# Essay Paper:
# Partial Identification in non-separable triangular models


Martial Toniotti

May 29, 2020

**Abstract**

We define an identification region and an estimator for the regression function $g(X, \varepsilon)$ of non-separable triangular models in the case where data is supposed to follow a general missing data pattern. We use fundamental concepts of partial identification, the generation of the unobservable random term $\varepsilon$ from within the model and assumptions on the instrumental variable to do so. General results are given and an empirical example is presented.

# 1 Introduction

In statistical inference, we make a lot of assumptions to be able to find efficient, convergent estimators or test statistics. One of the core assumptions is point identification. It means that we want to find the parameters of the distribution of interest in a single point, not having an interval. But it is not always justifiable to make the assumptions needed to point identify the parameters.

When we do sampling, we only analyze data from a sample of the population. We usually assume that our sample was randomly selected. Meaning each data point in the population had the same probability of being into the sample. The opposite way of seeing this is by saying each data point in the population had the same probability of not being in the sample: data points are missing at random. When we have incomplete data in our sample, we usually assume that each data point in the sample had the same probability of having missing information. Missing information is, once again, missing at random (MAR). But the literature shows several cases where this MAR assumption is not justifiable.

Marlin et al. (2012) analyzed it in the case of rating of online content ( i.e. songs), showing that in this case, it is not the case that ratings are missing at random, meaning the sample of ratings they get is not representative of the whole population, which can cause problems when doing inference or comparison between the different contents. Pedersen et al. (2017) have analyzed the validity of MAR in clinical epidemiological research, they have found that this assumption is not always justifiable and that a lot of previous results use this without clearly stating they use it.

After such results, we need to develop new tools to be able to do credible inferences. One way is to relax the assumptions needed for point identification. We need to examine how we can develop econometric modeling and inference when the parameter of interest is partially identified, meaning that only an interval is identified. Of course those results will not be as precise but might be more credible ( Law of decreasing credibility, Manski (2003) ). We will not be able to point identify parameters but will have identification intervals. This must be used to prevent errors in inference when we cannot,

under the sampling process, assume that data is missing at random. One way it is done in the literature is by using partial identification.

This is a significant and emerging field in econometrics because it helps to have more robust inference and help find some boundaries to problems that were previously disregarded as "impossible" to assess. Finding more restrictive theoretical boundaries in cases with missing data is something very important to be able to do inference in a lot of empirical cases.

For example in economics when assessing the effect of a welfare program where the population that uses the program is not random ( see Gundersen et al. (2017)) or some variable people will not admit ( i.e. academic fraud in Mavisakalyan and Meinecke (2016) or the wealth effect of some variables when data is not complete in Lusardi et al. (2016)). Those are some usage of partial identification in the literature, but this list is just a list of examples and is not even close to being exhaustive.

The fields in which partial identification is helpful are not limited to economics, some papers have found usage for it in epidemiology (i.e Agüero (2017)). Where the author tries to find a correlation between violence against women and the health of their offspring. In this case, data is not missing at random due to the relative cultural taboo on violence against women. Using this kind of analysis is indeed helpful in a lot of cases where we do not want to assume that data is MAR.

A lot has to be done in partial identification research due to the relatively new interest in the last years. This essay paper will start from the latest general research in non-separable triangular models (see D'Haultfœuille and Février (2015)).

Second part of this essay will focus on the identification and estimation of the regression function $g(X, \varepsilon)$. Of course without point identification. This means we will define an identification region for it and an estimator for this identification region. This will be done using concepts from Manski (2003) and generation of an unobservable random term from inside the model from Matzkin (2003).

Part three will focus on some assumption we can make on $g(X, \varepsilon)$ to be able to get analytical bounds and to do interpretation of the parameters. First assumption is that $g(.)$ is a polynomial function. It is motivated by a long tradition of such models since Gergonne (1815) due to their interpolation properties. Second assumption is that the function is a rational function. Usage of this assumption in a model is more recent ( see Billings and Zhu (1991)) but the interpolation properties of such models are way better.

Part four will focus on an empirical example to be able to see in a real case what the methods developed here are capable of and how to compute estimators. This example will be an analysis of statistics of the OECD about the number of practitioners in European Union countries and the percentage

of population holding a master's degree. It shows the usage of estimators developed in parts 2 and 3 in a case where there is a nonadditive random function.

## 2 Identification and estimation of $g(X, \varepsilon)$

the analysis focus on the partial identification and estimation of non-separable triangular models of form:

$$Y = g(X, \varepsilon)$$
$$X = h(Z, \eta)$$

Where Y is the outcome, X an endogenous variable, $\epsilon$, and $\eta$ are nonadditive random functions that represent heterogeneity, and Z is the instrument. We focus on the partial identification of the regression function $g(X, \varepsilon)$.
D'Haultfœuille and Février in their 2015 paper proved that a non-separable triangular model with discrete instrument had an identified regression function under some assumptions on the orbits of the support of the explanatory variables. Let's state it.

**Theorem 2.1**(from D'Haultfœuille and Février (2015) ): Suppose strong exogeneity of Z, dual strict monotonicity of $g(.)$ and $h(.)$, and $\varepsilon$ following an uniform distribution and some regularity conditions . Then g is identified if $\forall (x, x') \in \mathcal{X}, \exists x_1, ..., x_j \ s.t. \bar{\mathcal{O}}_{x_j} \cap \bar{\mathcal{O}}_{x_{j+1}} \neq \emptyset, \forall j \in 0, ..., J$ where $x_0 = x \ and \ x_{J+1} = x'$
The proof of this theorem and some more definitions of the concepts used in it are available in the mathematical appendix.
We try to achieve partial identification of such models here.
First, using the same model, but removing the missing-at-random assumption that is implied to be able to define an identification region for the function $g(X, \varepsilon)$.
The variable $o_G$ is here to be read as the observability of the random variable G. o is a binary random variable that equal 1 when G is observed and 0 when G isn't. The o random variables are always observed. I use here a different notation than Mansky due to the fact the letter z is already taken by the instrumental variable. We stop assuming in this model a missing at random hypothesis, that is:

$$\begin{aligned} P(Y = y|X = x) &= P(Y = y|X = x, o_X = 1, o_Y = 1) \\ &= P(Y = y|X = x, o_X = 0, o_Y = 0) \\ &= P(Y = y|X = x, o_X = 0, o_Y = 1) \\ &= P(Y = y|X = x, o_X = 1, o_Y = 0) \end{aligned} \tag{2.1}$$

We stop to use this assumption here, meaning we have now 4 different types of observation in our population, for each different combination of value for $o_X$ and $o_Y$. We keep here assuming the instrumental variable Z is always observed and missing at random.
After defining an identification region, we estimate it.

## 2.1 Identification region

**Definition and bounds of $\mathbf{H}[P(Y = y|X = x)]$**

This section is highly influenced by the general method used in Horowitz and Manski (2000) , Horowitz and Manski (1998) and Manski (2003). We can use in this case what is called a general missing data pattern, taking into accounts the 4 different types of observations, as defined in equation (2.1).
First, let us define an identification region, that is the set of all possible value of a parameter we cannot reject given empirical evidences alone, for $P(Y = y|X = x)$ with a general missing data pattern.

**Proposition 2.1**(3.8 in Manski (2003) ): Let $P_{ij} = P(o_X = i, o_Y = j)$ for $i, j = \{0, 1\}$. Then:

$$
\begin{aligned}
&\mathbf{H}[P(Y = y|X = x)] = \\
&\{[P(Y = y|X = x, o_X = 1, o_Y = 1)P(X = x|o_X = 1, o_Y = 1)P_{11} \\
&+ \eta_{10}P(X = x|o_X = 1, o_Y = 0)P_{10} + \eta_{00}p_0P_{00} + \eta_{01}p_1P_{01}] \\
&\div [P(X = x|o_X = 1, o_Y = 1)P_{11} + P(X = x|o_X = 1, o_Y = 0)P_{10} \\
&+ p_0P_{00} + p_1P_{01}] \\
&; \eta_{10}, \eta_{00} \in \Gamma_Y, p_0, p_1 \in [0, 1], \eta_{01} \in \Gamma_Y(p_1)\}
\end{aligned}
\tag{2.2}
$$

Where $\mathbf{H}[.]$ is the notation for the identification region of a parameter, $P_{jk}$ is $P(o_X = j, o_Y = k)$ , $\eta_{jk}$ is an element of the set of all possible values of $P(Y = y|X = x, o_X = j, o_Y = k)$ ( this notation is only used when we cannot compute an exact value but we have a set of possible value). $\Gamma_Y$ is the set of all possible values of $P(Y = y)$. $p_k$ is an element of the set of all possible values of $P(X = x|o_X = 0, o_Y = k)$.
**Proof:**
By law of total probability:

$$
\begin{aligned}
&P(Y = y|X = x) = \\
&\quad P(Y = y|X = x, o_X = 1, o_Y = 1)P(o_X = 1, o_Y = 1|X = x) \\
&+ P(Y = y|X = x, o_X = 1, o_Y = 0)P(o_X = 1, o_Y = 0|X = x) \\
&+ P(Y = y|X = x, o_X = 0, o_Y = 1)P(o_X = 0, o_Y = 1|X = x) \\
&+ P(Y = y|X = x, o_X = 0, o_Y = 0)P(o_X = 0, o_Y = 0|X = x)
\end{aligned}
$$

And, we know by Bayes rule that for $i, j = \{0, 1\}$:

$$
\begin{aligned}
&P(o_X = i, o_Y = j|X = x) = \\
&\qquad \frac{P(X = x|o_X = i, o_Y = j)P(o_X = i, o_Y = j)}{\sum_{i=\{0,1\}} \sum_{j=\{0,1\}} P(X = x|o_X = i, o_Y = j)P(o_X = i, o_Y = j)}
\end{aligned}
$$

