z}$). In our motivating example, R^2s for food Engel curves are typically between 50 and 70%, implying inflation factors of between 1.4 and 2 (or attenuation of between 30 and 50%).

As the attenuation in the **RP** procedure is an estimable quantity, it can be corrected. One can rescale $\hat{y}^{RP}$ by the estimated first stage (centered) $R^2_{y_1,Z_1}$, or, equivalently, rescale $\hat{\beta}^{RP}$ by the estimated first stage (centered) $R^2_{y_1,Z_1}$. We refer to this procedure as "Re-scaled Regression Prediction" (hereafter **RRP**), with the rescaled impute of y_2 denoted $\hat{y}_2^{RRP}$ and the resulting estimate of β denoted $\hat{\beta}^{RRP}$.

Proposition 2. *Assuming that both samples are random samples of the same population and (A1a), (A1b), (A1c), (A2a), (A2c), (A2d) and (A3) hold,*

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

Proof. Follows immediately from Proposition 1. □

Finally, consider the **BPP** and **AM** procedures, with resulting estimates $\hat{\beta}^{BPP}$ and $\hat{\beta}^{AM}$.

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

Proof. We have

$$\begin{aligned}\hat{\beta}^{RRP} &= (X_2^{'}X_2)^{-1}(X_2^{'}z_2)(z_1^{'}z_1)^{-1}(z_1^{'}y_1)\left[(y_1^{'}z_1)(z_1^{'}z_1)^{-1}(z_1^{'}y_1)\right]^{-1}y_1^{'}y_1 \\ &= (X_2^{'}X_2)^{-1}X_2^{'}z_2(y_1^{'}z_1)^{-1}y_1^{'}y_1 = \hat{\beta}^{BPP}. \end{aligned} \tag{8}$$

Thus, under the assumptions listed in Proposition 2, $\hat{\beta}^{BPP}$ is also consistent.

[3] Note that, because it is randomly drawn from a separate random sample, $\hat{v}_1$ is orthogonal to y_2. The problem with the composite error $\hat{v}_1 - \hat{v}_2$ lies in the prediction error $\hat{v}_2$.

The **AM** procedure takes the ratio of $\widehat{\beta\gamma} = (X_2^{'}X_2)^{-1}X_2^{'}z_2$ and $\hat{\gamma} = (y_1^{'}y_1)^{-1}y_1^{'}z_1$, to give $\hat{\beta}^{AM} = \widehat{\beta\gamma}/\hat{\gamma}$.

$$\begin{aligned}\hat{\beta}^{AM} &= \widehat{\beta\gamma}/\hat{\gamma} = (X_2^{'}X_2)^{-1}X_2^{'}z_2\left[(y_1^{'}y_1)^{-1}y_1^{'}z_1\right]^{-1} \\ &= (X_2^{'}X_2)^{-1}X_2^{'}z_2(y_1^{'}z_1)^{-1}y_1^{'}y_1 = \hat{\beta}^{RRP} = \hat{\beta}^{BPP}. \qquad (9)\end{aligned}$$

□

Consistency of $\hat{\beta}^{AM}$ follows either directly from the Slutsky theorem or by numerical equivalence to $\hat{\beta}^{RRP}$ and $\hat{\beta}^{BPP}$.

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) and Attanasio and Pistaferri (2014) study consumption inequality. We continue with the case of a single proxy to allow comparison of **BPP** to **RP** and **RRP**, and consider the case of a single x variable for ease of exposition (though the results extend naturally to a vector X). **AM** recovers β directly, and does not generate unit level estimates of y. The imputes $\hat{y}^{RP}$ and $\hat{y}^{RRP}$ are numerically different,

$$\hat{y}^{RP} = z_2(z_1^{\prime}z_1)^{-1}z_1^{\prime}y_1, \qquad (10)$$

$$\hat{y}^{RRP} = z_2(z_1^{\prime}z_1)^{-1}z_1^{\prime}y_1/R^2_{y_1,z_1} \qquad (11)$$

Algebra analogous to the proof of Proposition 3 shows that $\hat{y}^{RRP}$ and $\hat{y}^{BPP}$ 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 (A2a).

Denote population moments by $plim(\frac{1}{n}\sum y^2) = \sigma_{yy}$ and $plim(\frac{1}{n}\sum yx) = \sigma_{yx}$, again recalling that variables have been de-meaned. Denote sample moments based on $\hat{y}^{RP}$ by s_{yy}^{RP}

and s_{yx}^{RP}; and analogously for $\hat{y}^{RRP}$ and $\hat{y}^{BPP}$,

$$s_{yy}^{RP} = \frac{1}{n_2}\hat{y}^{RP\prime}\hat{y}^{RP} = \frac{1}{n_2}z_1'y_1(z_1'z_1)^{-1}z_2'z_2(z_1'z_1)^{-1}z_1'y_1, \tag{12}$$

$$plim\left(s_{yy}^{RP}\right) = \frac{(\gamma\sigma_{yy})^2}{\gamma\sigma_{yy}+\sigma_{uu}} = \sigma_{yy}\times\phi_{y,z}, \tag{13}$$

where again $\phi_{y,Z}$ is the population R^2 from the first stage regression. The sample variance of $\hat{y}^{RP}$ underestimates the population variance of y. A similar calculation gives:

$$plim\left(s_{yx}^{RP}\right) = \sigma_{yx}\times\phi_{y,z}. \tag{14}$$

Note that with a scalar x the OLS estimate of β is just s_{yx}^{RP}/s_{yy}^{RP} and this gives an additional intuition for the inconsistency of $\hat{\beta}^{RP}$ as an estimator of β: s_{yx}^{RP} is not a consistent estimator of σ_{yx}. Moreover, adding a residual to $\hat{y}^{RP}$, (the **RP+** procedure) does not correct this.

