Testing Conditional Moment Restrictions: A Partitioning Approach

Antonio Raiola*
University Carlos III of Madrid

March 20, 2024
[Click here for the latest version]

Abstract

This paper proposes χ^2-type tests for assessing the specification of regression models or general conditional moment restrictions. The data is partitioned according to the explanatory variables into several cells, and the tests evaluate whether the difference between the observed average of the dependent variable and its expected value under the model specification arises by chance. In contrast to existing omnibus procedures, χ^2 tests are asymptotically pivotal and fairly insensitive to the curse of dimensionality. The computation is straightforward and does not require bootstrapping or smoothing techniques. Importantly, the asymptotic properties of the test are invariant to sample-dependent partitions, which can be chosen to favor certain alternatives. A Monte Carlo study provides evidence of the good performance of the tests using samples of small or moderate size compared with existing omnibus alternatives, particularly when there are many explanatory variables. An empirical application regarding returns to education of African American students in the US complements the finite sample study.

JEL classification: C12, C31, C52

Keywords: Goodness-of-fit tests; Regression specifications; Distribution-free tests; Regressogram; Neyman-Pearson cells.

*I am grateful to Miguel Angel Delgado for his guidance and support, and to Juan Carlos Escanciano, Carlos Velasco, Juan Jose Dolado, and Nazarii Salish for the many insightful discussions. I also thank the participants at the ENTER seminar at Toulouse School of Economics, as well as the participants of the Econometrics Reading Group at the University Carlos III of Madrid for their helpful comments. I gratefully acknowledge support from the Ministerio de Ciencia e Innovación through research grants PRE2020-092409, the AEI (Spain) through research grant PID2019-110779GB-I00, and University Carlos III for the scholarship for Master's study. Email address: araiola@eco.uc3m.es.

1 Introduction

The identification and estimation of causal relationships often rely on models defined by conditional moment restrictions (CMR), typically in the form of parametric regression models. Ensuring the correct model specification is crucial for drawing valid inferences and providing reliable interpretations. This article introduces χ^2 tests to check the goodness-of-fit of CMR model specifications.

Existing tests, rooted in the definition of the integrated regression function (IRF), expand upon classical goodness-of-fit tests designed for cumulative density function (CDF) specifications and apply them to the regression model checking. The IRF generalizes the one-to-one relationship between density and the CDF to the regression framework, enabling the extension of all the tests proposed for CDFs to the context of regression. The alternative test statistics are characterized by a functional of the standard empirical process (SEP), which is defined as the difference between the empirical IRF/CDF and its restricted counterpart under the null hypothesis, suitably scaled.

Omnibus tests are designed to assess whether any deviations of the SEP from zero arise by chance rather than due to model misspecification. Classical test statistics for CDF specifications, which are based on a norm of the SEP, such as the well-known Kolmogorov-Smirnov, Cramér-von Mises, or Anderson-Darling tests, have been extended to the context of regression model specifications by Bierens (1982), Stute (1997), and Andrews (1997), among others. Several tests are based on transformations of the SEP. For instance, test statistics based on Fourier transforms of the SEP, which in fact compare characteristic functions (e.g., Koutrouvelis and Kellermeier 1981), have been adapted to the regression case by Bierens (1982) and Bierens and Ploberger (1997). Tests based on martingale transforms of the SEP (Khmaladze 1982) have been extended by Khmaladze and Koul (2004). Additionally, specification tests for Lebesgue density specification, which rely on the convolution of the SEP and a kernel function (e.g., Bickel and Rosenblatt 1973), have been extended to regression models by Hardle and Mammen (1993). All of these tests are known as minimum-distance tests (Pollard 1980).

When the model parameters are estimated from the sample, the SEP converges in distribution to a case-dependent process, and the tests are implemented using bootstrap

techniques. Two exceptions exist: (i) test statistics based on the SEP martingale transform, which converge in distribution to a Brownian motion with tabulated critical values; and (ii) test statistics based on the convolution of the SEP and a kernel function, suitably centered and standardized, with a standard normal limit distribution under the null when the bandwidth parameter vanishes at a suitable rate as the sample size diverges. However, the size properties of tests based on convolutions depend on the bandwidth choice, and in practice, the tests are implemented using bootstrap.

It is well known that tests based on the SEP exhibit poor power properties in the direction of high-frequency alternatives, as pointed out by Durbin and Knott (1972) based on the study of the SEP's principal components. Stute (1997) generalized these results to the regression case. Tests that rely on convolutions of the SEP and a kernel, with a vanishing bandwidth, are capable of detecting high-frequency alternatives but fail to detect the standard departures converging to the model in the null at the parametric rate. All of these tests are affected by the curse of dimensionality, showing limited power when the number of explanatory variables is large.

In practice, χ^2 tests, such as the Pearson (1900) primeval test and its subsequent versions, are among the most popular goodness-of-fit tests for CDF model specifications. These tests, after partitioning the data into a number of cells, say L, evaluates whether the difference between the observed and the expected frequencies within each cell occur by chance. The Pearson test statistics, for instance, can be represented as a quadratic form in the vector of differences between observed and expected frequencies. Note that the χ^2 tests are not omnibus: their goal is to detect alternatives where the true probability of one observation falling into a particular cell differs from the expected value specified by the model. However, they have important advantages with respect to tests based on SEP functionals. They are distribution-free and do not require the use of bootstrapping or smoothing techniques. Additionally, these tests are invariant to sample-dependent partitions, which can then be chosen to favor certain alternatives of interest. Also, they are computationally friendly.

Surprisingly, the χ^2 tests have not been extended to check the specification of CMR. This paper aims to fill this gap by proposing to group the data according to a partition of the explanatory variables into L classes, and considering χ^2 tests given by quadratic

forms in the vector of differences between observed and expected (according to the model) sample averages within each class. The Pearson test analog is a weighted sum of these squared differences, which is in fact a J test for the orthogonality conditions between regression errors and the explanatory variables in each of the L classes. The paper also considers a generalized Wald test statistic: a quadratic form in the vector of differences between observed and expected sample averages using any $\sqrt{n}$-consistent estimator of the parameters in the model.

Much like the classical χ^2 test statistic, these tests have a limiting null distribution which is invariant to sample-dependent partitions. I propose algorithms for the calculation of balanced cells, each designed to contain approximately the same number of observations, as well as unbalanced partitions intended to favor specific alternatives. I consider tests based on Neyman-Pearson (NP) partitions built upon deviations of the model fits under the null hypothesis from a predefined alternative, see Balakrishnan, Voinov, and Nikulin (2013). Under the correct specification of the alternative model, the NP partition maximizes the distance between the null and alternative models, thereby enhancing the power of the tests. In cases where there are no specific alternatives to prioritize, the recommendation is to utilize NP cells that compare the model under the null hypothesis with an auxiliary, flexible specification of the regression model.

Monte Carlo simulations show the good performances of the χ^2 tests compared to omnibus tests, particularly when the covariates dimension is high. The finite-sample study is complemented with an empirical illustration in which I apply the tests to analyze the returns of attending historically black college and universities (HBCU), relative to non-HCBU, for black students in the United States.

The structure of the paper is as follows. In the next section, I introduce the tests focusing on regression specifications. This facilitates the motivation and presentation and reduces the notational burden required for the more general case. In Section 3, I discuss the asymptotic properties of the tests using sample dependent partitions. Section 4 introduces partitioning algorithms designed to enhance the power of the tests. Section 5 analyzes Pitman's local power. In Section 6, I extend the discussion to general CMR. Section 7 presents a Monte Carlo study. In the last section, I provide a real-data application.

2 Chi-Squared Tests for CMR

Let $\{Z_i\}_{i=1}^n = \{(Y_i, X_i')'\}_{i=1}^n$ be an i.i.d sample from the $\mathbb{R}^{1+d_x}$-valued random vector $Z = (Y, X')'$ with distribution P, where Y is the response variable, and X is the d_x-dimensional vector of explanatory variables taking values in $\mathcal{X} \subset \mathbb{R}^{d_x}$. The correct specification hypothesis is stated as,

$$H_0 : \ m \in \mathcal{M} \tag{1}$$

with $m(\cdot) = \mathbb{E}[Y|X = \cdot]$ and $\mathcal{M} = \{m_\theta(\cdot) : \theta \in \Theta \subset \mathbb{R}^{d_\theta}\}$ being a family of parametric regression functions and a suitable parameter space, respectively. Thus, under H_0, there exists a $\theta_0 \in \Theta$ such that the IRF, $M(x) = \mathbb{E}\left[Y\mathbb{I}_{(-\infty,x]}(X)\right]$, with $\mathbb{I}_{(-\infty,x]}(X) = \prod_{j=1}^{d_x} \mathbb{I}_{(-\infty,x_j]}(X_j)$, and its version imposing the null hypothesis restrictions, $M_\theta(x) = \mathbb{E}\left[m_\theta(X)\mathbb{I}_{(-\infty,x]}(X)\right]$, are equal in all the Borel sets of $\mathcal{X}$, i.e.

$$M\{A\} = M_{\theta_0}\{A\} \ \text{ for all } A \in \mathcal{B}^{d_x}, \tag{2}$$

where $\mathcal{B}^{d_x}$ represent the Borel sigma-field of $\mathbb{R}^{d_x}$, $M\{A\} = \int_A M(dx)$ and $M_\theta\{A\} = \int_A M_\theta(dx)$ denote the Lebesgue-Stieltjes measure of M and M_θ over A, respectively. Recall that (2) is the definition of the regression function for the specification in H_0 (e.g., see definition 34.1 in Billingsley (2017)), which is equivalent to the following orthogonality conditions,

$$H_0 : \ \mathbb{E}[\varepsilon_{\theta_0}|X] = 0 \ \text{ a.s. for some } \theta_0 \in \Theta,$$

with $\varepsilon_\theta(z) = y - m_\theta(x)$, $z = (y, x')'$, denoting the regression error.
Intuitively, when X takes values in a finite set and θ_0 is known (i.e., under the simple hypothesis), a test for H_0, following proposal Pearson (1900) proposal, consists of comparing the conditional mean with the assumed regression model for all $x \in \mathcal{X}$. In this case, such a test is indeed omnibus (i.e. the test detect all the possible alternatives). For any type of covariates, once the data is partitioned into L cells, say, χ^2 tests assess whether the difference between the expected and observed averages in each cell arose by chance.

Consider the partition $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)$ of the covariates space $\mathcal{X}$ into L cells and let $\mathbf{I}_{\boldsymbol{\gamma}}(x) = (\mathbb{I}_{\gamma_1}(x), \ldots, \mathbb{I}_{\gamma_L}(x))'$ denote the vector of indicator functions over the sets within $\boldsymbol{\gamma}$. The building block of the χ^2 test statistics is the standardized vector of differences

between the observed averages of Y and the corresponding averages under the specification of H_0 in each cell,

$$\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\theta) = \sqrt{n}\left(\hat{\boldsymbol{\mu}}^0_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\gamma}}(\theta)\right) = \sqrt{n}\left(\hat{\mu}^0_1 - \hat{\mu}_1(\theta), \ldots, \hat{\mu}^0_L - \hat{\mu}_L(\theta)\right)', \tag{3}$$

where $\hat{\mu}^0_l = \hat{M}\{\gamma_l\} = n^{-1}\sum_{i=1}^n Y_i \mathbb{I}_{\gamma_l}(X_i)$ is the empirical IRF evaluated at γ_l, with $\hat{M}(x) = n^{-1}\sum_{i=1}^n Y_i \mathbb{I}_{(-\infty,x]}(X_i)$, and $\hat{\mu}_l(\theta) = \hat{M}_\theta\{\gamma_l\} = n^{-1}\sum_{i=1}^n m_\theta(X_i)\mathbb{I}_{\gamma_l}(X_i)$ is its restricted version under H_0, with $\hat{M}_\theta(x) = n^{-1}\sum_{i=1}^n m_\theta(X_i)\mathbb{I}_{(-\infty,x]}(X_i)$.

Under a simple hypothesis, i.e. when θ_0 is known, by the central limit theorem, under the null,

$$\hat{\Sigma}_{\boldsymbol{\gamma}}(\theta_0)^{-1/2}\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\theta_0) \xrightarrow{d} N(0, I_L), \tag{4}$$

where $\hat{\Sigma}_{\boldsymbol{\gamma}}(\theta) = n^{-1}\sum_{i=1}^n \varepsilon^2_\theta(Z_i)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'$ estimates,

$$\Sigma_{\boldsymbol{\gamma},0} = \text{Avar}\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\theta_0)\right) = \text{diag}\{\sigma^2_{0,1}, ..., \sigma^2_{0,L}\}, \tag{5}$$

under H_0, with $\sigma^2_{0,l} = \sigma^2_l(\theta_0)$, and $\sigma^2_l(\theta) = \mathbb{E}\left[\varepsilon^2_\theta(Z)\mathbb{I}_{\gamma_l}(X)\right]$. Thus, taking for granted that $\rho_l = \mathbb{E}\left[\mathbb{I}_{\gamma_l}(X)\right] > 0$ for all l, under H_0,

$$\hat{\chi}^2_{\boldsymbol{\gamma},0}(\theta_0) \xrightarrow{d} \chi^2_L, \tag{6}$$

where

$$\hat{\chi}^2_{\boldsymbol{\gamma},0}(\theta) = \hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\theta)'\hat{\Sigma}_{\boldsymbol{\gamma}}(\theta_0)^{-1}\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\theta) = n\sum_{l=1}^L \frac{\left(\hat{\mu}^0_l - \hat{\mu}_l(\theta)\right)^2}{\hat{\sigma}^2_{0,l}}, \tag{7}$$

and $\hat{\sigma}^2_{0,l} = n^{-1}\sum_{i=1}^n \varepsilon^2_{\theta_0}(Z_i)\mathbb{I}_{\gamma_l}(X_i)$. The test statistic $\hat{\chi}^2_{\boldsymbol{\gamma},0}(\theta_0)$ extends the classical Pearson's chi-squared test to check regression model specifications. It does so by replacing the comparison of observed frequencies with expected frequencies under the null hypothesis with a comparison of averages.

Of course, tests based on (3) are not omnibus, but designed for detecting deviations from H_0 of the type,

$$H_1(\boldsymbol{\gamma}) : \boldsymbol{\mu}^0_{\boldsymbol{\gamma}} \neq \boldsymbol{\mu}_{\boldsymbol{\gamma}}(\theta) \;\text{ for all } \theta \in \Theta, \tag{8}$$

where $\boldsymbol{\mu}^0_{\boldsymbol{\gamma}} = (\mu^0_{\gamma_1}, ..., \mu^0_{\gamma_L})' = (M\{\gamma_1\}, \ldots, M\{\gamma_L\})'$ is the vector of expected averages of

Y in each cell and $\boldsymbol{\mu}_{\boldsymbol{\gamma}}(\theta) = (\mu_{\gamma_1}(\theta), ..., \mu_{\gamma_L}(\theta))' = (M_\theta\{\gamma_1\}, \ldots, M_\theta\{\gamma_L\})'$ is its restricted version under H_0. The partitions cover a fundamental role in implementing the test and provide a flexible tool to exploit out-of-sample information on the possible alternatives. Neyman-Pearson (NP) cells (Balakrishnan, Voinov, and Nikulin 2013), for instance, consist of the points in $\mathcal{X}$ where the specification under the null and a pre-specified alternative coincide (see Example 1 below). Under the pre-specified alternative, NP classes maximize the L_2 norm of $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}$.

Example 1 (NP Classes)

Consider testing the linear model $m_\theta(X_i) = \theta_0 + \theta_1 X_i$ against the alternative specification,

$$H_1 : m_{1,\theta^*}(X_i) = \theta_0 + \theta_1 X_i + \theta_2^* \sin\left(\frac{50 X_i}{2\pi}\right)$$

where $\theta^* = (\theta, \theta_2^*)$ is a known vector. NP classes split $\mathcal{X}$ over the points where $\theta_2^* \sin\left(\frac{50x}{2\pi}\right) = 0$ and, hence, $m_\theta(\cdot) = m_{1,\theta^*}(\cdot)$. As a result, under H_1, $m_\theta(\cdot)$ is strictly bigger or strictly smaller than $m_{1,\theta^*}(\cdot)$ within each cell, and most cell-specific errors have the same sign, implying that the average error of a single cell is larger than the average error over the union of two contiguous cells (in absolute terms). See Figure 1 for a graphical representation of the partition.

When θ_0 is unknown, the criterion in (7) suggests the following minimum distance estimator, hereafter referred to as the grouped GMM estimator:

$$\hat{\theta}_{\boldsymbol{\gamma}} = \arg\min_{\theta \in \Theta} \chi^2_{\boldsymbol{\gamma}}(\theta) \tag{9}$$

where

$$\hat{\chi}^2_{\boldsymbol{\gamma}}(\theta) = \hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\theta)' \hat{\Sigma}_{\boldsymbol{\gamma}}(\tilde{\theta})^{-1} \hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\theta) = n \sum_{l=1}^{L} \frac{\left(\hat{\mu}_l^0 - \hat{\mu}_l(\theta)\right)^2}{\hat{\sigma}_l^2(\tilde{\theta})},$$

$\hat{\sigma}_l^2(\theta) = n^{-1} \sum_{i=1}^n \varepsilon_\theta^2(Z_i) \mathbb{I}_{\gamma_l}(X_i)$, and $\tilde{\theta}$ is some initial $\sqrt{n}$-consistent estimator of θ_0. The estimator $\hat{\theta}_{\boldsymbol{\gamma}}$ is analogous to the multinomial maximum-likelihood estimator (or minimum χ^2 estimator) in the classical case (see Cramér 1946). Under linear null hypothesis, i.e. $m_\theta(x) = x'\theta$, this corresponds to the feasible GLS estimator based on the aggregated data

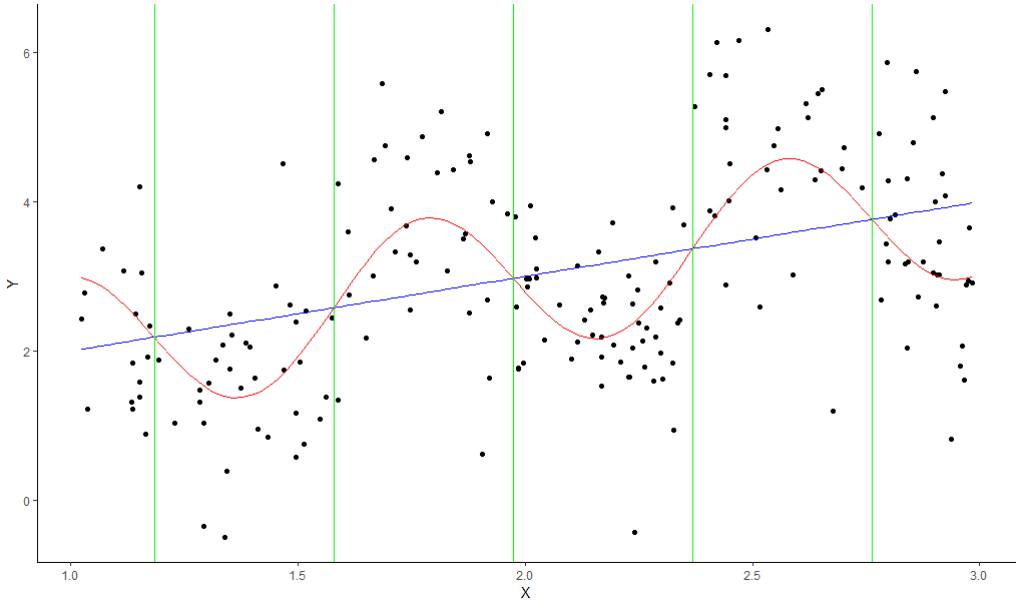


Figure 1: The graph depicts a random draw from the model under H_1 with $\theta_0^* = (1, 1, 1)$. The green lines depict the points where the model under the null (blue line) and under the alternative (red line) meet.