We can replace by the right-hand side of each iteration of Bayes rule the values of $P(o_X = i, o_Y = j|X = x)$.
every term is revealed by the sampling process except $P(Y = y|X = x, o_X = 0, o_Y = 1)$, $P(X = x|o_X = 0, o_Y = i)$ and $P(Y = y|X = x, o_X = i, o_Y = 0)$ for $i = \{0, 1\}$.
We write all possible values of $P(X = x|o_X = 0, o_Y = i)$ as $p_i$ ranging from 0 to 1.
We write all possible values of $P(Y = y|X = x, o_X = i, o_Y = 0)$ as $\eta_{i0} \in \Gamma_Y$. Those can take any value in the set of possible values of $P(Y = y|X)$.
For $P(Y = y|X = x, o_X = 0, o_Y = 1)$, we write it as $\eta_{01}$. But unlike other $\eta$, the set of possible values for it can be smaller. By law of total probability:

$$\begin{aligned} P(Y = y|o_X = 0, o_Y = 1) = \\ P(Y = y|X = x, o_X = 0, o_Y = 1)P(X = x|o_X = 0, o_Y = 1) \\ + P(Y = y|X \neq x, o_X = 0, o_Y = 1)P(X \neq x|o_X = 0, o_Y = 1) \end{aligned}$$

And if we isolate $P(Y = y|X = x, o_X = 0, o_Y = 1)$ and label terms not revealed by sampling process as $P(Y = y|X \neq x, o_X = 0, o_Y = 0)$ as $\gamma \in \Gamma_Y$ and $P(X = x|o_X = 0, o_Y = 1)$ as $p_1$, meaning $P(X = x|o_X = 0, o_Y = 1)$ is $(1 - p_1)$. Then we get:

$$P(Y = y|X = x, o_X = 0, o_Y = 1) = \frac{P(Y = y|o_X = 0, o_Y = 1) - (1 - p_1)\gamma}{p_1}$$

Denote the set of all possible value for this as:

$$\Gamma_Y(p1) = \Gamma_Y \bigcap [(P(Y = y|o_X = 0, o_Y = 1) - (1 - p_1))\gamma \div p_1; \gamma \in \Gamma_Y]$$

Rewrite the identification region. □

This identification region is true without any further assumption. We however add an assumption now about the instrumental variable. We assume statistical independance between Z and Y and a sequential cut between X and Y, that is:

**Assumption 2.1**:
2.1.1.Statistical Independance (SI) of Z and Y:

$$P(Y = y|X = x, Z = z) = P(Y = y|X = x), \forall z \in \mathcal{Z} \tag{2.3}$$

Where $\mathcal{Z}$ is the set of all possible values of Z.
2.1.2. Sequential cut between X and Y:
in $F^{\sigma}_{YX} = F^{\sigma_1}_{Y|X}\ F^{\sigma_2}_{X}$, $\sigma_1 \times \sigma_2$ is full rank.

Using assumption 2.1 we can further shrink the identification region with the following proposition.

**Proposition 2.2**(similar to 3.4 of Manski (2003)):
Assume assumption 2.1, Z countable then:

$$\begin{aligned}
&\mathbf{H}_{SI}[P(Y=y|X=x)] = \\
&\textstyle\bigcap_{z\in Z}\{[P(Y=y|X=x,o_X=1,o_Y=1,) \\
&P(X=x|o_X=1,o_Y=1,Z=z)P_{11|Z=z} \\
&+\eta_{10}P(X=x|o_X=1,o_Y=0,Z=z)P_{10|Z=z} \\
&+\eta_{00}p_0P_{00|Z=z}+\eta_{01}p_1P_{01|Z=z}] \\
&\div[P(X=x|o_X=1,o_Y=1,Z=z)P_{11|Z=z} \\
&+P(X=x|o_X=1,o_Y=0,Z=z)P_{10|Z=z}+p_0P_{00|Z=z}+p_1P_{01|Z=z}] \\
&;\eta_{10},\eta_{00}\in\Gamma_Y,p_0,p_1\in[0,1],\eta_{01}\in\Gamma_Y(p_1)\}
\end{aligned} \tag{2.4}$$

**Proof:**
Define for each value of Z, using proposition 2.1, $\mathbf{H}[P(Y=y|X=x,Z=z)]$. Since Z is countable, we have a finite number of different identification regions. Since we have assumption 2.1, it means the "real" value of $P(Y=y|X=x)$ is the same for all Z. It means it is in the identification region, and that thus we can define:

$$\mathbf{H}_{SI}[P(Y=y|X=x)] = \textstyle\bigcap_{z\in Z}\mathbf{H}[P(Y=y|X=x,Z=z)]$$

□

Z is discrete in d'Haultfœuille model, we make it countable here. Also it can be interpreted as combining data from different sampling processes. Assumption 2.1.1 is falsifiable by the following corrollary:

**Corollary 2.2.1:** If $\mathbf{H}_{SI}[P(Y=y|X=x)]=\varnothing$ then Assumption 2.1.1 is not correct.
**Proof:** By proposition 2.2:

$$\mathbf{H}_{SI}[P(Y=y|X=x)]=\varnothing \leftrightarrow \textstyle\bigcap_{z\in Z}\mathbf{H}[P(Y=y|X=x,Z=z)=\varnothing$$

It means that $\nexists n\in\Gamma_Y$ s.t. $n=P(Y=y|X=x),\forall z\in\mathcal{Z}$. Meaning assumption 2.1.1 does not hold. □

We can check our assumption here, that does not mean that statistical independance is true, but if the test fails then statistical independance is false.
Since the identification region is continuous over a compact set of values of parameters $p_0,p_1,\eta_{10},\eta_{00},\eta_{01}$, it means we can maximize and minimize it and get single-valued suprema. We get bounds for it. Let's define them:

**Proposition 2.3:**
Let B be a non-empty and proper subset of $\mathcal{Y}$, the support of Y. Let Assumption 2.1 hold. Let Z be countable. Let $\Gamma_Y$ be a bounded, continuous set. Then the upper bound of $P(Y \in B|X = x)$ over $\mathbf{H}_{SI}[P(Y = y|X = x)]$ is

$$U_{SI,y\in B}(X = x) = \min_{z\in\mathcal{Z}}(U^*_{y\in B}(X = x, Z = z)) \tag{2.5}$$

where

$$U^*_{y\in B}(X = x, Z = z) = \max_{p_0,p_1,\eta_{10},\eta_{00},\eta_{01}} \mathbf{H}[P(Y \in B|X = x, Z = z)] \tag{2.6}$$

**Proof:**
Define $\mathbf{H}[P(Y \in B|X = x, Z = z)$ using proposition 2.1. It depends on parameter $p_0, p_1 \in [0, 1]$ ; $\eta_{10}, \eta_{00} \in \Gamma_Y$ and $\eta_{01} \in \Gamma_Y(p_1)$.
All those sets are closed, bounded and continuous. They are also subsets of $\mathbb{R}$. $\mathbf{H}[P(Y \in B|X = x, Z = z)$ is linear. Meaning we can define a maximum for each value of Z:

$$U^*_{y\in B}(X = x, Z = z) = \max_{p_0,p_1,\eta_{10},\eta_{00},\eta_{01}} \mathbf{H}[P(Y \in B|X = x, Z = z)$$

the answer to this maximization problem is trivial and is:

$$\begin{aligned}
U^*_{y\in B}(X = x, Z = z) = & \\
& \{[P(Y \in B|o_X = 1, o_Y = 1, X = x) \\
& P(X = x|o_X = 1, o_Y = 1, Z = z)P(o_X = 1, o_Y = 1|Z = z) \\
& + P(X = x|o_X = 1, o_Y = 0, Z = z)P(o_X = 1, o_Y = 0|Z = z) \\
& + P(o_X = 0, o_Y = 0|Z = z) \\
& + P(y \in B|o_X = 0, o_Y = 1)P(o_X = 0, o_Y = 1|Z = z)] \\
& \div [P(X = x|o_X = 1, o_Y = 1, Z = z)P(o_X = 1, o_Y = 1|Z = z) \\
& + P(X = x|o_X = 1, o_Y = 0, Z = z)P(o_X = 1, o_Y = 0|Z = z) \\
& + P(o_X = 1, o_Y = 1|Z = z) \\
& + P(y \in B|o_X = 0, o_Y = 1)P(o_X = 0, o_Y = 1|Z = z)]\}
\end{aligned}$$

Compute it for each value of Z. We have a countable number of upper bounds. By assumption 2.1, knowing the "real" value of $P(Y \in B|X = x)$ is possible for all values of Z, meaning that the upper bound of $P(Y \in B|X = x)$ is:

$$U_{SI,y\in B}(X = x) = \min_{z\in\mathcal{Z}}(U^*_{y\in B}(X = x, Z = z))$$

Since $\mathbf{H}[P(Y \in B|X = x, Z = z)$ is an intersection over continuous sets where at least one value is common to all the sets, the upper bound of the

intersection must be minimal upper bound of the continuous sets. Meaning that:

$$\max_{\mathbf{H}_{SI}[P(Y=y|X=x)]}(P(Y \in B|X=x)) = U_{SI,y\in B}(X=x)$$

□

**Corollary 2.3.1:**
Let assumptions of Proposition 2.3 hold. Then the lower bound of $P(Y \in B|X=x)$ over $\mathbf{H}_{SI}[P(Y=y|X=x)]$ is

$$L_{SI,y\in B}(X=x) = \max_{z\in\mathcal{Z}}(L^*_{y\in B}(X=x, Z=z)) \tag{2.7}$$

where

$$L^*_{y\in B}(X=x, Z=z) = \min_{p_0,p_1,\eta_{10},\eta_{00},\eta_{01}} \mathbf{H}[P(Y \in B|X=x, Z=z)] \tag{2.8}$$

Corollary 2.3.1 is proven conversely to Proposition 2.3.
We thus have upper and lower bounds for $P(Y \in B|X=x)$ for any interval B. We can plot everything we have defined on the $Y \times P(Y|X=x)$ space as done in Figure 1. B is supposed to be a proper subset of $\mathcal{Y}$. Such a graphical representation is used throughout the analysis to help understand what we are doing.

