Partial Identification with Covariates

Armand Kapllani[†] and **Hector H. Sandoval**[‡]

November 9, 2021

JOB MARKET PAPER

Link to Most Current Version

Abstract

The missing outcome problem is a pervasive problem in economics that arises in many situations and hinders the researcher's ability to recover the population moments. The literature primarily focuses on the identifying power of shape restrictions which can be invoked in empirical studies in order to identify the statistics of interest. In this paper, we propose a novel approach of partial identification that does not rely on shape restrictions but instead explores the variation of the covariates in the sample. We illustrate our approach using the Index of Consumer Sentiment where the missing outcome problem resulted from the substitution of landlines with cellphones in telephone surveys. We construct sharp bounds on the Index of Consumer Sentiment and provide conditions under which the bounds are informative. We then extend our approach to the treatment effects literature by constructing bounds on the average treatment effect.

JEL-Codes: C4, C8, D12

Keywords: consumer confidence, coverage error, partial identification, RDD sampling

[†]PhD Candidate in Economics, Department of Economics, University of Florida, 322 MAT, Gainesville Florida 32611-7140. PO Box 117140. Telephone: +1 (352) 392-0382, e-mail: akapllani@ufl.edu.

[‡]Assistant Professor, Department of Economics, University of Florida, 325 MAT, Gainesville Florida 32611-7140. PO Box 117140. Telephone: +1 (352) 392-0475, e-mail: hsandoval@ufl.edu.

1 Introduction

Missing outcomes is a pervasive problem that arises in many situations hindering our capacity to recover population moments. The problem the researcher faces consists of learning the conditional or unconditional mean functions, or the distribution of an outcome of interest when its realizations are observed selectively Manski (1989, 2005). For example, missing outcomes appear as a result of survey nonresponse, attrition in longitudinal studies, or when population members do not appear in the sample frame. The latter problem is known in survey research as coverage bias, and it arises, for example, from the exclusion of cellphone-only population in standard landline telephone surveys. Hence, in order to point identify the population statistics of interest the researcher has to impose assumptions on the distribution of the missing outcomes. The literature has focused on the identifying power of shape restrictions assumptions combined with the sampling process which can be invoked in empirical studies (Manski 1995, 1997*b*; Manski and Pepper 2000; Manski 2003, 2009). Different from the current literature, we propose a novel form of partial identification that exploits variation in the data in order to construct sharp and informative bounds and we illustrate our approach using the University of Michigan Index of Consumer Sentiment. We also extend this approach to the treatment effects literature by constructing bounds on the average treatment effect and average treatment medium effect.

To motivate our paper and explain our main contribution, we formally present the structure of the problem as follows Manski (1989, 2009). Let $\mathcal{J}$ be our population of interest and let each member of the population be characterized by the vector $(Y, X, W, Z) \in \mathbb{Y} \times \mathbb{X} \times \mathbb{W} \times \mathbb{Z}$, where Y denotes the outcome variable of interest, (X, W) are vectors that measure the characteristics of a population member, and Z is a binary variable indicating the observability of the outcome ($Z = 1$ observable). The goal is to infer the population expectation $\mathbb{E}[Y]$. Using the law of iterated expectations, one can decompose the unconditional mean function as follows:[1]

$$\mathbb{E}[Y] = \mathbb{E}[Y|Z=1]\Pr(Z=1) + \mathbb{E}[Y|Z=0]\Pr(Z=0). \tag{1}$$

Suppose that the outcome Y is bounded in its support $\mathcal{Y}$ by a minimum value $\underline{y}$ and maximum value $\overline{y}$ which are both known. That is, $\underline{y} = \min\{Y\}$ and $\overline{y} = \max\{Y\}$, and this implies that both the unconditional and conditional population expectations will also be bounded. In most applications, the outcome Y will naturally bounded by definition. The sampling process reveals $\mathbb{E}[Y|Z=1]$ as we observe the conditional distribution $F_{Y|Z=1}$ but it says nothing about the probability distribution of the missing data $\Pr(Z)$ and the conditional mean function $\mathbb{E}[Y|Z=0]$. Depending on the context, the probability distribution $\Pr(Z)$ might or might not be known.[2] On the contrary, $\mathbb{E}[Y|Z=0]$ is unknown and may take any value in the known interval $[\underline{y}, \overline{y}]$. As a result, the identification region for $\mathbb{E}[Y]$, which considers all the possible values that the missing data can take, is as follows:

$$\mathcal{H}\{\,\mathbb{E}[Y]\} = \{\,\mathbb{E}[Y|Z=1]\Pr(Z=1) + \underline{y}\Pr(Z=0),\ \mathbb{E}[Y|Z=1]\Pr(Z=1) + \overline{y}\Pr(Z=0)\}. \tag{2}$$

The bandwidth of the identification region is $\mathcal{B} = (\overline{y}-\underline{y})\Pr(Z=0)$, a proper subset of $[\underline{y}, \overline{y}]$ when $\Pr(Z=0) > 0$. The bandwidth is a function of two components: (i) the range of Y, $(\overline{y}-\underline{y})$, which can also be seen as the interval corresponding to the identification region of the unobserved $\mathbb{E}[Y|Z=0]$; and (ii) the probability of the missing outcome, $\Pr(Z=0)$. The bandwidth is a singleton if $\mathbb{E}[Y|Z=1] = \mathbb{E}[Y|Z=0]$, that is, when Y is mean independent of Z (i.e., *missing at random*), or trivially when $\Pr(Z=0) = 0$.[3]

[1]The anatomy of the problem described in Manski (1989) considers the conditional expectation $\mathbb{E}[Y|X]$ because his focus is on prediction.

[2]For instance, if missing outcomes appear from attrition, this probability is known. On the contrary, if missing outcomes arise from the undercoverage of population in the sampling process, this probability is not directly revealed because the information is not contained in the sampling framework. It is worth noting that in the latter case not only the outcome is missing but also the covariates. Horowitz and Manski (2000) have studied randomized experiments with missing outcome and covariates and derived bounds on the population moments without imposing untestable assumptions.

[3]Throughout this paper the calligraphic letter $\mathcal{H}\{\cdot\}$ is reserved for identification regions, that is, sets that collect the feasible values of the quantity in the brackets. We say that point identification of the quantity in

In this paper we show how variation in covariates can be exploited in order to tighten the bandwidth $\mathcal{B}$ of the identification region (2) and thus improve the overall identification region $\mathcal{H}\{\mathbb{E}[Y]\}$. In particular, we show how to improve the identification region of $\mathbb{E}[Y|Z=0]$ such that $\mathcal{H}\{\mathbb{E}[Y|Z=0]\} \subset [\underline{y}, \overline{y}]$. Intuitively, our approach exploits the distinct bounds that an outcome can take across different strata or subpopulations. We illustrate our approach considering the coverage bias created from the substitution of landlines with cellphones in telephone surveys in the 2000s and 2010s and the University of Michigan Index of Consumer Sentiment (ICS). The ICS is constructed using consumers' responses from a monthly nationally representative telephone survey of 500 adults since 1978. Prior to July 2012, the sample of adults was selected using a landline random-digital dialing (RDD) sampling and between July 2012 and July 2015 from a dual-frame landline-cellular telephone design. After July 2015 the survey switched to a RDD cellular-only design. To motivate our problem we initially construct the identification region for the ICS, and show that the width of the region of ICS increased almost eight-fold between 2003 and 2012. Our results show that, depending on the covariate employed, the overall identification region for the ICS decreases up to a 32.3% relative to a region spanned by an observed bound and up to 71.2% relative to a theoretical bound.

The literature on partial identification has primarily focused on providing informative bounds on average treatment effects by exploring different shape restriction assumptions (Manski 1989, 1990, 1997*b*; Manski and Pepper 2000). Partial identification by imposing shape restrictions has been applied in different areas such as health economics (e.g. Gerfin and Schellhorn (2006); Kreider et al. (2012); Cygan-Rehm, Kuehnle and Oberfichtner (2017)) and labor economics (e.g. Pepper (2000); Gonzalez (2005); Lee and Wilke (2009); De Haan (2011)). The closest studies that are related to our study are Horowitz and Manski (1998,

the brackets is achieved if the identification region is a singleton and we have partial identification when the identification region contains many elements but it is smaller than all feasible values. We reserve Roman and Blackboard letters for random variables and their supports, respectively. For example, X denotes a random variable with support in space $\mathbb{X}$.

2000) where under different types of survey nonresponses (outcome censoring, joint censoring, regressor censoring, and a mixture of the previous cases) they construct informative bounds on unidentified population parameters but their focus is on bounding the asymptotic bias of estimates using imputations and weights and analyze the problems of inference where the outcome of interest varies with treatment and covariates.

The remainder of the paper is organized as follows. Section 2 develops the identification framework with additional covariates and contains our main contribution. Section 3 introduces the undercoverage bias problem and the ICS coverage bias in telephone surveys and illustrates our approach. In Section 4 we construct bounds on the conditional mean function and in Section 5 we conclude.

2 Partial Identification with Additional Covariates

The conditional mean $\mathbb{E}[Y|Z=0]$ in equation (1) is unknown and may take any value in the interval $[\underline{y}, \overline{y}]$. Thus without any further assumption, $\mathcal{H}\big\{\mathbb{E}[Y|Z=0]\big\} = [\underline{y}, \overline{y}]$, and its bandwidth is $\mathcal{B}_0 = \overline{y} - \underline{y}$. Considering this, the bandwidth of the identification region of $\mathbb{E}[Y]$, which is $\mathcal{B} = (\overline{y} - \underline{y})\Pr(Z=0)$, can be rewritten as $\mathcal{B} = \mathcal{B}_0 \Pr(Z=0)$. The latter expression makes it clear how improvements in the bandwidth $\mathcal{B}_0$ translates to $\mathcal{B}$ and ultimately to the overall identification region of $\mathbb{E}[Y]$. To illustrate the idea, let W denote a binary covariate such that $\Pr(W=1|Z=0) \in (0,1)$. Using the law of total probability we can decompose $\mathbb{E}[Y|Z=0]$ as follows,

$$\mathbb{E}[Y|Z=0] = \mathbb{E}[Y|Z=0, W=1]\Pr(W=1|Z=0) + \mathbb{E}[Y|Z=0, W=0]\Pr(W=0|Z=0) \quad (3)$$

Let $y^i = \mathbb{E}[Y|Z=0, W=i]$, $\underline{y}^i = \min\big\{\mathbb{E}[Y|Z=0, W=i]\big\}$, and $\overline{y}^i = \max\big\{\mathbb{E}[Y|Z=0, W=i]\big\}$ for $i \in \{0,1\}$. Using equation (3), the identification region for the conditional expectation $\mathbb{E}[Y|Z=0]$ is:

$$\begin{aligned}\mathcal{H}\{\,\mathbb{E}[Y|Z=0]\} = \{&\underline{y}^1 \Pr(W=1|Z=0) + \underline{y}^0 \Pr(W=0|Z=0), \\ &\overline{y}^1 \Pr(W=1|Z=0) + \overline{y}^0 \Pr(W=0|Z=0)\},\end{aligned} \tag{4}$$

and its corresponding bandwidth is:

$$\mathcal{B}_1 = (\overline{y}^1 - \underline{y}^1)\Pr(W=1|Z=0) + (\overline{y}^0 - \underline{y}^0)[1 - \Pr(W=1|Z=0)]. \tag{5}$$

By adding and subtracting $(\overline{y} - \underline{y})$, this bandwidth can be rewritten as follows:

$$\begin{aligned}\mathcal{B}_1 = (\overline{y} - \underline{y}) &+ [(\overline{y}^1 - \underline{y}^1) - (\overline{y} - \underline{y})]\Pr(W=1|Z=0) \\ &+ [(\overline{y}^0 - \underline{y}^0) - (\overline{y} - \underline{y})][1 - \Pr(W=1|Z=0)].\end{aligned} \tag{6}$$

Assumption 1. *Range of y^i is not a singleton.* $\underline{y} \leqslant \underline{y}^i < \overline{y}^i \leqslant \overline{y}$ *for* $i \in \{0,1\}$, *where* $\underline{y}$ *and* $\overline{y}$ *correspond to the lower and upper bounds of* Y, *respectively.*

Proposition 1. *If at least one of the weak inequalities in assumption 1 holds with strict inequality for any* $i \in \{0,1\}$, *then* $\mathcal{B}_1 < \mathcal{B}_0$.

Proof. *wlog* let $\underline{y} < \underline{y}^1 < \overline{y}^1 = \overline{y}$ and $\underline{y} = \underline{y}^0 < \overline{y}^0 = \overline{y}$. Hence, using equation (6),

$$\begin{aligned}\mathcal{B}_1 &= (\overline{y} - \underline{y}) - (\underline{y}^1 - \underline{y})\Pr(W=1|Z=0) \\ &< (\overline{y} - \underline{y}) = \mathcal{B}_0\end{aligned}$$

Similarly, let $\underline{y} = \underline{y}^1 < \overline{y}^1 < \overline{y}$ and $\underline{y} = \underline{y}^0 < \overline{y}^0 = \overline{y}$. Hence,

$$\begin{aligned}\mathcal{B}_1 &= (\overline{y} - \underline{y}) - (\overline{y} - \overline{y}^1)\Pr(W=1|Z=0) \\ &< (\overline{y} - \underline{y}) = \mathcal{B}_0\end{aligned}$$

□

Assumption 1 rules out the case where Y is a constant. Proposition 1 shows that if the range of Y conditional on W is contained within the range of Y, that is, $[\underline{y}^i, \overline{y}^i] \subset [\underline{y}, \overline{y}]$, then it is possible to tighten the bandwidth of $\mathbb{E}[Y|Z = 0]$, from $\mathcal{B}_0 = [\overline{y} - \underline{y}]$ to $\mathcal{B}_1$. As a result, the bandwidth of the identification region of $\mathbb{E}[Y]$ in equation (2) becomes $\mathcal{B} = \mathcal{B}_1 \Pr(Z = 0) < \mathcal{B}_0 \Pr(Z = 0)$. Intuitively, proposition 1 exploits the distinct bounds that the outcome can take across different strata or subpopulations. It is worth noting that this proposition shows that it is possible to tighten the bandwidth $\mathcal{B}_0$, but it does not show how it is identified in the data. We show this in the next section.