$\{\bar{Y}_l, \bar{X}_l\}_{l=1}^L$,

$$\hat{\theta}_{\boldsymbol{\gamma}} = \left[\sum_{l=1}^{L} \frac{\bar{X}_l \bar{X}_l'}{\hat{\sigma}_l^2(\tilde{\theta})}\right]^{-1} \sum_{l=1}^{L} \frac{\bar{X}_l \bar{Y}_l}{\hat{\sigma}_l^2(\tilde{\theta})},$$

with $\bar{Y}_l = n^{-1} \sum_{i=1}^n Y_i \mathbb{I}_{\gamma_l}(X_i)$ and $\bar{X}_l = n^{-1} \sum_{i=1}^n X_i \mathbb{I}_{\gamma_l}(X_i)$. In the non-linear case, we can iterate a feasible asymptotically efficient Gauss-Newton estimator starting from any preliminary $\sqrt{n}$-consistent estimator $\tilde{\theta}$,

$$\hat{\theta}_{\boldsymbol{\gamma}}^{(1)} = \tilde{\theta} + \left[\sum_{l=1}^{L} \frac{\hat{\mu}_l^*(\tilde{\theta}) \hat{\mu}_l^*(\tilde{\theta})'}{\hat{\sigma}_l^2(\tilde{\theta})}\right]^{-1} \sum_{l=1}^{L} \frac{\hat{\mu}_l^*(\tilde{\theta})(\hat{\mu}_l^0 - \hat{\mu}_l(\tilde{\theta}))}{\hat{\sigma}_l^2(\tilde{\theta})} \tag{10}$$

with $\hat{\mu}_l^*(\theta) = n^{-1} \sum_{i=1}^n \nabla m_\theta(X_i) \mathbb{I}_{\gamma_l}(X_i)$ and $\nabla m_{\bar{\theta}} = d/d\theta' m_\theta|_{\theta=\bar{\theta}}$. The estimator belongs to the class of minimum-distance estimators considered by Cristobal, Roca, and Manteiga (1987) and Koul and Ni (2004), with the main difference that the regressogram is used instead of kernels, and the weighting introduced to improve efficiency. Under suitable regularity conditions and under H_0,

$$\sqrt{n}(\hat{\theta}_{\boldsymbol{\gamma}} - \theta_0) \xrightarrow{d} N\left(0, \left[\boldsymbol{\mu}_{\boldsymbol{\gamma},0}^{*\prime} (\Sigma_{\boldsymbol{\gamma},0})^{-1} \boldsymbol{\mu}_{\boldsymbol{\gamma},0}^*\right]^{-1}\right),$$

where $\boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} = \boldsymbol{\mu}^*_{\boldsymbol{\gamma}}(\theta_0)$, with $\boldsymbol{\mu}^*_{\boldsymbol{\gamma}}(\theta) = (\mu^*_1(\theta), ..., \mu^*_L(\theta))'$ and $\mu^*_l(\theta) = \mathbb{E}\left[\hat{\mu}^*_l(\theta)\right]$ denoting the matrix of partial derivatives of $\boldsymbol{\mu}_{\boldsymbol{\gamma}}(\theta)$.

The $\hat{\chi}^2$ test statistics is, for $L > d_\theta$,

$$\hat{\chi}^2_{\boldsymbol{\gamma}} = \min_{\theta \in \Theta} \hat{\chi}^2_{\boldsymbol{\gamma}}(\theta) = \hat{\chi}^2_{\boldsymbol{\gamma}}(\hat{\theta}_{\boldsymbol{\gamma}}), \tag{11}$$

which is in fact a J test on the set of the L, out of the (possibly) many, orthogonality conditions implied by the null,

$$\mathbb{E}\big[Y \mathbb{I}_{\gamma_l}(X)\big] = \mathbb{E}\big[m_{\theta_0}(X) \mathbb{I}_{\gamma_l}(X)\big] \text{ for all } l \in \{1, 2, .., L\}.$$

Thus, under H_0, and for $L > d_\theta$,

$$\hat{\chi}^2_{\boldsymbol{\gamma}} \xrightarrow{d} \chi^2_{L - d_\theta}. \tag{12}$$

Is also well motivated, as suggested in classical goodness-of-fit χ^2 tests (e.g. Nikulin 1973 and Rao and Robson 1974), using the Wald testing principle based on $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})$, employing any $\sqrt{n}$-consistent estimator $\tilde{\theta}$,

$$\hat{\mathcal{W}}_{\boldsymbol{\gamma}}(\tilde{\theta}) = \hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta}) \widehat{\text{Avar}^-}\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) \hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta}), \tag{13}$$

where $\widehat{\text{Avar}^-}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta}))$ is a consistent estimator of some generalized inverse of $\text{Avar}\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$, $\text{Avar}^-\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$ say. Assuming $\widehat{\text{Avar}^-}\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) \xrightarrow{p} \text{Avar}^-\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$, and suitable regularity conditions, under H_0,

$$\hat{\mathcal{W}}_{\boldsymbol{\gamma}}(\tilde{\theta}) \xrightarrow{d} \chi^2_{r(\text{Avar}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})))}, \tag{14}$$

where for a given square matrix A, $r(A)$ denotes its rank.

Taking for granted the asymptotic linearity of the estimator (see Assumption 3 in the next section), the covariance matrix of $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})$ is characterized as,

$$\text{Avar}\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) = \Sigma_{\boldsymbol{\gamma},0} - \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} C'_{\boldsymbol{\gamma},0} - C_{\boldsymbol{\gamma},0} \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0} + \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} L_0 \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0}, \tag{15}$$

where $C_{\boldsymbol{\gamma},0} = \mathbb{E}\left[\varepsilon_{\theta_0}(Z) \mathbf{I}_{\boldsymbol{\gamma}}(X) l_{\theta_0}(Z)'\right]$, $L_0 = \mathbb{E}\left[l_{\theta_0}(Z) l_{\theta_0}(Z)'\right]$, and $l_{\theta_0}(\cdot)$ is the influence

function of $\tilde{\theta}$. When Avar $\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$ is full rank (e.g., if $\varepsilon_{\theta_0}(\cdot)\mathbf{I}_{\boldsymbol{\gamma}}(\cdot)$ and $l_{\theta_0}(\cdot)$ have linearly independent components), the Wald test can be performed on any finite splitting of the data. In this case, a valid choice of $\widehat{\text{Avar}^{-}}\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$ is given by the inverse of,

$$\hat{W}_{\boldsymbol{\gamma}}(\tilde{\theta}) = \hat{\Sigma}_{\boldsymbol{\gamma}}(\tilde{\theta}) - \hat{\boldsymbol{\mu}}^{*}_{\boldsymbol{\gamma}}(\tilde{\theta})\hat{C}_{\boldsymbol{\gamma}}(\tilde{\theta})' - \hat{C}_{\boldsymbol{\gamma}}(\tilde{\theta})\hat{\boldsymbol{\mu}}^{*'}_{\boldsymbol{\gamma}}(\tilde{\theta}) + \hat{\boldsymbol{\mu}}^{*}_{\boldsymbol{\gamma}}(\tilde{\theta})\hat{L}(\tilde{\theta})\hat{\boldsymbol{\mu}}^{*'}_{\boldsymbol{\gamma}}(\tilde{\theta}), \tag{16}$$

where $\hat{\boldsymbol{\mu}}^{*}_{\boldsymbol{\gamma}}(\theta) = (\hat{\mu}^{*}_{1}(\theta), ..., \hat{\mu}^{*}_{L}(\theta))'$, $\hat{C}_{\boldsymbol{\gamma}}(\theta) = n^{-1}\sum_{i=1}^{n}\varepsilon_{\theta}(Z_i)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)l_{\theta}(Z_i)'$, and $\hat{L}(\theta) = n^{-1}\sum_{i=1}^{n} l_{\theta}(Z_i)l_{\theta}(Z_i)'$. If the covariance matrix is rank deficient and the Moore-Penrose inverse of $\hat{W}_{\boldsymbol{\gamma}}(\tilde{\theta})$, denoted as $\hat{W}^{+}_{\boldsymbol{\gamma}}(\tilde{\theta})$, has rank converging in probability to the one of Avar $\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$, then $\hat{W}^{+}_{\boldsymbol{\gamma}}(\tilde{\theta}) \xrightarrow{p}$ Avar$^{+}\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$ (Theorem 2 of Andrews 1987). However, this need not be the case (Schott 2016, p. 222-224) and more complex methods might be required (see, e.g., Lütkepohl and Burda 1997). Consider, for instance, the nonlinear least squares (NLLS),

$$\tilde{\theta} = \arg\min_{\theta\in\Theta}\sum_{i=1}^{n}\varepsilon_{\theta}(Z_i)^2.$$

Under homoskedasticity, i.e. $\text{Var}(\varepsilon_{\theta_0}(Z)|X = x) = \sigma^2_0 = \sum_{l=1}^{L}\sigma^2_{0,l}$, it holds that, $l_{\theta_0}(z) = \Psi_0^{-1}\nabla m_{\theta_0}(x)\varepsilon_{\theta_0}(z)$, where $\Psi_0 = \mathbb{E}\left[\nabla m_{\theta_0}(X)\nabla m_{\theta_0}(X)'\right]$. Under H_0, when $\tilde{\theta}$ is asymptotically more efficient than $\hat{\theta}_{\boldsymbol{\gamma}}$, i.e., when $\text{Avar}(\hat{\theta}_{\gamma}) - \text{Avar}(\tilde{\theta})$ is p.d., Avar $\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) = \Sigma_{\boldsymbol{\gamma},0} - \sigma^2_0\boldsymbol{\mu}^{*}_{\boldsymbol{\gamma},0}\Psi_0^{-1}\boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0}$ is also p.d., and $\hat{\mathcal{W}}(\tilde{\theta}) \xrightarrow{d} \chi_L$. A similar reasoning holds for the probit/logit maximum likelihood estimator (MLE), as discussed in Section 8. Under heteroskedasticity, however, Avar $\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$ is not necessarily an invertible matrix.

Notice that the estimation of the covariance matrix, when Avar $\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$ is full rank, can be avoided by using a random normalizing weighting matrix, as in Kuan and Lee (2006) (see also Kiefer, Vogelsang, and Bunzel 2000 for an early reference). Similarly, the over-identification test of Lee, Kuan, and Hsu (2014) provides a robust version of the $\hat{\chi}^2$ test which does not require estimating $\hat{\Sigma}_{\boldsymbol{\gamma}}(\tilde{\theta})$. In these cases, the limit null distribution is non-standard, but pivotal.

3 Data-dependent Cells

When studying the large sample behavior of the statistics, it is crucial to address the inherent influence of the data on the selection of cells (Watson 1959). Moore and Spruill (1975) were among the first to address this concern in the distribution model check literature providing a rigorous derivation of the null distribution of χ^2 tests with rectangular data-dependent cells. In a more general setting, Pollard (1979) established the result for cells of arbitrary form using uniform results for empirical processes indexed by sets, while Andrews (1988a,b) and Delgado and Vainora (2023) applied the methodology to conditional distribution testing. This section provides a similar result for the more general CMR testing framework, showing that the grouped GMM estimator and the tests keep standard limit distribution when the partition is built with data-dependent cells.

A minimal set of assumptions, consisting of smoothness conditions and restrictions on the partitioning algorithm complexity, allows to state the convergence results in a self-contained fashion.

Assumption 1 *(a)* $\{Z_i = (Y_i, X_i')'\}_{i=1}^n$ *is a sequence of i.i.d. random vectors with* $\mathbb{E}|Y_i| < \infty$*; (b)* $\mathbb{E}\left[\varepsilon_{\theta_0}^2\right] < C$*, with* $C < \infty$*; (c)* Θ *is a compact subset of* $\mathbb{R}^{d_\theta}$ *and* θ_0 *is an interior point of* Θ*.*

Assumption 2 $m_\theta(\cdot)$ *is twice continuously differentiable in a neighborhood* $\mathcal{N}_{\theta_0}$ *of* θ_0*, with* $\mathcal{N}_{\theta_0} \subset \Theta$*. The gradient,* $\nabla m_\theta(\cdot) = d/d\theta m_\theta(\cdot)$*, is bounded by a square-integrable function* $R(\cdot)$ *such that* $\sup_{\theta\in\Theta_0} |\nabla^{(j)} m_\theta(\cdot)| \leq R(\cdot)$ *for all* $j \in \{1, .., d_\theta\}$*, where* $\nabla^{(j)}$ *denotes the* j*-th partial derivative, and* $\mathbb{E}\left[R(X)^2\right] < \infty$*.*

Assumption 3

(a) The estimator $\tilde{\theta}$ *satisfies the following asymptotic expansion under the null,*

$$\sqrt{n}(\tilde{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_{\theta_0}(Z_i) + o_p(1)$$

where $\mathbb{E}\left[l_{\theta_0}(Z)\right] = 0$ *and* $L_0 = \mathbb{E}\left[l_{\theta_0}(Z) l_{\theta_0}(Z)'\right]$ *is a finite and non-singular matrix.*

(b) The vector-valued function $l_\theta(\cdot)$ *is twice continuously differentiable in a neighborhood*

Θ_0 *of* θ_0 *with first partial derivatives bounded by a square-integrable function* $R_2(\cdot)$ *such that* $\sup_{\theta\in\Theta_0} |\nabla^{(j)} l_\theta(\cdot)| \leq R_2(\cdot)$ *for all* $j \in \{1, .., d_\theta\}$ *and* $\mathbb{E}\left[R_2(Z)^2\right] < \infty$.

Assumptions 1 and 2 are common in the model check literature (see Stute and Zhu 2002, for instance). Compared to papers developing χ^2 tests based on a probability model (e.g., Tauchen 1985), these require slightly higher smoothness conditions of the regression function, but leave completely unrestricted the data distribution. Assumption 3(a) holds for most of the estimators used in practice, such as least squares or GMM estimators, as well as for identification-robust minimum-distance estimators (see, e.g., Domínguez and Lobato 2004). While Assumption 3(b) is a technical requirement for the consistency of the plug-in estimator $\hat{W}_{\boldsymbol{\gamma}}(\tilde{\theta})$.

Let also state the necessary global identification and finite-variance conditions for the consistency and asymptotic normality of the grouped GMM estimator.

Assumption 2' *(a)* $\Sigma_{\boldsymbol{\gamma},0}$ *is positive definite (p.d.); (b)* $\mathbb{E}\left[Y\mathbf{I}_{\boldsymbol{\gamma}}\right] = \mathbb{E}\left[m_\theta(X)\mathbf{I}_{\boldsymbol{\gamma}}\right]$ *if and only if* $\theta = \theta_0$*; (c)* $\boldsymbol{\mu}^*_{\boldsymbol{\gamma},0}$ *is full rank.*

Notice that identification depends crucially on the partitioning choice $\boldsymbol{\gamma}$ and might not hold in practice. In the next section, I discuss instances where the assumption fails and how to avoid them.

Following Pollard (1979) and Andrews (1988a), the data-dependent partitions are modeled as random functions over a class of properly restricted measurable sets. Specifically, let $\mathbb{C}$ be a class of measurable sets in $\mathcal{X}$ from which the cells of each partition are drawn, and denote as $\mathbb{D}$ the class of partitions of $\mathcal{X}$ comprised of L sets from $\mathbb{C}$ (L is fixed for all n); that is,

$$\mathbb{D} = \left\{ \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_L) \in \mathbb{C}^L : \bigcup_{l=1}^{L} \gamma_l = \mathcal{X},\ \gamma_l \bigcap \gamma_f = \emptyset,\ \forall l \neq f \right\}, \tag{17}$$

where γ_l and γ_f denote sets of the partition $\boldsymbol{\gamma}$. We equip $\mathbb{C}$ with the topology generated by the $L_2(F_x)$ semi-norm, F_x being the distribution of X under P, and give $\mathbb{D}$ the corresponding product topology. This means that two set C_1, C_2 in $\mathcal{X}$ are close if $F_x\{C_1 \Delta C_2\}$ is small, where Δ denotes the symmetric difference operator, $C_1 \Delta C_2 = C_1 \cup C_2 \backslash C_1 \cap C_2$.

For each sample size n the corresponding partition is determined by a measurable mapping, denoted as $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_L)$, from the underlying probability space to $\mathbb{D}$, which converges in probability to some fixed partition of cells in $\mathbb{D}$. In other words, for all $\epsilon > 0$,

$$P\left(F_x\{\hat{\gamma}_l \Delta \gamma_l\} > \epsilon\right) \to 0 \text{ as } n \to \infty, \text{ for all } l = 1, 2, ..., L.$$

Assumption 4 *Under H_0, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}$ for some fixed set of cells $\boldsymbol{\gamma} \in \mathbb{D}$*

Assumption 4 represents a standard requirement in the literature concerning empirical processes indexed by sets, and it is satisfied by a broad range of partitioning algorithms. For example, this assumption is met when the cells depend on $\tilde{\theta}$ in a continuous fashion, and $\tilde{\theta}$ converges in probability to a non-random vector (Andrews 1988b). However, it is less clear whether cells generated by the comparison of different estimators meet this requirement, as discussed in the next section.

It is also possible to relax Assumption 4 to cases where $\hat{\boldsymbol{\gamma}}$ converges to a random element $\boldsymbol{\gamma}_0 \in \mathbb{D}$. This is achieved by replacing the convergence in probability with a uniform tightness condition and ensuring asymptotic independence between $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\theta_0)$. Refer to Section 6 of Pollard (1979) for further discussions.

Crucially, deriving the limit null distribution requires bounding the complexity of the partitions employed for test construction. This is achieved by assuming that the cells are drawn from a Vapnik-Cervonenkis (VC) class.

Assumption 5 $\mathbb{C}$ *is a VC class of sets.*

This assumption is convenient because it is independent of the data distribution P, yet it is general enough for our purposes. For example, algorithms that generate cells with a finite number of straight edges and the class of hyper ellipsoids are VC classes. Furthermore, unions, intersections, differences, and complements of VC classes are also VC classes (Andrews 1988a and Pollard 1984 provide a thorough discussion, see also Section 2.6 in Van Der Vaart 1996). A less stringent condition assumes that $\mathbb{C}$ is a Donsker class for the underlying probability measure (Pollard 1979). While this allows for a wider range of admissible partitions, verifying Donsker assumptions may prove more challenging in practice.

The following theorems show that the asymptotic distribution of the grouped GMM estimator and of the test statistics are unaffected by data-dependent cells. All the proofs are relegated to the appendix.

Theorem 1 *Let Assumptions 1, 2, 2', 4, 5 hold. Then, under the null hypothesis* H_0*,*

$$\sqrt{n}(\hat{\theta}_{\hat{\boldsymbol{\gamma}}} - \theta_0) \xrightarrow{d} N\left(0, \left[\boldsymbol{\mu}^{*\prime}_{\boldsymbol{\gamma},0}(\Sigma_{\boldsymbol{\gamma},0})^{-1}\boldsymbol{\mu}^{*}_{\boldsymbol{\gamma},0}\right]^{-1}\right),$$

Theorem 2 *Let Assumptions 1, 2, 4, 5, and the null hypothesis* H_0 *hold. Then,*

(a) Under Assumption 2' and $L > d_\theta$*,*

$$\hat{\chi}^2_{\hat{\boldsymbol{\gamma}}} \xrightarrow{d} \chi^2_{L-d_\theta}.$$

(b) Under Assumption 3, and $\widehat{\text{Avar}^-}\left(\hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})\right) \xrightarrow{p} \text{Avar}^-\left(\hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})\right)$*,*

$$\hat{\mathcal{W}}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta}) \xrightarrow{d} \chi^2_{r(\text{Avar}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})))}.$$

The upcoming section introduces a variety of partitioning procedures, discussing methods based on unsupervised clustering and model-based techniques that leverage information on the null hypothesis and pre-specified alternatives.

4 Partitioning Procedures

The partitioning algorithms can be divided into covariate-based methods and model-based methods, depending on the information used to split the sample. The first classifies the data exclusively based on the characteristics of the population and does not have a predictable effect on the power of the test. In contrast, model-based methods exploit information about the model under the null hypothesis and pre-specified alternatives to increase the power of the tests in those directions. The objective of these procedures is to construct partitions where the distance between the vectors of average predictions under the null and alternative hypotheses is maximized.

In this section, I first briefly explore the construction of partitions using covariate-based methods and the limitations related to the local identification of the $\hat{\chi}^2$ test. Next, the discussion turns to the construction of Neyman-Pearson (NP) classes. These partitioning algorithms aim to maximize the probability of rejecting the null hypothesis by comparing the parametric fits under the null and the pre-specified alternative. This method is particularly useful when the researcher's primary concern is rejecting a subset of deviations critical for the application at hand.

When the alternative is left unrestricted, the recommendation is to construct NP classes using an auxiliary flexible specification of the regression model (see Davidson, MacKinnon, et al. 2004 Ch. 15.2 for a review of tests based on artificial regressions). The procedure assumes a structured alternative model, which allows to capture deviations from the null hypothesis by exploiting the conditional mean dependence of the parametric residuals from the fitted values under the alternative.