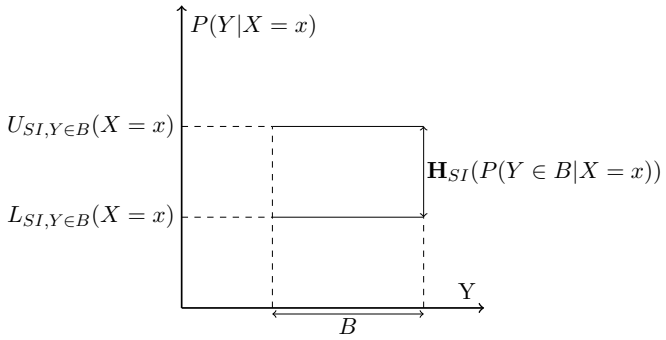



Figure 1: Graphical representation of previously defined concepts

We can choose B as small or as big as we want, we can thus define density-"like" curves if B tends to 0 and take the function of values of $U_{SI,y\in B}(X=x)$ and $L_{SI,y\in B}(X=x)$ for every proper subset B of $\mathcal{Y}$ that tends to zero. We will call them $U_{SI,\mathcal{Y}}(X=x)$ and $L_{SI,\mathcal{Y}}(X=x)$. They are to be seen similar to a classical density function ( as defined in Kolmogorov (1956) or Ord (1972) ). Each point has probability zero but we can work with integrals.

The integral of $U(.)$ over $\mathcal{Y}$ is superior to one and the integral of $L(.)$ over $\mathcal{Y}$ is inferior to one because both are bounds for each interval, not designed to be consistent density functions. As long as the number of intervals B is finite, the functions are functions evolving by discrete steps over $\mathcal{Y}$. For a number of intervals B sufficiently large, we can get a graph similar to figure 2.

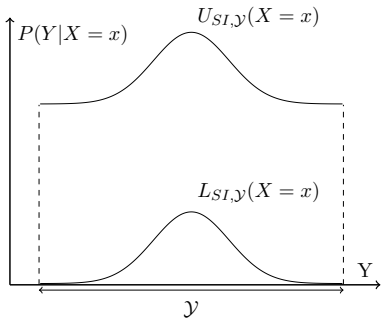



Figure 2: Graphical representation $U_{SI,\mathcal{Y}}(X=x)$ and $L_{SI,\mathcal{Y}}(X=x)$

The real density function of $P(Y|X=x)$ must be contained between those two functions. They are indeed bounds. Such bounds can be found for every value of X.

**Definition of H**$[g(X=x,\varepsilon)]$

In our paradigm, we consider $\varepsilon$ not as "what we cannot explain" defined after regression like in OLS and that must be additive, but as an idiosyncratic random variable that can intervene in $g(.)$ with any function. It is independent of both X and Z. It means that for any value of X, we cannot know in advance which value of $\varepsilon$ we get. This type of analysis comes from Matzkin (2003). To be able not to restrict our problem to additive $\varepsilon$ we define it differently.

There may be not an identified pair $(g(.),F_\varepsilon)$ from the data. That is what is called observational equivalence.

**Definition 2.1**(similar to its definition in Matzkin (2003)):
Any two functions $g(.),g'(.)$ are said to be observationally equivalent if $\exists F_\varepsilon, F'_\varepsilon$ such that $F_{Y,X}(g(.),F_\varepsilon)=F_{Y,X}(g'(.),F'_\varepsilon)$. That is if both give the same joint distribution of Y,X.

Here, since we work using partial identification, we have no interest in finding an unique pair $(g(.),F_\varepsilon)$, but we need to develop a criterion to see if a pair is not falsifiable without further assumptions.

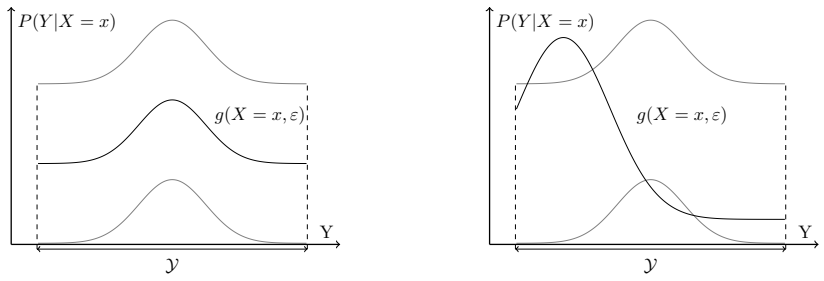
**Proposition 2.4**:
Assume conditions of Proposition 2.3. denote $\Gamma_\varepsilon(\subset \mathbb{R})$ the support of $\varepsilon$. Denote a bounded interval on $\Gamma_\varepsilon$, $E$ and its bound $\underline{E}, \bar{E}$. Denote a proper subset of $\mathcal{Y}$ defined by $E$, $B(X,E) = [g_1(X,\underline{E}); g_1(X,\bar{E})]$.
Then, The couple $(g_1(X,\varepsilon), F_\varepsilon)$ is not falsifiable without further assumptions $\leftrightarrow$

$$\forall x \in \mathcal{X}, \forall E \subseteq \Gamma_\varepsilon, L_{SI,y\in B(x,E)} \leq P(\varepsilon \in E) \leq U_{SI,y\in B(x,E)} \tag{2.9}$$

**Proof:**
$L_{SI,y\in B(x,E)}$ and $U_{SI,y\in B(x,E)}$ are defined by Proposition 2.3 and Corollary 2.3.1 and give us theoritical bounds for $P(y \in B(x,E)|X=x)$.
For any interval $E$, we can generate an interval in Y that represents all the value in the set $g_1(x, \varepsilon \in E)$. By definition, this is the interval $B(x,E)$.
If $g_1(X,\varepsilon)$ and $F_\varepsilon$ are actually how data is generated, then probabilty of Y being in interval $B(x,E)$ is what we would observe if $P(o_X) = P(o_Y) = 1$.
Here, we only have theoritical bounds on the possible values of $P(y \in B(x,E)|X=x)$. If they are correct, $g_1(X,\varepsilon)$ and $F_\varepsilon$ should give us a value of $P(y \in B(x,E)|X=x)$ between those theoritical bounds.

By construction of the interval $B(x,E)$, $P(y \in B(x,E)|X=x) = P(\varepsilon \in E)$. thus , $L_{SI,y\in B(x,E)} \leq P(\varepsilon \in E) \leq U_{SI,y\in B(x,E)}$ if the pair $(g_1(.), F_\varepsilon)$ is how data is indeed generated. Meaning that if it is not the case, then we can falsify it and say it is not how data is generated.

□




(a) $(g(.), F_\varepsilon)$ is not falsifiable (b) $(g(.), F_\varepsilon)$ is falsifiable

Figure 3: Graphical representation of proposition 2.4

Figure 3 shows graphically what it means to be falsifiable; it means there exist an interval where $P(Y|X=x)$ generated from $(g(.), F_\varepsilon)$ is higher than the upper bound or lower than the lower bound.

Proposition 2.4 can be useful in many ways. If we are, for some reasons, sure about the value of $g(.)$, we can test whether or not our $F_\varepsilon$ is correct or

simply use it as stated and test a pair. We can assume $F_\varepsilon$ and define from there an identification region for $g(X, \varepsilon)$.

**Assumption 2.2**: Assume $\varepsilon \sim F_\varepsilon$.

Sometimes researchers can make this assumption if they have good reasons to know how data is generated. From this assumption we can derive:

**Corollary 2.4.1**: Assume conditions of proposition 2.4. Assume assumption 2.2. Then

$$\mathbf{H}_{SI}[g(X,\varepsilon)] = [All\ g(X,\varepsilon)\ : \mathcal{X}, \Gamma_\varepsilon \to \mathcal{Y}\ s.t. \forall x \in X, \forall E \subseteq \Gamma_\varepsilon, L_{SI, y \in B(x,E)} \leq P(\varepsilon \in E) \leq U_{SI, y \in B(x,E)} where\ B(x,E) = [g_1(X = x, \underline{E}); g_1(X = x, \bar{E})]] \tag{2.10}$$

**Proof:** Define $\mathbf{H}_{SI}[g(X, \varepsilon)]$ as the set of all functions $g(X, \varepsilon)$ not falsifiable in the way defined by Theorem 2.4 given $F_\varepsilon$. □

We can also define a set of possible value of Y given X.

**Corollary 2.4.2**: Assume conditions of proposition 2.4. Assume assumption 2.2. Assume X=x. Then there exist a bounded set $[\underline{Y}, \bar{Y}]_{\varepsilon \in E} \in \mathcal{Y}$ for each interval E of $\Gamma_\varepsilon$ that is the bounded set of possible value of Y given X and $\varepsilon$.

Proof of this corollary is trivial, if it was not correct then no $g(.)$ would be possible. It is only a proof of existence, not a definition of such Y. The obvious assumption to hold to get a smaller identification region is to assume $\mathbf{E}[Y|X, Z] = g(X, \mathbf{E}(\varepsilon))$ or to assume the error term is additive, but we try for now to be as general as we can.

