Discussion Paper No. 16-092

IT Outsourcing and Firm Productivity: Eliminating Bias from Selective Missingness in the Dependent Variable

Christoph Breunig, Michael Kummer,
Jorg Ohnemus, and Steffen Viete

ZEW

Zentrum für Europäische
Wirtschaftsforschung GmbH

Centre for European
Economic Research

Discussion Paper No. 16-092

IT Outsourcing and Firm Productivity: Eliminating Bias from Selective Missingness in the Dependent Variable

Christoph Breunig, Michael Kummer,
Jorg Ohnemus, and Steffen Viete

Download this ZEW Discussion Paper from our ftp server:

http://ftp.zew.de/pub/zew-docs/dp/dp16092.pdf

IT Outsourcing and Firm Productivity: Eliminating Bias from Selective Missingness in the Dependent Variable[*]

Christoph Breunig [⋆]
Humboldt-Universität zu Berlin

Michael Kummer [1]
Georgia Institute of Technology
Centre for European Economic Research

Jorg Ohnemus [2]
Centre for European Economic Research

Steffen Viete [3]
Centre for European Economic Research

December 2016

Missing values are a major problem in all econometric applications based on survey data. A standard approach assumes data are *missing-at-random* and uses imputation methods, or even listwise deletion. This approach is justified if item non-response does not depend on the potentially missing variables' realization. However, assuming *missing-at-random* may introduce bias if non-response is, in fact, selective. Relevant applications range from financial or strategic firm-level data to individual-level data on income or privacy-sensitive behaviors.

In this paper, we propose a novel approach to deal with selective item non-response in the model's dependent variable. Our approach is based on instrumental variables that affect selection only through potential outcomes. In addition, we allow for endogenous regressors. We establish identification of the structural parameter and propose a simple two-step estimation procedure for it. Our estimator is consistent and robust against biases that would prevail when assuming missingness at random. We implement the estimation procedure using firm-level survey data and a binary instrumental variable to estimate the effect of outsourcing on productivity.

Keywords: endogenous selection, IV-estimation, inverse probability weighting, missing data, productivity, outsourcing, semiparametric estimation.

JEL-Codes: C14, C36, D24, L24

[*] We thank Irene Bertschek, Maximilian Kasy, Francois Laisney, Stephan Martin, Michael Ward for valuable suggestions. We also thank seminar participants at ZEW and Mines ParisTech for their questions and input.

[⋆] Spandauer Straße 1, 10178 Berlin, Germany, e-mail: christoph.breunig@hu-berlin.de

[1] 221, Bobby Dodd Way, # 222 Atlanta, GA, 30308, U.S.A., e-mail: michael.kummer@econ.gatech.edu

[2] P.O. Box 103443, D68034 Mannheim, Germany, e-mail: ohnemus@zew.de

[3] P.O. Box 103443, D68034 Mannheim, Germany, e-mail: viete@zew.de

1. Introduction

Missingness is a major problem in databases and survey-based data. While one well-known problem is recruiting a representative samples (unit non-response), a second major problem is incomplete answers (item non-response). This study focuses on item non-response, which arises when respondents to surveys prefer not to answer specific items or do not know the answer. Specifically, we focus on item non-response in the dependent variable. This problem particularly affects sensitive or specific information, which often are the outcome variables of interest at the heart of many economic studies. Examples are profits, turnovers, income, tax fraud, or the consumption of medications. A standard approach is assuming that data are *missing at random* (MAR) and then relying on listwise deletion of observations or on using imputation methods.[1] However, these practices may introduce a bias if the missingness is in fact selective, that is, when certain groups of observations are less likely to be reported/observed than others.

In this paper we propose a novel approach to correct for potential biases arising from missing data. Specifically, we study the estimation of averages of a selectively observed outcome in a cross-section context, when considering that certain characteristics are associated with higher frequencies of missing observations. The correction of the bias is based on an instrument that affects selection only through potential outcomes. In addition, we allow for endogeneity of regressors. We propose a simple two-step estimation procedure, and show that it is consistent and asymptotically normal.

We apply our estimator within a common class of econometric models, that is, production function estimation, and we study the effect of information technology (IT) outsourcing on productivity using a survey-based sample of German firms. Specifically, we aim to estimate the effect of IT outsourcing X on firm productivity Y^*, which is only partially observed. Commonly, in empirical studies the firms' sourcing decision is considered endogenous to the production process (e.g., Amiti and Konings [2007], Görg et al. [2008], Halpern et al. [2015]). However, focusing on endogeneity of the firms' sourcing activity is not sufficient in our application, since, in addition, the outcome Y^* is subject to selective missingness, as we illustrate in this paper. Indeed, firm productivity might directly influence the response behavior, for example, firms are less willing to report data after weak performance during the fiscal year. Consequently, we additionally have to correct for this selection error. To do so, we introduce an additional exclusion restriction on a control variable to account for selective item non-response. Specifically, we assume that this control variable does not contain any additional information on the missingness mechanism that is not already contained in the potential outcome Y^* and other controls. This exclusion restriction was recently considered by Ramalho and Smith [2013] and D'Haultfoeuille [2010]. This assumption is suitable in situations in which selection is driven by the outcome Y^* itself. We argue that this is likely the case in many applications that rely on firm-level survey data.

Probably the most common approach to deal with missing observations is to assume *missing at random* (MAR) (in the sense of Rubin [1976]), namely, that response depends only on observed covariates but not on potential outcomes. Unfortunately, the plausibility of this assumption may be questioned in many economic examples in which missing observations arise because of self-selection, or nonresponse, or because counterfactual variables are unobservable (for an analysis of sensitivity of MAR, see also Kline and Santos [2013]). In particular, when selection is driven by the underlying potential outcome itself, as

[1] e.g. Heckman [1974], Rubin [1976, 1987].

we argue is likely with applications such as ours, existing empirical strategies that assume MAR are infeasible. When response is driven directly by the outcome, it might be also difficult to find instruments that determine selection but not the outcome (see Heckman [1974]).

Contribution Using a similar exclusion restriction, Tang et al. [2003] and Zhao and Shao [2015] propose a Pseudo-Maximum Likelihood method to estimate a parametric conditional distribution of Y^*. In contrast to our paper, these papers do not address the issue of endogeneity of covariates. However, endogeneity of both selection and covariates was studied by Breunig et al. [2015]; in addition, they leave the functional form of the distribution of Y^* given covariates unrestricted. However, their approach relies on continuity of the instrumental variables that deal with selective non-response in the dependent variable. This is not the case in our and other applications. In Section 2 of this paper we treat selective non-response of outcome and endogeneity of covariates in a partial linear model and establish identification given discrete instruments. Being able to use dummy variables as instruments for selective non-response in an IV estimation adds the last missing piece to render estimators that correct for selective non-response in the dependent variable fully functional. We propose a simple two-step estimation procedure: First, we propose a constrained nonparametric least squares estimator for the conditional selection probability of observing Y^*. Second, we enter this estimator in a generalized method of moments (GMM) estimator to arrive at the structural parameter. We implement the estimation procedure in Section 3 and estimate the effect of IT outsourcing on productivity. In this application, the instrumental variable is binary. We find that our estimation procedure performs well and effectively corrects for biases that would prevail when MAR is assumed. The method can be easily adopted to most applications using survey data.

Non random missingness is an important problem in the estimation of production functions. However, production function estimation has thus far focused on bias due to endogenous input choice and on endogeneity through panel attrition and unit non-response (firm exit) (cf. Olley and Pakes [1996], Levinsohn and Petrin [2003]), Melitz and Polanec [2015]). We highlight that imposing MAR on missing values in the dependent variable is an additional source of biased estimates and propose a correction that is compatible with IV estimation.

2. Identification and Estimation of Structural Parameters

In this section, we provide assumptions under which the selection probability function $\mathbb{P}(\Delta = 1|Y^*)$ and the conditional mean $\mathbf{E}[Y^*|X = \cdot]$ are identified. We further motivate our estimation procedure. For the sake of simplicity we first consider the situation in which the parametric part of our model consists of only a scalar endogenous regressor. Thereafter we discuss the situation in which the parametric part coincides with a vector.

2.1. Model

Our aim is to identify the causal impact of a binary, potentially endogenous variable X on a selectively observed outcome Y^*. We consider a partially linear model

$$Y^* = X\beta_0 + m(W_1) + U \tag{2.1}$$

for some unknown structural scalar parameter β_0 and unknown nonparametric function m. A realization of (Δ, X, W) with $W = (W_1', W_2)'$ is observed for each individual in the random sample. However, Y^* may suffer from selective non-response: a realization of the dependent variable Y^* is observed when $\Delta = 1$ and missing when $\Delta = 0$. We write $Y = \Delta Y^*$. Additionally, we model the exogenous covariates as W_1 and to deal with potential endogeneity in the explanatory variable of interest we allow for an instrument W_2 such that $\mathbf{E}[U|W] = 0$. Here, the instrument W_2 is binary.

