Three Essays in Applied Microeconometrics

Stavros Poupakis

A thesis submitted for the degree of Doctor of Philosophy

Department of Economics

University of Essex

March 2018

Summary

Chapter 1 develops a specification test for a single index binary outcome model in semi-parametric estimation. The semiparametric estimator examined does not rely on any distributional assumption, but it still relies on the single-index assumption. The violation of this assumption creates a source of heteroscedasticity. I extend a set of attractive LM statistics, constructed using auxiliary regressions for the case of logit and probit models, to the semiparametric environment. I derive its asymptotic distribution and show that is has well-behaved finite properties in a Monte Carlo experiment. An empirical example is also provided.

Chapter 2 proposes a novel estimation strategy that accounts for asynchronous fieldwork, often found in multi-country surveys. The resulting biases are substantial and this is likely to provide misleading cross-country comparisons. I highlight the importance of accounting for the heterogeneity induced by seasonality in the context of regression modelling in order to obtain unbiased comparisons. This is illustrated with a comparison between a synchronous national survey and an asynchronous cross-national one.

Chapter 3, joint work with Thomas Crossley and Peter Levell, pro-

poses a novel estimator useful for data combination. Researchers are often interested in the relationship between two variables, with no available data set containing both. For example, surveys on income and wealth are often missing consumption data. A common strategy is to use proxies for the dependent variable that are common to both surveys to impute the dependent variable into the data set containing the independent variable. 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 an often used procedure is biased, and offer both a correction and refinements that improve precision. We illustrate these with a Monte Carlo study and an empirical application.

Acknowledgements

I owe many people thanks for their crucial support throughout my doctoral research. In particular, I wish to thank my supervisor Thomas Crossley, and my former supervisor João Santos Silva for their crucial guidance and support. Their role in my progress has been invaluable. In addition, I am grateful for useful comments to Abhimanyu Gupta, Ludovic Renou, Andrea Geraci, Simon Lodato, Luca Ferrari, Raluca Pahontu, Pierre Rialland and Federico Vaccari. I highly appreciate the support of Giovanni Mastrobuoni for the crucial role he played during the commencement of my studies in the department. Institutionally, I would like to thank the University of Essex for the financial support, and the department of Economics for providing a nurturing environment during all these years. Finally, I would like to thank my parents for their support throughout the years.

Contents

List of Tables

List of Figures

Chapter 1

A specification test for a single-index binary outcome model in semiparametric estimation

1.1 Introduction

Binary choice models are often employed in economics and statistics. A popular approach to estimate such model's parameters is maximum likelihood, which involves the imposition of some distributional assumptions in the error term. This becomes clear if one considers the latent variable model setting

$$y_i^* = x_i(\boldsymbol{\beta}) + \epsilon_i \tag{1.1}$$

where the response $y_i = 1$ if $y_i^* \geq 0$ and $y_i = 0$ if $y_i^* < 0$, and ϵ_i is assumed to be independent of $\boldsymbol{x}_i$. It is usually assumed that $\boldsymbol{x}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i'\boldsymbol{\beta}$ and this will be

the case throughout the paper. Then it is easy to see that

$$\Pr(y_i = 1|x) = \Pr(y_i^* \geq 0|x) = \Pr(\boldsymbol{x}_i'\boldsymbol{\beta}+\epsilon_i \geq 0) = \Pr(\epsilon_i \geq -\boldsymbol{x}_i'\boldsymbol{\beta}) = 1-F(-\boldsymbol{x}_i'\boldsymbol{\beta}).$$

where popular choices of $F(z)$ are usually link functions such as $\Phi(z)$ (i.e. probit) or $\exp(z)/(1+\exp(z))$ (i.e. logit). Then, the parameters can be estimated as the arguments that maximize the following log-likelihood

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \big\{ y_i \ln(\Pr(y_i = 1|\boldsymbol{x})) + (1 - y_i) \ln(1 - \Pr(y_i = 1|\boldsymbol{x})) \big\}. \quad (1.2)$$

The described procedure is attractive because, as a ML estimator, it is consistent and asymptotically efficient. However, it relies on the correct specification of the model – i.e. the distributional assumption to be correctly specified – which, if not, leads to inconsistent estimates. This limitation gave rise to a vast literature on semiparametric estimation of binary models with many popular alternatives suggested by Manski (1975), Gallant and Nychka (1987), Ichimura (1993), Klein and Spady (1993), among others.

The Klein and Spady (1993) estimator, the focus of this paper, is an appealing estimator as it is consistent, asymptotically normally distributed and achieves the semiparametric efficiency bound (Cosslett, 1983; Chamberlain, 1986):

$$\mathrm{E}\left(\frac{\mathrm{Var}\left(\frac{\partial x(\boldsymbol{\beta_0})}{\partial \boldsymbol{\beta}} | x(\boldsymbol{\beta}_0) \right) f^2\big(x(\boldsymbol{\beta}_0)\big)}{\Pr(y = 1|\boldsymbol{x})\big(1 - \Pr(y = 1|\boldsymbol{x})\big)} \right)^{-1}$$

The Klein and Spady (1993) approach assumes a single index on $E(y_i|\boldsymbol{x}) = m(\boldsymbol{x}_i'\boldsymbol{\beta})$ and suggests maximizing the quasi-likelihood

$$Q\big(\hat{P}(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} \left\{ y_i \ln\Big(\hat{P}_i(\boldsymbol{\beta})\Big) + (1-y_i)\ln\Big(\big(1-\hat{P}_i(\boldsymbol{\beta})\big)\Big) \right\} \tag{1.3}$$

where $\hat{P}_i$ is the estimated P_i, the conditional probability of y_i given $\boldsymbol{x}_i$, using the leave-one-out Nadaraya-Watson estimator

$$\hat{P}_i = \hat{m}_{-i}(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{\sum_{j\neq i} K\left(\frac{(\boldsymbol{x}_j'-\boldsymbol{x}_i')\boldsymbol{\beta}}{h}\right) y_i}{\sum_{j\neq i} K\left(\frac{(\boldsymbol{x}_j'-\boldsymbol{x}_i')\boldsymbol{\beta}}{h}\right)}.$$

In the context of binary choice models, such as this, the numerator represents the estimator of the conditional density $f(\boldsymbol{x}_i'\boldsymbol{\beta}|y_i=1)$ and the denominator represents the estimator of the unconditional density $f(\boldsymbol{x}_i'\boldsymbol{\beta})$. Klein and Spady (1993) show that the estimator that maximizes (1.3) behaves asymptotically as the estimator of

$$Q\big(P(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} \Big(y_i \ln\left(P_i\right) + (1-y_i)\ln(1-P_i)\Big)/n$$

Note that the slope coefficients are identified up to scale. Identification requires some normalization (standard approach is setting a coefficient equal to 1 and estimating all the other coefficients up to this scale), thus assuming the existence of a regressor that is continuous with non-zero coefficient.

In addition, the log likelihood requires a trimming function for the asymptotic properties of the estimator, as at the boundary points $\hat{P}$ will become either zero or one. However, as Klein and Spady (1993) show, trimming is not important in practice and may be ignored. Moreover, they recommend the use of a bias-reducing kernel or an adaptive (also called variable-bandwidth) kernel estimator. In the case of bias-reducing kernel, the estimated probabilities

can be negative, therefore the terms in the logarithmic functions in (1.3) need to be squared. Asymptotically this can be ignored, but in finite samples they may still occur. More recently, De Luca et al. (2008) show that a Gaussian Kernel with fixed bandwidth does not affect the performance in finite samples and this is used throughout the paper. Its function is given by

$$K\left(\frac{\boldsymbol{x}_j'\boldsymbol{\beta}-\boldsymbol{x}_i'\boldsymbol{\beta}}{h}\right)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\boldsymbol{x}_j'\boldsymbol{\beta}-\boldsymbol{x}_i'\boldsymbol{\beta}}{h}\right)^2}$$

An important decision, as with most nonparametric methods, is the choice of bandwidth which should satisfy the rate $n^{-1/6} < h < n^{-1/8}$ for the asymptotic properties (Klein and Spady, 1993). While the choice of bandwidth is normally left to the researcher's discretion, it is generally suggested to be estimated jointly with the parameters through cross-validation (Härdle et al., 1993).

Specification tests have been proposed to test the semiparametric model by comparing it with a parametric model, such as probit or logit, by measuring the deviation of the parametric expected value of y_i from the nonparametric, both conditional upon the single index $\boldsymbol{x}_i'\boldsymbol{\beta}$ (Klein, 1993; Horowitz and Härdle, 1994). These proposals could be used to test the distributional assumption only conditioning on the validity of the single-index assumption. This paper's aim is to propose a test for testing this assumption.

Violations of the single index assumption could be thought either as a case of omitted variables or as a case of presence of heteroskedasticity, and, as such, may be tested with a semiparametric LM test. Gonzalez-Rivera and Ullah (2001) have considered such developments for the linear regression model by constructing the statistic using nonparametric density estimators. This paper's contribution extends this approach in the binary choice models and

enables the construction of the statistic by means of an artificial regression. In that sense, specification tests by means of artificial regression, as the ones proposed by Davidson and MacKinnon (1984) are directly applicable in this context.

The paper is organized as follows: section 1.2 describes the set of the proposed specification tests, section 1.3 reports the results from the Monte Carlo experiments which emphasize the finite sample performance of these tests, section 1.4 presents an empirical application demonstrating the use of these tests and section 1.5 concludes.

1.2 The semiparametric LM test

Suppose in the model formulation of (1.1) one wishes to examine the case of linear omitted variables, which can be thought of as testing in

$$x_i'\boldsymbol{\beta} = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2$$

the restriction that $\boldsymbol{\beta}_2 = 0$. Similarly, for the case of heteroskedasticity of known form, the model now could have the form of

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta}_1 + \epsilon_i \tag{1.4}$$

where $\epsilon_i \sim N(0, \exp(2\boldsymbol{z}_i'\boldsymbol{\beta}_2))$ for a known $\boldsymbol{z}_i$ and $P(y_i = 1|\boldsymbol{x})$ is now $F(\boldsymbol{x}_i'\boldsymbol{\beta})$ with

$$x_i'\boldsymbol{\beta} = \frac{\boldsymbol{x}_i'\boldsymbol{\beta}_1}{\exp(\boldsymbol{z}_i'\boldsymbol{\beta}_2)}. \tag{1.5}$$

It can be easily seen that, if $\boldsymbol{\beta}_2 = 0$, the above model satisfies the homoskedasticity assumption, as the distribution of the error term becomes just

$N(0,1)$. Thus, both in the case of ommited variables or the presence of heteroskedasticity of known form, if one represents the restricted estimates by $\tilde{\boldsymbol{\beta}}^\top = [\tilde{\boldsymbol{\beta}}_1^\top \ \mathbf{0}^\top]$, one can use any of the trinity of tests for nested models to test such restrictions – i.e. the Likelihood Ratio (LR), the Wald and the Lagrange Multiplier (LM).[1] It is generally known that in the parametric setting these three tests are asymptotically equivalent (see a proof in Engle (1984)) and recently it has been shown that the three tests are equivalent in the increasing parameter setting (Gupta, 2018), but in finite samples this is not always the case.

The importance of testing the heteroskedasticity in the context of the Klein and Spady (1993) estimator comes from the single index assumption. In the DGP described in (1.5), the form of $x_i(\boldsymbol{\beta})$ is clearly a violation of the single index assumption. This means that the model leads to inconsistent estimates and should not be used. The fact that the form of $x_i(\boldsymbol{\beta})$ in (1.5) can satisfy the single index assumption under the restriction of $\boldsymbol{\beta}_2 = 0$, as in the case of linear omitted variables, motivates the use of likelihood-based tests for such restriction.

Under the null hypothesis, all three tests are asymptotically distributed as χ^2_r with degrees of freedom r equal to the number of restrictions. The construction of the LR test involves calculation of the likelihood under the null and the alternative as

$$LR = 2\big(\ln L(\hat{\boldsymbol{\beta}}) - \ln L(\tilde{\boldsymbol{\beta}})\big)$$

[1] Note that the Lagrange Multiplier test is also called Rao's score test.

whereas the Wald uses only the unrestricted

$$W = \hat{\boldsymbol{\beta}}_2^\top \left(\mathrm{Var}(\hat{\boldsymbol{\beta}}_2)\right)^{-1} \hat{\boldsymbol{\beta}}_2$$

and the LM, only the restricted

$$LM = g(\tilde{\boldsymbol{\beta}})^\top \tilde{I}^{-1} g(\tilde{\boldsymbol{\beta}}).$$

where the score vector $g(\tilde{\boldsymbol{\beta}})$ is $\frac{\partial \ln L(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}}$ and $\tilde{I}$ an estimator of the information matrix.

The construction of the LM statistic has the benefit of using only the restricted estimates, thus avoiding the need to use the unrestricted case of inconsistent parameters, estimated in a model that violates the single index assumption.

The asymptotic distribution of the LM under the null is easily derived – the gradient vector evaluated at the restricted estimates $g(\tilde{\boldsymbol{\beta}})$ has asymptotically a normal distribution with mean $\mathbf{0}$ and $\tilde{I}$ is a consistent estimate of the covariance matrix of that vector, which depends on $\tilde{\boldsymbol{\beta}}$. There are three ways to estimate that matrix: by minus the Hessian, using the outer product of the gradient or using the $I(\tilde{\boldsymbol{\beta}})$ with a typical element:

$$I_{jk}(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \frac{\frac{\partial F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}} \frac{\partial \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}_j} \frac{\partial F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}} \frac{\partial \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}_k}}{F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})(1 - F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}))} \tag{1.6}$$

However, LM statistics are rarely calculated this way. It is usually more convenient to compute them by means of artificial linear regressions. A set of such LM (or pseudo-LM) statistics based on artificial regressions proposed by Davidson and MacKinnon (1984) are considered in this paper, adapted in the

semiparametric context. I focus on artificial regressions constructed based on the latter estimate of the information matrix $I(\tilde{\boldsymbol{\beta}})$ as Davidson and MacKinnon (1984) showed they perform better in this context.

Consider the artificial linear regression

$$\boldsymbol{r}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{R}(\tilde{\boldsymbol{\beta}})\boldsymbol{b} + \text{errors} \tag{1.7}$$

where an element of the matrix $\boldsymbol{R}$ is

$$R_{ij}(\tilde{\boldsymbol{\beta}}) = \frac{\frac{\partial F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{x}_i'\boldsymbol{\beta}} \frac{\partial \boldsymbol{x}_i'\boldsymbol{\beta}}{\partial \boldsymbol{\beta}_j}}{\Big(F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big(1 - F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big)\Big)^{1/2}} = \frac{f(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})X_{ij}(\tilde{\boldsymbol{\beta}})}{\Big(F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big(1 - F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big)\Big)^{1/2}} \tag{1.8}$$

and an element of the vector $\boldsymbol{r}$ is

$$r_i(\tilde{\boldsymbol{\beta}}) = \frac{y_i - F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})}{\Big(F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big(1 - F(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big)\Big)^{1/2}} \tag{1.9}$$

where $f(\cdot)$ denotes the first derivative of $F(\cdot)$, and $X_{ij}(\tilde{\boldsymbol{\beta}})$ denotes the derivative of $\boldsymbol{x}_i'\boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}_j$ at $\tilde{\boldsymbol{\beta}}$. In the case of linear omitted variables we note that $X_{ij}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{X}_{ij}$, whereas in the case of heteroskedasticity, as in (1.5), the derivatives with respect to $\tilde{\boldsymbol{\beta}}_1$ is $\boldsymbol{x}_i$ and with respect to $\tilde{\boldsymbol{\beta}}_2$ is $-\boldsymbol{x}_i\tilde{\boldsymbol{\beta}}_1\boldsymbol{z}_i$.

The explained sum of squares of this artificial regression is

$$LM = \boldsymbol{r}^\top(\tilde{\boldsymbol{\beta}})\boldsymbol{R}(\tilde{\boldsymbol{\beta}})\big(\boldsymbol{R}^\top(\tilde{\boldsymbol{\beta}})\boldsymbol{R}(\tilde{\boldsymbol{\beta}})\big)^{-1}\boldsymbol{R}^\top(\tilde{\boldsymbol{\beta}})\boldsymbol{r}(\tilde{\boldsymbol{\beta}})$$

which is a LM statistic since $\boldsymbol{r}^\top(\tilde{\boldsymbol{\beta}})\boldsymbol{R}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{g}(\tilde{\boldsymbol{\beta}})$ and $\boldsymbol{R}^\top(\tilde{\boldsymbol{\beta}})\boldsymbol{R}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{I}(\tilde{\boldsymbol{\beta}})$.

This artificial regression also generates two more test statistics, the pseudo

F statistic

$$F = \frac{(\tilde{\boldsymbol{r}}^\top \tilde{\boldsymbol{r}} - SSR)/r}{SSS/(n-p)}$$

and the n times the uncentred R^2 of this regression

$$nR^2 = \frac{\boldsymbol{r}(\tilde{\boldsymbol{\beta}})^\top \boldsymbol{R}(\tilde{\boldsymbol{\beta}}) \big(\boldsymbol{R}(\tilde{\boldsymbol{\beta}})^\top \boldsymbol{R}(\tilde{\boldsymbol{\beta}})\big)^{-1} \boldsymbol{R}(\tilde{\boldsymbol{\beta}})^\top \boldsymbol{r}(\tilde{\boldsymbol{\beta}})}{\boldsymbol{r}(\tilde{\boldsymbol{\beta}})^\top \boldsymbol{r}(\tilde{\boldsymbol{\beta}})}$$

which are both asymptotically equivalent to the LM (see Engle (1984)).

It is natural to extend the Davidson and MacKinnon (1984) family of LM-type tests proposed for the probit and logit models to the semiparametric context, in particular to the case of Klein and Spady (1993) estimator. This is highly important as it can provide a test for the single index assumption.

