# Testing Missing at Random using Instrumental Variables

## Christoph Breunig*

* Humboldt-Universität zu Berlin, Germany

SFB 649

ECONOMIC RISK

BERLIN

# Testing Missing at Random using Instrumental Variables [*]

CHRISTOPH BREUNIG [⋆]

*Humboldt-Universität zu Berlin*

October 27, 2015

This paper proposes a test for *missing at random* (MAR). The MAR assumption is shown to be testable given instrumental variables which are independent of response given potential outcomes. A nonparametric testing procedure based on integrated squared distance is proposed. The statistic's asymptotic distribution under the MAR hypothesis is derived. We demonstrate that our results can be easily extended to a test of *missing completely at random* (MCAR). A Monte Carlo study examines finite sample performance of our test statistic. An empirical illustration concerns pocket prescription drug spending with missing values; we reject MCAR but fail to reject MAR.

*Keywords:* Incomplete data, missing-data mechanism, selection model, nonparametric hypothesis testing, consistent testing, instrumental variable, series estimation.

*JEL classification:* C12, C14

# 1. Introduction

When confronted with data sets with missing values it is often assumed in applied research that observations are *missing at random* (MAR) in the sense of Rubin [1976]. This condition requires that the probability of observing potential outcomes only depends on observed data. To help to decide whether MAR based techniques could be applied we develop in this paper a test for the MAR assumption. In general, MAR is not refutable without further assumptions and here we rely on instruments that are independent of the response mechanism given potentially observed outcomes. We show that this condition is sufficient to ensure testability of MAR and derive the asymptotic distribution under MAR of a proposed test statistic. We provide an extension of our testing procedure to asses the hypothesis *missing completely at random* (MCAR).

If the missing data mechanism does not follow MAR, a correction of the potential selection bias is necessary to ensure consistency of the estimation procedure. There exist two different instrumental variable approaches to overcome the problem of missing variables. The first approach relies on instruments that determine response but not the outcome and was pioneered by Heckman [1974]. Such instruments, however, are difficult to find when response is directly driven by the outcome. The second approach, also considered in this paper, relies on instruments that are independent of response given potential outcomes. This framework was used in parametric regression analysis by Chen [2001], Liang and Qin [2000], Tang et al. [2003], and Ramalho and Smith [2013]. A nonparametric extension was proposed by D'Haultfoeuille [2010] and Breunig et al. [2014]. While such instrumental variable methods reduce bias in general, if the data are MAR, they unnecessarily increase variance. Indeed, D'Haultfoeuille [2010] showed that estimation of the distribution of the potential outcome leads to a statistical inverse problem that is *ill-posed* in general. This implies that the variance of the estimator becomes arbitrarily large relative to the degree of ill-posedness.

We also provide a test for the MCAR assumption which imposes a stronger condition on the response mechanism than MAR. Indeed, MCAR rules out any correlation between response and outcome. When data are MAR but not MCAR various types of correction methods have been suggested so far and include weighted generalized estimating equations (Robins et al. [1994]), nonparametric estimation of the conditional estimating scores (Reilly and Pepe [1995]), and multiple imputation (Rubin [2004], Little and Rubin [2002]). For an overview and further references we refer to Ibrahim et al. [2005]. This literature makes either parametric model assumptions or has difficulties in dealing with continuous data. Using such correction methods reduces bias if MAR holds, under MCAR, however, this unnecessarily increases variance. Thus, it is of interest to examine the observed data for evidence whether the response mechanism satisfies not only MAR but also MCAR.

We show that the MAR hypothesis is equivalent to an identified conditional

moment equation. Based on this moment equation we construct our test statistic using a weighted integrated squared distance. Under the null hypothesis the test statistic converges to a series of independent, $\chi^2$–squared distributed random variables. The test statistic and its critical values can be easily implemented. Also only a slight modification is necessary to obtain a test for MCAR. Under a bounded completeness assumption, our testing procedure is shown to be consistent against fixed alternatives.

Besides a Monte Carlo simulation we demonstrate the finite sample properties in an empirical illustration using data from the Health and Retirement Study. In this survey, a fraction of participants does not report their exact expenditure of pocket prescription drugs. The assumption of MAR/MCAR seems problematic here as whether participants recall their exact expenditure might be related to the amount of expenditure itself. Using income as instrument we show that our test rejects the MCAR assumption; but fails to reject MAR.

In our instrumental variable framework, a test of MCAR has been proposed by Ramalho and Smith [2013]. Their Hausman type test statistic relies on a parametric model specification with discrete outcomes and differs form our method where no restriction on the marginal distribution of the outcome is imposed. Likelihood ratio tests to verify the hypothesis MCAR have been suggested by Fuchs [1982] and Little [1988], while Chen and Little [1999] considered a Wald-type test and Qu and Song [2002] proposed a generalized score type test based on quadratic inference functions. Kline and Santos [2013] develop a method for assessing the sensitivity of empirical conclusions to departures from MAR based on sharp bounds of conditional quantiles. As far as we know, a consistent test for MAR has not been proposed. We further emphasize that our testing procedure does not require knowledge of the conditional probability of observing potential outcomes up to a finite dimensional parameter.

The remainder of the paper is organized as follows. Section 2 provides sufficient conditions for testability of MAR and MCAR. The asymptotic distributions of the tests are derived and their consistency against fixed alternatives is established. Section 3 examines the finite sample performance of our test in a Monte Carlo simulations study while Section 4 illustrates the usefulness of our procedure in an empirical application.

## 2. The Test Statistic and its Asymptotic Properties

This section is about testability of missing at random assumptions and the asymptotic behavior of our proposed test statistics. First, we provide sufficient conditions on instruments to ensure testability of MAR and MCAR. Second, we build on identified conditional moment restrictions to construct

test statistics. Third, the test statistics' asymptotic distributions under the null hypotheses are derived and we establish consistency of the tests against fixed alternatives.

## 2.1. Testability

Let $Y^*$ denote a scalar, partially observed random variable and $X$ a $d_x$–dimensional vector of covariates which are always observed. Further, $\Delta$ is a binary missing–data indicator for $Y^*$, such that $\Delta = 1$ if a realization of $Y^*$ is observed and $\Delta = 0$ otherwise. We write $Y = \Delta Y^*$.

First, we consider hypothesis MAR, whether the response mechanism only depends on observed variables $X$. In this case, the null hypothesis under consideration is given by

$$MAR : \mathbb{P}(\Delta = 1 | Y^*, X) = \mathbb{P}(\Delta = 1 | X) \tag{2.1}$$

and the alternative by $\mathbb{P}\big(\mathbb{P}(\Delta = 1 | Y^*, X) = \mathbb{P}(\Delta = 1 | X)\big) < 1.$[1]

Second, we consider the MCAR hypothesis whether response is completely at random. As this hypothesis rules out any correlation between response and observed data, MCAR is stronger than MAR. The hypothesis under consideration is

$$MCAR : \mathbb{P}(\Delta = 1 | Y^*, X) = \mathbb{P}(\Delta = 1)$$

and the alternative is $\mathbb{P}\big(\mathbb{P}(\Delta = 1 | Y^*, X) = \mathbb{P}(\Delta = 1)\big) < 1.$

We now provide sufficient conditions for testability of the above hypotheses. A key requirement is that an additional vector $W$, an instrument, is available which satisfies the following conditions.

ASSUMPTION 1. *For each unit we observe $\Delta$, $Y$, $X$, and $W$.*

Assumption 1 is satisfied when only observations of $Y^*$ are missing. In the following, we assume that the random vector $W$ is independent of the response variable conditional on potentially observed variables $Y^*$ and covariates $X$.

ASSUMPTION 2. *It holds*

$$\Delta \perp\!\!\!\perp W \,|\, (Y^*, X).$$

Assumption 2 requires missingness to be primarily determined by $(Y^*, X)$. In particular, this exclusion restriction requires any influence of $W$ on $\Delta$ to be carried solely through $(Y^*, X)$. Conditional independence assumptions of this type are quite familiar in the econometrics and statistics literature. Examples are treatment effects (cf. Imbens [2004]) or non-classical measurement error

---

[1]Since conditional probabilities/expectations are defined only up to equality a.s., all equalities with conditional probabilities/expectations are understood as equalities a.s., even if we do not say so explicitly.

(cf. Hu and Schennach [2008]). For further discussion of Assumption 2 and illustrative examples we refer to Ramalho and Smith [2013] in case of nonresponse and D'Haultfoeuille [2010] in case of counterfactuals. As was shown by D'Haultfoeuille [2010], Assumption 2 is equivalent to a conditional moment restriction (see also Example 2.1 below) and thus testable under a completeness assumption (see Section 2.2 of D'Haultfoeuille [2010]).