For the rescaled impute $\hat{y}^{RRP}$, it follows from Equations (13) and (14) and the definition of $\hat{y}^{RRP}$ that

$$plim\left(s_{yy}^{RRP}\right) = \sigma_{yy}/\phi_{y,z} \tag{15}$$

and

$$plim\left(s_{yx}^{RRP}\right) = \sigma_{yx}. \tag{16}$$

Continuing with the bivariate regression intuition, the **RRP** procedure is consistent for β because it is consistent for σ_{yx}.

Finally, simple algebra establishes that

$$s_{yy}^{RRP} = s_{yy}^{BPP} \tag{17}$$

and

$$s_{yx}^{RRP} = s_{yx}^{BPP}, \tag{18}$$

This follows from the numerical equivalence of the de-meaned values of $\hat{y}^{RRP}$ and $\hat{y}^{BPP}$. Thus $plim\ s_{yy}^{BPP} = plim\ s_{yy}^{RRP} > \sigma_{yy} > plim\ s_{yy}^{RP}$. Turning again to our motivation 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 sample of $\hat{y}^{RP}$ gives an consistent (and unbiased) estimate of the population mean of y. However, if the **RPP** procedure is implemented by rescaling $\hat{y}^{RP}$ (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 regressand and for estimating quantities that depend on the first moment of y (poverty rates, for example).

Table 1: Summary of Imputation Methods (Consistency)

	μ_y	σ_{yy}	β
Regression Prediction (**RP**)	✓	×	×
Regression Prediction + $\hat{e}$ (**RP+**)	✓	✓	×
Rescaled Regression Prediction (**RRP**)	×	×	✓
Blundell et al., 2004; 2008 (**BPP**)	✓	×	✓
Arellano and Meghir, 1992 (**AM**)	-	-	✓

Notes: a ✓indicates that the procedure given by the row leads to a consistent estimate of the population parameter given by the column (μ_y, σ_{yy} or β). A × indicates that the procedure leads to an inconsistent estimate of the relevant parameter, and a dash indicates that the procedure does not provide an estimate via the analogous sample moment. The table assumes that the **RPP** procedure is implemented by rescaling $\hat{y}^{RP}$ (rather than rescaling β^{RP}).

Table 1 summarizes these consistency results. For the case of a single proxy any of the **RRP**, **BPP** and **AM** procedures give a consistent estimate of a regression coefficient β, but for estimating unconditional moments, imputations from **RP**, **BPP** and especially **RP+** are preferable.

2.2 *Hot-deck Imputation and Item-nonresponse*

We have taken the imputation of total consumption expenditure to an income or wealth survey with a continuous proxy (food expenditure) as our motivation and running example. However, the problem we highlight with regression prediction is more general. In particular, Lillard et al. (1986) and David et al. (1986) note that commonly employed hot-deck imputation procedures can be interpreted as regression predication plus an added residual. Such procedures draw a matched observation, $\hat{\hat{y}}_1$, of the missing variable from a cell defined by categorical variables (possibly generated by grouping continuous variables). $\hat{\hat{y}}_1$ can be viewed as a prediction using the coefficients from a saturated first stage regression on those categorical covariates (that is, one with a full set of interactions), plus a residual from the first stage regression,

$$\hat{\hat{y}} = \hat{y} + \hat{v}_1. \tag{19}$$

As above, the true value of (unobserved) y_2 can be decomposed into its projection onto the categorical matching variables and an orthogonal error

$$y_2 = \hat{y} + \hat{v}_2. \tag{20}$$

Then $\hat{\hat{y}}$ differs from y_2 by the error $\hat{v}_1 - \hat{v}_2$ which is by construction orthogonal to the matching variables, but not to y. Thus if the matching variables include some variables (proxies, Z) that are *not* included among the independent variables (X) in the regression of interest, the hot-deck procedure is identical to the **RP+** procedure described above, and our results apply.

We have also focused on the data combination problem: no single data set contains data on both y and X, but we do have data on (y_1, Z_1) and (X_2, Z_2). In this case y_2 must be *fully* imputed. However, the case of partial imputation, typically because of item-nonresponse, is also of interest. Suppose we have a data set (y_j, X_j, Z_j) with n cases but for a fraction m of cases y_j is missing at random (Rubin, 1976). For ease of exposition consider a single

x variable. Suppose we use a regression on Z_j or a hot deck procedure matching on Z_j to impute y_j from the complete cases to the missing cases (within the same data set). Reorder the data so that the cases with observed values for y_j come before the cases with imputed values for y_j. Denote the now "complete" y-vector by $[y : \hat{y}]_j$. The regression of $[y : \hat{y}]_j$ on x_j gives

$$\hat{\beta} = \left(\sum x_j^2\right)^{-1} \left(\sum_{1}^{mn_j} x_j y_j + \sum_{mn_j+1}^{n_j} x_j \hat{y}_j\right). \tag{21}$$

Using the results above it is easy to show that

$$plim\left(\hat{\beta}\right) = (1-m)\beta + m\beta\phi_{y,Z} = \beta\left(1 + m\left(\phi_{y,Z} - 1\right)\right). \tag{22}$$

Partial imputation with a proxy will suffer from the attenuation that we have highlighted for the case of full imputation, but with the bias depending on the fraction of cases with missing data (m), as well as the first stage population R^2.