I propose a semiparametric LM statistic in the usual form

$$LM = \left[\frac{\partial Q(\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right]^\top \tilde{\Sigma}^{-1} \frac{\partial Q(\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \tag{1.10}$$

where $\partial Q(\tilde{\boldsymbol{\beta}})/\partial \beta$ is the gradient of the quasilikelihood in (1.3) evaluated at the restricted estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\Sigma}$ is the consistent estimator of the covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, evaluated at the restricted estimates $\tilde{\boldsymbol{\beta}}$, defined as

$$\Sigma = E\left\{\frac{\partial P}{\partial \beta}\left[\frac{\partial P}{\partial \beta}\right]^\top \frac{1}{P(1-P)}\right\}_{\beta=\beta_0}. \tag{1.11}$$

Proposition 1. *Under the assumptions (C.1)–(C.9) in Klein and Spady (1993), the semiparametric LM statistic in Equation (1.10) convergences in distribution to a χ^2_r distribution under the null, with degrees of freedom equal to r (the number of restrictions).*

Proof. Proof in Appendix 1.6. □

Thus, one can proceed with the construction of the artificial regression as in (1.7) and construct the $\boldsymbol{r}$ vector and $\boldsymbol{R}$ matrix as in (1.9) and (1.8). These adapted versions will have the following form:

$$R_{ij}(\tilde{\boldsymbol{\beta}}) = \frac{\frac{\partial m(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{x}_i'\boldsymbol{\beta}} \frac{\partial \boldsymbol{x}_i'\boldsymbol{\beta}}{\partial \boldsymbol{\beta}_j}}{\Big(m(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big(1 - m(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big)\Big)^{1/2}} \tag{1.12}$$

and

$$r_i(\tilde{\boldsymbol{\beta}}) = \frac{y_i - m(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})}{\Big(m(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big(1 - m(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})\big)\Big)^{1/2}} \tag{1.13}$$

where, as before, $m(\cdot)$ is the leave-one-out Nadaraya-Watson estimator.

To construct the estimator $\Sigma(\tilde{\boldsymbol{\beta}})$ as in (1.6) I use the chain rule on:

$$\frac{\partial m(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial m(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{x}_i'\boldsymbol{\beta}} \frac{\partial \boldsymbol{x}_i'\boldsymbol{\beta}}{\partial \boldsymbol{\beta}}$$

obtaining

$$\Sigma(\tilde{\boldsymbol{\beta}}) = \frac{\partial m(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{x}_i'\boldsymbol{\beta}} \frac{\partial \boldsymbol{x}_i'\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} \left[\frac{\partial m(\boldsymbol{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{x}_i'\boldsymbol{\beta}} \frac{\partial \boldsymbol{x}_i'\boldsymbol{\beta}}{\partial \boldsymbol{\beta}}\right]^\top \frac{1}{m(\boldsymbol{x}_i'\boldsymbol{\beta})\big(1 - m(\boldsymbol{x}_i'\boldsymbol{\beta})\big)}.$$

Taking the analytical derivative yields

$$\frac{\sum y_j K^{(1)}\big(\frac{\boldsymbol{x}_j'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}}{h}\big) \sum K\big(\frac{\boldsymbol{x}_j'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}}{h}\big) - \sum y_j K\big(\frac{\boldsymbol{x}_j'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}}{h}\big) \sum K^{(1)}\big(\frac{\boldsymbol{x}_j'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}}{h}\big)}{\Big(\sum K\big(\frac{\boldsymbol{x}_j'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}}{h}\big)\Big)^2}$$

with the summation $\sum$ being $\sum_{j\neq i}$, or in a simpler form

$$\frac{\partial \hat{m}(z)}{\partial z} = \frac{\sum_{j\neq i} y_j K^{(1)}(z) - m(z)\sum_{j\neq i} K^{(1)}(z)}{\sum_{j\neq i} K(z)}$$

where $K^{(1)}(z) = \partial K(z)/\partial z$. It is easy to see that for the case of Gaussian kernel

$$K^{(1)}\Big(\frac{\boldsymbol{x}_j'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}}{h}\Big) = \frac{\boldsymbol{x}_j'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}}{h^2}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\boldsymbol{x}_j'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}}{h}\right)^2}\right) \tag{1.14}$$

1.3 Monte Carlo Results

A Monte Carlo simulation study was carried out to demonstrate the behaviour of the proposed test statistics in finite samples. I explore the performance of the tests under the two designs – one for testing for omitted variables and one testing for heteroskedasticity.

For the case of omitted variables I construct the latent variable as

$$y_i^* = x_i(\beta) + \epsilon_i$$

where the ϵ_i is drawn from a Student-t with 5 degrees of freedom and $x_i(\beta)$ as a linear index

$$x_i(\beta) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

where $x_{i1} \sim U(-1, 1)$, $x_{i2} \sim U(-1, 1)$ and $x_{i3} \sim N(0, 1)$. Then, y_i is created as an index variable as $y_i = I(y_i^* > 0)$ and this is what is observed by the researcher. I set $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_2 = 2$. The hypothesis of interest is

testing $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$. In the first set of experiments, the null hypothesis that $\beta_3 = 0$ was always true.

For the case of heterogeneity, the data generation process for the latent variable is

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

with $x_{i1} \sim U(-1, 1)$, $x_{i2} \sim U(-1, 1)$, and $\epsilon_i \sim N(0, \exp(2\beta_3 x_{i3}))$ with $x_{i3} \sim N(0, 1)$ and ϵ_i as before. The same rule is applied for the observed outcome $y_i = I(y_i^* > 0)$. Under this setting, the $P(y_i = 1|x)$ is equal to $F(x_i(\beta))$ with

$$x_i(\beta) = \frac{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}{\exp(\beta_3 x_{i3})}$$

being a specification that allows one to test for heteroskedasticity. As before, I set $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_2 = 2$. The hypothesis of interest is testing $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$. It is clear that if the null is true, the error term will be homoskedastic and the specification of the model will follow the single index assumption, as the denominator will reduce to unity. The alternative is a specification of heteroskedasticity and a clear violation of the single index assumption that the Klein and Spady (1993) semiparametric binary index model relies on. As before, in all the first set of experiments, the null was always true.[2]

[2]For each Monte Carlo experiment I conducted 10,000 replications. I assessed the performance of the tests for two different samples sizes – of 500 and of 2,000. The experiment was conducted using the R 3.3.1 statistical software. The estimation of the binary models was performed using the 0.60-2 version of the *np* package (Hayfield et al., 2008) available from CRAN. The bandwidth was fixed at $n^{-1/6.5}$ which satisfies the rate for the asymptotic properties discussed in the introduction. In each estimation step, 15 different starting points were chosen using the package's procedure. All the tests were constructed with author's coding, with trimming at 1e-5. As it is discussed in the introduction trimming is not important and does not changes the results. The code is available upon request.

For the omitted variables case I construct and compare the three statistics from the auxiliary regression, LM, F and nR^2 along with the Likelihood Ratio test statistic. For the heteroskedasticity design I only report the former three.[3] In general, one would prefer to calculate only LM-type tests as this requires estimation of the model under the null. Given that this is much easier to estimate than the estimation of the alternative, the focus of this study is on the statistics calculated from the auxiliary regression.

Results from the omitted variables design are presented in Table 1.1 for sample size 500 and 2,000. Table 1.1 reports the mean, the standard deviation and the relative frequency of rejecting the null at 5% significance level. For all the tests, as the sample size increases, their means and standard deviations approach their asymptotic values, showing the consistency of these tests. It is clear that among the LM-type tests, the LM is performing best. Its mean is close to one and has the smallest standard deviation.[4] The probability of rejecting the null is much smaller than the other two, and as good as the LR and the Wald.

Table 1.1: Performance of the Test Statistics for Omitted Variables

		LM	F	nR2	W	LR
500	Mean	1.102	1.190	1.189	1.178	1.164
	Std. Dev.	1.600	1.900	1.895	1.686	1.676
	Rej H_0 Freq	0.064	0.089	0.088	0.071	0.071
2000	Mean	1.108	1.128	1.127	1.116	1.117
	Std. Dev.	1.525	1.742	1.725	1.617	1.621
	Rej H_0 Freq	0.058	0.071	0.070	0.067	0.068

Results based on 10,000 replications.

The poor performance of the F and the nR^2 is not unexpected, even

[3]The comparison with the Wald test was not included since it is not attractive in this context.

[4]Note that the mean of χ^2_1 is 1 and the standard deviation is $\sqrt{2}$.

though they are based on the same auxiliary regression. In line with the results of Davidson and MacKinnon (1984) these two tests always have larger variances than the LM. As they note

$$nR^2 = \frac{n}{\tilde{r}^T\tilde{r}}LM$$

and

$$kF = \frac{n-m}{\tilde{r}^T\tilde{r} - LM}LM$$

where k is the number of restrictions and m the number of parameters in $\boldsymbol{\beta}$, here 1 and 4 respectively. It is clear from both equations, that the LM will have the smallest variance. In addition, it can be expected that F and $nR2$ will be relatively close, which is what Davidson and MacKinnon (1984) found and what also we found in our simulation study.

Table 1.2 reports the results from the heteroskedasticity design. In this case, the LR and the Wald are prohibited since they require evaluation of the model under the alternative, thus leaving the choice to the LM-type auxiliary tests. The results are very similar to the omitted variables case. As before, the LM test looks superior to all other LM-type tests: it has a smaller mean (closest to one), smaller standard deviation and rejects the null less times than any of the other two. Again, as the sample size increases, the measures reach the asymptotic ones, which indicates their consistency.

Figure 1.1 and 1.2 present the size of distortion for the omitted variables design and Figure 1.3 and 1.4 inform about the heteroskedasticity design. In all figures, the left hand side graphs show the empirical size of each test in the y-axis and the asymptotic size in the x-axis. The graphs on the right hand side have in the y-axis the difference of the empirical size and the asymptotic

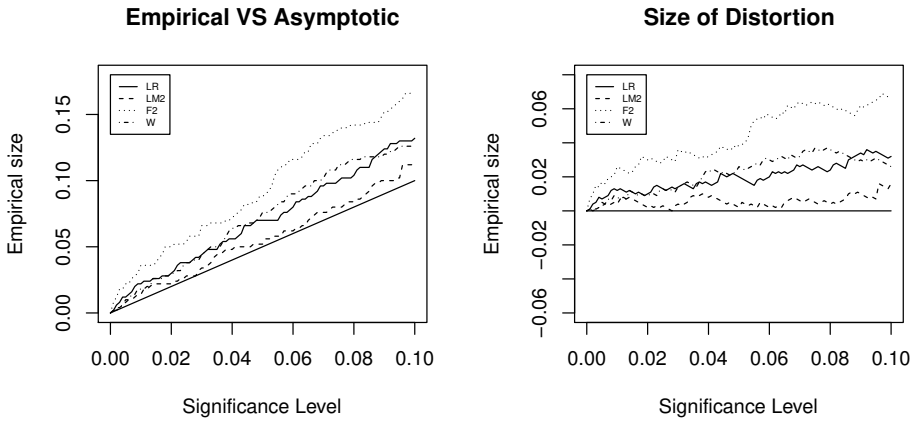
Table 1.2: Performance of the Test Statistics for Heteroskedasticity

		LM	F	nR2
500	Mean	1.156	1.192	1.195
	Std. Dev.	1.593	1.843	1.849
	Rej H_0 Freq	0.076	0.092	0.094
2000	Mean	1.021	1.062	1.063
	Std. Dev.	1.471	1.591	1.582
	Rej H_0 Freq	0.056	0.078	0.079

Results based on 10,000 replications.

and x-axis, as before, the asymptotic size.

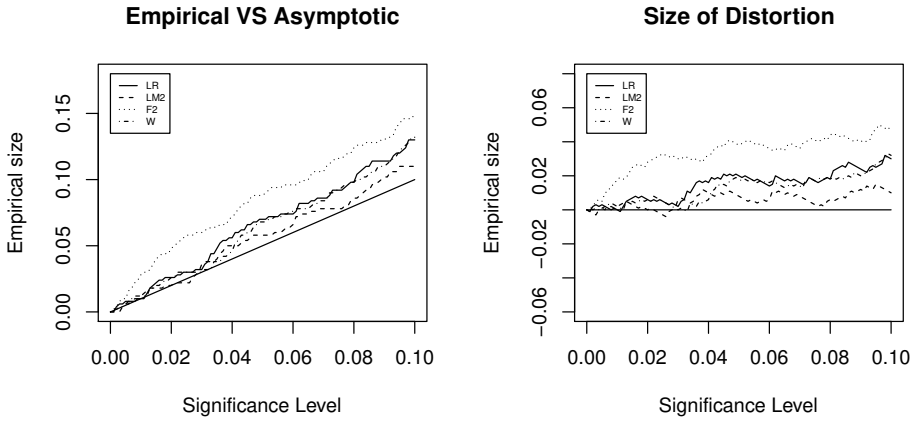
Figure 1.1: Size of Distortion for the Omitted Variables Test for $n = 500$



By plotting the whole range of significance level up to 0.10, one can further explore the rejection frequencies for different levels of significance. The $y = x$ line (45° line) on the left hand side graphs and the $y = 0$ line on the right hand side graphs denote the equality of the empirical and the asymptotic one. A divergence from these lines means that test is performing poorly, whereas when a test mimics them – or at least is as close as possible – implies the test is performing well. Of course, one should always expect the line to diverge as the significance level becomes greater, thus the size of distortion to increase.

One can clearly note that the best performing ones are the LM and the

Figure 1.2: Size of Distortion for the Omitted Variables Test for $n = 2000$



LR, with the LM behaving slightly better. Furthermore, for small values of the significance level, the LM test is overrejecting, but not to a higher extent than the underrejection of LR.

The F and nR^2 are behaving badly as the results from the Tables 1.1 and 1.2 have also indicated. Since they are almost numerically identical, as discussed above, the lines of the two tests overlap (resulting in identical rejection frequencies across the range of significance level). Thus the graphs include only the F test.

Figure 1.3: Size of Distortion for the Heteroskedasticity Test for $n = 500$

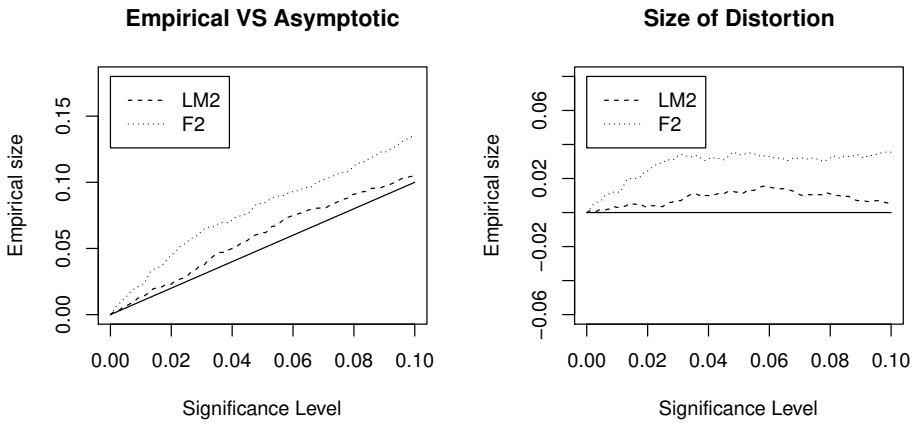


Figure 1.4: Size of Distortion for the Heteroskedasticity Test for $n = 2000$

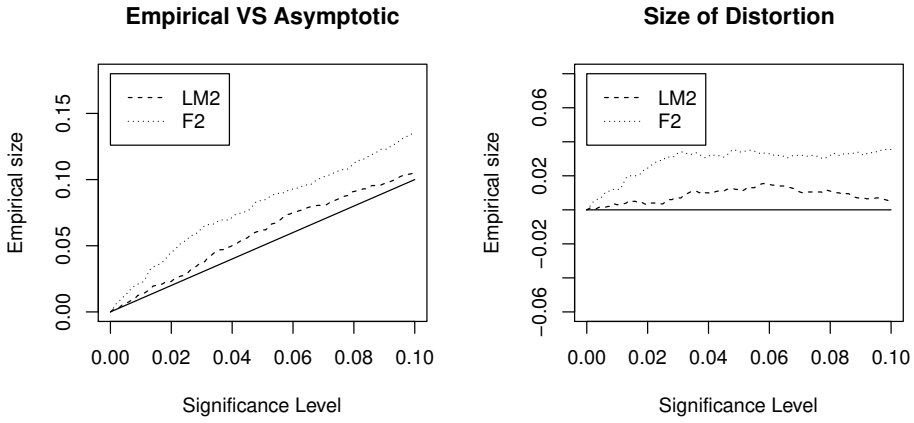


Table 1.3 presents some analysis of the power of the proposed test statistics. This is done for the case of heteroskedasticity, as this is one of main interest. The power performance of the three tests LM, F and nR2 is examined for a set of three different alternatives for β_3 (0.2, 0.4 and 0.6) as when β_3 is non-zero the error term is no more homoskedastic. The same sample sizes 500 and $2,000$ are considered as before. All the tests are consistent since the power goes to 1 as sample size increases. Moreover, all the three tests seem to have very similar power, with LM being slightly larger, even if under the null the LM was clearly better, as show above.

Table 1.3: Power of the Test Statistics for Heteroskedasticity

	β_3	LM	F	nR2
500	0.2	6.2%	5.8%	5.8%
	0.4	11.2%	10.7%	10.7%
	0.6	19.0%	18.8%	18.7%
2000	0.2	13.9%	11.4%	11.4%
	0.4	40.1%	39.8%	39.7%
	0.6	66.0%	64.8%	64.8%

Results based on 10,000 replications.

