Inverse Probability Weighted Generalised Empirical Likelihood Estimators

Inkmann, J.

Publication date:
2005

CentER

Discussion Paper

No. 2005–131

INVERSE PROBABILITY WEIGHTED GENERALISED EMPIRICAL LIKELIHOOD ESTIMATORS: FIRM SIZE AND R&D REVISITED

By Joachim Inkmann

December 2005

ISSN 0924-7815

TILBURG UNIVERSITY

Inverse Probability Weighted Generalised Empirical Likelihood Estimators: Firm Size and R&D Revisited

Joachim Inkmann*

Abstract: The inverse probability weighted Generalised Empirical Likelihood (IPW-GEL) estimator is proposed for the estimation of the parameters of a vector of possibly non-linear unconditional moment functions in the presence of conditionally independent sample selection or attrition. The estimator is applied to the estimation of the firm size elasticity of product and process R&D expenditures using a panel of German manufacturing firms, which is affected by attrition and selection into R&D activities. IPW-GEL and IPW-GMM estimators are compared in this application as well as identification assumptions based on independent and conditionally independent sample selection. The results are similar in all specifications.

Keywords: Generalised Empirical Likelihood, Inverse Probability Weighting, Propensity Score, Conditional Independence, Missing at Random, Selection, Attrition, R&D

JEL class.: C13, C33, O31

* Econometrics and Finance Group, CentER, Tilburg University and Netspar, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. An earlier version of this paper has benefited from comments of Irene Bertschek, Richard Blundell, Michael Lechner, Torsten Persson, Winfried Pohlmeier and Rainer Schnell. I am grateful to the Centre for European Economic Research (ZEW) for making available the Mannheim Innovation Panel (MIP), in particular to Sandra Gottschalk and Georg Licht for their expertise and support in the preparation of the data.

Introduction

Assume we are concerned with estimating the unknown qx1 parameter vector β_0, which describes some characteristics of the distribution of the random vector $Z_1' = (Y', X')$ satisfying the orthogonality condition $E[\psi_1(Z_1,\beta_0)] = 0$ where $\psi_1(Z_1,\beta)$ is a rx1 vector of unconditional moment functions (the use of the subscript 1 becomes obvious later). These moment functions may be derived from a structural or reduced form model of an economic theory of interest and are general enough to cover multiple equation models of linear and non-linear form. Suppose we have a random sample { Z_{1i}: i = 1,...,n } of Z_1 and the equation system is exactly identified, i.e. r = q. Then a method of moments estimator $\hat{\beta}_{MM}$ of β_0 can be derived from an application of the analogy principle outlined by Manski (1988) by solving

$$\frac{1}{n}\sum_{i=1}^{n}\psi_1\left(Z_{1i},\hat{\beta}_{MM}\right) = 0 . \tag{1}$$

The estimator is consistent and asymptotically normal under mild regularity conditions. Assume now that instead of observing a random sample of Z_1, we observe a random sample of { (Z_{1i},W_i,A_i): i = 1,...,n } from the distribution of Z_1 given A = 0 with mean $E[\psi_1(Z_1,\beta_0) \mid A = 0]$. A is an indicator variable which indicates selection by A = 0 and non-selection by A ≠ 0. W is a vector of attributes that is observed for any outcome of A. The sample average (1) computed on the basis of selected observations will not uniformly converge to the population moment of interest unless selection occurs independent of Z_1, i.e. $\Pr(A = 0 \mid Z_1) = \Pr(A = 0)$. However, working under the conditional independence assumption (CIA) that selection occurs independent of Z_1 given W, i.e. $\Pr(A = 0 \mid Z_1, W) = \Pr(A = 0 \mid W)$, implies that the population moments can be identified from $E[\psi_1(Z_1,\beta_0)] = E[\psi_1(Z_1,\beta_0)1(A = 0)/\Pr(A = 0 \mid W)]$ and suggests an inverse probability weighted (IPW) moment estimator $\hat{\beta}_{MM}^{IPW}$ of β_0 solving

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\hat{p}_i}\psi_1\left(Z_{1i},\hat{\beta}_{MM}^{IPW}\right)1(A_i = 0) = 0 , \tag{2}$$

where $\hat{p}_i$ denotes a consistent estimator of $\Pr(A_i = 0 \mid W_i)$. In Section 1 we will discuss CIA in more detail and provide references. For (2) the assumption is maintained that r = q. If we give up this assumption in favour of an over-identified equation system, i.e. r > q, (2) cannot be solved but an inverse probability weighted GMM estimator $\hat{\beta}_{GMM}^{IPW}$ (see Wooldridge (2002) for a general treatment of IPW-M estimators) follows straightforwardly. However, an IPW Generalised Empirical Likelihood (GEL) estimator $\hat{\beta}_{GEL}^{IPW}$ of β_0 can be obtained from (2) by replacing the empirical distribution probabilities 1/n with probabilities $\hat{\pi}_i$ satisfying

$$\sum_{i=1}^{n}\frac{\hat{\pi}_i}{\hat{p}_i}\psi_1\left(Z_{1i},\hat{\beta}_{GEL}^{IPW}\right)1(A_i = 0) = 0 . \tag{3}$$

Since there may be a large number of $\hat{\pi}_i$ for i = 1,...,n satisfying (3) a further criterion for selecting the probabilities is needed which consists of minimising an information criterion between $\hat{\pi}_i$ and 1/n (see, e.g., Newey and Smith (2004)). Section 2 overviews the recent literature on GEL estimation and emphasises possible advantages of GEL over GMM.

The econometric framework described so far is motivated by our attempt to estimate the relationship between firm size and R&D expenditures from the Mannheim Innovation Panel (MIP) of German manufacturing firms. In particular, we are concerned with a test of the Klepper (1996) and Cohen and Klepper (1996a, b) hypothesis that the firm size elasticity of R&D expenditures related to process innovations (in short: process R&D) should exceed the corresponding elasticity of R&D expenditures devoted to product innovations (product R&D). Section 3 introduces to the application, describes the data and the empirical specification. For the moment we note that selection affects the second wave of the MIP, where A = 0 describes firms observed in the first wave which stay in the panel and remain engaged in R&D.

The paper contributes to both the econometric literature on IPW and GEL estimators and the empirical industrial organisation literature on R&D, firm size and exit by

- introducing IPW-GEL estimators as a natural alternative to IPW-GMM estimators,
- comparing IPW-GEL and IPW-GMM estimators in an application to firm panel data,
- comparing outcomes under CIA and the stronger independence assumption (IA),
- mixing CIA and IA in an environment of sequential selection for a sensitivity analysis,
- investigating the joint distribution of the estimated probabilities $\hat{p}_i$ and $\hat{\pi}_i$ in (3),
- testing the Cohen and Klepper hypothesis regarding size and process and product R&D.

Estimation results are presented and discussed in Section 4. It turns out that

- GEL and GMM estimates are almost identical with slight efficiency advantages to GEL.
- Outcomes under CIA, IA and mixtures of both assumptions are very similar.
- Firms with a high conditional selection probability $\hat{p}_i$ tend to receive a slightly higher $\hat{\pi}_i$.
- The Cohen and Klepper hypothesis is clearly rejected in all empirical specifications.

We conclude in Section 5 with a discussion of the findings and an outlook on future research.

1. Identification: Conditional Independence Assumption

As already pointed out in the introduction, sample selection may not be harmful for any econometric analysis if one has reasons to assume that selection occurs independent of the variables of interest entering the orthogonality conditions $E[\psi_1(Z_1,\beta_0)] = 0$, i.e. $\Pr(A = 0 \mid Z_1)$ $= \Pr(A = 0)$. It has become standard to use the following notation for independence

$$Z_1 \coprod A = 0 . \tag{4}$$

This assumption is implicitly maintained in all empirical studies, which are based on selected observations only. In particular, all panel data applications, which do not address attrition as a selection mechanism, implicitly assume that unit non-response is independent of the variables of interest. This certainly applies to the overwhelming majority of panel data applications. Notable exceptions are the monographs by Lechner (1995), Rendtel (1995) and Schnell (1997) as well as work to which we will refer below. In a panel data framework, (4) is compatible with both balanced and unbalanced panel data analysis. The former simply conditions on sample units that are observed in all waves, while the latter conditions on observed units in each wave separately. The argument for using an unbalanced panel is one of achieving (asymptotic) efficiency gains not one of imposing weaker identification assumptions in the presence of sample selection. In panel data applications as well as in cross-sectional analysis, the independence assumption (IA) (4) usually appears to be very strong.

In order to weaken (4) econometricians increasingly refer to the statistics literature, in particular to Rubin (1976, 1977, 1978) and consider identification of the moments (and finally parameters) of interest on the basis of the conditional independence assumption (CIA) $\Pr(A = 0 \mid Z_1, W) = \Pr(A = 0 \mid W)$, or equivalently using the notation from (4) [1]

$$Z_1 \coprod A = 0 \mid W . \tag{5}$$

Under CIA independence is achieved by conditioning on a vector of conditioning variables or attributes (Holland (1986)) affecting both selection and the variables Z_1. Depending on the area of application CIA is also referred to as a Missing at Random assumption (Rubin (1976)), ignorability (Little and Rubin (1987)), unconfoundedness (Rubin (1978), Rosenbaum and Rubin (1983)) and selection on observables (Fitzgerald, Gottschalk and Moffitt (1998)).