2.1 Identification of Bandwidth

Proposition 1 shows how to improve the bandwidth $\mathcal{B}_0$ using covariate W. The condition is based on the bounds of the conditional expectations $\mathbb{E}[Y|Z = 0, W = i]$, which are not directly revealed by the data. In this section, we show how to recover these conditional means. We start our analysis using the strong assumption of conditional independence which is specified as follows:

Assumption 2. *Conditional independence.* $\mathbb{E}[Y|Z = 0, W] = \mathbb{E}[Y|Z = 1, W]$

Proposition 2. *If assumption 2 holds,* $\mathbb{E}[Y]$ *is point-identified (Manski (1989)).*

Proof. Under assumption 2 and applying the law of iterated expectations,

$$
\begin{aligned}
\mathbb{E}[Y|Z = 0] &= \sum_{i \in \{0,1\}} \mathbb{E}[Y|Z = 0, W = i] \Pr(W = i|Z = 0) \\
&= \sum_{i \in \{0,1\}} \mathbb{E}[Y|Z = 1, W = i] \Pr(W = i|Z = 0)
\end{aligned}
$$

□

Conditional mean independence point-identifies the expectation of interest $\mathbb{E}[Y]$ by itself. Under this assumption, there is no need to further recover the bandwidth $\mathcal{B}_1$. However, this is a strong assumption that cannot be justified in many settings. We propose a novel approach to identify this bandwidth and introduce it by considering the missing outcome problem that arises from coverage bias in telephone surveys. Nonetheless, the approach can be readily applied to other missing outcome problems as well.

Many telephone surveys use random-digital dialing (RDD) to select a sample of individuals via random selection of their telephone numbers.[4] In the early 2000s, standard RDD survey practices in the U.S. tended to exclude cellphones from their sampling frames (Ehlen and Ehlen (2007)). This exclusion created a type of missing outcome problem, which is known as coverage bias because a subset of the population is not surveyed. This undercoverage was of little concern in the early 2000s, and most importantly, it was not present prior to the cellphone era.[5]

Formally, let T be a binary variable indicating the occurrence of the undercoverage (missing outcome). In our example, $T = 1$ corresponds to the cellphone era, when the exclusion of cellphone-only population resulted in undercoverage. Accordingly, we rewrite the problem in equation (1) as follows:

$$\mathbb{E}[Y|T=1] = \mathbb{E}[Y|Z=1,T=1]\Pr(Z=1|T=1) + \mathbb{E}[Y|Z=0,T=1]\Pr(Z=0|T=1) \tag{7}$$

and its corresponding identification region as

$$\begin{aligned}\mathcal{H}\big\{\mathbb{E}[Y|T=1]\big\} = \big\{&\mathbb{E}[Y|Z=1,T=1]\Pr(Z=1|T=1) + \underline{y}_{t_1}\Pr(Z=0|T=1),\\ &\mathbb{E}[Y|Z=1,T=1]\Pr(Z=1|T=1) + \overline{y}_{t_1}\Pr(Z=0|T=1)\big\}\end{aligned} \tag{8}$$

Where $\underline{y}_{t_1} = \min\big\{\mathbb{E}[Y|Z=0,T=1]\big\}$ and $\overline{y}_{t_1} = \max\big\{\mathbb{E}[Y|Z=0,T=1]\big\}$. The corresponding bandwidth is $\mathcal{B}_{t_1} = (\overline{y}_{t_1} - \underline{y}_{t_1})\Pr(Z=0)$. That is, it depends on the range of Y at $T=1$. As before, we introduce the covariate W and let $\underline{y}^i_{t_1} = \min\big\{\mathbb{E}[Y|Z=0,T=1,W=i]\big\}$ and $\overline{y}^i_{t_1} = \max\big\{\mathbb{E}[Y|Z=0,T=1,W=i]\big\}$ for $i \in \{0,1\}$. In this case, the

[4] Random-digit dialing (RDD) is a probability sampling method that provides a sample of units by randomly selecting their telephone numbers. Wolter, Chowdhury and Kelly (2009) provide a discussion of RDD surveys in U.S.

[5] In 2003, the percentage of landline-only adults was 40.4%, while for cellphone-only and phoneless adults the percentages were 2.8% and 1.6%, respectively.

identification region of $\mathbb{E}[Y|Z=0,T=1]$ is

$$\mathcal{H}\{\,\mathbb{E}[Y|Z=0,T=1]\} = \{\underline{y}^1_{t_1}\Pr(W=1|Z=0,T=1) + \underline{y}^0_{t_1}\Pr(W=0|Z=0,T=1),\\ \overline{y}^1_{t_1}\Pr(W=1|Z=0,T=1) + \overline{y}^0_{t_1}\Pr(W=0|Z=0,T=1)\} \tag{9}$$

and its bandwidth is

$$\mathcal{B}_1 = (\overline{y}-\underline{y}) + [(\overline{y}^1_{t_1}-\underline{y}^1_{t_1})-(\overline{y}-\underline{y})]\Pr(W=1|Z=0,T=1)\\ + [(\overline{y}^0_{t_1}-\underline{y}^0_{t_1})-(\overline{y}-\underline{y})][1-\Pr(W=1|Z=0,T=1)] \tag{10}$$

Assumption 3. *Range extrapolation. For each $i \in \{0,1\}$, let*

(i) $\underline{y}^i_{t_1} = \min\{\,\mathbb{E}[Y|Z=0,W=i,T=1]\} = \min\{\,\mathbb{E}[Y|W=i,T=0]\} = \underline{y}^i_{t_0}$ *and*

(ii) $\overline{y}^i_{t_1} = \max\{\,\mathbb{E}[Y|Z=0,W=i,T=1]\} = \max\{\,\mathbb{E}[Y|W=i,T=0]\} = \overline{y}^i_{t_0}$

Proposition 3. *If assumption 3 holds, the bandwidth $\mathcal{B}_1$ is identified. If in addition the assumption 1 is satisfied, the overall identification region of $\mathbb{E}[Y]$ improves and is identified.*

Proof. Under assumption 3 the bandwidth $\mathcal{B}_1$ is

$$\begin{aligned}\mathcal{B} &= \mathcal{B}_1\Pr(Z=0)\\ &= \big[(\overline{y}^1_{t_1}-\underline{y}^1_{t_1})\Pr(W=1|Z=0) + (\overline{y}^0_{t_1}-\underline{y}^0_{t_1})\Pr(W=0|Z=0)\big]\Pr(Z=0)\\ &= \big[(\overline{y}^1_{t_0}-\underline{y}^1_{t_0})\Pr(W=1|Z=0) + (\overline{y}^0_{t_0}-\underline{y}^0_{t_0})\Pr(W=0|Z=0)\big]\Pr(Z=0)\end{aligned}$$

where all quantities in the last equality are known. □

Assumption 3 says that the range of Y is time invariant and thus it is possible to replace the unobserved range of $\mathbb{E}[Y|Z=0,W,T=1]$ with the observed range of $\mathbb{E}[Y|W,T=0]$. The latter conditional expectation is not censored by the undercoverage because it corresponds to the time period when the cellphone coverage bias did not exist, in our example. Is this a more sensible assumption than conditional independence? This assumption is more reasonable and in many cases the bounds of the variables of interest are restricted by definition and in advance of any realizations.

2.1.1 Example: Binary Response

Consider the case when then variable of interest Y is binary, thus $\mathbb{E}[Y] = \Pr(Y = 1)$ and $\underline{y} = 0$ and $\overline{y} = 1$. In this case, the empirical evidence shows that the bandwidth of the identification region of $\Pr(Y = 1)$ is

$$\mathcal{B} = \mathcal{B}_1 \Pr(Z = 0) = \Pr(Z = 0) \tag{11}$$

Where $\mathcal{B}_1 = (1-0)$ is the bandwidth of the identification region $\mathcal{H}\{\Pr(Y = 1|Z = 0)\}$. That is, $\mathcal{B}$ equals the probability of missing outcomes. When introducing the additional covariate W, $\mathcal{B}$ can be rewritten as

$$\begin{aligned}\mathcal{B} &= \mathcal{B}_1 \Pr(Z = 0) \\ &= \left[(\overline{y}^1 - \underline{y}^1) \Pr(W = 1|Z = 0) + (\overline{y}^0 - \underline{y}^0) \Pr(W = 0|Z = 0)\right] \Pr(Z = 0)\end{aligned} \tag{12}$$

2.2 Multiple Categorical Covariates

The previous results can be extended to the case of multiple categorical covariates. Consider the set of categorical variables $\boldsymbol{X} = \{X_1, ..., X_L\}$ and let the elements of its Cartesian product, $\mathcal{X} \equiv \underset{l\in\mathcal{L}}{\times} X_l$, be indexed by $i \in \mathcal{W} = \{1, \ldots, |\mathcal{X}|\}$ where $|\mathcal{X}|$ is the cardinality of set $\mathcal{X}$ and $\mathcal{L} = \{1, \ldots, L\}$. Define the covariate W as a categorical variable over the elements in the set $\mathcal{X}$. Using the law of total expectations, we can write $\mathbb{E}[Y|Z = 0]$ as follows

$$\mathbb{E}[Y|Z = 0] = \sum_{i\in\mathcal{W}} \mathbb{E}[Y|Z = 0, W = i] \Pr(W = i|Z = 0). \tag{13}$$

Let $\underline{y}^i = \min\{\mathbb{E}[Y|Z = 0, W = i]\}$ and $\overline{y}^i = \max\{\mathbb{E}[Y|Z = 0, W = i]\}$ for all $i \in \mathcal{W}$, as such the identification region of $\mathbb{E}[Y|Z = 0]$ is

$$\mathcal{H}\{\mathbb{E}[Y|Z = 0]\} = \left[\sum_{i\in\mathcal{W}} \underline{y}^i \Pr(W = i|Z = 0), \sum_{i\in\mathcal{W}} \overline{y}^i \Pr(W = i|Z = 0)\right] \tag{14}$$

with corresponding bandwidth $\mathcal{B}_1' = \sum_{i\in\mathcal{W}}(\overline{y}^i - \underline{y}^i) \Pr(W = i|Z = 0)$. Accordingly, assumption 1 and proposition 1 can be extended as follows.

Assumption 4. *Range of y^i is not a singleton. $\underline{y} \leqslant \underline{y}^i < \overline{y}^i \leqslant \overline{y}$ for $i \in \mathcal{W}$.*

Proposition 4. *If $\exists\, i \in \mathcal{W}$ such that at least one of the weak inequalities in assumption 4 holds with strict inequality, then $\mathcal{B}_1' < \mathcal{B}_0$.*

Proof. *wlog* let $\underline{y} < \underline{y}^j < \overline{y}^j = \overline{y}$ for some $j \neq i$ and $\underline{y} = \underline{y}^i < \overline{y}^i = \overline{y} \quad \forall i \neq j$, hence

$$\begin{aligned}\mathcal{B}_1' &= (\overline{y} - \underline{y}) + [(\overline{y}^j - \underline{y}^j) - (\overline{y} - \underline{y})] \Pr(W = j|Z = 0) + \\ &\qquad \sum_{i \neq j} [(\overline{y}^i - \underline{y}^i) - (\overline{y} - \underline{y})] \Pr(W = i|Z = 0) \\ &= (\overline{y} - \underline{y}) - (\underline{y}^j - \underline{y}) \Pr(W = j|Z = 0) \\ &< (\overline{y} - \underline{y}) = \mathcal{B}_0\end{aligned}$$

In a similar way, let $\underline{y} = \underline{y}^j < \overline{y}^j < \overline{y}$ for some $j \neq i$ and $\underline{y} = \underline{y}^i < \overline{y}^i = \overline{y} \quad \forall i \neq j$, hence

$$\begin{aligned}\mathcal{B}_1' &= (\overline{y} - \underline{y}) + [(\overline{y}^j - \underline{y}^j) - (\overline{y} - \underline{y})] \Pr(W = j|Z = 0) + \\ &\qquad \sum_{i \neq j} [(\overline{y}^i - \underline{y}^i) - (\overline{y} - \underline{y})] \Pr(W = i|Z = 0) \\ &= (\overline{y} - \underline{y}) - (\overline{y} - \overline{y}^j) \Pr(W = j|Z = 0) \\ &< (\overline{y} - \underline{y}) = \mathcal{B}_0\end{aligned}$$

□

Similarly, proposition 3 can be extended to identify the bandwidths.

2.2.1 Example: Two Binary Covariates

Let $\boldsymbol{X} = \{X_1, X_2\}$, such that $X_i \in \{0, 1\}$ for $i = 1, 2$. In this case, $\mathcal{X} = \{\{0, 0\}, \{0, 1\}, \{1, 0\}, \{1, 1\}\}$ and W is defined as

$$W = \begin{cases} 1 & \text{if} \quad X_1 = 0 \quad \text{and} \quad X_2 = 0 \\ 2 & \text{if} \quad X_1 = 0 \quad \text{and} \quad X_2 = 1 \\ 3 & \text{if} \quad X_1 = 1 \quad \text{and} \quad X_2 = 0 \\ 4 & \text{if} \quad X_1 = 1 \quad \text{and} \quad X_2 = 1 \end{cases} \tag{15}$$

Using these two categorical covariates and the law of total probability we can rewrite the conditional mean as

$$\mathbb{E}[Y|Z=0] = \sum_{i\in\{1,2,3,4\}} \mathbb{E}[Y|Z=0, W=i] \Pr(W=i|Z=0). \tag{16}$$

Again, let $\underline{y}^i = \min\{\mathbb{E}[Y|Z=0, W=i]\}$ and $\overline{y}^i = \max\{\mathbb{E}[Y|Z=0, W=i]\}$ for all $i \in \{1,2,3,4\}$, hence the identification region is

$$\mathcal{H}\{\mathbb{E}[Y|Z=0]\} = \left[\sum_{i\in\{1,2,3,4\}} \underline{y}^i Pr(W=i|Z=0), \sum_{i\in\{1,2,3,4\}} \overline{y}^i \Pr(W=i|Z=0) \right]. \tag{17}$$

The corresponding bandwidth of this region is

$$\mathcal{B}_1' = \sum_{i\in\{1,2,3,4\}} (\overline{y}^i - \underline{y}^i) \Pr(W=i|Z=0) \tag{18}$$

where we rewrite equation 18 as follows

$$\mathcal{B}_1' = (\overline{y} - \underline{y}) + \sum_{i\in\{1,2,3,4\}} [(\overline{y}^i - \underline{y}^i) - (\overline{y} - \underline{y})] \Pr(W=i|Z=0) \tag{19}$$

Using proposition 3, it is clear that $\mathcal{B}_1' < \mathcal{B}_0$.