Example 2.1 (IT Outsourcing and Productivity). In our application, we use an augmented production function model to estimate the effect of IT outsourcing on productivity. In its stylized version, we consider the following model (we abstract from additional dummy variables and other controls):

$$ln(Prod_i^*) = ITout_i\,\beta_0 + m\Big(\ln(K_i), \ln(L_i)\Big) + u_i.$$

In this empirical model $Prod_i^*$ denotes average labor productivity, which is only partially observed. We measure labor productivity by value added ($sales - costs\ of\ intermediates$) over labor. For the purpose of this application, capital K_i and labor L_i are considered as exogenous control variables.[2] The parameter of interest is β_0, the coefficient that measures the effect of IT outsourcing $ITout_i$. In this model, $ITout_i$ may be endogenous, and we use a standard instrumental variable strategy, based on the excluded instrument W_2 to account for this. In our application, W_2 is a variable measuring whether a firm sought '$Y2K$ consulting' to avoid the 'millennium bug.'[3] The novelty of this paper is that we can model productivity $Prod_i^*$ to be plagued with selective non-response, possibly because respondents avoid disclosing especially high (or low) sales. This is modeled by the response indicator Δ_i, which may depend on potential productivity itself. For instance, a company is more likely to report data if its sales (and hence measured productivity) are high than if sales are low. Our strategy allows that the firm's response could be a function of potential sales. We show below that β_0 cannot be estimated consistently without accounting for the selectivity in the non-response for $Prod_i^*$.

2.2. Identification

In what follows, we show which general assumptions allow to identify β_0. Conditioning model (2.1) on the exogenous covariates W_1 yields

$$\mathbf{E}(Y^*|W_1) = \mathbf{E}(X|W_1)\beta_0 + m(W_1). \tag{2.2}$$

Multiplying equations (2.1) and (2.2) by the binary instrument W_2 and taking expectations leads to

$$\begin{aligned} \mathbf{E}(Y^*W_2) &= \mathbf{E}(XW_2)\,\beta_0 + \mathbf{E}[m(W_1)W_2], \\ \mathbf{E}[\mathbf{E}(Y^*|W_1)W_2] &= \mathbf{E}[\mathbf{E}(X|W_1)W_2]\,\beta_0 + \mathbf{E}[m(W_1)W_2]. \end{aligned}$$

[2] Note that these variables are in fact strategic, and should be instrumented, e.g. by using investment Olley and Pakes [1996] or intermediate inputs [Levinsohn and Petrin, 2003, Petrin and Levinsohn, 2012, Wooldridge, 2009]. Yet, the purpose of our application is to implement our estimator that corrects for selective non-response in the dependent variable and its focus is the effect of IT outsourcing on productivity. Hence, while we deem it necessary to control for labour and capital, we avoid the additional expositional complexity that would arise from instrumenting for these variables as well.

[3] This bug threatened the IT systems of firms on January 1 2000, if they had relied on software that allocated only two 'year digits' when storing a date. See section 3 for more detail on this excluded instrument.

Now taking the difference of both equation yields

$$\mathbf{E}[(Y^* - \mathbf{E}(Y^*|W_1))W_2] = \mathbf{E}[(X - \mathbf{E}(X|W_1))W_2]\beta_0. \tag{2.3}$$

The parameter β_0 is not identified if $\mathbf{E}(XW_2) = \mathbf{E}[\mathbf{E}(X|W_1)W_2]$. Identification of β_0 thus requires the instrument W_2 to contain information about X which is not captured by the exogenous covariates W_1. The next assumption formalizes this restriction.

Assumption 1. *It holds* $\mathbf{E}[(X - \mathbf{E}(X|W_1))W_2] \neq 0$.

Under Assumption 1 we can write the structural β_0 as

$$\beta_0 = \frac{\mathbf{E}[(Y^* - \mathbf{E}(Y^*|W_1))W_2]}{\mathbf{E}[(X - \mathbf{E}(X|W_1))W_2]}.$$

In the following, we provide sufficient conditions to ensure identification of $\mathbf{E}[(Y^* - \mathbf{E}(Y^*|W_1))W_2]$.

Assumption 2 (Exclusion Restriction on Selection). *It holds* $\Delta \perp\!\!\!\perp X \,|\, (Y^*, W)$.

Assumption 2 requires that the covariate X has no direct effect on the response given potential outcome Y^* and W. As such, the covariate X serves as an instrumental variable for the selective response to Y^*. This assumption is well suited for our application but might also need to be modified to be appropriate for other particular applications. In fact, we may also assume that a subvector of W is independent of the response given the other potentially observed information. For instance, the instrument W_2 to account for endogeneity of X might also be used to account selective non-response of Y^* via the exclusion restriction $\Delta \perp\!\!\!\perp W_2 \,|\, (Y^*, X, W_1)$.

Example 2.2 (IT outsourcing and Productivity (cont'd)). In our application Assumption 1, is satisfied if $\mathbf{E}[(ITout - \mathbf{E}(ITout|Controls))Y2K] \neq 0$. This means that the instrument (which accounts for the endogeneity of IT outsourcing) contains information on IT outsourcing that is not captured by the other control variables. More importantly, Assumption 2 requires that $\Delta \perp\!\!\!\perp ITout \,|\, (Prod^*, ITout, L, K, Y2K)$, that is, IT outsourcing $ITout_i$ contains no information on the response Δ_i that is not already contained in the potential productivity of firms, $Prod_i^*$, and observed control variables such as labor or capital. While Assumption 1 is easily verified in practice, we emphasize that the exclusion restriction for selective non-response, imposed in Assumption 2, is also testable as shown by D'Haultfoeuille [2010].

Example 2.3 (Relation to Triangular Model). Assumption 2 can be justified in a triangular model as follows. Consider an equivalent formulation of model (2.2) as

$$Y^* = \mathbf{E}(X|W_1)\beta_0 + m(W_1) + \varepsilon, \tag{2.4}$$

where $\varepsilon = Y^* - \mathbf{E}(Y^*|W_1)$. In this case, the exclusion restriction on selection, i.e., $\Delta \perp\!\!\!\perp X \,|\, (Y^*, W)$, is satisfied when additionally

$$\begin{aligned} &\Delta = \phi(Y^*, W, \eta), \\ &\eta \perp\!\!\!\perp (X, \varepsilon)|W. \end{aligned}$$

Similarly, different types of exclusion restrictions can be justified.[4]

[4] For instance, the exclusion restriction $\Delta \perp\!\!\!\perp W_2 \,|\, (Y^*, X, W_1)$, mentioned above, is satisfied in model (2.4)

Let us denote $V^* \equiv (Y^*, X, W_1')'$, $V \equiv (Y, X, W_1')'$, and $Z = (X, W')'$. Assumption 2 implies $\mathbb{P}(\Delta = 1|V^*, Z) = \mathbb{P}(\Delta = 1|V^*)$ and hence, by the law of iterated expectations, we obtain the following conditional mean restriction:

$$\mathbf{E}\left[\frac{\Delta}{\mathbb{P}(\Delta = 1|V^*)}\Big|Z\right] = 1. \tag{2.5}$$

The following assumption enables us to identify the conditional probability $\mathbb{P}(\Delta = 1|V^*)$ via the previous moment restriction. Here, we denote $d_x = \dim(X)$ and $d_{w_1} = \dim(W_1)$.

Assumption 3. *(i) For all v in the support of V^*, $\mathbb{P}(\Delta = 1|V^* = v) = G(v'\vartheta)$ for some known strictly increasing function $G : \mathbb{R} \to (0,1)$ and some parameter $\vartheta \in \mathbb{R}^{d_x + d_{w_1}+1}$. (ii) It holds rank$(\mathbf{E}(\Delta Z V' g(V'\vartheta))) = d_x + d_{w_1} + 1$ where $g(\cdot) = G'(\cdot)/G^2(\cdot)$ is the derivative of $-1/G(\cdot)$.*

Assumption 3 (*i*) restricts the conditional probability of observing Y^* to be known up to a finite dimensional parameter. In particular, we model the selection probability in a single index framework. Typical examples are probit or logit models. Assumption 3 (*i*) also requires that the conditional probability of observing Y^* given (Y^*, X) is strictly positive. In particular, Assumption 3 can rule out certain types of selection, such as deterministic truncation models. Further, Assumption 3 (*ii*) ensures identification of the selection probability through equation (2.5) (see Theorem 2.1 of D'Haultfoeuille [2010]).