1.4 Empirical Application

In this section I discuss the importance of testing the single index assumption in the application of labour force participation selection models (Heckman, 1979). Following the work of Blundell and Meghir (1986) on the misspecification of female labour force supply models, semiparametric estimations have been increasingly employed in estimating the decision to participate in the labour market (see, for example, Martins (2001)). However, semiparametric estimation is not a panacea for misspecification.

Although many have discussed various problems in Martins (2001) – such as the significance and sign of the husband's wage on the wife's decision to get employed (see Coelho et al. (2005)) – my focus is on the potential violation of the single index assumption emerging even when employing a Klein and Spady (1993) estimation in the first stage of the selection models.

Martins (2001) estimates womens' wage in Portugal using the 1991 Portuguese Employment Survey.[5] In the first stage of the selection model she argues that a semiparametric specification should be used, rather than a probit, providing evidence against the normal distribution assumption. Although the data does indeed appear to suggest this is the case (as also argued in Coelho et al. (2005)), the more pressing issue I emphasize below is the violation of the single index assumption. This means that although the semiparametric estimation makes no assumption about the distribution of the error term, it does not overcome the single index assumption.

Table 1.4 presents the results of applying the LM test, discussed in this paper, to the covariates in the equation of labour force participation (Table

[5]Data is available on the JAE website, available to download at `http://qed.econ.queensu.ca/jae/2001-v16.1/martins/`.

III on page 31 in Martins (2001)). A replication of Table III showing Probit and Klein & Spady estimates is provided in the Appendix. Although the inclusion of CHILD, YCHILD and EDU does not reject the assumption of homoskedasticity (as indicated by the LM tests), this is not the case for HW and AGE2, which reject the assumption strongly.

Table 1.4: LM Test for the Martins (2001) paper

Variable	Description	LM test	d.f.	p-value
CHILD	Number of children younger than 18	0.665	1	0.414
YCHILD	Number of children younger than 3	2.266	1	0.132
HW	Husband's wage	7.273	1	0.007
EDU	Education, years of schooling	1.783	1	0.182
AGE2	Age squared	4.927	1	0.026

Each row is an independent test for the respective variable.

It is interesting to note that the results presented in Table 1.4 emphasize the unreliable estimates of husbands' wage on women's decision to enter the labour market. Since heteroskedasticity leads to inconsistent estimates, these results also speak directly to evidence presented in Coelho et al. (2005), further confirming estimation problems of husbands' wage effect.

1.5 Conclusion

In this paper, I propose a test for the single index assumption of the Klein and Spady (1993) semiparametric binary model. Although such models allow for a distribution-free maximization of the loglikelihood and, thus, assumed by many researchers as a way to overcome limitations of the fully parametric probit or logit models, they still rely on the single index assumption. Since violating the single index assumption can be directly seen as a form of heteroskedasticity, I propose a semiparametric approach to test this by a class of

LM-type tests by means of artificial regression as proposed by Davidson and MacKinnon (1984) for the case of parametric probit and logit.

Such tests are easy to implement and are based on simple auxiliary regressions. An adjustment is required to construct such tests in the semiparametric context. I also prove their consistency and show the asymptotic distribution of these new tests. Moreover, I compare their finite sample performance in a series of Monte Carlo simulations and find that the LM test proposed here is superior to other LM-type tests from the same auxiliary regressions and also from the LR test. In any case, LR is unattractive as it requires estimation of the unrestricted model, in addition to the restricted one. Moreover, in the Monte Carlo simulations, the LR calculation for the omitted variables case took substantially more time than the easily computable LM-type statistics.

Researchers are advised to be sceptical about the assumptions of the family of semiparametric models, as these also rely on important assumptions – although less restrictive, these should also be tested. This paper facilitates this process by proposing a method of testing the single index assumption. An extension of such LM-type tests to other semiparametric models, such as for the Gallant and Nychka (1987), should be straightforward.

1.6 Appendix

Proof. Proposition 1

$$\max Q(\beta) \quad \text{s.t. } c(\beta) = 0$$

F.O.C.

$$\frac{\partial Q(\tilde{\beta})}{\partial \beta} + \frac{\partial c(\beta_0)}{\partial \beta}\tilde{\lambda} = 0 \tag{1.15}$$

and

$$c(\tilde{\beta}) = 0 \tag{1.16}$$

By taking the Taylor expansion of $\partial Q(\hat{\beta})/\partial\beta$ around β_0, the equation becomes

$$\frac{\partial Q(\hat{\beta})}{\partial \beta} = \frac{\partial Q(\beta_0)}{\partial \beta} + \frac{\partial^2 Q(\beta_0)}{\partial \beta \partial \beta'}(\hat{\beta} - \beta_0)$$

$$0 = \frac{1}{\sqrt{n}}\frac{\partial Q(\beta_0)}{\partial \beta} + \frac{\partial^2 Q(\beta_0)}{\partial \beta \partial \beta'}\frac{1}{\sqrt{n}}(\hat{\beta} - \beta_0) \tag{1.17}$$

And a similar Taylor expansion of $\partial Q(\tilde{\beta})/\partial\beta$ around β_0 yields

$$\frac{1}{\sqrt{n}}\frac{\partial Q(\tilde{\beta})}{\partial \beta} = \frac{1}{\sqrt{n}}\frac{\partial Q(\beta_0)}{\partial \beta} + \frac{\partial^2 Q(\beta_0)}{\partial \beta \partial \beta'}\frac{1}{\sqrt{n}}(\tilde{\beta} - \beta_0) \tag{1.18}$$

subtracting (1.17) from (1.18), gives the following

$$\frac{1}{\sqrt{n}}\frac{\partial Q(\tilde{\beta})}{\partial \beta} = \frac{\partial^2 Q(\beta_0)}{\partial \beta \partial \beta'}\frac{1}{\sqrt{n}}(\tilde{\beta} - \hat{\beta}) = -\frac{\partial^2 Q(\beta_0)}{\partial \beta \partial \beta'}\frac{1}{\sqrt{n}}(\hat{\beta} - \tilde{\beta}). \tag{1.19}$$

By taking a Taylor expansion of $c(\hat{\beta})$ and $c(\tilde{\beta})$ around β_0,

$$c(\hat{\beta}) = c(\beta_0) + \frac{\partial c(\beta_0)}{\partial \beta}(\hat{\beta} - \beta_0) \tag{1.20}$$

$$c(\tilde{\beta}) = c(\beta_0) + \frac{\partial c(\beta_0)}{\partial \beta}(\tilde{\beta} - \beta_0) \tag{1.21}$$

Subtracting (1.21) from (1.20) gives

$$c(\hat{\beta}) = \frac{\partial c(\beta_0)}{\partial \beta}(\hat{\beta} - \tilde{\beta})$$

since from (1.16) we have that $c(\tilde{\beta}) = 0$. Multiplying both sides by $1/\sqrt{n}$ yields

$$\frac{1}{\sqrt{n}} c(\hat{\beta}) = \frac{\partial c(\beta_0)}{\partial \beta} \frac{1}{\sqrt{n}}(\hat{\beta} - \tilde{\beta}) \tag{1.22}$$

and combing (1.22) with (1.19) results to

$$\frac{1}{\sqrt{n}} c(\hat{\beta}) = \frac{\partial c(\beta_0)}{\partial \beta}' \left(\frac{\partial^2 Q(\beta_0)}{\partial \beta \partial \beta'} \right)^{-1} \frac{1}{\sqrt{n}} \frac{\partial Q(\tilde{\beta})}{\partial \beta}$$

where $\frac{\partial Q(\tilde{\beta})}{\partial \beta} = -\frac{\partial c(\beta_0)}{\partial \beta}\tilde{\lambda}$ and $\frac{\partial^2 Q(\beta_0)}{\partial \beta \partial \beta'} = \Sigma(\beta_0)$. Thus, this can be rewritten as

$$c(\hat{\beta}) = \frac{\partial c(\beta_0)}{\partial \beta}' \left(\Sigma(\beta_0)\right)^{-1} \frac{\partial c(\beta_0)}{\partial \beta}\tilde{\lambda} \tag{1.23}$$

Taking the Taylor expansion of $c(\hat{\beta})$ around β_0 shown in (1.20):

$$c(\hat{\beta}) = c(\beta_0) + \frac{\partial c(\beta_0)}{\partial \beta}(\hat{\beta} - \beta_0)$$

and by the asymptotic normality of the Klein & Spady estimator, that is

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma^{-1}(\beta_0))$$

yields

$$\sqrt{n}(c(\hat{\beta}) - c(\beta_0)) \xrightarrow{d} N\left(0, \frac{\partial c(\beta_0)}{\partial \beta}' \Sigma^{-1}(\beta_0) \frac{\partial c(\beta_0)}{\partial \beta}\right)$$

Under the null $c(\beta_0) = 0$, this becomes just

$$\sqrt{n}c(\hat{\beta}) \xrightarrow{d} N\left(0, \frac{\partial c(\beta_0)}{\partial \beta}' \Sigma^{-1}(\beta_0) \frac{\partial c(\beta_0)}{\partial \beta}\right) \tag{1.24}$$

Hence, from (1.24) and (1.23), we obtain

$$\sqrt{n}\tilde{\lambda} \xrightarrow{d} N\left(0, \left(\frac{\partial c(\beta_0)}{\partial \beta}' \Sigma^{-1}(\beta_0) \frac{\partial c(\beta_0)}{\partial \beta}\right)^{-1}\right).$$

Therefore, in the quadratic form (see Theorem B.11 in Greene (2003) p.1086) we find that

$$\frac{1}{n}\tilde{\lambda}' \left(\frac{\partial c(\beta_0)}{\partial \beta}' \Sigma^{-1}(\beta_0) \frac{\partial c(\beta_0)}{\partial \beta}\right) \tilde{\lambda} \xrightarrow{d} \chi^2_r$$

or equivalently, we have the asymptotic distribution of the LM statistic, when the null is true:

$$\frac{1}{n} \frac{\partial Q(\hat{\beta})}{\partial \beta}' \Sigma^{-1}(\beta_0) \frac{\partial Q(\hat{\beta})}{\partial \beta} \xrightarrow{d} \chi^2_r.$$

□

Table 1.5: Replication of Table III in Martins (2001)

	Unrestricted Probit		Probit (β_{AGE} set to 1)		Klein & Spady	
	Coef/Coef$_{Age}$	SE	Coef	SE	Coef	SE
CHILD	-0.137***	(.038)	-0.129***	(0.026)	-0.166***	(0.036)
YCHILD	-0.070	(.088)	-0.058	(0.071)	-0.145	(0.094)
HW	-0.084	(.087)	-0.080	(0.078)	-0.153	(0.104)
EDU	0.153***	(.044)	0.141***	(0.009)	0.203***	(0.039)
AGE2	-0.150***	(.009)	-0.148***	(0.004)	-0.173***	(0.010)
AGE	1		1		1	
CONSTANT	-1.081	(.943)	-1.119	(0.861)		
LogLik	-1372.3		-1,372.4		-1,371.6	
Observations	2,339		2,339		2,339	

Chapter 2

Controlling for asynchronous fieldwork in cross-national surveys

2.1 Introduction

While most social scientists employ cross-national research with the aim to compare events and processes across countries, a number of important aspects hinder correctly-derived comparisons in their analyses. Unbiased estimation of between-country differences has been a topic of intense discussion in the literature (Hakim, 2000; Brislin et al., 1973; Teune, 1977).

Although there exists no general framework for successful cross-national survey research, attempts to file quality standards that should be followed have been proposed in the survey methodology literature (Harkness, 1999; Lynn, 2003). One important recommendation urges survey agencies to conduct the surveys within a common fieldwork period across countries (i.e. common start

and end date). Even when this recommendation is implemented, biases may still arise in cases where fieldwork is asynchronous.

The purpose of this paper is to outline biases arisen when comparing quantities of interest across countries using multi-country survey data when fieldwork is not synchronized across countries. This means that although there is a common fieldwork period, the distribution of the monthly interviews is very different across countries. I first motivate the occurrence of this seasonality in responses using a national survey with monthly samples, and then examine a frequently-studied cross-national survey where the asynchronicity of fieldwork leads to biased comparisons.

Essentially, the problem arises when researchers make comparisons based on data drawn from different seasons or, even when drawn from the same seasons, the proportion of respondents in one country (or one year) within that season is not the same as in another country (or year).[1] Regrettably, in the former case, the researcher cannot account for seasonality. Thankfully though, in the latter case, solutions can be found. As such, I propose that accounting for the heterogeneity induced due to seasonality, by controlling for seasonal variation in the context of regression modelling, leads, *ceteris paribus*, to unbiased comparisons.

This potential source of bias *within country* has been noticed, and accounted for, in a number of within-country studies on consumption (Blundell et al., 1993; Longhi, 2014), health economics (Clemes et al., 2011; Kimura et al., 2015; McCormack et al., 2010; Visscher and Seidell, 2004), happiness

[1]Suppose you wish to compare food consumption in Spain and the UK, with both countries following a common-fieldwork period regulation (say, March to December of each year). The comparison might still be biased if, in the UK, most respondents are surveyed during winter, while most Spanish in the summer. An analogous example applies to time in national surveys.

economics (Connolly, 2013).[2] Despite a recent focus on cross-country modelling (see Skinner and Mason (2012); Kaminska and Lynn (2017) for sample design weights, Bryan and Jenkins (2016) for multilevel modelling), there is still no guidance on dealing with country comparison biases arisen from seasonality.

Seasonality can alter the outcome of interest via two different effects. A *direct* effect, in which seasonality changes the actual outcome (e.g. questions about present ice-cream consumption, when the respondent's true consumption is affected by seasonality) and an *indirect* effect in which seasonality changes the reported outcome (e.g. questions about past ice-cream consumption, say over the past year - the respondent's recall is affected by seasonality[3]). More broadly, failing to account for seasonality in the presence of asynchronous fieldwork can be considered a case of omitted variable bias.

In what follows, I elaborate on the sources and consequences of asynchronous fieldwork in the study of consumption. The paper is organized as a comparison between a UK synchronous within country fieldwork (Section 2.2) and a European asynchronous cross-country survey (Section 2.3). Section 2.4 concludes this paper.

2.2 Understanding Society - A success story

This section presents an empirical illustration emphasizing how synchronous fieldwork (a proportional distribution of respondents across months)

[2]Surprisingly, seasonal variation can exist even in outcomes that one would not normally expect, such as body mass index (BMI) or waist circumference. These measures are considered objective, and hence often used in comparative analysis among countries, but, as existing research has shown, these indices are not free from seasonality (Visscher and Seidell, 2004).

[3]One would expect respondents to over-report their yearly ice-cream consumption if asked during summer and under-report it during winter – for seasonal variation in reported food consumption see (Subar et al., 1994).

overcomes problems of seasonality, by comparison to an asynchronous survey in Section 2.3. This is illustrated by using the Understanding Society, which is a rare case in which such fieldwork is conducted. In this case, if one is interested in estimating the average yearly consumption in a particular year, despite potential observable seasonal variations in consumption, the estimate will be correct as long as the survey includes a representative sample of respondents within each month.[4]

Understanding Society[5] is a household longitudinal survey capturing yearly data about the social and economic circumstances and attitudes of UK household members aged 16 or over. This surveys expands the former British Household Panel Survey, thus incorporating many appealing features, such as a large sample size (about 40,000 households in Wave 1).

The survey design is such that each wave is, in fact, conducted over a two-year period, with the first wave being conducted over 2009 and 2010. Understanding Society has a stratified-clustered design selected through probability proportionate to size (PPS) methods, making it a representative sample of the UK population. For the existing waves (i.e. up until 2015), the survey was conducted using face-to-face computer assisted personal interviewing (CAPI), following a letter inviting the household to participate in the survey and a phone call detailing the procedure and the members' acceptance to take part in the survey.

[4] Note that the correct estimate is achieved, not due to the lack of seasonal variation *per se*, but due to a survey design with a representative sample of respondents within each month. For example, the average yearly ice-cream consumption in the UK in 2017 is estimated correctly given the survey design of Understanding Society, *even though* it is probably the case that ice-cream consumption during summer is higher than during winter.

[5] Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service.

Although the survey includes several subpanels, I use the only monthly representative subset -i.e. General Population Sample (GPS) subset - which is crucial for illustrating seasonal variation. The subset is divided into 24 monthly samples in each wave, with around 1,000 households each month leading to an overall sample of around 24,000 households per wave. To achieve yearly interview follow-ups, subsequent waves are overlapped, with the target to interview respondents in the same month as in the initial wave.

An impressive 40% of respondents were successfully interviewed in the targeted period (i.e. same month), with as much as 80% being interviewed within an interval of one month from their target, and almost all (95%) within an interval of two months. Overall, this offers a sufficiently large and well-structured sample size which allows one to further examine the effects of seasonality (for a detailed description of the survey design see Lynn (2009); Buck and McFall (2011)).

Throughout the analysis I use all available waves to date, covering the period between 2009 to 2015 (Wave 1 -Wave 6). This allows me to estimate average yearly consumption and consumption patterns across months. Consumption is chosen as an application in this paper as it exhibits straigthforward seasonal variation. Other applications may include doctor visits, physical activity, self-reported health, or life-satisfaction.

2.2.1 Consumption Seasonality

The study of consumption and its relationship to income has been at the centre of economic research following the seminal papers of Engel (1895) (the so-called Engel Curve) along with key contributions from Working (1943); Leser (1963); Deaton and Muellbauer (1980). A large body of research ex-

tends the simple log-log linear model (allowing direct calculation of elasticity) with polynomial or non-parametric specifications (Blundell et al., 1993; Banks et al., 1997; Blundell et al., 2003) often found in cross-country comparisons (see De Luca and Peracchi (2012)).