3 Application: Undercoverage in the U.S. and Consumer Confidence

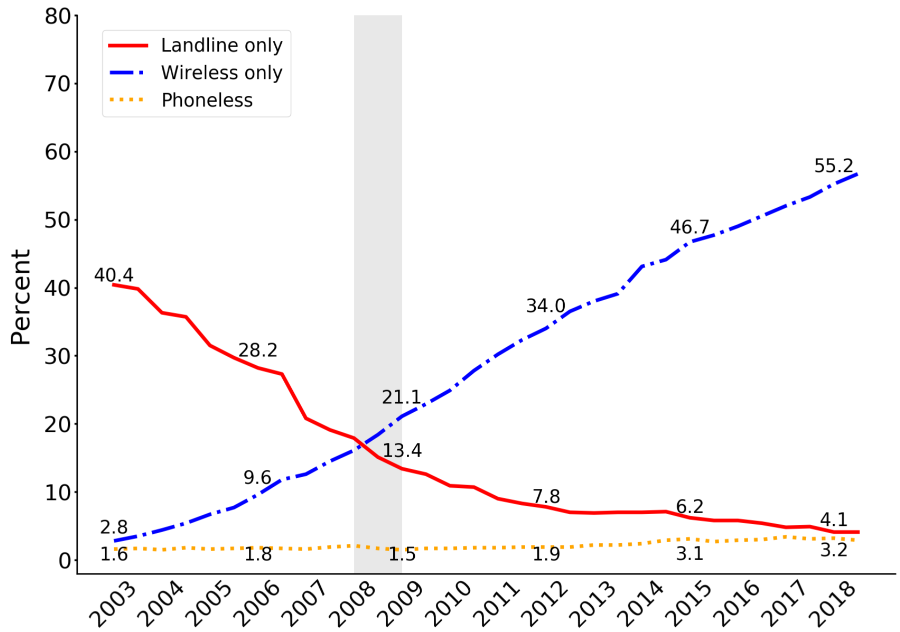
3.1 Undercoverage in U.S.

Undercoverage occurs when population members do not appear in the sample frame, for example, as a result of the exclusion of cellphone-only population in random digit dialing (RDD) landline samples and more recently due to the exclusion of landline-only population in RDD cellphone samples. The undercoverage of cellphone-only population was of little concern in the early 2000s; however, as this population increased, the difference between having cellphone-only or landline-only along with the corresponding characteristics of each group became a potential source of coverage bias.

Figure 1 shows the percentage of adults broken down by landline-only, cellphone-only, and phoneless using data from the National Center for Health Statistics, which releases telephone coverage estimates for the U.S. from a national representative sample, that is, the National Health Interview Survey (NHIS) (Blumberg and Luke 2006, 2010, 2014, 2018). In 2003, the percentage of landline-only adults was 40.4%, while for cellphone-only and phoneless adults the percentages were 2.8% and 1.6%, respectively. By 2009, the percentage of landline-only adults dropped to 13.4% and laid below the percentage of cellphone-only which reached 21.1%. The phoneless adults remained at 1.5% during this period. In 2012, the percentage of landline-only adults decreased to 7.8% and to 4.1% by 2018, while the percentage of cellphone-only adults increased from 34% to 55.2% during this period. Finally, although the percentage of phoneless adults remained low and constant in the 2000s, it increased slightly between 2012 and 2018 from 1.9% to 3.2%.

As a result of the changes in telephone service preferences, the undercovered population in RDD landline sampling (phoneless and cellphone-only) increased monotonically in the last two decades. In 2003, the undercovered population was 4.4%, increasing to 35.9% in 2012, and reaching 58.4% by 2018. On the contrary, the undercovered population in RDD cellphone sampling (phoneless and landline-only) has trended downwards, starting at 42% in 2003 and reaching 7.3% by 2018.

Figure 1: Percentage of Adults by Telephone Status



Source: National Health Interview Survey (NHIS). Shaded area denotes NBER-defined recession

Table 1 shows that the demographic characteristics of cellphone-only and landline-only adults' populations in the U.S. between 2006 and 2018 tend to be different. For instance, looking at the age distribution in 2006, 64.3% of the cellphone-only adult population was aged 18-34, 33.9% aged 35-64, and 1.8% were 65 and older, while the distribution for landline-only was 25.4%, 48.1%, and 26.5%, respectively. This contrast in the age distributions persisted in 2018. The distribution by race across the two populations was similar in 2006, with each population having 50% White, around 26% Hispanic, and 16% Black. By 2018, however, the distributions diverted, and 60% of the cellphone-only population were White and 21.5% Hispanic, while 70% of the landline-only were White and 10.9% Hispanic. Similarly, important differences appear in terms of education achievement. A greater share of landline-only adults has a 4-year college degree or higher compared to cellphone-only adults across the years in the table.

Table 1: Demographics by Telephone Service Status

	Cellphone-only			Landline-only		
	Jan-June 2006	Jan-June 2012	Jan-June 2018	Jan-June 2006	Jan-June 2012	Jan-June 2018
Age						
18-24	33.54	20.16	13.81	10.07	9.55	7.35
25-29	19.25	15.58	11.65	7.31	5.27	3.26
30-34	11.52	13.60	11.40	8.03	6.03	3.63
35-44	16.83	20.08	19.13	16.47	16.15	11.34
45-64	17.09	25.68	31.97	31.67	39.70	37.75
65+	1.77	4.90	12.04	26.45	23.30	36.66
Race						
White	50.64	46.61	59.85	48.78	59.90	70.03
Hispanic	25.54	29.48	21.45	27.94	18.15	10.86
Black	16.17	16.09	11.39	16.85	13.85	11.42
Asian	6.83	6.45	5.60	5.49	7.20	6.38
Other	0.82	1.38	1.70	0.94	0.91	1.31
Education						
Some high school or less	14.46	14.80	8.55	9.19	9.44	5.70
High school graduate or GED	29.35	26.63	22.47	23.37	22.30	18.55
Some post-high school, no degree	31.48	26.66	22.18	21.90	21.43	17.46
4-year college degree or higher	24.72	31.91	46.79	45.54	46.83	58.29
Gender						
Female	46.97	50.93	50.11	52.12	52.87	52.73
Male	53.03	49.07	49.89	47.88	47.13	47.27
Sample Size	30,971	40,929	30,810	30,971	40,929	30,810

Source: Own calculations based on the National Health Interview Surveys (NHIS). Our calculations use the sample weights and closely match the official figures, with differences at the second decimal place. The sample size consists of American adults age 18 and older.

3.2 Consumer Sentiment

Consumer confidence measures are leading indicators that are based on questions relating to households' expectation for changes in business conditions and their financial situation. In particular, these measures have been associated with household spending and future economic activity.[6]

In this section, we illustrate the construction of the identification region using con-

[6]For instance, the study by Carroll, Fuhrer and Wilcox (1994) and Bram and Ludvigson (1998) found that after controlling for economic fundamentals, consumer confidence still has value for predicting household spending in the U.S. Similarly, Barsky and Sims (2012) shows that consumer confidence has predictive implication for the future paths of macroeconomic variables. More recently, Gillitzer and Prasad (2018) and Benhabib and Spiegel (2019) assessed the causal effect of consumer confidence on economic activity using an instrumental variables approach.

sumer confidence. Consumer sentiment surveys are regularly conducted in at least forty-five countries (Curtin 2007). In particular, consumer confidence measures are seen as leading indicators that are associated with consumption expenditure and future economic activity, thus providing an early signal about the strength of the economy (see, for instance, Blanchard (1993); Carroll, Fuhrer and Wilcox (1994); and more recently Benhabib and Spiegel (2019)).

In the U.S., consumer confidence has been measured by the University of Michigan Index of Consumer Sentiment (UMICS).[7] The index is constructed using consumers' responses to five questions, which have remained unchanged since their inception and which are part of a broader survey of consumer attitudes. The responses come from a monthly nationally-representative telephone survey of 500 adults. The survey has been conducted since the 1940s and it is available monthly since 1978. Prior to July 2012, the sample of adults was selected using RDD landline and between July 2012 and July 2015 from a dual-frame landline-cellular telephone design. After July 2015 the survey switched to a cellular-only design[8].

In our application, the *missing at random* assumption is not credible because not only the characteristics of the covered and uncovered population differ (see Table 1), but the consumer confidence levels tend to differ as well. For example, when looking at data prior to 2000, that is, prior to the undercoverage due to the exclusion of cellphones, consumer confidence in the U.S. among those aged 18-34 and 35-54 were each significantly higher than confidence among those aged 55 and older. The latter subpopulation of seniors tends to have more landlines, while the younger subpopulations are typically undercovered with RDD landline sampling.

Our analysis covers the period between 2003 and 2018 due to the availability of tele-

[7] Another national index is the Conference Board's Consumer Confidence Index which uses a mail out survey, hence exempt from the issues discussed here.

[8] Another index, the University of Florida Consumer Sentiment Index (UFCSI) also comes from a monthly telephone survey from around 500 randomly selected adult residents of Florida. Since its inception, the survey has used RDD landline, but switched to RDD cellphone in January 2015. The index is made up of the same five questions as the Michigan index but it is available since 1985 and monthly since 1991[9] We construct the identification region for the Florida index also. The results can be provided upon request.

phone coverage data. This information comes from the National Health Interview Surveys (NHIS) and is available every six months; thus we calculate a semiannual measure of consumer confidence for each index by averaging the corresponding months. In Figure 2 we plot the consumer sentiment index of the University of Michigan and its identification region following the bounds provided in equation (2). We use the historical observed minimum of 55.3 and maximum of 112 to construct the identification region.[10] Table A1 in Appendix A contains all the information behind the figures and our analysis.

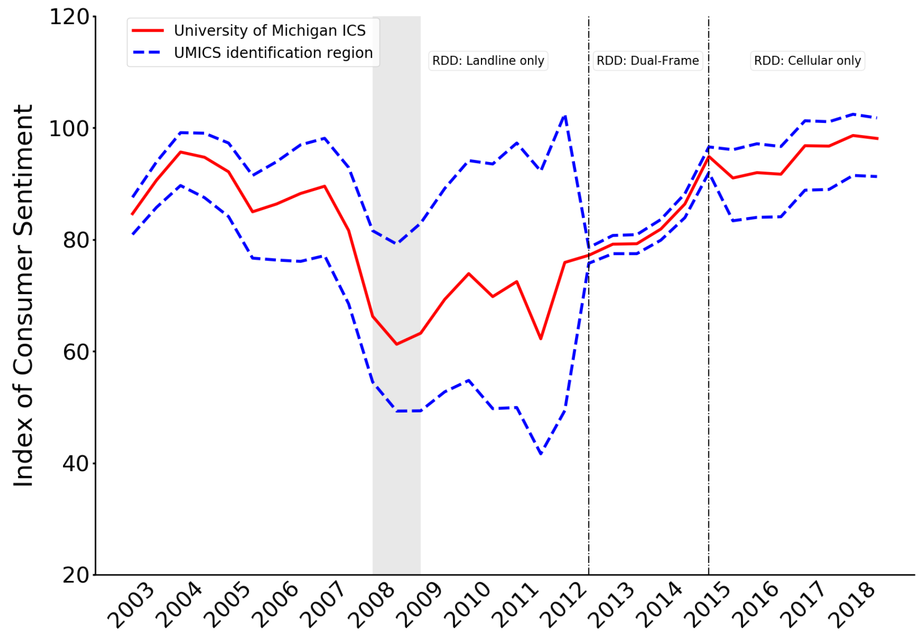
Looking at UMICS in Figure 2, our calculations show that in the early 2000s the identification region is relatively small, as expected, since only 4.5% of the adult population was undercovered. In the late 2000s, the region increases considerably, particularly during the recession years, and it continues to widen until July 2012. In the second half of 2008, when the undercovered population was 20%, UMICS reached its lowest semiannual value of 61.3 with corresponding bounds of 60 and 71.5 (a width of 11.5). The upper bound sets consumer confidence above the levels observed six months before the recession. By 2012, the undercovered population reached 36%, and the bounds are 68.5 and 88.9 (a width of 20.4). It is worth noting that the published index tends to be closer to its lower bound until the switch to cellular-only, thus potentially providing a more pessimistic outlook. After July 2012, the width of the region collapses sharply since only the phoneless adults, 2% of the population, were undercovered when the survey switched to dual-frame landline-cellular design. The size of the region remains stable until the first half of 2015. In the second half of 2015, the region increased when the survey finally switched to RDD cellphone and the uncovered population, in this case, reached 8.6%.

Finally, Figure 3 summarizes our findings by plotting the width of the identification region for the index. As shown before, this width also corresponds to the width of the identification region for the coverage error. The widths increase monotonically since 2003

[10] In the Michigan index, the minimum corresponds to the reading observed in November 2008 and the maximum to January 2000

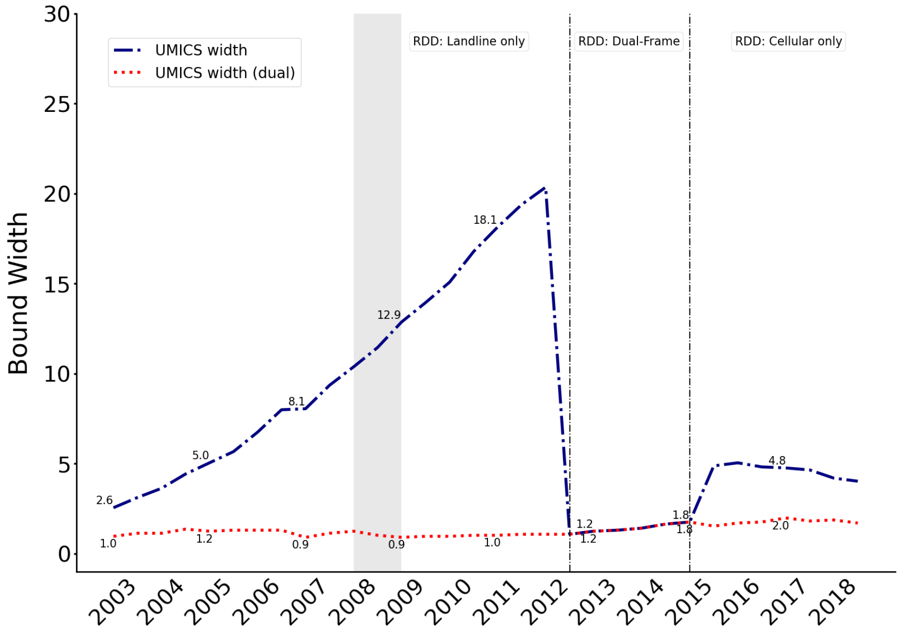
up until July 2012 and January 2015 for UMICS. The width of the index increases from 2.6 in 2003 to 20.4 in 2012, that is, almost eightfold. A dual-frame design clearly reduces the width of the region considerably. For instance, the switch from RDD landline to a dual-frame reduced the width of the index from 20.4 to 1.1 in 2012. As expected, this width increases when the survey switches to RDD cellphone, however, the magnitude of the change is noticeably small from 1.8 to 4.9 in 2015. In addition, the figure plots the width for a hypothetical region in which the index adopts a dual-frame during the whole period. This width remains low and constant over time, as expected. Notably the difference between using a dual-frame and RDD cellphone are small in the later years. For instance, in the second-half of 2015, the difference was 3.4, and by the end of 2018, this gap reduces to 2.3.