ASSUMPTION 3. *For all bounded measurable functions $\phi$, $\mathbf{E}[\phi(Y^*, X)|X, W] = 0$ implies that $\phi(Y^*, X) = 0$.*

Assumption 3 is known as bounded completeness. In contrast, to ensure identification in nonparametric instrumental variable models, stronger versions of Assumption 3, such as $L^2$–completeness, are required. This type of completeness condition requires Assumption 3 to hold for any measurable function $\phi$ with $\mathbf{E}|\phi(Y^*, X)|^2 < \infty$. $L^2$–completeness is also a common assumption in nonparametric hypothesis testing in instrumental variable models, see, for instance, Blundell and Horowitz [2007] or Fève et al. [2012]. There are only a few examples in the nonparametric instrumental regression literature where it is sufficient to assume completeness only for bounded functions. One example is estimation of Engel curves as in Blundell et al. [2007] which, by definition, are bounded between zero and one. We emphasize that bounded completeness is much less restrictive than $L^2$ completeness. Sufficient conditions for bounded completeness have been provided by Mattner [1993] or D'Haultfoeuille [2011] among others. We see below that inference under the considered hypotheses does not require bounded completeness. On the other hand, we need to impose Assumption 3 to ensure consistency against fixed alternatives.

If a valid instrumental variable $W$ is available then consistent density estimation and regression is possible even if MAR does not hold true. On the other hand, using instrumental variable estimation methods when MAR holds can be inappropriate as the following two examples illustrate.

EXAMPLE 2.1. [Density Estimation] The joint probability density function of $(Y^*, X)$ satisfies

$$p_{Y^*X}(y, x) = \frac{\mathbb{P}(\Delta = 1, Y^* = y, X = x)}{\mathbb{P}(\Delta = 1|Y^* = y, X = x)}$$

assuming that the conditional probability in the denominator is bounded away from zero. The conditional probability $\mathbb{P}(\Delta = \delta|Y^* = y, X = x)$ is not identified in general. On the other hand, if instrumental variables $W$ are available, satisfying more restrictive completeness assumptions than Assumption 3, then this probability is identified (see D'Haultfoeuille [2010]) through the conditional moment restriction

$$\mathbf{E}\left(\frac{\Delta}{\mathbb{P}(\Delta = 1|Y^*, X)}\Big|X, W\right) = 1. \tag{2.2}$$

Estimating $\mathbb{P}(\Delta = 1|Y^*, X)$ via this equation leads to a large variance relative to the ill-posedness of the underlying inverse problem and the accuracy of this estimator can be very low (see D'Haultfoeuille [2010] and Breunig et al. [2014]). If the data, however, reveals that MAR holds true then $\mathbb{P}(\Delta = 1|Y^*, X)$ can be directly estimated from the data. $\qquad\square$

EXAMPLE 2.2 (Regression under Univariate Nonresponse). Consider estimation of $\mathbf{E}(\phi(Y^*)|X)$ for some known function $\phi$ and $Y^*$ scalar. Either $\phi$ is the identity function in case of mean regression or $\phi(Y^*) = \mathbb{1}\{Y^* \leqslant q\}$ in quantile regression for some quantile $q \in (0, 1)$. Let the conditional probability $\mathbb{P}(\Delta = 1|Y^*, X)$ be bounded away from zero. As in Breunig et al. [2014] (p. 5) it holds

$$\mathbf{E}(\phi(Y^*)|X) = \mathbf{E}\left(\frac{\Delta\phi(Y^*)}{\mathbb{P}(\Delta = 1|Y^*, X)}\Big|X\right)$$

where $\mathbb{P}(\Delta = 1|Y^*, X)$ can be estimated via the conditional mean restriction (2.2). As shown in Breunig et al. [2014], the first step estimation of $\mathbb{P}(\Delta = 1|Y^*, X)$ leads to an additional bias term which can reduce accuracy of estimation. Also in this case, imposing MAR is desirable to simplify the estimation procedure and increase estimation precision. $\qquad\square$

EXAMPLE 2.3 (Relation to Triangular Models). Assumptions 2 and 3 hold true in the triangular model

$$\Delta = \varphi(Y^*, X, \eta) \quad \text{with} \quad \eta \perp\!\!\!\perp (W, \varepsilon)$$
$$Y^* = \phi(\psi(X, W) + \varepsilon) \quad \text{with} \quad W \perp\!\!\!\perp \varepsilon$$

under a large support condition of $\psi(X, W)$, regularity assumptions for $\varepsilon$, and if the conditional characteristic function of $\varepsilon$ given $X$ is infinitely often differentiable and does not vanish on the real line. See D'Haultfoeuille [2011] page 462–463 for further details. Requiring this characteristic function to be nonvanishing is a standard assumption in the deconvolution literature. The normal, Student, $\chi^2$, gamma, and double exponential distributions all satisfy this assumption while the uniform and the triangular distributions are the only common distributions to violate this restriction.
In this triangular model, MCAR requires the function $\varphi$ in the selection equation to be dependent on $X$ and $\eta$ only; that is, $\Delta = \varphi(X, \eta)$. Under MCAR, $\varphi$ depends neither on $Y^*$ nor on $X$ and hence, the structural equation simplifies to $\Delta = \varphi(\eta)$. The triangular model illustrates the difference to Heckman's approach (cf. its nonparametric version in Das et al. [2003]) where an instrument enters the selection equation. $\qquad\square$

The following result states that the null hypothesis MAR is testable under the previous conditions. Further, exploiting the properties of the instrument $W$ shows that MAR is equivalent to an identified conditional moment restriction.

THEOREM 2.1. *Under Assumptions 1–3 the null hypothesis MAR is testable.*

*Proof.* Due to Assumption 3 the null hypothesis MAR is equivalent to

$$\mathbf{E}\left[\mathbb{P}(\Delta = 1|Y^*, X) - \mathbb{P}(\Delta = 1|X)\big|X, W\right] = 0.$$

By Assumption 2 we have $\mathbb{P}(\Delta = 1|Y^*, X) = \mathbb{P}(\Delta = 1|Y^*, X, W)$ and thus, the law of iterated expectations yields $\mathbf{E}[\mathbb{P}(\Delta = 1|Y^*, X)|X, W] = \mathbf{E}[\Delta|X, W]$. Hence, hypothesis MAR is equivalent to

$$\mathbf{E}\left[\Delta - \mathbb{P}(\Delta = 1|X)\big|X, W\right] = 0$$

where the left hand side is point identified. $\qquad\square$

The following corollary provides a testability result for the hypothesis MCAR. The result follows as in the proof of Theorem 2.1 by replacing $\mathbb{P}(\Delta = 1|X)$ with $\mathbb{P}(\Delta = 1)$.

COROLLARY 2.2. *Let Assumptions 1–3 hold true. Then the null hypothesis MCAR is equivalent to* $\mathbf{E}[\Delta - \mathbb{P}(\Delta = 1)|X, W] = 0$ *and hence, is testable.*

## 2.2. The Test Statistic

In the previous section, we observed that each null hypothesis is equivalent to a conditional moment restriction

$$\mathbf{E}\left[r(\Delta, X)\big|X, W\right] = 0$$

for some bounded function $r$. Equivalently, by considering the squared integrated distance we obtain

$$\int \mathbf{E}\left[r(\Delta, X)\big|X = x, W = w\right]^2 \pi(x, w)d(x, w) = 0$$

for some weight function $\pi$ which is strictly positive almost surely (a.s.) on $\mathcal{X} \times \mathcal{W}$ ($\mathcal{X}$ and $\mathcal{W}$ denote the supports of $X$ and $W$, respectively). Let $p_{XW}$ denote the joint probability density function of $(X, W)$. Further, let $\nu$ be an a.s. strictly positive density function on $\mathcal{X} \times \mathcal{W}$. Let us introduce approximating functions $\{f_j\}_{j \geqslant 1}$ which are assumed to form an orthonormal basis in the Hilbert space $L_\nu^2 := \left\{\phi : \int |\phi(x, w)|^2 \nu(x, w)d(x, w) < \infty\right\}$. Now choosing $\pi(x, w) = p_{XW}^2(x, w)/\nu(x, w)$ together with Parseval's identity yields

$$0 = \int \left|\mathbf{E}\left[r(\Delta, X)\big|X = x, W = w\right]p_{XW}(x, w)/\nu(x, w)\right|^2 \nu(x, w)d(x, w)$$

$$= \sum_{j=1}^{\infty} \left(\mathbf{E}\left[r(\Delta, X)f_j(X, W)\right]\right)^2.$$

Given a strictly positive sequence of weights $(\tau_j)_{j \geqslant 1}$ the last equation is equivalent to

$$\sum_{j=1}^{\infty} \tau_j \left( \mathbf{E}\left[ r(\Delta, X) f_j(X, W) \right] \right)^2 = 0. \tag{2.3}$$

Our test statistics below are based on an empirical version of the left hand side of (2.3). To do so, we truncate the infinite sum at some finite integer. Below we choose $(\tau_j)_{j \geqslant 1}$ to be a strictly increasing which implies that we reduce weight to those generalized Fourier coefficients as basis functions are becoming more nonlinear. Additional weighting of the testing procedure was also used by Horowitz [2006], Blundell and Horowitz [2007], and Breunig [2015].