I use four expenditure items found in Understanding Society. These include food expenditure at home and outside home over the last 4 weeks (revealing the direct seasonal effect - Figures 2.1a and 2.1b) and gas and electricity expenditure over the last 12 months (illustrating the indirect effect - Figures 2.2a and 2.2b).

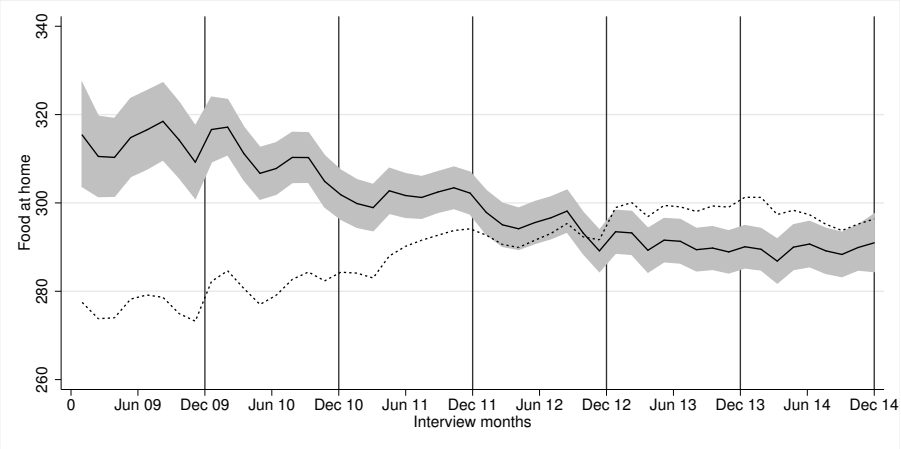
Each graph is estimated by a local polynomial smooth kernel density.[6] In each graph, the y-axis denotes the expenditure in pounds sterling and the x-axis reports the number of months following the first interview in the sample. The vertical lines are drawn at December of each year (at 12, 24, 36, 48, 60 and 72) to ease the graphs' readability. The dashed line represents the reported expenditure and the solid line represents the real expenditure in constant prices (2015 July). The shaded area indicates the 95% confidence interval of the solid line.

It is clear from Figures 2.1a and 2.1b that food expenditure exhibits seasonal variation. As one might expect, food expenditure at home has much less variation, whereas food expenditure outside home is affected much more by seasonality - the kinks in the expenditure are obvious during the summer, along with its cyclicality. This pattern is maintained even when looking at the adjusted expenditure (dashed line). From this, one may notice the overall trend in the time-series of national food consumption which, as expected, increases

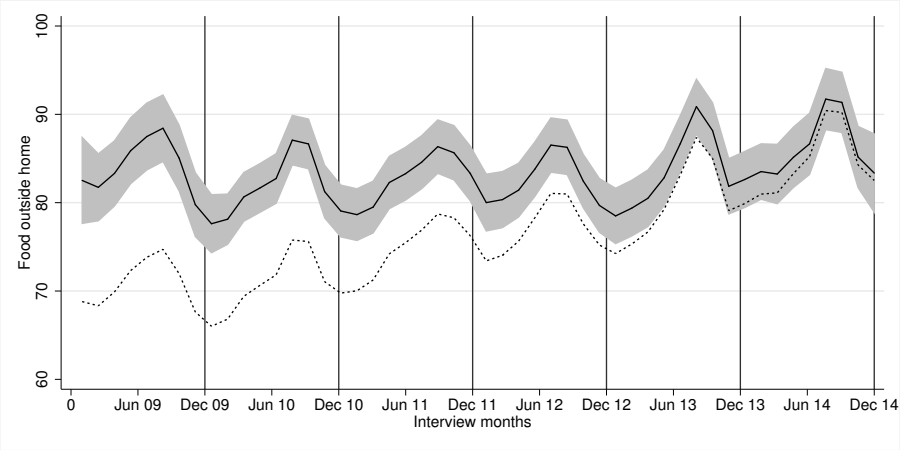
[6]The GPS subset, on which the graphs are based, does not include any weights. Consequently, the estimation is unweighted. For robustness, overall weights were used and results remain unchanged.

Figure 2.1: Direct Effect - Actual Seasonal Variation

(a) Food Home



(b) Food Out



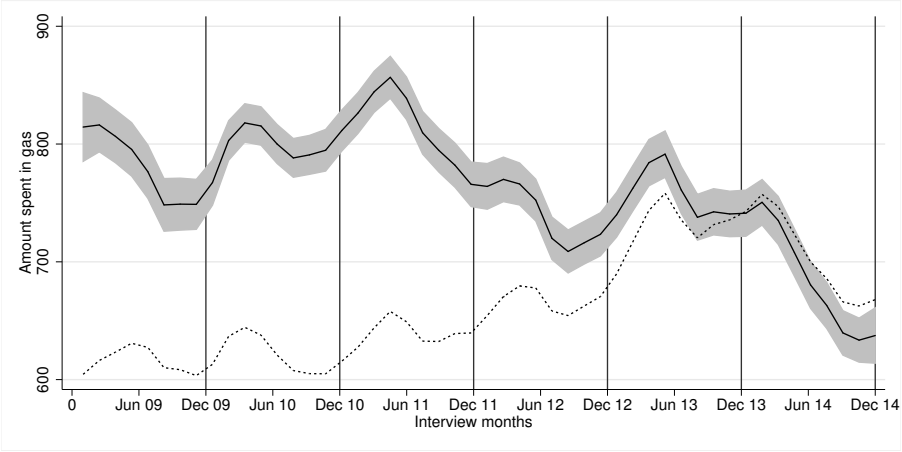
over time.

The results reported in Figure 2.2a are striking. Even if gas is probably an expenditure category expected to have one of the highest seasonal variation, the respondents in this survey are asked about their gas expenditure over the *last 12 months* rather than current consumption.

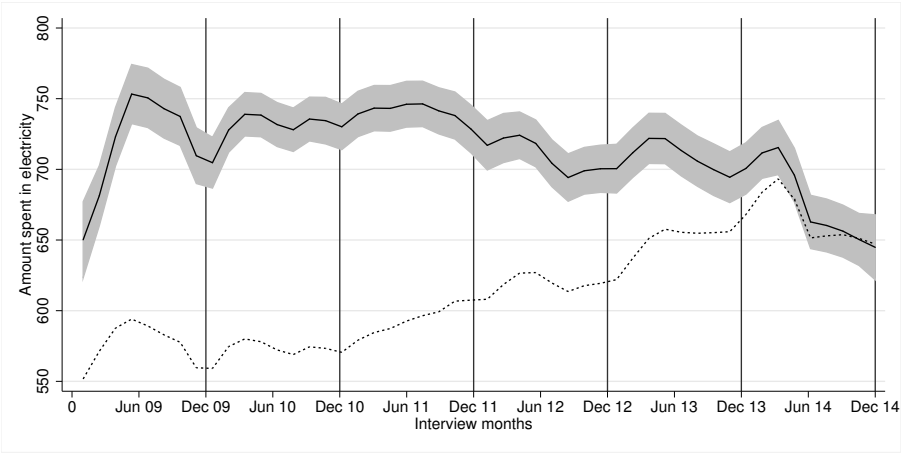
As such, even though gas expenditure in itself should vary across the year, one should not find any effect of seasonality, based on the respondents' assessment of their consumption over the last 12 months. Nevertheless, responses are highly affected by seasonality, with respondents reporting higher

Figure 2.2: Indirect Effect - Recall Seasonal Variation

(a) Gas



(b) Electricity



expenditure after each winter and lower expenditure before each winter.[7]

By contrast, electricity expenditure displays a milder seasonal pattern. This is not unexpected, given that electricity consumption should not very much between summer and winter (e.g. daylight variation between the two seasons may have a moderate impact on this expenditure, which would explain the mild seasonal pattern identified in Figure 2.2b). As before, the dashed lines

[7]This occurs most likely because the most recent gas expenditure affects respondents' reporting behaviour. Respondents interviewed at the beginning of the calendar year most likely recall their most recent (winter) expenditure, which is naturally higher during winter than summer, leading them to use this as a guidance for their reported consumption.

show the upward trend in gas and electricity expenditure in real prices.

Turning to the relationship between consumption and income, I estimate the simple linear Engel curve specification for the food expenditure at home and outside home combined – i.e. I estimate the effect of the natural logarithm of the real income (constant prices) on the natural logarithm of food consumption.[8]

Table 2.1 presents the results of three regression models. Model 1 assumes no seasonality, hence no month dummies are included in the model. The second model assumes only shifts (due to seasonality) in food expenditure, and, as such, month dummies are included to capture this. Model 3 assumes that the income elasticity is changing due to seasonality and includes the interaction of month and the natural logarithm of income to estimate month-specific income elasticity. A joint test of all interaction terms is performed in order to test whether they are jointly zero. The test rejects this and thus favours the specification with month-specific income elasticities ($F_{22,23443} = 3.28$, p-value< 0.001).

All models include year fixed effects to capture year-specific variation and controls for the respondent's gender, age,[9] education and employment status, household size and number of under-aged children. All covariates are significant and the sign of coefficients is in the expected direction. Additionally, the income elasticity for food is positive, significant and smaller than 1.

Table 2.2 shows the Average Marginal Effects (AMEs) based on the regression models 1 and 3 from Table 2.1. Model 1 AMEs corresponds to Model

[8]Table2.8 in the Appendix reports results for food away only, with a greater income elasticity than the combined food. This is not surprising given that food expenditure at home is generally more inelastic than food expenditure outside home.

[9]The partner's age was excluded from the model as it is highly correlated with the respondent's age and did not improve the model. The results remain unchanged when this additional covariate is included.

Table 2.1: Regression Log Food Home and Out

	Model 1		Model 2		Model 3	
	Coef	SE	Coef	SE	Coef	SE
Log Income	0.248***	(0.005)	0.248***	(0.005)	0.257***	(0.011)
Log Prices	0.169	(0.157)	0.401**	(0.158)	0.405**	(0.158)
HH Size	0.250***	(0.003)	0.250***	(0.003)	0.250***	(0.003)
Female	-0.007	(0.005)	-0.007	(0.005)	-0.007	(0.005)
Age	0.003***	(0.000)	0.002***	(0.000)	0.002***	(0.000)
N Children	-0.069***	(0.004)	-0.069***	(0.004)	-0.069***	(0.004)
Unemployed	-0.164***	(0.011)	-0.164***	(0.011)	-0.164***	(0.011)
Retired	-0.020**	(0.008)	-0.020**	(0.008)	-0.020**	(0.008)
Homemaker	0.012	(0.009)	0.013	(0.009)	0.013	(0.009)
Disabled	-0.159***	(0.013)	-0.159***	(0.013)	-0.159***	(0.013)
Other empl.	-0.052***	(0.012)	-0.051***	(0.012)	-0.052***	(0.012)
Other qual.	0.105***	(0.007)	0.105***	(0.007)	0.105***	(0.007)
Degree qual.	0.184***	(0.008)	0.184***	(0.008)	0.184***	(0.008)
Year FE	✓		✓		✓	
Month FE	X		✓		✓	
Month*Income	X		X		✓	
Constant	1.464	(1.424)	-0.643	(1.439)	-0.750	(1.441)
Observations	96,738		96,738		96,738	
R-squared	0.414		0.414		0.414	

*Clustered SE in brackets; *** p<0.01, ** p<0.05, * p<0.1; Ref for categorical: Male, Employed, No qualifications*

1, which calculates a single income elasticity, denoted as 'overall'. Model 3 AMEs reports the income elasticity for each month taken from Model 3. In this case, the 'overall' income elasticity is the average of the 12 months. Given the theoretical expectations laid out above (footnote 4), it is not surprising that the average income elasticities from both models are equal, even if one model did not account for monthly changes in the income elasticity.

As this section is intended to provide an illustration of synchronous fieldwork - i.e. the analytical sample has a representative and equal proportion of people in each month - it is clear that when one calculates the 12 month-specific average, even by omitting the interaction terms, the estimate is not affected

Table 2.2: AMEs Log Food Home and Out

	Model 1		Model 3	
	Coef	SE	Coef	SE
January			0.257***	(0.011)
February			0.236***	(0.011)
March			0.229***	(0.011)
April			0.255***	(0.011)
May			0.225***	(0.011)
June			0.256***	(0.011)
July			0.266***	(0.011)
August			0.276***	(0.012)
September			0.239***	(0.012)
October			0.244***	(0.011)
November			0.254***	(0.011)
December			0.248***	(0.012)
Overall	0.249***	(0.005)	0.249***	(0.005)

*Clustered SE in brackets; *** p<0.01, ** p<0.05, * p<0.1*

by seasonality. As explained more in depth above, the (yearly) average income elasticity is correctly estimated by the simple model (i.e. Model 1) in this survey. If, however, the distribution of the sample would have differed across months, Model 1 would have provided biased estimates - that is, the estimate would have been affected by seasonality. This is illustrated in the next section.

2.3 Cross-Country Comparisons Biases

2.3.1 Controlling for asynchronicity

The foregoing discussion clarified that estimates based on a survey in which fieldwork is synchronous overcomes the problem of seasonality. Asynchronicity, however, tends to occur less in national surveys than in cross-country ones.[10] Cross-national surveys are more susceptible to asynchronic-

[10]This may happen because of coordination problems (i.e. it is easier to coordinate the fieldwork within a single country than across several countries), organizational capacity (i.e.

ity, not only because of the coordination problems or organizational capacity discussed above, but also because of budget limitations, different time constraints, national-specific guidelines on survey data collection. As such, erroneous comparisons are more likely to occur, the estimates representing, in fact, a comparison between a country's population in winter with another country's population in summer.

In order to correct these comparisons and estimate country-specific coefficients, as in Equation (2.1), regardless if seasonality is correlated with X or not, as long as it is correlated with C (e.g. asynchronous fieldwork) leads to omitted variables bias. Then the recommendation is to include a triple interaction of seasonal dummies, country and the independent variable X.[11] Clearly, when the researcher estimates separate regressions for each country, consistent estimates are obtained by interacting only X with seasonal dummies.

$$y = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 M + \beta_4 XCM + \epsilon \tag{2.1}$$

where y is the dependent variable, X is the regressor of interest, M is the month dummies and C is the country dummies.

Consequently, inferences based on cross-national surveys are likely to be biased, as long as asynchronicity is omitted. This is likely the case of most European-based surveys employed in the social sciences, such as European Social Survey (ESS), European Values Survey (EVS), Survey of Health, Ageing and Retirement in Europe (SHARE).[12]

one organization is likely to lead the fieldwork in a single-country survey, while this is not the case when several countries are surveyed).

[11] In order to obtain the country-specific coefficient, one can simply take the average over M. This is easily obtained in statistical software like Stata. For example, one can use `margins, dydx(x) over(country) at((asbalanced) M)`.

[12] A notable exception is the Eurobarometer study which usually conducts all the interviews within a month.

2.3.2 Data Description

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a leading example of cross-country survey, often employed in the social sciences. As such, it is an appealing survey to showcase the importance of accounting for asynchronous fieldwork. This, in turn, may have positive implications on future studies exploring this survey by providing an analytical framework for cross-country comparisons, accounting for seasonality.

SHARE is a cross-national panel survey exploring various socio-economic circumstances and attitudes of Europeans aged 50 or older.[13] The survey is conducted every two years, with the initial wave beginning in 2004 and the 7th wave being currently collected. The first wave covers about 19,500 households and about 28,500 individuals in 11 European countries (Austria, Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden and Switzerland). Throughout the survey 27 countries participated for at least one wave.[14]

The SHARE sample is a probability sample drawn for each participating country, covering different sampling methods, ranging from simple random selection of households at the national level to multi-stage design at the regional level (Klevmarken et al., 2005). A household is selected as long as it has at least one member aged 50 or older. Interviews are normally conducted using

[13]The SHARE data collection has been primarily funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812) and FP7 (SHARE-PREP: N211909, SHARE-LEAP: N227822, SHARE M4: N261982). Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

[14]Apart from some countries in wave 1 which participated in later waves, additional countries include: Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Finland, Hungary, Israel, Latvia, Lithuania, Malta, Poland, Portugal, Romania, Slovakia, Slovenia.

face-to-face computer assisted personal interviewing (CAPI), in addition to a self-administered paper-and-pencil questionnaire and show-cards.

Like any panel survey, attrition may be a problem, however, the sampling design includes refreshments every wave in order to maintain a representative sample of the targeted populations.[15]

Despite challenges arising in conducting cross-national surveys, SHARE manages to implement a harmonised cross-national survey which maintains a high-quality survey design, fieldwork monitoring and survey management since the baseline (2004) wave (de Luca and Lipps, 2005).

Since then, compliance profiles are regularly published to inform the researchers about the fieldwork periods, interviewers composition, contact and response rates. Table 2.3 shows the number of participating countries and the specific fieldwork periods within each wave (see de Luca and Lipps (2005); Schroder (2011); Malter and Börsch-Supan (2013a,b)).

Table 2.3: Fieldwork Periods in SHARE

Wave	Number of countries	Fieldwork Period
1	12	September 2004 - December 2005
2	15	October 2006 - September 2007 †
4	16	November 2010 - March 2012 ‡
5	15	February 2013 - November 2013
6	18	January 2015 - November 2015

†*Except Ireland that lasted until December 2007.* ‡*Although fieldwork period was longer due to funding limitations, the core of SHARE countries had it between Feb-Dec 2011.*

[15]Given its appealing characteristics, first and foremost, one of the few European panel surveys, between the start of the survey (2004) until August 2017, more than 1,800 publications of books, book chapters, journal articles and working papers were based on the SHARE data (see `http://www.share-project.org/share-publications/user-publications-statistics.html`).

2.3.3 Descriptive Statistics

Focusing on the measure of consumption, as in the previous section, the respondent is asked the following question regarding his or her household's food consumption behaviour at home: *Thinking about the last 12 months, how much did your household spend in a typical month on food to be consumed at home?* and outside the house *Thinking about the last 12 months: how much did your household spend in a typical month on food to be consumed outside home?*. Clearly, these questions have the potential indirect seasonality effect discussed in the previous section.