## 2.2 Estimation

### Estimation of $\mathbf{H}_{SI}[g(.)]$

From now on we assume we have n data points. First, let's define estimators for the lower and upper bounds on some intervals of B.

**Definition 2.2:**

$$\begin{aligned}\hat{U}_{n,IS,y\in B}(X=x,Z=z) = \{&[\hat{P}_n(Y\in B|o_X=1,o_Y=1,X\in N_n(x))\\ &\hat{P}_n(X\in N_n(x)|o_X=1,o_Y=1,Z=z)\hat{P}_n(o_X=1,o_Y=1|Z=z)\\ &+\hat{P}_n(X\in N_n(x)|o_X=1,o_Y=0,Z=z)\hat{P}_n(o_X=1,o_Y=0|Z=z)\\ &+\hat{P}_n(o_X=0,o_Y=0|Z=z)\\ &+\hat{P}_n(y\in B|o_X=0,o_Y=1)\hat{P}_n(o_X=0,o_Y=1|Z=z)]\\ &\div[\hat{P}_n(X\in N(x)|o_X=1,o_Y=1,Z=z)\hat{P}_n(o_X=1,o_Y=1|Z=z)\\ &+\hat{P}_n(X\in N_n(x)|o_X=1,o_Y=0,Z=z)\hat{P}_n(o_X=1,o_Y=0|Z=z)\\ &+\hat{P}_n(o_X=1,o_Y=1|Z=z)\\ &+\hat{P}_n(y\in B|o_X=0,o_X=1)\hat{P}_n(o_X=0,o_Y=1|Z=z)]\}\end{aligned} \tag{2.11}$$

where $\hat{P}_n$ of a probability is the sample equivalent of a probability and $N_n(x)$ is a neighborhood of x as small as we want defined by some $\delta \in \mathbb{R}_+$ : $N_n(x) = [x - \frac{\delta}{n}; x + \frac{\delta}{n}]$. We use here a neighborhood of x and not directly x for convergence reasons.

Let's state the Lehmann–Scheffé theorem.

**Theorem 2.2**(Lehmann-Scheffé Theorem, adapted for binomial distribution, from Lehmann and Scheffé (1950) and Lehmann and Scheffé (1955)): Assume k data points generated from a bernoulli distribution with probability of success p($\in [0,1]$). Define estimator of p, $\bar{p} = \frac{\sharp of success}{k}$. Then $\bar{p}$ is the unique uniformly minimum-variance unbiased estimator of p.
**Proof:**

$$\begin{aligned}\mathbb{E}(\bar{p}) &= \mathbb{E}(\frac{\sharp of success}{k})\\ &= \frac{p*k}{k}\\ &= p\end{aligned}$$

By Rao-Blackwell Theorem ( RAO (1945)) since the number of success over the total sample size is a sufficient statistic and $\bar{p}$ is an unbiased estimator of p, All other estimators have at least same variance.
But is it unique ? Another unbiased estimator with same variance $p'$. Then $\mathbb{E}(p' - \bar{p}) = 0$.
And since both are only defined by the parameter p (= both are complete family over p),

$$\begin{aligned}p' - \bar{p} &= 0\\ p' &= \bar{p}\end{aligned}$$

□

Remember that a sequence of Bernoulli processes is a binomial distribution. We've showed in theorem 2.2 the optimal estimator for any binomial distribution. We can continue and show convergence of our estimator:

**Proposition 2.5:** Assume n data points. Assume assumptions of proposition 2.3. Then $\hat{U}_{n,IS,y\in B}(X=x,Z=z)-U_{IS,y\in B}(X=x,Z=z) \xrightarrow{p} 0$ when $n\to\infty$.
**Proof:** In $\hat{U}_{n,IS,y\in B}(X=x,Z=z)$ expression, each term can be interpreted as a binomial distribution. Using Theorem 2.2 we know the sample equivalents, meaning the value of those probabilities in the sample, are the uniformly minimum-variance unbiased estimator the parameter of interest. It means that all terms that are sample equivalent in $\hat{U}(.)$ will converge to their equivalent in $U(.)$.
The only difference is for terms with N(x) in them in place of x.But: $N(x)=[x-\frac{\delta}{n};x+\frac{\delta}{n}]$ meaning that once n tends to infinity, $N(x)\to x$. We can then reuse Lehmann-Scheffé Theorem to show it is asymptotically unbiased and converge to the real value. It is not the minimum-variance estimator though. Overall it means that when $n\to\infty$ , all terms of the estimator $\hat{U}_{n,IS,y\in B}(X=x,Z=z)$ converges to their equivalent in $U_{IS,y\in B}(X=x,Z=z)$, meaning $\hat{U}_{n,IS,y\in B}(X=x,Z=z)-U_{IS,y\in B}(X=x,Z=z)\xrightarrow{p}0$. □

**Corollary 2.5.1:**
Assume assumptions of proposition 2.5.We can define $\hat{L}_{n,IS,y\in B}(X=x,Z=z)$ conversely to $\hat{U}_{n,IS,y\in B}(X=x,Z=z)$. Then $\hat{L}_{n,IS,y\in B}(X=x,Z=z)-L_{IS,y\in B}(X=x,Z=z)\xrightarrow{p}0$ when $n\to\infty$.
The Proof is an emulation of proof of proposition 2.5.

Why do we use $N(x)$ in the estimators and not x? It is due to properties of $\hat{P}_n(Y\in B|o_X=1,o_Y=1,X\in N_n(x))$ in finite samples. The sample equivalent of $P(Y\in B|o_X=1,o_Y=1,X=x))$, due to the fact that X is continuous, will always be either 0 or 1. It will never be another value. Even if it will asymptotically converge to the actual value of the parameter, in finite sample it can causes problems. That is why $N(x)$ is used here, because in a neighborhood of x, the value of $\hat{P}_n(Y\in B|o_X=1,o_Y=1,X\in N_n(x))$ can be anywhere on $\Gamma_Y$, the support of the measure. And it will also asymptotically converges to the real value of the parameter.

We need to define in a finite sample n all intervals with different values of the bounds. Let's do it.

**Lemma 2.1:**
Assume a finite sample n. Assume general missing data pattern. Assume

h ($0 < h < n$) data points of type $o_X = 0, o_Y = 1, Z = z$. Then for any complete observation where $X = x, Z = z, o_X = 1, o_Y = 1$, there exist a set $H_n(X = x, Z = z, h)$, the set having all the intervals where we have one interval for each possible pair $(\hat{U}(.), \hat{L}(.))$.

**Proof:**

We will work with a complete observation (X=x,Y=y, Z=z, o_X=,o_Y=1) and h uncomplete observations $(o_X = 0, o_Y = 1, Z = z, Y)$.

Sort the h+1 observations we have by their values of Y (observed for all of them). Name each observation by their place in this sequence. we have $1, 2, ..., c, c+1, ..., h, h+1$ . where c is the number in the sequence given to the complete observation. The values of the bounds will change only when a new observation is in the interval of Y we take.

Basic combinatorics tells us than in our situation we have (c-1)(h-c)+h intervals on $\mathcal{Y}$ with distinct bounds and c in them. In those intervals, the value of $\hat{U}(.)$ will of course be 1 (because all that we observe is in the interval, meaning the upper bound will of course be one). But the value of $\hat{L}(.)$ will change, meaning we will have $(c-1)(h-c)+h$ intervals with $\hat{U}(.) = 1$ and a different value of the lower bound. Let's call the set of all those intervals $H_L(X = x, Z = z, h)$. This set is always definable as long as h is non-null.

Now for all the sets without c, we get $(c-1)!+(h-c)!$ different intervals on $\mathcal{Y}$ with distinct bounds if $c < h$. In those intervals, the value of $\hat{L}(.)$ will of course be 0 (because we can assume none of those have the same value of X as our complete observation, meaning on this interval lower bound of Y being in it can be 0). But the value of $\hat{U}(.)$ will change. Let's call the set of all intervals here $H_U(X = x, Z = z, h)$. This set is always definable as long as h is non-null.

let's denote $H_n(X = x, Z = z, h) = H_L(X = x, Z = z, h) \bigcap H_U(X = x, Z = z, h)$. It is the set of intervals having for each of them a different pair $(\hat{U}(.), \hat{L}(.))$. and have $(c-1)(h-c)+h+(c-1)!+(h-c)!$ components when $c < h$. When $c \geq h$, we can still define $H_n$ but it will have a different number of intervals.

□

Of course, each pair is true over more than one interval, but over a continuous interval of values of Y ( between which there are no new data points). Let's define it more formally:

**Definition 2.3:**

for each interval $j \in H(X = x, Z = z, h)$, we can identify a pair $(L_j, U_j)$. This pair holds on an interval of values of Y. The smallest interval on which the pair of bound is the same is interval j as defined in lemma 2.1. Because it is a bounded interval, if we shrink it, we lose data points and the pair of bound change too.

The largest interval on which the pair of bounds is identified is an open interval. It is by definition the largest one with the same data points as

j and only those, we will call it $\bar{j}$. When we sort by values of Y the h+1 observations used in lemma 2.1, if j, interval on Y, is defined by bound (on j) observations f and f+1, then the interval $\bar{j}$ is the open interval $]f-1, f+2[$.