Our test statistic is based on an empirical analog of the left hand side of (2.3) given $(\Delta_1, Y_1, X_1, W_1), \ldots, (\Delta_n, Y_n, X_n, W_n)$ of independent and identical distributed (iid.) copies of $(\Delta, Y, X, W)$ where $Y = \Delta Y^*$. Let us introduce the notation for the conditional probability function $h(\cdot) = \mathbb{P}(\Delta = 1 | X = \cdot)$. We estimate $h$ by the series least square estimators

$$\widehat{h}_n(\cdot) = e_{\underline{k_n}}(\cdot)^t \left( \mathbf{X}_n^t \mathbf{X}_n \right)^- \mathbf{X}_n^t \Delta_n$$

where $\Delta_n = (\Delta_1, \ldots, \Delta_n)$, $e_{\underline{k_n}}(\cdot) := (e_1(\cdot), \ldots, e_{k_n}(\cdot))^t$ is a vector of basis functions, and $\mathbf{X}_n := \left( e_{\underline{k_n}}(X_1), \ldots, e_{\underline{k_n}}(X_n) \right)^t$. In the multivariate case, we consider a tensor-product linear sieve basis, which is the product of univariate linear sieves. The dimension parameter $k_n$ increases with sample size $n$.

Consider the null hypothesis MAR. From the proof of Theorem 2.1, we deduce $r(\Delta, X) = \Delta - h(X)$. Replacing $h$ by the proposed estimator $\widehat{h}_n$ we obtain the test statistic

$$S_n^{MAR} = \sum_{j=1}^{m_n} \tau_j \left| n^{-1} \sum_{i=1}^{n} \left( \Delta_i - \widehat{h}_n(X_i) \right) f_j(X_i, W_i) \right|^2 \tag{2.4}$$

where $m_n$ increases with sample size $n$ and $(\tau_j)_{j \geqslant 1}$ is a strictly positive sequence of weights which is nonincreasing. We reject null hypothesis MAR if the test statistic $S_n^{MAR}$ becomes too large.

For the null hypothesis MCAR, Corollary 2.2 gives $r(\Delta, X) = \Delta - \mathbb{P}(\Delta = 1)$. Again, following the derivation of the statistic $S_n^{MAR}$ we obtain a statistic for MCAR given by

$$S_n^{MCAR} = \sum_{j=1}^{m_n} \tau_j \left| n^{-1} \sum_{i=1}^{n} \left( \Delta_i - \widehat{\Delta}_n \right) f_j(X_i, W_i) \right|^2 \tag{2.5}$$

where $\widehat{\Delta}_n = n^{-1} \sum_{i=1}^{n} \Delta_i$. We reject MCAR if the test statistic $S_n^{MCAR}$ becomes too large.

## 2.3. Assumptions for Inference

In the following, $\mathcal{Y}$, $\mathcal{X}$, and $\mathcal{W}$ denote the supports of $Y$, $X$, and $W$, respectively. The usual Euclidean norm is denoted by $\|\cdot\|$ and $\|\cdot\|_\infty$ is the supremum norm.

ASSUMPTION 4. *(i) We observe a sample* $(\Delta_1, Y_1, X_1, W_1), \ldots, (\Delta_n, Y_n, X_n, W_n)$ *of independent and identical distributed copies of* $(\Delta, Y, X, W)$ *where* $Y = \Delta Y^*$. *(ii) The functions* $\{f_j\}_{j \geqslant 1}$ *form an orthonormal basis in* $L^2_\nu$. *(iii) There exists some constant* $C > 0$ *such that* $\sup_{(x,w) \in \mathcal{X} \times \mathcal{W}} \{p_{XW}(x,w)/\nu(x,w)\} \leqslant C$.

In our simulations, we used trigonometric basis functions or orthonormalized Hermite polynomials where Assumption 4 (*ii*) is automatically satisfied if, respectively, $\nu$ is Lebesque measure on $[0,1]$ or $\nu$ is the standard normal density. Assumption 4 (*iii*) is a mild restriction on the density of $(X, W)$ relative to $\nu$. Assumption 4 implies $\mathbf{E}|f_j(X, W)|^2 \leqslant C$. The next assumption involves the linear sieve space $\mathcal{H}_n := \left\{ \phi : \phi(\cdot) = \beta_n' \underline{e_{k_n}}(\cdot) \text{ where } \beta_n \in \mathbb{R}^{l_n} \right\}$ where the dimension of the sieve space $k_n$ increases with sample size $n$.

ASSUMPTION 5. *(i) There exists* $E_{k_n} h \in \mathcal{H}_n$ *such that* $\|E_{k_n} h - h\|^2_\infty = O(1/\gamma_{k_n})$ *for some increasing sequence* $(\gamma_l)_{l \geqslant 1}$. *(ii) It holds* $\sup_{x \in \mathcal{X}} \|\underline{e_{k_n}}(x)\|^2 = O(k_n)$ *such that* $k_n^2 \log(n) = o(n)$. *(iii) The smallest eigenvalue of* $\mathbf{E}[\underline{e_{k_n}}(X)\underline{e_{k_n}}(X)^t]$ *is bounded away from zero uniformly in* $n$.

Assumption 5 (*i*) determines the sieve approximation error for estimating the function $h$ in the supremum norm and is used to control the bias of the estimator of $h$. For the relation to $L^2$ approximation conditions see Belloni et al. [2015] p. 10–16 or Chen and Christensen [2015]. An excellent review of approximating properties of different sieve bases is given in Chen [2007]. Assumption 5 (*ii*) and (*iii*) restrict the magnitude of the approximating functions $\{e_j\}_{j \geqslant 1}$ and impose nonsingularity of their second moment matrix (cf. Newey [1997]).

Consistent estimation of critical values requires the following additional assumption where we use the notation $\psi_j(\cdot) = \mathbf{E}[f_j(X, W)|X = \cdot]$.

ASSUMPTION 6. *There exists* $E_{k_n} \psi_j \in \mathcal{H}_n$ *such that* $\max_{1 \leqslant j \leqslant m_n} \|E_{k_n}\psi_j - \psi_j\|^2_\infty = O(1/\gamma_{k_n})$.

Assumption 6 ensures that the basis functions $\{f_j\}_{1 \leqslant j \leqslant m_n}$ can be as well approximated as $h$ by using the sieve space $\mathcal{H}_n$. Below we choose for $f_j$ Hermite polynomials or cosine basis functions which, as analytic functions, can be sufficiently well approximated.

## 2.4. Asymptotic Distribution under MAR

Before establishing the asymptotic distribution of the test statistic $S_n^{MAR}$ under MAR, we require the following definition. For any realization $(\delta, x, w)$ of $(\Delta, X, W)$, let $\varepsilon(\delta, x, w)$ be an infinite dimensional vector with $j$-th entry given

by

$$\varepsilon_j(\delta, x, w) = \sqrt{\tau_j}\Big(\delta - h(x)\Big)\Big(f_j(x, w) - \sum_{l=1}^{\infty} \mathbf{E}\Big[f_j(X, W)e_l(X)\Big]e_l(x)\Big).$$

It holds $\mathbf{E}[\varepsilon(\Delta, X, W)] = 0$ under MAR. We assume $\mathbf{E}\,|\varepsilon_j(\Delta, X, W)|^2 < \infty$, which is automatically satisfied if $\{e_l\}_{l \geqslant 1}$ forms an orthonormal basis. Thereby, under MAR the covariance matrix given by $\Sigma = \mathbf{E}[\varepsilon(\Delta, X, W)\varepsilon(\Delta, X, W)^t]$ of $\varepsilon(\Delta, X, W)$ is well defined. The ordered eigenvalues of $\Sigma$ are denoted by $(\lambda_j)_{j \geqslant 1}$. Furthermore, we introduce a sequence $\{\chi^2_{1j}\}_{j \geqslant 1}$ of independent random variables that are distributed as chi-square with one degree of freedom. The proof of the next theorem can be found in the appendix.