The latter designation emphasises the fact that W has to be observed for both selected and non-selected units, which demands for rich data sets in many applications, and also serves to separate CIA from more traditional econometric techniques in the selection on unobservables framework of Heckman (1979) and Hausman and Wise (1979). CIA rules out correlation between the error terms of the selection equation and the equations of interest after conditioning on W. Identification on the basis of the selection on unobservable assumption, however, depends either on the functional form (Heckman (1979)) or on exclusion restrictions (Newey (1999)). These requirements are certainly not without problems as well. Another frequently cited argument against CIA (e.g. Heckman (2001)) is the problem that CIA is usually not testable. Hirano, Imbens, Ridder and Rubin (2001), however, present a test for CIA with respect to attrition in panel data that can be applied if refreshment samples are available. A strong argument in favour of CIA against selection on unobservables is the limited applicability of the latter assumption in more complicated non-linear and multiple equation models as multi-period panel data models (see Ridder (1990) and Verbeek and Nijman (1996)). As already noted in the introduction and shown below in more detail, CIA suggests feasible estimators in a general non-linear multiple equation modelling framework.[2]

These estimators make use of the conditional selection probability[3] $\Pr(A = 0 \mid W)$ which has been central to the analysis of Rosenbaum and Rubin (1983) who show that

$$W \coprod A = 0 \mid \Pr(A = 0 \mid W) \tag{6}$$

$$Z \coprod A = 0 \mid \Pr(A = 0 \mid W). \tag{7}$$

[1] See Dawid (1979), Holland (1986) and Angrist (1997) for a statistics perspective on CIA.
[2] Since we emphasise CIA's applicability to non-linear functions of Z, we refrain from describing the conditional mean independence assumption (CMIA, see, e.g. Hirano, Imbens, and Ridder (2003)) as a somewhat weaker identification condition in models based on orthogonality conditions which are linear in Z.
[3] Also known as the propensity score or balancing score (Rosenbaum and Rubin, 1983).

The so-called balancing property (6) of the conditional selection probability is sometimes used to assess the quality of an estimator of $\Pr(A=0 \mid W)$ in applied work (see Dehejia and Wahba (1999)) but follows mechanically from the definition of the conditional selection probability. (7) is implied by CIA and reduces the dimension of the conditioning set in (5) to one.

As noted in the introduction CIA and the conditional selection probability $\Pr(A=0 \mid W)$ can be used to derive the following important result

$$
\begin{aligned}
&E\left[\psi_1(Z_1,\beta_0)\frac{1(A=0)}{\Pr(A=0 \mid W)}\right] = E\left[E\left[\psi_1(Z_1,\beta_0)\frac{1(A=0)}{\Pr(A=0 \mid W)} \mid W\right]\right] \\
&= E\left[E\left[\frac{\psi_1(Z_1,\beta_0)}{\Pr(A=0 \mid W)} \mid W, A=0\right]\Pr(A=0 \mid W)\right] \\
&= E\left[E\left[\frac{\psi_1(Z_1,\beta_0)}{\Pr(A=0 \mid W)} \mid W\right]\Pr(A=0 \mid W)\right] = E\left[\psi_1(Z_1,\beta_0)\right] = 0
\end{aligned}
\tag{8}
$$

where CIA is exploited in the transition to the last row. Thus, the orthogonality conditions of interest are recovered from weighting the selected unconditional moment functions with the inverse of the conditional selection probability. This result has led to a large number of publications advocating the use of inverse probability weighting (IPW) approaches – in the tradition of Horvitz and Thompson (1952) – to correct for sample selection and unit non-response bias under CIA as practised in equation (2). Examples include Robins and Rotnitzky (1995), Robins, Rotnitzky and Zhao (1995), Fitzgerald, Gottschalk and Moffitt (1998), Horowitz and Manski (1998), Hirano, Imbens, Ridder and Rubin (2001), Abowd, Crepon and Kramarz (2001) and Wooldridge (2002).[4]

2. Estimation: IPW Generalised Empirical Likelihood

If the moment functions $\psi_1(Z_1,\beta)$ are exactly identified, i.e. r = q, the result (8) can be used to construct the IPW-MM estimator described in (2) by replacing the unknown conditional selection probabilities $\Pr(A=0 \mid W)$ with a consistent parametric or non-parametric first step estimator $\hat{p}$. If there are over-identifying restrictions, i.e. r > q, a GMM (Hansen (1982)) estimator follows straightforwardly from using the same weighted moment functions as in (2). In both cases, the first step estimator $\hat{p}$ affects the variance-covariance matrix of the limiting distribution of the stabilising transformation of the second step IPW estimator of β_0. The necessary variance-covariance adjustments are described by Newey (1984) and Newey and McFadden (1994) for a parametric first estimation step and by Newey (1994) and Newey and McFadden (1994) for a non-parametric first estimation step. Wooldridge (2002) considers a class of IPW-M estimators, which are based on a first step Maximum Likelihood (ML) estimator $\hat{p}$ and shows that the unadjusted variance-covariance matrix is larger (in a matrix sense) than the correctly adjusted matrix.[5]

[4] A closely related literature considers IPW estimation of average treatment effects using an estimator of the propensity score (see, e.g., Hirano, Imbens, and Ridder (2003)).
[5] Wooldridge argues that the resulting conservative inference may be desirable. It may also be misleading.

Wooldridge also shows that using an estimator of the conditional selection probability instead of the true (but usually unknown) probability $\Pr(A=0\mid W)$ helps improving upon the asymptotic efficiency of the IPW-M estimator. The semi-parametric efficiency bound for a class of IPW-GEE estimators (Generalised Estimating Equations; GMM estimators, which employ a particular set of instruments) suggested by Robins, Rotnitzky and Zhao (1995), which also rely on a first step ML estimator of the conditional selection probability, is derived by Robins and Rotnitzky (1995).[6] Robins, Rotnitzky and Zhao (1995) and Wooldridge (2002) show that the bound can be approached by enlarging the parametric model of the conditional selection probability with additional variables (like quadratic terms of continuous variables and interactions), even if they have a true coefficient of zero.[7] Robins, Rotnitzky and Zhao (1995) propose first step ML logit estimators, Fitzgerald, Gottschalk and Moffitt (1998) first step ML probit estimators for the conditional selection probability in order to correct for attrition in a panel data framework using an IPW estimator exploiting (8). The parametric approach is maintained in the following discussion by specifying $\Pr(A=0\mid W)=p(W'\gamma_0)$, where a logit (probit) model results for p(.) equal to the c.d.f. of the logistic (standard normal) distribution.[8]

Instead of sequentially estimating γ_0 and β_0 we follow Newey (1984) and consider joint estimation of $\theta_0'=(\beta_0',\gamma_0')$ by stacking the weighted moments of interest with the vector of moment functions describing the first estimation step. Let $\psi_2(Z_2,\gamma)$ be the score for the ML estimator of γ_0 where $Z_2'=(A,W')$. The score satisfies $E[\psi_2(Z_2,\gamma_0)]=0$ which suggests computing a moment based estimator of θ_0 from a random sample of $Z'=(Z_1',Z_2')$ on the basis of the following extended set of moment functions

$$\psi(Z,\theta)=\begin{pmatrix}\psi_1(Z_1,\beta)\dfrac{1(A=0)}{p(W'\gamma)}\\ \psi_2(Z_2,\gamma)\end{pmatrix}. \tag{9}$$

As a result of this joint estimation approach, the variance-covariance formula of the estimator being used (e.g. GMM for r > q) automatically performs the necessary variance-covariance matrix adjustment for the estimator of β_0 which we discussed before. Abowd, Crepon and Kramarz (2001) consider an IPW-GMM estimator on the basis of (9) in a multi-period panel data framework. In such an application moment functions like (9) are defined for each period and stacked in one vector, which is used for the GMM estimator.[9]

---

[6] See Newey (1990) for an introduction to the concept of semi-parametric efficiency. In our context, a semi-parametric efficient estimator reaches the Cramér-Rao bound for any consistent and asymptotically normal (regular) estimator of a parametric sub-model satisfying the orthogonality conditions.

[7] Hahn (1998) derives the semi-parametric efficiency bound for the related class of propensity score weighted estimators of average treatment effects and Hirano, Imbens and Ridder (2003) and Wang, Linton, and Härdle (2004) propose estimators, which attain this bound.

[8] Semi- and non-parametric estimators of binomial choice models are discussed in Horowitz (1993). Hahn (1998) and Hirano, Imbens and Ridder (2003) employ non-parametric first step estimators of the propensity score, Wang, Linton and Härdle (2004) discuss semi-parametric estimation of the propensity score.