With respect to income, SHARE offers two measures: a generated one (composed from 19 items, including rent, pensions, interest from assets, etc.) and a self-reported one (which includes the respondent's answer to the question *How much was the overall income, after tax, that your entire household had in an average month in the previous year?*). Both measures have advantages and limitations. For example, the generated measure is likely to be more accurate given its wide coverage of income sources, but it is more prone to item non-response, whereas a single, self-reported, income is more likely to suffer from measurement error (De Luca and Celidoni, 2015).

Overall, the self-reported income measure appears superior, for the purposes of this paper, compared to the generated one simply because the latter exhibits frequent item non-response and thus many missing values (an appropriate treatment of selection and item non-repsponse is beyond the scope of this paper - see De Luca and Peracchi (2012) for a detailed treatment of these issues).

Wave 3 (2008/2009) SHARELIFE is not considered given that it is a retrospective survey about the respondents' life histories and does not ask

respondents about their 2008/2009 consumption. Wave 1 is also excluded because it does not include the self-reported measure of income. The resulting sample thus covers waves 2, 4, 5 and 6 (DOIs: 10.6103/SHARE.w2.600, 10.6103/SHARE.w3.600, 10.6103/SHARE.w4.600, 10.6103/SHARE.w5.600, and 10.6103/SHARE.w6.600) and includes 55,784 household-year-observations drawn from eight countries (Austria, Belgium, Denmark, France, Germany, Italy, Sweden, Switzerland).[16]

Crucially, all waves considered include a variable denoting the month in which each respondent was interviewed and this is used in order to capture asynchronicity. Table 2.4 reports the number of households available for each month across countries.

Table 2.4: Sample Size across Country and Wave

	Austria	Germany	Sweden	Italy	France	Denmark	Swi.	Belgium
Jan	173	205	293	507	205	192	205	308
Feb	336	2,007	806	1,115	135	546	1,297	1,198
Mar	914	1,834	853	1,466	2,326	1,340	1,766	2,026
Apr	1,126	1,264	898	1,359	2,005	1,204	1,208	1,854
May	897	917	1,093	1,224	2,384	1,220	966	2,045
Jun	928	471	1,059	1,358	1,161	942	569	1,893
Jul	836	245	518	1,340	499	414	350	1,064
Aug	1,002	866	506	386	163	407	214	661
Sep	1,094	479	693	497	158	525	114	450
Oct	980	213	701	245	153	590	137	217
Nov	574	319	963	48	1,011	475	280	298
Dec	546	144	239	18	287	133	171	180

In addition, Figure 2.3 reports the distribution of interview months across countries in percentages. Figure 2.4 shows the cumulative frequency of interview months, which represents the rate of the data collection.

[16]The choice of countries followed a two-step procedure. First, I select only the countries surveyed in all the four waves considered in this paper. Out of the 11 countries that participated in all waves, I exclude Greece, the Netherlands and Spain, as interviews did not take place in all months in these countries.

A steep line at the beginning and then flat suggests the country collected many interviews at the start of the fieldwork period, whereas an initially flat and progressively steep line suggests that the country belated the collection of most interviews at the start of the fieldwork period.

Figure 2.3: Sample Coverage by Month in SHARE WAVE 6

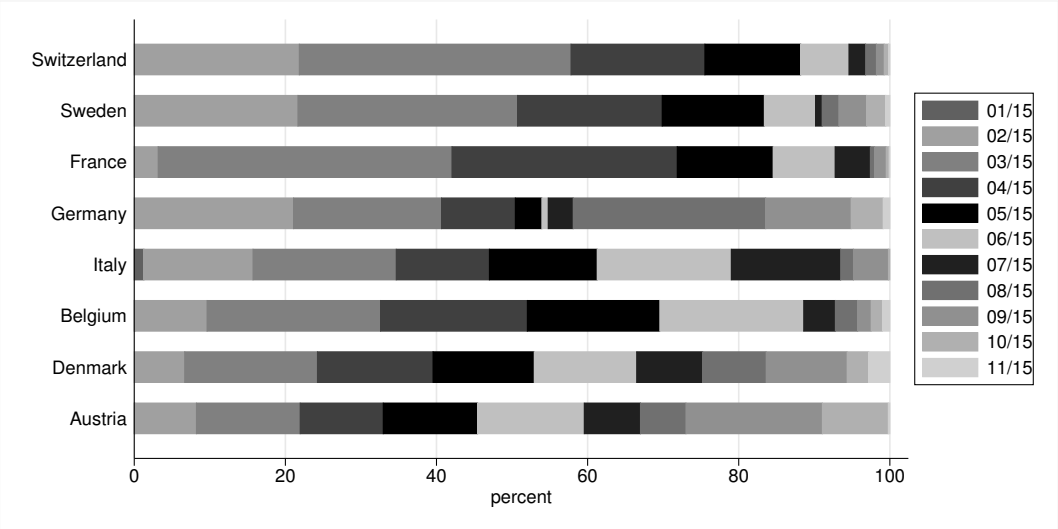
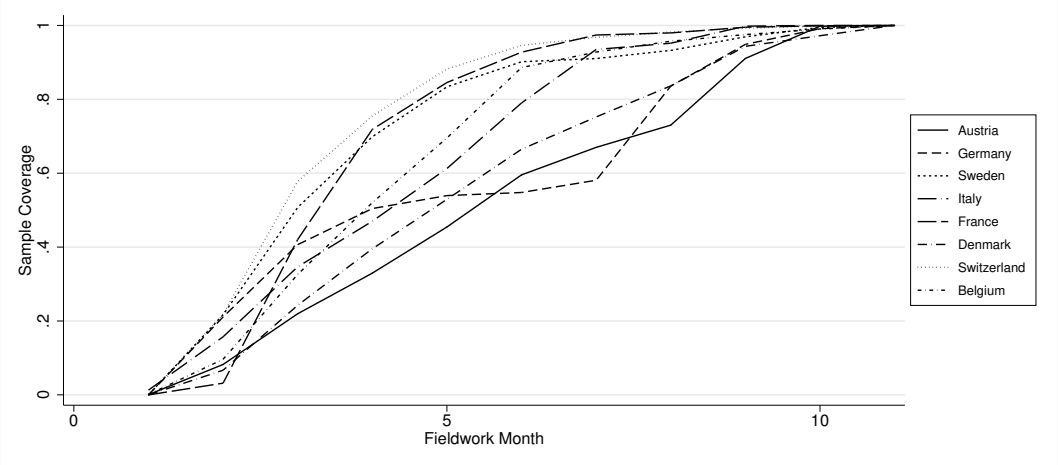


Figure 2.4: Sample Coverage by Month in SHARE WAVE 6 - Cumulative



It is evident that, although within a common fieldwork period, countries performed the fieldwork in very different patterns. For example, in Wave 6,

Switzerland and Sweden started the data collection process fast and within 3 months covered over 60% of their target sample, whereas Denmark and Austria had only about 20% in the same period. The patterns for all other waves are included in the appendix (Figures 2.5 and 2.6) and they all exhibit wide variations in the data collectin patterns.

Table 2.5 reports descriptive statistics, including the number of observations, mean and standard deviation for each variable included in the analysis.

Table 2.5: Descriptive Statistics

Variable	Observations	Mean	Std. Dev.
Log Food	68,854	6.147	0.641
Log Food Out	50,057	4.473	1.036
Log Food Home	69,680	5.973	0.613
Log Income	65,861	7.885	0.999
HH Size	77,520	1.951	0.902
Age	77,519	66.027	10.198
Retired	76,580	0.574	0.495
Employed	76,580	0.295	0.456
Unemployed	76,580	0.028	0.164
Sick/Disabled	76,580	0.031	0.173
Homemaker	76,580	0.073	0.260
N Children	75,230	2.074	1.377
Years Education	74,340	11.195	4.447
Female	77,520	1.559	.4972
First in Wave 1	77,520	0.355	0.479
First in Wave 2	77,520	0.128	0.334
First in Wave 4	77,520	0.316	0.465
First in Wave 5	77,520	0.176	0.381
First in Wave 6	77,520	0.025	0.155

2.3.4 Main Results

This section tests empirically the proposed solution in obtaining country-specific income elasticities, accounting for seasonality emerged from the asynchronous fieldwork. In order to obtain the following Engel curve, I estimate

the following model

$$\log F_{ict} = \beta_0 + \beta_1 \log Y_{ict} + \beta_3 M_{ict} + \beta_4 C_{ict} M_{ict} \log Y_{ict} + \beta_5 X_{ict} + \delta_c + \lambda_t + \epsilon_{ict} \quad (2.2)$$

where for each individual i in country c and year t, F_{ict} denotes food expenditure, Y_{ict} denotes income, λ_t is the year fixed effect, δ_c is the country fixed effect, M_{ict} is the month dummies and X_{ict} includes other covariates. The model is estimated using OLS with clustered standard errors, since the same households are observed over time.

Table 2.6 includes two models to examine the relationship between food expenditure (home and out combined) and income. As illustrated above in the case of Understanding Society, food outside home exhibits greater seasonality than at home.[17]

Both models include the interaction between country and income thus allowing the estimation of country-specific income elasticities. In addition, both models include commonly used covariates (gender, age, household size, education, number of children, labour market status) and they are all significant and in the expected direction. Moreover, wave and country fixed effects and an index for the first appearance in the panel are included in both models.

The key difference between Model 1 and Model 2 is that the former ignores the month of the interview whereas the latter includes the triple interaction of month dummies, country and income ($F_{96,22138} = 9.56$, p-value< 0.001).

[17]See Table 2.8 in the Appendix for the estimates of the relationship between the food outside home and income.

Table 2.6: Regression Log Food Home and Out

	Model 1 Without		Model 2 With	
	Coef	SE	Coef	SE
Log Income	0.092***	(0.008)	0.274***	(0.103)
HH Size	0.234***	(0.004)	0.234***	(0.004)
Age	-0.004***	(0.000)	-0.004***	(0.000)
Employed	-0.020**	(0.008)	-0.022***	(0.008)
Unemployed	-0.231***	(0.017)	-0.227***	(0.016)
Sick/Disabled	-0.180***	(0.017)	-0.178***	(0.017)
Homemaker	-0.026**	(0.011)	-0.026**	(0.011)
N children	-0.011***	(0.002)	-0.011***	(0.002)
Years Education	0.016***	(0.001)	0.016***	(0.001)
Female	-0.088***	(0.006)	-0.088***	(0.006)
First in Wave 2	0.018**	(0.009)	0.016*	(0.010)
First in Wave 4	-0.012	(0.009)	-0.012	(0.009)
First in Wave 5	-0.002	(0.009)	-0.017*	(0.009)
First in Wave 6	-0.052***	(0.017)	-0.063***	(0.017)
Wave FE	✓		✓	
Country FE	✓		✓	
Country*Log Income	✓		✓	
Month FE	X		✓	
Country*Month	X		✓	
Month*Log Income	X		✓	
Country*Month*Log Income	X		✓	
Constant	5.065***	(0.068)	3.723***	(0.753)
Observations	55,784		55,784	
R-squared	0.315		0.322	

*Clustered SE in brackets; *** p<0.01, ** p<0.05, * p<0.1; Ref for categorical: Retired, Male, First in Wave 1*

Table 2.7 presents the income elasticity (average marginal effects of income) by country, calculated from the regression model presented in Table 2.6. As expected, the income elasticity for food expenditure is positive (and smaller than 1) and statistically significant at the conventional levels of 5%.

Between the two models one can identify changes in the estimated coefficient when accounting for seasonality by as much as 0.04 in Sweden. The extent to which the coefficients for each country between the two models differ

Table 2.7: AMEs Log Food Home and Out

	Without seasonality		With seasonality	
Country	Coef	SE	Coef	SE
Austria	0.092***	(0.008)	0.104***	(0.011)
Germany	0.138***	(0.009)	0.130***	(0.010)
Sweden	0.205***	(0.013)	0.223***	(0.012)
Italy	0.097***	(0.007)	0.085***	(0.011)
France	0.150***	(0.009)	0.172***	(0.016)
Denmark	0.142***	(0.009)	0.139***	(0.011)
Switzerland	0.102***	(0.008)	0.107***	(0.013)
Belgium	0.117***	(0.009)	0.119***	(0.013)
Observations	55,784		55,784	

Clustered SE in brackets; *** p<0.01, ** p<0.05, * p<0.1

is driven by the fieldwork characteristics in each country (i.e. more equally distributed across months or more frequent in some months than others). The implication of these results may be significant for comparisons across countries. For example, in Model 1, Austria appears to have a more inelastic income elasticity than Germany. However, when accounting for the fieldwork characteristics, the two countries appear to have, in fact, the same income elasticity (in Model 2, Austria's income elasticity is bigger but not statistically different than in Germany).

Looking more broadly, the ranking of income elasticities across countries may change significantly, as shown in Model 2. As such, any cross-country comparisons based on Model 1 would be misleading, as the coefficients reflect the asynchronous fieldwork.

2.4 Conclusion

This paper emphasizes the importance of controlling for asynchronicity and the likely biases arising when this data-collection feature is omitted from

analyses. The main application of this paper focuses on food consumption, though, as explained previously, asynchronicity is likely to affect quantities of interest one may not necessarily expect to be seasonal (such as waist circumference, BMI). Further applications may include doctor visits or psychical activity relevant to health economics or life-satisfaction and happiness relevant to happiness economics.

Broadly speaking, cross-country surveys are more susceptible to asycnrhonicity, be it direct or indirect. Cross-country biases are illustrated using the SHARE data in a direct comparison with a national survey (Understanding Society). An important difference between the two surveys employed is the target populations, with the former including only those aged 50 or older and the latter including representative samples across all age groups. Interestingly, the results report substantial differences between models accounting for and those omitting seasonality, *even when* these are based on reported food consumption outside the home for elderly people. Clearly one would expect this to vary a lot less for older age groups. Consequently, the results could have been even more striking had the survey included more varied age groups.

Overall, this paper emphasizes two areas of recommendation. First, survey designers are encouraged to organize the fieldwork not only within a common period but, crucially, with a similar pattern across months within this period. Ideally, this should be done by collecting monthly samples. However, more flexible alternatives may be considered, such as imposing various benchmarks by which a certain proportion of the total sample should be interviewed.

A second recommendation concerns researchers who are advised to include seasonal dummies in their estimations in order to avoid erroneous cross-country comparisons. This solution is not unique and alternative methods

accounting for seasonality may be considered, such as weighting or matching. A likely limitation researchers may face is of course the under-sampling of respondents in each given month. This, however, may be corrected depending on the application of interest. For example, for labour-market applications researchers may group adjacent months into quarters, while for health or health-utilization outcomes (such as doctor visits or waiting times) one may simply consider two seasons, one including the months expected to be busiest and one including the remaining months.

This paper makes a contribution to the literature on cross-country modelling and provides a sensible recommendation in dealing with seasonality. Although the proposed solution is trivial, seasonal-induced biases have been largely neglected in the literature. The conclusions drawn in this paper are generalizable to most other European cross-national surveys. Crucially, this paper makes evident that not accounting for seasonality might lead to comparisons between some nations during summer and others during winter.