Lastly, I discuss the conditions for convergence in probability of NP classes to fixed cells in $\mathbb{D}$. Since these partitioning methods rely on comparing different estimators, understanding the convergence to fixed cells is not straightforward and requires additional investigation. Conditions for covariate-based partitioning methods are already discussed in Andrews (1988a) and Andrews (1988b).

4.1 Covariate-based Partitioning Methods

Partitioning methods based on covariate encompass both straightforward approaches, such as dividing the data into cubic cells, and more sophisticated techniques like k-means clustering, hierarchical clustering, and nearest-neighbor clustering, among others (refer to Chapter 14 of Hastie, Tibshirani, and Friedman 2009 for a review of unsupervised clustering methods).

In the testing context, statistical practice involves dividing data into equiprobable cells when testing against unknown probability distributions (see, for instance, Greenwood and Nikulin 1996). This can be achieved through various methods, including partitioning into unions of statistically equivalent blocks (SEB), as discussed by Gessaman (1970). The procedure, applied to an $n \times d_F$ design matrix, involves sorting observations by column

values and grouping them into S blocks at each iteration. Here, S is a user-chosen parameter, resulting in a partition with S^{d_F} cells. To avoid generating excessively fine partitions, instead of applying the algorithm directly to the design matrix, $\mathbb{X}$, one can apply the algorithm to a lower-dimensional projection of the data, such as the first q principal components of $\mathbb{X}$ or the fitted values of the parametric model under the null hypothesis. In the latter case, the algorithm is equivalent to splitting the data based on the S quantile of $m_{\hat{\theta}}(X_i)_{i=1}^n$.

In practice, as discussed in the next sub-section, these methods are implemented in conjunction with other model-based partitioning algorithms. Their primary objective is to generate additional splits for the preliminary two-cell partition created by the model-based methods, which are subsequently used to construct the final partition.

However, it's important to note that the $\hat{\chi}^2$ requires the L moment conditions generated by the partition to identify θ_0, as outlined in Assumption 2'. This imposes additional restrictions on the set of feasible partitioning algorithms. For example, consider a scenario where $X = (X_1, \ldots, X_{d_x})'$, with $d_x > 2$, and X_j being mutually independent and identically distributed random variables for each $j = 1, ..., d_x$. In the linear setting, $H_0 : m(x) = \theta_0' x$, Assumption 2'(c) - i.e., the matrix of partial derivatives $\mu^*_{\gamma,0}$ is of full rank - is equivalent to assuming that the vector of grouped covariates,

$$\mu^*_{\gamma,0} = \mathbb{E}\left[\mathbf{I}_{\boldsymbol{\gamma}}(X)X'\right],$$

is linearly independent. Then, it is easy to show, that if the partition is based only on one dimension of the data - e.g., $\gamma_l = \{x \in \mathcal{X} : x_1 \in [a_{l-1}, a_l]\}$ and $\cup_{l=1}^L [a_{l-1}, a_l] = \mathbb{R}$, for $l = 1, \ldots, L$ - the matrix $\mu^*_{\gamma,0}$ is rank deficient. This is because the partition cells are uninformative about the values the other covariates can take and, by independence and identical distribution, it follows that $\mathbb{E}\left[X_j \mathbb{I}_{\gamma_l}(X_1)\right] = \mathbb{E}\left[X_k \mathbb{I}_{\gamma_l}(X_1)\right]$ for any $j, k \neq 1$ and all l. Even when the partitioning utilizes all covariates, $\mu^*_{\gamma,0}$ might be rank deficient, especially when cells are based on one-dimensional projections of the data. For instance, in the previous scenario, consider $\gamma_l = \{x \in \mathcal{X} : c'x \in [a_{l-1}, a_l]\}$ for $l = 1, \ldots, L$, and $c = (c_1, \cdots, c_{d_x})$. Then, it holds that, $r(\mu^*_{\gamma,0}) < d_\theta$ whenever $c_l = c_k$ for some $k \neq l$, or $c_l = 0$ for some l.

In the Monte Carlo simulations of Section 7 and the empirical illustration of Section 8, the data is partitioned using the following three covariate-based methods: SEB on the first principal component of $\mathbb{X}$ (**PCS**), SEB on the vector of fitted values (**FS**), and k-means clustering on $\mathbb{X}$ (**KM**). The former methods use a one-dimensional projections of the data and might be problematic for the implementation of the $\hat{\chi}^2$ test. The SEB on fitted values has the additional advantage of exploiting the dependence between the residuals and the fitted values under the alternative to generate large aggregate squared residuals. The k-means algorithm, on the other hand, partitions by maximizing the between-cluster distance, and should be less affected by the rank deficiency of $\mu^*_{\gamma,0}$.

4.2 Neyman-Pearson (NP) Classes

Frequently, the researcher's primary concern is rejecting a subset of deviations critical for the application at hand. For instance, when estimating Mincer's earning regression, it is common to compare it with alternative specifications of the log-income profile (e.g. Polachek et al. 2008). Similarly, in the Cox (1972) model, a parsimonious parametric specification of the baseline hazard is often compared to more flexible models (e.g. Seetharaman and Chintagunta 2003).

In this case, a procedure that splits the data where the deviations from the null toward the alternative hypothesis are the largest is preferable. To this end, consider the construction of Neyman-Pearson (NP) classes built over the intersection points between the null and the alternative specification. These classes, in the probability distribution model checking literature, increase (or even maximize) Pearson's measure of discrepancy between the null and the alternative hypothesis (see Remark 3.3 in Balakrishnan, Voinov, and Nikulin 2013).

Let H_1 be a given alternative parametric specification of the regression function,

$$H_1 : m \in \mathcal{M}_1$$

with $\mathcal{M}_1 = \{m_{1,\theta^*}(\cdot) : \theta^* \in \Theta^*\}$ being a family of parametric regression functions and $\Theta^* \subset \mathbb{R}^{d_{\theta^*}}$ a suitable parameter space. Thus, under H_1, $m_{1,\theta^*_0}(X) = m(X)$ a.s. for some $\theta^*_0 \in \Theta^*$. Neyman-Pearson classes partition the data splitting over the points where the

models under H_0 and under H_1 meet, i.e. where $m_{1,\tilde{\theta}^*}(\cdot) = m_{\tilde{\theta}}(\cdot)$ a.s., with $\tilde{\theta}^* = \theta_0^* + o_p(1)$ under H_1 and $\tilde{\theta}$ the corresponding consistent estimator of θ_0 under H_0. By doing so, the difference between the average prediction under the null and the alternative model in each class is the largest possible, making it easier to detect deviations from the null toward H_1.

If X is a univariate variable, finding the intersection set, $\{x \in \mathcal{X} : m_{1,\tilde{\theta}^*}(x) - m_{\tilde{\theta}}(x) = 0\}$, and dividing the data accordingly is a straightforward process. However, when X is multivariate, the intersection set includes surfaces, and splitting $\mathcal{X}$ across different dimensions becomes significantly more complex. In line with the original proposal of Balakrishnan, Voinov, and Nikulin (2013), rather than dividing the data at these intersections, $\mathcal{X}$ is partitioned into two classes based on whether the difference between the models is greater or less than zero.

$$\hat{\gamma}_1 = \{x \in \mathcal{X} : m_{1,\tilde{\theta}^*}(x) - m_{\tilde{\theta}}(x) > 0\},$$
$$\hat{\gamma}_2 = \{x \in \mathcal{X} : m_{1,\tilde{\theta}^*}(x) - m_{\tilde{\theta}}(x) \leq 0\},$$

Notice that, if the alternative model is correctly specified, i.e. $\mathbb{E}[Y|X] = m_{1,\theta_0^*}(X)$ a.s. for some $\theta_0^* \in \Theta^*$, then $\hat{\gamma}_1$ is the set of points in $\mathcal{X}$ where the null model underestimates the regression function, and $\hat{\gamma}_2$ is the set of points in $\mathcal{X}$ where the null model overestimates the regression function. In this case, χ^2 tests with NP classes detect deviations of the type,

$$H_1(\gamma): \ \mathbb{E}\left[\left(m_{1,\theta^*}(X) - m_\theta(X)\right)\mathbb{I}_{\gamma_l}(X)\right] \neq 0 \text{ for some } l \in \{1,2\} \text{ and for all } \theta \in \Theta.$$

where $\gamma_1 = \{x \in \mathcal{X} : m_{1,\tilde{\theta}^*}(x) - m_\theta(x) > 0\}$ and $\gamma_2 = \{x \in \mathcal{X} : m_{1,\tilde{\theta}^*}(x) - m_\theta(x) \leq 0\}$. That is, if the alternative model is correct, χ^2 tests with Neyman-Pearson classes are capable of detecting any deviations from it, at least asymptotically.

It is convenient to further divide the data into additional cells to enhance the power of the tests and, ensuring $L > d_\theta$, perform the J test. This is done by employing one of the covariate-based procedures discussed above - **PCS**, **FS**, or **KM** - to split each of the preliminary two-cell partitions into $L/2$ cells.

Algorithm 1 (Parametric Neyman-Pearson (PNP) Algorithm)

(0) Fix a number of cells, L, and a minimum number of observations in each cell, n_{min}.

(1) Compute the fitted values under the null $\{m_{\tilde{\theta}}(X_i)\}_{i=1}^n$.

(2) Compute the fitted values under the alternative $\{m_{1,\tilde{\theta}^*}(X_i)\}_{i=1}^n$.

(3) Split the data into two cells, $\hat{\gamma}_1$ and $\hat{\gamma}_2$, where $\hat{\gamma}_1 = \{x \in \mathcal{X} : m_{1,\tilde{\theta}^*} - m_{\tilde{\theta}}(x) > 0\}$ and $\hat{\gamma}_2 = \{x \in \mathcal{X} : m_{1,\tilde{\theta}^*} - m_{\tilde{\theta}}(x) \leq 0\}$.

*(4) If $L > 2$, split the largest and smallest cells into $\lceil L/2 \rceil$ and $\lfloor L/2 \rfloor$ cells, respectively, by applying one of the following methods: **PCS**, **FS**, or **KM**.*

This is not the only solution to determine additional splits. The key point for the NP procedure is that, in each cell, the differences between the two models have the same sign. Once this is granted, the two-cell NP partition can be split with any other procedure.

Example 2 (PNP Classes)

Consider the same scenario as in Example 1, but now under a composite hypothesis, where the model parameters are unknown both under the null hypothesis and under the alternative. After obtaining the OLS estimates for both the null and alternative models, the Neyman-Pearson critical regions are defined based on the points $x \in \mathcal{X}$ where the equation $(\tilde{\theta}_{00} - \tilde{\theta}^*_{00}) + (\tilde{\theta}_{01} - \tilde{\theta}^*_{01})x - \tilde{\theta}^*_{02} \sin(50x/2\pi) = 0$ holds. The resulting partition is depicted in Figure 2.

4.3 Flexible Neyman-Pearson (FNP) Classes

When considering the unrestricted alternative, Neyman-Pearson classes correspond to subsets of points where the model under the null hypothesis consistently overestimates and underestimates the (unknown) true regression model. In other words, these classes help us distinguish situations where $m(x)$ is bigger or smaller than $m_{\theta_0}(x)$, which characterize the alternative scenario.

Additionally, differences between the parametric model and the regression function become evident when we examine how the parametric residual relates to various function of the covariates vector, such as the model fitted values. This connection is harnessed with the adoption of a structured alternative "index" model of the form $\mathbb{E}[\varepsilon_{\theta_0}(Z)|X] =$

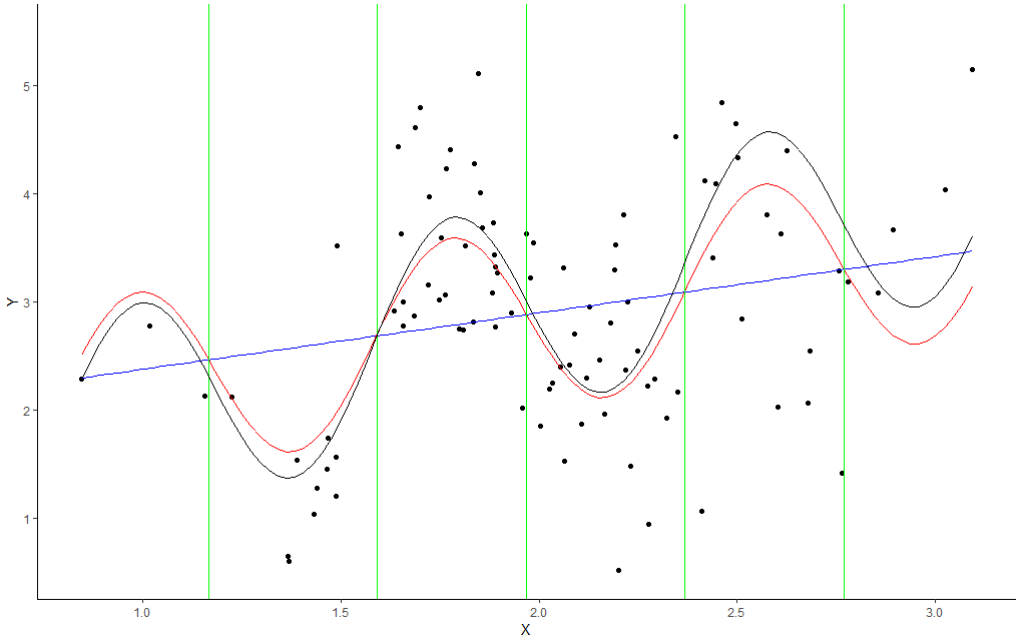


Figure 2: Neyman-Pearson classes with parametric estimate of the regression function. The data is generated under H_1 (black line) with $\theta^* = (1, 1, 1)$. The green line depicts the points where the estimates under the null (blue line) and the estimates under the alternative (red line) meet.

$f(m_{\theta_0}(X))$ a.s., for some $\theta_0 \in \Theta$, where $f(\cdot)$ represents an unknown smooth function (see, e.g., Han, Ma, Ren, and Wang 2023 for a similar approach). Consequently, under the null hypothesis, the regression error is independent in mean from the assumed regression function, i.e. $\mathbb{E}\left[\varepsilon_{\theta_0}(Z)|m_{\theta_0}(X)\right] = 0$. In contrast, under the alternative hypothesis, a systematic dependence in the means of the errors emerges.

This motivates the adoption of an auxiliary regression of $\varepsilon_{\theta_0}(Z)$ on a fixed-order polynomial expansion of the fitted values,

$$\varepsilon_{\tilde{\theta}}(Z_i) = \beta_0 + \sum_{j=1}^{q} \beta_j m_{\tilde{\theta}}(X_i)^j + \epsilon_i,$$

and use it to split where the predicted residuals are greater or smaller than zero. The procedure consists of using the auxiliary model

$$\mathcal{M}_{BP} = \left\{\beta_0 + \beta_1 m_{\tilde{\theta}}(x) + \beta_2 m_{\tilde{\theta}}(x)^2 + \cdots + \beta_q m_{\tilde{\theta}}(x)^q : \beta \in \mathbb{B} \subset \mathbb{R}^q\right\},$$

to split the sample. Denote as $\tilde{\beta} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_q)$ the OLS estimates of β. Under the null

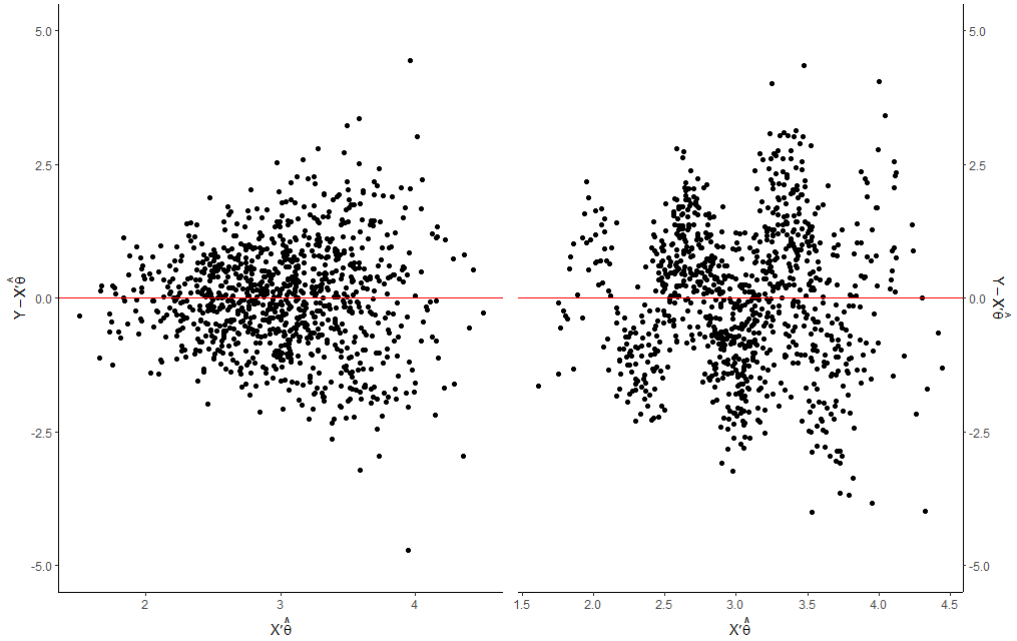


Figure 3: Residual-Fitted values scatter plot under the null (left panel) and under the alternative (right panel). Data generated using the model in (28) under homoskedasticity, $a = 0$, with $c = 0$ (left panel) and $c = 50$ (right panel).

hypothesis, $\tilde{\theta} = \theta_0 + o_p(1)$, and $\tilde{\beta}_0 = \tilde{\beta}_1 = \cdots = \tilde{\beta}_q = o_p(1)$, implying that the fitted values do not have predictive power over the residuals. Under the structured alternative, some of the estimated coefficients converge to a non-zero limit, and the fitted values exhibit predictive power on the residuals. The dependence, then, is used to split cases based on the predicted residuals to generate cells with residuals of the same sign and, thus, larger aggregate residuals. Tests built with these partitions can be seen as a formalization of the statistical practice of looking at the residual-fitted values scatter plot to capture model mispecification and heteroskedasticity (e.g., Cook 1994). Remarkably, when $m_\theta(x)$ is linear, the auxiliary regression consists of the artificial regression used in the Ramsey (1969) RESET test for goodness-of-fit of the linear regression model.

Algorithm 2 (Flexible Neyman-Pearson (FNP) Algorithm)

(0) Fix a number of cells, L, and a minimum number of observations in each cell, n_{min}.

(1) Compute the fitted values under the null $\{m_{\tilde{\theta}}(X_i)\}_{i=1}^n$ and obtain the residuals, $\{\varepsilon_{\tilde{\theta}}(Z_i)\}_{i=1}^n$.

(2) Perform an auxiliary regression of $\varepsilon_{\tilde{\theta}}(Z_i)$ on a fixed-order polynomial expansion of

the fitted values and obtain the vector of predicted residuals, $\{\hat{\varepsilon}_{\tilde{\theta}}(Z_i)\}_{i=1}^n$.

(3) Split the data into two cells, $\hat{\gamma}_1$ *and* $\hat{\gamma}_2$*, where* $\hat{\gamma}_1 = \{x \in \mathcal{X} : \hat{\varepsilon}_{\tilde{\theta}}(x) > 0\}$ *and* $\hat{\gamma}_2 = \{x \in \mathcal{X} : \hat{\varepsilon}_{\tilde{\theta}}(x) \leq 0\}$.

(4) If $L > 2$*, split the largest and smallest cells into* $\lceil L/2 \rceil$ *and* $\lfloor L/2 \rfloor$ *cells, respectively, by applying one of the following methods:* ***PCS****,* ***FS****, or* ***KM****.*

4.4 Convergence of NP Classes

I examine the conditions under which NP (PNP and FNP) partitions converge to fixed cells in $\mathbb{D}$. The validity of Assumption 4 for NP partitions may not be immediately apparent, considering that the splitting points in this case are determined through the comparison of different estimators of the regression function. However, as discussed below, a significant class of NP partitions converges to fixed cells in $\mathbb{D}$.