Example 2.4 (IT outsourcing and Productivity (cont'd)). In our application, $\mathbb{P}(\Delta = 1|V^*)$ denotes the probability that a company reports sales or costs of intermediates given potential productivity $Prod^*$ and other controls (that is $\mathbb{P}(\Delta = 1|V^*) = \mathbb{P}(\Delta = 1|Prod^*, ITout, L, K, Y2K)$). As we show below, identification of the function $v \mapsto \mathbb{P}(\Delta = 1|V^* = v)$ is key to identifying the structural parameter through inverse probability weighting but is also of interest on its own, because it illustrates whether the MAR assumption is violated (see also Breunig [2015] for a formal test of it). In our application, we see that the conditional probability depends on the potential productivity realizations in a nonlinear fashion (see chapter 3). We also show that reporting $Prod^*$ does depend on $Prod^*$ itself even if other important control variables are included, that is, the assumption of MAR does not hold true.

With Assumption 3, we can conclude the identification argument. Applying Assumption 2 together with the law of total expectation yields

$$\begin{aligned}\mathbf{E}(Y^*|X, W) &= \mathbf{E}\left[\mathbf{E}\left(\frac{\Delta Y^*}{G(\vartheta_0' V^*)}\Big|V^*\right)\Big|X, W\right] \\ &= \mathbf{E}\left[\frac{Y}{G(V'\vartheta_0)}\Big|X, W\right],\end{aligned}$$

where we hold that $G(\vartheta_0' V) = G(\vartheta_0' V^*)$ whenever Δ is different from zero. In particular, by conditioning both sides of the previous conditional mean equation by W we obtain

$$\mathbf{E}(Y^*|W) = \mathbf{E}\left[\frac{Y}{G(V'\vartheta_0)}\Big|W\right]. \tag{2.6}$$

when additionally

$$\begin{aligned}&\Delta = \phi(Y^*, X, W_1, \eta)\\ &\eta \perp\!\!\!\perp (W_2, \varepsilon)|(X, W_1).\end{aligned}$$

This shows that after the parameter ϑ_0 is identified through the instrumental variable restriction (2.5), we can identify the conditional mean $\mathbf{E}(Y^*|W)$ through inverse probability weighting. Now since $\mathbf{E}[X - \mathbf{E}(X|W_1)W_2] \neq 0$, due to Assumption 1 we obtain identification of the parameter β_0 as summarized in the next result, which is our main identification result.

THEOREM 2.1. *Let Assumptions 1–3 be satisfied. Then, the structural parameter in model* (2.1) *is identified through*

$$\beta_0 = \frac{\mathbf{E}\Big[\Big(Y/G(V'\vartheta_0) - \mathbf{E}(Y/G(V'\vartheta_0)|W_1)\Big)W_2\Big]}{\mathbf{E}\Big[(X - \mathbf{E}(X|W_1))W_2\Big]}.$$

This concludes the identification argument, and we now move on to derive an appropriate estimator for our setting.

2.3. A closed form Estimator of the structural Parameter

Our estimator of the structural parameter β_0 is based on the previous constructive identification results. We estimate the nuisance parameter ϑ and the nonparametric functions $g(w_1, \vartheta) = \mathbf{E}(Y/G(V'\vartheta)|W_1 = w_1)$ and $h(w_1) = \mathbf{E}(X|W_1 = w_1)$ in a first step. We replace ϑ_0 by a generalized method of moments (GMM) estimator $\widehat{\vartheta}_n$ based on an empirical analog of the conditional moment equation (2.5).

We propose the following new estimator of the structural parameter β_0 given by

$$\widehat{\beta}_n = \frac{\sum_{i=1}^n W_{2i}\Big(Y_i/G(V_i'\widehat{\vartheta}_n) - \widehat{g}_n(W_{1i}, \widehat{\vartheta}_n)\Big)}{\sum_{i=1}^n W_{2i}\Big(X_i - \widehat{h}_n(W_{1i})\Big)}.$$

where we replaced the nonparametric functions g and h by the series least squares estimators, as follows.

Let $L \geqslant 1$ denote the number of basis functions used to approximate these functions, where L increases with sample size n. We then introduce a vector of basis functions denoted by $p^L(w) = (p_1(w), \dots, p_L(w))'$. Further, the matrix of basis vectors evaluated at the sample points of W_1 is denoted by $\mathbf{W}_n \equiv (p^L(W_{11}), \dots, p^L(W_{1n}))'$. We follow Breunig et al. [2015] and consider the following series least squares estimator with inverse probability weighting

$$\widehat{g}_n(w, \widehat{\vartheta}_n) \equiv p^L(w)'(\mathbf{W}_n'\mathbf{W}_n)^{-1}\sum_{i=1}^n \frac{Y_i}{G(V_i'\widehat{\vartheta}_n)}p^L(W_{1i}).$$

Moreover, we replace h by the series least squares estimator (see e.g., Newey [1997])

$$\widehat{h}_n(w) \equiv p^L(w)'(\mathbf{W}_n'\mathbf{W}_n)^{-1}\sum_{i=1}^n X_i p^L(W_{1i}).$$

The next result establishes the asymptotic distribution of the estimator $\widehat{\beta}_n$. We show its consistency with the identified structural parameter and asymptotic normality. In applications, such asymptotic distribution results can be useful in constructing approximate confidence intervals. The next theorem also uses Assumption 5, which gathers the technical conditions to ensure the asymptotic distribution result and is discussed in the Appendix, where also the proof is provided.

Theorem 2.2. *Let Assumptions 1–5 be satisfied. Then we have*

$$\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

where σ^2 denotes the variance of the random variable

$$W_2\Big(\frac{Y}{G(V'\vartheta_0)} - g(W_1)\Big) - \Big(\frac{\Delta}{G(V'\vartheta_0)} - 1\Big)Z'\,A(A'A)^{-1}\,\mathbf{E}\Big[(W_2 - \mathbf{E}[W_2|W_1])YV\frac{G_\vartheta(V'\vartheta_0)}{G^2(V'\vartheta_0)}\Big]$$

where $A = \mathbf{E}\Big[ZV'G_\vartheta(V'\vartheta_0)/G^2(V'\vartheta_0)\Big]$.

The asymptotic result of Theorem 2.2 remains valid if the variance σ^2 is replaced by its empirical analog. This can be also used to construct pointwise confidence intervals for $\widehat{\beta}_n$. In our application, however, we rely on resampling methods.

2.4. Extension: Multivariate Control Variables

In the following extension, we lay out how our identification and estimation strategy carries over to models with multivariate endogenous regressors.[5] Again let (Δ, Y^*, X', W') be a jointly distributed random vector in which (Y^*, X, W) is a random vector that takes values in $\mathbb{R}^{1+d_x+d_w}$, and Δ is a random variable that takes values in $\{0, 1\}$. So, in contrast to the previous case, X is not scalar but a random vector and may also include exogenous covariates. As above, a realization of (Δ, X, W) is observed for each individual in the random sample while a realization of the dependent variable Y^* is observed when $\Delta = 1$ and missing when $\Delta = 0$ (again we let $Y = \Delta Y^*$). We consider a partially linear model

$$Y^* = X'\boldsymbol{\beta}_0 + m(W_1) + U \tag{2.7}$$

for some unknown parameter vector $\boldsymbol{\beta}_0$ and unknown nonparametric function m. In addition, to account for endogeneity of X, we assume that a multivariate instrument W_2 is available such that $\mathbf{E}[U|W] = 0$ where $W = (W_1', W_2')'$. As in the derivation of (2.3), conditioning model (2.7) on the exogenous covariates W_1 and/or the instruments W_2 yields

$$\mathbf{E}[(Y^* - \mathbf{E}(Y^*|W_1))W_2] = \mathbf{E}[W_2(X - \mathbf{E}(X|W_1))']\,\boldsymbol{\beta}_0 \tag{2.8}$$

The parameter vector $\boldsymbol{\beta}_0$ is not identified if $\mathbf{E}(X|W_2) = \mathbf{E}(\mathbf{E}(X|W_1)|W_2)$. Intuitively, this means that W_1 has no additional information for explaining variation in X that is already available from W_2. Put differently, identification of β_0 requires the instrument W_2 to contain additional information in explaining variations of X that is not included in the exogenous covariates W_1. The next assumption formalizes this restriction.

Assumption 4. *The matrix $\mathbf{E}[(X - \mathbf{E}(X|W_1))W_2']\,\mathbf{E}[W_2(X - \mathbf{E}(X|W_1))']$ is invertible.*

Assumption 4 can be easily verified in practice by calculating the minimal eigenvalue of the empirical counterpart of $\mathbf{E}[X - \mathbf{E}(X|W_1)|W_2]\,\mathbf{E}[(X - \mathbf{E}(X|W_1))W_2]'$ and checking its distance to zero. In the following, let us introduce the vector valued function $\mathbf{h}(w_1) = \mathbf{E}(X|W_1 = w_1)$. Assumption 4 ensures that the $\boldsymbol{\beta}_0$ is identified through equation (2.3) (given that the left-hand side is identified), and we can write

$$\boldsymbol{\beta}_0 = \Big(\mathbf{E}[(X - \mathbf{h}(W_1))W_2']\,\mathbf{E}[W_2(X - \mathbf{h}(W_1))']\Big)^{-1}\,\mathbf{E}[(X - \mathbf{h}(W_1))W_2']\,\mathbf{E}[(Y^* - \mathbf{E}(Y^*|W_1))W_2]$$

[5] Note that this extension is beyond the scope of our specific application.