2.3 Practicalities

The analysis above is trivially extended to handle additional covariates. If additional covariates W are added to both the first stage regression and regression of interest, then the results above hold by straightforward application of the Frisch-Waugh-Lovell theorem (y, X and Z can be "residualized" and then the results apply directly to the residualized variables). There are two points to note: First, the additional covariates W must be added to both the first stage regression and regression of interest. Second, if covariates are added, then the relevant first stage R^2 is the *partial* R^2 associated with Z.

Often a researcher will want to estimate a panel version of Equation (1): $\Delta y = \Delta X\beta + \Delta\epsilon$ where $\Delta y = y^1 - y^0$ and superscripts denote time (and similarly for X and ϵ). As before β is the main object of interest and could be estimated consistently by OLS if we had complete data (that is, $plim(\Delta X \times \Delta\epsilon) = 0$). Suppose we have no data from which to compute $\frac{1}{n}\sum \Delta y \times \Delta X$, but do have have some data on $(y_1^1, Z_1), (y_2^0, Z_2), (\Delta X_3, Z_3)$. In our running example, one often wants to estimate the effect of income or wealth changes on consumption and the available data would be a repeated cross-sectional household budget survey combined with a panel survey on income or wealth. Then y_3 can be imputed year by year. It is straightforward extension of the results above to show that $\hat{\beta}^{RRP}$ is consistent in this case, and with one proxy $\hat{\beta}^{BPP}$ remains numerically identical to $\hat{\beta}^{RRP}$.

Finally, suppose that y is measured with error. This would be a natural concern in our running example, as consumption expenditure is a difficult quantity to measure, even in a detailed household budget survey. Even if this measurement error is classical, it is obvious that the both the **BPP** and **AM** procedures require an instrument for y, as both involve a regression on Z_1 on y_1 to get $\hat{\gamma}$. If $plim(y_1' u_1) \neq 0$ because u_1 contains the measurement error in y_1, then an instrument for y is required to obtain a consistent estimate of γ. With the **RRP** procedure, y_1 is the independent variable in the first stage imputation regression, so that classical measurement error in y_1 does not lead to an inconsistent estimate of the regression slope. However, classical measurement error in y_1 leads to an inconsistent estimate of the population first stage R^2 $(\phi_{y,Z})$. With a single proxy, this can be overcome by estimating $\phi_{y,Z}$ as the product of the Engel curve and inverse Engel curve regression slopes, where the latter can be estimated by OLS but the former must be estimated by IV (because y_1 is the independent variable).

2.4 Related Literature

In this paper we study the use of proxies to predict a dependent variable.[4] 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 innocuous cases of Berkson measurement error in an independent

[4] 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.

variable, or classical measurement error in a dependent variable.[5] 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.[6]

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

$$z = Y\gamma + u \tag{23}$$

and the second stage is not (1) but rather

$$Y^k\gamma^k = X\beta^k + \epsilon^k \tag{24}$$

where the k superscript denotes the kth element of a vector. Here $Y^k\gamma^k$ is not observed and so is replaced by the first stage estimate $Y^k\hat{\gamma}^k$. Of course the vector $\hat{\gamma}$ differs from γ by an estimation error $(Y'Y)^{-1}Y'\hat{u}$, but, given the set-up, the stochastic element $\hat{u}$ is orthogonal to Y, and so also X, and thus causes problems for inference but not bias. Although the

[5] Berkson measurement error in an independent variable is also a problem in nonlinear models. See for example Blundell et al. (2019).

[6] They illustrate their results using self-reported data on consumption expenditure.

motivation and second-stage regressand are different, this procedure essentially regresses z on Y, analogously to the **BPP** procedure, rather than y on Z as in the **RP** procedure, so the Berkson measurement error problem does not arise.

We now relate our results to two further literatures: on item nonresponse and partial imputation (Hirsch and Schumacher (2004) and Bollinger and Hirsch (2006)) and on 2-sample Instrumental Variables and 2SLS procedures (Klevmarken (1982), Angrist and Krueger (1992)). To do so, it is useful to consider a more general set up than the one analyzed above (A1, A2, and A3). We retain assumptions (A1 and A3) and continue to treat x as the exogenous independent variable of interest, but replace (A2a) with

$$Z = X\theta + \nu \tag{25}$$

where

$$plim\left(\frac{1}{n_j}x_j'\epsilon_j\right) = plim\left(\frac{1}{n_j}x_j'\nu_j\right) = 0 \tag{26}$$

and

$$plim\left(\frac{1}{n_j}\epsilon_j'\nu_j\right) = \Sigma_{\epsilon\nu}. \tag{27}$$

Note that $\Sigma_{\epsilon\nu}$ is $L \times 1$ with elements $\sigma_{\epsilon\nu_l}$. If $\sigma_{\epsilon\nu_l} \neq 0$ then z_l predicts variation in y that is not predicted by X, and so z_l is in that sense a proxy.

The set-up studied above (A1, A2 and A3) is a special case of this more general set-up, with the following restrictions

$$\theta = \beta\gamma, \tag{28}$$

and

$$\nu = \epsilon\gamma + u, \tag{29}$$

so that

$$\Sigma_{\epsilon\nu} = \gamma\sigma_{\epsilon\epsilon} \tag{30}$$

and

$$\Sigma_{\nu\nu} = \gamma^{'}\gamma\sigma_{\epsilon\epsilon} + \Sigma_{uu} \tag{31}$$

This restricted set-up can be motivated by economic theory in some applications (for example, by two-stage budgeting in the case of food consumption and total consumption, or more generally by the idea that z_l is a simple indicator for y, and affected by X only *through* y).