[9] In this framework, the estimator of the conditional response probabilities usually cannot be interpreted as a ML estimator unless selection is independent across time. The score of a ML estimator, which allows for correlation over time may become computationally unattractive. Bertschek and Lechner (1998) and Inkmann (2000) propose computationally less demanding alternative specifications of unconditional moment functions for binary choice panel data models. Robins, Rotnitzky, and Zhao (1995) propose conditioning on lagged response in multi-period panel data applications. In this case the conditional selection probability for period t factorises into a product of lagged probabilities of the form $\Pr(A_t = 0\mid A_{t-1} = 0, W)\Pr(A_{t-1} = 0\mid A_{t-2} = 0, W) \ldots \Pr(A_0 = 0\mid W)$ where W consists of time-invariant attributes. A similar specification will be used in Section 3 to account for sequential selection.

Recently, a number of different moment based estimators have been proposed as an alternative to GMM estimation. These include the Empirical Likelihood (EL) estimator of Owen (1988), Qin and Lawless (1994) and Imbens (1997), the Exponential Tilting (ET) estimator of Kitamura and Stutzer (1997) and Imbens, Spady and Johnson (1998) and the Continuous Updating (CU) estimator of Hansen, Heaton, Yaron (1996).[10] It has been noted by Smith (1997), Brown, Newey and May (1998), Imbens, Spady and Johnson (1998) and Newey and Smith (2004) that all these estimators belong to the same family of estimators meanwhile known as Generalised Empirical Likelihood (GEL) estimators. Like GMM, these estimators require a set of orthogonality conditions $E[\psi(Z,\theta_0)]=0$ as the only substantial prerequisite for the consistent estimation of θ_0. The reason for considering alternatives to GMM is the sometimes unsatisfying small sample performance of the semi-parametric efficient GMM estimator for the given set of unconditional moment conditions, which employs an inefficient first step estimator to construct an optimal weight matrix used in the second estimation step. These small sample shortcomings become increasingly serious with an increasing number of overidentifying restrictions $r-q$ (see Inkmann (2001), ch. 7, for a review of the extensive literature on small sample properties of GMM). Newey and Smith (2004) show that the GEL estimators share the first order asymptotic properties of the efficient GMM estimator, while improving upon the higher order asymptotic bias properties of GMM. The authors show that the reasons are twofold: in contrast to GMM, GEL does not require a first step weight matrix estimator and GEL implies a semi-parametric efficient estimator of the first derivative of the population moments with respect to the unknown parameter vector, which eliminates the correlation between the Jacobian and the moment functions, which both appear in the first order conditions of the optimisation programme. Since the dimension of the Jacobian matrix increases with an increasing number of moment conditions, GEL is expected to perform better than GMM if $r-q$ is large.[11]

The GEL estimator is defined by solving the information-theoretic optimisation problem (see Corcoran (1998), Newey and Smith (2004))

$$\hat{\theta}=\underset{\theta\in\Theta,\pi_1\dots,\pi_n}{\arg\min}\sum_{i=1}^{n}h_\tau(\pi_i) \qquad \text{subject to} \qquad \sum_{i=1}^{n}\pi_i=1 \tag{10}$$

$$\sum_{i=1}^{n}\pi_i\psi(Z_i\theta)=0 \tag{11}$$

where $h_\tau(\pi)=\left(\tau^2+\tau\right)^{-1}\left((n\pi)^{\tau+1}-1\right)/n$ is a member of the family of power divergence statistics indexed by τ (see Cressie and Read (1984)) between the scalar probability π_i and the empirical distribution weights $1/n$. The interpretation of the GEL optimisation program is very intuitive: simultaneously chose θ and π_i, for i = 1,...,n, such that the weighted sample counterpart (11) of the population orthogonality condition is satisfied and the probabilities[12] π_i are as similar as possible to the empirical distribution according to some information criterion. Note that

---

[10] See Inkmann (2001), ch. 7, for an introduction to these estimators.

[11] Note, however, that the argument refers to a higher order asymptotic result for GEL while the GMM shortcomings are known to exist in small samples.

[12] Interpreting the weights π as probabilities requires that they are non-negative. This is satisfied for small $\lambda'\psi(Z,\theta)$ in equations (12) and (13) below as pointed out by Newey and Smith (2004). All weights estimated for the applied part of this paper fall into the interval (0,1).

(11) is solved regardless of the presence of over-identifying restrictions. The EL (ET) and CU estimators result from (10) and (11) with $\tau = 0$ ($\tau = -1$) and $\tau = 1$, respectively, using limit representations for $\tau = 0$ ($\tau = -1$) to obtain $h_\tau(\pi) = -\ln\pi$ ($h_\tau(\pi) = \pi\ln\pi$).

Newey and Smith (2004) show that the solution to the program (10) and (11) is identical to the solution to the following saddle point problem in θ and the rx1 parameter vector λ

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{\lambda\in\Gamma(\theta)} \sum_{i=1}^{n} \rho\left(\lambda'\psi\left(Z_i,\theta\right)\right) \tag{12}$$

where $\rho(\nu) = -(1+\tau\nu)^{1+1/\tau}/(1+\tau)$ is a function in ν which is concave over its domain V and $\Gamma(\theta) = \{\lambda : \lambda'\psi(Z_i,\theta) \in V, i = 1,\ldots,n\}$. The EL (ET) and CU estimators result from (12) from choosing $\rho(\nu) = \ln(1+\nu)$ $(\rho(\nu) = -\exp(\nu))$ and $\rho(\nu) = -(1+\nu)^2$. Since the dimension of the optimisation problem is reduced from n + q to r + q, solving (12) in applied work should be computationally more attractive than solving (10) and (11), in particular using the penalty function optimisation approach suggested by Imbens, Spady and Johnson (1998). Newey and Smith (2004) derive an explicit solution for the estimators $\hat{\pi}_i$ of the probabilities π_i, i = 1,...,n, in (10) and (11) in terms of the estimators $\hat{\theta}$ and $\hat{\lambda}$ satisfying (12), which can be written as (using the notation $\nabla\rho(\nu) = \partial\rho/\partial\nu$)

$$\hat{\pi}_i = \frac{\nabla\rho\left(\hat{\lambda}'\psi\left(Z_i,\hat{\theta}\right)\right)}{\sum_{j=1}^{n}\nabla\rho\left(\hat{\lambda}'\psi\left(Z_j,\hat{\theta}\right)\right)}. \tag{13}$$

The CU estimator is a computationally particularly attractive special case of (12) and (13) because an explicit solution for the optimal $\hat{\lambda}$ in (12) is available

$$\hat{\lambda} = -\hat{V}^{-1}\overline{\psi} \qquad \text{with} \qquad \hat{V} = \frac{1}{n}\sum_{i=1}^{n}\psi\left(Z_i,\hat{\theta}\right)\psi\left(Z_i,\hat{\theta}\right)' \qquad \text{and} \qquad \overline{\psi} = \frac{1}{n}\sum_{i=1}^{n}\psi\left(Z_i,\hat{\theta}\right), \tag{14}$$

which simplifies the optimisation problem (12) from a saddle point problem to a more standard minimisation problem for $\hat{\theta}$ and implies the following probability estimators from (13)

$$\hat{\pi}_i = \frac{1-\overline{\psi}'\hat{V}^{-1}\psi\left(Z_i,\hat{\theta}\right)}{n\left(1-\overline{\psi}'\hat{V}^{-1}\overline{\psi}\right)} \tag{15}$$

for i = 1,...,n, which were formerly derived by Back and Brown (1993). Brown and Newey (1998) show that an estimator of any expectation, which is based on a weighted sample equivalent using probabilities (15), is efficient. Brown, Newey and May (1998) extend this result to any estimator based on probabilities (13).

The IPW-GEL estimator results from using the particular moment functions (9). Evaluated at the GEL estimator, (11) collapses to equation (3) in the introduction if the estimators $\hat{p}_i$, for i = 1,...,n, of the conditional selection probabilities are obtained in a first estimation step. We are aware of two references discussing IPW estimators in relation the GEL: Hirano, Imbens and Ridder (2003) interpret an IPW estimator that is based on an estimated propensity score as an empirical likelihood estimator, which efficiently incorporates the information about the propensity score. Nevo (2002) derives a weighted estimator, which is identical to an ET estimator if based on a representative sample and to an IPW moment estimator if based on a selected sample. Previous work discussing an estimator like (3) and its extension using the moment functions (9) considered in this section, which makes use of both estimated conditional selection probabilities $\hat{p}_i$ and estimated empirical distribution weights $\hat{\pi}_i$ is not known to us. This IPW-GEL estimator is applied in the remaining part of the paper to estimate the firm size elasticity of product and process R&D and compared to the IPW-GMM estimator using (9) proposed by Abowd, Crepon and Kramarz (2001). We will use the CU version of GEL with weights (15) for this purpose because the saddle point problem simplifies to a minimisation problem in this case and work by Donald, Imbens and Newey (2002) indicates that CU has higher order asymptotic efficiency advantages over competing GEL estimators.