In the following, we provide sufficient conditions to ensure identification of $\mathbf{E}[(Y^* - \mathbf{E}(Y^*|W_1))W_2]$. Since, by Assumption 3, it holds

$$\mathbf{E}(Y^*|W) = \mathbf{E}\left[\frac{Y}{G(V'\vartheta_0)}\middle|W\right] \tag{2.9}$$

where the right-hand side is identified, we obtain the following identification result of the multivariate structural parameter $\boldsymbol{\beta}_0$.

Proposition 2.3. *Let Assumptions 1–3 hold true. Then, in model (2.7), the parameter* $\boldsymbol{\beta}_0$ *is identified through*

$$\boldsymbol{\beta}_0 = \Big(\mathbf{E}[(X - \mathbf{h}(W_1))W_2'] \, \mathbf{E}[W_2(X - \mathbf{h}(W_1))']\Big)^{-1} \\ \times \mathbf{E}[(X - \mathbf{h}(W_1))W_2'] \, \mathbf{E}\Big[\Big(Y/G(V'\vartheta_0) - \mathbf{E}(Y/G(V'\vartheta_0)|W_1)\Big)W_2\Big].$$

In order to estimate the parameter vector of interest $\boldsymbol{\beta}$, we need to estimate the nuisance parameter ϑ and the functions $g(w_1, \vartheta) = \mathbf{E}(Y/G(V'\vartheta)|W_1 = w_1)$ as above and the vector valued function $\mathbf{h}(w_1) = \mathbf{E}(X|W_1 = w_1)$ in a first step. We replace ϑ_0 with a GMM estimator $\widehat{\vartheta}_n$ based on an empirical analog of the conditional moment equation (2.5). Further, we replace the function $\mathbf{h}$ with the series least squares estimator

$$\widehat{\mathbf{h}}_n(w) \equiv p^L(w)'(\mathbf{W}_n'\mathbf{W}_n)^{-1}\sum_{i=1}^n X_i p^L(W_{1i})$$

In a second step, we propose the estimator of $\boldsymbol{\beta}$ given by

$$\widehat{\boldsymbol{\beta}}_n = \left(\sum_{i=1}^n W_{2i}\left(X_i - \widehat{\mathbf{h}}_n(W_{1i})\right)\sum_{i=1}^n W_{2i}\left(X_i - \widehat{\mathbf{h}}_n(W_{1i})\right)\right)^{-1} \\ \times \sum_{i=1}^n W_{2i}\left(X_i - \widehat{\mathbf{h}}_n(W_{1i})\right)\sum_{i=1}^n W_{2i}\left(Y_i/G(V_i'\widehat{\vartheta}_n) - \widehat{g}_n(W_{1i})\right).$$

This concludes our extension for multivariate control variables, and we now turn to applying our estimator to a production function setting using survey data.

3. Application: The Impact of IT Outsourcing on Firm Success

3.1. Setting and Motivating Question

We now apply our estimation procedure developed in section 2 to study the effects of IT outsourcing on firm performance using firm-level micro data. We follow the empirical literature on services outsourcing (see below) and study IT outsourcing using an extended production function framework.

Selective item non-response in firm-level data: In settings like ours, high item non-response rates in particular variables complicate identification of the model parameters. Items that are plagued by a considerable share of non-response are typically monetary values, such as sales, or costs of intermediate inputs, which are required to construct key variables of interest. This problem has been documented for many business surveys that are fundamental to economic research. An important example is the US Census Bureau's Census of Manufacturers (CM), which is the main data source for much of the research on US plant-level productivity.[6] White et al. [2012] document shares of imputed values in 2007 for the items' total value of shipments, cost of electricity, and cost of material inputs of 27%, 37%, and 42%, respectively. The situation is similar for the establishment panel of the Institute for Employment Research of the Federal Labor Service in Germany (IAB), which is a cornerstone database for firm-level research in Germany. Some of the highest rates of non-response in the 2007 wave of the survey arise for key variables such as payroll (14.4%), intermediate inputs' share of revenue (17.4%), and last year's annual revenue (18.6%) [Drechsler, 2010]. These high rates of item non-response in key variables highlight the scope of the problems that item non-response might cause if it depended on the undisclosed variable's value.

In addition to firm-level survey data, the problem of missing values in items referring to monetary values is also well documented for individual and household surveys.[7] Particulary in the context of the firm, non-response might result from the lack of the right information for the individual respondent being surveyed. In addition, non-response, specifically for monetary values, is frequently related to the perceived sensitivity of the information (Drechsler [2010], Kennickell [1998]). In dealing with item non-response, applied empirical research based on firm-level survey data commonly rests on assuming MAR and pursues listwise deletion or is based on imputed data.

We argue that assuming MAR likely results in biased estimates in applications such as ours for two reasons. The probability of response in business surveys can often be related to factors such as unit size or industry. Small firms may not keep track of requested items, because of lower reporting obligations (see, e.g., Thompson and Washington [2013]). What is more, apart from the relevance of possibly unobserved firm characteristics, we stress that, in many cases, item non-response is likely to be heavily driven by the underlying values themselves. For example, at firms that underwent negative shocks and generated low sales over the fiscal year, the respondents might wish to keep poor performance confidential. Consequently, they might be less likely to disclose this information. In such cases, the MAR assumption will be violated, and commonly used strategies in applied empirical research (listwise deletion and imputation) will yield biased estimates.

Related Literature: The theoretical literature on outsourcing dates back to the seminal work by Coase [1937] and his theory of the firm. Traditionally, this literature focuses on transaction costs and incomplete contracts (Grossman and Hart [1986], Williamson [1979, 1981, 1989]) to explain vertical integration. More recent literature focuses on the rise in services outsourcing in response to a rapid expansion of the business services sector and trade (see, e.g., Abraham and Taylor [1996], Feenstra [1998], Grossman and Helpman [2005]). While theory motivates international outsourcing primarily by differentials in factor prices, it explains domestic services outsourcing by scale economies of specialized input providers. Outsourcing might also help to even out the workload of the workforce when demand is volatile (Abraham and Taylor [1996]).

[6] For instance, Black and Lynch [2001], Foster et al. [2008], Olley and Pakes [1996].

[7] See, e.g., Frick and Grabka [2010] for its discussion of non-response issues in earnings and wealth variables in the German Socio-Economic Panel (GSOEP), the British Household Panel Survey (BHPS), and the Household, Income and Labor Dynamics in Australia (HILDA) survey. Kennickell [1998] examines the issue in the context of the Survey of Consumer Finances (SCF).

IT outsourcing has been a key dimension of business services outsourcing, at least since Eastman Kodak handed its entire data and microcomputer operations to an IBM-led consortium (Loh and Venkatraman [1992]). This is not surprising, given the great importance of information technology for productivity, which has been widely documented both for the wider economy [Brynjolfsson and Hitt, 2003], and particularly for information intensive sectors such as the health sector [Lee et al., 2013].[8] The importance of IT outsourcing is reflected in its steady growth over the past few decades (Han et al. [2011], ZEW [2010]). Outsourcing IT services can be an attractive way to leverage cost advantages, and it can also facilitate the restructuring of production such that the remaining workers become more efficient (Amiti and Wei [2009]). Moreover, drawing on more specialized providers can increase the quality of IT services and thus improve input quality (Lacity et al. [2009]).

Against this background, we investigate whether IT outsourcing increases labor productivity at German manufacturing and services firms. This question has important policy implications for both investment in more powerful IT infrastructure and labor policy.

3.2. Empirical strategy

In order to investigate the effect of IT outsourcing on firm-level average labor productivity, we estimate a production function augmented by firms' IT outsourcing activities. In particular, we model labor productivity at firm i, $Prod_i^**$, as a function of capital, K_i, and labor inputs, L_i. IT outsourcing, $ITout_i$, is a binary variable indicating whether the firm subcontracted IT services and enters our production function as a shift parameter (alongside other controls W_1):

$$ln(Prod_i^*) = m(ln(K_i), ln(L_i)) + \beta_0 ITout_i + W_{1,i}'\beta_1 + u_i. \qquad (3.1)$$

In line with Equation (2.1), we allow our production function to be flexible with respect to capital and labor inputs. Assuming $m(.)$ to be linear in K and L gives the empirical production function based on an augmented Cobb-Douglas production function as a special case.[9] We estimate both the partially linear and the linear Cobb-Douglas model (the latter using two-stage least squares). In all estimations we allow for endogeneity of IT outsourcing.