Returning to the more general set-up, the **RP** procedure is

$$\hat{\beta}^{RP} = \left(\frac{X_2^{'}X_2}{n_2}\right)^{-1}\frac{X_2^{'}Z_2}{n_2}\left(\frac{Z_1^{'}Z_1}{n_1}\right)^{-1}\frac{Z_1^{'}y_1}{n_1}. \tag{32}$$

Then,

$$plim(\hat{\beta}^{RP}) = \theta\ \left(\theta^{'}\Sigma_{XX}\theta + \Sigma_{\nu\nu}\right)^{-1}\left(\theta^{'}\Sigma_{XX}\beta + \Sigma_{\nu\epsilon}\right) \tag{33}$$

In the more restricted case studied above (A1 and A2), Equation (33) reduces to Proposition 1. Thus the inconsistency $\hat{\beta}^{RP}$ arising from the Berkson measurement errors in the regressand is quite general, but the **RRP**, **BPP** and **AM** solutions depend on the specific structure assumed in equations (A1a), (A1b), (A1c), (A2a), (A2c) and (A2d).

Consider an alternative restriction on Equation (25) where $Z = X\theta$ (so that all elements of $\Sigma_{\epsilon\nu}$ and $\Sigma_{\nu\nu}$ are zero.) Then from Equation (33)

$$plim\left(\hat{\beta}^{RP}\right) = \beta. \tag{34}$$

This special case, in which Z is linear combination of X, is ruled out above by strictly positive asymptotic variances in (A1d) and (A2e). It is also not very interesting in data combination problem that is our main focus. However, it is potentially more interesting in the the case partial imputation of a y vector in response to item nonresponse. Hirsch and Schumacher (2004) and Bollinger and Hirsch (2006) study the case of partial imputation of a y vector using a hot-deck procedure matched on a subset of the X variables in the main regression of interest. They show that this leads to biased estimates of the regression coefficients of interest. As noted above, hot-deck procedures map onto the **RP+** procedure we describe. However, we study the case in which imputation is based (additionally) on variables (the proxies Z) that are excluded from X. Interestingly, both analyses have an unbiased limit case in which the proxies Z span X. Taking our results and theirs together demonstrates that (partial) imputation of y will lead to inconsistent regression coefficient if the variables Z either predict variation in y that is not predicted by X or fail to predict variation in y that is explained by variation in X. Counter-intuitively, the only procedure that is consistent for β without a correction is to predict y with the same variables that will be used in the second stage regression (effectively, to regress a particular linear combination of X on itself).

It is also useful to contrast the imputation procedures studied in this paper with the 2-sample IV (2SIV) and 2-sample 2SLS approaches (Klevmarken (1982), Angrist and Krueger (1992) and Inoue and Solon (2010)) applied to the combination of CE consumption data and PSID income data by Lusardi (1996). In the general set-up above suppose that $\sigma_{\epsilon\nu} = 0$ so that Z is related to y only through X. Thus Z is not a proxy in the sense given above. From Equation (33) the **RP** procedure remains inconsistent for β.

However, the 2-sample-2SLS estimator is

$$\hat{\beta}^{2S2SLS} = (\hat{X_1}'\hat{X_1})^{-1}\hat{X_1}'y_1 \tag{35}$$

where $\hat{X}_1 = Z_1(Z_2'Z_2)^{-1}Z_2'X_2$. It is straightforward to show that if a standard rank condition holds, and under the assumption that $\sigma_{\epsilon\nu} = 0$, the 2-sample 2SLS estimator 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 the key assumption is that Z affects y only through X, which is the polar opposite to the assumption that Z is a proxy or proxies (as noted above, a useful proxy must have information about y over and above the information in X). 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).[7] An additional virtue of this procedure is that inherent measurement error in y poses no additional difficulties as long as that measurement error is uncorrelated with Z. However, it is important to note that, as the key assumption that supports the use of Z as an instrument contradicts the assumption required to use Z as a proxy (and vice-versa), a variable may be a plausible instrument or a plausible proxy, or neither; but never both.[8]

3 Inference and Precision

3.1 Asymptotic Standard Errors - One Proxy

The direct estimation of (1) on complete data, under the assumptions listed in (A1), would result in an asymptotic variance for $\hat{\beta}$ of $\left(\Sigma_{XX}\right)^{-1}\sigma_{\epsilon}^{2}$.[9] 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 case to illustrate these points, and then give a correct formula for the asymptotic standard errors with possibly multiple proxies.

With a single proxy, $\hat{\beta}^{AM}$, $\hat{\beta}^{RRP}$ and $\hat{\beta}^{BPP}$ are numerically identical, so we derive the asymptotic variance from the **AM** approach. The first stage (A2a) and reduced form (4)

[7] Inoue and Solon (2010) show that 2SIV is not in general efficient because it does not take account of fact that Z_1 and Z_2 will be different in finite samples. They suggest the 2-Sample Two-Stage Least Squares procedure is therefore preferred.

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

[9] We have assumed homoscedasticity in A1 and A2 but the inference results presented here could be extended to the heteroscedastic case following an approach similar to that for 2-sample 2SLS presented in Pacini and Windmeijer (2016).

give two moments

$$plim\left(\frac{1}{n_1}\sum y_1'(z_1-\gamma y_1)\right)=plim\left(\frac{1}{n_1}\sum y_1'u_1\right)=0,$$

$$plim\left(\frac{1}{n_2}\sum X_2'(z_2-\gamma\beta X_2)\right)=plim\left(\frac{1}{n_2}\sum X_2'(\gamma\epsilon_2+u_2)\right)=0$$

which identify the parameters γ and β.