3. Application: Firm Size and R&D Expenditures

Studying the relationship between firm size and R&D input and output variables has a long tradition in theoretical and empirical industrial organisation, which dates back to the claim, usually attributed to Schumpeter, that large firms have advantages over small firms in conducting successful research and development (see Cohen and Klepper, 1996b for an overview of the literature). This literature was enriched in the 1990s by distinguishing separate types of R&D, namely process R&D, which is targeted to cost-reducing process innovations, and product R&D, which is conducted with the goal to introduce product innovations that may lead to temporary monopoly gains until imitation occurs. Klepper (1996) shows in a theoretical model that different types of R&D may be predominantly associated with different stages of a product life-cycle and correspondingly, with firms of a certain age and size. The model implies that early and new entrants compete in heterogeneous product innovations in an emerging product market but engage increasingly in cost-reducing process R&D with increasing size to create a cost advantage that prevents possible competitors from entry. Since at the same time unsuccessful innovators are forced to leave the market, a mature market will involve a small number of producers with stable market shares. This model as well as the related models of Cohen and Klepper (1996a, b) depend on two main assumptions: product innovations are heterogeneous, which leads to firms producing variants of the same product at the same time, and most importantly, that the returns to process innovations are proportional to output and size since process innovations can be spread over a larger number of applications. These assumptions are responsible for the implication that the share of process R&D in the sum of process and product R&D increases with increasing firm size, which we will call the Cohen and Klepper hypothesis in short. The empirical analysis of Scherer (1991), which precedes the theoretical contributions of Cohen and Klepper, provides support for this hypothesis on the basis of US data.

A simple way of testing the Cohen and Klepper hypothesis was suggested by Fritsch and Meschede (2001) who estimate the firm size elasticity of process and product R&D. A firm size elasticity of process R&D exceeding the corresponding elasticity of product R&D would confirm the Cohen and Klepper hypothesis. The elasticities can be readily estimated from a linear model of log R&D expenditures in log firm size as measured by the number of employees. This empirical approach is adopted in the following. Fritsch and Meschede report the results from separate OLS estimations of the firm size elasticity of process and product R&D using a cross section of 627 German manufacturing firms (on enterprise level). They show that R&D increases less than proportional with firm size and that the firm size elasticity of process R&D (0.706) slightly exceeds the corresponding elasticity of product R&D (0.689). Fritsch and Meschede refer to a Wald test of the equality of both elasticities, which does not reject the null.[13] Hence, the results neither provide support for the Cohen and Klepper hypothesis nor reverse its content. Firm size has the same impact on both process and product R&D, a result which is in line with findings by Arvanitis (1997) for Swiss firms. We enrich the specification of Fritsch and Meschede by a) treating process and product R&D equations as seemingly unrelated regression, b) adding a random effects error term structure which demands for panel data, c) accounting for possible sample selection bias. The latter extension is the one of main interest for an application of the IPW-GEL estimator, course.

Separate information on R&D expenditures related to process and product innovations is available in the first two waves of the Mannheim Innovation Panel (MIP, 1993-1994), which is collected by the Centre for European Economic Research (ZEW) in Mannheim (see Harhoff and Licht (1994) for details on this data source, which serves as Germany's contribution to the European Community Innovation Surveys, CIS). We define the population of interest as consisting of firms, which engage in R&D activities in 1993 because the Cohen and Klepper theory does not apply to firms, which do not attempt to innovate.[14] The sample includes 1063 firms from this population observed in the first wave with full information.

The source of sample selection becomes obvious from investigating these firms in the second wave. Figure 1 provides an overview of the 1994 sample decomposition. More than 50% ((180+374)/1063) of the original 1993 sample is missing in 1994 because of (unit) non-response or attrition. Among the respondents, 60 firms were not engaged in R&D activities in 1994, which leaves a number of 449 firms with full information on R&D and firm size. Figure 1 also sheds some light on the decomposition of non-respondents: following their traces throughout the waves 1995 to 1997, it turns out that 180 non-responding firms return in later waves while 374 firms do not come back. Without further information, it cannot be ruled out that the latter group includes true market exits and not only panel exits.[15]

Figure 1 also defines the indicator variable A used before in Sections 1 and 2. Only firms with A = 0 are observed with full information while firms with A = 1 do not perform R&D and firms with A = 2 or A = 3 do not respond. Thus, there are three selection mechanisms denoted Attrition, R&D and Exit in the figure. They are related to each other in a hierarchical structure

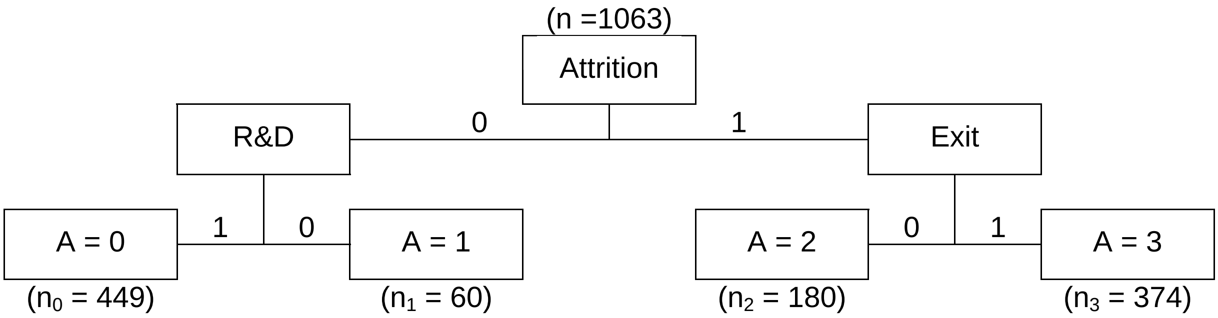
[13] It remains somewhat unclear, however, how the test statistic was computed since separate OLS regressions were performed to estimate the size elasticity of process and product R&D. Most likely, the Wald test was based on a diagonal variance-covariance matrix of the two elasticity estimators, which may be misleading. Given the magnitude of the parameter and standard deviation estimates in the present case, however, it is very likely that equality cannot be rejected.

[14] More precisely, these firms pass the so called innovation filter asking for realised and expected innovative activities in the three years preceding the interview and the three years following the interview, respectively.

[15] See Abowd, Crepon and Kramarz (2001) for another attempt to gain information on non-responding firms.

since R&D conditions on Attrition = 0 and Exit conditions on Attrition = 1. Thus, the denominator in (8) factorises into $\Pr(A = 0 \mid W) = \Pr(A = 0 \mid A = 0 \vee A = 1 \mid W)\Pr(A = 0 \vee A = 1 \mid W)$.

Figure 1: Second Period Sample Decomposition



Notes: Numbers in parentheses refer to (sub-) sample sizes.

We consider a number of alternative identification assumptions based on the discussion in Section 1. As a reference, we assume in Model A that selection occurs independent of the variables of interest, i.e. $\Pr(A = 0 \mid Z) = \Pr(A = 0)$. The independence assumption is dropped in Model D in favour of conditional independence, i.e. $\Pr(A = 0 \mid Z, W) = \Pr(A = 0 \mid W)$. In addition, we consider two specifications in between the pure IA and CIA models, which exploit the sequential structure of selection depicted in Figure 1: Model B assumes that attrition is independent of Z and the decision to engage in R&D is independent on Z given W, Model C reverses Model B's order of CIA and IA with respect to the two selection stages. In both cases, one of the right hand side probabilities in $\Pr(A = 0 \mid W) = \Pr(A = 0 \mid A = 0 \vee A = 1 \mid W)$ $\Pr(A = 0 \vee A = 1 \mid W)$ can be extracted from the expectation (8) without affecting the orthogonality condition. The idea behind Models B and C is to gain more insight into the relevance of IA versus CIA and to provide a kind of sensitivity analysis for the two more extreme models, A and B. The Appendix contains a proof that the specification employed by Models B and C identifies the moment conditions of interest along the lines of the result (8). Table 1 summarises the identification assumptions employed in Models A-D.

Table 1: Model Specifications

Model	IA/ CIA	Selection Probabilities
A	IA	$\Pr(A = 0 \mid A = 0 \vee A = 1, Z) = \Pr(A = 0 \mid A = 0 \vee A = 1)$
	IA	$\Pr(A = 0 \vee A = 1 \mid Z) = \Pr(A = 0 \vee A = 1)$
B	CIA	$\Pr(A = 0 \mid A = 0 \vee A = 1, Z, W) = \Pr(A = 0 \mid A = 0 \vee A = 1, W)$
	IA	$\Pr(A = 0 \vee A = 1 \mid Z) = \Pr(A = 0 \vee A = 1)$
C	IA	$\Pr(A = 0 \mid A = 0 \vee A = 1, Z) = \Pr(A = 0 \mid A = 0 \vee A = 1)$
	CIA	$\Pr(A = 0 \vee A = 1 \mid Z, W) = \Pr(A = 0 \vee A = 1 \mid W)$
D	CIA	$\Pr(A = 0 \mid A = 0 \vee A = 1, Z, W) = \Pr(A = 0 \mid A = 0 \vee A = 1, W)$
	CIA	$\Pr(A = 0 \vee A = 1 \mid Z, W) = \Pr(A = 0 \vee A = 1 \mid W)$

Notes: IA: independence assumption, CIA: conditional independence assumption. Indicator A defined in Figure 1.