2.5 Appendix

Table 2.8: Understanding Society Regression Log Food Out

	Model 1		Model 2		Model 3	
	Coef	SE	Coef	SE	Coef	SE
Log Income	0.409***	(0.009)	0.409***	(0.009)	0.392***	(0.020)
Log Prices	1.867***	(0.483)	-1.080	(1.835)	-1.072	(1.835)
HH Size	0.106***	(0.006)	0.106***	(0.006)	0.106***	(0.006)
Female	-0.105***	(0.009)	-0.107***	(0.009)	-0.107***	(0.009)
Age	-0.003***	(0.000)	-0.003***	(0.000)	-0.003***	(0.000)
N Children	-0.088***	(0.007)	-0.088***	(0.007)	-0.087***	(0.007)
Unemployed	-0.215***	(0.020)	-0.216***	(0.020)	-0.216***	(0.020)
Retired	-0.025	(0.016)	-0.025	(0.016)	-0.025	(0.016)
Homemaker	-0.074***	(0.018)	-0.073***	(0.018)	-0.073***	(0.018)
Disabled	-0.365***	(0.024)	-0.364***	(0.024)	-0.364***	(0.024)
Other empl.	-0.055***	(0.021)	-0.052**	(0.021)	-0.052**	(0.021)
Other qual.	0.192***	(0.012)	0.192***	(0.012)	0.192***	(0.012)
Degree qual.	0.360***	(0.015)	0.361***	(0.015)	0.361***	(0.015)
Year FE	✓		✓		✓	
Month FE	X		✓		✓	
Month*Income	X		X		✓	
Constant	-7.605***	(2.142)	5.418	(8.104)	5.517	(8.102)
Observations	80,070		80,070		80,070	
R-squared	0.186		0.187		0.187	

*Clustered SE in brackets; *** p<0.01, ** p<0.05, * p<0.1; Ref for categorical: Male, Employed, No qualifications*

Table 2.9: Understanding Society AMEs Log Food Out

	Model 1		Model 3	
	Coef	SE	Coef	SE
January			0.392***	(0.020)
February			0.376***	(0.022)
March			0.402***	(0.022)
April			0.417***	(0.022)
May			0.369***	(0.023)
June			0.415***	(0.021)
July			0.428***	(0.022)
August			0.443***	(0.023)
September			0.399***	(0.022)
October			0.424***	(0.021)
November			0.434***	(0.021)
December			0.417***	(0.024)
Overall	0.410***	(0.009)	0.410***	(0.009)

*Clustered SE in brackets *** p<0.01, ** p<0.05, * p<0.1*

Figure 2.5: Sample Coverage by Month in SHARE WAVE 1, 2, 4 and 5

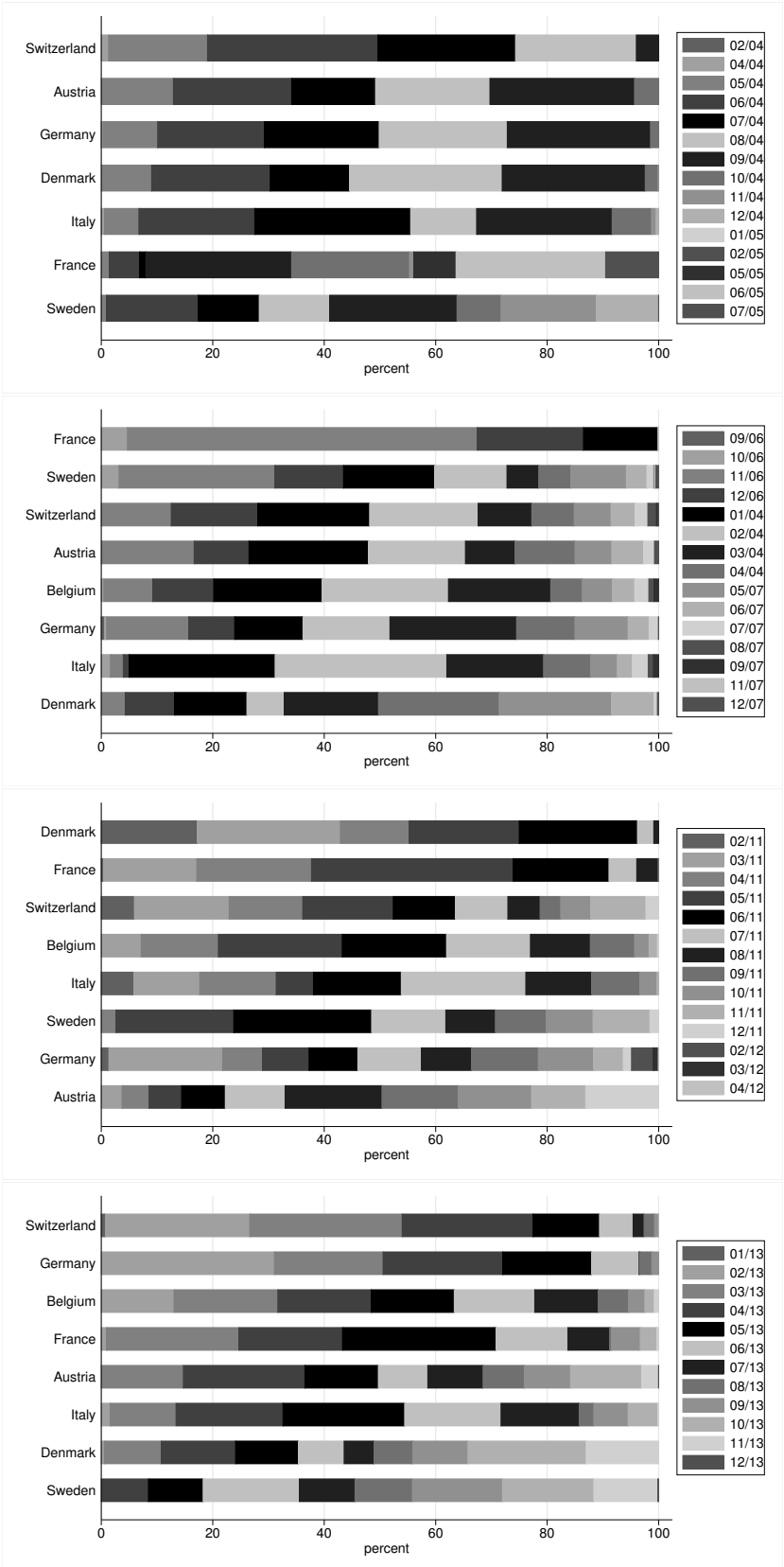


Figure 2.6: Sample Coverage by Month in SHARE WAVE 1, 2, 4 and 5 Cumulative

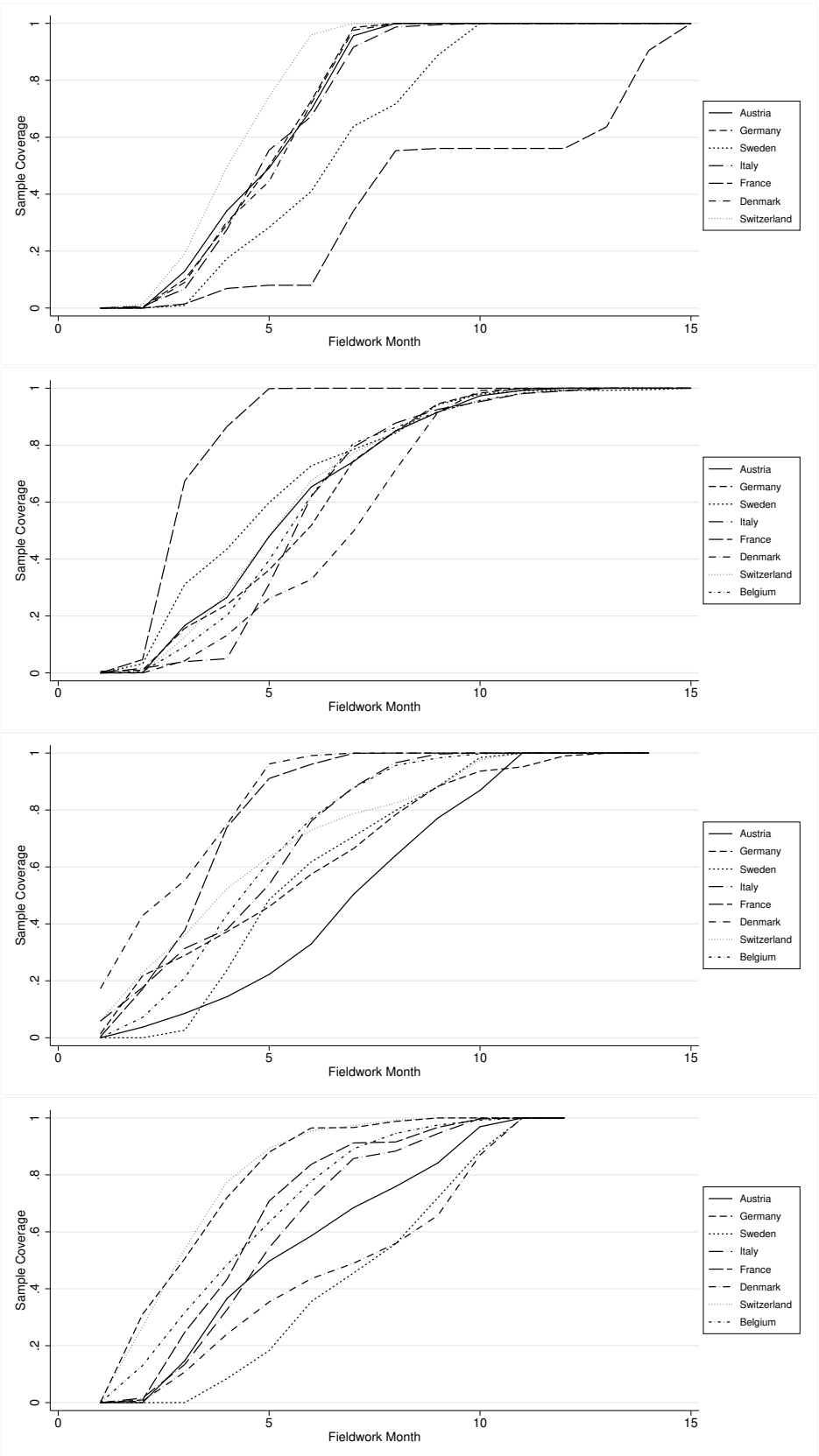


Table 2.10: SHARE Regression Log Food Out

	Model 1 Without		Model 2 With	
	Coef	SE	Coef	SE
Log Income	0.153***	(0.015)	0.522***	(0.174)
HH Size	0.094***	(0.007)	0.093***	(0.007)
Age	-0.004***	(0.001)	-0.004***	(0.001)
Employed	0.139***	(0.016)	0.136***	(0.016)
Unemployed	-0.236***	(0.036)	-0.234***	(0.036)
Sick/Disabled	-0.170***	(0.034)	-0.173***	(0.034)
Homemaker	0.083***	(0.026)	0.078***	(0.026)
N Children	-0.037***	(0.005)	-0.037***	(0.005)
Years Education	0.021***	(0.001)	0.020***	(0.001)
Female	-0.208***	(0.011)	-0.208***	(0.011)
First in Wave 2	0.020	(0.020)	0.009	(0.021)
First in Wave 4	0.041**	(0.018)	0.021	(0.018)
First in Wave 5	0.044**	(0.018)	0.023	(0.019)
First in Wave 6	-0.001	(0.033)	-0.041	(0.034)
Wave FE	✓		✓	
Country FE	✓		✓	
Country*Log Income	✓		✓	
Month FE	X		✓	
Country*Month	X		✓	
Month*Log Income	X		✓	
Country*Month*Log Income	X		✓	
Constant	3.288***	(0.138)	0.631	(1.296)
Observations	39,975		39,975	
R-squared	0.165		0.173	

*Clustered SE in brackets; *** p<0.01, ** p<0.05, * p<0.1; Ref for categorical: Retired, Male, First in Wave 1*

Table 2.11: SHARE AMEs Log Food Out

	Without seasonality		With seasonality	
Country	Coef	SE	Coef	SE
Austria	0.153***	(0.015)	0.181***	(0.021)
Germany	0.202***	(0.016)	0.173***	(0.019)
Sweden	0.312***	(0.027)	0.355***	(0.026)
Italy	0.125***	(0.018)	0.142***	(0.028)
France	0.155***	(0.021)	0.142***	(0.038)
Denmark	0.114***	(0.016)	0.108***	(0.021)
Switzerland	0.154***	(0.016)	0.135***	(0.021)
Belgium	0.159***	(0.016)	0.160***	(0.021)

*Clustered SE in brackets; *** p<0.01, ** p<0.05, * p<0.1*

Chapter 3

Regression with Imputed Dependent Variables

3.1 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 of 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 to impute that dependent variable into the data set containing the independent variable. For example, in a very well known paper, Skinner (1987) (hereafter **SK**) proposed

a method for using the U.S Consumer Expenditure Survey (CE) to impute a consumption measure into the Panel Survey of Income Dynamics (PSID).[1] 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 **SK** procedure leads to an inconsistent estimate of the regression coefficient of interest. We show that the asymptotic bias is equal to the R^2 of the first stage regression of the variable to be imputed on the proxy or proxies. This leads us to suggest a 'rescaled-Skinner' (hereafter **RSK**) procedure.

We then show that with a single proxy, the **RSK** procedure is numerically identical to a procedure developed by Blundell et al. (2004, 2008) (hereafter **BPP**) in which the first stage involves, in contrast to **SK**, regressing the proxy on the variable to be imputed, and then inverting. We further show that the usual OLS standard errors from a regression of an imputed dependent variable (derived from the **RKS** or **BPP** procedures) are incorrect, and provide an estimator of the asymptotic standards errors of the regression coefficient of interest.

Lusardi (1996) combines CE consumption data with PSID income data with the 2-sample IV approach proposed by Angrist and Krueger (1992). We clarify the relationship between that approach and the imputation procedures

[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 investment spending can be important (Crossley et al., 2017). Second, for the application we have in mind, this procedure results in income or wealth being on both the right and left-hand side of the equation so that any measurement error can cause quite serious problems (Browning et al., 2014).

we study. We also show how the precision of imputation procedures can be improved using an adjustment for finite sample differences between the data sets that is analogous to the advantage of 2-sample-2-stage least squares over 2-sample-IV described in Inoue and Solon (2010).

We illustrate these points with a Monte Carlo study and with an empirical example with using the CE and PSID.

In the next section we lay out our basic framework, derive the main results, and relate them to the prior literature. Section 3 takes up the question of inference. Section 4 illustrates with a small Monte Carlo experiment and an application to the CE and PSID. Finally, Section 5 concludes.

3.2 Basic Setup And Results

3.2.1 Basic Setup

Consider estimating the regression

$$C = X\beta + \epsilon \tag{3.1}$$

where β is the parameter of interest. To make things concrete, C could be (nondurable) consumption, and X a vector including 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) makes no difference to the analysis that follows.

Assume that the usual regression assumptions hold, so that β could be consistently estimated by Ordinary Least Squares if we had complete data. In particular, $plim(\frac{1}{n_j}X_j'X_j) = \Sigma_{XX}$ and $plim(\frac{1}{n_j}X_j'\epsilon_j) = 0$ for any representative

sample j.[2] However, we have no data that allows us to calculate the empirical analogue $(X_j'C_j)$ of the population covariances $X'C$. Subscripts $j = 1, 2...$ index the data set (or sample); absence of a subscript indicates a population quantity.

We do have data on (C_1, Z_1) and (X_2, Z_2). Z is our proxy for C. Both data sets are 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 proxy and the dependent variable of interest. With total nondurable consumption, as our quantity of interest, and food consumption, as a proxy, this relationship is an Engel Curve:

$$Z = C\gamma + u. \tag{3.2}$$

This implies a reduced form relationship between food spending and income:

$$Z = X\beta\gamma + \epsilon\gamma + u. \tag{3.3}$$

Note that Equation (3.2) makes clear that Z *must* depend on ϵ (Z has some information about C that is not contained in X). As we will elaborate below, this is the opposite to what is required for two-sample IV. Given Z with these properties, one can *impute* C *using* Z.

For clarity of exposition, we begin with the cross-sectional case and abstract from additional covariates in either the consumption function (3.1) or Engel curve (3.2). We will also initially assume that $plim(\frac{1}{n}C'u) = 0$. This would fail, for example, if there were measurement error in in C. Below, we

[2] If the samples are not drawn from population in the same way, this might be overcome by inverse probability weighting.

expand on additional covariates, panel data, and measurement error in C.

3.2.2 Alternative Imputation Strategies

In an early paper, Skinner (1987), suggested regressing C_1 on Z_1 in the CE and using the resulting coefficient to predict $\hat{C}_2$ in the PSID (and then regressing $\hat{C}_2$ on X_2). Note the first stage here is an 'inverse' Engel curve. This procedure was advocated by Browning et al. (2003) and recent applications include Attanasio and Pistaferri (2014) and Arrondel et al. (2015).

Alternatively, Blundell et al. (2004, 2008), again using the CE and PSID, first regress Z_1 on C_1 to get $\hat{\gamma}$, and then predict $\hat{C}_2 = Z_2\frac{1}{\hat{\gamma}}$. That is, they estimate an Engel curve and then invert it to predict consumption. This procedure has also recently been employed by Attanasio et al. (2012).

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, one could regress Z_1 on C_1 to get $\hat{\gamma}$, then regress Z_2 on X_2 to get $\widehat{\beta\gamma}$ (Equation (3.3)), and take ratio of the two to estimate β.

We consider first the **SK** procedure. Note that by iterative expectations, the **SK** procedure gives an unbiased estimate of the first moment of consumption $E[\hat{C}_2^{SK}] = E[C]$. However, the regression of $\hat{C}_2^{SK}$ on X_2 does not give a consistent (or unbiased) estimate of β.

Proposition 1. *Under the assumptions that both samples are random samples drawn from the same population, the usual regression assumptions and that $Cov(C, Z) \neq 0$: $plim(\hat{\beta}^{SK}) = \beta R^2$ where R^2 is the (population) R^2 corresponding to the first-stage imputation regression of C on Z.*

Proof. Note that the imputed dependent variable is $\hat{C}_2^{SK} = X_2' Z_2 (Z_1' Z_1)^{-1} Z_1' C_1$ and the regression of $\hat{C}_2^{SK}$ on X_2 gives:

$$\hat{\beta}^{SK} = (X_2' X_2)^{-1} X_2' Z_2 (Z_1' Z_1)^{-1} Z_1' C_1 \tag{3.4}$$

and with simple algebra the probability limit of this is:

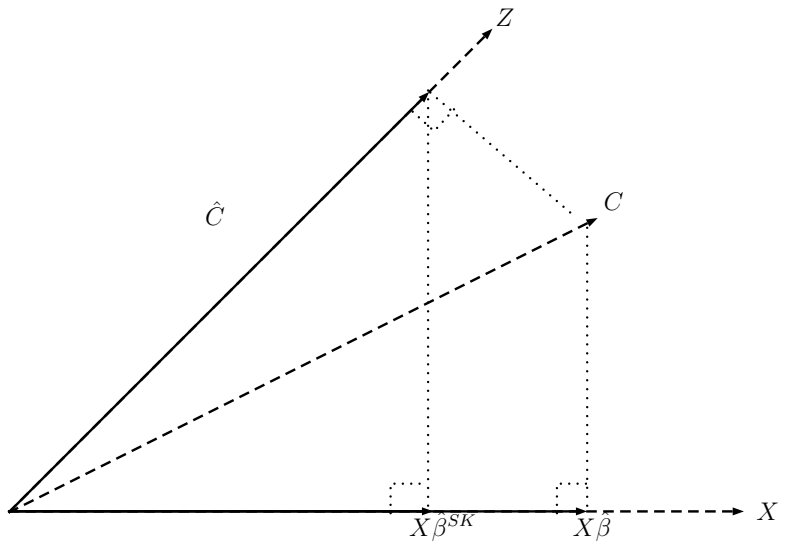
$$\begin{aligned} plim(\hat{\beta}^{SK}) &= plim \left\{ \frac{(X_2' X_2)^{-1}}{n_2^{-1}} \frac{X_2' Z_2}{n_2} \right\} \times plim \left\{ \frac{(Z_1' Z_1)^{-1}}{n_1^{-1}} \frac{Z_1' C_1}{n_1} \right\} \\ &= \beta\gamma \times \frac{1}{\gamma} R_{zc}^2 \\ &= \beta R_{zc}^2 \neq \beta \end{aligned}$$

where the first part comes directly from Equation (3.3) as $plim(\widehat{\beta\gamma}) = \beta\gamma$ and the second part is the reverse regression of a simple regression. Using the fact that the product of the two coefficients from the two regressions is the R^2 from either of them, then the probability limit of the reverse regression is simply $\frac{1}{\gamma} R_{zc}^2$, where R_{zc}^2 denotes the population R-squared. □

Figure 3.1 illustrates the problem graphically. The dashed vectors C, Z and X represent data. The solid black vector $X\beta$ is the orthogonal projection of C onto X (which would be obtained by regression with complete data). The **SK** procedure first projects C onto Z, giving the solid black vector $\hat{C}$, and then projects this vector onto X giving the solid black vector $X\beta^{SK}$. Note that $X\beta^{SK} < X\beta$.