First, consider the following setting, where the null hypothesis is given by $H_0 : m(X) = \theta_0' X$ a.s., for some $\theta_0 \in \mathbb{R}^{d_x}$, whereas the alternative of interest is $H_1 : m(X) = \theta_0^{*\prime} X + \theta_1^* f(X)$ a.s., for some $(\theta_0^*, \theta_1^*) \in \mathbb{R}^{d_x+1}$, with $f : \mathbb{R}^{d_x} \to \mathbb{R}$ being a known function of the regressors. Parametric NP partitions are constructed around the points where the two models fit are equal; that is,

$$x \in \mathcal{X} : x'(\tilde{\theta}_0 - \tilde{\theta}_0^*) = f(x)\tilde{\theta}_1^*,$$

where $\tilde{\theta}_0$ is the OLS estimator of Y on X, and $(\tilde{\theta}_0^*, \tilde{\theta}_1^*)$ is the OLS estimator of Y on $\tilde{X} = (X, f(X))$. This is equivalent to stating that the set of solutions, x_0, satisfies

$$x_0'(\tilde{\theta}_0 - \tilde{\theta}_0^*) = f(x_0)\tilde{\theta}_1^*.$$

Working out the OLS algebra, we obtain the following representations under H_0,

$$\sqrt{n}(\tilde{\theta}_0 - \tilde{\theta}_0^*) = A_n \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{X}_i e_i,$$
$$\sqrt{n}\tilde{\theta}_1^* = B_n \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{X}_i e_i,$$

where $A_n = d_n^{-1}\begin{bmatrix} A_{1,n} & A_{2,n} \end{bmatrix}$, $B_n = d_n^{-1}\begin{bmatrix} 1 & A_{1,n}' \end{bmatrix}$, and,

$$A_{1,n} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i f(X_i)\right),$$

$$A_{2,n} = -\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) X_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n} X_i f(X_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1},$$
$$d_n = \left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) X_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i f(X_i)\right).$$

It follows that the set of splitting points is given by the solutions of the following equation,

$$(x' A_n - f(x) B_n)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{X}_i e_i = 0,$$

which holds for any $x_0 \in \mathcal{X}$ such that $x_0' A_n = f(x_0) B_n$. Thus, x_0 converge in probability to the fixed solutions of $x_0' A = f(x_0) B$, where A and B are the probability limits of A_n and B_n, respectively.

Thus, even when the partition depends on the comparison of estimated parameters, the splitting points can converge to fixed points in $\mathcal{X}$. The following proposition generalizes this result by incorporating specific conditions on the behavior of the parameter estimators under the null and alternative hypotheses to ensure the convergence of NP classes.

Proposition 1 *Let $\mathcal{M} = \{m_\theta(\cdot) : \theta \in \Theta\}$ and $\mathcal{M}_1 = \{m_{1,\theta^*}(\cdot) : \theta^* \in \Theta^*\}$ represent the models under H_0 and H_1, respectively. Further, denote with $\tilde{\theta}$ and $\tilde{\theta}^*$ the respective $\sqrt{n}$-consistent estimators of θ_0 and θ_0^* under H_0 and H_1, where $m_{\theta_0} = m$ a.s. under H_0, and $m_{1,\theta_0^*} = m$ a.s. under H_1. Then, if under H_0:*

(a) There exists a pseudo-parameter $\bar{\theta}^*$, such that $\tilde{\theta}^* \xrightarrow{p} \bar{\theta}^*$ under H_0.

(b) $m_{\theta_0} = m_{1,\bar{\theta}^*}$ a.s.

(c) $\sqrt{n}(\tilde{\theta} - \theta_0) = C n^{-1/2}\sum_{i=1}^{n} h(X_i, \varepsilon_{\theta_0}) + o_p(1)$, $\sqrt{n}(\tilde{\theta}^* - \theta_1^*) = D n^{-1/2}\sum_{i=1}^{n} h(X_i, \varepsilon_{\theta_0}) + o_p(1)$, where C and D are $c \times d_h$ and $d \times d_h$ constant matrices, and $h(\cdot,\cdot)$ is some $\mathbb{R}^{d_h}$-valued function of regressors and errors such that $n^{-1/2}\sum_{i=1}^{n} h(X_i, \varepsilon_{\theta_0}) = O_p(1)$.

Then, the NP partition splitting points converge in probability to fixed points in $\mathcal{X}$.

The most crucial conditions for the convergence to fixed cells are Conditions (a) and (b), which require that the model under the alternative encompasses the model under the null hypothesis. These conditions are typically satisfied in the context of both PNP partitions and FNP partitions. In the case of FNP partitions this is particularly easy to check, as the alternative regression model corresponds to a polynomial expansion of the fitted values, thus nesting the model under the null.

In the appendix, I provide Monte Carlo evidence demonstrating NP partitions converging to fixed cells in $\mathbb{D}$.

5 Local Power

To analyze the local power of the χ^2 tests under the alternative hypothesis, let's consider the following sequence of local alternatives,

$$H_{1,n} : m(x) = m_{\theta_0}(x) + \frac{1}{\sqrt{n}} h(x) \text{ a.s.}, \tag{18}$$

where $h(X)$ is a random variable representing departures from the null hypothesis with $\mathbb{E}\left[h(X)^2\right] < \infty$ and $0 \leq P(h(X) = 0) < 1$.

Denote as $\delta_{\boldsymbol{\gamma}} = (\delta_1, \ldots, \delta_L)$, with $\delta_l = \mathbb{E}\left[h(X)\mathbb{I}_{\gamma l}(X)\right]$ for $l = 1, \ldots, L$. Under $H_{1,n}$ and θ_0 known, the test statistics converge to a noncentral chi-squared distribution with L degrees of freedom,

$$\begin{aligned} \hat{\chi}^2_{\hat{\boldsymbol{\gamma}},0} &\xrightarrow{d} \chi^2_L(\lambda), \\ \hat{\mathcal{W}}_{\hat{\boldsymbol{\gamma}}}(\theta_0) &\xrightarrow{d} \chi^2_L(\lambda), \end{aligned} \tag{19}$$

where,

$$\lambda = \delta'_{\boldsymbol{\gamma}} (\Sigma_{\boldsymbol{\gamma},0})^{-1} \delta_{\boldsymbol{\gamma}} \tag{20}$$

is the non-centrality parameter.

Equations (19) and (20) highlight an important trade-off when choosing L. On one

hand, the inequality,

$$\frac{\delta_l^2}{\sigma_{0,l}^2} + \frac{\delta_f^2}{\sigma_{0,f}^2} \geq \frac{(\delta_l + \delta_f)^2}{\sigma_{0,l}^2 + \sigma_{0,f}^2},$$

shows that λ is non-decreasing for nested partitions. That is, as the partition becomes finer with more cells, the test becomes more capable of detecting smaller deviations from the null hypothesis, resulting in increased power. On the other hand, the (global) power of χ^2 tests under the alternative hypothesis $H_{1,n}$ decreases as the number of cells increases. This decline in power is due to the higher variability of the limit distribution in (19) associated with a larger number of cells. As the number of cells grows, the distribution becomes more spread out, leading to a higher chance of observing test statistics falling in less extreme regions and, thus, not rejecting. Similar trade-offs have already been noted in the classical goodness-of-fit testing literature (see, e.g., Kallenberg, Oosterhoff, and Schriever 1985).

When θ_0 is unknown, under $H_{1,n}$, the estimators of θ_0 typically follow the asymptotic expansion in Assumption 3' below (see, e.g., Newey 1985).

Assumption 3' *Under $H_{1,n}$, the estimator $\tilde{\theta}$ satisfies,*

$$\sqrt{n}(\tilde{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_{\theta_0}(Z_i) + \delta + o_p(1)$$

where the function $l_\theta(\cdot)$ is as defined in Assumption 3 and $\delta \in \mathbb{R}^{d_\theta}$ is a non-zero vector.

It's easy to show, for instance, that Assumption 3' is satisfied by the grouped GMM estimator, with

$$\delta = -\left[\boldsymbol{\mu}^*_{\boldsymbol{\gamma},0}(\Sigma_{\boldsymbol{\gamma},0})^{-1}\boldsymbol{\mu}^{*\prime}_{\boldsymbol{\gamma},0}\right]^{-1} \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0}(\Sigma_{\boldsymbol{\gamma},0})^{-1}\delta_{\boldsymbol{\gamma}}.$$

It follows that under $H_{1,n}$ and the same assumption of Theorem 2,

$$\begin{aligned} \hat{\chi}^2_{\hat{\gamma}} &\xrightarrow{d} \chi^2_{L-d_\theta}(\lambda_1), \\ \hat{\mathcal{W}}_{\hat{\gamma}}(\tilde{\theta}) &\xrightarrow{d} \chi^2_{r(\text{Avar}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})))}(\lambda_2), \end{aligned} \tag{21}$$

where

$$\lambda_1 = \delta'_{\boldsymbol{\gamma}} \left[(\Sigma_{\boldsymbol{\gamma},0})^{-1} - (\Sigma_{\boldsymbol{\gamma},0})^{-1} \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} \left(\boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0} (\Sigma_{\boldsymbol{\gamma},0})^{-1} \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} \right)^{-1} \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0} (\Sigma_{\boldsymbol{\gamma},0})^{-1} \right] \delta_{\boldsymbol{\gamma}} \tag{22}$$

$$\lambda_2 = \left(\delta_{\boldsymbol{\gamma}} - \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} \delta \right)' \operatorname{Avar}^- (\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})) \left(\delta_{\boldsymbol{\gamma}} - \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} \delta \right) \tag{23}$$

are the respective non-centrality parameters. Comparing the power of these two tests is not a straightforward task, given the differences in both degrees of freedom and non-centrality parameters in their respective distributions. In fact, scenarios can be constructed where the $\hat{\chi}^2$ test exhibits greater power than the Wald test, and vice versa, reasoning as Moore and Spruill (1975) and Moore (1977), for the classical case.

Instead of the non-centrality parameters, consider now the squared L2 norm of the drift,

$$||\delta_{\boldsymbol{\gamma}}||_2^2 = \delta'_{\boldsymbol{\gamma}} \delta_{\boldsymbol{\gamma}} = \sum_{l=1}^{L} \delta_l^2 = \sum_{l=1}^{L} \mathbb{E} \left[h(X) \mathbb{I}_{\gamma_l}(X) \right]^2 .$$

Then, for any pair of δ_l and δ_f such that $\operatorname{sgn}(\delta_l) = \operatorname{sgn}(\delta_f)$, the inequality $(\delta_l + \delta_f)^2 \geq \delta_l^2 + \delta_f^2$ shows that the optimal partitioning for $||\delta_{\boldsymbol{\gamma}}||_2^2$ consists of two cells: one containing the points where $h(\cdot) = \sqrt{n} \left(m(\cdot) - m_{\theta_0}(\cdot) \right)$ takes only positive values, and the other where it takes only negative values. Essentially, the optimal partitioning of $||\delta_{\boldsymbol{\gamma}}||_2^2$ is achieved with two Neyman-Pearson classes.

Proposition 2 *The euclidean norm of the drifts, $||\delta_{\boldsymbol{\gamma}}||_2$, is maximized by two Neyman-Pearson classes, $\gamma^* = \{\gamma_i^*\}_{i=1}^2$,*

$$\gamma_1^* = \{x \in \mathcal{X} : h(x) \geq 0\} \quad \gamma_2^* = \{x \in \mathcal{X} : h(x) < 0\}$$

This suggests that, in some sense, Neyman-Pearson classes correspond to optimization criteria for the non-centrality parameter. Of course, a rigorous argument to justify efficiency should account for the dependence of the non-centrality parameter denominators on the cell boundaries.

6 General CMR

The analysis is extended to general moment restrictions with the introduction of a response variables vector, Y, taking values in $\mathcal{Y} \subset \mathbb{R}^{d_y}$, $d_y \geq 1$, and a generalized residual vector (Wooldridge 1990), $\boldsymbol{\varepsilon}_\theta : (\mathcal{Y}, \mathcal{X}) \to \mathbb{R}^{d_\varepsilon}$ with $\boldsymbol{\varepsilon}_\theta(\cdot) = (\varepsilon_{1,\theta}(\cdot), ..., \varepsilon_{d_\varepsilon,\theta}(\cdot))'$, defining parametric relationships between Y and X. The null hypothesis is defined as before, i.e. $H_0 : \mathbb{E}\left[\boldsymbol{\varepsilon}_{\theta_0}|X\right] = 0$ a.s. for some $\theta_0 \in \Theta$.

The generality of this framework allows testing for a wide range of econometric models such as simultaneous equation model identified by instrumental variables (Newey 1990) or nonlinear in parameters and endogenous variables models, e.g. Box-Cox transform.

When the dimension of the generalized residual is bigger than one, it might be optimal to consider a partition for each component of $\boldsymbol{\varepsilon}_\theta(\cdot)$. In particular, for each $j \in \{1, ..., d_\varepsilon\}$, let $\mathbb{D}_j$ be a class of partitions of $\mathcal{X}$ comprised of L_j sets from $\mathbb{C}$ (L_j is fixed for all n); that is,

$$\mathbb{D}_j = \left\{ \boldsymbol{\gamma}_j = (\gamma_{j,1}, ..., \gamma_{j,L_j})' \in \mathbb{C}^{L_j} : \cup_{l=1}^{L_j} \gamma_{j,l} = \mathcal{X},\ \gamma_{j,l} \cap \gamma_{j,f} = \emptyset,\ \forall l \neq f \right\}. \tag{24}$$

The partition corresponding to the j-th component of the generalized residual is an element of $\mathbb{D}_j$, $\boldsymbol{\gamma}_j \in \mathbb{D}_j$, or a random element $\hat{\boldsymbol{\gamma}}_j \in \mathbb{D}_j$, with probability limit $\boldsymbol{\gamma}_j \in \mathbb{D}_j$.

Denote as $\mathcal{E}_\theta(\cdot)$ the $\bar{L} \times \bar{L}$ block diagonal matrix of generalized residuals with main diagonal elements given by $\{\varepsilon_{j,\theta}(\cdot) I_{L_j}\}_{j=1}^{d_\varepsilon}$, where $\bar{L} = \sum_j L_j$. If $L_1 = L_2 = \cdots = L_{d_\varepsilon} = L$, then $\mathcal{E}_\theta(\cdot) = \text{diag}[\boldsymbol{\varepsilon}_\theta(\cdot)] \otimes I_L$, where $\text{diag}[\boldsymbol{\varepsilon}_\theta(\cdot)] = \text{diag}\{\varepsilon_{1,\theta}(\cdot), ..., \varepsilon_{d_\varepsilon,\theta}(\cdot)\}$ is the $d_\varepsilon \times d_\varepsilon$ diagonal matrix with the components of $\boldsymbol{\varepsilon}_\theta(\cdot)$ on the main diagonal and $\otimes$ denotes the Kronecker product.

The χ^2 test statistics can be expressed as quadratic forms of

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{E}_\theta(Z_i) \mathbf{I}_{\boldsymbol{\gamma}}(X_i), \tag{25}$$

where $\mathbf{I}_{\boldsymbol{\gamma}}(\cdot) = (\mathbf{I}'_{\boldsymbol{\gamma}_1}, ..., \mathbf{I}'_{\boldsymbol{\gamma}_{d_\varepsilon}})'$ is the vector of indicator functions over all the partitions. While the covariance matrices of $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\theta_0)$ and $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})$ under the null are given by,

$$\Sigma_{\boldsymbol{\gamma},0} = \mathbb{E}\left[\mathcal{E}_{\theta_0}(Z) \mathbf{I}_{\boldsymbol{\gamma}}(X) \mathbf{I}_{\boldsymbol{\gamma}}(X)' \mathcal{E}_{\theta_0}(Z)'\right], \tag{26}$$

and,

$$\text{Avar}\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) = \Sigma_{\boldsymbol{\gamma},0} - \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} C'_{\boldsymbol{\gamma},0} - C_{\boldsymbol{\gamma},0} \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0} + \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} L_0 \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0}, \tag{27}$$

where $\boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} = \mathbb{E}\left[\nabla \mathcal{E}_{\theta_0}(Z) \mathbf{I}_{\boldsymbol{\gamma}}(X)\right]$ is the Jacobian matrix, $C'_{\boldsymbol{\gamma},0} = \mathbb{E}\left[\mathcal{E}_{\theta_0}(Z) \mathbf{I}_{\boldsymbol{\gamma}}(X) l_{\theta_0}(Z)'\right]$, and L_0 is defined as before.

In general, $\Sigma_{\boldsymbol{\gamma},0}$ is not diagonal, unless $\boldsymbol{\varepsilon}_\theta(\cdot)$ consists of orthogonal components or $d_\varepsilon = 1$. However, when $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \ldots = \boldsymbol{\gamma}_{d_\varepsilon}$, $\Sigma_{\boldsymbol{\gamma},0}$ is block diagonal with L blocks of size $d_\varepsilon \times d_\varepsilon$ corresponding to the covariance matrix of $\boldsymbol{\varepsilon}_\theta(\cdot)$ in each cell of the partition.

7 Monte Carlo Study

This section presents a Monte Carlo study to evaluate the finite sample performance of χ^2 tests for testing the null hypothesis of linearity within a regression model. To establish a benchmark for comparison with existing tests for H_0, the χ^2 tests are compared with those proposed by Stute (1997) and Stute and Zhu (2002) for regression specifications based on marked residuals processes.

The data generating process (DGP) is given by,

$$Y_i = \sum_{j=1}^{d_x} X_{j,i} + b \sin\left(\frac{c \sum_{j=1}^{d_x} X_{j,i}}{2\pi}\right) + \sigma(X_i)\epsilon_i \tag{28}$$

where $X_i = (X_{1,i}, ..., X_{d_x,i})$ is a vector of mutually independent covariates distributed uniformly over [0,1], $X_{i,j} \sim U[0,1]$, the error distributes normally and independently from X, $\epsilon | X \sim N(0,1)$,

$$\sigma^2(X) = \frac{g(X)}{\mathbb{E}[g(X,a)]},$$

and $g(X,a) = e^{aX_1}$, with $\mathbb{E}[\sigma^2(X)] = 1$. The model under the null corresponds to $b = 0$, while for the alternative models $b = 0.5$. The parameter a controls the heteroskedasticity severity, with $a = 0$ corresponding to homoskedasticity, while c governs the extent of departures from linearity. In the first part of the simulations, I provide evidence for the size of the test for $a \in \{0, 3\}$, I then fix $a = 3$ and proceeds to examine the tests power for $c \in \{10, 50\}$. Notably, with $c = 50$, the deviations manifest at higher frequencies making them more challenging to discern from sampling error. The generated samples

have size $n \in \{100, 200, 500, 1000\}$ with a dimension of the covariate vector $d_x \in \{5, 10\}$. The number of Monte Carlo repetition, R, is set at $R = 3000$ for all sample sizes. The rejection rates are reported at the 5% nominal level, results at 1% and 10% are similar.

The $\hat{\chi}^2$ test is built as described in Section 2 using the GMM estimator,

$$\hat{\theta}_{\boldsymbol{\gamma}} = \left[\sum_{l=1}^{L} \frac{\bar{X}_l \bar{X}_l'}{\hat{\sigma}_l^2(\tilde{\theta})}\right]^{-1} \sum_{l=1}^{L} \frac{\bar{X}_l \bar{Y}_l}{\hat{\sigma}_l^2(\tilde{\theta})}.$$

where $\tilde{\theta}$ is the OLS estimator. Given the data-generating process, one can easily show that the influence function of $\tilde{\theta}$, $l_{\theta_0}(z) = \mathbb{E}\left[XX'\right]^{-1} x\varepsilon_{\theta_0}(z)$, and $\varepsilon_{\theta_0}(z)\mathbf{I}_{\boldsymbol{\gamma}}(x)$ have linearly independent components (however, the independence may not hold when X includes a constant term). Thus, Avar $\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$, is full rank and the Wald test can be implemented using the plug-in estimator $\hat{W}_{\boldsymbol{\gamma}}(\tilde{\theta})$,

$$\begin{aligned}\hat{W}_{\boldsymbol{\gamma}}(\tilde{\theta}) = {} & \mathbb{E}_n\left[\varepsilon_{\tilde{\theta}}(Z_i)^2 \mathbf{I}_{\boldsymbol{\gamma}}(X_i)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\right] - \mathbb{E}_n\left[\mathbf{I}_{\boldsymbol{\gamma}}(X_i)X_i'\right]\mathbb{E}_n\left[X_iX_i'\right]^{-1}\mathbb{E}_n\left[\varepsilon_{\tilde{\theta}}(Z_i)^2 X_i\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\right] - \\ & - \mathbb{E}_n\left[\varepsilon_{\tilde{\theta}}(Z_i)^2 \mathbf{I}_{\boldsymbol{\gamma}}(X_i)X_i'\right]\mathbb{E}_n\left[X_iX_i'\right]^{-1}\mathbb{E}_n\left[X_i\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\right] + \\ & + \mathbb{E}_n\left[\mathbf{I}_{\boldsymbol{\gamma}}(X_i)X_i'\right]\mathbb{E}_n\left[X_iX_i'\right]^{-1}\mathbb{E}_n\left[\varepsilon_{\tilde{\theta}}(Z_i)^2 X_iX_i'\right]\mathbb{E}_n\left[X_iX_i'\right]^{-1}\mathbb{E}_n\left[X_i\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'\right].\end{aligned}$$

I assess the test's performance using the model-based partitioning methods described in Section 4. For both Parametric NP (**PNP**) classes and Flexible NP (**FNP**) classes, the division into L cells of the initial two-cell partition is done using SEB on the first principal component (**PCS**), SEB on the vector of fitted values (**FS**), and K-means clustering (**KM**). When $d_x = 5$, the partitioning is done using $L \in \{4, 6, 8\}$ cells for the Wald test and $L^* \in \{9, 11, 13\}$ for the J test, whereas when $d_x = 10$, partitions with $L \in \{8, 10, 12\}$ and $L^* \in \{18, 20, 22\}$ cells are examined. Notice that L^* is selected to ensure that the limit null distribution of the two χ^2 tests is the same, enabling a fair comparison between them.