THEOREM 2.3. *Let Assumptions 1, 2, 4, and 5 hold true. If*

$$\sum_{j=1}^{m_n} \tau_j = O(1), \quad n = o(\gamma_{k_n}), \quad \text{and} \quad m_n^{-1} = o(1) \tag{2.6}$$

*then under MAR*

$$n\, S_n^{MAR} \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j\, \chi^2_{1j}.$$

The rate $n = o(\gamma_{k_n})$ ensures that bias for estimating the function $h$ vanishes sufficiently fast. Below, we show that under classical smoothness assumptions this rate requires an undersmoothed estimator for $h$. We also like to emphasize that for the asymptotic result in Theorem 2.3, the bounded completeness condition stated in Assumption 3 is not required. Below we write $a_n \sim b_n$ when there exist constants $c, c' > 0$ such that $cb_n \leqslant a_n \leqslant c'b_n$ for all sufficiently large $n$.

EXAMPLE 2.4. Let $X$ be continuously distributed. Let $e_1, \ldots, e_{k_n}$ be spline basis functions and $p$ be the number of continuous derivatives of $h$. Then Assumption 5 (*i*) holds true with $\gamma_j \sim j^{2p/d_x}$ (see Newey [1997]). Condition $n = o(\gamma_{k_n})$ and Assumption 5 (*ii*) is satisfied if $k_n \sim n^{\kappa}$ with $d_x/(2p) < \kappa < 1/(2 + \varepsilon)$ for any small $\varepsilon > 0$. Here, the required smoothness of $h$ is $p > (2 + \varepsilon)(1 + d_x)/2$. Hence, the estimator of $h$ needs to be undersmoothed. $\square$

REMARK 2.1 (Estimation of Critical Values). The asymptotic distribution of our test statistic derived in Theorem 2.3 depends on unknown population quantities. As we see in the following, the critical values can be easily estimated. Let us define

$$\widehat{\varepsilon}_j(\delta, x, w) = \sqrt{\tau_j}\Big(\delta - \widehat{h}_n(x)\Big)\Big\{f_j(x, w) - \sum_{j'=1}^{k_n}\Big(\frac{1}{n}\sum_{i=1}^{n}f_j(X_i, W_i)e_{j'}(X_i)\Big)e_{j'}(x)\Big\}.$$

We replace $\Sigma$ by the $m_n \times m_n$ dimensional matrix

$$\widehat{\Sigma}_n = n^{-1} \sum_{i=1}^n \left( \widehat{\varepsilon}_1(\Delta_i, X_i, W_i), \dots, \widehat{\varepsilon}_{m_n}(\Delta_i, X_i, W_i) \right)^t \left( \widehat{\varepsilon}_1(\Delta_i, X_i, W_i), \dots, \widehat{\varepsilon}_{m_n}(\Delta_i, X_i, W_i) \right).$$

Let $(\widehat{\lambda}_{jn})_{1 \leqslant j \leqslant m_n}$ denote the ordered eigenvalues of $\widehat{\Sigma}_n$. To obtain the critical values, we replace $\sum_{j=1}^\infty \lambda_j \chi_{1j}^2$ by the finite sum $\sum_{j=1}^{m_n} \widehat{\lambda}_{jn} \chi_{1j}^2$. $\qquad\square$

The following result establishes consistency of the empirical critical values as introduced in the previous remark. In contrast to Theorem 2.3, we have to impose an upper bound on the size of the dimension parameter $m_n$.

PROPOSITION 2.4. *Let the conditions of Theorem 2.3 and Assumption 6 be satisfied. Assume*

$$\sum_{j \geqslant m_n} \lambda_j \chi_{1j}^2 = o_p(1) \quad and \quad m_n k_n^2 = o(n).$$

*Then for all z it holds*

$$\mathbb{P}\left( \sum_{j=1}^{m_n} \widehat{\lambda}_{jn} \chi_{1j}^2 \leqslant z \right) - \mathbb{P}\left( \sum_{j=1}^\infty \lambda_j \chi_{1j}^2 \leqslant z \right) = o(1).$$

## 2.5. Asymptotic Distribution under MCAR

We now derive the asymptotic distribution of the statistic for testing $S_n^{MCAR}$ under the null hypothesis MCAR. For any realization $(\delta, x, w)$ of $(\Delta, X, W)$, let us introduce an infinite dimensional vector $v(\delta, x, w)$ with $j$-th entry

$$v_j(\delta, x, w) = \sqrt{\tau_j} \left( \delta - \mathbb{P}(\Delta = 1) \right) \left( f_j(x, w) - \mathbb{E}[f_j(X, W)] \right).$$

We have $\mathbb{E}[v_j(\Delta, X, W)] = 0$ under MCAR. Let $\Sigma_v$ be the covariance matrix of $v(\Delta, X, W)$; that is, $\Sigma_v = \mathbb{E}[v(\Delta, X, W)v(\Delta, X, W)^t]$.
In this subsection, the ordered eigenvalues of $\Sigma_v$ are denoted by $(\lambda_l)_{l \geqslant 1}$. The next result is a direct consequence of Theorem 2.3 and hence, we omit its proof.

COROLLARY 2.5. *Let Assumptions 1, 2, and 4 hold true. If*

$$\sum_{j=1}^{m_n} \tau_j = O(1) \quad and \quad m_n^{-1} = o(1)$$

*then under MCAR*

$$n S_n^{MCAR} \xrightarrow{d} \sum_{j=1}^\infty \lambda_j \chi_{1j}^2.$$

11

REMARK 2.2 (Estimation of Critical Values). Estimation of critical values in case of Corollary 2.5 follows easily from Remark 2.1. Let us define

$$\widehat{v}_j(\delta, x, w) = \sqrt{\tau_j}\left(\delta - \widehat{\Delta}_n\right)\left\{f_j(x, w) - n^{-1}\sum_{i=1}^n f_j(X_i, W_i)\right\}$$

with $\widehat{\Delta}_n = n^{-1}\sum_{i=1}^n \Delta_i$. We replace $\Sigma_v$ by the $m_n \times m_n$ dimensional matrix

$$\widehat{\Sigma}_n = n^{-1}\sum_{i=1}^n \left(\widehat{v}_1(\Delta_i, X_i, W_i), \ldots, \widehat{v}_{m_n}(\Delta_i, X_i, W_i)\right)^t \left(\widehat{v}_1(\Delta_i, X_i, W_i), \ldots, \widehat{v}_{m_n}(\Delta_i, X_i, W_i)\right).$$

Let $(\widehat{\lambda}_{jn})_{1 \leqslant j \leqslant m_n}$ denote the ordered eigenvalues of $\widehat{\Sigma}_n$. To obtain empirical critical values, we replace $\sum_{j=1}^{\infty}\lambda_j \chi_{1j}^2$ by the finite sum $\sum_{j=1}^{m_n}\widehat{\lambda}_{jn}\chi_{1j}^2$. Consistency follows as in Proposition 2.4. □

## 2.6. Consistency against Fixed Alternatives

Under each null hypothesis, the asymptotic distribution results remain valid if $(Y^*, X)$ is not bounded complete for $(X, W)$; that is, Assumption 3 does not hold true. On the other hand, we show that under bounded completeness consistency of our tests against fixed alternatives can be obtained. To establish this property we require the following additional assumption.

ASSUMPTION 7. *The function $p_{XW}/v$ is uniformly bounded away from zero.*

If MAR fails, Assumption 7 together with Assumption 3 ensures that the generalized Fourier coefficients $\mathbf{E}[r(\Delta, X)f_j(X, W)]$ are non-zero for some integer $j \geqslant 1$. The following proposition shows that our test has the ability to reject a false null hypothesis with probability 1 as the sample size grows to infinity. For the next results, let us introduce a sequence $(a_n)_{n \geqslant 1}$ satisfying $a_n = o(n)$. The proof of the next proposition can be found in the appendix.

PROPOSITION 2.6. *Assume that MAR does not hold. Let Assumptions 1–7 be satisfied. Then*

$$\mathbb{P}\left(n\,S_n^{MAR} > a_n\right) = 1 + o(1).$$

The rate $(a_n)_{n \geqslant 1}$ is arbitrarily close to the parametric rate $n^{-1}$ which is due the weighting sequence $(\tau_j)_{j \geqslant 1}$ with $\sum_{j=1}^{m_n}\tau_j = O(1)$. The next two result is a direct consequence of Proposition 2.6 and hence, its proof is omitted.

COROLLARY 2.7. *Assume that MCAR does not hold. Let Assumptions 1–4 and 7 be satisfied. Then*

$$\mathbb{P}\left(n\,S_n^{MCAR} > a_n\right) = 1 + o(1).$$

# 3. Monte Carlo Simulations

In this section, we study the finite-sample performance of our test by presenting the results of a Monte Carlo simulation. The experiments use a sample size of 500 and there are 1000 Monte Carlo replications in each experiment. Results are presented for the nominal level 0.05.