Apart from the moment functions $\psi_1(Z_1,\beta)$ representing the seemingly unrelated regressions (SUR) of (log) product and process R&D in the 1994 wave of the panel $\left(Y_{94}^{prod}, Y_{94}^{proc}\right)$ and the moment functions $\psi_2(Z_2,\gamma)$ representing the score of the (relevant for the respective model of interest from Table 1) conditional selection probability in terms of the 1994 indicator A and the 1993 attributes W, the set of moment functions contains $\psi_0(Z_1,\beta)$ representing the SUR of product and process R&D in the 1993 wave of the panel $\left(Y_{93}^{prod}, Y_{93}^{proc}\right)$. We employ a log linear model for the R&D equations and a logit model with $p(W'\gamma) = 1/(1+\exp(-W'\gamma))$ for the selection equation, which results in the over-identified system of moment functions

$$\psi(Z,\theta) = \begin{pmatrix} \psi_0(Z_1,\beta) \\ \psi_1(Z_1,\beta)1(A=0) \end{pmatrix} \text{ for Model A}$$

$$\text{with } \psi_0(Z_1,\beta) = \begin{pmatrix} X_{93}\left(Y_{93}^{prod} - X_{93}'\beta^{prod}\right) \\ X_{93}\left(Y_{93}^{proc} - X_{93}'\beta^{proc}\right) \end{pmatrix}, \ \psi_1(Z_1,\beta) = \begin{pmatrix} X_{94}\left(Y_{94}^{prod} - X_{94}'\beta^{prod}\right) \\ X_{94}\left(Y_{94}^{proc} - X_{94}'\beta^{proc}\right) \end{pmatrix} \text{ and}$$

$$\psi(Z,\theta) = \begin{pmatrix} \psi_0(Z_1,\beta) \\ \psi_1(Z_1,\beta)\dfrac{1(A=0)}{p(W'\gamma)} \\ \psi_2(Z_2,\gamma) \end{pmatrix} \text{ with } \psi_2(Z_2,\gamma) = \begin{cases} [W(1(A=0) - p(W'\gamma))]1(A=0 \vee A=1) & \text{Model B} \\ W(1(A=0 \vee A=1) - p(W'\gamma)) \quad \text{for} & \text{Model C} \\ W(1(A=0) - p(W'\gamma)) & \text{Model D.} \end{cases}$$

The parameter and data vectors of main interest are defined as $\beta' = \left(\beta'^{prod}, \beta'^{proc}\right)$ and $Z_1' = \left(Y_{93}^{prod}, Y_{93}^{proc}, X_{93}', Y_{94}^{prod}, Y_{94}^{proc}, X_{94}'\right)$. Note that the distinction between A = 2 and A = 3 in Figure 1 is not exploited in any specification. The reason for differentiating between these temporary and permanent non-response categories is purely motivated by the economic theory outlined at the beginning of this section: since we cannot rule out that A = 3 firms are no longer operating in their market, and the product life-cycle theory predicts that exit occurs because of unsuccessful innovation attempts, it seems natural to exclude that non-response is independent of the residuals of the product and process R&D equations. In this case, Models C and D would be appropriate if we are able to find the relevant attributes W.

Table 2: Descriptive Statistics for R&D and Firm Size Variables

1993 (n = 1063)				1994 (n_0 = 449)			
Product R&D	Process R&D	Firm size	R&D dept.	Product R&D	Process R&D	Firm size	R&D dept.
-0.9133 (2.271)	-1.8733 (1.918)	5.4381 (1.634)	0.5513 (0.498)	-0.7494 (2.219)	-1.7798 (1.859)	5.3476 (1.641)	0.5969 (0.491)

Notes: Means with standard deviations in parentheses. R&D and firm size in logs, R&D department is binary.

Descriptive statistics for the variables entering the R&D equations are given in Table 2. Comparing the 1993 sample to the 1994 sub-sample with A = 0 it turns out that firms in 1994 tend to be slightly smaller and invest slightly less in both product and process R&D. The share of firms with a dedicated R&D department, however, increases. The latter variable will be included into the R&D equations to account for abnormal R&D expenditures of high tech companies. Apart from this variable we keep the R&D equations as simple as possible. We experimented with ten industry dummies and with a dummy indicating a firm's location in the

Eastern part of Germany (the area of the former GDR). Both extensions were clearly rejected in all specifications by the D-test of Newey and West (1987).[16]

The East German indicator is included in the set of attributes. Remember that attributes have to be observed for both respondents and non-respondents in the second wave. Thus, we are left with time-invariant information and variables observed in the first wave. Given our argument that non-response may be explained in part by market exits, which are caused by unsuccessful innovation attempts, a lagged R&D expenditures (sum of process and product R&D) variable and the R&D department indicator, are obvious candidates for W. Since the product life-cycle theory predicts that this argument is weakened once firm size is controlled for, 1993 firm size is included as well in W. In addition, we consider a dummy variable with value one, if the firm reports in 1993 that it is constrained in its innovation activities by the lack of qualified staff as an attribute. The variable might capture a trivial cause for attrition, namely that non-response simply occurs because of a high opportunity cost of filling out the questionnaire. It also should affect R&D efforts, although is turns out insignificant, when included among the regressors of the R&D equations, regardless of the specification being used. Finally, we try to approximate demand expectations by a variable with value one if a firm expects in 1993 a serious decline in demand, which should influence both R&D efforts and the willingness to participate in surveys. Table A1 in the Appendix displays summary statistics of the conditioning variables W for all relevant sub-samples defined by the outcome of the selection indicator A. Comparing responding (Attrition = 0 in Figure 1) and non-responding firms (Attrition = 1) it turns out that the latter are more likely to be located in the Western part of Germany, constrained by the lack of qualified staff and affected by negative demand expectations. Both groups of companies are of similar average size and have very similar R&D characteristics. However, these variables help to differentiate between A = 0 and A = 1 companies (conditional on Attrition = 0) and between A = 2 and A = 3 firms (conditional on Attrition = 1). Responding firms, which no longer engage in R&D efforts in 1994, are much smaller than the responding R&D performing companies, have much smaller R&D expenditures in 1993 and are much less likely to have a specialized R&D department. Similarly, permanent non-respondents spent much less on innovative activities in 1993 than temporary non-respondents. Overall, the descriptive statistics certainly support the reasons outlined above for considering these variables as attributes. The estimation results presented in the next section will show if these variables are sufficiently informative to obtain substantially different results under IA and CIA identification.

4. Estimation Results

We focus on the estimated firm size elasticity of process and product R&D since these variables are relevant for a test of the Cohen and Klepper hypothesis, which suggests that the firm size elasticity of process R&D should exceed the corresponding elasticity of product R&D. Table 3 presents estimation results for Models A-D in combination with the CU version of the IPW-GEL estimator and the IPW-GMM estimator.

[16] Inclusion of the latter dummy variable even led to a rejection of the over-identifying restrictions.

Table 3: Estimated Firm Size Elasticity of Product and Process R&D Expenditures

	GEL				GMM			
Model	A	B	C	D	A	B	C	D
Product R&D	0.7691 (20.48)	0.7683 (20.52)	0.7763 (20.89)	0.7711 (20.80)	0.7674 (20.42)	0.7682 (20.52)	0.7759 (20.87)	0.7704 (20.77)
Process R&D	0.6773 (18.13)	0.6770 (18.21)	0.6824 (18.29)	0.6795 (18.31)	0.6757 (18.08)	0.6769 (18.20)	0.6824 (18.29)	0.6794 (18.32)
$\chi^2(1)$ (equality)	7.6064 (0.006)	7.6238 (0.006)	7.8776 (0.005)	7.6391 (0.006)	7.5764 (0.006)	7.6375 (0.006)	7.8305 (0.005)	7.5308 (0.006)

Notes: $\chi^2(1)$ refers to the Wald test statistic for the null hypothesis of equal firms size elasticity of process and product related R&D. Numbers in parentheses are t-values except for the Wald test row where the numbers refer to p-values. Complete estimation results are given in Table A2 in the Appendix. Model specifications A-D are defined in Table 1.