Figure 2: University of Michigan Index of Consumer Sentiment and Identification Region



Note: Shaded area denotes NBER-defined recession and vertical lines mark changes in RDD sampling design. The bounds are constructed using Table 1 and equation (2).

Figure 3: Widths of the Identification Region



Note: Shaded area denotes NBER-defined recession and vertical lines mark changes in RDD sampling design. The widths of the identification region are constructed using Table 2 and equation (2)).

Table 2 shows the descriptive statistics of ICS by age, gender, and education, and by period before and after the year 2000. The observed bound of the index can be restricted to the interval $[54.3, 117, 3]$. The lower bound is observed among respondents 55 years and older and the upper bound among those with a college degree or more.

Table 2: Index of Consumer Sentiment (ICS) Descriptive Statistics

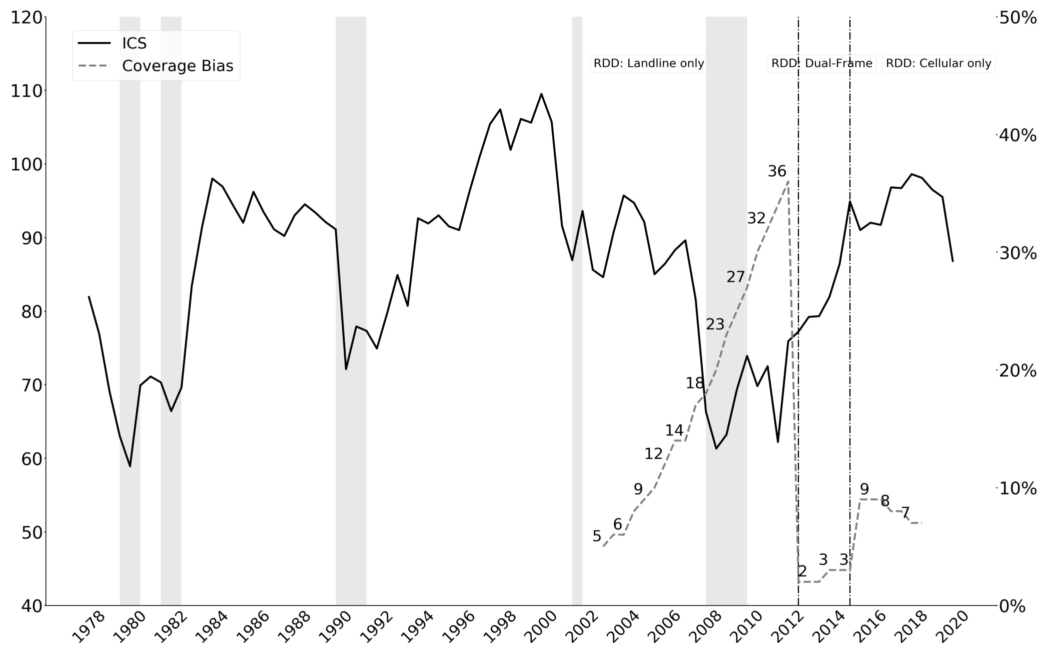
	Period: 1978-2000				Period: 2001-2018				Period: 1978-2018			
	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean
ICS	58.92	109.45	91.43	87.70	61.25	98.63	86.37	84.00	58.92	109.45	90.39	86.08
Age groups												
18-34	66.60	116.28	100.31	96.16	69.22	108.72	97.25	94.49	66.60	116.28	99.18	95.43
35-54	55.13	111.35	92.20	87.65	61.75	101.72	87.98	85.84	55.13	111.35	91.52	86.86
55+	54.28	101.58	80.61	79.36	57.32	96.27	78.13	77.66	54.28	101.58	80.08	78.61
Gender												
Male	65.30	112.22	97.35	93.30	63.98	105.93	91.90	88.90	63.98	112.22	95.54	91.37
Female	54.52	107.22	86.80	83.40	57.00	91.48	81.57	79.45	54.52	107.22	85.07	81.67
Education level												
High school or less	56.03	101.43	85.38	82.43	58.73	99.87	80.47	78.21	56.03	101.43	82.35	80.57
Some college	63.08	110.37	94.17	90.95	59.53	101.55	86.78	83.43	59.53	110.37	89.03	87.65
College degree or more	62.78	117.30	99.02	94.84	62.95	101.85	92.30	88.86	62.78	117.30	96.09	92.22

Note: The table shows the descriptive statistics of University of Michigan Index of Consumer Sentiment (ICS) by age, gender, and education using half-year data for the period 1978 - 2018.

Prior to July 2012, the sample of adults was selected using RDD landline and between July 2012 and July 2015 the design changed to a dual-frame landline-cellular. After July 2015 the survey switched to a cellular-only design. As a result of the changes in the design, different members of the population have been excluded from the sampling frame at different moments, giving rise to varying degrees of coverage bias over time.

Figure 4 plots consumer sentiment between 1978 and 2018 and the evolution of its coverage bias since 2013 using a half-year frequency. The vertical lines in the figure correspond to the changes in the sampling design. The undercovered population is calculated using the National Health Interview Surveys (NHIS), which provides telephone coverage estimates for the U.S. every six months between 2003 and 2018. For this reason, we calculate a half-year measure of consumer confidence by averaging the corresponding months. Table A1 in Appendix A contains the percentage for each population, the percentage of undercovered population (coverage bias), as well as the half-year ICS. Prior to 2012, the phoneless and only-wireless populations were excluded from the sampling frame. However, as a result of the rapid substitution of landline with cellphones in the 2000s, the percentage of undercovered population increased monotonically from 5% to 36% between 2003 and 2012. Between 2012 and 2015, only the phoneless population was excluded from the sampling frame. Consequently, during this period of time, the undercovered population declined sharply, remaining around 2.5% on average. Finally, in the second-half of 2015, both the phoneless and only-landline populations were excluded, and thus the undercovered population reached 9% and has slowly declined since, setting at 7% by 2018.

Figure 4: Consumer Sentiment and Coverage Bias



Note: The figure plots the University of Michigan Index of Consumer Sentiment (ICS) between 1978 and 2018 and the percent of undercovered population between 2003 and 2018 using half-year data. The undercovered population corresponds to the members of the population excluded from the sample frame. The shaded areas denotes NBER-defined recessions and the vertical lines denote changes in the sampling design.

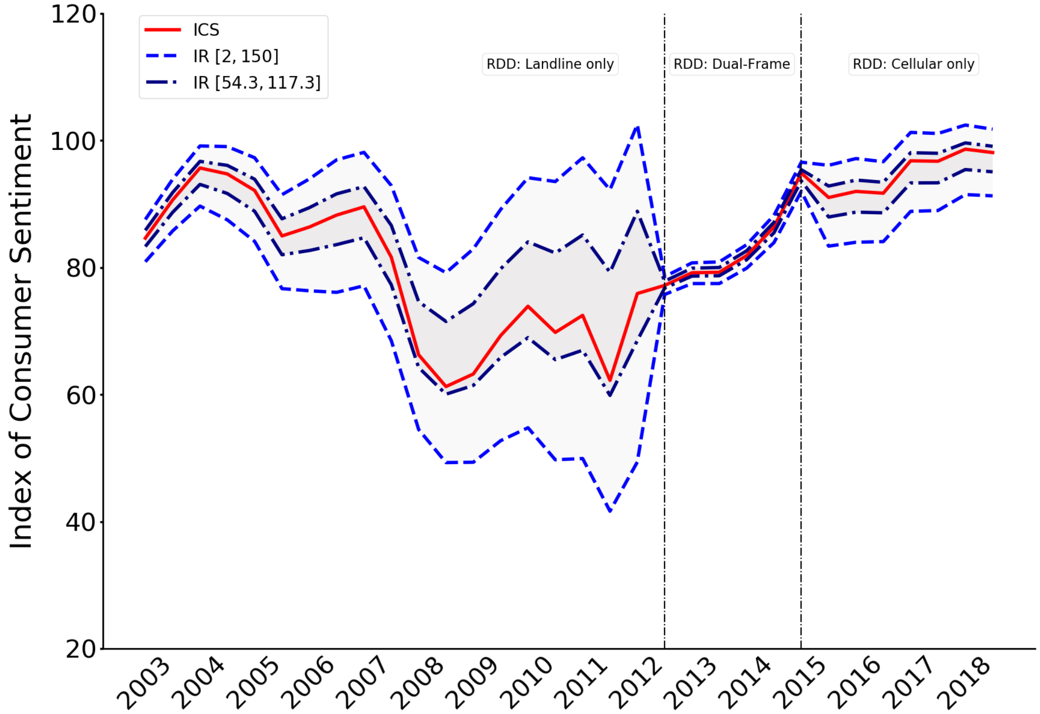
3.3 ICS Identification Region

Figure 5 plots consumer sentiment and its identification region described by equation (2) using two different bounds for the index. The wider region considers the bound $[2, 150]$, while the tighter one uses $[54.3, 117.3]$. The former bound is based on the potential values that the index can take in theory and the latter corresponds to the observed range of the index between 1978 and 2000.[11] The identification region spanned by the empirical bound is 42.6% of the region spanned by the theoretical bound, and it is strictly contained in it.[12] Regardless of the bound considered, our calculations show that in the early 2000s the identification region of the index was relatively small, as expected, since only 4.5% of the adult population was undercovered. In the late 2000s, the region increased considerably, particularly during the recession years, and it continued to widen until July 2012. In the second half of 2008, the ICS reached 61.3 points, its lowest value in our time frame, and the undercovered population reached 20%. As a results of this undercoverage, the index could have taken any value in the interval $[49.3, 79.2]$, if the bound $[2, 150]$ is employed, or any value in the interval $[59.8, 72.6]$, if the bound $[54.3, 117.3]$ is considered. By the first-half of 2012, consumer sentiment reached 75.9 points and the percent of undercovered population peaked at 36%. However, the index could have taken any value in the intervals $[49.4, 102.5]$ or $[68.2, 90.8]$ depending on the bound considered. It is worth noticing that the published index tends to be closer to its lower bound. After July 2012, the width of both identification regions collapse sharply since only the phoneless adults, 2% of the population, were excluded. The size of the regions remained small and stable until the first half of 2015. In the second half of 2015, when the survey made a final switch to the RDD cellphone design, the identification region widened as a results of the increase in the undercovered population, which set at around 8.6%.

[11] We consider the period 1978-2000 to avoid any coverage bias from the substitution of landlines with cellphones. Nonetheless, the interval remains the same when considering the whole period, 1978-2018.

[12] At every point in time, the bandwidths are $\mathcal{B}_{[2,150]} = (150 - 2)\Pr(Z = 0)$ and $\mathcal{B}_{[54.3,117.3]} = (117.3 - 54.3)\Pr(Z = 0)$. Hence $\mathcal{B}_{[54.3,117.3]}/\mathcal{B}_{[2,150]} = (117.3 - 54.3)/(150 - 2) = 0.426$.

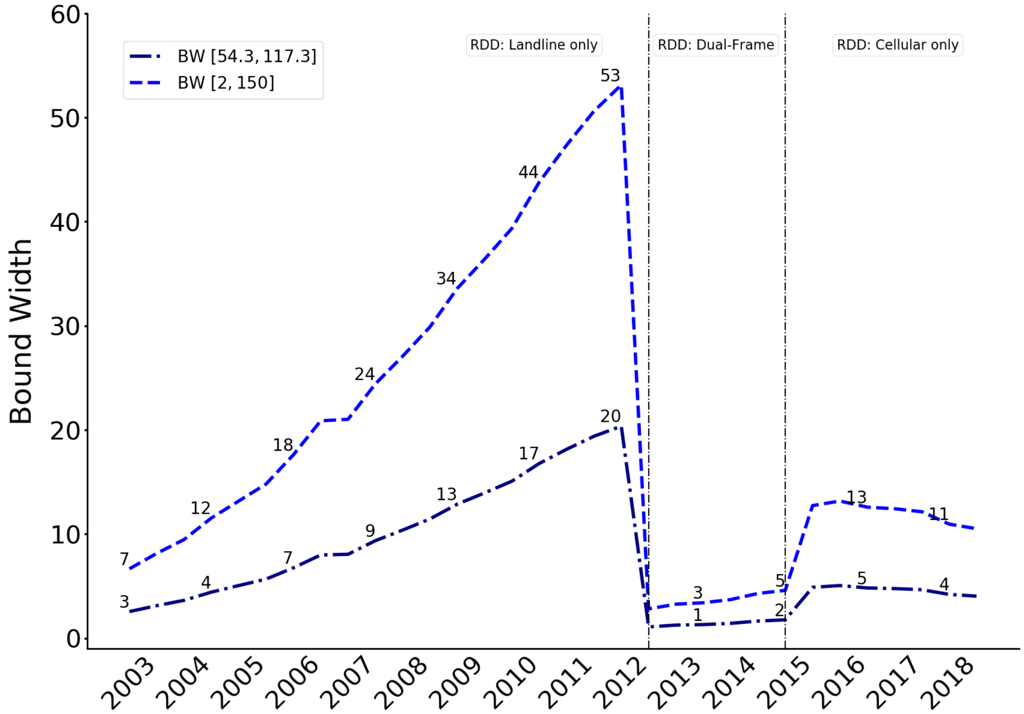
Figure 5: Identification Region of Consumer Sentiment



Note: The figure plots the identification region (IR) of the University of Michigan Index of Consumer Sentiment (ICS). The regions are constructed using equation (2) and the bounds $[2, 150]$ and $[54.3, 117.3]$. The shaded area denotes NBER-defined recession and the vertical lines denote changes in the sampling design.