The parametric **PNP** cells are constructed using the correctly specified alternative model (with known parameter c). These tests exploit additional information about the alternative specifications of the regression function. Including them allows for a comprehensive comparison with the other feasible partitioning methods. The **FNP** partition is created using predictions of a regression of parametric residuals on a polynomial expansion

of the fitted values with order $q = 3$.

The performance of the χ^2 tests are compared with two minimum-distance tests for regression specifications based on marked residuals processes indexed by real vectors and real numbers; cf. Stute (1997) and Stute and Zhu (2002), respectively.

$$R_{1,n}(x_1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Y_i - X_i'\tilde{\theta} \right) \mathbb{I}\{X_i \leq x_1\}, \; x_1 \in \mathbb{R}^{d_x}$$
$$R_{2,n}(x_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Y_i - X_i'\tilde{\theta} \right) \mathbb{I}\{X_i'\tilde{\theta} \leq x_2\}, \; x_1 \in \mathbb{R}.$$

Incidentally, the marked residual process indexed by partitions generated with the **FS** method,

$$\hat{\mathbf{\Phi}}_{\hat{\gamma}_l}(\tilde{\theta}) = n^{-1/2} \sum_{i=1}^{n} \varepsilon_{\tilde{\theta}}(Z_i) \mathbb{I}\{X_i'\tilde{\theta} \in C_l\},$$

where $\hat{\gamma}_l = \{x \in \mathcal{X} : x'\tilde{\theta} \in C_l\}$ and $\bigcup_{l=1}^{L} C_l = \mathbb{R}$, is a finite cells version of the process used in the Stute and Zhu (2002) test. The test statistics consist of a functional of $R_{1,n}(\cdot)$ and $R_{2,n}(\cdot)$. In these simulations, we only consider the Kolmogorov-Smirnov functional,

$$KS1 = \sup_{x_1 \in \mathbb{R}^{d_x}} |R_{1,n}(x_1)|,$$
$$KS2 = \sup_{x_2 \in \mathbb{R}} |R_{2,n}(x_2)|.$$

Since these tests have non-pivotal limiting distribution, the critical values are estimated with Wild bootstrap (see Stute, Manteiga, and Quindimil 1998) using $B = 999$ bootstrap repetitions for each replication.

The simulation results for the linear model, as presented in Tables 1 (size) and 2 (power), reveal several key findings. In Table 1, both $\hat{\mathcal{W}}$ and $\hat{\chi}^2$ tests generally exhibit excellent size accuracy. Some relatively small size distortions of the J test occur when the sample size is small and the number of cells large; however, they get better as the sample size increases. Notably, the $KS1$ test experiences pronounced size distortions when $d_x = 10$, while the $KS2$ test demonstrates greater robustness under higher covariate dimensions, as expected.

Turning to Table 2, our focus shifts to assessing the tests' ability to detect deviations

Table 1: Linear Model: Size ($b = 0$)

					PCS				**FS**				**KM**					
					FNP		**PNP**		**FNP**		**PNP**		**FNP**		**PNP**			
n	d_x	a	L	L^*	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	KS1	KS2
100	5	0	4	9	6.70	7.13	4.66	4.73	6.30	5.26	5.13	5.20	7.06	6.43	4.56	4.43	5.14	6.40
			6	11	6.10	6.10	4.23	4.26	5.00	4.90	4.46	4.70	6.30	4.39	3.90	4.06		
			8	13	5.40	4.53	4.26	4.03	4.53	4.30	3.63	3.86	4.50	4.13	4.36	3.73		
100	5	3	4	9	5.80	6.60	4.30	4.10	5.33	4.76	4.36	4.73	5.80	6.13	4.53	5.03	4.86	7.12
			6	11	4.80	4.43	4.03	3.90	4.56	3.80	4.39	4.33	5.13	5.00	4.46	3.46		
			8	13	4.53	5.06	3.26	3.83	4.16	3.63	3.56	3.23	4.46	4.13	3.70	3.50		
100	10	0	8	18	6.23	5.53	5.53	5.06	6.23	5.06	5.33	6.03	7.03	5.46	5.33	4.70	1.46	8.14
			10	20	5.30	5.73	5.53	4.83	5.33	4.56	5.03	4.16	6.13	4.53	5.33	3.96		
			12	22	5.23	5.06	4.03	3.93	4.50	4.33	4.39	4.66	5.09	3.70	4.03	3.36		
100	10	3	8	18	5.86	5.26	3.86	4.03	5.26	4.63	5.13	4.33	5.50	4.53	4.16	3.90	0.78	9.16
			10	20	4.46	3.70	3.50	3.80	3.73	4.20	3.59	4.03	4.56	3.36	4.10	3.20		
			12	22	3.73	3.86	3.30	3.59	3.26	3.10	3.20	2.29	4.16	3.80	3.23	2.90		
200	5	0	4	9	6.90	7.13	5.13	4.70	6.90	5.13	5.46	4.73	7.00	6.66	5.23	4.50	5.56	6.75
			6	11	6.03	6.80	4.20	5.03	5.26	4.73	5.46	4.53	5.70	5.33	4.30	4.80		
			8	13	5.00	6.13	4.73	4.50	5.00	4.66	4.56	4.70	5.00	5.63	4.63	4.53		
200	5	3	4	9	6.53	6.80	4.93	4.46	6.60	5.13	4.96	4.39	6.63	6.36	5.06	4.23	5.56	6.56
			6	11	5.20	6.03	4.70	3.90	5.13	5.09	4.63	4.59	6.20	6.03	4.20	4.43		
			8	13	4.70	4.96	4.39	3.86	3.93	4.13	4.16	3.73	4.76	4.39	3.86	3.96		
200	10	0	8	18	6.53	6.36	5.33	4.96	6.56	5.53	5.40	4.90	6.73	6.33	5.60	4.86	2.80	6.60
			10	20	6.53	6.30	4.50	4.96	5.00	4.93	4.63	4.50	5.86	5.13	4.93	3.90		
			12	22	5.20	5.93	5.23	4.73	4.76	4.23	3.96	4.43	5.33	5.23	4.86	4.46		
200	10	3	8	18	5.83	5.93	4.50	4.43	5.16	5.09	4.36	4.26	5.76	6.00	5.03	3.76	1.40	6.04
			10	20	5.20	4.63	4.66	4.13	4.39	4.73	4.30	4.70	5.23	5.03	4.66	4.33		
			12	22	5.09	4.90	4.50	4.06	4.26	4.43	4.00	4.50	5.60	4.13	4.26	4.63		
500	5	0	4	9	7.36	7.83	5.73	4.86	6.56	5.93	5.33	5.00	7.13	7.43	5.30	4.96	6.40	7.28
			6	11	6.46	7.30	5.00	4.86	5.96	5.46	5.13	4.23	6.53	6.00	5.16	5.26		
			8	13	5.76	5.63	5.09	5.36	4.46	5.46	4.50	5.53	5.76	6.26	4.30	4.20		
500	5	3	4	9	6.16	7.10	4.86	5.33	5.73	5.09	4.93	5.46	7.03	6.46	4.36	4.76	5.76	5.84
			6	11	5.66	5.96	4.83	5.46	6.36	4.70	5.46	5.26	5.70	5.93	4.33	4.90		
			8	13	5.23	5.86	4.36	4.20	4.83	4.76	4.23	4.33	5.86	5.53	4.39	4.20		
500	10	0	8	18	6.80	5.70	5.50	4.80	4.66	4.59	5.60	5.26	6.96	6.03	5.63	4.66	4.32	6.32
			10	20	5.13	6.63	4.50	5.40	5.60	4.93	4.90	5.56	6.20	5.86	4.46	4.70		
			12	22	6.50	6.16	5.00	4.63	5.73	5.33	5.00	5.20	6.00	5.26	4.70	4.86		
500	10	3	8	18	5.73	5.36	4.26	5.56	4.63	5.16	4.36	5.06	5.53	5.23	5.16	5.00	3.20	5.60
			10	20	5.09	6.30	4.96	4.46	5.16	4.66	4.96	4.56	5.33	4.73	4.96	4.56		
			12	22	5.53	4.96	4.26	5.13	5.20	5.13	5.36	4.80	5.06	4.80	5.26	3.59		
1000	5	0	4	9	7.10	7.46	4.23	5.06	6.56	5.56	5.53	5.53	6.36	7.43	4.53	5.53	3.40	5.92
			6	11	6.70	5.90	4.80	5.13	5.09	5.40	5.00	5.70	5.46	5.00	4.33	4.86		
			8	13	5.83	6.33	4.80	4.76	5.70	5.03	5.40	4.36	5.76	5.56	4.56	5.23		
1000	5	3	4	9	6.63	7.50	4.86	4.93	6.46	5.76	4.93	4.76	6.73	6.53	4.46	5.00	3.55	7.11
			6	11	5.43	6.43	4.36	5.06	5.40	5.33	5.40	5.30	6.13	5.63	3.93	5.09		
			8	13	5.13	6.80	4.80	4.30	5.40	5.40	5.06	5.46	4.90	6.06	4.23	4.43		
1000	10	0	8	18	5.90	6.60	5.23	4.90	5.80	6.10	5.06	5.43	5.80	6.43	5.86	4.56	6.37	4.88
			10	20	5.03	5.63	5.03	4.73	5.50	4.59	4.80	5.33	4.96	6.03	4.70	5.30		
			12	22	5.30	6.13	5.03	5.03	4.26	5.06	4.93	5.00	5.00	5.80	5.36	4.86		
1000	10	3	8	18	5.86	5.66	4.56	5.06	5.50	4.53	4.66	5.26	5.93	6.16	5.43	5.46	6.22	5.62
			10	20	5.53	6.03	5.30	5.16	5.16	5.40	4.90	4.13	5.00	6.13	4.56	4.56		
			12	22	6.03	5.09	4.70	4.80	5.09	5.23	5.30	5.30	6.10	5.36	4.83	4.86		

Percentage of rejections at the nominal level $\alpha = 0.05$ of the $\hat{\chi}^2$ and the Wald test under different partitioning methods. n, d_x, a, L, and L^* denote the sample size, the number of covariates, the degree of heteroskedasticity, and the number of cells of the Wald and J test, respectively. The last two columns report the rejection rates of the Kolmogorov-Smirnov tests.

Table 2: Linear Model: Power ($b \neq 0$)

					PCS				**FS**				**KM**					
					FNP		**PNP**		**FNP**		**PNP**		**FNP**		**PNP**			
n	d_x	a	L	L^*	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	KS1	KS2
100	5	10	4	9	46.6	22.4	43.6	29.3	46.8	19.7	46.0	33.8	46.1	39.7	41.9	39.6	31.5	35.8
			6	11	37.8	19.3	37.5	25.4	37.7	18.8	38.6	29.7	38.6	33.6	34.5	33.9		
			8	13	32.5	18.3	29.7	20.0	33.1	16.6	33.1	24.5	32.5	27.2	31.4	27.5		
100	5	50	4	9	7.4	7.9	74.0	56.9	8.5	7.8	72.2	52.3	7.9	7.4	74.2	73.2	5.2	8.0
			6	11	5.6	6.4	65.5	51.9	7.4	7.4	65.1	50.8	5.7	6.0	69.0	68.3		
			8	13	4.8	4.2	59.0	47.2	6.9	6.9	56.1	46.2	5.0	5.3	61.6	61.3		
100	10	10	8	18	28.4	14.1	41.8	28.6	30.2	15.4	40.8	27.9	28.4	22.6	41.9	37.3	1.6	28.2
			10	20	22.4	12.9	36.3	25.6	22.9	15.0	34.8	25.6	23.7	18.6	36.1	33.1		
			12	22	19.6	11.2	31.0	21.3	20.4	12.7	29.0	21.7	19.9	16.1	31.1	26.8		
100	10	50	8	18	5.7	5.5	53.0	38.7	5.9	4.6	53.7	36.9	5.5	5.5	57.1	53.2	0.7	8.6
			10	20	5.5	4.9	49.0	35.7	4.6	4.9	46.4	33.6	5.7	4.1	50.7	47.4		
			12	22	4.6	4.2	41.7	30.5	4.1	3.6	40.6	30.7	4.6	3.6	45.3	41.4		
200	5	10	4	9	80.7	43.2	78.5	57.7	81.8	36.4	80.5	64.5	77.7	73.3	75.2	74.3	59.8	65.1
			6	11	76.3	44.3	76.1	54.5	79.4	38.6	79.5	61.4	76.1	69.5	72.8	71.4		
			8	13	70.2	41.8	71.3	51.3	72.9	41.4	74.0	60.3	70.1	65.5	67.7	66.6		
200	5	50	4	9	10.0	9.0	97.3	87.0	17.2	19.7	97.2	84.0	9.3	9.7	97.3	97.3	5.0	10.2
			6	11	8.2	8.9	95.9	86.2	19.3	21.5	95.1	87.2	9.2	8.6	96.5	96.7		
			8	13	8.7	8.1	94.1	85.4	22.4	22.9	93.1	87.1	8.0	7.6	94.8	96.5		
200	10	10	8	18	69.4	40.1	81.5	65.2	74.0	43.7	81.5	64.9	70.3	65.5	80.4	79.2	4.0	50.4
			10	20	64.1	38.6	79.1	61.1	69.8	41.8	78.2	62.3	63.5	60.1	78.0	75.3		
			12	22	61.3	36.6	76.0	59.9	67.9	39.6	75.6	61.0	62.3	57.0	75.3	73.5		
200	10	50	8	18	5.6	5.4	92.1	73.6	7.6	7.3	91.8	72.9	5.2	5.7	93.1	92.4	1.4	7.4
			10	20	5.0	5.6	88.9	72.8	7.8	7.0	87.9	73.6	6.1	5.3	90.6	90.1		
			12	22	4.8	4.6	87.1	72.0	8.4	6.6	86.2	71.7	5.3	4.8	89.4	87.3		
500	5	10	4	9	99.7	78.9	99.7	94.1	99.8	69.4	99.7	94.6	99.5	98.8	99.4	99.3	94.7	98.2
			6	11	99.7	82.2	99.8	94.9	99.9	77.5	99.8	96.3	99.4	98.9	99.3	99.1		
			8	13	99.1	83.8	99.4	95.6	99.5	81.6	99.5	97.5	99.2	99.0	99.2	99.2		
500	5	50	4	9	24.3	17.5	100	99.0	52.7	63.5	100	98.4	22.0	18.6	100	100	5.4	19.6
			6	11	22.8	19.9	100	99.6	66.4	72.4	100	99.6	20.7	17.0	100	100		
			8	13	26.4	19.4	100	99.7	82.5	77.5	100	99.8	18.8	14.6	100	100		
500	10	10	8	18	99.8	84.8	99.9	97.6	99.9	87.5	99.9	97.3	99.6	99.6	100	99.9	17.5	94.4
			10	20	99.7	87.1	99.9	97.9	99.8	89.1	99.9	97.9	99.6	99.5	99.9	99.8		
			12	22	99.5	88.6	99.8	97.6	99.9	90.8	99.9	98.1	99.4	99.6	99.8	99.8		
500	10	50	8	18	8.1	6.8	100	98.9	36.0	35.1	99.9	98.7	7.9	6.7	100	99.9	4.4	10.0
			10	20	8.4	7.2	100	99.1	44.3	39.1	99.9	99.1	8.9	6.4	100	100		
			12	22	7.9	6.8	100	99.2	49.2	38.9	99.9	99.3	8.2	6.7	100	100		
1000	5	10	4	9	100	94.8	100	99.4	100	87.5	100	99.2	100	100	100	100	99.9	100
			6	11	100	97.8	100	99.8	100	94.1	100	99.9	100	99.9	100	100		
			8	13	100	97.9	100	99.9	100	96.4	100	100	100	100	100	100		
1000	5	50	4	9	46.5	37.0	100	99.9	81.1	91.0	100	99.5	40.9	33.4	100	100	4.6	45.0
			6	11	47.4	42.0	100	99.9	92.5	96.6	100	99.9	40.7	33.0	100	100		
			8	13	53.9	42.4	100	99.9	99.7	98.4	100	100	40.1	29.1	100	100		
1000	10	10	8	18	100	98.2	100	99.9	100	98.4	100	99.9	100	100	100	100	37.9	100
			10	20	100	99.3	100	100	100	99.3	100	100	100	100	100	100		
			12	22	100	99.5	100	99.9	100	99.4	100	100	100	100	100	100		
1000	10	50	8	18	15.2	11.3	100	99.9	83.2	84.2	100	99.9	17.2	11.3	100	100		
			10	20	17.6	12.5	100	99.9	91.2	87.4	100	99.9	17.3	11.5	100	100		
			12	22	16.5	11.7	100	100	96.2	90.3	100	100	16.0	11.1	100	100	4.3	20.0

Percentage of rejections at the nominal level $\alpha = 0.05$ of the $\hat{\chi}^2$ and the Wald test under different partitioning methods. n, d_x, a, L, and L^* denote the sample size, the number of covariates, the degree of heteroskedasticity, and the number of cells of the Wald and J test, respectively. The last two columns report the rejection rates of the Kolmogorov-Smirnov tests.

from linearity when the null hypothesis is false ($b = 0.5$). The $\hat{\mathcal{W}}$ test generally exhibits higher power compared to the $\hat{\chi}^2$ test, especially in cases with larger values of c. However, this difference varies depending on the choice of the partitioning method. For example, under **PNP** partitioning with k-means, the two tests perform the same. The χ^2 tests outperform the omnibus tests when the split into L cells of the initial two-cell partition is done with the **FS** method, particularly against high-frequency deviations ($c = 50$). When the additional splits are done using one of the other two methods, the performance varies.

As expected, both χ^2 tests using **PNP** partitions consistently exhibit the highest power across all scenarios, given their ability to leverage out-of-sample information. In this case, the performance of the χ^2 tests is similar across the different covariate-based partitioning methods.

8 Empirical Application

I present an empirical application of the testing procedures, analyzing the impact of attending Historically Black Colleges and Universities (HBCU) versus Traditional White Institutions (TWI) on the labor and educational outcomes of African American students, based on Fryer and Greenstone (2010). As discussed by Price and Viceisza (2023), HBCU play a pivotal role in enhancing the identity, confidence, and self-esteem of students, particularly black individuals. These institutions have higher ratios of black students, making it easier for students to foster a sense of belonging. HBCU also offer innovative programs, such as college preparation summer programs and Black-centered curricula, encouraging students to actively engage in academic experiences.

The question of whether HBCU are more effective than TWI for the academic and economic achievement of black students holds significant policy implications. HBCU have relied heavily on federal funds, and despite the recent commitment by the US administration to invest over 7 billion dollars in HBCU[1], they have consistently faced underfunding compared to TWI. This financial disparity can significantly impact the quality of education provided by HBCU, potentially limiting opportunities for their students. Thus,

[1] https://www.ed.gov/news/press-releases/fact-sheet-biden-harris-administration-highlights-record-championing-historically-black-colleges-and-universities-hbcus

the suggestion of merging HBCU with other institutions or redirecting resources toward existing TWI could lead to a more efficient and cost-effective higher education system.