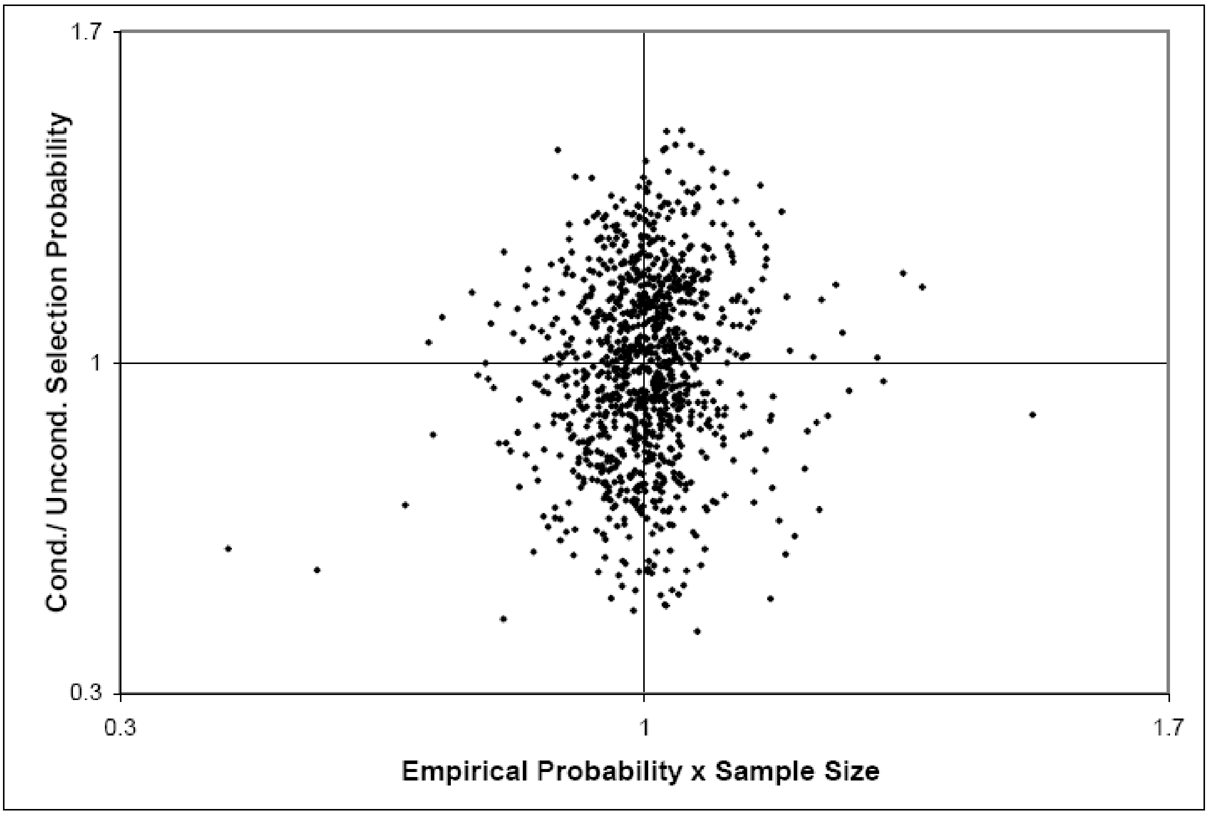
Without going into the details for the moment, it becomes immediately obvious from Table 3 that the firm size elasticity of product R&D (around 0.77) exceeds the firm size elasticity of process R&D (around 0.68), which serves as clear evidence against the Cohen and Klepper hypothesis in this sample of German manufacturing firms. A Wald test of the hypothesis of an equal elasticity rejects the null in all specifications at the 1% level as obvious from the last row of the table. We also note that R&D expenditures increase less than proportional with an increase in firm size contrary to the findings of many studies overviewed by Cohen and Klepper (1996b). The average estimate of the firm size elasticity of process R&D is very similar to the cross-section estimate of 0.706 obtained by Fritsch and Meschede (2001) for German companies using another data source. The average estimate of the firm size elasticity of product R&D, however, significantly exceeds the Fritsch and Meschede estimate of 0.689.

The striking result from Table 3 is certainly the stability of the parameter estimates across model specifications A-D and estimators GEL/ GMM. A comparison of the GEL elasticity estimates under pure IA identification (A) with those obtained under pure CIA identification (D) reveals completely negligible differences in the magnitude of 0.002. The difference of the corresponding GMM estimates is slightly larger but still below 0.004. The results of the mixed IA/ CIA models B and C are again very similar. The only notable difference between IPW-GEL and IPW-GMM concerns the transition from Model A to Model B: weakening IA in Model A in favour of CIA in Model B with respect to the decision to engage in R&D leads to a slight decrease of the GEL parameter estimates, but to a slight increase of the GMM estimates. A substitution of IA in Model A with CIA in Model C with respect to the attrition stage in Figure 1 slightly increases the elasticity estimates for both GEL and GMM.

Full estimation results are presented in Table A2 in the Appendix. The last row of this table contains p-values from a Hansen-Sargan test of the over-identifying restrictions, which emerge from imposing the usual panel data parameter restriction, that β is the same in the moment functions $\psi_0(Z_1,\beta)$ for the 1993 wave and in the moment functions $\psi_1(Z_1,\beta)$ for the 1994 wave. This restriction is not rejected regardless of the model specification and the estimator being used. It should be emphasised again that this test is not a test of the identification conditions underlying the different specifications. Table A2 also reveals that all parameters of the R&D equations remain largely unaffected from different specifications and estimators, not only the firm size elasticity. Overall, the CU version of GEL is slightly more efficient than GMM as suggested by Donald, Imbens and Newey (2002).

The parameter estimates of the logit selection equation do, of course, vary between specifications B-D because they refer to different selection events: $\Pr(A = 0 \mid A = 0 \vee A = 1, W)$ in Model B, $\Pr(A = 0 \vee A = 1 \mid W)$ in Model C and $\Pr(A = 0 \mid W)$ in Model D. It turns out that lagged information on the R&D department, R&D expenditures, an East German location, the dummy variable indicating lack of qualified staff and the dummy variable indicating negative demand expectations serve as significant predictors for the probability to engage in R&D given response, $\Pr(A = 0 \mid A = 0 \vee A = 1, W)$, in Model B. The R&D variables have a positive sign, the staff missing and demand expectation indicators a negative sign as one would expect. Less obvious, East German firms are more likely to engage in R&D. Predicting the response probability, $\Pr(A = 0 \vee A = 1 \mid W)$, in Model C seems more difficult. Smaller firms are significantly less likely to respond than larger firms, while the impact of R&D expenditures is only positive significant at the 10% level and the demand indicator negative significant at the same level. With the exception of the R&D department indicator, all variables turn out significant at the 5% level with the aforementioned signs in the prediction of the joint conditional probability of response and continued engagement in R&D, $\Pr(A = 0 \mid W)$, in Model D.

Figure 2: Estimated Empirical Distribution and Selection Probabilities (GEL/ Model D)



Finally, we would like to investigate the joint distribution of the estimates $\hat{\pi}_i$ in (15) and $\hat{p}_i$ for i = 1,...,n in order to get more insight into the way the IPW-GEL estimator works in this application and solves, e.g., equation (3) in the introduction. Figure 2 depicts the joint distribution for the pure CIA specification in Model D. The weights $\hat{\pi}_i$ were multiplied with the sample size 1063 and the conditional selection probabilities $\hat{p}_i$ were divided by the unconditional selection probability 449/1063. Thus, an observation with both probabilities equal to one is average with respect to its conditional selection propensity and receives an empirical distribution weight of 1/n. Figure 2 suggests a random distribution of firms around this (1,1) point, although the coefficient of correlation in the magnitude of 0.11 reveals that firms with a high conditional selection probability tend to receive a slightly higher weight. The variation in the vertical (selection probability) dimension is larger than the variation in the horizontal (empirical probability) dimension with the exception of a few companies, who receive either a very low or very high empirical weight. The two extreme observations which are closest to the left and right hand side borders of the chart (say, company L and company R) receive weights of 0.44 and 1.55 times 1/n. Company L is a respondent with R&D activities (A = 0) while company R is a permanent non-respondent (A = 3). Although company L belongs to the A = 0 group, its predicted selection probability is only 0.61 times the unconditional selection probability 449/1063 while the non-responding company R achieves 0.89 times 449/1063. Looking at the characteristics of these firms, Company L (a relatively large firm) is exceptional in reducing its process R&D expenditures from well above average to well below average from 1993 to 1994 while company R (a relatively small firm) is exceptional in reporting a zero to all dummy variables in W.

5. Conclusion

The paper introduces the IPW-GEL estimator of the parameters of a vector of possibly nonlinear unconditional moment functions from a selected sample under the assumption that selection occurs conditionally independent of the variables of interest. The estimator has been motivated by an attempt to estimate the firm size elasticity of product and process R&D from a sample of German manufacturing firms that is affected from both attrition and selection into R&D activities. Underlying this application is a test of the Cohen and Klepper hypothesis that process R&D should increase with increasing firm size to a larger extent than product R&D. The IPW-GEL estimator is applied and compared to an IPW-GMM estimator. The estimates turn out to be very similar with very slight efficiency advantages to IPW-GEL. The Cohen and Klepper hypothesis is clearly rejected in all specifications of the two estimators.

A comparison of the IA and CIA identification assumptions and mixtures of these assumptions is conducted. Despite an obvious availability of good predictors of the respective selection probabilities, IA and CIA outcomes are hardly distinguishable. One may ask if this is good news or bad news for the outcome of the empirical analysis. On the one hand, one might argue that the result should be interpreted as good news because selection does not seem to affect the variables of interest in any notable magnitude. On the other hand, one might argue that we were not able to identify the relevant attributes W, which would be bad news. Remember that the task of these variables is not prediction of the selection indicator but providing conditioning information, which achieves independence between selection and the variables of interest. Unfortunately, the identification assumptions are not subject to a conventional hypothesis test, which renders it difficult to decide on the success of the empirical specifications. However, since all attributes were carefully chosen, we certainly have gained trust in the magnitude of the parameter estimates and the rejection of the Cohen and Klepper hypothesis.