Figure 6 plots the bandwidths of the two identification regions and provides a visual summary of the behavior of each region over time. The magnitude of the bandwidths increases monotonically since 2003 up until July 2012. The bandwidth of the region using the bound $[2, 150]$ increases from 7 to 53 points, while the one using the bound $[54.3, 117, 3]$ increases at a slower pace, from 3 to 20 points. As expected, the bandwidths became smaller between 2012 and 2015, and increased again after the second-half of 2015.

Figure 6: Bandwidth of the Identification Region



Note: The figure plots the bandwidth (BW) of the identification region of the ICS using the bounds $[2, 150]$ and $[54.3, 117.3]$. The bandwidths are constructed using equation (2). The vertical lines denote changes in the sampling design.

3.4 ICS Identification Region with Covariates

Consumer confidence is split by demographics such as age, gender, or education. For instance, across all data points in our time-series, consumer confidence differs by gender, with women reporting less confidence. The descriptive statistics in Table 1 shows that consumer confidence for women is 83.4 points on average between 1978 and 2000, while for men is 93.3 points.[13] Furthermore, the bound for women is contained in the bound of the whole index, that is, $[54.5, 107.2] \subset [54.3, 117.3] \subset [2, 150]$. Similarly, the bound for men is $[65.3, 112.2] \subset [54.3, 117.3] \subset [2, 150]$. Figure 9 in Appendix A provides a visual inspection of the evolution of consumer confidence by gender. Notably, a similar pattern occurs when considering consumer confidence by age or education, with respondents age 55 and

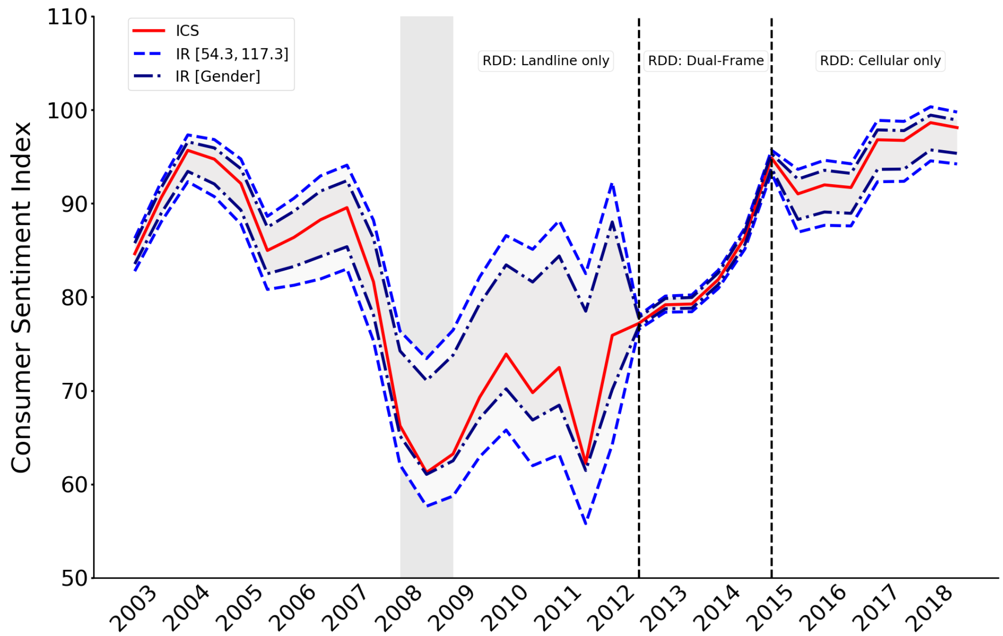
[13] We consider the period 1978-2000 to avoid any potential coverage bias.

older or with high school or less reporting less confidence (Figure 10 and 11 in Appendix A, respectively).

Following proposition 1, it is possible to tighten the identification region of the index using a covariate. Using equation (5), we recalculate the identification region considering gender as covariate. We consider the bounds $(\underline{y}^0, \overline{y}^0) = (54.5, 107.2)$ for women and $(\underline{y}^1, \overline{y}^1) = (65.3, 112.2)$ for men. We calculate the $\Pr(W = 0 \mid Z = 0)$ and $\Pr(W = 1 \mid Z = 0)$ using the NHIS.

Figure 7 plots the identification region using the observed bound $[54.3, 117.3]$ and the improved region using the covariate gender. As anticipated, the latter region is contained in the former one. For example, in the first-half of 2012, when the undercovered population peaked at 36%, the index could have taken any value in the interval $[68.1, 90.8]$, if the bound $[54.3, 117.3]$ is considered, but the region improves to $[70.1, 88]$ with the covariate gender. In other words, the region shrinks by 21.1%. Figures 12, 13, and 14 in Appendix B show the identification region using age, education, and the interacion of education and gender, respectively. We provide a summary of the results in Table 2 and a visual inspection in Figure 4.

Figure 7: Bounds using gender as covariate



Note: The figure plots the identification region (IR) of the University of Michigan Index of Consumer Sentiment (ICS). The regions are constructed using (2) and the interval $[54.3, 117.3]$, and (6) and gender as covariate. The shaded area denotes NBER-defined recession and the vertical lines denote changes in the design.

Table 3 summarizes the bandwidth improvements of the identification region using different covariates. In addition, we present the improvements across the different sampling designs that occurred over time. The average bandwidth using the theoretical bound $[2, 150]$ during the RDD landline period was 26.8. This average bandwidth declines to 3.7 when the survey switched to a dual-frame and it increases to 12.1 in the RDD cellphone period after the second-half of 2015. A similar pattern is observed for the average bandwidth of the identification region spanned by the observed bound $[54.3, 117.3]$. When using the covariate gender to improve the region, the average bandwidth during the RDD landline period is 9, and drops to 1.2 and 4.1 in the dual-frame and RDD cellphone periods, respectively. As a result, the overall identification region of the index decreases by 66.4% with respect to the region spanned by the bound $[2, 150]$ and by 21.2% with respect to the region using the observed bound $[54.3, 117.3]$.[14] Finally, it is worth noting that the covariate education

[14]We calculate the percentage change for each time point within the corresponding period and then we

produces the greatest improvement, reducing the identification region by 67.8% or by 24.4% depending on the baseline region considered. Furthermore, when the interaction of gender and education is considered, the identification region of the index declines up to 71.2% or up to 32.3%. Finally, Figure 8 provides a visual inspection of the improvements in the bandwidths. The bandwidth improvements are computed as the absolute difference between the improved bounds using the covariates and the bounds constructed using the minimum value of the index 54.3 and maximum of of the index 117.3.

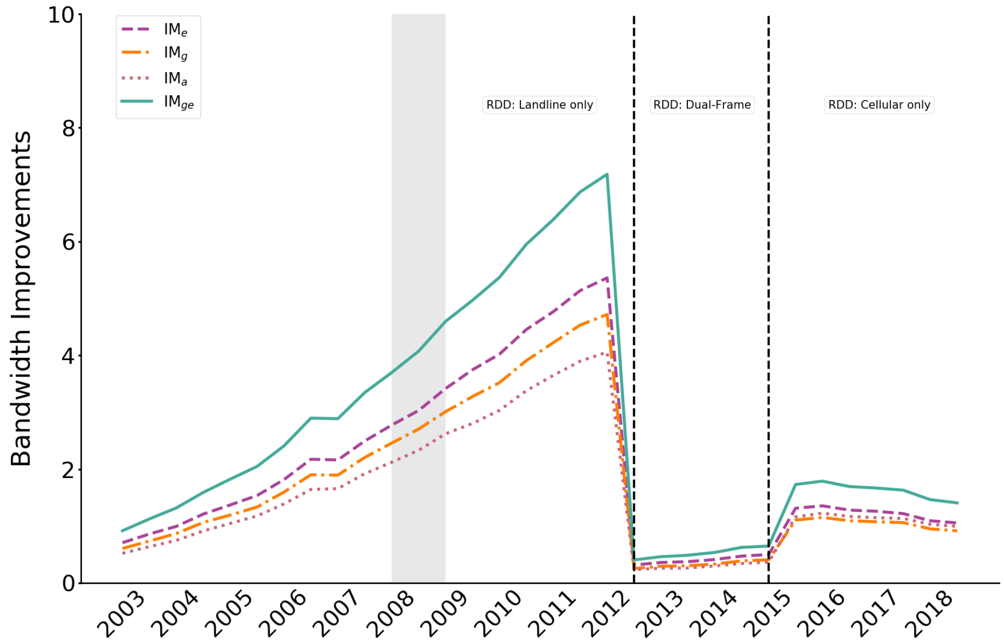
compute the average by period.

Table 3: Summary of Bandwidth Improvements

	RDD Landline-only (2003h1-2012h1)			**RDD Dual-frame** (2012h2-2015h1)			**RDD Cellphone-only** (2015h2-2018h2)			**All** (2003h1-2018h2)		
	Average BW	Improvement [1]	Improvement [2]	Average BW	Improvement [1]	Improvement [2]	Average BW	Improvement [1]	Improvement [2]	Average BW	Improvement [1]	Improvement [2]
IR [2,150]	26.8			3.7			12.1			19.2		
IR [54.3,117.3]	11.4			1.6			5.1			8.2		
IR [Gender]	9.0	-21.2	-66.4	1.2	-21.0	-66.4	4.1	-20.4	-66.1	6.5	-21.0	-66.4
IR [Age]	9.3	-18.3	-65.2	1.3	-18.8	-65.4	4.0	-21.9	-66.7	6.7	-19.2	-65.6
IR [Education]	8.7	-24.2	-67.7	1.2	-25.9	-68.5	3.9	-23.8	-67.6	6.2	-24.4	-67.8
IR [Education & Gender]	7.8	-32.2	-71.1	1.0	-33.7	-71.8	3.5	-31.6	-70.9	5.6	-32.3	-71.2

Note: The table summarizes the bandwidth improvements. The average bandwidth (BW) from the identification regions (IR) is shown in the first column of every different type of RDD. The average improvements of IR [Gender], IR [Age], IR [Education], and IR [Education & Gender] relative to IR[54.3,117.3] and IR[2,150] are shown in columns [1] and [2] for every different type of RDD, respectively.

Figure 8: Bound improvements



Note: Key to symbols: ICS = University of Michigan Index of Consumer Sentiment, IM_g = Bound improvement using gender as covariate, IM_a = Bound improvement using age as covariate, IM_{ge} = Bound improvement using gender and education as covariate. All bound improvements are relative to the IR$[54.3, 117.3]$

4 Bounds on Conditional Mean Function

We now consider the conditional mean function. Let the vector (Y, X, Z, W) characterize some of member of the population of interest, where Y is the outcome of interest, X is vector of covariates, and Z is a binary variable taking unity if the outcome Y can be observed by the researcher and zero otherwise. Using the law of iterated expectations (LIE), we can decompose the conditional mean as follows:

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, Z = 0] \Pr(Z = 0|X) + \mathbb{E}[Y|X, Z = 1] \Pr(Z = 1|X) \tag{20}$$

where the latent conditional mean $\mathbb{E}[Y|X, Z = 0]$ is unobserved by the researcher and cannot be point identified, while the other components of equation (20) can be point identified using the sampling process. We assume that Y conditional on X is bounded within the interval

$[\underline{y}^x, \overline{y}^x]$, where the bounds of the latter are known. Using this fact the identification region of the conditional mean is as follows:

$$\begin{aligned}\mathcal{H}\{\mathbb{E}[Y|X]\} = [&\underline{y}^x \Pr(Z=0|X) + \mathbb{E}[Y|X, Z=1]\Pr(Z=1|X), \\ &\overline{y}^x \Pr(Z=0|X) + \mathbb{E}[Y|X, Z=1]\Pr(Z=1|X)],\end{aligned} \tag{21}$$

and the bandwidth of the identification region is $\mathcal{B}_1^x = \mathcal{B}_0^x \Pr(Z=0|X)$ with $\mathcal{B}_0^x = (\overline{y}^x - \underline{y}^x)$. The tightness of the identification region is a function of $\mathcal{B}_0^x$ and $\Pr(Z=0|X)$, but we can further refine the bounds on $\mathbb{E}[Y|X]$ if we can tighten $\mathcal{B}_0^x$.