Let's define our estimator for the identification region of $g(.)$:
**Definition 2.4:**

$$\begin{aligned}\hat{\mathbf{H}}_{n,SI}[g(X,\varepsilon)] =& [All\ g(X,\varepsilon)\ : X, \Gamma_\varepsilon \to Y\ s.t. \forall X_i \in X \\ & from\ complete\ observations, \forall j \in H_n(X = x, Z = z, h), \\ & \hat{L}_{n,SI,y\in j} \leq P(\varepsilon \in E(j)) \leq P(\varepsilon \in E(\bar{j})) \leq \hat{U}_{n,SI,y\in j}]\end{aligned} \tag{2.12}$$

where $E(j) = [\underline{\varepsilon}, \bar{\varepsilon}]$, $E(\bar{j}) = [\underline{e}, \bar{e}]$ where $g(X = x, \varepsilon = \underline{\varepsilon}) = Y_f$, $g(X = x, \varepsilon = \bar{\varepsilon}) = Y_{f+1}$, $g(X = x, \varepsilon = \underline{e}) = Y_{f-1}$, $g(X = x, \varepsilon = \bar{e}) = Y_{f+2}$. We can define $E(\bar{j})$ as a closed interval and not an open one as it should be because $F_\varepsilon$ is a density function so the fact that the interval is open or closed does not matter.

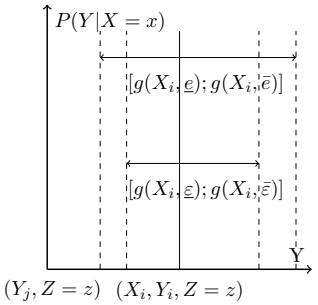



Figure 4: Graphical representation of values of interval used in definition of $\hat{\mathbf{H}}_{n,SI}[g(X,\varepsilon)]$

Let's prove asymptotic convergence of $\hat{\mathbf{H}}_{n,SI}[g(X,\varepsilon)]$ towards $\mathbf{H}_{SI}[g(X,\varepsilon)]$.

**Theorem 2.3:**(convergence of the estimator of the identification region)
Let assumptions of Lemma 2.1. Let assumptions of Corollary 2.4.1. Let $F_\varepsilon$ be a proper density function.Let $P(o_X = 0, o_Y = 1, Z = z)$ and $P(o_X = 1, o_Y = 1, Z = z)$ be non-null and independant of the size of the sample $\forall Z \in \mathcal{Z}$. Then

$$\hat{\mathbf{H}}_{n,SI}[g(X,\varepsilon)] - \mathbf{H}_{SI}[g(X,\varepsilon)] \xrightarrow{p} 0 \quad when \quad n \to \infty. \tag{2.13}$$

**Proof:**
The proof will be in three part. First, we will show that there exist $H_n(X = x, Z = z, h), \forall X \in \mathcal{X}$ and it tends to the set defined by $\forall j \subseteq \mathcal{Y}$ and thus $\forall E \subseteq \Gamma_\varepsilon$ if $F_\varepsilon$ is correct. The second one will show convergence of

$P(\varepsilon \in E(j))$ and$P(\varepsilon \in E(\bar{j}))$. Last part will show total convergence.
1.
As n tends to infinity, h, the number of observations in the sample of type $(o_X = 0, o_Y = 1, Z = z)$ will will tend also to infinity if assumption on $P(o_X = 0, o_Y = 1, Z = z)$ is respected. The same is true for n, the number of observations of type $o_X = 1, o_Y = 1, Z = z$. Meaning that eventually, we will get a complete observation for all possible values of X. We can thus define $H_n(X = x, Z = z, h)$ for all values of X eventually.
Also, since h tends to infinity with eventually values $\forall y \in \mathcal{Y}$, it means all closed intervals on $\mathcal{Y}$ will eventually be in $H_n(X = x, Z = z, h)$, which in turn, by definition of $E$, will be equivalent to all intervals on $\Gamma_\varepsilon$
2.
As h tends to infinity, the distance between the values of Y of the sorted set of h+1 observations (h observations of type $o_X = 0, o_Y = 1, Z = z$ and one observation of type $o_X = 1, o_Y = 1, Z = z, X = x$ ) will tend to zero. meaning $P(\varepsilon \in E(j))$ and $P(\varepsilon \in E(\bar{j}))$ tends to the same values.
3.
we showed in lemma 2.1 that estimators for the bounds are convergent on any set. We showed in point 2 that $P(\varepsilon \in E(j))$ and $P(\varepsilon \in E(\bar{j}))$ are assymptotically convergent. Meaning that for any set B we check on, $\hat{L}_{n,SI,y\in j} \leq P(\varepsilon \in E(j)) \leq P(\varepsilon \in E(\bar{j})) \leq \hat{U}_{n,SI,y\in j}$ will tend to $L_{SI,y\in B} \leq P(\varepsilon \in E) \leq U_{SI,y\in B}$ for some E.
Now, point one shows that $\bigcup_{x\in\mathcal{X}} H_n(X = x, Z = z, h)$ tends to all the intervals checked in the definition of $\mathbf{H}_{SI}[g(X, \varepsilon)]$. Meaning our criterion is assymptotically correct and we assymptotically check on all needed intervals. It means there exist a sample size $\hat{n}$ for which any function $g(.)$ that is not in $\mathbf{H}_{SI}[g(X, \varepsilon)]$ will be out of $\hat{\mathbf{H}}_{\hat{n}^{sup},SI}[g(X, \varepsilon)]$ for $\hat{n}^{sup} \geq \hat{n}$.There also exist a sample size $\hat{n}^2$ so that a function in $g(.)$ in $\mathbf{H}_{SI}[g(X, \varepsilon)]$ will be in $\hat{\mathbf{H}}_{\hat{n}^{2,sup},SI}[g(X, \varepsilon)]$ for $\hat{n}^{2,sup} \geq \hat{n}^2$. Those two conditions are the definition of convergence of a set toward another, meaning:
$\hat{\mathbf{H}}_{n,SI}[g(X, \varepsilon)] - \mathbf{H}_{SI}[g(X, \varepsilon)] \xrightarrow{p} 0 \quad when \quad n \to \infty$

□

This theorem is really important because now we have a definition to determine empirically in finite sample the identification region. We can further assume form of $g(.)$ to be able to have analytical bounds for its identification region. First let's take a look at the problem of estimation of $g(.)$ itself.

### Estimation of $g(X, \varepsilon)$

We can also, without even assuming $F_\varepsilon$ , compute a set of function g for which, for each value of X, the value of $g(X = x, \varepsilon)$ will be the smallest defined subset of Y on which $U_{SI,y\in B(x,E)}$ is at its maximum value. It is a maximum likelihood estimator. To be precise, the maximum of possible likelihood estimator.

Conversely, we can define a "maximum of minimum likelihood" estimator, using the maximum values of the lower bound of $P(y \in B|X = x)$. Those two are equivalent in the sense that they will give the same value of g(.). It is because lower bounds and upper bounds are just the same information translated, so the maximum does not change of place.
The problem with this approach is that it will give us the same value as if we assumed that data is missing-at-random in the sample because we don't model explicitly that what we don't observe could change the maximum of likelihood. It is therefore not interesting to continue in this way of maximum of likelihood estimators. The point is that wanting to point identify in a situation of partial observability will create an assumption of data missing at random no matter how we do it.

# 3 Analytical assumptions on $g(X, \varepsilon)$

Previous definition and estimation give a criterion to check if a function $g(X, \varepsilon)$ is possible or not, but does not give a way to define bounds. Having analytical bounds for the function would be convenient to be able to assess the effect of the covariate on Y. It is particularly interesting compared to the analysis of, for instance, $E(Y|X = x_1)$ and $E(Y|X = x_2)$ because we can interpret parameters.

## 3.1 Parametrization

We can decompose g(.) in:

$$g(X, \varepsilon) = g_1(g_2(X), \varepsilon) \tag{3.1}$$

It is not an assumption in that all functions g(.) are decomposable like that. We can remark $g_1(.)$ is not unique because multiple decomposition are possible, but a couple $(g_1, g_2)$ is unique.
**Assumption 3.1**: Assume we know $g_1(.)$.

It is a big assumption if, as we will do, we assume after a form on $g_2(.)$. Sometimes, if the data generating process is well known in its form, we can make assumption 3.1 . It needs anyway to be thoroughly justified.
Coupled with an assumption on the form of $g_2(.)$ and its parametrization (for example assume $g_1(X) = \beta^X$ or $g_1(X) = \beta X$), we can find an identification region for those parameters, which is very convenient when we want to interpret the parameters.
**Definition 3.1**: The identification region of a parameter $\beta$ is:

$$\mathbf{H}_{SI}[\beta] = \bigcap_{x \in \mathcal{X}} \mathbf{H}_{X=x,SI}[\beta] \tag{3.2}$$

Where $\mathbf{H}_{X=x,SI}[\beta]$ is the set of value of the parameter so that $g_1(g_{2,\beta}(X = x), \varepsilon) \in \mathbf{H}_{SI}[g(.)]$.

Since we have an asymptotically convergent estimator $\mathbf{H}_{SI}[g(.)]$, we can conversely define an asymptotically convergent estimator of $\mathbf{H}_{SI}[\beta]$:
**Definition 3.1**: An asymptotically convergent estimator for $\mathbf{H}_{SI}[\beta]$ is:

$$\hat{\mathbf{H}}_{n,SI}[\beta] = \bigcap_{x \in \chi} \hat{\mathbf{H}}_{n,X=x,SI}[\beta] \tag{3.3}$$

Where $\hat{\mathbf{H}}_{n,X=x,SI}[\beta]$ is the set of value of the parameter so that $g_1(g_{2,\beta}(X = x), \varepsilon) \in \hat{\mathbf{H}}_{n,SI}[g(.)]$ and $\chi$ is the set of all value of X from complete observation ( $o_X = o_Y = 1$) in the sample.

**Proposition 3.1:** Let assumptions of Theorem 2.3. Let assumption 3.1. Let some parametrization $g_2(.)$. Let $P(o_Y = 1, o_X = 1)$ be stable no matter the size of the sample. Then

$$\hat{\mathbf{H}}_{n,SI}[\beta] - \mathbf{H}_{SI}[\beta] \xrightarrow{p} 0 \quad when \quad n \to \infty \tag{3.4}$$

**Proof:**
as n tends to infinity, $\chi$ tends to $\mathcal{X}$. Meaning we take the intersect asymptotically on all the sets needed. The rest has already been proven to converge in Theorem 2.3. □
This method would work for any parametrization of $g_2(.)$. But which parametrization should be used when we don't know which to use ?