The analysis is based on the National Longitudinal Survey of the High School Class of 1972 (NLS) used by Fryer and Greenstone (2010), which is publicly available on OpenICPSR[2]. The survey is a nationally representative sample of 23,451 high school seniors in 1972, with follow-up interviews conducted in 1973, 1974, 1976, 1979, and 1986. This dataset provides rich information concerning these students, including details about their demographics, academic performance, and labor market outcomes.

The sample comprises 624 African American students who pursued higher education after high school and were tracked throughout the 1972-1986 period. Among them, 260 attended an HBCU, while 364 attended a TWI. The considered outcomes include the logarithm of hourly wage in 1986, the probability of attaining a bachelor's degree, and the probability of obtaining a graduate degree.

Table 3: Summary Statistics of Black Students (NLS)

	HBCU			**TWI**		
	Mean	St. deviation	NA	Mean	St. deviation	NA
ln(wage)	2.16	0.64	57	2.11	0.51	71
Bachelor's degree	0.67	0.46	0	0.62	0.48	2
Graduate degree	0.14	0.35	0	0.16	0.37	2
SAT	685.55	133.46	143	793.68	183.39	212
ACT	12.86	4.02	201	14.73	5.31	269
Hs GPA	3.20	1.19	27	3.24	1.33	19
Family income	3.45	2.50	59	3.89	2.70	75
Father education	1.71	0.96	4	1.96	1.16	9
Mother education	1.95	1.08	3	2.08	1.11	4
Private high school	0.03	0.17	0	0.05	0.23	0
Female	0.66	0.47	0	0.64	0.47	0
South	0.87	0.33	0	0.48	0.50	0

The table report mean, standard deviation, and number of missing values of the outcomes and characteristics of African American who attended college. The columns HBCU and TWI refer to the type of college attended. In the first three rows, the outcomes are the logarithm of hourly wage in 1986, the probability of receiving a bachelor's degree, and the probability of enrolling in graduate school. The remaining rows report the covariates. The total sample size is $n = 624$, with $n_1 = 260$ students enrolled in HBCU, and $n_2 = 364$ in TWI.

Following Fryer and Greenstone (2010), I examine a set of covariates, encompassing home environment variables and pre-college characteristics. The former includes family

[2]Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-10-12. https://doi.org/10.3886/E113736V1

income, mother's and father's education, and two dummy variables indicating the student's sex and residence in the south. The latter comprises SAT and ACT scores, high school GPA, and attendance at a private school. Unlike Fryer and Greenstone (2010), I neither categorize the covariates nor replace missing values with 0s. Instead, I address missing values through multivariate imputation by chained equations (Van Buuren and Groothuis-Oudshoorn 2011).

The identification of the enrollment effect in an HBCU on outcomes relies on two sets of assumptions. First, I estimate the wage effect under the assumption of conditional mean independence (CMIA) for potential outcomes. Next, I estimate the effect on all outcomes under the unconfoundedness framework (see, e.g., Rosenbaum and Rubin 1983), where potential outcomes are assumed to be conditionally independent of treatment status.

Let $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$, where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes under treatment and control, respectively, and D_i is a binary variable indicating whether the respondent attended an HBCU. In the CMIA framework, the following assumptions allow the identification of the average treatment effect (ATE).

Assumption 6 ***CMIA***

(a) $\mathbb{E}[Y(j)|D, X] = \mathbb{E}[Y(j)|X]$ *for* $j \in \{0, 1\}$.

(b) $E[Y(j)|X] = E[Y(j)] + m_{\theta_0}(X)$ *for some* $\theta_0 \in \Theta$ *and* $j \in \{0, 1\}$.

Assumption 6 (a) states that once we condition on the set of characteristics, the choice of enrolling in an HBCU is, on average, random. While Assumption 6 (b) supposes a parametric specification of the potential outcomes regression and homogeneous treatment effects among the two groups. Under this assumption, the ATE is identified as the parameter δ_1 in the following regression:

$$\mathbb{E}[Y|D, X] = \delta_0 + \delta_1 D + m_{\theta_0}(X).$$

I consider two different specifications of the potential outcomes regression model. The first is a simple linear specification where, in addition to the listed covariates, a dummy variable for family income above the third quartile, denoted as $\mathbb{I}\{faminc > q_3\}$, is added. This additional covariate accommodates for the potential differential effect of a richer

family background on the future wages of students. While the second is a nonlinear index model of unknown order.

- **Specification (A1)**: $m_\theta(X) = \theta_1' X + \theta_2 \mathbb{I}\{faminc > q_3\}$.
- **Specification (B1)**: $m_\theta(X) = (|\theta_1' X + \theta_2 \mathbb{I}\{faminc > q_3\}|)^{\theta_3}$.

In the unconfoundedness framework, Assumption 6 is replaced by the following assumptions.

Assumption 6' $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i | X_i$ *and* $0 < p(x) < 1 \ \ \forall x \in \mathbb{R}^{d_x}$.

Here $p(X) = P(D = 1|X)$ denotes the propensity score. Under Assumption 6', the ATE and the average treatment of treated (ATT) are identified (Rosenbaum 1987) by,

$$\text{ATE} = \mathbb{E}\left[\left(\frac{D}{p(X)} - \frac{1-D}{1-p(X)}\right) Y\right] \quad \text{and} \quad \text{ATT} = \frac{\mathbb{E}\left[\left(D - \frac{p(X)(1-D)}{1-p(X)}\right) Y\right]}{\mathbb{E}\left[D\right]}.$$

This motivates the following inverse probability weighting (IPW) estimators,

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{w_1(Z_i)}{\bar{w}_1(Z_i)} - \frac{w_0(Z_i)}{\bar{w}_0(Z_i)}\right) Y_i \quad \text{and} \quad \widehat{\text{ATT}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{w_1^T(Z_i)}{\bar{w}_1^T(Z_i)} - \frac{w_0^T(Z_i)}{\bar{w}_0^T(Z_i)}\right) Y_i,$$

where $w_1(Z_i) = D_i/\hat{p}(X_i)$, $w_0(Z_i) = (1 - D_i)/(1 - \hat{p}(X_i))$, $w_1^T(Z_i) = D_i$, $w_0^T(Z_i) = (1-D_i)\hat{p}(X_i)/(1-\hat{p}(X_i))$, and $\bar{w}_d$ and $\bar{w}_d^T$ are the sample means of w_d and w_d^T, respectively, for $d = 0, 1$. The normalized weights reduce the variance and instability of the estimates (Hirano, Imbens, and Ridder 2003).

The estimation of the propensity score is done by fitting a probit model, i.e., $p(X) = \Phi(X'\theta_0)$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. I consider two different specifications of the propensity score model,

- **Specification A2**: X includes all the covariates listed in Table 3.
- **Specification B2**: X includes all the covariates listed in Table 3 and an interaction between the variables "south" and "family income".

Specification (B2) is more comprehensive and takes into account the varied effects on family income for Southern students, given the prevalence of HBCUs in the area. This implies reduced mobility costs for attendees.

To ensure the consistency of the estimated causal parameters, it is crucial to verify the correct specification of the regression models. Table 4 below presents the $\hat{\chi}^2$, the Wald test, and omnibus tests for each of the two regression and propensity score specifications. Then, Table 5 and 6 compare the ATE and ATT estimates obtained from these specifications within the two causal frameworks.

Note that, under the null hypothesis for the probit model, the maximum likelihood estimator (MLE) is the most efficient, and Avar $\left(\mathbf{\Phi}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) = \Sigma_{\boldsymbol{\gamma},0} - \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0}\mathcal{I}^{-1}\boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0}$ is positive definite. Here $\boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} = \mathbb{E}\left[\phi(X'\theta_0)\mathbf{I}_{\boldsymbol{\gamma}}(X)X'\right]$, $\phi(x)$ is the density of the standard normal at x, and, $\mathcal{I} = \mathbb{E}\left[\phi(X'\theta_0)^2 XX'/\Phi(X'\theta_0)(1-\Phi(X'\theta_0))\right]$ is the Fisher information matrix.

The **PNP** partitions are built using the opposite specification as reference, while **FNP** partition uses a polynomial expansion of the fitted values of order $q = 3$. Additional splits of the initial two-cell partition are performed using the fitted values (**FS**) for the Wald test and k-means (**KM**) for the J test.

The comparison is done with the Kolmogorov-Smirnov test statistics of Stute and Zhu (2002) and Sant'Anna and Song (2019) (SS) for the regression model and for the propensity score specifications, respectively. Both omnibus tests are implemented using $B = 10,000$ bootstrap replications. Following common practice, the ATE and ATT are estimated from the subsample of observations with estimated propensity score outside of the $[0.05, 0.95]$ interval (Crump, Hotz, Imbens, and Mitnik 2009).

The results reported in Table 4 suggest that specification (A1) is more suitable for modeling the true regression model than the polynomial specification (B1). This is particularly evident from the J tests with **FNP** cells, which frequently reject the correct specification of model (B1) at relatively low significance levels. The test proposed by Stute and Zhu (2002) is unable to detect deviations from the null hypothesis in both cases, although the test statistic in specification (A1) is significantly larger than in (B1).

On the other hand, specification (B2) is not rejected at the 9% significance level, suggesting a better fit for the propensity score model than specification (A2). Even in this case, the omnibus test does not detect any significant mispecification for the two models.

Table 4: Regression and Pscore Specification Tests

(Regression)

			FNP		PNP		
Spec.	L	L^*	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	KS2
(A1)	2	14	0.68	0.87	0.20	0.59	0.97
	4	16	0.89	0.94	0.52	0.27	
	6	18	0.53	0.56	0.39	0.41	
	8	20	0.36	0.61	0.67	0.44	
	10	22	0.36	0.59	0.35	0.31	
(B1)	2	15	0.80	0.03	0.05	0.20	0.25
	4	17	0.96	0.04	0.50	0.21	
	6	19	0.86	0.06	0.26	0.57	
	8	21	0.94	0.17	0.56	0.15	
	10	23	0.95	0.05	0.50	0.16	

(PScore)

			FNP		PNP		
Spec.	L	L^*	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	$\hat{\mathcal{W}}$	$\hat{\chi}^2$	SS
(A2)	2	12	0.50	0.81	0.26	·	0.29
	4	14	0.06	0.78	0.53	·	
	6	16	0.02	0.66	0.26	0.29	
	8	18	0.17	0.85	0.36	0.37	
	10	20	0.47	0.30	0.34	·	
(B2)	2	13	0.09	0.80	0.94	0.62	0.37
	4	15	0.30	0.97	0.86	0.84	
	6	17	0.38	0.99	0.89	·	
	8	19	0.39	0.58	0.29	0.86	
	10	21	0.17	0.74	0.56	0.58	

The table reports p-values of the $\hat{\chi}^2$ and the Wald test, under **FNP** and **PNP**, for the null hypothesis of correct specification for the regression and propensity score models. L and L^* denote the number of cells used for the Wald and J test, respectively. Additional splits are done using the **FS** algorithm for the Wald tests and **KM** for the $\hat{\chi}^2$ test. The columns KS2 and SS report the p-values of the Kolmogorov-Smirnov statistics of Stute and Zhu (2002) and Sant'Anna and Song (2019) tests, respectively. Tests for which the estimated parameter reached the boundary of the parameter space are not reported.

Table 5: Estimates of ATE on Wages Using Regressions

	Specification (A1)	Specification (B1)
ATE	12.96	8.14
Std. Error	(5.57)	(5.27)
Pr($>\|t\|$)	[0.02]	[0.12]

The table reports estimates of the ATE for the the logarithm of hourly wage in 1986 using two different specifications of the regression model. Estimates and standard errors are multiplied by 100.

The estimated Average Treatment Effect (ATE) for wages, derived from the potential outcomes regression model, is 50% larger than the estimate under specification (B1), while the standard errors from both models are quite similar. The ATE for wages of the IPW estimator using (B2) falls between the best and worst estimates from the regression models. Notably, the estimated ATTs for all outcomes using the IPW estimator are statistically negligible, suggesting homogeneous treatment effects across HBCU and TWI students.

These findings are consistent with the results of Fryer and Greenstone (2010) and a substantial portion of the literature on HBCUs (Price and Viceisza 2023), which identified similar positive effects of HBCU attendance on the wages of black students and the likelihood of obtaining a college degree.

Table 6: ATE and ATT Estimates Using IPW Estimators

	Specification (A2)			Specification (B2)		
	ln(wage)	Col. Degree	Grad. Degree	ln(wage)	Col. Degree	Grad. Degree
ATE	12.77	9.20	-0.43	10.25	8.75	-0.90
Std. Error	(6.30)	(4.73)	(3.76)	(6.12)	(4.93)	(3.63)
$\Pr(>\|t\|)$	[0.04]	[0.05]	[0.90]	[0.09]	[0.07]	[0.80]
ATT	5.65	4.00	-0.43	5.29	3.98	-0.61
Std. Error	(5.60)	(4.94)	(3.46)	(5.64)	(4.95)	(3.56)
$\Pr(>\|t\|)$	[0.31]	[0.41]	[0.90]	[0.34]	[0.42]	[0.86]

The table reports the IPW estimates of the ATE and ATT for the following outcomes: the logarithm of hourly wage, the probability of obtaining a bachelor's degree, and the probability of obtaining a graduate degree. Estimates and standard errors (in parentheses) are multiplied by 100, and p-values are enclosed in brackets. The estimates are obtained by excluding observations with a propensity score outside the interval $[0.05, 0.95]$.

It is worth noticing that the validity of the unconfoundedness and CMIA assumptions rely on the comprehensive inclusion of all relevant factors in the student's decision-making process regarding enrollment in an HBCU institution. When uncertainties surround the validity of the independence assumptions, the estimates of ATE and ATT presented below should be regarded as statistical associations rather than causal effects.

9 Conclusion

In conclusion, this article introduces a novel approach for validating the correct specification of conditional moment restriction (CMR) models, which are fundamental for identifying and estimating causal relationships. The proposed method harnesses the well-established chi-squared (χ^2) tests, commonly used in classical goodness-of-fit contexts, to goodness-of-fit checks for CMR specifications. Traditionally, CMR model checks have been adapted from tests for the cumulative density function (CDF) based on functionals of the standard empirical process (SEP), but these often exhibit limitations, especially when dealing with high-dimensional data and high-frequency alternatives. The introduced χ^2 tests, on the other hand, are distribution-free, do not necessitate complex bootstrapping or smoothing techniques, and offer flexibility in partitioning the data to favor specific alternative hypotheses. Thus, providing a valid complementary instruments to the existing model checking procedures. Monte Carlo simulations and empirical evidence highlight the effectiveness of these χ^2 tests, especially in scenarios involving a large number of

covariates. The empirical analysis demonstrates the practical application of these tests in assessing the returns of attending historically black colleges and universities (HBCU) for black students in the United States. The results suggest that, in line with the literature, HBCU are more effective than traditional white institutions (TWI) in advancing the economic and academic success of black students.

A Appendix A

A.1 Lemmas

I first state auxiliary lemmas for the propositions and theorems in the main text. Let $\rightsquigarrow$ denote weak convergence on $l^\infty(\mathbb{D})$ (see definition 13.3 in Van Der Vaart 1996, hereafter VW), where $l^\infty(\mathbb{D})$ is the space of all real-valued functions that are uniformly bounded on $\mathbb{D}$, and $\rightarrow_d$ denote convergence of real-valued random variables. Troughout, to highlight the dependency on the partition, denote as $\hat{\mathbf{\Phi}}_\theta(\boldsymbol{\gamma}) = \hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\theta)$ and $\hat{\mathbf{\Phi}}_0(\boldsymbol{\gamma}) = \hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\theta_0)$.

Lemma 1 *Under the null H_0,*

(a) if Assumption 1, 2, and 2' hold, then $\Sigma_{\boldsymbol{\gamma}}(\tilde{\theta}) = \Sigma_{\boldsymbol{\gamma},0} + o_p(1)$.

(b) if Assumption 1, 2, 3 hold, and $\text{Avar}\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$ is p.d., then $\hat{W}_{\boldsymbol{\gamma}}(\tilde{\theta}) = \text{Avar}\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) + o_p(1)$

Lemma 2 *Under the null hypothesis H_0, Assumptions 1, and 5,*

$$\hat{\mathbf{\Phi}}_0(\cdot) \rightsquigarrow \Phi_0(\cdot) \text{ as a process on } l^\infty(\mathbb{D}),$$

where $\Phi_0(\cdot)$ is an $\mathbb{R}^L$-valued Gaussian process with zero mean vector and covariance structure given by,

$$\mathbb{E}\left[\Phi_0(\boldsymbol{\gamma})\Phi_0(\tilde{\boldsymbol{\gamma}})\right] = \mathbb{E}\left[\varepsilon_{\theta_0}(Z)^2 \mathbf{I}_{\boldsymbol{\gamma}}(X)\mathbf{I}_{\tilde{\boldsymbol{\gamma}}}(X)'\right] \quad \forall \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}} \in \mathbb{D}.$$

Lemma 3 *Under the null hypothesis H_0, and Assumptions 1-5, it holds that:*

(a) $\sup_{\boldsymbol{\gamma}\in\mathbb{D}} \left| \hat{\mathbf{\Phi}}_{\tilde{\theta}}(\boldsymbol{\gamma}) - (\hat{\mathbf{\Phi}}_0(\boldsymbol{\gamma}) - \hat{\boldsymbol{\mu}}^{*\prime}_{\boldsymbol{\gamma}}(\theta_0)\sqrt{n}(\hat{\theta} - \theta_0)) \right| = o_p(1)$.

(b) $\hat{\boldsymbol{\mu}}^*_{\tilde{\boldsymbol{\gamma}}}(\theta_0) = \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} + o_p(1)$.

Lemma 4 *Under the null H_0, and Assumptions 4,5,*

(a) if Assumption 1, 2, and 2' hold, then $\Sigma_{\tilde{\boldsymbol{\gamma}}}(\tilde{\theta}) = \Sigma_{\boldsymbol{\gamma},0} + o_p(1)$.

(b) if Assumption 1, 2, 3 hold, and $\text{Avar}\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)$ is p.d., then $\hat{W}_{\tilde{\boldsymbol{\gamma}}}(\tilde{\theta}) = \text{Avar}\left(\hat{\mathbf{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right) + o_p(1)$

A.2 Proofs

For any class of functions $\mathcal{F}$, denote as $\{P_n f : f \in \mathcal{F}\}$ the empirical measure indexed by $\mathcal{F}$, such that $P_n f = n^{-1} \sum f(Z_i)$; alike, Pf denotes the population measure, $Pf = \int f(Z) dP$. We say that a class of functions is: i) Glivenko-Cantelli for P (hereafter, P-GC) whenever $\sup_{f \in \mathcal{F}} |P_n - P| f = o_p(1)$; ii) P-Donsker if $\{\sqrt{n}(P_n - P) f : f \in \mathcal{F}\}$ converge in distribution to a tight random element in the space $l^\infty(\mathcal{F})$. Throughout, we refer to both classes of sets with finite VC dimension and classes of functions with finite VC subgraph dimension as VC classes. These classes, having uniformly bounded covering numbers (Theorem 2.6.7 in VW), are Glivenko-Cantelli and Donsker (see Theorem 2.4.3 and 2.5.2 in VW) for any probability measure on the sample space, provided that they have integrable and square-integrable envelope function, respectively.