To illustrate the idea consider a discrete random variable, W, with support $\mathcal{W}$ and realized values i, then by the LIE we can decompose the latent conditional mean $\mathbb{E}[Y|X, Z=0]$ in (20) as follows:

$$\mathbb{E}[Y|X, Z=0] = \sum_i \mathbb{E}[Y|X, Z=0, W=i]\Pr(W=i|X, Z=0). \tag{22}$$

Suppose that Y conditional of X, $Z=0$, and $W=i$ is bounded between $\underline{y}^{xi}$ and $\overline{y}^{xi}$, where $-\infty < \underline{y}^{xi} \leqslant \overline{y}^{xi} < \infty$, then, the identification region of $\mathbb{E}[Y|X, Z=0]$ is as follows:

$$\mathcal{H}\{\mathbb{E}[Y|X, Z=0]\} = \left[\sum_i \underline{y}^{xi} P(W=i|Z=0, X), \sum_i \overline{y}^{xi} P(W=i|Z=0, X)\right] \tag{23}$$

where the bandwidth of the identification region is

$$\mathcal{B}_0^{xi} = \sum_w (\overline{y}^{xi} - \underline{y}^{xi}) P(W=i|Z=0, X) \tag{24}$$

Proposition 5. *Suppose that the range of y^{xi} is not a singleton and $\underline{y}^x \leqslant \underline{y}^{xi} < \overline{y}^{xi} \leqslant \overline{y}^x$ for i, then if there exists $W=i$ such that at least one of the weak inequalities holds with strict inequality, then $\mathcal{B}_0^{xi} < \mathcal{B}_0^x$.*

Proof. *wlog* let $\underline{y}^x \leqslant \underline{y}^{xj} < \overline{y}^{xj} = \overline{y}^x$ for $j \neq i$ and $\underline{y}^x = \underline{y}^{xi} < \overline{y}^{xi} \leqslant \overline{y}^x$ for all $i \neq j$, hence

$$\begin{aligned}
\mathcal{B}_0^{xi} &= (\overline{y}^x - \underline{y}^x) + [(\overline{y}^{xj} - \underline{y}^{xj}) - (\overline{y}^x - \underline{y}^x)] \Pr(W = j|Z = 0, X) + \\
&\qquad \sum_{i \neq j} [(\overline{y}^{xi} - \underline{y}^{xi}) - (\overline{y}^x - \underline{y}^x)] \Pr(W = i|Z = 0, X) \\
&= (\overline{y}^x - \underline{y}^x) - (\underline{y}^{xj} - \underline{y}^x) \Pr(W = j|Z = 0, X) \\
&< (\overline{y}^x - \underline{y}^x) = \mathcal{B}_0^x
\end{aligned}$$

In a similar fashion, let $\underline{y}^x = \underline{y}^{xj} < \overline{y}^{xj} < \overline{y}^x$ for some $j \neq i$ and $\underline{y}^x = \underline{y}^{xi} < \overline{y}^{xi} = \overline{y} \quad \forall i \neq j$, hence

$$\begin{aligned}
\mathcal{B}_0^{xi} &= (\overline{y}^x - \underline{y}^x) + [(\overline{y}^{xj} - \underline{y}^{xj}) - (\overline{y}^x - \underline{y}^x)] \Pr(W = j|Z = 0, X) + \\
&\qquad \sum_{i \neq j} [(\overline{y}^{xi} - \underline{y}^{xi}) - (\overline{y}^x - \underline{y}^x)] \Pr(W = i|Z = 0, X) \\
&= (\overline{y}^x - \underline{y}^x) - (\overline{y}^x - \overline{y}^{xj}) \Pr(W = j|Z = 0, X) \\
&< (\overline{y}^x - \underline{y}^x) = \mathcal{B}_0^x
\end{aligned}$$

□

4.1 Bounds Using Level-set Restrictions

Manski (1990) explored the identifying power of level set restrictions and in this section we extend his work in refining the bounds on the conditional mean function. Our conditional moment of interest is the conditional mean function, $\mathbb{E}[Y|X]$, and by the LIE we can decompose it as follows:

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, Z = 0]Pr(Z = 0) + \mathbb{E}[Y|X, Z = 1]Pr(Z = 1) \tag{25}$$

Suppose that the outcome Y conditional of X is bounded within some known interval $[\underline{y}^x, \overline{y}^x]$, where $-\infty < \underline{y}^x \leqslant \overline{y}^x < \infty$ and let $\mu(X) \equiv \mathbb{E}[Y|X]$.

$$\mathbb{E}[Y|X] \in \mathcal{M}(X) \equiv \big[\underline{y}^x Pr(Z=0|X) + \mathbb{E}[Y|X, Z=1]Pr(Z=1|X), \\ \overline{y}^x Pr(Z=0|X) + \mathbb{E}[Y|X, Z=1]Pr(Z=1|X)\big] \tag{26}$$

Now suppose that $\mathbb{E}[Y|X]$ is constant on some set $X_0 \subset \mathcal{X}$, then the collection of bounds $\mathcal{M}(X)$ with $X \in X_0$, has a non-empty intersection which contains the common value of $\mathbb{E}[Y|X]$. That is, for each $\kappa \in X_0$ we have:

$$\begin{aligned} \mu(\kappa) \in \mathcal{M}(\kappa) &\equiv \bigcap_{x \in X_0} \mathcal{M}(X=x) \\ &\equiv \Bigg[\sup_{x \in X_0} \Big\{ \underline{y}^x Pr(Z=0|X=x) + \mathbb{E}[Y|X=x, Z=1]Pr(Z=1|X=x) \Big\}, \\ &\qquad \inf_{x \in X_0} \Big\{ \overline{y}^x Pr(Z=0|X=x) + \mathbb{E}[Y|X=x, Z=1]Pr(Z=1|X=x) \Big\} \Bigg] \end{aligned} \tag{27}$$

In order to refine the bounds on equation (27), let W be a discrete random variable, then by the LIE we can decompose $\mathbb{E}\left[Y|X, Z=0\right]$ as follows:

$$\mathbb{E}\left[Y|X, Z=0\right] = \sum_w \mathbb{E}[Y\big|X, Z=0, W=w]\,\Pr(W=w|X, Z=0) \tag{28}$$

The following assumption says that the conditional mean function, $\mathbb{E}\left[Y\big|X, Z=0, W\right]$, in equation (28), has a bounded support.

Assumption 5. *Conditional on X, $Z=0$, and $W=w$, Y is bounded between $\underline{y}^{x0w}$ and $\overline{y}^{x0w}$, for each w, where $-\infty < \underline{y}^{x0w} \leqslant \overline{y}^{x0w} < \infty$.*

Then, it follows directly from Assumption 5 that $\underline{y}^{x0w} \leqslant \mathbb{E}[Y\big|X, Z=0, W=w] \leqslant \overline{y}^{x0w}$ for every w and suppose that the conditional mean function, $\mathbb{E}[Y\big|X, Z=0]$, is constant over some space $X_{x0} \in X$. Then, for any $\kappa \in \mathcal{X}_{x0}$ we have:

$$
\begin{aligned}
\mu(\kappa) \in \mathcal{M}(\kappa) &\equiv \bigcap_{x \in X_{x0}} \mathcal{M}(X = x, Z = 0) \\
&\equiv \Bigg[\sup_{x \in X_{x0}} \Big\{ \sum_w \underline{y}^{x0w} Pr(W = w | X = x, Z = 0) \Big\}, \\
&\qquad \inf_{x \in X_{x0}} \Big\{ \sum_w \overline{y}^{x0w} Pr(W = w | X = x, Z = 0) \Big\} \Bigg]
\end{aligned} \tag{29}
$$

and the width of identification region is

$$
\begin{aligned}
&\inf_{x \in X_{x0}} \Big\{ \sum_w \overline{y}^{x0w} Pr(W = w | X = x, Z = 0) \Big\} - \\
&\sup_{x \in X_{x0}} \Big\{ \sum_w \underline{y}^{x0w} Pr(W = w | X = x, Z = 0) \Big\}
\end{aligned} \tag{30}
$$

Using the lower and upper bounds obtained in equation (29) and plugging them to (27), then the refined identification region on $\mathbb{E}[Y|X]$ is as follows:

$$
\begin{aligned}
\mu(\kappa) \in \mathcal{M}(\kappa) &\equiv \bigcap_{x \in X_0} \mathcal{M}(X = x) \\
&\equiv \Bigg[\sup_{x \in X_0} \Big\{ \sup_{x \in X_{x0}} \Big\{ \sum_w \underline{y}^{x0w} Pr(W = w | X = x, Z = 0) \Big\} Pr(Z = 0 | X = x) + \\
&\qquad \mathbb{E}[Y | X = x, Z = 1] Pr(Z = 1 | X = x) \Big\}, \\
&\qquad \inf_{x \in X_0} \Big\{ \inf_{x \in X_{x0}} \Big\{ \sum_w \overline{y}^{x0w} Pr(W = w | X = x, Z = 0) \Big\} Pr(Z = 0 | X = x) + \\
&\qquad \mathbb{E}[Y | X = x, Z = 1] Pr(Z = 1 | X = x) \Big\} \Bigg]
\end{aligned} \tag{31}
$$

Bounds using level-set restrictions can be expanded to the treatment effects literature and we intend to address it in a later version of this paper where we will consider cases with and without imposing shape restrictions.

5 Concluding Remarks

Missing outcomes is a pervasive problem that arises in many situations hindering our capacity to recover population moments. For instance, missing outcomes appear as a result of survey nonresponse, attrition in longitudinal studies, or when population members do not appear in the sample frame. The latter problem is known as coverage bias and it arises from the exclusion of cellphone-only population in standard landline telephone surveys, for example. The literature has focused on the identifying power of shape restrictions assumptions which can be invoked in empirical studies. This paper propose a novel way to improve the identification regions by using adding covariates. Using the different sampling designs employed by the University of Michigan Index of Consumer Sentiment over time, we show that the identification region varies with the coverage bias.

In particular, our results show that the width of the identification region increased substantially between 2003 and 2012, when the undercovered population peaked. Nonetheless, by exploiting the additional restrictions imposed by the covariates, this region can be reduced up to 32.3% relative to the region spanned by the empirical bound $[54.3, 117.3]$ and up to 71.2% relative to the theoretical bound $[2, 150]$. Hence, by exploiting variation in the covariates we are able to tighten the bounds without imposing any shape restrictions. In Appendix C we extend our approach to the treatment effects literature and show that under some very mild conditions the bounds can be tightened. We intend to expand our approach to the treatment effects literature using level-set restrictions. Our partial identification approach which exploits variation across strata of the sample can improve the bounds and provides the researcher with information on the magnitude of undercoverage bias.

References

Barsky, Robert B and Eric R Sims. 2012. "Information, animal spirits, and the meaning of innovations in consumer confidence." *American Economic Review* 102(4):1343–77.

Benhabib, Jess and Mark M Spiegel. 2019. "Sentiments and economic activity: Evidence from US states." *The Economic Journal* 129(618):715–733.

Blanchard, Olivier. 1993. "Consumption and the Recession of 1990-1991." *The American Economic Review* 83(2):270–274.

Blumberg, Stephen J and Julian V Luke. 2006. "Wireless substitution: early release of estimates from the National Health Interview Survey, January–June 2006.".

Blumberg, Stephen J and Julian V Luke. 2010. "Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2010.".

Blumberg, Stephen J and Julian V Luke. 2014. "Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2014.".

Blumberg, Stephen J and Julian V Luke. 2018. "Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2018.".

Bram, Jason and Sydney C Ludvigson. 1998. "Does consumer confidence forecast household expenditure? A sentiment index horse race." *Economic Policy Review* 4(2).

Carroll, Christopher D, Jeffrey C Fuhrer and David W Wilcox. 1994. "Does consumer sentiment forecast household spending? If so, why?" *The American Economic Review* 84(5):1397–1408.

Curtin, Richard. 2007. "Consumer sentiment surveys." *Journal of Business Cycle Measurement and Analysis* 2007(1):7–42.

Cygan-Rehm, Kamila, Daniel Kuehnle and Michael Oberfichtner. 2017. "Bounding the causal effect of unemployment on mental health: Nonparametric evidence from four countries." *Health Economics* 26(12):1844–1861.

De Haan, Monique. 2011. "The effect of parents' schooling on child's schooling: a nonparametric bounds analysis." *Journal of Labor Economics* 29(4):859–892.

Ehlen, John and Patrick Ehlen. 2007. "Cellular-only substitution in the United States as lifestyle adoption: Implications for telephone survey coverage." *Public Opinion Quarterly* 71(5):717–733.

Fan, Yanqin and Sang Soo Park. 2010. "Sharp bounds on the distribution of treatment effects and their statistical inference." *Econometric Theory* 26(3):931–951.

Gerfin, Michael and Martin Schellhorn. 2006. "Nonparametric bounds on the effect of deductibles in health care insurance on doctor visits–Swiss evidence." *Health economics* 15(9):1011–1020.

Gillitzer, Christian and Nalini Prasad. 2018. "The effect of consumer sentiment on consumption: cross-sectional evidence from elections." *American Economic Journal: Macroeconomics* 10(4):234–69.

Gonzalez, Libertad. 2005. "Nonparametric bounds on the returns to language skills." *Journal of Applied Econometrics* 20(6):771–795.

Heckman, James J, Hidehiko Ichimura and Petra E Todd. 1997. "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme." *The review of economic studies* 64(4):605–654.

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

Kreider, Brent, John V Pepper, Craig Gundersen and Dean Jolliffe. 2012. "Identifying the effects of SNAP (food stamps) on child health outcomes when participation is endogenous and misreported." *Journal of the American Statistical Association* 107(499):958–975.

Lee, Sokbae and Ralf A Wilke. 2009. "Reform of unemployment compensation in Germany: A nonparametric bounds analysis using register data." *Journal of Business & Economic*

Statistics 27(2):193–205.

Manski, Charles F. 1989. “Anatomy of the selection problem.” *Journal of Human Resources* 24(3):343–360.

Manski, Charles F. 1990. “Nonparametric bounds on treatment effects.” *The American Economic Review* 80(2):319–323.

Manski, Charles F. 1995. *Identification problems in the social sciences.* Harvard University Press.

Manski, Charles F. 1997*a*. “The mixing problem in programme evaluation.” *The Review of Economic Studies* 64(4):537–553.

Manski, Charles F. 1997*b*. “Monotone Treatment Response.” *Econometrica* 65(6):1311–1334.

Manski, Charles F. 2003. *Mnotone Treatment Response.* Springer.

Manski, Charles F. 2005. “Partial identification with missing data: concepts and findings.” *International Journal of Approximate Reasoning* 39(2-3):151–165.

Manski, Charles F. 2009. *Identification for Prediction and Decision.* Harvard University Press.

Manski, Charles F and John V Pepper. 2000. “Monotone Instrumental Variables: With an Application to the Returns to Schooling.” *Econometrica* 68(4):997–1010.

Pepper, John V. 2000. “The intergenerational transmission of welfare receipt: A nonparametric bounds analysis.” *Review of Economics and Statistics* 82(3):472–488.

Wolter, Kirk, Sadeq Chowdhury and Jenny Kelly. 2009. Design, conduct, and analysis of random-digit dialing surveys. In *Handbook of Statistics.* Vol. 29 Elsevier pp. 125–154.