As basis functions $\{f_j\}_{j \geq 1}$ used to construct our test statistic, we use throughout the experiments orthonormalized Hermite polynomials. Hermite polynomials form an orthonormal basis of $L_{\bar{\omega}}^2$ with a weighting function being the density of the standard normal distribution; that is, $\bar{\omega}(x) = \exp(-x^2)/\sqrt{2\pi}$. They can be obtained by applying the Gram–Schmidt procedure to the polynomial series $1, x, x^2, \ldots$ under the inner product $\langle \phi, \psi \rangle_{\bar{\omega}} = (2\pi)^{-1/2} \int \phi(x)\psi(x) \exp(-x^2)dx$. That is, $H_1(x) = 1$ and for all $j = 2, 3, \ldots$

$$H_j(x) = \frac{x^{j-1} - \sum_{k=1}^{j-1}\langle id^{j-1}, p_j \rangle_\omega p_j(x)}{\int \left(x^{j-1} - \sum_{k=1}^{j-1}\langle id^{j-1}, p_j \rangle_\omega p_j(x)\right)\omega(x)dx}. \tag{3.1}$$

Our testing procedure is now build up on the basis functions

$$f_j(\cdot) = \frac{H_{j+1}(\cdot)}{\sqrt{\langle H_j, H_j \rangle_\omega}}$$

for all $j = 1, 2, \ldots$ If the support of the instrument $W$ or its transformation lies in the interval $[0, 1]$ then one could also use, for instance, cosine basis functions

$$f_j(x) = \sqrt{2}\cos(\pi j x)$$

for $j = 1, 2, \ldots$. We also implemented our test statistic with these cosine functions in the settings studied below. But as the results are very similar to the ones with Hermite polynomials presented below we do not report them here. Throughout our simulation study, the number of orthonormalized Hermite polynomials is 10. In the multivariate case, we consider a tensor product of those weighted Hermite polynomials. Due to the weighting sequence $(\tau_j)_{j \geq 1}$, results not too sensitive to the number of Hermite polynomials. When implementing the test with cosine basis functions we used 100 basis functions. In contrast, results might be more sensitive to the choice of basis functions used to estimate $h$. Below we use cross validation to choose the appropriate number of basis functions for this function.

**Testing MCAR** Realizations of $(Y^*, W)$ were generated by $W \sim \mathcal{N}(0, 1)$ and $Y^* \sim \rho W + \sqrt{1 - \rho^2}\,\varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 1)$. The constants $\rho$ characterizes the "strength" of the instrument $W$ and is varied in the experiments. For a random variable $V$, introduce the function $\phi_2(V) = \mathbb{1}\{V \geq q\} + 0.1 * \mathbb{1}\{V \leq q\}$ where

$q$ is the 0.2 quantile of the empirical distribution of $V$. In each experiment,

| Model | | Empirical Rejection probability of $nS_n$ with | | | Little's test |
|---|---|---|---|---|---|
| $\rho$ | $v$ | $\tau_j = j^{-2}$ | $\tau_j = j^{-3}$ | $\tau_j = j^{-4}$ | |
| 0.2 | 0.0 | 0.055 | 0.057 | 0.056 | 0.062 |
| | 0.3 | 0.148 | 0.159 | 0.162 | 0.155 |
| | 0.5 | 0.290 | 0.297 | 0.304 | 0.338 |
| | 0.7 | 0.505 | 0.529 | 0.530 | 0.568 |
| 0.3 | 0.0 | 0.055 | 0.057 | 0.056 | 0.062 |
| | 0.3 | 0.253 | 0.265 | 0.268 | 0.273 |
| | 0.5 | 0.559 | 0.588 | 0.592 | 0.611 |
| | 0.7 | 0.839 | 0.855 | 0.857 | 0.853 |
| 0.4 | 0.0 | 0.055 | 0.057 | 0.056 | 0.062 |
| | 0.3 | 0.387 | 0.393 | 0.398 | 0.430 |
| | 0.5 | 0.813 | 0.831 | 0.840 | 0.843 |
| | 0.7 | 0.985 | 0.988 | 0.986 | 0.988 |

Table 1: Empirical Rejection probabilities for Testing MCAR

realizations of the response variable $\Delta$ were generated by

$$\Delta \sim \mathrm{Bin}\left(1, \phi_{.2}\left(v\,Y^* + \sqrt{1 - v^2}\,\xi\right)\right)$$

for some constant $0 \leqslant v \leqslant 1$ and where $\xi \sim \mathcal{N}(0,1)$. If $v = 0$ then response $\Delta$ does not depend on $Y^*$ and hence the null hypothesis MCAR holds true.
The critical values are estimated as in Remark 2.2. For $m = 100$ we observed that the estimated eigenvalues $\widehat{\lambda}_j$ are sufficiently close to zero for all $j \geqslant m$. To provide a basis for judging whether the power of our test is high or low, we also provide the empirical rejection probabilities when using a test of MCAR for normal data proposed by Little [1988].
The empirical rejection probabilities of test statistic $S_n^{MCAR}$ using different weightings and Little's test are depicted in Table 1. First, we observe, not surprisingly, that the power of all tests increase as the correlation between $Y^*$ and $W$ (measured by $\rho$) becomes larger. Second, power also increases with constant $v$. From Table 1 we also see that our tests with different weighting sequences have similar power properties and our tests behave similar as Little's test.

**Testing MAR** Realizations of $(Y^*, X, W)$ were generated by $W \sim \mathcal{N}(0,1)$, $X \sim 0.2\,W + \sqrt{1 - 0.2^2}\,\xi$ and $Y^* \sim \rho\,W + \sqrt{1 - \rho^2}\,\xi + \varepsilon$ where $\xi, \varepsilon \sim \mathcal{N}(0, 0.25)$. The constant $\rho$ is varied in the experiments. The critical values are estimated as described in Remark 2.1.

In each experiment, realizations of response $\Delta$ were generated by

$$\Delta \sim \text{Binomial}\left(1, \phi_{.2}\left(v\, Y^* + \sqrt{1 - v^2}\, X\right)\right)$$

for some constant $0 \leqslant v \leqslant 1$. Clearly, if $v = 0$ then the null hypothesis MAR holds true. We estimate the function $h$ using B-splines. The number of

| Model | | Empirical Rejection probability of $nS_n$ with | | |
|:---:|:---:|:---:|:---:|:---:|
| $\rho$ | $v$ | $\tau_j = j^{-2}$ | $\tau_j = j^{-3}$ | $\tau_j = j^{-4}$ |
| 0.3 | 0.0 | 0.041 | 0.044 | 0.046 |
| | 0.3 | 0.059 | 0.059 | 0.059 |
| | 0.5 | 0.134 | 0.133 | 0.127 |
| | 0.7 | 0.231 | 0.230 | 0.220 |
| 0.5 | 0.0 | 0.041 | 0.044 | 0.046 |
| | 0.3 | 0.204 | 0.198 | 0.192 |
| | 0.5 | 0.646 | 0.622 | 0.603 |
| | 0.7 | 0.926 | 0.912 | 0.895 |
| 0.7 | 0.0 | 0.041 | 0.044 | 0.046 |
| | 0.3 | 0.479 | 0.466 | 0.449 |
| | 0.5 | 0.953 | 0.942 | 0.924 |
| | 0.7 | 0.994 | 0.995 | 0.996 |

Table 2: Empirical Rejection probabilities for Testing MAR

knots and orders is chosen via cross validation. Computational procedures were implemented using the statistical software R using the crs Package Hayfield and Racine [2007]. In our experiments, cross validation tended to undersmooth the estimator of $h$ which implied a sufficiently small bias of this estimator. On the other hand, to obtain appropriate undersmoothing one could also use a data driven choice of basis functions suggested by Picard and Tribouley [2000]. The critical values are estimated as described in Remark 2.1. Table 2 depicts the empirical rejection probabilities of the test $S_n^{MAR}$ when using different weightings.

# 4. Empirical Illustration

We now apply our testing procedure to analyze response mechanisms in a data set from the Health and Retirement Study (HRS). In this survey, respondents were asked about their out of pocket prescription drug spending. Those who were not able to report point values for these were asked to provide brackets, giving point values for some observations and intervals of

different sizes for others. This censoring problem might violate the MAR hypothesis: the variable is censored only for those who do not recall how much they spent, and remembering how much one spent might be correlated with the level of spending itself. We refer to Armstrong [2015] who constructed confidence intervals for partially identified regression of prescription drug spending on income.

In this empirical illustration, we consider the 1996 wave of the survey and restrict attention to women with less than 25,000$ of yearly income who report using prescription medications. This results in a data set with 943 observations. Of these observations, participants do not report the exact prescription expenditures but rather an interval of nonzero width with finite endpoints or provide no information on their prescription expenditures at all. The length of these intervals is 10, 80, and 400. If the participants do not report their exact amount but an interval of length 10 we consider the center of this interval as observation (these are only six observations). Intervals with width larger or equal to 80 are treated as missing values. Thereby, 55 observations are missing (roughly 6%). The results presented below are essentially the same when intervals with length 10 are treated as missing observations.