Further trust may be gained in future work by considering a more general semi- or non-parametric specification of the conditional selection probability in order to rule out a possible source of misspecification. Using replacement samples as suggested by Hirano, Imbens, Ridder and Rubin (2001) to test the identification assumptions employed in this paper is another worthwhile task for future research. Finally, it should be interesting to apply the asymptotic bias corrections derived by Newey and Smith (2004) for GEL and GMM estimators.

Appendix: Proofs

Moment identification in Model B:

Assumptions:
(CIA): $\Pr(A = 0 \mid A = 0 \vee A = 1, Z, W) = \Pr(A = 0 \mid A = 0 \vee A = 1, W)$
(IA): $\Pr(A = 0 \vee A = 1 \mid Z) = \Pr(A = 0 \vee A = 1)$

$$E\left[\psi(Z_1,\beta_0)\frac{1(A=0)}{\Pr(A=0 \mid A=0 \vee A=1, W)}\right]$$

$$= E\left[E\left[\psi_1(Z_1,\beta_0)\frac{1(A=0)}{\Pr(A=0 \mid A=0 \vee A=1, W)} \mid W\right]\right]$$

$$= E\left[E\left[\frac{\psi_1(Z_1,\beta_0)}{\Pr(A=0 \mid A=0 \vee A=1, W)} \mid W, A=0\right]\Pr(A=0 \mid W)\right]$$

$$= E\left[E\left[\frac{\psi_1(Z_1,\beta_0)}{\Pr(A=0 \mid A=0 \vee A=1, W)} \mid W\right]\Pr(A=0 \mid W)\right] \qquad \text{(CIA)}$$

$$= E\left[E\left[\frac{\psi_1(Z_1,\beta_0)}{\Pr(A=0 \mid A=0 \vee A=1, W)} \mid W\right]\Pr(A=0 \mid A=0 \vee A=1, W)\Pr(A=0 \vee A=1 \mid W)\right]$$

$$= E\left[E\left[\psi_1(Z_1,\beta_0) \mid W\right]\Pr(A=0 \vee A=1 \mid W)\right]$$

$$= E\left[E\left[\psi_1(Z_1,\beta_0) \mid W, A=0 \vee A=1\right]\Pr(A=0 \vee A=1 \mid W)\right] \qquad \text{(IA)}$$

$$= E\left[E\left[\psi_1(Z_1,\beta_0)1(A=0 \vee A=1) \mid W\right]\right]$$

$$= E\left[\psi_1(Z_1,\beta_0)1(A=0 \vee A=1)\right]$$

$$= E\left[\psi_1(Z_1,\beta_0)\right] = 0 \qquad \text{(IA)}$$

Moment identification in Model C:

Assumptions:
(IA): $\Pr(A = 0 \mid A = 0 \vee A = 1, Z) = \Pr(A = 0 \mid A = 0 \vee A = 1)$
(CIA): $\Pr(A = 0 \vee A = 1 \mid Z, W) = \Pr(A = 0 \vee A = 1 \mid W)$

$$E\left[\psi(Z_1,\beta_0)\frac{1(A=0)}{\Pr(A=0 \vee A=1 \mid W)}\right]$$

$$= E\left[E\left[\psi_1(Z_1,\beta_0)\frac{1(A=0)}{\Pr(A=0 \vee A=1 \mid W)} \mid W\right]\right]$$

$$= E\left[E\left[\psi_1(Z_1,\beta_0)\frac{1(A=0 \mid A=0 \vee A=1)}{\Pr(A=0 \vee A=1 \mid W)} \mid W, A=0 \vee A=1\right]\Pr(A=0 \vee A=1 \mid W)\right]$$

$$= E\left[E\left[\psi_1(Z_1,\beta_0)\frac{1(A=0 \mid A=0 \vee A=1)}{\Pr(A=0 \vee A=1 \mid W)} \mid W\right]\Pr(A=0 \vee A=1 \mid W)\right] \qquad \text{(CIA)}$$

$$= E\left[E\left[\psi_1(Z_1,\beta_0)1(A=0 \mid A=0 \vee A=1) \mid W\right]\right]$$

$$= E\left[\psi_1(Z_1,\beta_0)1(A=0 \mid A=0 \vee A=1)\right]$$

$$= E\left[\psi_1(Z_1,\beta_0)\right] = 0 \qquad \text{(IA)}$$

Appendix: Tables

Table A1: Descriptive Statistics for 1993 Conditioning Variables W

A ∈	0	1	0 1	0 1 2 3	2 3	2	3
R&D expenditures	-0.1439 (2.204)	-1.5153 (2.162)	-0.3055 (2.241)	-0.2988 (2.177)	-0.2925 (2.118)	-0.0141 (2.118)	-0.4265 (2.012)
Firm size	5.4301 (1.672)	4.7686 (1.640)	5.3522 (1.680)	5.4381 (1.634)	5.5171 (1.588)	5.7246 (1.588)	5.4173 (1.590)
R&D department	0.5880 (0.493)	0.2667 (0.446)	0.5501 (0.498)	0.5513 (0.498)	0.5523 (0.498)	0.5944 (0.498)	0.5321 (0.500)
East German firm	0.3341 (0.472)	0.2000 (0.403)	0.3183 (0.466)	0.2935 (0.456)	0.2708 (0.445)	0.2111 (0.445)	0.2995 (0.459)
Qualified staff missing	0.2004 (0.401)	0.4000 (0.494)	0.2240 (0.417)	0.2399 (0.427)	0.2545 (0.436)	0.2722 (0.436)	0.2460 (0.431)
Demand decreasing	0.2227 (0.417)	0.3500 (0.481)	0.2377 (0.426)	0.2690 (0.444)	0.2978 (0.458)	0.2889 (0.458)	0.3021 (0.460)
Sub-sample size	449	60	509	1063	554	180	374

Notes: Means with standard deviations in parentheses. R&D and firm size in logs, all other variables are binary.

Table A2: GEL and GMM Estimation Results (n = 1063)

	GEL				GMM			
Product	A	B	C	D	A	B	C	D
Intercept	-5.81 (-32.3)	-5.81 (-32.5)	-5.85 (-32.9)	-5.84 (-32.9)	-5.80 (-32.2)	-5.81 (-32.5)	-5.85 (-32.9)	-5.83 (-33.0)
Firm size	0.77 (20.5)	0.77 (20.5)	0.78 (20.9)	0.77 (20.8)	0.77 (20.4)	0.77 (20.5)	0.78 (20.9)	0.77 (20.8)
R&D dept.	1.52 (16.6)	1.51 (16.5)	1.51 (16.5)	1.50 (16.3)	1.52 (16.6)	1.51 (16.5)	1.51 (16.5)	1.51 (16.4)
Process	A	B	C	D	A	B	C	D
Intercept	-5.87 (-33.0)	-5.87 (-33.1)	-5.90 (-33.2)	-5.89 (-33.3)	-5.86 (-32.9)	-5.87 (-33.1)	-5.90 (-33.2)	-5.89 (-33.3)
Firm size	0.68 (18.1)	0.68 (18.2)	0.68 (18.3)	0.68 (18.3)	0.68 (18.1)	0.68 (18.2)	0.68 (18.3)	0.68 (18.3)
R&D dept.	0.64 (7.97)	0.63 (7.81)	0.64 (7.86)	0.63 (7.73)	0.64 (7.94)	0.63 (7.80)	0.64 (7.85)	0.63 (7.74)
Probability	A	B	C	D	A	B	C	D
Intercept	-	1.90 (2.63)	0.68 (1.93)	0.36 (1.00)	-	1.94 (2.66)	0.68 (1.93)	0.35 (0.96)
Firm size	-	0.01 (0.11)	-0.12 (-2.14)	-0.12 (-2.06)	-	0.01 (0.06)	-0.12 (-2.13)	-0.12 (-2.01)
R&D dept.	-	0.83 (2.51)	-0.04 (-0.29)	0.13 (0.88)	-	0.83 (2.48)	-0.04 (-0.27)	0.13 (0.89)
R&D exp.	-	0.27 (2.56)	0.08 (1.69)	0.13 (2.56)	-	0.27 (2.56)	0.08 (1.70)	0.13 (2.53)
East	-	1.01 (2.80)	0.16 (1.13)	0.35 (2.38)	-	1.00 (2.75)	0.16 (1.11)	0.35 (2.35)
Staff miss.	-	-0.79 (-2.69)	-0.15 (-0.99)	-0.32 (-2.12)	-	-0.78 (-2.65)	-0.14 (-0.98)	-0.31 (-2.08)
Demand -	-	-0.45 (-1.48)	-0.24 (-1.65)	-0.33 (-2.21)	-	-0.45 (-1.47)	-0.23 (-1.61)	-0.32 (-2.16)
J-test	A	B	C	D	A	B	C	D
$\chi^2(6)$	6.78 (0.34)	6.22 (0.40)	6.00 (0.42)	6.63 (0.36)	6.49 (0.37)	6.23 (0.40)	6.03 (0.42)	6.67 (0.35)

Notes: Numbers in parentheses are t-values except for the J-test row where the numbers refer to p-values. The J-test is the Hansen-Sargan Test of over-identifying restrictions. For full length variable names confer to Table A1. Model specifications A-D are defined in Table 1. The “product” part of the table refers the product R&D equation, the “process” part to the process R&D equation and the “probability” part to the conditional selection probability.