Note that the first stage R^2 for food Engel curves and 'inverse' Engel curves are typically between 50 and 70%. In terms of the size of the bias, this implies inflation factors of between 1.4 and 2 (or downward bias of between 30 and 50%).

Figure 3.1: Skinner Imputation Procedure as Projections



As the bias in the **SK** procedure is an estimable quantity, it can be corrected. One can rescale $\hat{\beta}^{SK}$ by the estimated first stage R^2_{ZC}. We refer to the resulting estimate of β as the 're-scaled Skinner' (hereafter **RSK**, $\hat{\beta}^{RSK}$).

Proposition 2. *Under the assumptions that both samples are random samples drawn from the same population, the usual regression assumptions and that* $Cov(C, Z) \neq 0$*: The* ***RSK*** *procedure results in a consistent estimate of* β*.*

$$plim(\hat{\beta}^{RSK}) = \beta \tag{3.5}$$

Proof. Note that

$$\hat{\beta}^{RSK} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'C_1[C_1'Z_1(Z_1'Z_1)^{-1}Z_1'C_1]^{-1}C_1'C_1 \tag{3.6}$$

Because $Z_1'C_1$ and $Z_1'Z_1$ are scalars, this reduces to:

$$\hat{\beta}^{RSK} = (X_2'X_2)^{-1}X_2'Z_2(C_1'Z_1)^{-1}C_1'C_1 \tag{3.7}$$

which has probability limit:

$$plim(\hat{\beta}^{RSK}) = \beta \frac{\gamma\beta\Sigma_{XX} + \gamma\sigma_\epsilon^2}{\gamma\beta\Sigma_{XX} + \gamma\sigma_\epsilon^2} = \beta. \tag{3.8}$$

Note that this rescaling of $\hat{\beta}^{SK}$ is equivalent to rescaling the predicted consumption vector $\hat{C}_2^{SK}$ by $1/\widehat{R^2_{ZC}}$ prior to using it as the dependent variable in regression. Of course, the resulting re-scaled vector of imputed values does not have the correct first moment.

$$E\left[\frac{1}{\widehat{R^2_{ZC}}}\hat{C}_2^{SK}\right] \neq E[C] \tag{3.9}$$

□

Now consider the **BPP** procedure, with resulting estimate $\hat{\beta}^{BPP}$.

Proposition 3. *In the case of one proxy and under the assumptions that both samples are random samples drawn from the same population, the usual regression assumptions and that $Cov(C, Z) \neq 0$: The **BPP** procedure is numerically identical to **RSK**. It is therefore also consistent.*

Proof.

$$\hat{\beta}^{BPP} = (X_2'X_2)^{-1}X_2'Z_2(C_1'Z_1)^{-1}C_1'C_1 = \hat{\beta}^{RSK} \tag{3.10}$$

□

Finally, the **AM** procedure takes the ratio of $\widehat{\beta\gamma} = (X_2'X_2)^{-1}X_2'Z_2$ and $\hat{\gamma} = (C_1'C_1)^{-1}C_1'Z_2$, to give $\hat{\beta}^{AM} = \widehat{\beta\gamma}/\hat{\gamma}$.

Proposition 4. *In the case of one proxy and under the assumptions that both samples are random samples drawn from the same population, the usual*

regression assumptions and that $Cov(C, Z) \neq 0$*:* $\hat{\beta}^{AM}$ *is numerically identical to* $\hat{\beta}^{RSK}$ *and* $\hat{\beta}^{BPP}$ *and consistent.*

Proof.

$$\begin{aligned}\hat{\beta}^{AM} = \widehat{\beta\gamma}/\hat{\gamma} &= (X_2'X_2)^{-1}X_2'Z_2\left[(C_1'C_1)^{-1}C_1'Z_2\right]^{-1} \\ &= (X_2'X_2)^{-1}X_2'Z_2(C_1'Z_1)^{-1}C_1'C_1 = \hat{\beta}^{RSK} = \hat{\beta}^{BPP} \end{aligned} \tag{3.11}$$

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

Note that because it is numerically identical to the **RSK** procedure, the **BPP** procedure gives biased estimates of the first moment. **AM** recovers β directly, and does not generate unit level estimates of C.

It is also useful to think about second moments, as these imputation procedures have been used to study consumption inequality, as well as wealth and income effects (**BPP**, for example, consider consumption inequality). Simple algebra (similar to above) shows that:

$$Asymp\ Var(\hat{C}^{SK}) = Asymp\ Var(C) \times R^2_{C,Z} \tag{3.12}$$

and:

$$Asymp\ Cov(\hat{C}^{SK}, X) = Asymp\ Cov(C, X) \times R^2_{C,Z}. \tag{3.13}$$

Note that with a scalar X the OLS estimate of β is just $Cov(\hat{C}, X)/Var(X)$ this gives an additional intuition for the bias in $\hat{\beta}^{SK}$. For the **BPP** or **RSK** estimates of C we have:

$$Asymp\ Var(\hat{C}^{RSK}) = Asymp\ Var(\hat{C}^{BPP}) = Asymp\ Var(C)/R^2 \tag{3.14}$$

and:

$$Asymp\ Cov(\hat{C}^{RSK}, X) = Asymp\ Cov(\hat{C}^{BPP}, X) = Asymp\ Cov(C, X) \quad (3.15)$$

Thus,

$Asymp\ Var(\hat{C}^{RSK}) = Asymp\ Var(\hat{C}^{BPP}) > Asymp\ Var(C) > Asymp\ Var(\hat{C}^{SK})$.

Attanasio and Pistaferri (2014) show that trends in $Var(\hat{C}^{BPP})$ and $Var(C)$ are similar, but there is a level difference. The similarity in trends suggests that the first-stage (imputation) R^2_{ZC} is roughly constant across years in their data.

3.2.3 Related Problems

As is well known, classical measurement error in the independent variable causes attenuation bias in simple regression, but classical measurement error in the dependent variable does not.[3] The measurement error induced by imputation is instead a Berkson measurement error (or prediction error). It is also widely recognized that Berkson measurement error in an independent variable does not cause bias in a simple regression (Berkson, 1950; Wansbeek and Meijer, 2000)[4] What appears to be much less known is that Berkson errors in a dependent variable do cause bias.

Hyslop and Imbens (2001) show attenuation bias in a regression of $\hat{C}$ on X where $\hat{C}$ is an optimal linear prediction generated by a survey respondent (not the econometrician). Relative to the imputation problem we study, key

[3]The intuition is that if $\tilde{X} = X + \tilde{v}$, with $\tilde{v} \perp X$, then $C = \tilde{X}\beta - \tilde{v}\beta + \epsilon$ and $\tilde{v}$ is correlated with $\tilde{X}$ by construction; in contrast if $\tilde{C} = C + \tilde{v}$, with $\tilde{v} \perp C$ then $\tilde{C} = X\beta + \tilde{v} + \epsilon$ and the error terms is uncorrelated with X.

[4]If $X = \hat{X} + \hat{v}$, but $\hat{v} \perp \hat{X}$, then $C = \hat{X}\beta - \hat{v}\beta + \epsilon$ but the error term is not correlated with the right hand side observable.

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 $Z = C + u$; ($\gamma = 1$). Hoderlein and Winter (2010) study a similar problem to Hyslop and Imbens, but in a nonparametric setting. Again, in their model it is the survey respondent, rather than the econometrician doing the predicting.[5]

The case of proxies for independent/explanatory variables has been studied by Lubotsky and Wittenberg (2006) and Bollinger and Minier (2015). These studies discuss the case of using multiple proxies for an unobserved variable of interest within one single sample.

It is also useful to contrast the imputation procedures studied in this paper with the two-sample IV (2SIV) approach first suggested by Angrist and Krueger (1992) and applied to the combination of CE consumption data and PSID income data by Lusardi (1996). The estimator is:

$$\hat{\beta}^{2SIV} = \left(\frac{Z_2' X_2}{n_2}\right)^{-1} \left(\frac{Z_1' C_1}{n_1}\right) \tag{3.16}$$

and Z is typically a grouping variable (e.g.. birth cohort, occupation, birth cohort x education).

The key assumption is that $E[Z'\epsilon] = 0$ (Z affects C only through X), which is the polar opposite to the assumption necessary to use Z as a proxy (as noted above, a useful proxy must have information about C over and above the information in X). With 2SIV, we effectively use Z to impute X.

One virtue of this procedure is that measurement error in C 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

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

the use 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.

3.2.4 Extensions

Covariates

Additional covariates can be added to both the imputation equation and the equation of interest (in our example, the Engel curve and the consumption equation). Let the additional covariates be W_j and $M_{WJ} = I_j - W_j'(W_j'W)_j^{-1}W_j$ is the orthogonal projection matrix for W_j (where I_j is an identity matrix and again $J = 1, 2$ indexes samples). Following the Frisch-Waugh-Lovell theorem we could purge Z_j, C_j and X_j of W_j. Using these adjusted variables (and noting that M_{WJ}, $\hat{C_2}^{SK} = X_2'M_{W2}Z_2(Z_1'M_{W1}Z_1)^{-1}Z_1'M_{W1}C_1$) and the regression of $\hat{C_2}^{SK}$ on $M_{W2}X_2$ gives:

$$\hat{\beta}^{SK} = (X_2'M_{W2}X_2)^{-1}X_2'M_{W2}Z_2(Z_1'M_{W1}Z_1)^{-1}Z_1'M_{W1}C_1 \tag{3.17}$$

Rescaling by the first stage R^2 gives

$$\hat{\beta}^{RSK} = (X_2'M_{W2}X_2)^{-1}X_2'M_{W2}Z_2(C_1'M_{W1}Z_1)^{-1}C_1'M_{W1}C_1 \tag{3.18}$$

Denoting $plim(\frac{X_j'M_{Wj}X_j}{n_j}) = \Sigma_{XX|W}$, and taking the probability limit:

$$plim(\hat{\beta}^{RSK}) = \beta\frac{\gamma\beta\Sigma_{XX|W} + \gamma\sigma_\epsilon^2}{\gamma\beta\Sigma_{XX|W} + \gamma\sigma_\epsilon^2} = \beta. \tag{3.19}$$

Inspection of the algebra above makes clear that consistency follows only if C_1 and X_2 are both purged of W_j, or equivalently, if the same covariates are added to both the Engel curve and the consumption equation.[6]

Panel case

Often a researcher wishes to estimate $\Delta C = \Delta X\beta + \Delta\epsilon$ where $\Delta C = C^1 - C^0$ (superscripts denote time). As before, β is the main object of interest and could be estimated consistently by OLS if we had complete data (that is, $plim(\Delta X \Delta\epsilon) = 0$). Suppose we have no data from which to compute $\frac{1}{N}\sum \Delta C * \Delta X$, but do have have some data on $(C_1^1, Z_1), (C_2^0, Z_2), (\Delta X_3, Z_3)$. An obvious example is a repeated cross-sectional household budget survey combined with a panel survey on income and wealth survey. C_3 can then be imputed year by year. It is easy to show that $\hat{\beta}^{RSK}$ and $\hat{\beta}^{BPP}$ remain consistent and numerically identical in this case.

Measurement error in C

Suppose that C is measured with error. This would be a natural concern if C is consumption expenditure, which is a difficult quantity to measure, even in a detailed household budget survey. Even if this measurement error is classical, it is obvious that both the **BPP** and the **AM** procedures require an instrument for C, as both involve a regression of Z_1 on C_1 to get $\hat{\gamma}$. If $plim(\frac{C_1' u_1}{n_1}) \neq 0$ because u_1 contains incorporates the measurement error in C_1, then an instrument for C is required to obtain a consistent estimate of γ.

With the **RSK** procedure, C_1 is the independent variable in the first-stage imputation regression, so that classical measurement error in C_1 does

[6] In the same way that, with IV, exogenous covariates in the equation of interest should be added to the first stage.

not lead to an inconsistent estimate of the regression slope. However, classical measurement error in C_1 does still cause a problem because it leads to an inconsistent estimate of the population first-stage R^2_{ZC}. This can be overcome by estimating R^2_{ZC} as the produce 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 C_1 is the independent variable).[7]

3.3 Inference and precision.

3.3.1 Asymptotic Standard Errors

The direct estimation of (3.1) on complete data would result in an asymptotic variance for $\hat{\beta}$ of $(\Sigma_{XX})^{-1}\sigma^2_\epsilon$. When we impute the dependent variable, $\hat{\beta}^{AM}$ $\hat{\beta}^{RSK}$ and $\hat{\beta}^{BPP}$ are numerically identical, so we derive the asymptotic variance from the **AM** approach. The Engel curve (3.2) and reduced form (3.3) give two moments:

$$plim\left(\frac{C_1'(Z_1-\gamma C_1)}{n_1}\right) = plim\left(\frac{C_1'u_1}{n_1}\right) = 0$$

$$plim\left(\frac{X_2'(Z_2-\gamma\beta X_2)}{n_2}\right) = plim\left(\frac{X_2'(\gamma u_2+\epsilon)}{n_2}\right) = 0$$

which identify the parameters γ and β.

It is informative to first consider implementing $\hat{\beta}^{AM}$ (or equivalently $\hat{\beta}^{BPP}$ or $\hat{\beta}^{RSK}$) on a single sample, containing all of C, Z, X (of course, a researcher would have no reason to do this, but it delivers a useful intuition).

[7] Of course, given the numerical equivalence of the **RSK**, **BPP** and **AM** procedures, it cannot be that one offers an advantage over the other two in dealing with measurement error in C_1.

Denote $plim(\frac{C_1'C_1}{n_1}) = \Sigma_{CC}$. In this one-sample case, the asymptotic variance-covariance matrix of the moments is:

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

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

$$Asymp\ Var(\hat{\beta}) = \frac{(\Sigma_{XX})^{-1} \sigma_\epsilon^2}{R_{ZC}^2} \tag{3.20}$$

Thus the loss of asymptotic precision, due to imputation and relative to the direct estimation of (3.1), is proportional to the first stage R_{ZC}^2. Note the similarity in the loss of precision with linear IV estimation relative to OLS instrumental variables, which is also proportional to the first stage R^2 (Shea, 1997).

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

$$F = \begin{bmatrix} \sigma_u^2 \Sigma_{CC} & 0 \\ 0 & \left(\gamma^2\sigma_u^2 + \sigma_\epsilon^2\right) \Sigma_{XX} \end{bmatrix}$$

where 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 $(\Sigma_{CC})^{-1} \sigma_u^2$. The

asymptotic variance of $\hat{\beta}$ is:

$$(\Sigma_{XX})^{-1}\left(\sigma_\epsilon^2+\gamma^{-2}\sigma_u^2\right)+(\Sigma_{CC})^{-1}\beta^2\gamma^{-2}\sigma_u^2.$$

$$=(\Sigma_{XX})^{-1}\sigma_\epsilon^2+\gamma^{-2}(\Sigma_{XX})^{-1}\sigma_u^2+\beta^2\gamma^{-2}(\Sigma_{CC})^{-1}\sigma_u^2$$

This can be written as:

$$=\frac{(\Sigma_{XX})^{-1}\sigma_\epsilon^2}{R_{ZC}^2}+2\beta^2\left(\frac{1-R_{ZC}^2}{R_{ZC}^2}\right) \tag{3.21}$$

The second term in this expression represents a *second* loss of asymptotic precision, due to the use of two different samples.

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

$$\begin{aligned}\hat{V}^{OLS}(\hat{\beta}^{BPP}) &= (X_2'X_2)^{-1}\left(\hat{C}_2-X_2\hat{\beta}\right)'\left(\hat{C}_2-X_2\hat{\beta}\right)\\ &= (X_2'X_2)^{-1}\left[\hat{C}_2'\hat{C}_2-\hat{C}_2'X_2(X_2'X_2)X_2'\hat{C}_2\right]\\ &= (X_2'X_2)^{-1}\big[C_1'C_1(Z_1'C_1)^{-1}Z_2'Z_2(Z_1'C_1)^{-1}C_1'C_1\\ &\quad -C_1'C_1(Z_1'C_1)^{-1}Z_2'X_2(X_2'X_2)X_2'Z_2(Z_1'C_1)^{-1}C_1'C_1\big]\end{aligned}$$

With some algebra, it is straightforward to show that:

$$\begin{aligned}plim\left[\hat{V}^{OLS}(\hat{\beta}^{BPP})\right] &= \left[\frac{(\Sigma_{XX})^{-1}\sigma_\epsilon^2}{R_{ZC}^2}+\beta^2\left(\frac{1-R_{ZC}^2}{R_{ZC}^2}\right)\right]\\ &= Asym\ Var(\hat{\beta}^{BPP})-\beta^2\left(\frac{1-R_{ZC}^2}{R_{ZC}^2}\right)\end{aligned} \tag{3.22}$$

So the usual OLS standard errors are too small, by the factor $\beta^2 \left(\frac{1-R^2_{ZC}}{R^2_{ZC}}\right)$ but can be corrected using the available consistent estimates of β and R^2_{ZC}.[8]

3.3.2 Finite-sample improvement

For the case where Z is an instrument, 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 a Two-stage Least Squares (finite-sample) improvement. Their estimator is

$$\hat{\beta}^{TS2SLS} = \left(\hat{X}_1'\hat{X}_1\right)^{-1}\hat{X}_1'C_1 \tag{3.23}$$

where $\hat{X}_1 = Z_1(Z_2'Z_2)^{-1}Z_2'X_2$. We can express this as:

$$\hat{\beta}^{TS2SLS} = \left(\frac{Z_2'X_2}{n_2}\right)^{-1} W_{12}\left(\frac{Z_1'C_1}{n_1}\right) \tag{3.24}$$

where $W_{12} = (Z_2'Z_2/n_2)(Z_1'Z_1/n_1)^{-1}$.