It is informative to first consider implementing $\hat{\beta}^{AM}$ (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, the asymptotic variance-covariance matrix of the moments is

$$F=\begin{bmatrix}\sigma_u^2\Sigma_{yy} & \beta\sigma_u^2\Sigma_{XX}\\ \beta\sigma_u^2\Sigma_{XX} & \left(\gamma^2\sigma_\epsilon^2+\sigma_u^2\right)\Sigma_{XX}\end{bmatrix} \tag{36}$$

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 the parameters. The asymptotic variance of $\hat{\gamma}$ is of course $\left(\Sigma_{yy}\right)^{-1}\sigma_u^2$. The asymptotic variance of $\hat{\beta}$ is

$$Asymp\ Var(\hat{\beta})=\frac{\left(\Sigma_{XX}\right)^{-1}\sigma_\epsilon^2}{\phi_{y,Z}}. \tag{37}$$

Thus the loss of asymptotic precision due to imputation (relative to the direct estimation of (A1a)), is proportional 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 also proportional to a first stage R^2 (Shea, 1997).

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

$$F = \left[\begin{array}{cc} \sigma_u^2 \Sigma_{yy} & 0 \\ 0 & \alpha \left(\gamma^2 \sigma_\epsilon^2 + \sigma_u^2 \right) \Sigma_{XX} \end{array} \right]$$

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 $(G' F^{-1} G)^{-1}$ where G is the gradient of the moments with respect the parameters. The asymptotic variance of $\hat{\gamma}$ is still $(\Sigma_{yy})^{-1} \sigma_u^2$. The asymptotic variance of $\hat{\beta}$ is

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

This can be written as

$$Asymp\ Var(\hat{\beta}) = \alpha^{-1} \left(\frac{(\Sigma_{XX})^{-1} \sigma_\epsilon^2}{\phi_{y,Z}} + 2\beta^2 \left(\frac{1 - \phi_{y,Z}}{\phi_{y,Z}} \right) \right) \tag{38}$$

The second term inside the brackets represents the loss of asymptotic precision, due to the use of two different samples. This loss of precision is because the covariances between moments in equation (36) 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 the **RRP** or **BPP** procedures) are incorrect, but can easily be corrected. The OLS standard errors (as produced by standard software packages) are

$$\begin{aligned} \hat{V}^{OLS}(\hat{\beta}^{BPP}) &= (X_2 X_2)^{-1} \left(\hat{y}_2 - X_2 \hat{\beta} \right)' \left(\hat{y}_2 - X_2 \hat{\beta} \right) = (X_2 X_2)^{-1} \left(\hat{y}_2' \hat{y}_2 - \hat{y}_2' X_2 (X_2' X_2) X_2' \hat{y}_2 \right) \\ &= (X_2' X_2)^{-1} \left[y_1' y_1 (z_1' y_1)^{-1} z_2' z_2 (z_1' y_1)^{-1} y_1' y_1 - y_1' y_1 (z_1' y_1)^{-1} z_2' X_2 (X_2' X_2) X_2' z_2 (z_1' y_1)^{-1} y_1' y_1 \right]. \end{aligned}$$

With some algebra, it is straightforward to show that

$$plim\left(\hat{V}^{OLS}(\hat{\beta})\right) = \frac{\left(\Sigma_{XX}\right)^{-1}\sigma_{\epsilon}^{2}}{\phi_{y,Z}} + \beta^{2}\left(\frac{1-\phi_{y,Z}}{\phi_{y,Z}}\right)$$
$$= \alpha \times Asym\,Var(\hat{\beta}) - \beta^{2}\left(\frac{1-\phi_{y,Z}}{\phi_{y,Z}}\right). \qquad (39)$$

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

3.2 Asymptotic Standard Errors - General Case

If there is more than one proxy $\hat{\beta}^{RRP} \neq \hat{\beta}^{AM}$. To derive the asymptotic variance of $\hat{\beta}^{RRP}$ we write the **RRP** procedure as a set of moment conditions. The parameters are β^{RRP} and g, where g is the vector of coefficients from a linear projection of y on the matrix Z.

Proposition 4. *Assuming that both samples are independent, random samples of the same population and A1a, A1b, A1c, A1d, A2a, A2c, A2d and A3 hold, the $\hat{\beta}^{RRP}$ has asymptotic variance*

$$Asym\,Var(\hat{\beta}^{RRP}) = \Sigma_{XX}^{-1}\left[\alpha\Sigma_{XX}\sigma_e^2 + \frac{\Sigma_{XZ}}{\phi_{y,Z}}\left(\sigma_\delta^2\Sigma_{ZZ}^{-1}\right)\frac{\Sigma_{ZX}}{\phi_{y,Z}}\right]\Sigma_{XX}^{-1}.$$

Proof. Consider the two regressions (dropping sample subscripts for convenience) $y = Zg + \delta$ and $Z\hat{g}/R^2_{y,Z} = X\beta + e$. The moments are:

$$E[Z'(y - Zg)] = 0$$
$$E\left[X'\left(\frac{Z\hat{g}}{R^2_{y,Z}} - X\beta\right)\right] = 0$$

Then we have

$$G = \begin{bmatrix} 0 & -Z'Z \\ -X'X & X'Z\frac{1}{\phi_{y,Z}} \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} Z'Z\sigma_\delta^2 & 0 \\ 0 & \alpha X'X\sigma_e^2 \end{bmatrix}$$