Whether we observe prescription drug expenditure $Y^*$ is assumed to be independent of the yearly income $W$ conditional on $Y^*$. We thus use yearly income as instrumental variable. The exclusion restriction might be criticized since unobserved individual characteristics, such as cognitive ability, might be strongly related to income and could directly influence the recalling behavior of women. Below we also introduce additional covariates to reduce the effect of such a potential bias. Moreover, we directly perform a test of assumption $\Delta \perp\!\!\!\perp W|Y^*$ based on Breunig et al. [2014] which builds up on D'Haultfoeuille [2010]'s observation that $\Delta \perp\!\!\!\perp W|Y^*$ is equivalent to $\mathbf{E}[\Delta/\mathbb{P}(\Delta = 1|Y^*)|W] = 1$. The test is based on checking whether there exists a smooth function $\phi$ with $\phi(\cdot) \geqslant 1$ satisfying $\mathbf{E}[\Delta\phi(Y^*)|W] = 1$. The value of the $L^2$ test statistic proposed by Breunig et al. [2014] is $-1.1830$ (when using (B-spline basis functions for $\phi$ with order 2 and 3 knots so $k_n = 2 + 3 + 1 = 6$ and the criterion function with with order 3 and 7 knots so $m_n = 3 + 7 + 1 = 11$). At 0.05 nominal level the p-value is 0.2368 and thus we fail to reject the exclusion assumption of the instrument $W$.

As additional covariates $X$, we include birth year of participants. In this case the milder version of exclusion restriction $\Delta \perp\!\!\!\perp W|(Y^*, X)$ has to be imposed. The test statistics are constructed as described in the previous section. More precisely, we truncate $W$ to $[0, 1]$ and use cosine basis functions with $m_n = 100$. Table 3 depicts the values of our test statistics for testing MCAR with and without covariates $X$ and MAR. As we see from this table, our test statistics reject the hypothesis MCAR (with and without $X$) but fail to reject the hypothesis MAR.

In Figure 1, we estimate the conditional probability $\mathbb{P}(\Delta = 1|Y^*)$. This conditional probability is identified through the conditional mean equation $\mathbf{E}[\Delta/\mathbb{P}(\Delta = 1|Y^*)|X] = 1$ (cf. D'Haultfoeuille [2010]). We use a sieve minimum

|  |  | $\tau_j = j^{-2}$ | $\tau_j = j^{-3}$ | $\tau_j = j^{-4}$ |
|---|---|---|---|---|
| MCAR: $\mathbb{P}(\Delta = 1\|Y^*) = \mathbb{P}(\Delta = 1)$ | Value of $nS_n^{MCAR}$ | 0.273 | 0.244 | 0.233 |
|  | Critical Values | 0.214 | 0.179 | 0.169 |
| MCAR: $\mathbb{P}(\Delta = 1\|Y^*, X) = \mathbb{P}(\Delta = 1)$ | Value of $nS_n^{MCAR}$ | 0.156 | 0.050 | 0.020 |
|  | Critical Values | 0.100 | 0.033 | 0.014 |
| MAR: $\mathbb{P}(\Delta = 1\|Y^*, X) = \mathbb{P}(\Delta = 1\|X)$ | Value of $nS_n^{MAR}$ | 0.085 | 0.028 | 0.012 |
|  | Critical Values | 0.101 | 0.038 | 0.021 |

Table 3: Values of $nS_n^{MCAR}$ and $nS_n^{MAR}$ together with their critical values

distance estimator based on B-splines as in Breunig et al. [2014] to estimate $\mathbb{P}(\Delta = 1|Y^*)$. From Figure 1 we see that the estimator for $\mathbb{P}(\Delta = 1|Y^*)$ is strictly decreasing with $Y^*$. On the other hand, if data is MCAR then $\mathbb{P}(\Delta = 1|Y^*)$ would be constant. Significant statements about the shape of the curve, however, are hard to make as the solution to the conditional moment restriction is an ill-posed problem and hence confidence intervals can be very wide.
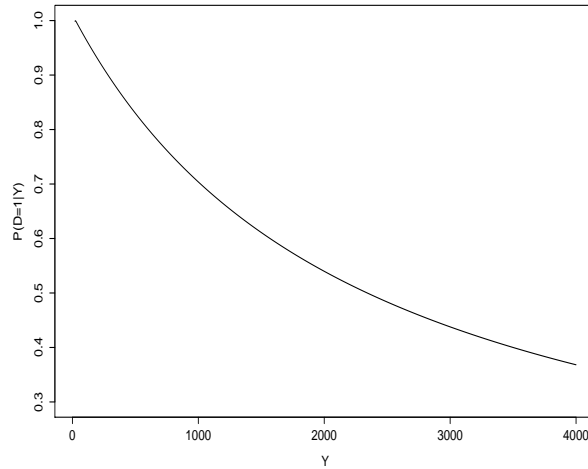


Figure 1: Graph of $\mathbb{P}(\Delta = 1|Y^* = \cdot)$

# A. Appendix

## A.1. Proofs of Section 2.

Throughout the Appendix, let $C > 0$ denote a generic constant that may be different in different uses. For ease of notation let $\sum_i = \sum_{i=1}^n$ and $\sum_{i'<i} = \sum_{i=1}^n \sum_{i'=1}^{i-1}$. Further, to keep notation simple we define $Z := (X, W)$. In the following, $\varepsilon_{\underline{m_n}}(\delta, x, w)$ and $\varepsilon_{\underline{m_n}}^h(\delta, x)$ denote $m_n$–dimensional vectors with $j$– th entries given by $\varepsilon_j(\delta, x, w)$ and $\varepsilon_j^h(\delta, x) := \sqrt{\tau_j}\left(\delta - h(x)\right)\sum_{l=1}^\infty \mathbf{E}[f_j(Z)e_l(X)]e_l(x)$, respectively. In the appendix, $f_{\underline{m_n}}^\tau$ denotes a $m_n$ dimensional vector with entries $\sqrt{\tau_j}f_j$ for $1 \leq j \leq m_n$.

PROOF OF THEOREM 2.3. The proof is based on the decomposition

$$n^{-1/2} \sum_i \left(\Delta_i - \widehat{h}_n(X_i)\right) f_{\underline{m_n}}^\tau(Z_i)$$

$$= n^{-1/2} \sum_i \varepsilon_{\underline{m_n}}(\Delta_i, Z_i)$$

$$+ n^{-1/2} \sum_i \left(\left(h(X_i) - \widehat{h}_n(X_i)\right)f_{\underline{m_n}}^\tau(Z_i) + \varepsilon_{\underline{m_n}}^h(\Delta_i, X_i)\right)$$

$$= I_n + II_n \quad \text{(say)}.$$

Consider $I_n$. Consider some fixed integer $m \geq 1$. Using Cramer Wold device it is easily seen that

$$n^{-1/2} \sum_i \varepsilon_{\underline{m}}(\Delta_i, Z_i) \xrightarrow{d} \mathcal{N}(0, \Sigma_m)$$

where $\Sigma_m$ is the upper $m \times m$ submatrix of $\Sigma$. Hence, we have

$$\sum_{j=1}^m \left| n^{-1/2} \sum_i \varepsilon_j(\Delta_i, Z_i) \right|^2 \xrightarrow{d} \sum_{j=1}^m \lambda_j \chi_{1j}^2$$

with $\lambda_j, 1 \leq j \leq m$, being eigenvalues of $\Sigma_m$. On the other hand, observe

$$\sum_{j>m} \mathbf{E} \left| n^{-1/2} \sum_i \varepsilon_j(\Delta_i, Z_i) \right|^2 = \sum_{j>m} \mathbb{V}\mathrm{ar}\left(n^{-1/2} \sum_i \varepsilon_j(\Delta_i, Z_i)\right) = \sum_{j>m} \mathbf{E}\,\varepsilon_j^2(\Delta, Z)$$

which becomes sufficiently small for large $m$ as $\mathbf{E}\,\varepsilon_j^2(\Delta, Z)/\tau_j \leq C$ for all $j \geq 1$. Hence, from page 199 in Serfling [1981] we infer that $I_n \xrightarrow{d} \sum_{j \geq 1} \lambda_j \chi_{1j}^2$.