Similarly, we can improve the finite sample precision of the RSK estimator ($\hat{\beta}^{RSK}$) by accounting for the fact that Z_1 and Z_2 are different in finite samples. Define the corrected-RSK estimator as:

$$\hat{\beta}^{cRSK} = (X_2'X_2)^{-1}X_2'Z_2W_{12}(Y_1'Z_1)^{-1}C_1'C_1 \tag{3.25}$$

where W_{12} is the correction matrix for differences between the two samples as defined above. Note that with a single proxy, W_{12} is a scalar. We illustrate this finite-sample improvement in the next section.

[8] A Stata command that implements the correct standard errors is available from the authors.

3.4 Illustrations

3.4.1 Monte Carlo Experiment

To illustrate the points made above we first present a small Monte Carlo Experiment. There is a single regressor (income) $X \sim U(-2, 2)$. The dependent variable of interest is $C = 1.0X + \epsilon$ with $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\sigma_\epsilon^2 = 2$. The parameter of interest is $\beta = 1.0$.

We will consider the case where we cannot regress C on X directly, because information on these quantities is collected in separate surveys (we only observe C_1 and X_2, so that we cannot calculate the empirical covariance, $Cov(C_1, X_1)$ or $Cov(C_2, X_2)$). However, both surveys contain a potential proxy for C, Z. We generate this as follows:

$$Z_1 = 0.5C_1 + u_1$$

$$Z_2 = 0.5C_1 + u_2$$

with $u_1, u_2 \sim N(0, \Sigma)$ where $\Sigma = (\sigma_{u_1}^2, 0.6 \backslash 0.6, 3)$. We consider two cases, with $\sigma_{u_1}^2 = 2$ and $\sigma_{u_1}^2 = 4$. These imply a first stage R^2 of respectively 0.4 and 0.25. We simulate this population multiple times, each time implementing the **SK** and **RSK** procedures, with and without the Inoue-Solon-type finite sample correction. The results are presented in Table 3.1.

The first row of Table 3.1 reports, for comparison, the estimates we obtain if we do have full data that contain both regress C and X, so that we can estimate the regression model of interest directly. As expected, on average OLS recovers $\beta = 1.0$ exactly.

The second row of Table 3.1 considers the original SK procedure. It is

Table 3.1: Monte Carlo Experiment

	$\sigma^2_{u_1} = 2$				$\sigma^2_{u_1} = 4$			
	$n = 250$		$n = 1,000$		$n = 250$		$n = 1000$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
β Full	1.000	(0.111)	1.000	(0.055)	1.000	(0.111)	1.000	(0.055)
β^{SK}	0.401	(0.090)	0.400	(0.044)	0.250	(0.071)	0.250	(0.035)
β^{RSK}	1.004	(0.211)	1.001	(0.103)	1.010	(0.275)	1.002	(0.136)
β^{cRSK}	1.000	(0.193)	1.001	(0.095)	1.005	(0.256)	1.001	(0.127)

Note: 10,000 Replications

immediately apparent the regression of C imputed this way on X does not recover $\beta = 1.0$ but instead recovers βR^2 (1.0×0.4 when $\sigma^2_u = 2$ and 1.0×0.25 when $\sigma^2_u = 4$).

The third row reports results for the RSK procedure. It is obvious that it is consistent, though there is a loss of precision relative to the case of full data reported in the first row.

The final row of Table 3.1 reports the results of trying to improve the precision of $\hat{\beta}^{RSK}$ with the Inoue-Solon-type finite sample correction. The correction improves the finite sample precision (and accuracy) of $\hat{\beta}^{RSK}$.

3.4.2 Empirical Application

As a further illustration, we implemented these procedures using the PSID (from 1979 to 1992) and CE (1980 to 1992) data employed in **BPP**. Following BPP we use food at home as our proxy (Z) for total nondurable consumption (C). Our equation of interest is in the spirit of the excess sensitivity tests in Ziliak (1998).

$$\Delta \widehat{\ln c} = \beta \Delta \ln y + \epsilon$$

where $\Delta\widehat{\ln c}$ is the change in imputed log nondurable consumption and $\Delta \ln y$ is the change in log net income, and we instrument $\Delta \ln y$ with $\Delta \ln y_{-1}$. $\Delta\widehat{\ln c}$ is constructed by the **RSK** procedure. We regress $\ln C$ on Z in the CEX, use the resulting coefficients to predict $\ln C$ in the PSID, and rescale those predictions by $R^2_{C,Z}$ (from the first stage in the PSID). The results are presented in Table 3.2.

The elasticity of consumption growth with respect to income growth is about 50% larger after rescaling, though, even so, the elasticity with respect to predicted income (i.e., the IV estimates) are not significantly different from zero. The insensitivity of consumption growth to predicted income growth is predicted by the theory.

Table 3.2: Empirical Example: Log Consumption Growth on Log Income

	Skinner	Rescaled-Skinner	Rescaled-Skinner (IS correction)
OLS			
β	0.068	0.099	0.086
(SE)	(0.010)	(0.014)	(0.012)
IV			
β	0.006	0.009	0.008
(SE)	(0.039)	(0.057)	(0.050)

In the first stage regressions we control for the age of head, the age of head squared, family size, race of head, dummies for the number of children, region and year dummies. In the second stage regressions we include changes in these demographic and region variables as well as year dummies. Income is instrumented with its lag.

3.5 Conclusion

Although imputation of the dependent variable in a regression induces error '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).

The inconsistency can be overcome by rescaling by the first-stage (imputation) R^2 (the **RSK** 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.

These procedures employ two samples and we have shown how a refinement analogous to the Inoue-Solon refinement to Angrist and Krueger's Two-Sample IV procedure can be used to improve finite-sample precision. 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.

Finally we note again that the key assumption that supports the use of a variable, Z, as an imputation proxy exactly contradicts the assumption required to use Z as an instrument (and vice-versa). A variable may be a plausible instrument or a plausible proxy, or neither, but never both. If, in a given application, the assumption required to use Z as an instrument are more plausible than those required to use Z as an imputation proxy, then the Inoue-Solon refinement to Angrist and Krueger's Two-Sample IV procedure should be employed, rather than the procedures outlined in this paper.

Bibliography

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. *ECB Working Paper Series*, (1817).

Attanasio, O., Hurst, E., and Pistaferri, L. (2012). The Evolution of Income, Consumption, and Leisure Inequality in the US, 1980-2010. *National Bureau of Economic Research*, (17982).

Attanasio, O. and Pistaferri, L. (2014). Consumption Inequality over the Last Half Century: Some Evidence using the New PSID Consumption Measure. *The American Economic Review: Papers & Proceedings*, 104(5):122–126.

Banks, J., Blundell, R., and Lewbel, A. (1997). Quadratic Engel Curves and Consumer Demand. *The Review of Economics and Statistics*, 79(4):527–539.

Berkson, J. (1950). Are There Two Regressions? *Journal of the American Statistical Association*, 45(250):164–180.

Blundell, R. and Meghir, C. (1986). Selection Criteria for a Microeconometric Model of Labour Supply. *Journal of Applied Econometrics*, 1(1):55–80.

Blundell, R., Pashardes, P., and Weber, G. (1993). What do we Learn About Consumer Demand Patterns from Micro Data? *The American Economic Review*, pages 570–597.

Blundell, R., Pistaferri, L., and Preston, I. (2004). Imputing consumption in the PSID using food demand estimates from the CEX. *The Institute for Fiscal Studies, UCL (University College London), The Institute for Fiscal Studies*.

Blundell, R., Pistaferri, L., and Preston, I. (2008). Consumption Inequality and Partial Insurance. *The American Economic Review*, pages 1887–1921.

Blundell, R. W., Browning, M., and Crawford, I. A. (2003). Nonparametric Engel Curves and Revealed Preference. *Econometrica*, 71(1):205–240.

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.

Börsch-Supan, A. (2017a). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2. Release version: 6.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w2.600.

Börsch-Supan, A. (2017b). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 6.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w4.600.

Börsch-Supan, A. (2017c). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release version: 6.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w5.600.

Börsch-Supan, A. (2017d). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 6. Release version: 6.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w6.600.

Brislin, R. W., Lonner, W. J., Thorndike, R. M., et al. (1973). *Cross-cultural Research Methods*. J. Wiley New York, NY.

Browning, M., Crossley, T. F., and Weber, G. (2003). Asking Consumption Questions in General Purpose Surveys. *The Economic Journal*, 113(491):540–567.

Browning, M., Crossley, T. F., and Winter, J. (2014). The Measurement of Household Consumption Expenditures. *Annual Review Economics*, 6(1):475–501.

Bryan, M. L. and Jenkins, S. P. (2016). Multilevel Modelling of Country Effects: A Cautionary Tale. *European Sociological Review*, 32(1):3–22.

Buck, N. and McFall, S. (2011). Understanding Society: Design Overview. *Longitudinal and Life Course Studies*, 3(1):5–17.

Chamberlain, G. (1986). Asymptotic Efficiency in Semi-parametric Models with Censoring. *Journal of Econometrics*, 32(2):189–218.

Clemes, S. A., Hamilton, S. L., and Griffiths, P. L. (2011). Summer to Winter Variability in the Step Counts of Normal Weight and Overweight Adults Living in the UK. *Journal of Physhical Activity and Health*, 8(1):36–44.

Coelho, D., Veiga, H., and Veszteg, R. (2005). Comment on Parametric and Semiparametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal. UFAE and IAE working papers, Unitat de Fonaments de l'Anlisi Econmica (UAB) and Institut d'Anlisi Econmica (CSIC).

Connolly, M. (2013). Some Like It Mild and Not Too Wet: The Influence of Weather on Subjective Well-Being. *Journal of Happiness Studies*, 14(2):457–473.

Cosslett, S. R. (1983). Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model. *Econometrica*, 51(3):765–782.

Crossley, T., Levell, P., and Low, H. (2017). Consumption Spending, Housing Investments, and the Role of Leverage.

Davidson, R. and MacKinnon, J. G. (1984). Convenient Specification Tests for Logit and Probit Models. *Journal of Econometrics*, 25(514):241–262.

De Luca, G. and Celidoni, M. (2015). Item Nonresponse and Imputation Strategies in SHARE Wave 5. In Börsch-Supan, A. and Malter, F., editors, *SHARE Wave 5: Innovations and Methodology*, pages 85–100. MEA, Mannheim.

De Luca, G. et al. (2008). SNP and SML Estimation of Univariate and Bivariate Binary-choice Models. *The Stata Journal*, 8(2):190–220.

de Luca, G. and Lipps, O. (2005). Fieldwork and survey management in share. In Börsch-Supan, A. and Jürges, H., editors, *The Survey of Health, Aging, and Retirement in Europe Methodology*, pages 75–81. MEA, Mannheim.

De Luca, G. and Peracchi, F. (2012). Estimating Engel Curves under Unit and Item Nonresponse. *Journal of Applied Econometrics*, 27(7):1076–1099.

Deaton, A. and Muellbauer, J. (1980). An Almost Ideal Demand System. *The American Economic Review*, 70(3):312–326.

Engel, E. (1895). *Die Lebenskosten belgischer Arbeiter-Familien früher und jetzt*. C. Heinrich.

Engle, R. F. (1984). Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. In Griliches, Z. and Intriligator, M., editors, *Handbook of Econometrics*, volume 2, pages 775–826. Elsevier, North-Holland, Amsterdam.

Gallant, A. R. and Nychka, D. W. (1987). Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica*, 55(2):363–390.

Gonzalez-Rivera, G. and Ullah, A. (2001). Rao's Score Test with Nonparametric Density Estimators. *Journal of Statistical Planning and Inference*, 97:85–100.

Greene, W. H. (2003). *Econometric Analysis*. Pearson, seventh edition.

Gupta, A. (2018). Nonparametric Specification Testing via the Trinity of Tests. *Journal of Econometrics*, 203(1):169–185.

Hakim, C. (2000). *Research Design: Successful Designs for Social and Economic Research*. Routledge Taylor & Francis, London and New York, second edition.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal Smoothing in Single-Index Models. *The Annals of Statistics*, 21(1):157–178.

Harkness, J. (1999). In Pursuit of Quality: Issues for Cross-national Survey Research. *International Journal of Social Research Methodology*, 2(2):125–140.

Hayfield, T., Racine, J. S., et al. (2008). Nonparametric Econometrics: The *np* Package. *Journal of Statistical Software*, 27(5):1–32.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161.

Horowitz, J. L. and Härdle, W. (1994). Testing a Parametric Model Against a Semiparametric Alternative. *Econometric Theory*, 10(5):821–848.

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.

Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-index Models. *Journal of Econometrics*, 58(1-2):71–120.

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

Kaminska, O. and Lynn, P. (2017). Survey-based Cross-country Comparisons where Countries vary in Sample Design: Issues and Solutions. *Journal of Official Statistics*, 33(1):123–136.

Kimura, T., Kobayashi, H., Nakayama, E., and Kakihana, W. (2015). Seasonality in Physical Activity and Walking of Healthy Older Adults. *Journal of Physiological Anthropology*, 34(1):1–6.

Klein, R. and Spady, R. H. (1993). An Efficient Semiparametric Estimator for Binary Response Models. *Econometrica*, 61(2):387–421.

Klein, R. W. (1993). Specification Tests for binary Choice Models based on Index Quantiles. *Journal of Econometrics*, 59(3):343–375.

Klevmarken, N. A., Swensson, B., and Hesselius, P. (2005). The share sampling procedures and calibrated design weights. In Börsch-Supan, A. and Jurges, H., editors, *The Survey of Health, Aging, and Retirement in Europe - Methodology*, pages 28–69. MEA, Mannheim.

Leser, C. E. V. (1963). Forms of Engel Functions. *Econometrica: Journal of the Econometric Society*, pages 694–703.

Longhi, S. (2014). Residential Energy Use and the Relevance of Changes in Household Circumstances. Technical Report 2014-22, ISER Working Paper Series.

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.

Lynn, P. (2003). Developing Quality Standards for Cross-national Survey Research: Five Approaches. *Inernational Journal of Social Research Methodology*, 6(4):323–336.

Lynn, P. (2009). Sample design for Understanding Society. *Understanding Society Working Paper Series*, 1.

Malter, F. and Börsch-Supan, A. (2013a). *SHARE Compliance Profiles Wave 4.* MEA, Max Planck Institute for Social Law and Social Policy., Munich.

Malter, F. and Börsch-Supan, A. (2013b). *SHARE Compliance Profiles Wave 5.* MEA, Max Planck Institute for Social Law and Social Policy., Munich.

Manski, C. F. (1975). Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics*, 3:205–228.

Martins, M. F. O. (2001). Parametric and Semiparametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal. *Journal of Applied Econometrics*, 16(1):23–39.

McCormack, G. R., Friedenreich, C., Shiell, A., Giles-Corti, B., and Doyle-Baker, P. K. (2010). Sex-and Age-specific Seasonal Variations in Physical Activity among Adults. *Journal of Epidemiology and Community Health*, 64(11):1010–1016.

Schroder, M. (2011). *Retrospective Data Collection in the Survey of Health, Ageing and Retirement in Europe. SHARELIFE Methodology.* MEA, Mannheim.

Shea, J. (1997). Instrument Relevance in Multivariate Linear Models: A Simple Measure. *The Review of Economics and Statistics*, 79(2):348–352.

Skinner, C. and Mason, B. (2012). Weighting in the Regression Analysis of Survey Data with a Cross-national Application. *Canadian Journal of Statistics*, 40(4):697–711.

Skinner, J. (1987). A Superior Measure of Consumption from the Panel Study of Income Dynamics. *Economics Letters*, 23(2):213–216.

Subar, A. F., Frey, C. M., Harlan, L. C., and Kahle, L. (1994). Differences in Reported Food Frequency by Season of Questionnaire Administration: the 1987 National Health Interview Survey. *Epidemiology*, pages 226–233.

Teune, H. (1977). Analysis and interpretation in cross-national survey research. In A. Szalai, R. Petrella, S. R. and Scheuch, E., editors, *Cross-National Comparative Survey Research: Theory and Practice*, pages 49–93. Pergamon, Oxford.

University of Essex (2016). Institute for Social and Economic Research, NatCen Social Research and Kantar Public, [producers]: Understanding Society: Waves 1-6, 2009-2015 [computer file]. 8th Edition. Colchester, Essex: UK Data Service [distributor], November 2016. SN: 6614.

Visscher, T. and Seidell, J. (2004). Time Trends (1993–1997) and Seasonal Variation in Body Mass Index and Waist Circumference in the Netherlands. *International Journal of Obesity*, 28(10):1309–1316.

Wansbeek, T. and Meijer, E. (2000). Measurement Error and Latent Variables in Econometrics.

Working, H. (1943). Statistical Laws of Family Expenditure. *Journal of the American Statistical Association*, 38(221):43–56.

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.