Then

$$\begin{aligned} G'V^{-1}G &= \begin{bmatrix} 0 & -X'X \\ -Z'Z & Z'X\frac{1}{\phi_{y,Z}} \end{bmatrix} \begin{bmatrix} (Z'Z)^{-1}\sigma_\delta^{-2} & 0 \\ 0 & \alpha^{-1}(X'X)^{-1}\sigma_e^{-2} \end{bmatrix} \begin{bmatrix} 0 & -Z'Z \\ -X'X & X'Z\frac{1}{\phi_{y,Z}} \end{bmatrix} \\ &= \begin{bmatrix} \alpha^{-1}X'X\sigma_e^{-2} & -\alpha^{-1}X'Z\sigma_e^{-2}\frac{1}{\phi_{y,Z}} \\ -\alpha^{-1}Z'X\sigma_e^{-2}\frac{1}{\phi_{y,Z}} & Z'Z\sigma_\delta^{-2} + \alpha^{-1}Z'X(X'X)^{-1}X'Z\sigma_e^{-2}\frac{1}{\phi_{y,Z}}\frac{1}{\phi_{y,Z}} \end{bmatrix} \end{aligned}$$

Then, the variance of $\hat{\beta}$ is the upper left of $(G'V^{-1}G)^{-1}$ which is

$$= \alpha(X'X)^{-1}\sigma_e^2 + (X'X)^{-1}\frac{X'Z}{\phi_{y,Z}}(Z'Z)^{-1}\sigma_\delta^2\frac{Z'X}{\phi_{y,Z}}(X'X)^{-1}$$

which can be re-written as

$$= (X'X)^{-1}\left[\alpha X'X\sigma_e^2 + \frac{X'Z}{\phi_{y,Z}}(Z'Z)^{-1}\sigma_\delta^2\frac{Z'X}{\phi_{y,Z}}\right](X'X)^{-1}$$

where Z is a matrix $n \times k$ where k is the number of proxies. □

A STATA package that implements the **RRP** procedure and provides the correct standard errors is available from the authors at https://github.com/spoupakis/rrp.

4 Monte Carlo Experiments

To illustrate the points made above we first present a small Monte Carlo study. The baseline data generating process is as follows. There is a single regressor $x \sim N(0, 2)$. The dependent variable of interest is $y = 1 + \beta \times x + \epsilon$ with $\sigma_\epsilon = 1$. The parameter of interest is $\beta = 1$.

We cannot regress y_2 on x_2 directly, because information on these quantities is collected in separate surveys (We only observe y_1 and x_2, so that we cannot calculate the empirical covariance, $\frac{1}{n}\sum y_1 \times x_1$ or $\frac{1}{n}\sum y_2 \times x_2$). However, both surveys contain a potential proxies for y. We begin with the case of a single proxy, z, which we generate as follows,

$$z_1 = 1 + 0.5 \times y_1 + u_1 \qquad \text{and} \qquad z_2 = 1 + 0.5 \times y_2 + u_2$$

with $u_1, u_2 \sim N(0, \sigma_u^2)$. We consider the case where $\sigma_u = 1$ and a first stage R^2 of 0.56.

We simulate this population multiple times, each time drawing two data sets (y_1, z_1) and (x_2, z_2), and implementing the **RP**, **RP+**, **RRP**, **BPP** and **AM** procedures. Sample size is 500 for both samples (so that $\frac{n_1}{n_2} = 1$) and we perform 10,000 replications.

The results are presented in Table 2. The first column shows the case of complete data (OLS on a data set with both y and x); the remaining columns display results for different imputation procedures. The first row gives the mean over 10,000 replications of the estimate of β. With complete data OLS is unbiased for β. The **RP** and **RP+** procedures are systematically biased and the mean attenuation factor is equal to the population first stage R^2 of 0.56. The **RRP**, **BPP** and **AM** procedures (which are numerically identical here) are approximately unbiased for β.

Table 2: Monte Carlo Experiment: One proxy

	FULL	RP	RP+	RRP	BPP	AM
Mean of $\hat{\beta}$	1.000	0.556	0.555	1.002	1.002	1.002
Std. Dev. of $\hat{\beta}$	0.022	0.036	0.049	0.065	0.065	0.065
Mean of SE($\hat{\beta}$)	0.022	0.028	0.043	0.050	0.050	
Mean of Corrected SE($\hat{\beta}$)				0.064		
Mean of $\frac{1}{n}\sum \hat{y}_i$	1.000	1.000	0.999	1.805	1.000	
Mean of $\frac{1}{n-1}\sum(\hat{y}_i - \bar{\hat{y}})^2$	4.999	2.784	5.000	9.048	9.048	

Note: Results based on 10,000 replications, $n = 500$, $\beta = 1$, $E(y) = 1$, $V(y) = 5$.

Rows two through four show the standard deviation of $\hat{\beta}$ across replications along with the mean of the OLS standard error across replications and (in the case of the **RRP** procedure) the mean of the corrected standard error. When regressing an imputed y on x, the usual OLS variance formula leads to a standard error that is too small, but the corrected standard error correctly captures the variation of $\hat{\beta}$ in repeated sampling.

Finally, rows five and six consider estimating the first two unconditional moments of y. The mean of imputed $\hat{y}$ from the **RP** procedure is unbiased for the population mean of y but of course the variance of $\hat{y}$ from this procedure is not unbiased for the population variance. Adding a stochastic residual from the first stage (the **RP+** procedure) corrects this. Because the **RRP** and **BPP** procedures amount to an upward rescaling of $\hat{y}$, the variances of the resulting imputations are quite biased estimates of the population variance of y. However, in the case of **BPP**, there is no bias in the mean.

Table 3 illustrates the case when there are two proxies available. We generate these as:

$$\begin{aligned} z_{a,1} &= 1 + \gamma_A \times y_1 + u_{A,1} \\ z_{b,1} &= 1 + \gamma_B \times y_1 + u_{B,1} \end{aligned} \qquad \text{and} \qquad \begin{aligned} z_{a,2} &= 1 + \gamma_A \times y_2 + u_{A,2} \\ z_{b,2} &= 1 + \gamma_B \times y_2 + u_{B,2} \end{aligned}$$

where $\gamma_A = 0.4$, $\gamma_B = 0.3$ and $u_A, u_B \sim MVN(0, \Sigma_u)$ with $\Sigma_u = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$.