Proof of Lemma 1. For the first part of the Lemma, by the weak law of large numbers (WLLN) and a mean value theorem argument (MVT), suffices to show that

$$\frac{1}{n} \sum_{i=1}^{n} \left(\varepsilon^2_{\tilde{\theta}}(Z_i) - \varepsilon^2_{\theta_0}(Z_i) \right) \mathbb{I}_{\gamma_l}(X_i) = I + II + III = o_p(1)$$

for each $l \in 1, 2, ..., L$, where,

$$\begin{aligned} I &= (\tilde{\theta} - \theta_0)' \frac{1}{n} \sum_{i=1}^{n} \nabla m_{\bar{\theta}}(X_i) \nabla m_{\bar{\theta}}(X_i)' \mathbb{I}_{\gamma_l}(X_i)(\hat{\theta} - \theta_0), \\ II &= \frac{2}{n} \sum_{i=1}^{n} m_{\theta_0}(X_i) \nabla m_{\bar{\theta}}(X_i)' \mathbb{I}_{\gamma_l}(X_i)(\tilde{\theta} - \theta_0), \\ III &= \frac{2}{n} \sum_{i=1}^{n} Y_i (m_{\tilde{\theta}}(X_i) - m_{\theta_0}(X_i)) \mathbb{I}_{\gamma_l}(X_i), \end{aligned}$$

and $|\bar{\theta} - \theta_0| \leq |\tilde{\theta} - \theta_0|$. The triangle inequality, Assumption 2, and the consistency of $\tilde{\theta}$ show that, $|I| \leq d^2_{\bar{\theta}} \left\| \tilde{\theta} - \theta_0 \right\|^2 n^{-1} \sum_{i=1}^{n} R(X_i)^2 = o_p(1)$, where $\|\cdot\|$ denotes the euclidean

norm. By a similar reasoning,

$$\begin{aligned}|II| &\leq d_\theta \left\|\tilde{\theta} - \theta_0\right\| \frac{2}{n}\sum_{i=1}^{n} m_{\theta_0}(X_i)R(X_i) \\ &\leq d_\theta \left\|\tilde{\theta} - \theta_0\right\| \left(\mathbb{E}\left[Y^2\right]\right)^{1/2} \left(\mathbb{E}\left[R(X)^2\right]\right)^{1/2} + o_p(1) = o_p(1)\end{aligned}$$

where the last inequality follows from the WLLN, the law of iterated expectation, and Cauchy-Schwarz inequality. Finally, after expanding again around θ_0 it is easy to see that $|III| \leq d_\theta \left\|\tilde{\theta} - \theta_0\right\| 2n^{-1}\sum_{i=1}^{n} Y_i R(X_i)\mathbb{I}_{\gamma_l}(X_i) = o_p(1)$.

For the second part of the lemma, we need to show that $\hat{L}(\tilde{\theta}) = L_0 + o_p(1)$, $\hat{C}_{\boldsymbol{\gamma}}(\tilde{\theta}) = C_{\gamma,0} + o_p(1)$, and $\hat{\boldsymbol{\mu}}^*_{\boldsymbol{\gamma}}(\tilde{\theta}) = \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} + o_p(1)$. By the usual MVT argument and the law of large numbers,

$$L_n = L_0 + I + II + II' + o_p(1)$$

with,

$$\|I\| = \left\|\frac{1}{n}\sum_{i=1}^{n} \nabla l_{\bar{\theta}}(Z_i)\left(\hat{\theta} - \theta_0\right)\left(\hat{\theta} - \theta_0\right)' \nabla l_{\bar{\theta}}(Z_i)'\right\| \leq d_\theta^4 \left\|\hat{\theta} - \theta_0\right\|^2 \frac{1}{n}\sum_{i=1}^{n} R_2^2(Z_i) = o_p(1),$$

$$\begin{aligned}\|II\| &= \left\|\frac{1}{n}\sum_{i=1}^{n} \nabla l_{\bar{\theta}}(Z_i)\left(\hat{\theta} - \theta_0\right) l'_{\theta_0}(Z_i)\right\| \leq d_\theta^2 \left\|\hat{\theta} - \theta_0\right\| \frac{1}{n}\sum_{i=1}^{n} \|l_{\theta_0}(Z_i)\| R_2(Z_i) \\ &\leq d_\theta^2 \left\|\hat{\theta} - \theta_0\right\| \left(\frac{1}{n}\sum_{i=1}^{n} \|l_{\theta_0}(Z_i)\|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} R_2^2(Z_i)\right)^{1/2} = o_p(1).\end{aligned}$$

Alike, we write $\hat{C}_{\boldsymbol{\gamma}}(\tilde{\theta})$ as,

$$\hat{C}_{\boldsymbol{\gamma}}(\tilde{\theta}) = I - II - III + C_{\gamma,0} + o_p(1),$$

where,

$$I = \frac{1}{n}\sum_{i=1}^{n} l_{\theta_0}(Z_i)\nabla m_{\bar{\theta}}(X_i)'(\hat{\theta}-\theta_0)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'$$
$$II = \frac{1}{n}\sum_{i=1}^{n} \nabla l_{\bar{\theta}}(Z_i)'(\hat{\theta}-\theta_0)\varepsilon_{\theta_0}(Z_i)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'$$
$$III = \frac{1}{n}\sum_{i=1}^{n} \nabla l_{\bar{\theta}}(Z_i)(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)'\nabla m_{\bar{\theta}}(X_i)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'$$

By Assumptions 2, and 3,

$$\begin{aligned}\|I\| &\leq \left\|\hat{\theta}-\theta_0\right\| \frac{1}{n}\sum_{i=1}^{n} \|l_{\theta_0}(Z_i)\| \, \|\nabla m_{\bar{\theta}}(X_i)\| \, \|\mathbf{I}_{\boldsymbol{\gamma}}(X_i)\| \\ &\leq \sqrt{L}\left\|\hat{\theta}-\theta_0\right\| \left(\frac{1}{n}\sum_{i=1}^{n} \|l_{\theta_0}(Z_i)\|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} R^2(X_i)\right)^{1/2} = o_p(1),\end{aligned}$$

An analogous reasoning shows that $\|II\| = o_p(1)$, and,

$$\begin{aligned}\|III\| &\leq \left\|\hat{\theta}-\theta_0\right\|^2 \frac{1}{n}\sum_{i=1}^{n} \|\nabla l_{\bar{\theta}}(Z_i)\| \, \|\nabla m_{\bar{\theta}}(X_i)\| \, \|\mathbf{I}_{\boldsymbol{\gamma}}(X_i)\| \\ &\leq \sqrt{L} d_\theta^3 \left\|\hat{\theta}-\theta_0\right\|^2 \left(\frac{1}{n}\sum_{i=1}^{n} R^2(X_i)\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} R_2^2(Z_i)\right)^{1/2} = o_p(1).\end{aligned}$$

Finally, $\hat{\boldsymbol{\mu}}^*_{\boldsymbol{\gamma}}(\tilde{\theta}) = \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} + o_p(1)$ follows from the proof of Lemma 3 below. ■

Proof of Lemma 2. By Lemma 2.6.17 in VW and Assumption 5, both $\mathbb{D}$ and $\{\mathbf{I}_{\boldsymbol{\gamma}}(x) : \gamma \in \mathbb{D}\}$ are VC classes. Therefore, $\mathcal{F} = \{\varepsilon_{\theta_0}(z)\mathbf{I}_{\boldsymbol{\gamma}}(x) : \gamma \in \mathbb{D}\}$ is a VC class (Lemma 2.6.18 in VW), with square integrable envelope function $F = |\varepsilon_{\theta_0}|$, and, hence, is P-Donsker. The convergence of the finite-dimensional distributions (fidis) of $\hat{\boldsymbol{\Phi}}_0(\cdot)$ to those of $\boldsymbol{\Phi}_0(\cdot)$, by the multivariate central limit theorem, characterize the limit process. ■

Proof of Lemma 3. By an MVT argument,

$$\hat{\boldsymbol{\Phi}}_{\hat{\theta}}(\boldsymbol{\gamma}) = \hat{\boldsymbol{\Phi}}_0(\boldsymbol{\gamma}) - I'\sqrt{n}(\hat{\theta}-\theta_0) - \hat{\boldsymbol{\mu}}^{*'}_{\boldsymbol{\gamma}}(\theta_0)\sqrt{n}(\hat{\theta}-\theta_0)$$

where,

$$I = \hat{\boldsymbol{\mu}}^*_{\boldsymbol{\gamma}}(\bar{\theta}) - \hat{\boldsymbol{\mu}}^*_{\boldsymbol{\gamma}}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\big(\nabla m_{\bar{\theta}}(X_i) - \nabla m_{\theta_0}(X_i)\big)\mathbf{I}_{\boldsymbol{\gamma}}(X_i)'$$

and $|\bar{\theta} - \theta_0| \leq |\tilde{\theta} - \theta_0|$. The class $\{\nabla m_\theta(x) : \theta \in \Theta\}$ is a collection of continuous mapping, $\theta \to \nabla m_\theta$, over the compact metric space Θ with integrable envelope function $R(\cdot)$ and, therefore, is P-GC (e.g., Example 19.8 in Van der Vaart 2000). Thus,

$$\begin{aligned}\sup_{\boldsymbol{\gamma}\in\mathbb{D}}\|I\| &\leq \sup_{\gamma\in\mathbb{D}}\frac{1}{n}\sum_{i=1}^{n}\|\nabla m_{\bar{\theta}}(X_i) - \nabla m_{\theta_0}(X_i)\|\,\|\mathbf{I}_{\boldsymbol{\gamma}}(X_i)\| \\ &\leq \sqrt{L}\frac{1}{n}\sum_{i=1}^{n}\|\nabla m_{\bar{\theta}}(X_i) - \nabla m_{\theta_0}(X_i)\| = o_p(1).\end{aligned}$$

where the last equality follows from an application of the uniform law of large numbers (e.g., Davidson 1994, Theorem 21.6). For the second part of the Lemma is sufficient to prove that,

$$|II| = \left|\frac{1}{n}\sum_{i=1}^{n}\nabla^{(j)}m_{\theta_0}(X_i)\big(\mathbf{I}_{\hat{\boldsymbol{\gamma}}}(X_i) - \mathbf{I}_{\boldsymbol{\gamma}}(X_i)\big)\right| = o_p(1),$$

for each $j \in \{1, .., d_\theta\}$. To see this is true, notice that by Assumption 5, $\mathbb{D}\Delta\mathbb{D} = \{\boldsymbol{\gamma}_1\Delta\boldsymbol{\gamma}_2 : \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{D}\}$ is a class of subsets of unions of VC classes, and hence is VC. Therefore, $\{|\nabla^{(j)}m_{\theta_0}(x)|\mathbf{I}_{\tilde{\boldsymbol{\gamma}}}(x) : \tilde{\boldsymbol{\gamma}} \in \mathbb{D}\Delta\mathbb{D}\}$ is also VC with integrable envelope $R(\cdot)$ and, hence, P-GC. Thus, for each $j \in \{1, .., d_\theta\}$,

$$\begin{aligned}|II| &\leq \frac{1}{n}\sum_{i=1}^{n}|\nabla^{(j)}m_{\theta_0}(X_i)|\mathbf{I}_{\hat{\boldsymbol{\gamma}}\Delta\boldsymbol{\gamma}}(X_i) \\ &\leq \sup_{\tilde{\gamma}\in\mathbb{D}\Delta\mathbb{D}}(P_n - P)|\nabla^{(j)}m_{\theta_0}|\mathbf{I}_{\tilde{\gamma}} + \mathbb{E}\left[|\nabla^{(j)}m_{\theta_0}(X)|\mathbf{I}_{\hat{\boldsymbol{\gamma}}\Delta\boldsymbol{\gamma}}(X)\right] \\ &= o_p(1) + \mu_R(\hat{\boldsymbol{\gamma}}\Delta\boldsymbol{\gamma}) = o_p(1)\end{aligned}$$

where $\mu_R(\hat{\boldsymbol{\gamma}}\Delta\boldsymbol{\gamma}) = \big(\mu_R(\hat{\gamma}_1\Delta\gamma_1), ..., \mu_R(\hat{\gamma}_L\Delta\gamma_L)\big)'$, and $\mu_R(A) = \int_A R(X)dP$ is a (signed) measure absolutely continuous with respect to P. The last equality follows from Assumption 4. ■

Proof of Lemma 4. For each element on the main diagonal of $\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta}) - \hat{\Sigma}_{\boldsymbol{\gamma}}(\tilde{\theta})$ write,

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_{\hat{\theta}}(Z_i)^2(\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i)) = I + II + III$$

where

$$\begin{aligned} I &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{\theta_0}(Z_i)^2(\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i)) \\ II &= (\hat{\theta} - \theta_0)' \frac{1}{n}\sum_{i=1}^{n} \nabla m_{\bar{\theta}}(X_i)\nabla m_{\bar{\theta}}(X_i)'(\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i))(\hat{\theta} - \theta_0) \\ III &= -(\hat{\theta} - \theta_0)' \frac{2}{n}\sum_{i=1}^{n} \varepsilon_{\theta_0}(Z_i)\nabla m_{\bar{\theta}}(X_i)(\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i)) \end{aligned}$$

The class $\{\varepsilon_{\theta_0}(z)^2\mathbb{I}_{\tilde{\gamma}_l} : \tilde{\gamma} \in \mathbb{C}\Delta\mathbb{C}\}$ is VC with integrable envelope function $\varepsilon_{\theta_0}^2$ and, hence, is P-GC. Therefore, $|I| \leq \mu_\sigma(\hat{\gamma}_l \Delta \gamma_l) + o_p(1) = o_p(1)$, by Assumption 4. Also, $|II| \leq \sqrt{L}d_\theta^2 \left\|\hat{\theta} - \theta_0\right\|^2 n^{-1}\sum_{i=1}^n R(X_i)^2 = o_p(1)$, by Assumptions 1-2 and the consistency of $\hat{\theta}$, and $|III| \leq \sqrt{L}d_\theta \left\|\hat{\theta} - \theta_0\right\| n^{-1}\sum_{i=1}^n \varepsilon_{\theta_0} R(X_i) = o_p(1)$ by Cauchy-Schwarz inequality. Thus, $\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta}) = \hat{\Sigma}_{\boldsymbol{\gamma}}(\tilde{\theta}) + o_p(1)$, and the first part of the lemma follows from Lemma 1(a).

For the second part of the lemma, notice that by Lemma 3(b) (and the proof of the first part of Lemma 3), $\hat{\boldsymbol{\mu}}^*_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}) = \boldsymbol{\mu}^*_{\boldsymbol{\gamma},0} + o_p(1)$, and for each element of $C_n(\hat{\gamma}) - C_n(\gamma_0)$ it holds that,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{\hat{\theta}}(Z_i) l_{\hat{\theta},j}(\mathbb{I}_{\hat{\gamma}_l}(X_i) - \mathbb{I}_{\gamma_l}(X_i)) &\leq \left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_{\hat{\theta}}(Z_i)^2 \mathbb{I}_{\hat{\gamma}_l \Delta \gamma_l}(X_i)\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} l_{\hat{\theta},j}\right)^{1/2} \\ &= o_p(1)O_p(1), \end{aligned}$$

where $l_{\theta,j}$ denotes the j-th component of l_θ and the last equality follows from the first part of this proof and Lemma 1. ■

Proof of Theorem 1. The consistency of the grouped GMM estimator follows from $\sup_{\theta\in\Theta} |Q_n(\theta) - Q_0(\theta)| = o_p(1)$, where $Q_0(\theta) = \mathbb{E}\left[\varepsilon_\theta(Z)\mathbf{I}_{\boldsymbol{\gamma}}(X)\right]' (\Sigma_{\boldsymbol{\gamma},0})^{-1} \mathbb{E}\left[\varepsilon_\theta(Z)\mathbf{I}_{\boldsymbol{\gamma}}(X)\right]$ and $Q_n(\theta) = n^{-1}\hat{\chi}^2_{\hat{\boldsymbol{\gamma}},\tilde{\theta}}(\theta)$ (see Theorem 2.1 in Newey and McFadden 1994, for instance).

To see that the condition holds, notice that,

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_\theta(Z_i)\mathbf{I}_{\hat{\boldsymbol{\gamma}}}(X_i)-\mathbb{E}\left[\varepsilon_\theta(Z)\mathbf{I}_{\boldsymbol{\gamma}}(X)\right]\right|\leq\sup_{\theta\in\Theta}|I|+\sup_{\theta\in\Theta}|II|+\sup_{\theta\in\Theta}|III|$$

where

$$\begin{aligned}I&=(P_n-P)\varepsilon_\theta\mathbf{I}_{\boldsymbol{\gamma}},\\ II&=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{\theta_0}(Z_i)(\mathbf{I}_{\hat{\boldsymbol{\gamma}}}(X_i)-\mathbf{I}_{\boldsymbol{\gamma}}(X_i)),\\ III&=\frac{1}{n}\sum_{i=1}^{n}\left[\nabla m_{\bar{\theta}}(Z_i)(\mathbf{I}_{\hat{\boldsymbol{\gamma}}}(X_i)-\mathbf{I}_{\boldsymbol{\gamma}}(X_i))'\right]'(\theta-\theta_0).\end{aligned}$$

The mapping $\theta\to\varepsilon_\theta$ is continuous over the compact Θ, with

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\varepsilon_\theta(Z)\right]\leq\mathbb{E}\left[\sup_{\bar{\theta},\theta\in\Theta}\varepsilon_{\theta_0}(Z)-\nabla m_{\bar{\theta}}(X)'(\theta-\theta_0)\right]\leq d_\theta\mathbb{E}\left[R(Z)\right]D<\infty,$$

by Assumption 2 and the compactness of Θ, where D denotes the diameter of Θ. Therefore, both $\{\varepsilon_\theta(z):\theta\in\Theta\}$ and $\{\varepsilon_\theta(z)\mathbf{I}_{\boldsymbol{\gamma}}(X):\theta\in\Theta\}$ are P-GC classes (see Corollary 8.6 in Giné and Zinn 1984, for instance) and, hence, $\sup_{\theta\in\Theta}|I|=o_p(1)$. Then, by Cauchy-Schwarz inequality and Assumptions 1 and 4,

$$\sup_{\theta\in\Theta}|II|\leq\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{\theta_0}^2(Z_i)\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{\hat{\boldsymbol{\gamma}}\Delta\boldsymbol{\gamma}}(X_i)\right)^{1/2}=o_p(1).$$

Finally,

$$\begin{aligned}\sup_{\theta\in\Theta}\|III\|&\leq\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{I}_{\hat{\boldsymbol{\gamma}}\Delta\boldsymbol{\gamma}}(X_i)\|\,\|\nabla m_{\bar{\theta}}(X_i)\|\,\|\theta-\theta_0\|\\ &\leq d_\theta D\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{I}_{\hat{\boldsymbol{\gamma}}\Delta\boldsymbol{\gamma}}(X_i)\|\,R(X_i)=o_p(1),\end{aligned}$$

where the second inequality follows from Assumption 2 and the compactness of Θ. This result, together with the consistency of $\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})$ implies that $\sup_{\theta\in\Theta}|Q_n(\theta)-Q_0(\theta)|=o_p(1)$ and, therefore, $\hat{\theta}_{\hat{\boldsymbol{\gamma}}}=\theta_0+o_p(1)$. For the asymptotic normality: by Assumptions 2 and

1(c), the first-order conditions of the minimization problem are satisfied with probability approaching one, $\hat{\boldsymbol{\mu}}^*_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})^{-1}n^{-1/2}\hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}}) = 0$. Expanding $\hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})$ around θ_0 and solving gives the Bahadur representation,

$$\sqrt{n}(\hat{\theta}_{\boldsymbol{\gamma}} - \theta_0) = -\left[\hat{\boldsymbol{\mu}}^*_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})^{-1}\hat{\boldsymbol{\mu}}^{*'}_{\hat{\boldsymbol{\gamma}}}(\bar{\theta})\right]^{-1}\hat{\boldsymbol{\mu}}^*_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta})^{-1}\hat{\boldsymbol{\Phi}}_0(\hat{\boldsymbol{\gamma}})$$

where $|\bar{\theta} - \theta_0| \leq |\hat{\theta}_{\hat{\boldsymbol{\gamma}}} - \theta_0|$.