Estimation of production functions such as equation (3.1) is often complicated by considerable item non-response in the output measure, which is typically constructed from data on firms' financial performance. As we expect item non-response in measures for labor productivity to be driven by the underlying value, we expect that MAR is commonly violated in comparable empirical applications using survey data. We therefore resort to our estimation strategy developed in Section 2. To do so, we impose an exclusion restriction that relies on independence between firms' outsourcing status ($ITout_i$) and Δ_i conditional on ($Prod_i^*, ITout_i, K_i, L_i, Y2K_i$). As it is unlikely that the firms' outsourcing status carries additional information about the interview partner's response behavior beyond

[8] Earlier studies have found similarly important productivity contributions in the health sector [Menon et al., 2000], or in retailing [Reardon et al., 1997, Schreyer and Pilat, 2001].

[9] We start with a standard Cobb-Douglas production function $Y_i^* = A_i K_i^{\alpha_K} L_i^{\alpha_L} ITout_i^{\beta_0}$ with output Y_i^* being a function of capital K_i, labor inputs L_i, and a Hicks-neutral efficiency term A_i. The binary variable $ITout_i$, indicates use of IT outsourcing and enters the production function as a shift parameter. Dividing by L_i, taking logs on both sides and adding an i.i.d. error term u_i gives the linear version of empirical model.

our control variables and our performance measure, we expect our exclusion restriction to hold.

In addition to accounting for selective item non-response in $Prod_i^*$, we allow for endogeneity of the outsourcing decision in Equation (3.1) via a standard IV approach. For that we use *Y2K consulting* as excluded instrument, which measures whether a firm resorted on external consultancy for the year 2000 problem (also known as the Y2K problem, or the millennium bug) (Ohnemus [2007]). The year 2000 problem was due to "short sighted" early computer programming, which stored only the last two digits of a year. This practice would have caused some date-related processes to operate incorrectly from January 1, 2000, onwards. Virtually all firms were equally confronted with the year 2000 problem. The extent of consulting services depended on how seriously the Y2K problem affected the firm's workflow. The instrument is valid if the year 2000 problems are unrelated to a firm's productivity in 2004. This would be violated if management focused on an operating system's use of two or four digits to store years when the purchase decision was made. Because this is a deep feature of programming, which did not receive broad media attention until the end of the 1990s, such a managerial focus seems unlikely. However, suffering from the Y2K problem may increase the likelihood of using further IT outsourcing services. After a firm gains experience in using external help to solve IT problems, management might be more inclined to outsource other IT activities as well.

3.3. Implementation Details

We implement our semiparametric estimator, which we derived in the previous section. In the first step of our estimation procedure, we estimate the selection probability function, which is used in the second step to weight the observations in the actual production function estimation. The second-step estimation applies these weights, but otherwise uses only those observations for which the dependent variable $Prod^*$ is observed, i.e., when $\Delta = 1$.

For the first step of the estimation, we need to introduce a link function G for a parametric model of the response mechanism Δ. The function G chosen coincides with the cumulative standard normal distribution Φ. Further, because of the estimation of the conditional probability $\mathbb{P}(\Delta = 1|V^* = v) = \Phi(v'\vartheta_0)$ we face a nonlinear optimization problem. To do so, we adopt the following choice of the starting value.

1. Estimate the parameter ϑ_s under missing completely at random (MCAR), i.e., the first entry of the parameter vector is the empirical analog of $\Phi^{-1}\Big(\mathbb{P}(\Delta = 1)\Big)$ and all other parameters are set to zero. In our application, we chose the first entry of ϑ_s somewhat smaller to ensure convergence of our optimization routine.

2. Linearize the estimation problem through a first-order Taylor approximation around ϑ_s, i.e.,

$$\mathbf{E}\left[\Big(\frac{\Delta}{\Phi(V'\vartheta)} - 1\Big)Z\right] \approx \mathbf{E}\left[\Big(\frac{\Delta}{\Phi(V'\vartheta_s)} - 1\Big)Z\right] - \mathbf{E}\left[\Big(\frac{\Delta V'\varphi(V'\vartheta_s)}{\Phi^2(V'\vartheta_s)}\Big)Z\right](\vartheta - \vartheta_s),$$

where φ is the standard normal probability density function. The norm of the linearization is minimized when ϑ coincides with

$$\vartheta^* \approx \vartheta_s + \mathbf{E}\left[\frac{\Delta V'Z\varphi(V'\vartheta_s)}{\Phi^2(V'\vartheta_s)}\right]^{-1} \mathbf{E}\left[\Big(\frac{\Delta}{\Phi(V'\vartheta_s)} - 1\Big)Z\right],$$

where we used that $\dim(V) = \dim(Z)$, which is satisfied in our application.

Moreover, our semiparametric estimation approach relies on the choice of smoothing parameter L used in our estimator $\widehat{g}_n$ and $\widehat{h}_n$ (see Section 2.3), which is implemented via cross validation. For the estimation of the finite sample variance of our estimator we use the bootstrap.

3.4. Data Description and Summary Statistics

We use data from a firm survey conducted via computer-aided telephone interviews by the Centre for European Economic Research (ZEW). The survey has a special focus on the diffusion and the use of information and communication technologies (ICT) at German companies. For our application, we use the 2004 wave of the data, which contain information on firms' IT outsourcing activities. The sample is drawn using a stratified sampling design, with stratification cells being defined by size class of the firm, industry affiliation, and two regions (East/West Germany).[10]

Following much of the literature we measure firms *average labor productivity* by total sales minus costs of intermediate inputs (in euros) per employee, $Prod_i^* = (sales - costs\,of\,intermediates)/L$. Missing values in $Prod_i^*$ stem from considerable item non-response to the survey questions on total sales, as well as costs of intermediates during the fiscal year.

The survey questionnaire covered the entire range of IT services companies that might need to operate their business, asking further whether the firms had outsourced each specific activity to an external service provider in whole or in part. We restrict the analysis to services that are required at every firm using computer technology in its business operations, namely, the (i) installation of hardware and software, (ii) computer system maintenance, and (iii) user assistance and support.[11] The constructed dummy variable for *IT outsourcing* used in our estimation takes the value of 1 if a firm outsources at least one of those three basic IT services completely and 0 otherwise.

As is the common case in respective firm-level survey data, no information is available to directly measure the physical capital stock of the firms. We therefore assume investment to be proportional to the capital stock and use *gross investment* figures as an empirical proxy for capital K (see, e.g., Bertschek and Kaiser [2004]). We measure *labor L* in full-time equivalent terms, assuming that a part-time employee represents half of a full-time employee. The instrumental variable chosen for *Y2K consulting* is a dummy variable that equals 1 if a firm resorted on external consultancy for the year 2000 problem (0 otherwise). We additionally control the firms' overall IT intensity. Thus, we include the share of employees working predominantly with personal computers in the model (*pcwork*). This measure is a common proxy for 'general purpose' IT and has been widely used in the IT productivity literature (e.g., Bloom et al. [2012], Bresnahan et al. [2002]). Moreover, we include 13 industry dummies constructed from two-digit standard industry codes (NACE)[12] and a dummy indicating whether the firm is located in Eastern Germany.

[10] As a sampling frame, the survey uses the data pool of "Verband der Vereine Creditreform" (CREDITREFORM), a credit rating agency, which provides the largest database on firms available in Germany.

[11] We thereby disregard more sophisticated IT services, which most of the firms in our sample do not need, such as software programming.

[12] See Table 1 for the industry distribution of the estimation sample.

The raw data for this paper consist of 3,801 observations.[13] In most items the share of missing values is well below 5%. In addition to our dependent variable, reported investments stands out, with about 27% missing observations in the raw data (see Table 2). For variables with modest missingness (rates below 5%), we regard the assumption of MCAR and applying listwise deletion as innocent. Additionally correcting for item non-response in an independent variable is possible, but would add considerable weight to our exposition. Hence, for this application, we assume MCAR for all independent variables and perform a complete case analysis in (X, W).[14] This simplification allows us to focus on the potential bias due to item non-response in the dependent variable *Prod** and how it can be corrected. We stress, however, that MCAR in investments is a strong assumption. A full estimation should correct for potential non-selective missingness in this variable.

Table 3 reports summary statistics of the resulting estimation sample. Our sample consists of 2,631 observations. The total number of observations for which *Prod** is missing is 537. As the number of employees is fully observed, the missing observations in *Prod** stem from item non-response to survey questions on sales and costs of intermediate inputs. The incidence of item non-response in our estimation sample totals about 20%. For reasons outlined above, we expect the missingness in *Prod** not to be random and the resulting bias to be far from negligible, given the considerable share of item non-response in our dependent variable.