Appendix

Partial Identification with Covariates

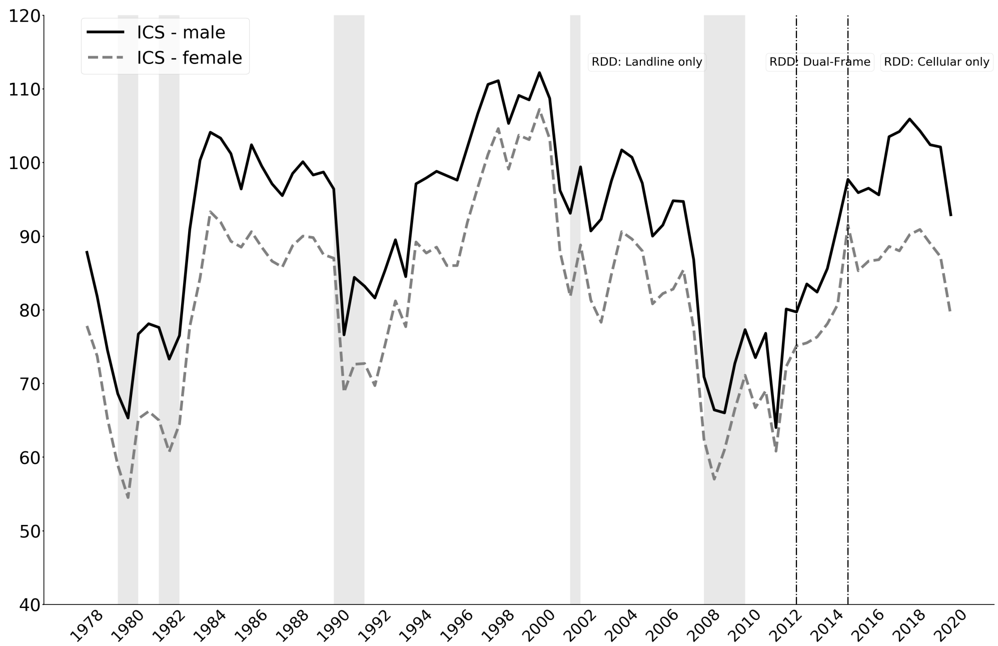
A Consumer Sentiment by Covariates

Table A1: Telephone Service Coverage and Consumer Confidence

Date	Landline with cellphone	Landline without cellphone	Landline with unknown cellphone	Nonlandline with unknown cellphone	Cellphone-only	Phoneless	UFCSI 6-month average	UMICS 6-month average
Jan - Jun 2003	45.70	40.40	9.40	0.20	2.80	1.60	86.00	84.60
Jul - Dec 2003	45.20	39.80	9.50	0.20	3.50	1.70	92.80	90.60
Jan - Jun 2004	46.90	36.30	10.40	0.50	4.40	1.50	95.20	95.70
Jul - Dec 2004	46.80	35.70	9.70	0.50	5.40	1.80	93.30	94.70
Jan - Jun 2005	46.10	31.50	13.50	0.70	6.70	1.60	93.30	92.10
Jul - Dec 2005	46.40	29.70	13.90	0.70	7.70	1.70	86.10	85.00
Jan - Jun 2006	49.50	28.20	10.40	0.60	9.60	1.80	88.90	86.40
Jul - Dec 2006	48.10	27.30	10.50	0.70	11.80	1.70	86.80	88.30
Jan - Jun 2007	63.30	20.80	1.70	0.10	12.60	1.60	85.90	89.60
Jul - Dec 2007	63.20	19.10	1.20	0.10	14.50	1.90	77.60	81.60
Jan - Jun 2008	63.00	17.90	0.80	0.00	16.10	2.10	66.80	66.30
Jul - Dec 2008	63.70	15.10	1.00	0.00	18.40	1.70	64.50	61.30
Jan - Jun 2009	63.50	13.40	0.40	0.00	21.10	1.50	67.40	63.20
Jul - Dec 2009	62.50	12.60	0.30	0.00	22.90	1.70	70.00	69.30
Jan - Jun 2010	62.20	10.90	0.30	0.00	24.90	1.70	72.00	73.90
Jul - Dec 2010	59.40	10.70	0.30	0.10	27.80	1.80	69.10	69.80
Jan - Jun 2011	58.80	9.00	0.20	0.00	30.20	1.80	71.20	72.50
Jul - Dec 2011	57.30	8.30	0.20	0.00	32.30	1.90	65.50	62.20
Jan - Jun 2012	56.10	7.80	0.20	0.00	34.00	1.90	75.50	75.90
Jul - Dec 2012	54.40	7.00	0.20	0.10	36.50	1.90	77.10	77.20
Jan - Jun 2013	52.80	6.90	0.10	0.00	38.00	2.20	77.60	79.20
Jul - Dec 2013	51.50	7.00	0.10	0.10	39.10	2.20	75.90	79.30
Jan - Jun 2014	47.30	7.00	0.10	0.10	43.10	2.40	79.10	81.90
Jul - Dec 2014	45.80	7.10	0.10	0.10	44.10	2.90	84.10	86.40
Jan - Jun 2015	43.90	6.20	0.10	0.00	46.70	3.10	92.80	94.90
Jul - Dec 2015	43.70	5.80	0.10	0.00	47.70	2.70	90.40	91.00
Jan - Jun 2016	42.10	5.80	0.10	0.00	49.00	2.90	91.50	92.00
Jul - Dec 2016	41.00	5.40	0.00	0.00	50.50	3.00	91.70	91.70
Jan - Jun 2017	39.60	4.80	0.10	0.00	52.00	3.40	96.10	96.80
Jul - Dec 2017	38.50	4.90	0.10	0.10	53.30	3.10	96.50	96.70
Jan - Jun 2018	37.40	4.10	0.00	0.10	55.20	3.20	98.80	98.60
Jul - Dec 2018	36.20	4.10	0.00	0.00	56.70	2.90	98.00	98.10

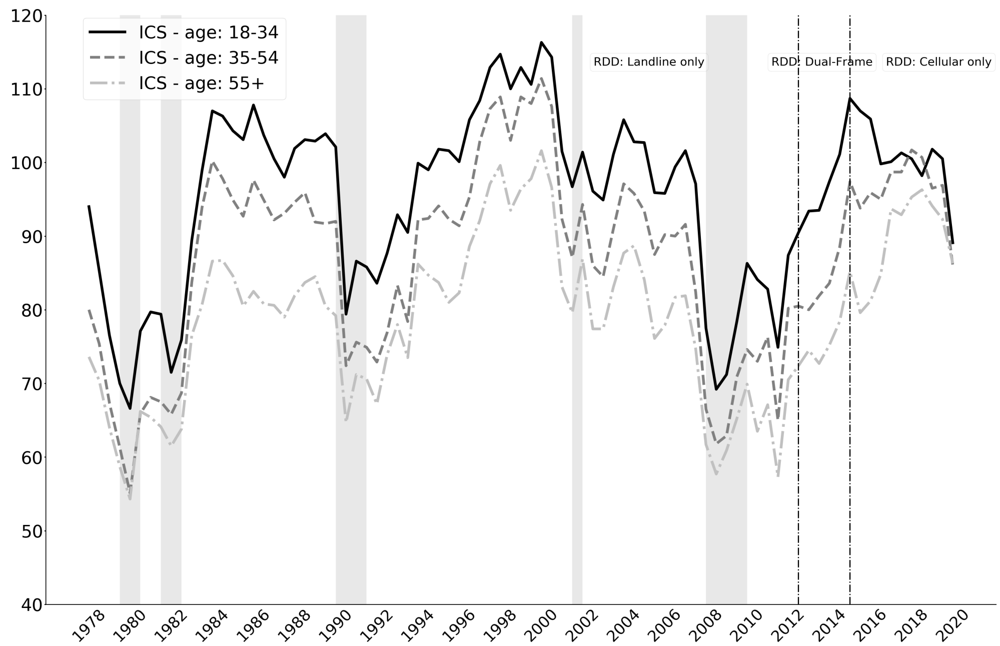
Source: Blumberg and Luke (2006, 2010, 2014, 2018), University of Michigan Survey Research Center and University of Florida Bureau of Economic and Business Research.

Figure 9: Consumer Sentiment by Gender



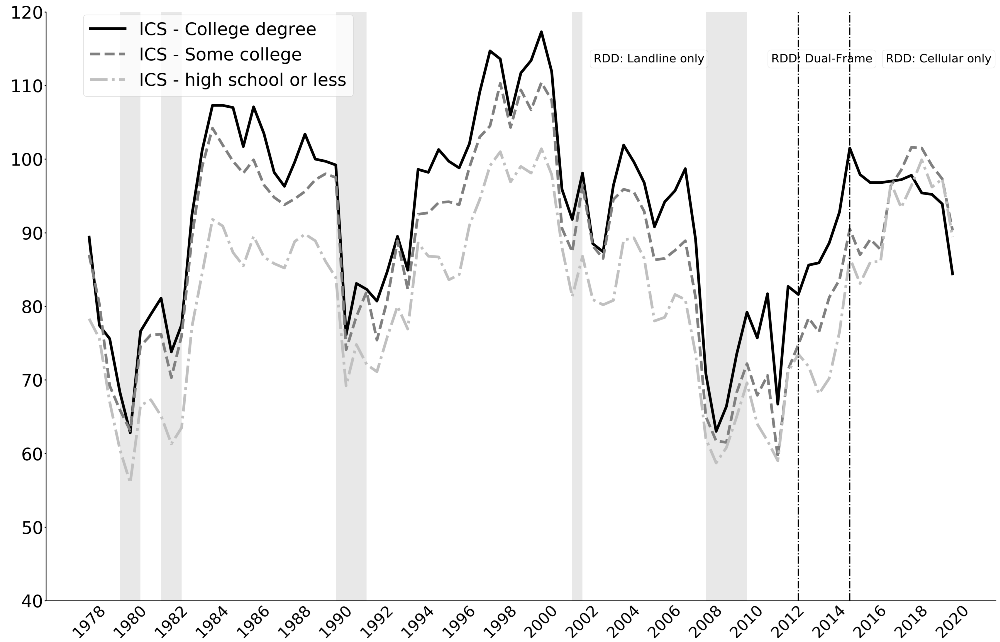
Source: University of Michigan Survey Research Center. Shaded areas denote NBER-defined recessions. The graph displays the breakdown of consumer sentiment by gender from 2003 to 2020. The consumer sentiment for males is higher than for females.

Figure 10: Consumer Sentiment by Age



Source: University of Michigan Survey Research Center. Shaded areas denotes NBER-defined recessions. The graph displays the breakdown of consumer sentiment by age groups from 2003 to 2020. As shown the age group 18-34 has a higher sentiment when compared with the other two age groups.

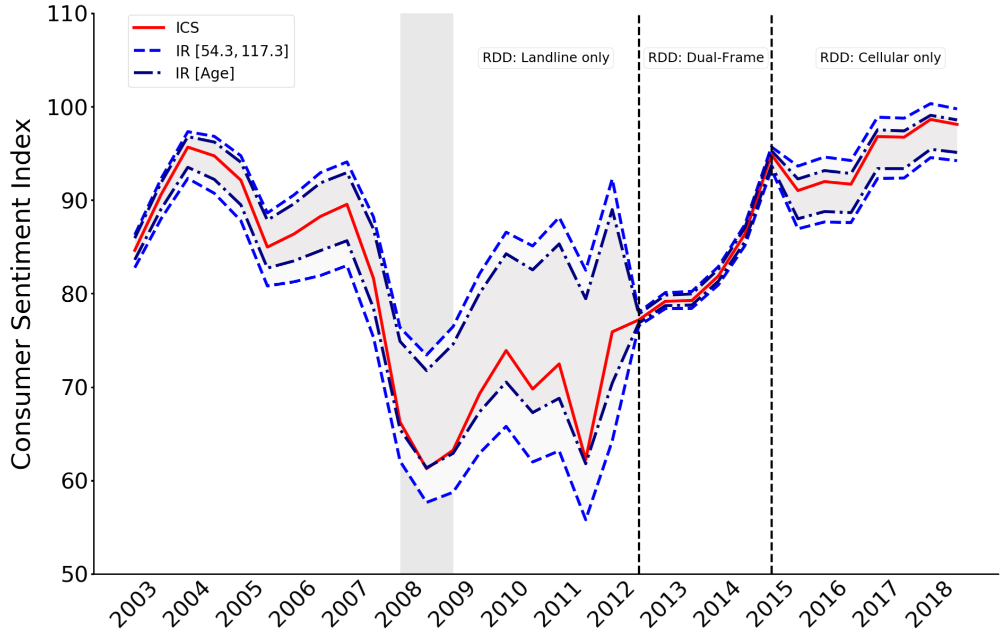
Figure 11: Consumer Sentiment by Education Levels



Source: University of Michigan Survey Research Center. Shaded areas denotes NBER-defined recessions. The graph displays the breakdown of consumer sentiment by education levels from 2003 to 2020. As shown consumers with college degree or higher have a higher sentiment when compared to consumers with some college degree or high school.

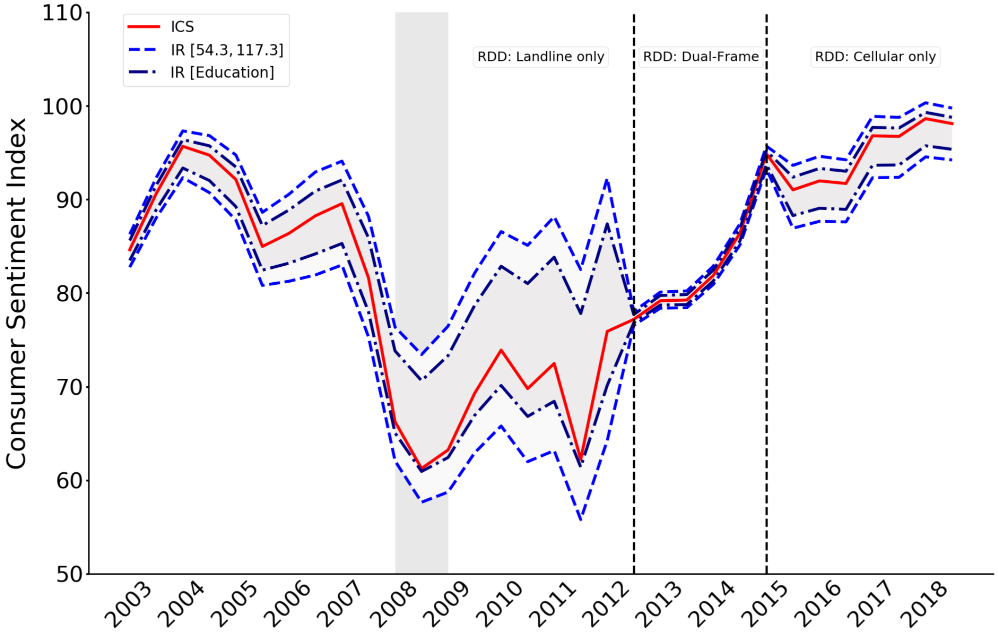
B Bound Improvement Using Different Covariates

Figure 12: Bounds using age as covariate



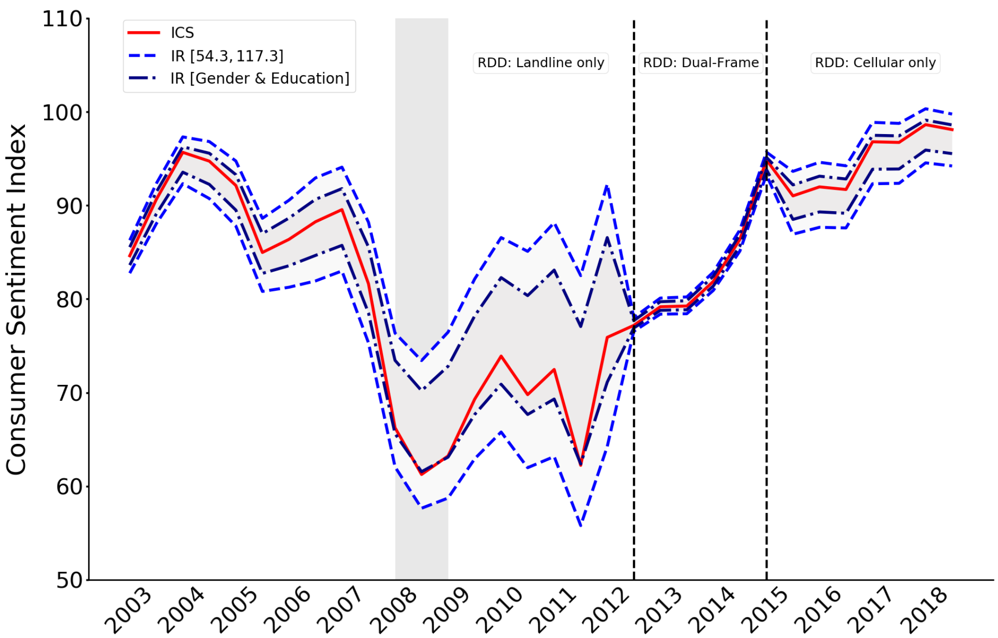
Note: Key to symbols: ICS = University of Michigan Index of Consumer Sentiment, and IR = Identification region conditioning on age. The shaded area denotes NBER-defined recession and the dashed vertical lines mark changes in the RDD sampling design.