## 3.2 Polynomial function modeling

Let's make the following assumption on the form of $g_2(X)$:
**Assumption 3.2**:

$$g_2(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_k X^k \tag{3.5}$$

for some k and for some parameters $\beta_0, ..., \beta_k \in \mathbb{R}^k$.

This assumption is non-falsifiable of course. It can be a good assumption because curve fitting properties of those functions are well known and have been used for a long time ( since Gergonne (1815) ). Also, interpretation of the parameters is possible. But a big problem with polynomial functions is that they fit poorly oscillating functions. When we suspect the data generating process to be of this type, next section may be useful.

## 3.3 Rational function modeling

In the same fashion, we can make the following assumption on the form of $g_2(X)$:
**Assumption 3.3**:

$$g_2(X) = \frac{\beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_k X^k}{\beta_{k+1} + \beta_{k+2} X + \beta_{k+3} X^2 + ... + \beta_{2k+1} X^k} \tag{3.6}$$

for some k and for some parameters $\beta_0, ..., \beta_{2k+1} \in \mathbb{R}^{2k+1}$.
This assumption may be relevant when we need to fit more closely oscillating phenomenons. Interpolatory properties of rational functions are better. It also already has been used in econometric models for a long time (since Billings and Zhu (1991)). It may be very difficult to interpret the parameters due to the rational form, which makes it less popular and less useful in a wide range of cases.

Other forms are possible for $g_2(.)$ and should not be avoided when the researcher has strong reasons to think $g_2(.)$ is otherwise, but the previously presented assumptions help gather a wide range of phenomenon and are the ones used in most econometric models.

# 4 Empirical Example

This section will focus on the usage of the new tools defined before to estimate bounds on parameters in an empirical case. First Data will be presented, then the model will be specified. In a third section, the actual estimation will occur and then results will be presented. All computation will be commented on to be sure to present a comprehensive and understandable way of computing the estimators.

## 4.1 Data

We will work with data from OECD about the number of doctors per 1000 inhabitants ( OECD (2019)) and the percentage of the population with a master equivalent degree ( OECD (2016)). For this analysis, all data that is an estimation in the OECD report will be considered to be missing. Also, we will work with all countries in the European Union, but some are not affiliated with the OECD. We will interpret data for those countries to be missing too. From doctors per 1000 inhabitants, it is straightforward to find the percentage of population composed of doctors, it is what will be used here, to be in the same metric ( percentage of the total population) for the two data used.
A table with everything can be found in the appendix.

## 4.2 Modeling

We denote from here the percentage of doctor in population by Y and the percentage of master equivalent holder in population by X. No instrumental variables will be used here. To use notation from last sections, we will assume $g_1(.)$ of form $\varepsilon X\beta$, or to be more clear, that data is generated following:

$$Y = \varepsilon(X\beta) \tag{4.1}$$

Where X, Y are defined as before, $\varepsilon$ is an homogeneity parameter between countries. $\beta$ is a parameter common to all countries that will express the country-mean of population having a master having this master in medecine. we need to do some assumptions:

- We can always write $F^{\sigma}_{YX} = F^{\sigma_1}_{Y|X}\ F^{\sigma_2}_{X}$, but we assume furthermore here that there is a sequential cut between $F^{\sigma_1}_{Y|X}$ and $F^{\sigma_2}_{X}$, that is $\sigma_1 \times \sigma_2$ is full rank.
- We assume $\varepsilon \in [0, 2]$ following an euclidian density rising linearly between 0 and 1 and decreasing linearly between 1 and 2, like in figure 5.
- We assume here data is generated from a general missing data pattern.

- X is discrete and not continuous due to rounding to the percent in data. We will use it here but keep in mind it would be the same with a continuous X.

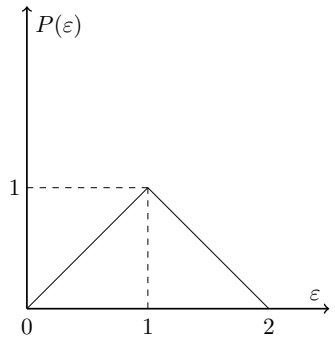



Figure 5: Density of $\varepsilon$

### 4.3 Estimation of bounds

We will try to find here the identification region of parameter $\beta$ according to our sample. First, we can compute some sample equivalent probabilities we will need during estimation. Since data is assumed to be generated from a general missing data pattern, we can always compute:

$$\hat{P}_n(o_X = 1, o_Y = 1) = \frac{13}{27}$$

$$\hat{P}_n(o_X = 1, o_Y = 0) = \frac{8}{27}$$

$$\hat{P}_n(o_X = 1, o_Y = 0) = \frac{1}{27}$$

$$\hat{P}_n(o_X = 0, o_Y = 0) = \frac{5}{27}$$

All those are different from 0, which mean that the assumption that data is generated from a general missing data pattern may be not a strong one.
Knowing that, we can start our estimation by defining, according to notation of Lemma 2.1, the set of all intervals B on $\mathcal{Y}$ (which here is $[0, 100]$ ), where, for the value of X of some complete observation (here we will do the complete observation of Austria), the values of the bounds of $P(Y \in B|X = x)$ are different. One will remember that it is all possible combination of all the observations of type $(o_X = 1, o_Y = 0)$ and of the complete observation. Here we have only one observation of type $(o_X = 1, o_Y = 0)$ (France) and the complete observation (Austria), meaning we have 3 different intervals:

$$H_n(X = 12) = \{[0.337; 0.512], [0.337; 0.512[, [0.512; 100]\} \tag{4.2}$$

There exist also other intervals but they are not informative ( the upper bound is equal to 1 and the lower bound equal to 0). We need also to keep in mind that here are one interval with the values of the bound, but there exist an infinity of them (all intervals on the value of Y where only the observations in each interval is). For each of those interval B, we will compute the upper and lower bound of $P(Y \in B|X = 12)$. All bounds defined here after are given X=12, but it is omitted to have lighter notation.

**[0.337;0.512]**

For this interval, since it contains the only complete observation where X=12, We can compute the upper bound but some logic will tell us that it is 1 since all observations are in it.

$$\hat{U}_{n,Y\in[0.337;0.512]} = 1$$

For the lower bound, we just need to apply the formula defined before. We don't use the estimator with the neighborhood of X but replace all instance of $N(x)$ by $X = x$ due to the fact X is here discrete, meaning the reasons why the neighborhood of x was used in place of $X = x$ makes no sense here.

$$\begin{aligned}\hat{L}_{n,Y\in[0.337;0.512]} = [&\hat{P}_n(Y \in [0.337; 0.512]|X = 12, o_X = 1, o_Y = 1)\\ &\hat{P}_n(X = 12|o_X = 1, o_Y = 1)\hat{P}_n(11)]\div\\ &[\hat{P}_n(X = 12|o_X = 1, o_Y = 1)\hat{P}_n(11)+\\ &\hat{P}_n(X = 12|o_X = 1, o_Y = 0)\hat{P}_n(10) + \hat{P}_n(00)+\\ &\hat{P}_n(Y \notin [0.337; 0.512]|o_X = 0, o_Y = 1)\hat{P}_n(01)]\end{aligned}$$

Where $\hat{P}_n(ij) = \hat{P}_n(o_X = i, o_Y = j)$. We compute from data the different value needed and get:

$$\begin{aligned}\hat{L}_{n,Y\in[0.337;0.512]} &= \frac{1 * 0.076923 * 0.48148}{0.076923 * 0.48148 + 0.25 * 0.29629 + 0.18518 + 0 * 0.037}\\ &= 0.125\end{aligned}$$

Remember the value of those bounds are correct for all sets with the same observations inside it. Meaning all sets from $Y \in [0.337; 0.512]$ to $Y \in [0, 100]$. We can thus write, in the fashion of proposition 2.4:

$$0.125 \leq P(12\varepsilon\hat{\beta} \in [0.377, 0.518]) \leq P(12\varepsilon\hat{\beta} \in [0, 100]) \leq 1 \tag{4.3}$$

$\hat{\beta}$ are used and not regular $\beta$ because we are here estimating. We will not get a point-identified estimation for $\beta$ of course since we are in partial identification.