3.5. Results

This section outlines the application of our estimation procedure. We also evaluate our estimator against the assumption of MAR. As a benchmark, we use listwise deletion as well as multiple imputation (Rubin [1978]), which in practice is the most commonly used strategy in dealing with item non-response in practice.

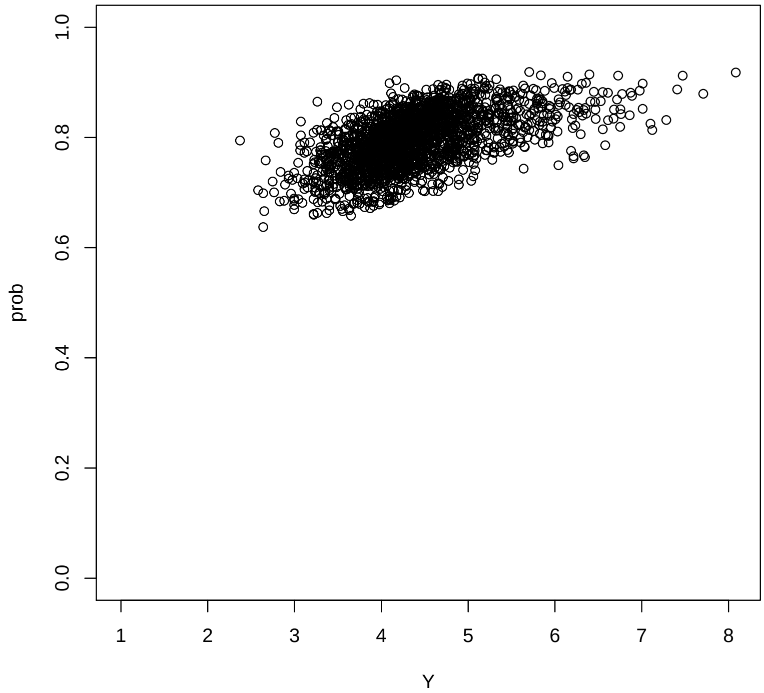
The first step of our estimation procedure accounts for non-random missingness of the dependent variable. In this step, we estimate the selection probability function, which makes it the critical step of our method. This estimation step involves the outcome measure as well as the indicator for IT outsourcing, *ITout*, which we use as our instrument for selection to model the response mechanism Δ. In addition, we include capital K, labor L, and $Y2K$, the indicator for Y2K-consulting, into the first-step estimation. In the second step, we then use the parameter estimates $\widehat{\vartheta}_n$ to compute the weighing factor of each observation. The weighing factors are then used to weight each observation in the actual production function estimation, which uses only observations for which the dependent variable *Prod** is observed.

Based on our first-step estimation results, Figure 1 shows how the probability of item non-response is related to observed realizations of the latent dependent variable *Prod**. For instance, if productivity is below 4, the estimated probability of reporting, given other controls is always below 1.Most importantly, the non-response is not random, but becomes less likely for larger values in *Prod*. Therefore, Figure 1 clearly suggests that MAR is violated with the data at hand. The first-step estimation results highlight the need to account for

[13] The complete survey data include 4,252 observations. We drop 369 observations from the sector 'electronic processing and telecommunication', because firms providing IT services to other companies typically belong to this sector and cannot be meaningfully included in the analysis. We further removed 82 observations with illogical values in input and output measures to arrive at a dataset of 3,801 valid observations.

[14] We could easily mirror datasets, such as the Census of Manufacturers (CM), and impute missing values in the investment variable [White et al., 2012], but this would still require assuming MAR.

Figure 1: Conditional Probability of Item Nonresponse and Observed Outcomes.



Notes: Figure 1 illustrates the first-step estimation using all 2,631 observations. It displays the estimator of the function $v \mapsto \mathbb{P}(\Delta = 1|V^* = v)$ evaluated at the realizations $(Prod_i, ITout_i, L_i, K_i, Y2K_i)$ when productivity is observed, i.e., $\Delta_i = 1$. The figure plots the estimated conditional probability against the observed realizations of $Prod^*$.

the nonignorable nonresponse in $Prod^*$ in the second step of our estimation procedure. In particular, as we observe a positive relationship between the probability of observing $Prod^*$ and the underlying value itself, estimation strategies relying on MAR will be based on too large values of $Prod^*$ leading us to underestimate the true population parameter β_0.

Table 5 shows our main results from the second step of our estimation procedure. All specifications estimate variants of the model discussed in Section 2. We show two groups of three columns. Columns 1-3 show the linearized version of the production function model in Equation (3.1). We estimate the model by two-stage least squares. Columns 4-6 correspond to the more flexible, and preferred partially linear model. This specification does not impose linearity on $m(\cdot)$. Columns 1 and 4 estimate the model, assuming that the dependent variable was MCAR and deleting the entire observation from the estimation (listwise deletion). Columns 2 and 5 use imputation techniques which assume the variable was missing at random (MAR) to keep the observation in the dataset. Columns 3 and 6 apply the correction developed in Section 2. Hence, column 6 shows our preferred estimator, which combines the correction for selective missingness with the inclusion of K and L in the non-parametric component of the model.

In all specifications, we control for labor and capital, and include 13 sector dummies, an indicator for a firm's location in Eastern Germany, as well as the measure for the firm's IT intensity in the second-stage estimation of the production function. We report bootstrap standard errors obtained using 500 repetitions. As suggested by the first-step results in Figure 1, we expect that assuming MCAR (columns 1 and 4) leads to much smaller coefficient estimates. In columns 2 and 5, we attempt to correct for the bias of

MCAR by using multiple imputation. We impute $ln(Prod^*)$ using all variables available in our estimation sample as predictors.[15] However, when MAR is violated, the imputation insufficiently corrects for the bias. In fact, we find the estimation results based on multiple imputations of $Prod^*$ to be close to the results obtained by listwise deletion.

In columns 3 and 6, we use the full estimation procedure we propose in this paper. While, in column 3, we apply the correction within the linear specification of the production model, column 6 shows the results for the partially linear model, which underlies the discussion in Section 2. Our correction for the selective missingness leads to considerably larger coefficient estimates in both the linear and the partially linear model. Given our finding of a positive relation between the response and the underlying value of the outcome, as well as the positive relationship of the outcome and outsourcing, this leads estimation based on listwise deletion (MCAR) or imputation (MAR) to underestimate the effect of outsourcing. In all our estimations, the standard errors are reasonably small to guarantee meaningful inference.

Regarding the interpretation of our result, we find positive and economically meaningful productivity returns to IT outsourcing (in all specifications). However, these positive returns are underestimated under the MAR assumption and when employing standard methodologies, such as listwise deletion (MCAR) and multiple imputation.

4. Conclusions

Selective item non-response is a major problem in all survey-based data. We propose a novel approach to correct for potential biases in the estimation of econometric models when the dependent variable is subject to missing data. Prevalent strategies in applied empirical research to deal with missing data rely on MAR, namely, listwise deletion or multiple imputation. We show that these approaches can lead to biased estimates of the central coefficients. The bias is most likely when the missingness is related to the independent variables in systematic ways, and its sign depends on this relationship.

We develop a new estimation approach that can be used in IV estimation and is robust to selectively missing realizations of the dependent variable. The approach is based on a second set of instrumental variables that affect selection only through potential outcomes. We apply our proposed method to revisit the estimation of productivity returns to IT outsourcing. We argue that in such settings, that is, production function estimation based on survey data, MAR is likely violated. Our empirical application in fact supports this hypothesis. Our estimator is easily applied, and we find positive and economically meaningful productivity returns to IT outsourcing. Importantly, the positive returns are underestimated when standard methodologies are employed that assume MAR (listwise deletion and multiple imputation).

Our results highlight the consequences of the widely used MAR assumption within a broadly applied class of empirical models (production function estimation). The literature dealing with estimation of production functions has so far focused on bias due to endogenous input choice and on endogeneity through panel attrition and unit non-response (firm exit) (cf. Olley and Pakes [1996], Levinsohn and Petrin [2003], Melitz and Polanec [2015]). We highlight that, in addition, imposing MAR on missing values in the *dependent* variable is likely to yield biased estimates in this context.

[15] The imputation was conducted using the *R* package *mi* (https://cran.r-project.org/web/packages/mi/). We generate $m = 5$ datasets and combine the individual estimation results according to Rubin [1978].

Finally, our new estimator can be fruitfully used in applied empirical research with either continuous or binary instruments. We provide a semiparametric version, and a version for linear IV (2SLS) of the estimator, and we show an application for the broad class of production function estimation models. However, we note that the relevance of selective missingness of the dependent variable in our application carries over to many other applications and important datasets, such as the US Census Bureau's Census of Manufacturers, the IAB establishment panel, or other firm-, individual-, and household-level surveys.