Consider $II_n$. We have

$$\|II_n\|^2 \leqslant 2 \sum_{j=1}^{m_n} \left| \tau_j^{1/2} n^{-1/2} \sum_i \left( \widehat{h}_n(X_i) - (E_{k_n}h)(X_i) \right) f_j(Z_i) - \varepsilon_j^h(\Delta_i, X_i) \right|^2$$

$$+ 2 \sum_{j=1}^{m_n} \tau_j \left| n^{-1/2} \sum_i (E_{k_n}h - h)(X_i) f_j(Z_i) \right|^2$$

$$=: A_{n1} + A_{n2}.$$

Consider $A_{n1}$. In the following, we denote $Q_n := n^{-1} \sum_i e_{\underline{k_n}}(X_i) e_{\underline{k_n}}(X_i)^t$. It holds $\widehat{h}_n(\cdot) = e_{\underline{k_n}}(\cdot)^t (nQ_n)^{-1} \sum_i \Delta_i e_{\underline{k_n}}(X_i)$. By Assumption 5, the eigenvalues of $\mathbf{E}[e_{\underline{k_n}}(X) e_{\underline{k_n}}(X)^t]$ are bounded away from zero and hence, it may be assumed that $\mathbf{E}[e_{\underline{k_n}}(X) e_{\underline{k_n}}(X)^t] = \mathbf{I}_{k_n}$ where $\mathbf{I}_{k_n}$ is the $k_n$ dimensional identity matrix (cf. Newey [1997], p. 161). We observe

$$A_{n1} \leqslant 2 \sum_{j=1}^{m_n} \left| \tau_j^{1/2} \sum_{l=1}^{k_n} \mathbf{E}[f_j(Z)e_l(X)] Q_n^{-1} n^{-1/2} \sum_i \left( \Delta_i - E_{k_n}h(X_i) \right) e_{\underline{k_n}}(X_i) - \varepsilon_j^h(\Delta_i, X_i) \right|^2$$

$$+ 2\|E_{k_n}h - \widehat{h}_n\|_X^2 \sum_{j=1}^{m_n} \tau_j \sum_{l=1}^{k_n} \left| n^{-1/2} \sum_i e_l(X_i) f_j(Z_i) - \mathbf{E}[e_l(X)f_j(Z)] \right|^2$$

$$=: 2B_{n1} + 2B_{n2} \quad \text{(say). (A.1)}$$

For $B_{n1}$ we evaluate due to the relation $Q_n^{-1} = \mathbf{I}_{k_n} - Q_n^{-1}(Q_n - \mathbf{I}_{k_n})$ that

$$B_{n1} \leqslant 2 \left\| \mathbf{E}[f_{\underline{m_n}}^\tau(Z)e_{\underline{k_n}}(X)^t] n^{-1/2} \sum_i (\Delta_i - E_{k_n}h(X_i))e_{\underline{k_n}}(X_i) - \varepsilon_{\underline{m_n}}^h(\Delta_i, X_i) \right\|^2$$

$$+ 2 \left\| \mathbf{E}[f_{\underline{m_n}}^\tau(Z)e_{\underline{k_n}}(X)^t] \right\|^2 \|Q_n - \mathbf{I}_{k_n}\|^2 \|Q_n^{-1}\|^2 \left\| n^{-1/2} \sum_i \left( \Delta_i - E_{k_n}h(X_i) \right) e_{\underline{k_n}}(X_i) \right\|^2$$

$$= 2C_{n1} + 2C_{n2} \quad \text{(say)}.$$

Further, from $\mathbf{E}\left[ \left( \Delta - E_{k_n}h(X) \right) e_{\underline{k_n}}(X) \right] = 0$, $\mathbf{E}\left[ (h - E_{k_n}h)(X)e_{\underline{k_n}}(X) \right] = 0$, and

$\mathbf{E}[\varepsilon_{\underline{m_n}}(\Delta, Z)] = 0$ we deduce

$$C_{n1} \leqslant 2 \sum_{j=1}^{m_n} \tau_j \, \mathbf{E} \Big| \sum_{l > k_n} \mathbf{E}[f_j(Z)e_l(X)](h(X) - \Delta)e_l(X) \Big|^2$$

$$+ 2 \sum_{j=1}^{m_n} \tau_j \, \mathbf{E} \Big| \sum_{l=1}^{k_n} \mathbf{E}[f_j(Z)e_l(X)](E_{k_n}h - h)(X)e_l(X) \Big|^2$$

$$\leqslant 2 \sum_{j=1}^{m_n} \tau_j \, \mathbf{E} \Big| \sum_{l > k_n} \mathbf{E}[f_j(Z)e_l(X)]e_l(X) \Big|^2$$

$$+ C \, k_n \, \|E_{k_n}h - h\|_X^2 \sum_{j=1}^{m_n} \tau_j \sum_{l=1}^{k_n} \mathbf{E}[f_j(Z)e_l(X)]^2$$

$$= o(1)$$

using that $\mathbf{E}[(h(X) - \Delta)^2|X]$ is bounded, $\sum_{j=1}^{m_n} \tau_j \sum_{j=1}^{k_n} \mathbf{E}[f_j(Z)e_l(X)]^2 = O(1)$, and by assumption

$$k_n^2 \, \|E_{k_n}h - h\|_X^2 = O(k_n^2/\gamma_{k_n}) = O(k_n^2/n) = o(1).$$

Consider $C_{n2}$. Further, by Rudelson's matrix inequality (see Rudelson [1999] and also Lemma 6.2 of Belloni et al. [2015]) it holds

$$\big\| Q_n - \mathbf{I}_{k_n} \big\|^2 = O_p\big(n^{-1}\log(n)\,k_n\big).$$

Moreover, since the difference of eigenvalues of $Q_n$ and $\mathbf{I}_{k_n}$ is bounded by $\|Q_n - \mathbf{I}_{k_n}\|$, the smallest eigenvalue of $Q_n$ converges in probability to one and hence, $\|Q_n^{-1}\|^2 = 1 + o_p(1)$. Further,

$$\sum_{l=1}^{k_n} \mathbf{E} \Big| n^{-1/2} \sum_i \big( \Delta_i - E_{k_n}h(X_i) \big)e_l(X_i) \Big|^2 = \sum_{l=1}^{k_n} \mathbf{E} \Big| \big( \Delta - E_{k_n}h(X) \big)e_l(X) \Big|^2 = O(k_n)$$

and hence $C_{n2} = O_p\big(n^{-1}\log(n)\,k_n^2\big) = o_p(1)$. Consequently, $B_{n1} = o_p(1)$. Consider $B_{n2}$. It holds

$$\sum_{l=1}^{k_n} \Big| n^{-1} \sum_i e_l(X_i)f_j(Z_i) - \mathbf{E}[e_l(X)f_j(Z)] \Big|^2 = O_p(k_n/n).$$

Since $\|E_{k_n}h - \widehat{h}_n\|_X^2 = O_p(k_n/n)$ (cf. Theorem 1 of Newey [1997]) and $k_n^2/n = o(1)$ it follows that $B_{n2} = o(1)$. Thus, we conclude $A_{n1} = o_p(1)$. For $A_{n2}$ we observe

that

$$\mathbf{E}\,A_{n2} \leqslant n \sum_{j=1}^{m_n} \tau_j\, \mathbf{E}\,|(E_{k_n}h - h)(X)f_j(Z)|^2$$

$$\leqslant n\|E_{k_n}h - h\|_\infty^2 \sum_{j=1}^{m_n} \tau_j\, \mathbf{E}\,f_j^2(Z) = O\big(n/\gamma_{k_n}\big) = o(1).$$

Consequently, we have $II_n = o_p(1)$ which completes the proof. $\qquad\square$

For the next proof we introduce the following notations. Let us denote $\psi_{jn}(x, w) = f_j(x, w) - \mathbf{E}[f_j(Z)|X = x]$. With $\gamma_{jn} = \mathbf{E}[f_j(Z)e_{k_n}(X)]$ and $\widehat{\gamma}_{jn}$ its empirical version, we further define $\psi_{jn}(x, w) = f_j(x, w) - \gamma_{jn}^t e_{k_n}(x)$ and $\widehat{\psi}_{jn}(x, w) = f_j(x, w) - \widehat{\gamma}_{jn}^t e_{k_n}(x)$. Moreover, let $\widetilde{\Sigma}_n$ denote a $m_n \times m_n$ matrix with $j, l$–th entry $n^{-1} \sum_{i=1}^n \sqrt{\tau_j \tau_l}\,(\Delta_i - E_{k_n}h_n(X_i))^2 \psi_{jn}(Z_i)\psi_{ln}(Z_i)$. Let $\widehat{\beta}_n$ and $\beta_n$ be such that $\widehat{h}_n(\cdot) = \widehat{\beta}_n' e_{\underline{k}_n}(\cdot)$ and $E_{k_n}h(\cdot) = \beta_n' e_{\underline{k}_n}(\cdot)$.