Table 3: Monte Carlo Experiment: Two proxies

	FULL	RP	RP+	RRP	AM
Mean of $\hat{\beta}$	1.000	0.712	0.712	1.000	1.001
Std. Dev. of $\hat{\beta}$	0.022	0.034	0.044	0.048	0.048
Mean of SE($\hat{\beta}$)	0.022	0.028	0.039	0.039	
Mean of Corrected SE($\hat{\beta}$)				0.048	

Note: Based on 10,000 replications, $n = 500$, $\beta = 1$, $E(y) = 1$, $V(y) = 5$.

The key points are that the **RRP** and **AM** procedures remain approximately unbiased for β, and that the additional proxy improves precision.

Table 4 repeats the study of two proxies for different values of $\sigma_{u,B}$ equal to 1, 2 and 4. and reports the standard deviation of $\hat{\beta}$. This illustrates that in finite samples, the **RRP** procedure can be more efficient than **AM**.

Table 4: Monte Carlo Experiment: Two proxies, varying $\sigma_{u,B}$

Mean of $\hat{\beta}$	FULL	RP	RP+	RRP	AM
For $\sigma_{u,A} = 1$ and $\sigma_{u,B} = 1$	0.022	0.034	0.044	0.048	0.048
For $\sigma_{u,A} = 1$ and $\sigma_{u,B} = 2$	0.022	0.036	0.048	0.060	0.066
For $\sigma_{u,A} = 1$ and $\sigma_{u,B} = 4$	0.022	0.036	0.050	0.067	0.089

Note: Based on 10,000 replications, $n = 500$, $\beta = 1$, $E(y) = 1$, $V(y) = 5$.

Table 5 considers a hot-deck imputation. Here z_1 and z_2 are partitioned into bins, and the missing y_2 is imputed by drawing a y_1 from the relevant z-bin. As demonstrated above, this is formally equivalent to **RP+**. The results are in column 2 (titled "Hot-deck"). As expected, estimates of β are significantly biased, with bias equal to the first stage R^2. Estimates of the unconditional mean and variance are unbiased. In column 3 we rescale the donated y_1 by the first stage R^2, and refer to this a "rescaled hot-deck" (RHD). The result is very limited empirical bias in estimates of β but significant bias in estimates of the unconditional mean and variance.

Table 5: Monte Carlo Experiment: Hot-deck Imputation

	FULL	Hot-deck (HD)	RHD
Mean of $\hat{\beta}$	1.000	0.532	0.986
Std. Dev. of $\hat{\beta}$	0.022	0.049	0.088
Mean of $\frac{1}{n}\sum \hat{y}_i$	1.000	1.001	1.858
Mean of $\frac{1}{n-1}\sum(\hat{y}_i - \bar{\hat{y}})^2$	4.999	4.990	17.218

Note: Based on 10,000 replications, $n = 500$, $\beta = 1$, $E(y) = 1$, $V(y) = 5$. The imputation is based on 1 proxy, partitioned into 10 bins.

5 Empirical Illustrations

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

5.1 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 was only available for a subset of goods, we also impute nondurable consumption spending from the CE Survey into the PSID.[10] 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), log utility spending and the number of cars owned by the household (up to a maximum of two). 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 2005-2013, who do not move, are not observed with zero incomes and who are not observed

[10] Prior to 1999 the PSID only included food and utility spending, which was then broadened to include health expenditures, gasoline, car maintenance, transportation, child care and education. In 2005, additional categories for clothing and entertainment were added.

renting over the sample period. To prevent our results being driven by extreme values, we also 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 a 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 measure 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 the fact that changes in relative prices might change the relationship between food and total spending from one period to the next.[11]

Table 6 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.

Table 7 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 **RP** procedure employed by Skinner, and the third column shows results using the **RRP**

[11] 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 utilty 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 throw out observations who do not report 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 without imputed spending measures with those in the PSID.

Table 6: Imputing nondurable consumption spending using CES

	(1) 2005	(2) 2007	(3) 2009	(4) 2011	(5) 2013
log Food	0.562***	0.545***	0.541***	0.549***	0.555***
	(0.012)	(0.008)	(0.008)	(0.008)	(0.008)
log Utilities	0.377***	0.382***	0.410***	0.384***	0.389***
	(0.015)	(0.010)	(0.011)	(0.010)	(0.011)
Cars	0.035***	0.036***	0.027***	0.042***	0.031***
	(0.005)	(0.004)	(0.003)	(0.004)	(0.004)
Partial R^2	0.728	0.755	0.751	0.761	0.753
N	1,590	2,896	2,759	2,668	2,470

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. 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 year dummies. "Cars" refers to the number of cars capped at a maximum of two. 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 inital regression.

approach we set out above.

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 the **RP** procedure, we underestimate the effects of housing wealth on consumption (with the estimated effect falling to 0.83%). Using the **RRP** procedure, we obtain a value of 1.11% which is very similar to that obtained using the complete data in the PSID. This illustrates the theoretical results of Section 2. It also suggests that the assumed Engel curve model (A2) underpinning our imputation procedure is reasonable in this context and moreover that our demographic covariates adequately control for any sample differences between the PSID and the CE Interview Survey.

5.2 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).