A. Summary Statistics and Estimation Tables

Table 1: Industry Distribution

	Obs.	Percent
consumer goods	251	9.54
chemical industry	138	5.25
other raw materials	239	9.08
metal and machine construction	309	11.74
electrical engineering	177	6.73
precision instruments	230	8.74
automobile	167	6.35
wholesale trade	135	5.13
retail trade	199	7.56
transportation & postal services	202	7.68
banks & insurances	154	5.85
technical services	230	8.74
other business-related services	200	7.60
Total	2631	100.00

Notes: This table shows the number of firms in the estimation sample by industry. Source: ZEW ICT-Survey 2004.

Table 2: Summary Statistics: Raw Data

	ln(prod)	ln(Capital)	ln(Labor)	outsource	y2k-Consult	pcwork	ost
nobs	3801	3801	3801	3801	3801	3801	3801
NAs	1389	1058	0	34	172	13	0
Mean	4.399	5.124	4.075	0.361	0.521	44.990	0.246
Stdev	0.720	2.583	1.750	0.480	0.500	32.115	0.431
Median	4.306	5.017	3.951			40.000	
Minimum	2.372	0.000	0.000			0.000	
Maximum	8.083	14.809	11.002			100.000	

Notes: Summary statistics of the main variables in the raw dataset. The variables are shown in the columns, the descriptive statistics in the rows.
Source: ZEW ICT-Survey 2004.

Table 3: Summary Statistics: Estimation Sample

	ln(prod)	ln(Capital)	ln(Labor)	outsource	y2k-Consult	pcwork	ost
nobs	2631	2631	2631	2631	2631	2631	2631
NAs	535	0	0	0	0	0	0
Mean	4.388	5.129	3.889	0.394	0.521	42.966	0.261
Stdev	0.710	2.570	1.707	0.489	0.500	31.726	0.439
Median	4.287	5.081	3.761			33.000	
Minimum	2.372	0.000	0.000			0.000	
Maximum	8.083	14.809	11.002			100.000	

Notes: Summary statistics for the main variables of interest in the estimation sample. The variables are shown in the columns, the descriptive statistics in the rows.
Source: ZEW ICT-Survey 2004.

Table 4: First Stage Results (two-stage least squares)

	Dependent variable	
	outsource	
	(1)	(2)
y2k-Consult		0.210*** (0.019)
ln(Labor)	−0.046*** (0.009)	−0.059*** (0.008)
ln(Capital)	−0.003 (0.006)	−0.002 (0.005)
pcwork	−0.002*** (0.0003)	−0.002*** (0.0003)
ost	0.044** (0.021)	0.046** (0.021)
consumer goods	−0.049 (0.046)	−0.063 (0.045)
chemical industry	−0.036 (0.054)	−0.060 (0.053)
other raw materials	−0.073 (0.047)	−0.093** (0.046)
metal and machine construction	−0.129*** (0.044)	−0.128*** (0.043)
electrical engineering	−0.237*** (0.050)	−0.220*** (0.049)
precision instruments	−0.122*** (0.047)	−0.125*** (0.045)
automobile	−0.060 (0.052)	−0.072 (0.051)
wholesale trade	0.010 (0.053)	−0.013 (0.052)
retail trade	−0.076 (0.049)	−0.099** (0.048)
transportation and postal services	−0.067 (0.049)	−0.069 (0.048)
banks and insurances	0.029 (0.051)	0.006 (0.050)
electronic processing and telecommunication	−0.192*** (0.046)	−0.176*** (0.045)
other business-related services	0.731*** (0.045)	0.675*** (0.044)
Observations	2,631	2,631
Adjusted R^2	0.055	0.098
F Statistic	10.545*** (df = 16; 2614)	17.833*** (df = 17; 2613)

Notes: This table shows the first stage results of the two stage least squares regression using the full estimation sample. The coefficient for the excluded instrument, *Y2K*, is shown in the first row. Column 1 shows estimation results without the excluded instrument, and column 2 provides regression result including the instrument. Significant at 1% ***, significant at 5% **, significant at 10% *.

Table 5: Main Results.

	Linear Estimation			Semiparametric Estimation		
	Listw. Deletion(MCAR)	Imputation(MAR)	Correction	Listw. Deletion(MCAR)	Imputation(MAR)	Correction
outsource	0.431*** (0.144)	0.457*** (0.151)	0.591* (0.309)	0.461*** (0.145)	0.499*** (0.157)	0.591* (0.315)
controls	YES	YES	YES	YES	YES	YES
N	2096	2631	2631	2096	2631	2631

NOTES: This table shows our main estimation results for the effect of IT outsourcing on productivity. The dependent variable is log(sales-costs of intermediates/employees). Standard errors are in parenthesis. The coefficient of interest is the effect of IT outsourcing on productivity. We use the "Y2K bug consulting" as the IV for IT outsourcing. Columns 1-3 show the results of the fully parametric model, using two-stage least squares and instrumenting IT outsourcing using the Y2K bug. The specifications in Columns 4-6 nonparametrically control for investment and firm size (number of employees), and parametrically for the share of employees working predominantly with personal computers, sector, location in Eastern Germany, and whether the company indicated interest in the survey results. Columns 4-6 use non-parametric Kernel estimation. Columns 1 and 4 show the biased benchmark using listwise deletion (MCAR). Columns 2 and 5 correct for missingness using imputation methods that assume missing at random (MAR). Columns 3 and 6 apply the correction for selective non-response that we present in this paper. All standard errors are obtained using bootstrap with 200 repetitions. Significant at 1% ***, significant at 5% **, significant at 10% *.

B. Appendix

B.1. Asymptotic Distribution of the Estimator

We begin this Appendix by giving the conditions under which the asymptotic distribution result summarized in Theorem 2.2 is valid.

Assumption 5. *(i) We observe a sample $((\Delta_1, Y_1, X_1, W_{11}, W_{21}), \ldots, (\Delta_n, Y_n, X_n, W_{1n}, W_{2n}))$ of independent and identical distributed (i.i.d.) copies of (Δ, Y, X, W_1, W_2) where $Y = \Delta Y^*$. (ii) It holds $\sup_w \|p^L(w)\|^2 = O(L)$ with $L \equiv L(n)$ and $L^2/n = o(1)$. (iii) The smallest eigenvalue of $\mathbf{E}[p^L(W_1)p^L(W_1)']$ is bounded away from zero uniformly in n. (iv) There exists $\gamma, \kappa \in \mathbb{R}^L$ such that $n \sup_w |\gamma' p^L(w) - g(w, \vartheta_0)| = o(1)$ and $n \sup_w |\kappa' p^L(w) - h(w)| = o(1)$. (v) The parameter space Θ is compact; the function G is differentiable and $\|G_\vartheta(v'\theta)\|$ is bounded for every v and $\theta \in \Theta$.*

Assumption 5 (*ii*) – (*iii*) restricts the magnitude of the approximating functions $\{p_j\}_{j\geqslant 1}$ and imposes nonsingularity of their second moment matrix. It is a standard assumption for series estimators (cf., *e.g.*, Assumption 2 in Newey [1997]). Assumption 5 (*ii*) holds for instance for polynomial splines, Fourier series and wavelet bases. Assumption 5 (*iv*) determines imposes an undersmoothing condition on the sieve approximation errors which characterize the bias of the estimated regression functions g and h. This ensures that these sieve approximation biases in our estimation procedures become asymptotically negligible. In addition to this, we require smoothness of the function G.

Proof of Theorem 2.2. We observe

$$\sqrt{n}\,\mathbf{E}\Big[(X - \mathbf{E}(X|W_1))W_2\Big]\Big(\widehat{\beta}_n - \beta_0\Big)$$

$$= \underbrace{n^{-1/2}\sum_{i=1}^n \Big(W_{2i}\Big(Y_i/G(V_i'\vartheta_0) - \Pi_L g(W_{1i})\Big) - \mathbf{E}\Big[W_2\Big(Y/G(V'\vartheta_0) - \Pi_L g(W_1)\Big)\Big]\Big)}_{I}$$

$$+ \underbrace{n^{-1/2}\sum_{i=1}^n Y_i\Big(W_{2i} - p^L(W_{1i})'\,\mathbf{E}[W_2 p^L(W_1)]\Big)\Big(1/G(V_i'\widehat{\vartheta}_n) - 1/G(V_i'\vartheta_0)\Big)}_{II}$$

$$+ \underbrace{n^{-1/2}\sum_{i=1}^n W_{2i}\Big(\Pi_L g(W_{1i}) - g(W_{1i})\Big)}_{III} + o_p(1).$$