<span style="color:#8B2020">PROOF OF PROPOSITION</span> 2.4. Observe that

$$\sum_{j=1}^{m_n} \widehat{\lambda}_{jn}\chi_{1j}^2 = \sum_{j=1}^{\infty} \lambda_j \chi_{1j}^2 + \sum_{j=1}^{m_n}(\widehat{\lambda}_{jn} - \lambda_j)\chi_{1j}^2 + o_p(1).$$

Let $\chi_{m_n}^2$ denote the chi-square distribution with $m_n$ degrees of freedom. We have

$$\sum_{j=1}^{m_n}(\widehat{\lambda}_{jn} - \lambda_j)\chi_{1j}^2 \leqslant \chi_{m_n}^2 \max_{1 \leqslant j \leqslant m_n} |\widehat{\lambda}_{jn} - \lambda_j|.$$

It holds $\chi_{m_n}^2 = O_p(m_n)$ and it thus remains to show $\max_{1 \leqslant j \leqslant m_n} |\widehat{\lambda}_{jn} - \lambda_j| = O_p(k_n^2/n)$. Let $\Sigma_{m_n}$ denote the upper $m_n \times m_n$ matrix of $\Sigma$. We have

$$\max_{1 \leqslant j \leqslant m_n} |\widehat{\lambda}_{jn} - \lambda_j| \leqslant \|\widehat{\Sigma}_n - \Sigma_{m_n}\| \leqslant \|\widehat{\Sigma}_n - \widetilde{\Sigma}_n\| + \|\widetilde{\Sigma}_n - \Sigma_{m_n}\|$$

21

where

$$\|\widehat{\Sigma}_n - \widetilde{\Sigma}_n\|^2$$

$$\leqslant \sum_{j,l=1}^{m_n} \tau_j \tau_l \Big| \frac{1}{n} \sum_i \Big( 2(\Delta_i - E_{k_n} h(X_i))(\widehat{h}_n(X_i) - E_{k_n} h(X_i)) + (\widehat{h}_n(X_i) - E_{k_n} h(X_i))^2 \Big) \psi_{jn}(Z_i)\psi_{ln}(Z_i)$$

$$+ (\Delta - \widehat{h}_n(X_i))^2 \Big( 2(\widehat{\psi}_{jn} - \psi_{jn})(Z_i)\psi_{ln}(Z_i) + (\widehat{\psi}_{jn} - \psi_{jn})(Z_i)(\widehat{\psi}_{ln} - \psi_{ln})(Z_i) \Big) \Big|^2$$

$$\leqslant C \sum_{j,l=1}^{m_n} \tau_j \tau_l \Big\{ \Big| \mathbf{E}[(\Delta - E_{k_n} h(X))\psi_{jn}(Z)\psi_{ln}(Z)e_{\underline{k_n}}(X)](\widehat{\beta}_n - \beta_n) \Big|^2$$

$$+ \Big| (\widehat{\beta}_n - \beta_n)^t \mathbf{E}[e_{\underline{k_n}}(X)\psi_{jn}(Z)\psi_{ln}(Z)e_{\underline{k_n}}(X)^t](\widehat{\beta}_n - \beta_n) \Big|^2$$

$$+ \Big| \mathbf{E}[\psi_{ln}(Z)e_{\underline{k_n}}(X)^t](\widehat{\gamma}_{jn} - \gamma_{jn}) \Big|^2 + \Big| (\widehat{\gamma}_{jn} - \gamma_{jn})^t(\widehat{\gamma}_{ln} - \gamma_{ln}) \Big|^2 \Big\} + O_p(k_n^4/n^2)$$

$$= O_p(k_n^4/n^2).$$

Here we used that under *MAR* it holds

$$\| \mathbf{E}[(\Delta - E_{k_n} h(X))\psi_{jn}(Z)\psi_{ln}(Z)e_{\underline{k_n}}(X)]\| = \| \mathbf{E}[(h(X) - E_{k_n} h(X))\psi_{jn}(Z)\psi_{ln}(Z)e_{\underline{k_n}}(X)]\|$$

$$\leqslant \|h - E_{k_n} h\|_\infty \max_{1 \leqslant j \leqslant m_n} \mathbf{E}[\psi_{jn}^2(Z)] \sup_{x \in \mathcal{X}} \|e_{\underline{k_n}}(x)\|$$

$$= O\Big( \sqrt{k_n/\gamma_{k_n}} \Big) = o\Big( \sqrt{k_n/n} \Big).$$

Finally, is easily seen that $\|\widetilde{\Sigma}_n - \Sigma_{m_n}\| = O_p(k_n^2/n)$, which proves the result. □

PROOF OF PROPOSITION 2.6. Let us introduce a smoothing operator $K$ which has an eigenvalue decomposition $\big\{ \sqrt{\tau_j}, f_j \big\}_{j \geqslant 1}$ and a conditional expectation operator $T$ defined by $T\phi = \mathbf{E}[\phi(\Delta, X)|Z]$ for any bounded function $\phi$. Since $p_Z/\nu$ is uniformly bounded away from zero by some constant $C > 0$ we obtain

$$S_n = \sum_{j=1}^{m_n} \tau_j \Big| n^{-1} \sum_i r(\Delta_i, X_i)f_j(Z_i) \Big|^2 + o_p(1)$$

$$= \sum_{j=1}^{m_n} \tau_j \Big| \mathbf{E}\big[ r(\Delta, X)f_j(Z) \big] \Big|^2 + o_p(1)$$

$$= \sum_{j=1}^{\infty} \Big| \int_{\mathcal{X} \times \mathcal{W}} \sqrt{\tau_j}\, \mathbf{E}(r(\Delta, X)|Z = z)\frac{p_Z(z)}{\nu(z)}f_j(z)\nu(z)d(z) \Big|^2 + o_p(1)$$

$$= \int_{\mathcal{X} \times \mathcal{W}} \Big| (KTr)(z)\frac{p_Z(z)}{\nu(z)} \Big|^2 \nu(z)dz + o_p(1)$$

$$\geqslant C\, \mathbf{E}\big| (KTr)(Z) \big|^2 + o_p(1).$$

Since $K$ is nonsingular by construction it follows from the proof of Theorem

2.1 that $\mathbf{E}\,|(KTr)(Z)|^2 > 0.$                                          □

# References

T. B. Armstrong. Asymptotically exact inference in conditional moment inequality models. *Journal of Econometrics*, 2015.

A. Belloni, V. Chernozhukov, D. Chetverikov, and K. Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

R. Blundell and J. Horowitz. A nonparametric test of exogeneity. *Review of Economic Studies*, 74(4):1035–1058, Oct 2007.

R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

C. Breunig. Goodness-of-fit tests based on series estimators in nonparametric instrumental regression. *Journal of Econometrics*, 184(2):328–346, 2015.

C. Breunig, E. Mammen, and A. Simoni. Semi/nonparametric estimation in case of endogenous selection. Technical report, University of Mannheim (submitted.), 2014.

H. Y. Chen and R. Little. A test of missing completely at random for generalised estimating equations with missing data. *Biometrika*, 86(1):1–13, 1999.

K. Chen. Parametric models for response-biased sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):775–789, 2001.

X. Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 2007.

X. Chen and T. M. Christensen. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 2015.

M. Das, W. K. Newey, and F. Vella. Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70(1):33–58, 2003.

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

X. D'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(03):460–471, 2011.

F. Fève, J.-P. Florens, I. Van Keilegom, et al. Estimation of conditional ranks and tests of exogeneity in nonparametric nonseparable models. Technical report, UCL, 2012.

C. Fuchs. Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, 77 (378):270–278, 1982.

T. Hayfield and J. S. Racine. np: Nonparametric kernel smoothing methods for mixed datatypes. *R package version 0.13-1*, 2007.

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

J. L. Horowitz. Testing a parametric model against a nonparametric alternative with identification through instrumental variables. *Econometrica*, 74(2):521–538, 2006.

Y. Hu and S. M. Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

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

K.-Y. Liang and J. Qin. Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):773–786, 2000.

R. J. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404): 1198–1202, 1988.

R. J. Little and D. B. Rubin. Statistical analysis with missing data. 2002.

L. Mattner. Some incomplete but boundedly complete location families. *The Annals of Statistics*, 21(4):2158–2162, 1993.

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

D. Picard and K. Tribouley. Adaptive confidence interval for pointwise curve estimation. *Annals of statistics*, pages 298–335, 2000.

A. Qu and P. X.-K. Song. Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika*, 89(4):841–850, 2002.

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

M. Reilly and M. S. Pepe. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314, 1995.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.

S. Seaman, J. Galati, D. Jackson, J. Carlin, et al. What is meant by missing at random? *Statistical Science*, 28(2):257–268, 2013.

R. J. Serfling. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 1981.

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

# SFB 649 Discussion Paper Series 2015

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.