**[0.337;0.512[**

For the second interval, since it does not contain the only complete observation we have, of course the lower bound possible is 0, we can compute it analytically but it will of course be the same value.

$$\hat{L}_{n,Y\in[0.337;0.512[} = 0$$

For the upper bound, we need to apply formula defined in definition 2.2. It is not copied here to keep the document readable.

$$\begin{aligned}\hat{U}_{n,Y\in[0.337;0.512[} &= \frac{0 + 0.25 * 0.29626 + 0.18518 + 0.037037}{0.076923 * 0.48148 + 0.25 * 0.29629 + 0.18518 + 0.037} \\ &= 0.88888\end{aligned}$$

We can thus write, defining the smallest set with those bounds $[0.377, 0.518[$ and the biggest one $[0, 0.518[$ :

$$0 \leq P(12\varepsilon\hat{\beta} \in [0.377, 0.518[) \leq P(12\varepsilon\hat{\beta} \in [0, 0.518[) \leq 0.88888 \tag{4.4}$$

**]0.337;0.512]**

Once again, since this interval does contain the only observation given X=12, we can say that the upper bound, obviously, will be one.

$$\hat{U}_{n,Y\in]0.337;0.512]} = 1$$

the lower bound will be:

$$\begin{aligned}\hat{L}_{n,Y\in]0.337;0.512]} &= \frac{1 * 0.076923 * 0.48148}{0.076923 * 0.48148 + 0.25 * 0.29629 + 0.18518 + 1 * 0.037} \\ &= 0.11111\end{aligned}$$

Meaning we can write:

$$0.11111 \leq P(12\varepsilon\hat{\beta} \in ]0.377, 0.518]) \leq P(12\varepsilon\hat{\beta} \in ]0.377, 100]) \leq 1 \tag{4.5}$$

## 4.4 Estimation of $\hat{\mathbf{H}}_n[\beta]$

We can write, using equation (4.3).

$$0.125 \leq P(12\varepsilon\hat{\beta} \in [0.377, 0.518]) \tag{4.6}$$

Define the identification region of parameter $\beta$.

$$\hat{\mathbf{H}}_n[\beta] = [\underline{\hat{\beta}}, \bar{\hat{\beta}}]$$

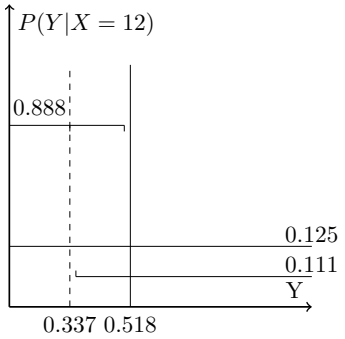



Figure 6: Bounds in the $Y \times P(Y|X=12)$ space

where

$$P(12\varepsilon\bar{\beta} \in [0.377, 0.518]) = 0.125$$

Let's focus on $\bar{\beta}$. We can define, following definition 2.4, $\underline{\varepsilon}$ and $\bar{\varepsilon}$ so that:

$$\int_{\underline{\varepsilon}}^{\bar{\varepsilon}} \varepsilon dF_\varepsilon = 0.125$$

where $12\underline{\varepsilon}\bar{\beta} = 0.337$ and $12\bar{\varepsilon}\bar{\beta} = 0.518$. Since we assumed $F_\varepsilon$, we get ($F_\varepsilon$ is discontinuous, we write here that the interval of $\varepsilon$ is strictly below one, it is not an assumption but just a way to write less useless computation given the solution is unique):

$$\int_{\underline{\varepsilon}}^{\bar{\varepsilon}} \varepsilon d\varepsilon = 0.125$$

After integration and replacement of both bounds on $\varepsilon$ by their expression in term of $\bar{\beta}$, we get the following second degree equation:

$$-\frac{5.373437504 * 10^{-4}}{\bar{\beta}^2} - 1.125 = 0$$

This equation has two solution, but one gives a negative value to $\bar{\beta}$, which of course since the interpretation we give of it is that it's the percentage of population having a master being a doctor, it is strictly positive. We get:

$$\bar{\beta} = 0.0655647$$

If $F_\varepsilon$ is correct, then we have at most 6.55 percent of the people having a master having it in medicine. To get the estimation region of $\beta$, we should check all conditions on all complete observations and take the intersection of those. If this intersection is null then $F_\varepsilon$ is wrong. We must keep in mind we only proved that our estimator is asymptotically correct, and here, with n=27, it is really a strong assumption to say it is correct. The point of this section is just to show how to compute bounds on parameters with real data.

# 5 Conclusion

This work define an estimator of the identification region of the regression function. This is really incomplete and is based on very strong assumptions that are not always justifiable. We managed to remove the Missing-at-random assumption at, mainly, the cost of adding an assumption on $F_\varepsilon$. It may not always worth it. It must be thoroughly justified, from other observations of the phenomenon of interest for example.
Everything here stay as general as possible. Another assumption is that, even if we use an instrumental variable, we also assume that we could work without at the price of a larger identification region, by assuming a sequential cut. Future research may be done in trying to define an IV estimator in cases where we cannot assume a sequential cut nor that data is missing-at-random. Future research may also want to find a more efficient way to do estimations, because when n ( and particularly h) go up, computation necessarly become factorially more complex. In large data sets it may be highly inefficient.
Also, we had to add an assumption after removing one, it shows how data itself say nothing, and that it is data and some assumptions about it that make any econometrics possible.

# Appendix A

# Appendix

## 1 D'Haultfœuille and Février theorem

Let's first define some concepts used in the proof and then let's do the actual proof.

### 1.1 Definitions

**Orbit of a point:**

$$\mathcal{O}_x = \{g \cdot x | g \in G\}$$

The orbit of a point is the set of element g(x) that are possible under g being in a group G.
Without lose of generality, it means, for example in the case of the group of inversible matrix and matrix multiplication:

$x' \in \mathcal{O}_x$ means $\exists s_{ij} = \begin{pmatrix} s_{i_1 j_1} & \dots & s_{i_p j_1} \\ \dots & & \dots \\ s_{i_1 j_p} & \dots & s_{i_p j_p} \end{pmatrix} s.t.\ x' = s_{ij} \cdot x$

There exists an inversible matrix of dimension [p x p] that can gives x' from x in a matrix multiplication. In this document, it is how $\mathcal{O}_x$ is used. In the same fashion, here, $\bar{\mathcal{O}_x}$ means here the closure of $\mathcal{O}_x$ in $\mathcal{X}$ , meaning every possible value of x ( $x \in \mathcal{X}$ ) , the matrices in $\bar{\mathcal{O}_x}$ produces value that are also in $\mathcal{X}$ .

**identification:**
We say something is identified if we can find it in an unique way from observable variables.
In a model of type Y=m(X, $\epsilon$), the pair ( m, $F_\epsilon$ ) is identified in the set of their support ( M , $\Gamma$) if two conditions are verified:

(i) $(m, \epsilon) \in (M, \Gamma)$

(ii) $[F_{YX}(m, F_\epsilon) = F_{YX}(m', F'_\epsilon)] \rightarrow [(m, F_\epsilon) = (m', F'_\epsilon)], \quad \forall (m', F'_\epsilon) in (M, \Gamma)$

**observational equivalence:**
In a model of type Y=m(X, $\epsilon$),Two functions m' and m" $\in M$ are observationally equivalent if
$\exists F'_\epsilon, F"_\epsilon \in \Gamma \ s.t \ \forall (x, y), \ F_{YX}(y, x; m', F'_\epsilon) = F_{YX}(y, x; m", F"_\epsilon)$

**Lemma 1 of Matzkin (2003):**
In a model of type Y=m(X, $\epsilon$), $v, \tilde{v} \in V$ defined as $v = m^{-1}, v' = m'^{-1}$ are observationally equivalent if and only if there exists a strictly increasing function $g : v(A, R) \rightarrow R$ so that $\tilde{v} = g \circ v$ on $A \times R$ .

## 1.2 The model

The model here is defined in a bit of a different way as in D'Haultfœuille and Février (2015). It is just a matter of notation and the results, assumptions and so on are the same. In general, we can always define the joint CDF of Y, our independant variable, and X, our explanatory variable. From here, we can always define the distribution of Y given X:

$$F^{\sigma}_{YX} = F^{\sigma_1}_{Y|X} \ F^{\sigma_2}_{X}$$

The problem here is that what we want to measure (typically elements in $\sigma_1$ ) is dependant with X. It means we have endogeneity. To circumvent this problem, we will use an instrument Z. The instrument will make us able to get parameters independant of X.

$$F^{\sigma'}_{YXZ} = F^{\sigma_3}_{Y|XZ} \ F^{\sigma_4}_{X|Z} F^{\sigma_5}_{Z}$$

Here we got rid of endogeneity. What we want to measure (typically elements in $\sigma_3$) is independant of X. We got back to exogeneity, which is needed to do the modeling here.
From here we can define X and Y:

$$X = h(Z, \eta)$$

$$Y = g(X, \epsilon)$$

Where h(.) and g(.) functions ,$\epsilon$ and $\eta$, random terms (not necessarily the error terms).
The generalization of $\epsilon$ and $\eta$ as being random terms, and not necessarily errors terms that need to be in Y in a nonadditive way comes from Matzkin (2003). $\epsilon$ and $\eta$ can be interpreted as different things depending on the model ( heterogeneity of the sampling for example) and shouldn't be interpreted as an error as like in OLS models. The use of those two random variables as argument of g(.) and h(.) here is there to generalize this to all possible roles of $\epsilon$ and $\eta$.

**Assumption 1**

Strong exogeneity of Z.

$$Z \perp\!\!\!\perp (\epsilon, \eta)$$

We cannot define here strong exogeneity using sequential cut because $\epsilon$ and $\eta$ are generic random variables, not error terms.

**Assumption 2**

Dual Strict monotonicity.

$$\forall (x, z) \in \mathcal{X} \times \{1, ..., K\}, \tau \Rightarrow g(x, \tau),\ \upsilon \Rightarrow h(z, \upsilon)$$

It creates a one to one mapping between ( X, Y) and $(\epsilon, \eta)$ for Z fixed.

**Assumption 3**

$\epsilon$ has a uniform distribution.

$$\epsilon \sim \mathcal{U}(0, 1)$$

It is just a normalization. Using lemma 1 of Matzkin, we know that we just need to find a observationally equivalent functions, not the real ones. Hence, we just need $F_\epsilon$ to be monotonically increasing. We choose thus an uniform distribution for simplicity purpose.

**Assumption 4**

Regularity conditions. Those are really straight forward.
1.

$$support\ of\ X|Z = [\underline{x}, \bar{x}] \perp\!\!\!\perp Z\ with\ -\infty \leq \underline{x} < \bar{x} \leq \infty$$

2.

$$F_\eta\ is\ continuous\ and\ strictly\ increasing\ on\ the\ support\ of\ \eta$$

3.