Figure 13: Bounds using education as covariate



Note: Key to symbols: ICS = University of Michigan Index of Consumer Sentiment, and IR = Identification region conditioning on education. The shaded area denotes NBER-defined recession and the dashed vertical lines mark changes in the RDD sampling design.

Figure 14: Bounds using gender and education covariates



Note: Key to symbols: ICS = University of Michigan Index of Consumer Sentiment, and IR = Identification region conditioning on gender and education. The shaded area denotes NBER-defined recession and the dashed vertical lines mark changes in the RDD sampling design.

C Bounds on treatment effects

C.1 Treatment Effect Framework

An economic agent can be either be in the treated state or the untreated state, however he can not occupy both states at the same time. Let Y_1 be the potential outcome when the agent is in the treated state and Y_0 the potential outcome in the untreated state. The agent's gain or net utility from participating in the program is $\Delta = Y_1 - Y_0$ and the average treatment effect (ATE) is given by

$$\mathbb{E}[\Delta] = \mathbb{E}[Y_1 - Y_0]$$

where $\mathbb{E}[\Delta]$ exists and is finite a.e since Y_1 and Y_0 have finite first moments.

C.1.1 Nonparametric Bounds on ATE

Manski (1990) derived non-parametric bounds on the average treatment effect without imposing any assumptions or shape restrictions.

Assumption 6. *Suppose that the supports $\mathcal{Y}_1$ and $\mathcal{Y}_0$ of the potential outcomes Y_1 and Y_0, respectively, are bounded. That is, $\underline{y}^1 = \min(\mathcal{Y}_1)$, $\overline{y}^1 = \max(\mathcal{Y}_1)$ and $\underline{y}^0 = \min(\mathcal{Y}_0)$, $\overline{y}^0 = \max(\mathcal{Y}_0)$.*

Then, by the LIE we can compose both components of the average treatment effect as follows:

$$\mathbb{E}[Y_1] = \mathbb{E}[Y_1|Z=1]\Pr(Z=1) + \mathbb{E}[Y_1|Z=0]\Pr(Z=0) \tag{32}$$

$$\mathbb{E}[Y_0] = \mathbb{E}[Y_0|Z=1]\Pr(Z=1) + \mathbb{E}[Y_0|Z=0]\Pr(Z=0) \tag{33}$$

where $\mathbb{E}[Y_i|Z=1]$ for all $i = 0, 1$ exists and is finite a.e. $F_{Z=1}$. Under Assumption 6 the lower and upper bounds on Y_1 and Y_0 are as follows:

$$\mathrm{LB}(\mathbb{E}[Y_1]) = \mathbb{E}[Y_1|Z=1]\Pr(Z=1) + \underline{y}^1\Pr(Z=0) \tag{34}$$

$$\mathrm{UB}(\mathbb{E}[Y_1]) = \mathbb{E}[Y_1|Z=1]\Pr(Z=1) + \overline{y}^1\Pr(Z=0) \tag{35}$$

$$\mathrm{LB}(\mathbb{E}[Y_0]) = \underline{y}^0\Pr(Z=1) + \mathbb{E}[Y_0|Z=0]\Pr(Z=0) \tag{36}$$

$$\mathrm{UB}(\mathbb{E}[Y_0]) = \overline{y}^0\Pr(Z=1) + \mathbb{E}[Y_0|Z=0]\Pr(Z=0) \tag{37}$$

Then the bounds on $\mathbb{E}[\Delta]$ is as follows:

$$\begin{aligned}\mathcal{H}[\mathbb{E}[\Delta]] = \Big[&\mathbb{E}[Y_1|Z=1]\Pr(Z=1) + \underline{y}^1\Pr(Z=0) - \overline{y}^0\Pr(Z=1) - \mathbb{E}[Y_0|Z=0]\Pr(Z=0),\\ &\mathbb{E}[Y_1|Z=1]\Pr(Z=1) + \overline{y}^1\Pr(Z=0) - \underline{y}^0\Pr(Z=1) - \mathbb{E}[Y_0|Z=0]\Pr(Z=0)\Big]\end{aligned} \tag{38}$$

and the width of ATE is:

$$width(\mathbb{E}[\Delta]) = (\overline{y}^1 - \underline{y}^1)\Pr(Z=0) + (\overline{y}^0 - \underline{y}^0)\Pr(Z=1) \tag{39}$$

C.1.2 Improvement Using Covariates

We can further improve the bounds if we reduce the differences $(\overline{y}^1 - \underline{y}^1)$ and $(\overline{y}^0 - \underline{y}^0)$. We decompose $\mathbb{E}[Y_0|Z=1]$ and $\mathbb{E}[Y_1|Z=0]$ as follows

$$\mathbb{E}[Y_i|Z=0] = \mathbb{E}\Big[\,\mathbb{E}\big[Y_i|Z=0,W\big]\big|Z=0\Big] \quad \text{for} \quad i=0,1. \tag{40}$$

Assumption 7. *Conditional on $Z = 0$ and W, Y_0 and Y_1 respectively have bounded supports for each $W = w$. Let $\mathcal{Y}_{1w}$ and $\mathcal{Y}_{0w}$ denote each support respectively then $\underline{y}^{1w} = \min(\mathcal{Y}_{1w})$, $\overline{y}^{1w} = \max(\mathcal{Y}_{1w})$ and $\underline{y}^{0w} = \min(\mathcal{Y}_{0w})$, $\overline{y}^{0w} = \max(\mathcal{Y}_{0w})$.*

Under Assumption 7 the bounds on $\mathbb{E}[Y_0|Z = 1]$ and $\mathbb{E}[Y_1|Z = 0]$ are as follows:

$$\mathcal{H}\big[\,\mathbb{E}[Y_0|Z = 1]\big] = \left[\sum_w \underline{y}^{0w} \Pr(W = w|Z = 1), \sum_w \overline{y}^{0w} \Pr(W = w|Z = 1)\right] \tag{41}$$

$$\mathcal{H}\big[\,\mathbb{E}[Y_1|Z = 0]\big] = \left[\sum_w \underline{y}^{1w} \Pr(W = w|Z = 0), \sum_w \overline{y}^{1w} \Pr(W = w|Z = 0)\right] \tag{42}$$

Then the new bounds on $\mathbb{E}[Y_0]$ and $\mathbb{E}[Y_1]$:

$$\mathrm{LB}(\mathbb{E}[Y_1]) = \mathbb{E}[Y_1|Z = 1]\Pr(Z = 1) + \sum_w \underline{y}^{1w} \Pr(W = w, Z = 0) \tag{43}$$

$$\mathrm{UB}(\mathbb{E}[Y_1]) = \mathbb{E}[Y_1|Z = 1]\Pr(Z = 1) + \sum_w \overline{y}^{1w} \Pr(W = w, Z = 0) \tag{44}$$

$$\mathrm{LB}(\mathbb{E}[Y_0]) = \sum_w \underline{y}^{0w} \Pr(W = w, Z = 1) + \mathbb{E}[Y_0|Z = 0]\Pr(Z = 0) \tag{45}$$

$$\mathrm{UB}(\mathbb{E}[Y_0]) = \sum_w \overline{y}^{0w} \Pr(W = w, Z = 1) + \mathbb{E}[Y_0|Z = 0]\Pr(Z = 0) \tag{46}$$

The new identification region of $\mathbb{E}[\Delta]$ is:

$$\begin{aligned}\mathcal{H}[\Delta] = \Bigg[\, &\mathbb{E}[Y_1|Z = 1]\Pr(Z = 1) - \mathbb{E}[Y_0|Z = 0]\Pr(Z = 0) + \\ &\qquad \sum_w \big[\underline{y}^{1w} \Pr(W = w, Z = 0) - \overline{y}^{0w} \Pr(W = w, Z = 1)\big], \\ &\mathbb{E}[Y_1|Z = 1]\Pr(Z = 1) - \mathbb{E}[Y_0|Z = 0]\Pr(Z = 0) + \\ &\qquad \sum_w \big[\overline{y}^{1w} \Pr(W = w, Z = 0) - \underline{y}^{0w} \Pr(W = w, Z = 1)\big]\Bigg]\end{aligned} \tag{47}$$

The width of the new identification region is:

$$width(\mathbb{E}[\Delta]) = \sum_{w} \left[(\overline{y}^{1w} - \underline{y}^{1w}) \Pr(W = w, Z = 0) + (\overline{y}^{0w} - \underline{y}^{0w}) \Pr(W = w, Z = 1) \right] \quad (48)$$

Assumption 8. $\underline{y}^{i} \leqslant \underline{y}^{iw} < \overline{y}^{iw} \leqslant \overline{y}^{i}$ *for* $i \in \{0, 1\}$ *and for all* w.

Proposition 6. *If at least one of the inequalities in Assumption 8 holds than the width of the new bounds is tighter.*

C.1.3 Heterogeneous Treatment Effects

Let the assignment to treatment D be a binary variable and let the potential outcomes Y_1 and Y_0 be continuous outcomes. Let $F_1(\cdot)$ and $F_0(\cdot)$ be the marginal distributions of Y_1 and Y_0.

Theorem C.1. *Sklar Theorem (1959): Let* F *be the distribution function with univariate marginal distribution functions* F_0 *and* F_1*, then there exists a copula* $C(a, b) : (a, b) \in [0, 1] \times [0, 1]$ *such that*

$$F(y_0, y_1) = C(F_1(y_0), F_0(y_1))$$

for all y_0, y_1*.*

When the marginal distribution functions F_0 and F_1 are continuous then C is the unique copula of F and it characterizes its dependence structure, otherwise C is only uniquely determined only on $ran(F_0) \times ran(F_1)$, where $ran(F_i)$ denotes the range of the cumulative distribution function F_i.

For $(u, v) \in [0, 1] \times [0, 1]$ then the Fréchet-Hoeffding lower and upper bounds for the copula are:

$$\max(u + v - 1, 0) \leqslant C(u, v) \leqslant \min(u, v)$$

Hence for any (y_0, y_1) we have

$$\max(F_0(y_0) + F_1(y_1) - 1, 0) \leqslant F(y_0, y_1) \leqslant \min(F_0(y_0), F_1(y_1))$$

The lower and upper bounds are the Fréchet-Hoeffding lower and upper bounds for the bivariate distributions with fixed marginal distribution functions F_0 and F_1. Heckman, Ichimura and Todd (1997) and Manski (1997*a*) applied this result in the treatment effects literature.

C.1.4 Bounds on the Distribution of Treatment Effects with Covariates

We know consider bounds on the distribution of treatment effects by exploiting variation in the covariates. We build our work on Fan and Park (2010) who provided bounds on the distribution of treatment effects. However, different from Fan and Park (2010) we do not impose their assumption $C1$ that requires that the potential outcomes (Y_0, Y_1) to be jointly independent of the treatment assignment D conditional on covariates X. Then

$$\begin{aligned} F_1(y|x) &= \Pr(Y_1 \leqslant y|X = x) \\ &= \Pr(Y_1 \leqslant y|X = x, Z = 1)\Pr(Z = 1|X = x) + \Pr(Y_1 \leqslant y|X = x, Z = 0)\Pr(Z = 0|X = x) \end{aligned} \tag{49}$$

where $\Pr(Y_1 \leqslant y|X = x, Z = 0)$ is unobserved, but we can rewrite as follows

$$\begin{aligned} \Pr(Y_1 \leqslant y|X = x, Z = 0) &= \frac{\Pr(Y_1 \leqslant y, X = x, Z = 0)}{\Pr(X = x, Z = 0)} \\ &= \frac{\Pr(Y_1 \leqslant y, X = x) - \Pr(Y_1 \leqslant y, X = x, Z = 1)}{\Pr(X = x, Z = 0)} \\ &= \left[\frac{1}{1 - p(x)}\right]\left[F_1(y|x) - p(x)F_Y(y|X = x, Z = 1)\right] \end{aligned} \tag{50}$$

and in a similar fashion we have

$$\begin{aligned}
\Pr(Y_0 \leqslant y | X = x, Z = 1) &= \frac{\Pr(Y_0 \leqslant y, X = x, Z = 1)}{\Pr(X = x, Z = 1)} \\
&= \frac{\Pr(Y_0 \leqslant y, X = x) - \Pr(Y_0 \leqslant y, X = x, Z = 0)}{\Pr(X = x, Z = 1)} \\
&= \left[\frac{1}{1 - p(x)}\right]\left[F_1(y|x) - (1 - p(x))F_Y(y|X = x, Z = 0)\right]
\end{aligned} \tag{51}$$

where $p(x) = \Pr(Z = 1|X = x)$ and $Y = ZY_1 + (1 - Z)Y_0$. Combining equations 50, 51, and Theorem C.1 one can obtain the upper and lower bounds on distribution of the average treatment effects.