References

Abowd, J. M., B. Crepon, and F. Kramarz (2001): "Moment Estimation with Attrition: An Application to Economic Models", *Journal of the American Statistical Association*, 96 (456), 1223-1231.

Angrist, J. D. (1997): "Conditional Independence in Sample Selection Models", *Economics Letters*, 54, 103-112.

Arvanitis, S. (1997): "The Impact of Firm Size on Innovative Activity – An Empirical Analysis Based on Swiss Firm Data", *Small Business Economics*, 9, 473-490.

Back, K., and D. P. Brown (1993): "Implied Probabilities in GMM Estimators", *Econometrica*, 61 (4), 971-975.

Bertschek, I., and M. Lechner (1998): "Convenient Estimators for the Panel Probit Model", *Journal of Econometrics*, 87, 329-371.

Brown, B. W., and W. K. Newey (1998): "Efficient Semiparametric Estimation of Expectations", *Econometrica*, 66 (2), 453-464.

Brown, B. W., W. K. Newey, and S. May (1998): "Efficient Bootstrapping for GMM", unpublished manuscript, Rice University and MIT.

Cohen, W. M., and S. Klepper (1996a): "Firm Size and the Nature of Innovation within Industries: The Case of Process and Product R&D", *Review of Economics and Statistics*, 78, 232-243.

Cohen, W. M., and S. Klepper (1996b): "A Reprise of Size and R&D", *The Economic Journal*, 106 (437),925-951.

Corcoran, S. A. (1998): "Bartlett Adjustment of Empirical Discrepancy Statistics", *Biometrika*, 85, 967-072.

Cressie, N., and T. R. C. Read (1994): "Multinomial Goodness-of-Fit Tests", *Journal of the Royal Statistical Society,* Series B, 46, 440-464.

Dawid, A. P. (1979): "Conditional Independence in Statistical Theory", *Journal of the Royal Statistical Society*, Series B, 41, 1-31.

Dehejia, R. H., and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94 (448), 1053-1062.

Donald, S. G., G. W. Imbens, and W. K. Newey (2002): "Choosing the Number of Instruments for GMM and GEL Estimators", working paper, Department of Economics, MIT.

Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998): "An Analysis of Sample Attrition in Panel Data", *Journal of Human Resources*, 23 (2), 251-299.

Fritsch, M., and M. Meschede (2001): "Product Innovation, Process Innovation, and Size", *Review of Industrial Organization,* 19 (3), 335-350.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, 66, 315-331.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50 (4), 1029-1054.

Hansen, L. P., J. Heaton, and A. Yaron (1996): "Finite-Sample Properties of Some Alternative GMM Estimators", *Journal of Business & Economic Statistics*, 14 (3), 262-280.

Harhoff, D., and G. Licht (1994): "Das Mannheimer Innovationspanel", in: U. Hochmuth and J. Wagner (eds.), *Firmenpanelstudien in Deutschland, konzeptionelle Überlegungen und empirische Analysen*, 255-284, Tübinger Volkswirtschaftliche Schriften, Tübingen.

Hausman, J. A., and D. A. Wise (1979): "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment", *Econometrica*, 47 (2), 455-473.

Heckman, J. J. (1979): “Sample Selection Bias as a Specification Error”, *Econometrica*, 47, 153-162.

Heckman; J. J. (2001): “Econometrics and Empirical Economics”, *Journal of Econometrics*, 100, 3-5.

Hirano, K., G. W. Imbens, G. Ridder (2003): “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score”, *Econometrica*, 71 (4), 1161-1189.

Hirano, K., G. W. Imbens, G. Ridder, and D. B. Rubin (2001): “Combining Panel Data Sets with Attrition and Refreshment Samples”, *Econometrica*, 69 (6), 1645-1659.

Holland, P. W. (1986): “Statistics and Causal Inference”, *Journal of the American Statistical Association*, 81, 945-970.

Horowitz, J. L. (1993): “Semiparametric and Nonparametric Estimation of Quantal Response Models”, in: G. S. Maddala, C. R. Rao, and H. D. Vinod (eds.), *Handbook of Statistics*, vol. 11, Amsterdam: North-Holland.

Horowitz, J. L., and C. F. Manski (1998): “Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputation”, *Journal of Econometrics*, 84, 37-58.

Horvitz, D., and D. Thompson (1952): “A Generalization of Sampling Without Replacement from a Finite Population”, *Journal of the American Statistical Association*, 47, 663-685.

Imbens, G. W. (1997): “One-Step Estimators for Over-Identified Generalized Method of Moments Models”, *Review of Economic Studies*, 64, 359-383.

Imbens, G. W., R. H. Spady, and P. Johnson (1998): “Information-Theoretic Approaches to Inference in Moment Condition Models”, *Econometrica*, 66 (2), 333-358.

Inkmann, J. (2000): “Misspecified Heteroskedasticity in the Panel Probit Model: A Small Sample Comparison of GMM and SML Estimators”, *Journal of Econometrics*, 97, 227-259.

Inkmann, J. (2001): *Conditional Moment Estimation of Nonlinear Equation Systems - With an Application to an Oligopoly Model of Cooperative R&D*, Lecture Notes in Economics and Mathematical Systems, Vol. 497, Heidelberg: Springer.

Kitamura, Y, and M. Stutzer (1997): “An Information-Theoretic Alternative to Generalized Method of Moments Estimation”, *Econometrica*, 65 (4), 861-874.

Klepper, S. (1996): “Entry, Exit, Growth, and Innovation over the Product Life Cycle”, *American Economic Review*, 86 (3), 562-583.

Lechner, M. (1995): *Ökonometrische Modelle für qualitative Paneldaten: Das Beispiel der selbständigen Erwerbstätigkeit in den neuen Bundesländern*, Frankfurt: Campus.

Little, R. J. A., and D. B. Rubin (1987): *Statistical Analysis with Missing Data*, New York: Wiley.

Manski, C. F. (1988): *Analog Estimation Methods in Econometrics*, Chapman & Hall, New York.

Nevo, A. (2002): “Sample Selection and Information-Theoretic Alternatives to GMM”, *Journal of Econometrics*, 107, 149-157.

Newey, W. K. (1984): “A Method of Moments Interpretation of Sequential Estimators”, *Economics Letters*, 14, 201-206.

Newey, W. K. (1990): “Semiparametric Efficiency Bounds”, *Journal of Applied Econometrics*, 5, 99-135.

Newey, W. K. (1994): “The Asymptotic Variance of Semiparametric Estimators”, *Econometrica*, 62 (6), 1349-1382.

Newey, W. K. (1999): “Consistency of Two-Step Sample Selection Estimators Despite Misspecification of Distribution”, *Economics Letters*, 63, 129-132.

Newey, W. K., and D. L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing", in: R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, Amsterdam: North-Holland, 2111-2245.

Newey, W. K., and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators", *Econometrica*, 72 (1), 219-255.

Newey, W. K., and K. D. West (1987): "Hypothesis Testing with Efficient Method of Moments Estimation", *International Economic Review*, 28, 777-787.

Owen, A. B. (1988): "Empirical Likelihood Ratio Confidence Intervals for a Single Functional", *Biometrika*, 75, 237-249.

Qin, J., and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations", *The Annals of Statistics*, 22 (1), 300-325.

Rendtel, U. (1995): *Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität*, Frankfurt: Campus.

Ridder, G. (1990): "Attrition in Multi-Wave Panel Data", in: J. Hartog, G. Ridder, J. Theeuwes (eds.): *Panel Data and Labor Market Studies*, Amsterdam: North-Holland, 45-68.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data", *Journal of the American Statistical Association*, 90 (429), 122-129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data", *Journal of the American Statistical Association*, 90 (429), 106-121.

Rubin, D. B. (1976): "Inference and Missing Data", *Biometrika*, 63, 581-592.

Rubin, D. B. (1977): "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, 66, 688-701.

Rubin, D. B. (1978): "Baysian Inference for Causal Effects: The Role of Randomization", *Annals of Statistics*, 6, 34-58.

Scherer, F. M. (1991): "Changing Perspectives on the Firm Size Problem"; in: Z. J. Acs and D. B. Audretsch (eds.), *R&D, Patents and Productivity*, Chicago: Chicago University Press, 417-461.

Schnell, R. (1997): *Nonresponse in Bevölkerungsumfragen*, Opladen: Leske + Budrich.

Smith, R. J. (1997): "Alternative Semiparametric Likelihood Approaches to Generalized Method of Moments Estimation", *Economic Journal*, 107, 503-519.

Verbeek, M., and T. Nijman (1996): "Incomplete Panels and Selection Bias", in: L. Mátyás and P. Sevestre (eds.), *The Econometrics of Panel Data*, 2nd edition, ch. 18, Dordrecht: Kluwer.

Wang, Q., O. Linton, and W. Härdle (2004): "Semiparametric Regression Analysis with Missing Response at Random", *Journal of the American Statistical Association*, 99 (466), 334-345.

Wooldridge, J. M. (2002): "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification", *Portuguese Economic Journal*, 1, 117-139.