$(u, \nu) \Rightarrow F_{\epsilon|\eta=\nu}(u)$ is continuous on $[0, 1] \times \mathcal{H}$ where $\mathcal{H}$ is the interior of support of $\eta$

$$u \Rightarrow F_{\epsilon|\eta=\nu} is\ strictly\ increasing\ on\ (0, 1)\ \forall \nu \in \mathcal{H}$$

4.

$$g(.,.)\ is\ continuous\ on\ \mathcal{X} \times (0, 1)$$
$$h(z,.)\ is\ continuous\ on\ \mathcal{H}$$

### 1.3 Theorem

**Suppose Assumptions 1 to 4 hold. Then g is identified if** $\forall(x,x') \in \mathcal{X}, \exists x_1,...,x_j \; s.t. \bar{\mathcal{O}_{x_j}} \cap \bar{\mathcal{O}_{x_{j+1}}} \neq \emptyset, \forall j \in 0,...,J$ **where** $x_0 = x$ *and* $x_{J+1} = x'$

**Proof:** We want to prove g(.,.) is identified. It means, by lemma 1 of Matzkin and definition of identification that we need to prove:

$$\exists Q_{x',x} \text{ strictly increasing s.t.}$$
$$g(x',\tau) = Q_{x',x} \circ g(x,\tau), \quad \forall(x,x',\tau) \in \mathcal{X}^2 \times (0,1)$$

**1. prove it is the case when** $x' \in \mathcal{O}_x$

By definition of the CDF, while fixing values of Z and X (and thus $\eta$ since X= h(Z, $\eta$ ) and Z, h(.,.) and X don't change):

$$F_{Y|X=x,Z=i}(g(x,\tau)) = P(Y \leq g(x,\tau)|\eta = h^{-1}(i,x), Z = i)$$

Since g(x, $\tau$ ) is strictly increasing in $\tau$ , due to assumption 2:

$$= P(\epsilon \leq \tau|\eta = h^{-1}(i,x), Z = i)$$

and, by assumption 1, we know $\eta \perp\!\!\!\perp Z$ , so:

$$= P(\epsilon \leq \tau|\eta = h^{-1}(i,x), Z = j)$$

By definition of a CDF function and define $s_{ij}(x) = h(j, h^{-1}(i,x))$:

$$= F_{Y|X=s_{ij}(x),Z=j}(g(s_{ij}(x),\tau)$$

to sum up, we have:

$$F_{Y|X=x,Z=i}(g(x,\tau)) = F_{Y|X=s_{ij}(x),Z=j}(g(s_{ij}(x),\tau)$$

we know, by the fact that $F_{Y|X=x,Z=i}$ is monotonic, that it is inversible. so:

$$g(x,\tau) = F^{-1}_{Y|X=x,Z=i} \circ F_{Y|X=s_{ij}(x),Z=j}(g(s_{ij}(x),\tau)$$

Now, we can define a monotonic function $Q_{s_{ij}(x),x} = F^{-1}_{Y|X=x,Z=i} \circ F_{Y|X=s_{ij}(x),Z=j}$. we know $x' \in \mathcal{O}_x$, meaning in our case that there exist a function s that can be represented as an inversible function:

$$\exists s_{ij} = \begin{pmatrix} s_{i_1 j_1} & ... & s_{i_p j_1} \\ ... & & ... \\ s_{i_1 j_p} & ... & s_{i_p j_p} \end{pmatrix} \; s.t. \; x' = s_{ij}(x)$$

We can represent it without matrices for the univariate case:

$$\exists s_{ij} = s_{i_1,j_1} \circ ... \circ s_{i_p j_p} \; s.t. \; x' = s_{ij}(x)$$

Which means that the proposition is true for $Q'_x x = Q_{i_1 j_1 x} \circ ... \circ Q_{i_p j_p x}$.

**2. prove it is the case when $x' \in \bar{\mathcal{O}}_x$**

We define by $\bar{\mathcal{O}}_x$, the closure of the orbit of x in $\mathcal{X}$ . Since g is continuous in all its argument ( see assumption 4.4) , we can define a sequence of n x :

$$x_1, ..., x_n s.t \lim_{n\to\infty} x_n = x' \quad with\ \forall x_j \in x_1, ..., x_n : x_j \in \mathcal{O}_x$$

It means that for each term of the sequence $x_j$, by the first point of this proof, $\exists Q_{x',x}$ *strictly increasing s.t.* $g(x_j, \tau) = Q_{x_j,x} \circ g(x, \tau)$ . Even when n tends to infinity. It exists thus too for x' in the closure of the orbit of x.

**3. prove it $\forall (x, x', \tau) \in \mathcal{X}^2 \times (0, 1)$**

By assumption of the theorem, $\exists x_1, ..., x_J$ *s.t* $\bar{\mathcal{O}_{x_j}} \cap \bar{\mathcal{O}_{x_{j+1}}} \neq \emptyset$.
Let $x_j^* \in \bar{\mathcal{O}_{x_j}} \cap \bar{\mathcal{O}_{x_{j+1}}}$
By previous discussion, we have:
$g(x_j^*, \tau) = Q_{x_j^*,x_j} \circ g(x_j, \tau)$ and $g(x_j^*, \tau) = Q_{x_j^*,x_{j+1}} \circ g(x_{j+1}, \tau)$
So:

$$Q_{x_j^*,x_j} \circ g(x_j, \tau) = Q_{x_j^*,x_{j+1}} \circ g(x_{j+1}, \tau)$$

And since $Q_{x_j^*,x_j}$ is strictly increasing, we can invert it:

$$g(x_j, \tau) = Q^-1_{x_j^*,x_j} \circ Q_{x_j^*,x_{j+1}} \circ g(x_{j+1}, \tau)$$

by iteration, for any x' and x:

$$g(x', \tau) = (Q^-1_{x_J^*,x'} \circ Q_{x_J^*,x_J}) \circ ... \circ (Q^-1_{x_0^*,x_1} \circ Q_{x_0^*,x}) \circ g(x, \tau)$$

or in a more compact form using some notation:

$$g(x', \tau) = \mathcal{Q}_{x',x} \circ g(x, \tau)$$

It holds for any $(x, x', \tau)$

**4. Identify g(.)**

Let's define:

$$G_x(u) = E(F_{Y|X} \circ \mathcal{Q}_{Xx}(u))$$

By normalization:

$$G_x(g(x, \tau) = \tau$$

Since the Q and F functions are strictly increasing, $G_x$ is strictly increasing too. Thus it has only one solution.
$g(x, \tau)$ is identified as the solution in u to:

$$G_x(u) = \tau$$

□

## 2 Data table

| **Countries** | **Doctors in %** | **Master equivalent in %** |
|---|---|---|
| Austria | 0.518 | 12 |
| Belgium | 0.308 | 15 |
| Bulgaria | / | / |
| Croatia | / | / |
| Cyprus | / | / |
| Czech Republic | / | 16 |
| Denmark | 0.4 | 11 |
| Estonia | 0.347 | 20 |
| Finland | / | 14 |
| France | 0.337 | / |
| Germany | 0.425 | 11 |
| Greece | / | 2 |
| Hungary | 0.332 | 9 |
| Ireland | / | 8 |
| Italy | / | 14 |
| Latvia | 0.321 | 11 |
| Lithuania | 0.456 | 15 |
| Luxembourg | 0.298 | 18 |
| Malta | / | / |
| Netherlands | / | 12 |
| Poland | 0.238 | 21 |
| Portugal | / | 17 |
| Romania | / | / |
| Slovakia | 0.342 | 17 |
| Slovenia | 0.31 | 15 |
| Spain | 0.388 | 14 |
| Sweden | / | 12 |

# Contents

# Bibliography


Agüero, J. M. (2017). Using partial identification methods to estimate the effect of violence against women on their children's health outcomes. *Applied Economics Letters*, 24(15):1057–1060.

Billings, S. and Zhu, Q. (1991). Rational model identification using an extended least-squares algorithm. *International Journal of Control*, 54(3):529–546.

D'Haultfœuille, X. and Février, P. (2015). Identification of nonseparable triangular models with discrete instruments. *Econometrica*, 83(3):1199–1210.

Gergonne, M. (1815). l'interpolation des suites. In *Annales de Mathématiques pures et appliquées*, volume 6, pages 242–252.

Gundersen, C., Kreider, B., Pepper, J., and Tarasuk, V. (2017). Food assistance programs and food insecurity: implications for canada in light of the mixing problem. *Empirical Economics*, 52(3):1065–1087.

Horowitz, J. L. and Manski, C. F. (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics*, 84(1):37–58.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association*, 95(449):77–84.

Kolmogorov, A. N. (1956). Foundations of the theory of probability.

Lehmann, E. L. and Scheffé, H. (1950). Completeness, similar regions, and unbiased estimation: Part i. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 10(4):305–340.

Lehmann, E. L. and Scheffé, H. (1955). Completeness, similar regions, and unbiased estimation: Part ii. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(3):219–236.

Lusardi, A., Christelis, D., and de Bassa Scheresberg, C. (2016). Entrepreneurship among baby boomers: Recent evidence from the health and retirement study. *Available at SSRN 2898910*.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Marlin, B., Zemel, R. S., Roweis, S., and Slaney, M. (2012). Collaborative filtering and the missing at random assumption.

Matzkin, R. L. (2003). Nonparametric estimation of nonadditive random functions. *Econometrica*, 71(5):1339–1375.

Mavisakalyan, A. and Meinecke, J. (2016). The labor market return to academic fraud. *European Economic Review*, 82:212–230.

OECD (2016). *Education at a Glance 2016*. OECD.

OECD (2019). *Health at a Glance 2019*.

Ord, J. K. (1972). Families of frequency distributions.

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., and Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, 9:157.

RAO, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91.