For some $\overline{\vartheta}_n$ between ϑ_0 and $\widehat{\vartheta}_n$ we have

$$II = \sqrt{n}(\widehat{\vartheta}_n - \vartheta_0)' n^{-1}\sum_{i=1}^n \Big(W_{2i} - p^L(W_{1i})'\,\mathbf{E}[W_2 p^L(W_1)]\Big) Y_i V_i\, G_\vartheta(V_i'\overline{\vartheta}_n)/G^2(V_i'\overline{\vartheta}_n)$$

$$= \sqrt{n}(\widehat{\vartheta}_n - \vartheta_0)'\,\mathbf{E}\Big[(W_2 - \mathbf{E}[W_2|W_1]) Y V G_\vartheta(V'\vartheta_0)/G^2(V'\vartheta_0)\Big] + o_p(1)$$

by the uniform law of large numbers. Since $\widehat{\vartheta}_n$ is the GMM estimator of ϑ_0 it is well known that

$$\sqrt{n}(\widehat{\vartheta}_n - \vartheta_0) = (A'A)^{-1}A'\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i\Big(\Delta_i/G(V_i'\vartheta_0) - 1\Big) + o_p(1)$$

where $A = \mathbf{E}\Big[ZV'G_\vartheta(V'\vartheta_0)/G^2(V'\vartheta_0)\Big]$. This computation yields

$$\begin{aligned}\frac{\sqrt{n}}{\sigma}\Big(I+II\Big) = \sum_{i=1}^{n} \frac{}{\sqrt{n}\sigma}\Big(W_{2i}\Big(Y_i/G(V_i'\vartheta_0) - g(W_{1i})\Big) - \mathbf{E}\Big[W_2\Big(Y/G(V'\vartheta_0) - g(W_1)\Big)\Big] \\ - \Big(\Delta_i/G(V_i'\vartheta_0) - 1\Big)Z_i'\,A(A'A)^{-1}\,\mathbf{E}\Big[(W_2 - \mathbf{E}[W_2|W_1])YVG_\vartheta(V'\vartheta_0)/G^2(V'\vartheta_0)\Big]\Big) + o_p(1) \\ = \sum_{i=1}^{n} s_{in} + o_p(1).\end{aligned}$$

Moreover, s_{in}, $1 \leqslant i \leqslant n$ satisfy the Lindeberg conditions. The central limit theorem of Lindeberg-Feller thus implies $\sum_{i=1}^{n} s_{in} \xrightarrow{d} \mathcal{N}(0,1)$. Finally, $nIII = o_p(1)$ due to undersmooting, which completes the proof. □

References

K. G. Abraham and S. K. Taylor. Firms' use of outside contractors: Theory and evidence. *Journal of Labor Economics*, 14(3):394–424, 1996.

M. Amiti and J. Konings. Trade liberalization, intermediate inputs, and productivity: Evidence from Indonesia. *The American Economic Review*, 97(5):1611–1638, 2007.

M. Amiti and S.-J. Wei. Service offshoring and productivity: Evidence from the United States. *The World Economy*, 32(2):203–220, 2009.

I. Bertschek and U. Kaiser. Productivity effects of organizational change: microeconometric evidence. *Management Science*, 50(3):394–404, 2004.

S. E. Black and L. M. Lynch. How to compete: the impact of workplace practices and information technology on productivity. *Review of Economics and Statistics*, 83(3):434–445, 2001.

N. Bloom, R. Sadun, and J. Van Reenen. Americans do IT better: Us multinationals and the productivity miracle. *The American Economic Review*, 102(1):167–201, 2012.

T. F. Bresnahan, E. Brynjolfsson, and L. M. Hitt. Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The Quarterly Journal of Economics*, 117(1):339–376, 2002.

C. Breunig. Testing missing at random using instrumental variables. Technical report, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany, 2015.

C. Breunig, E. Mammen, and A. Simoni. Nonparametric estimation in case of endogenous selection. Technical report, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany, 2015.

E. Brynjolfsson and L. M. Hitt. Computing productivity: Firm-level evidence. *Review of economics and statistics*, 85(4):793–808, 2003.

R. H. Coase. The nature of the firm. *Economica*, 4(16):386–405, 1937.

X. D'Haultfoeuille. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154(1):1–15, 2010.

J. Drechsler. Multiple imputation of missing values in the wave 2007 of the IAB establishment panel. Technical report, IAB discussion paper, 2010.

R. C. Feenstra. Integration of trade and disintegration of production in the global economy. *The Journal of Economic Perspectives*, pages 31–50, 1998.

L. Foster, J. Haltiwanger, and C. Syverson. Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *The American Economic Review*, 98(1):394–425, 2008.

J. R. Frick and M. M. Grabka. Item non-response and imputation of annual labor income in panel surveys from a cross-national perspective. *Survey Methods in multicultural, multinational, and multiregional Contexts. Wiley, New York*, 2010.

H. Görg, A. Hanley, and E. Strobl. Productivity effects of international outsourcing: evidence from plant-level data. *Canadian Journal of Economics/Revue canadienne d'Économique*, 41(2):670–688, 2008.

G. M. Grossman and E. Helpman. Outsourcing in a global economy. *Review of Economic Studies*, 1(72):135–159, 2005.

S. J. Grossman and O. D. Hart. The costs and benefits of ownership: A theory of vertical and lateral integration. *The Journal of Political Economy*, 4(94):691–719, 1986.

L. Halpern, M. Koren, and A. Szeidl. Imported inputs and productivity. *The American Economic Review*, 105(12):3660–3703, 2015.

K. Han, R. J. Kauffman, and B. R. Nault. Returns to information technology outsourcing. *Information Systems Research*, 22(4):8241–7840, 2011.

J. Heckman. Shadow prices, market wages, and labor supply. *Econometrica: Journal of the Econometric Society*, pages 679–694, 1974.

A. B. Kennickell. Multiple imputation in the survey of consumer finances. In *Proceedings of the Section on Survey Research Methods*, 1998.

P. Kline and A. Santos. Sensitivity to missing data assumptions: Theory and an evaluation of the US wage structure. *Quantitative Economics*, 4(2):231–267, 2013.

M. C. Lacity, S. A. Khan, and L. P. Willcocks. A review of the IT outsourcing literature: Insights for practice. *The Journal of Strategic Information Systems*, 18(3):130–146, 2009.

J. Lee, J. S. McCullough, and R. J. Town. The impact of health information technology on hospital productivity. *The RAND Journal of Economics*, 44(3):545–568, 2013.

J. Levinsohn and A. Petrin. Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies*, 70(2):317–341, 2003.

L. Loh and N. Venkatraman. Diffusion of information technology outsourcing: influence sources and the kodak effect. *Information Systems Research*, 3(4):334–358, 1992.

M. J. Melitz and S. Polanec. Dynamic olley-pakes productivity decomposition with entry and exit. *The RAND Journal of Economics*, 46(2):362–375, 2015.

N. M. Menon, B. Lee, and L. Eldenburg. Productivity of information systems in the healthcare industry. *Information Systems Research*, 11(1):83–92, 2000.

W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147 – 168, 1997.

J. Ohnemus. Does IT outsourcing increase firm success? an empirical assessment using firm-level data. *ZEW Discussion Paper*, 07–087, 2007.

G. S. Olley and A. Pakes. The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297, 1996.

A. Petrin and J. Levinsohn. Measuring aggregate productivity growth using plant-level data. *The RAND Journal of Economics*, 43(4):705–725, 2012.

E. A. Ramalho and R. J. Smith. Discrete choice non-response. *The Review of Economic Studies*, 80(1):343–364, 2013.

J. Reardon, R. Hasty, and B. Coe. The effect of information technology on productivity in retailing. *Journal of retailing*, 72(4):445–461, 1997.

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

D. B. Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association, 1978.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley & Sons, 1987.

P. Schreyer and D. Pilat. Measuring productivity. *OECD Economic studies*, 33(2):127–170, 2001.

G. Tang, R. J. Little, and T. E. Raghunathan. Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90(4):747–764, 2003.

K. J. Thompson and K. T. Washington. Challenges in the treatment of unit nonresponse for selected business surveys: a case study. *Survey Methods: Insights from the Field (SMIF)*, 2013. URL http://surveyinsights.org/?p=2991.

T. K. White, J. P. Reiter, and A. Petrin. Plant-level productivity and imputation of missing data in us census manufacturing data. NBER Working Paper 17816, National Bureau of Economic Research, 2012. URL http://www.nber.org/papers/w17816.

O. E. Williamson. Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics*, 2(22):233–261, 1979.

O. E. Williamson. The economics of organization: The transaction cost approach. *American Journal of Sociology*, 3(87):548–577, 1981.

O. E. Williamson. *Transaction cost economics*, volume 1, chapter Handbook of Industrial Organization, pages 135–182. North-Holland, 1989.

J. M. Wooldridge. On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, 104(3):112–114, 2009.

ZEW. *Interaktiv, mobil, international 1717 Unternehmen im Zeitalter von Web 2.0, IKT-Report*. Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim, 2010.

J. Zhao and J. Shao. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110(512): 1577–1590, 2015.