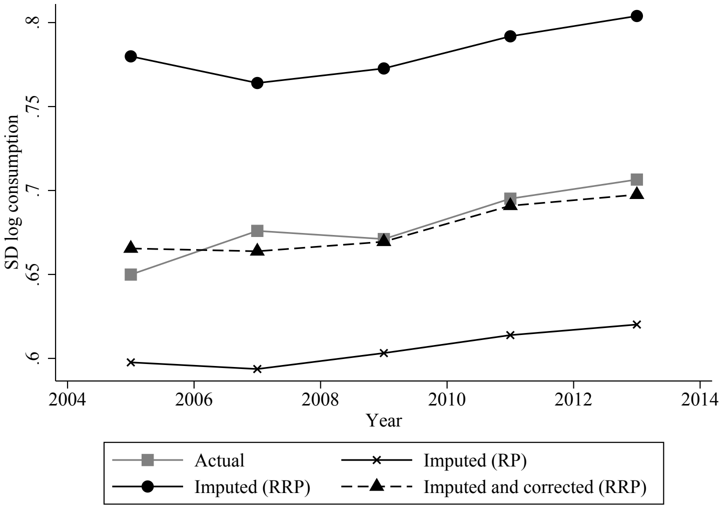
Table 7: Empirical Example: Log nondurable consumption on house values

	(1)	(2)	(3)
	PSID	CE (RP)	CE (RRP)
$\log \text{Income}_{t-3}$	0.047**	0.036*	0.048
	(0.016)	(0.015)	(0.030)
$\log \text{Income}_{t-2}$	0.064***	0.043*	0.057
	(0.018)	(0.017)	(0.034)
$\log \text{Income}_{t-1}$	0.040*	0.024	0.032
	(0.017)	(0.016)	(0.032)
$\log \text{Income}_{t}$	0.109***	0.080***	0.107**
	(0.020)	(0.019)	(0.038)
$\log \text{Income}_{t+1}$	0.105***	0.075***	0.099**
	(0.016)	(0.015)	(0.030)
log House value	0.114***	0.083***	0.111***
	(0.010)	(0.010)	(0.020)
N	5,406	5,406	5,406

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. 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 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 (RP) 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 the re-scaled regression prediction (RRP) procedure.

To do this we impute consumption measures for all households in the PSID (this time including non-homeowners) and plot the standard deviation over time for imputed consumption from the **RP** procedure and from the **RRP** procedure. 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 also trim the top and bottom 1% of consumption spending in the PSID. The results are shown in Figure 2.

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 re-scaled regression prediction corrected using the relationship in equation (15) ("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 **RP** and **RRP** imputed measures reflects the stability of the first stage R^2 over time.

We also note that the **RP** measure tends to understate the *level* of consumption inequality, while the re-scaled (**RRP**) procedure tends to overstate it. This was shown analytically in Section 2. This example reinforces the point made in Table 1 that while the **RRP** procedure 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 in equation (15) to the RRP estimate of the standard deviation we obtain roughly the correct standard deviation. Once again, this suggests that the linear Engel curve relationship we assumed for our imputation procedure (ie., A2) is appropriate in this application.

6 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). Common hot-deck imputation procedures have the same structure, when the matching variables include variables beyond those in included among the independent variables in the regression. This inconsistency can be overcome by rescaling by the first stage (imputation) R^2 (the **RRP** procedure) or by employing reverse regression in the first stage (the **BPP** procedure). Even then, we have shown that the usual OLS standard errors are not correct, but they can be corrected with estimable quantities.

Our analysis demonstrates that the preferred method of imputation may depend on the intended application. This poses a challenge to data providers who may wish to include imputed variables in a standardized data set for multiple users.

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

References

Angrist, J. D. and 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. and 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., and Savignac, F. (2015). Wealth effects on consumption across the wealth distribution: Empirical evidence. Technical Report 1817, ECB Working Paper Series.

Attanasio, O., Hurst, E., and Pistaferri, L. (2015). The evolution of income, consumption, and leisure inequality in the United States, 1980–2010. In Carroll, C. D., Crossley, T. F., and Sabelhaus, J., editors, *Improving the Measurement of Consumer Expenditures*, pages 100–140. National Bureau of Economic Research, University of Chicago Press.

Attanasio, O. and 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.

Baker, S. R., Kueng, L., Meyer, S., and Pagel, M. (2018). Measurement error in imputed consumption. Working Paper 25078, National Bureau of Economic Research.

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

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

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

Bollinger, C. R. and Hirsch, B. T. (2006). Match bias from earnings imputation in the Current Population Survey: The case of imperfect matching. *Journal of Labor Economics*, 24(3):483–519.

Bollinger, C. R. and 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., and Weber, G. (2003). Asking consumption questions in general purpose surveys. *Economic Journal*, 113(491):F540–F567.

Browning, M., Crossley, T. F., and Winter, J. (2014). The measurement of household consumption expenditures. *Annual Review of Economics*, 6(1):475–501.

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

Dumont, M., Rayp, G., Thas, O., and 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. and 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.

Hirsch, B. T. and Schumacher, E. J. (2004). Match bias in wage gap estimates due to earnings imputation. *Journal of Labor Economics*, 22(3):689–722.

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

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

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

Kaplan, G., Mitman, K., and Violante, G. L. (2016). Non-durable consumption and housing net worth in the great recession: Evidence from easily accessible data. Technical Report 22232, National Bureau of Economic Research.

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., and 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. and 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.

Pacini, D. and Windmeijer, F. (2016). Robust inference for the Two-Sample 2SLS estimator. *Economics Letters*, 146(C):50–54.

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.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

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

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

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.

Ziliak, J. P. (1998). Does the choice of consumption measure matter? An application to the permanent-income hypothesis. *Journal of Monetary Economics*, 41(1):201–216.