From the proof of Lemma 3, it follows that $\hat{\boldsymbol{\mu}}^*_{\hat{\boldsymbol{\gamma}}}(\theta) - \hat{\boldsymbol{\mu}}^*_{\hat{\boldsymbol{\gamma}}}(\theta_0) = o_p(1)$ for any $\theta \underset{p}{\to} \theta_0$. Thus by consistency of $\hat{\theta}_{\hat{\boldsymbol{\gamma}}}$ and Lemma 1(b), it follows that both $\hat{\boldsymbol{\mu}}^*_{\hat{\boldsymbol{\gamma}}}(\hat{\theta}_{\boldsymbol{\gamma}})$ and $\hat{\boldsymbol{\mu}}^*_{\hat{\boldsymbol{\gamma}}}(\bar{\theta})$ converge in probability to $\boldsymbol{\mu}^*_{\boldsymbol{\gamma},0}$. Finally, by Assumptions 4-5, and the uniform continuity of the sample paths of $\Phi_0(\cdot)$,

$$\Phi_0(\hat{\boldsymbol{\gamma}}) \xrightarrow{d} N\left(0, \Sigma_{\boldsymbol{\gamma},0}\right).$$

The result follows by Slutsky's lemma. ■

Proof of Theorem 2. By Lemma 3, 4, and Assumption 3 (or Theorem 1 and Assumption 2' for the $\hat{\chi}^2$ tests) both test statistics are asymptotically equivalent to the following quadratic form,

$$q(\hat{\theta}, W, \hat{\gamma}) = \left(\Phi_0(\hat{\boldsymbol{\gamma}}) - \sqrt{n}\boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0}\bar{l}_{\theta_0}\right)' W^{-1}\left(\Phi_0(\hat{\boldsymbol{\gamma}}) - \sqrt{n}\boldsymbol{\mu}^*_{\boldsymbol{\gamma},0}\bar{l}_{\theta_0}\right), \tag{A1}$$

where $\bar{l}_{\theta_0} = n^{-1}\sum_{i=1}^n l_{\theta_0}(Z_i)$ and W^{-1} is the probability limit of the weighting matrix W_n^{-1}. In particular, the couple $(\hat{\theta}, W_n^{-1})$ is equal to $(\hat{\theta}_{\hat{\boldsymbol{\gamma}}}, \hat{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\tilde{\theta}))$ in the $\hat{\chi}^2$ test and $\left(\tilde{\theta}, \widehat{\text{Avar}^-}\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right)\right)$ in the Wald test. The functional

$$\phi(z, w, \gamma) = \left(z(\gamma) - \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0}w\right)' W^{-1}\left(z(\gamma) - \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0}w\right),$$

mapping $(\hat{\boldsymbol{\Phi}}_0(\cdot), \sqrt{n}\bar{l}_{\theta_0}, \hat{\boldsymbol{\gamma}})$ into $q(\hat{\theta}, W, \hat{\boldsymbol{\gamma}})$ is continuous with respect to the product topology on $l^\infty(\mathbb{D}) \times \mathbb{R}^L \times \mathbb{D}$ (see Lemma 4 in Andrews 1988a). Thus, Theorem 2 follows by establishing the limit null distribution of $(\Phi_0(\hat{\boldsymbol{\gamma}}) - \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma},0}\sqrt{n}\bar{l}_{\theta_0})$ and an application of the continuous mapping theorem (e.g., Theorem 1.3.6 in VW). Lemma 2, Assumptions 1, 4, 5, and the central limit theorem imply that $(\Phi_0(\cdot), \sqrt{n}\bar{l}_{\theta_0}, \hat{\gamma})$ is a uniformly tight process

on $\mathbb{D}$ with fidis converging weakly to those of $(\Phi_0(\cdot), l_0, \gamma_0)$, where $l_0 \overset{d}{=} N(0, L_0)$ and $\mathbb{E}\left[l_0 \Phi_0(\boldsymbol{\gamma})\right] = \mathbb{E}\left[l_{\theta_0}(Z)\varepsilon_{\theta_0}(Z)\mathbf{I}_{\boldsymbol{\gamma}}(X)'\right]$. Thus,

$$(\hat{\boldsymbol{\Phi}}_0(\cdot), \sqrt{n}\bar{l}, \hat{\boldsymbol{\gamma}}) \rightsquigarrow (\Phi_0(\cdot), l_0, \gamma_0) \text{ on } l^\infty(\mathbb{D}),$$

and by the continuous mapping theorem,

$$q(\hat{\theta}, W, \hat{\gamma}) \xrightarrow{d} \tilde{Y}' W^{-1} \tilde{Y},$$

where $\tilde{Y} \overset{d}{=} N(0, \Sigma_{\tilde{Y}})$ and $\Sigma_{\tilde{Y}} = \text{Avar}\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\hat{\theta})\right)$. In the Wald test, where W^{-1} is a generalized inverse of $\Sigma_{\tilde{Y}}$, $\tilde{Y}' W^{-1} \tilde{Y} \overset{d}{=} \chi^2_{r(\text{Avar}\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\tilde{\theta})\right))}$ by Theorem 7.3(i) in Rao and Mitra (1972). In the $\hat{\chi}^2$ test, $W = \Sigma_{\boldsymbol{\gamma},0}$, and $\Sigma_{\tilde{Y}} = \text{Avar}\left(\hat{\boldsymbol{\Phi}}_{\boldsymbol{\gamma}}(\hat{\theta}_{\hat{\gamma}})\right) = \Sigma_{\boldsymbol{\gamma},0} - \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma}}(\boldsymbol{\mu}^*_{\boldsymbol{\gamma}} \Sigma^{-1}_{\boldsymbol{\gamma},0} \boldsymbol{\mu}^{*'}_{\boldsymbol{\gamma}}) \boldsymbol{\mu}^*_{\boldsymbol{\gamma}}$. Since $\text{Var}\left(\Sigma^{-1/2}_{\boldsymbol{\gamma},0}\tilde{Y}\right) = \Sigma^{-1/2}_{\boldsymbol{\gamma},0}(\Sigma_{\tilde{Y}})\Sigma^{-1/2}_{\boldsymbol{\gamma},0}$ is idempotent with rank equal to $L - d_\theta$, it follows that $\tilde{Y}' W^{-1} \tilde{Y} = (\Sigma^{-1/2}_{\boldsymbol{\gamma},0}\tilde{Y})'(\Sigma^{-1/2}_{\boldsymbol{\gamma},0}\tilde{Y}) \overset{d}{=} \chi^2_{L-d_\theta}$, proving the theorem. ■

Proof of Proposition 1. By an MVT argument, for fixed x, one can rewrite the difference between the two models fit as,

$$m_{\theta_0}(x) + \Delta m_{\bar{\theta}}(x)(\tilde{\theta} - \theta_0) = m_{1,\theta^*_1}(x) + \Delta m_{1,\bar{\theta}^*}(x)(\tilde{\theta}^* - \theta^*_1).$$

where $|\bar{\theta} - \theta_0| \leq |\tilde{\theta} - \theta_0|$ and $|\bar{\theta}^* - \theta^*_1| \leq |\tilde{\theta}^* - \theta^*_1|$. Then using conditions (a) and (b) of Proposition 2, we obtain, that the splitting points, x_0, solve,

$$\Delta m_{\bar{\theta}}(x_0) C n^{-1/2} \sum_{i=1}^{n} g(X_i, \varepsilon_{\theta_0}) - \Delta m_{1,\bar{\theta}^*}(x_0) D n^{-1/2} \sum_{i=1}^{n} g(X_i, \varepsilon_{\theta_0}) + o_p(1) = 0,$$

and, thus,

$$\Delta m_{\bar{\theta}}(x_0) C - \Delta m_{1,\bar{\theta}^*}(x_0) D = o_p(1).$$

■

Proof of Proposition 2. Let

$$d_{\boldsymbol{\gamma}}(I_L) = \delta(\boldsymbol{\gamma})'\delta(\boldsymbol{\gamma}) = \sum_{l=1}^{L} \delta_l^2 = \sum_{l=1}^{L} \mathbb{E}\left[h(X)\mathbb{I}_{\gamma_l}(X)\right]^2.$$

If $h(x) > 0$ ($h(x) \leq 0$) for all $x \in \mathcal{X}$, then $\delta_l \geq 0$ ($\delta_l \leq 0$) for all $l = 1, 2, ..., L$, and, since $(\delta_l + \delta_f)^2 \geq (\delta_l^2 + \delta_f^2)$ for each δ_l, δ_f such that $sign(\delta_l) = sign(\delta_f)$, it follows that,

$$\sum_{l=1}^{L} \delta_l^2 \leq \left(\sum_{l=1}^{L} \delta_l \right)^2 .$$

Thus, a unique cells maximize the drifts norm.

Consider, instead, the case $h(x) = 0$ for some $x \in \mathcal{X}$, then for any finite split of $\mathcal{X}$ with $L > 1$, we have that

$$\delta_l^2 + \delta_f^2 = (\delta_l^+ + \delta_l^-)^2 + (\delta_f^+ + \delta_f^-)^2 \leq (\delta_l^+ + \delta_f^+)^2 + (\delta_l^- + \delta_f^-)^2 = \delta_l^{*2} + \delta_f^{*2},$$

where δ_l^+ and δ_l^- denote the positive and negative part of δ_l, respectively. Therefore, any partition is dominated by a partition with the same number of cells obtained by merging the positive and negative part of each δ_l into a new cell. Since the δ_l^* are either non-negative or non-positive, by merging them in two cells containing only the positive and negative δ_l^* we obtain a partition with two cells that dominates any partition with $L > 1$ cells. Of course, since

$$(\delta_l^+ + \delta_l^-)^2 \leq (\delta_l^+)^2 + (\delta_l^-)^2,$$

the two-cells partition dominates the one-cell partition as well. Thus, the two-cells partition into positive and negative values maximizes the drifts norm. ■

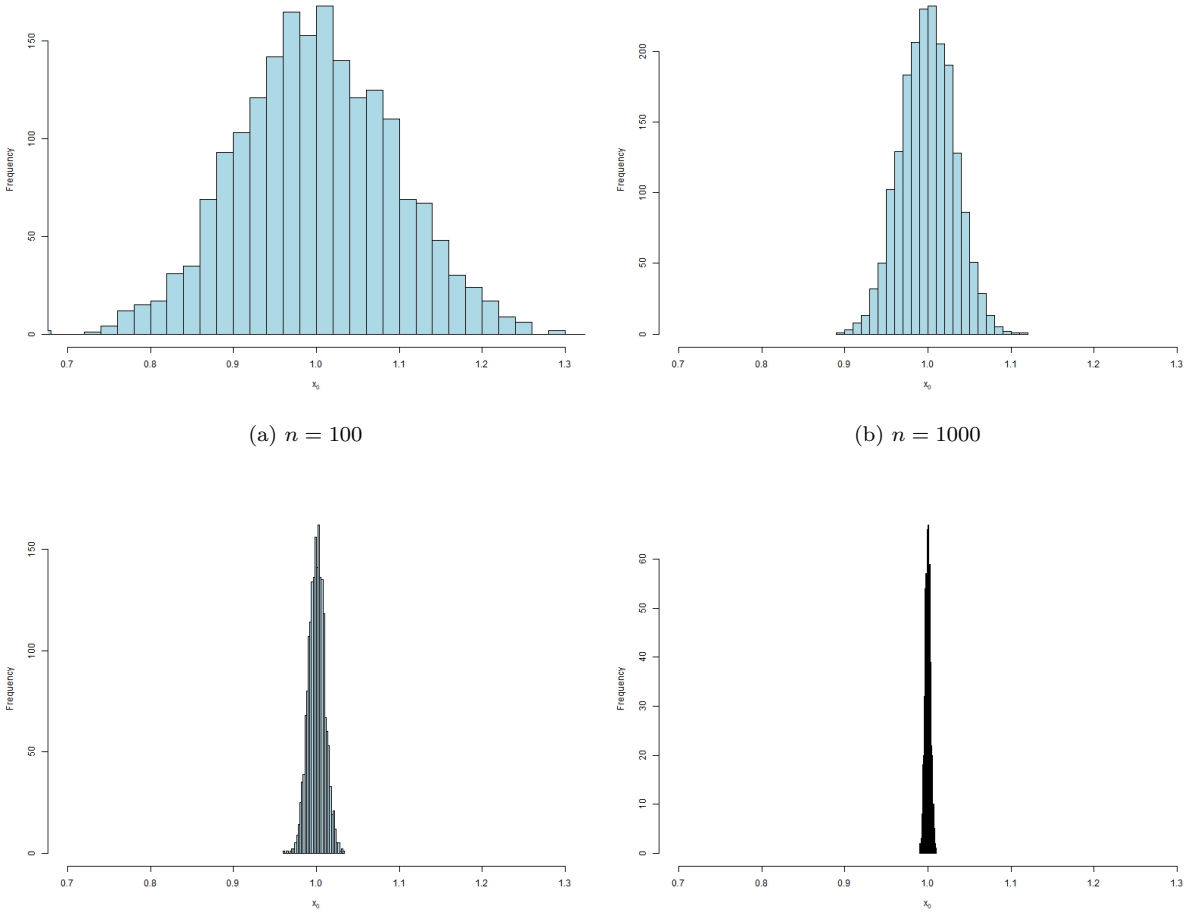
B Appendix B

B.1 Convergence of NP Cells: Simulations

I present Monte Carlo evidence demonstrating NP partitions converging to fixed cells in $\mathbb{D}$. In Figure 1, we generate data as $Y = 1 + \varepsilon$, where $\varepsilon \sim N(0, 1)$. Under the null hypothesis, $H_0 : m(X) = c$ a.s., with c as a constant. The alternative model, $H_1 : m(X) = c + \beta X$ a.s., includes a constant c and slope β, where X is generated as $X \sim N(1, 1)$. In Figure 2, we create a linear model, $Y = 1 + X + \varepsilon$, with $\varepsilon \sim N(0, 1)$. The null hypothesis is $H_0 : m(X) = c + \beta X$ a.s., with c and β as constants. The alternative

model, $H_1 : m(X) = c + \beta X + \gamma X^2$ a.s., introduces a curvature term γ. I consider sample sizes of $n = 100$, $n = 1000$, $n = 10000$, and $n = 10000$, to compute the splitting points using the NP partition algorithm over $R = 1000$ replications. The results, depicted in the figures below, align with theoretical findings, illustrating the convergence of splitting points to fixed points in $\mathcal{X}$.

Figure 4: Convergence of x_0 for the constant model under linear alternative.



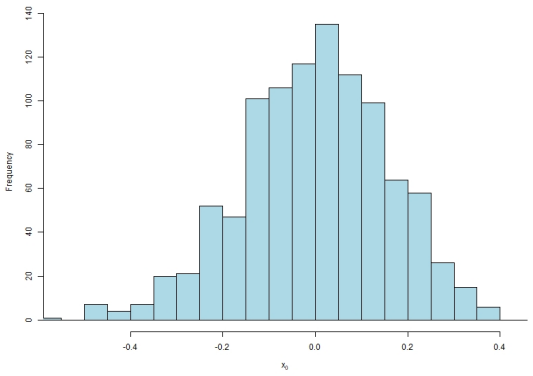
(a) $n = 100$

(b) $n = 1000$

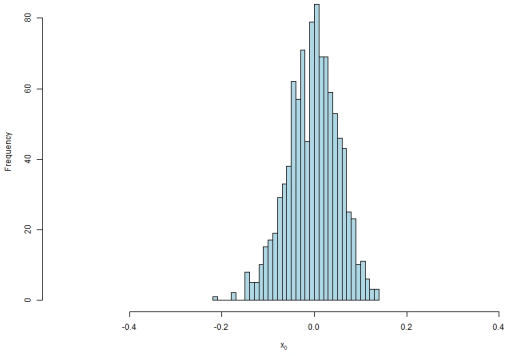
(c) $n = 10,000$

(d) $n = 100,000$

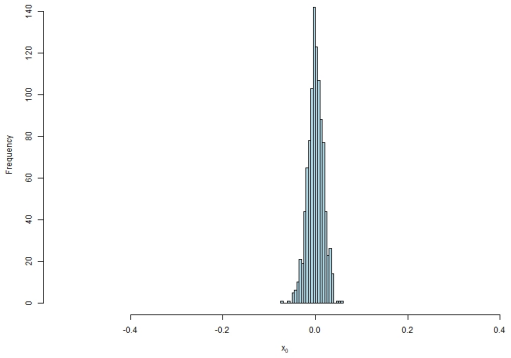
Figure 5: Convergence of x_0 for the linear model under quadratic alternative. The first four graphs and the last four refer to the two roots of the quadratic equation.



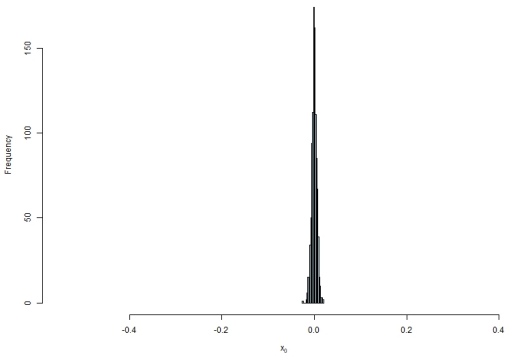
(a) $n = 100$



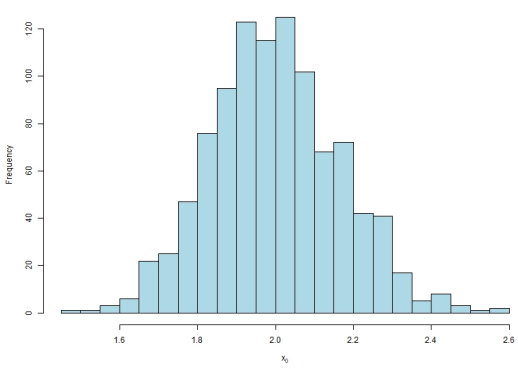
(b) $n = 1000$



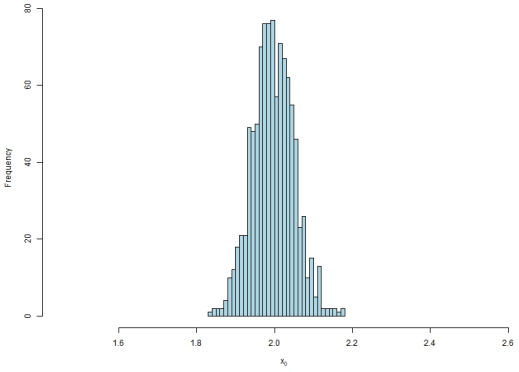
(c) $n = 10,000$



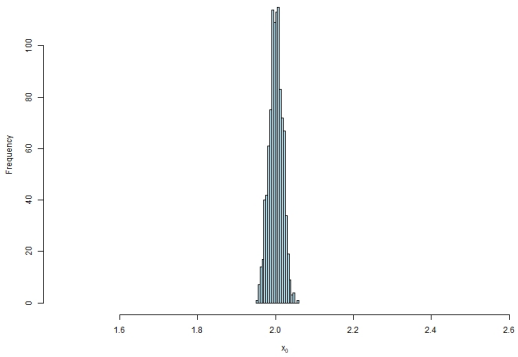
(d) $n = 100,000$



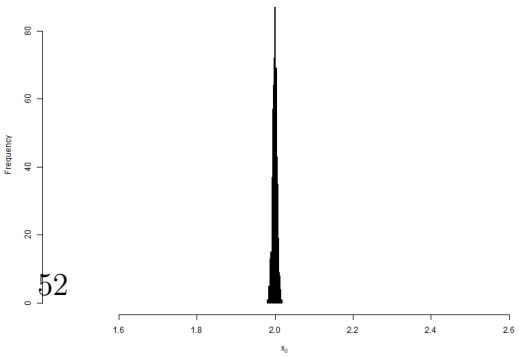
(e) $n = 100$



(f) $n = 1000$



(g) $n = 10,000$



(h) $n = 100,000$

References

Andrews, D. W. (1987): "Asymptotic results for generalized Wald tests," *Econometric Theory*, 3(3), 348–358.

——— (1988a): "Chi-square diagnostic tests for econometric models: theory," *Econometrica: Journal of the Econometric Society*, pp. 1419–1453.

——— (1988b): "Chi-square diagnostic tests for econometric models: Introduction and applications," *Journal of Econometrics*, 37(1), 135–156.

——— (1997): "A conditional Kolmogorov test," *Econometrica: Journal of the Econometric Society*, pp. 1097–1128.

Balakrishnan, N., V. Voinov, and M. S. Nikulin (2013): *Chi-squared goodness of fit tests with applications*. Academic Press.

Bickel, P. J., and M. Rosenblatt (1973): "On some global measures of the deviations of density function estimates," *The Annals of Statistics*, pp. 1071–1095.

Bierens, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20(1), 105–134.

Bierens, H. J., and W. Ploberger (1997): "Asymptotic theory of integrated conditional moment tests," *Econometrica: Journal of the Econometric Society*, pp. 1129–1151.

Billingsley, P. (2017): *Probability and measure*. John Wiley & Sons.

Cook, R. D. (1994): "On the interpretation of regression plots," *Journal of the American Statistical Association*, 89(425), 177–189.

Cox, D. R. (1972): "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202.

Cramér, H. (1946): *Mathematical methods of statistics*, vol. 26. Princeton university press.

Cristobal, J. C., P. F. Roca, and W. G. Manteiga (1987): "A class of linear regression parameter estimators constructed by nonparametric estimation," *The Annals of Statistics*, 15(2), 603–609.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96(1), 187–199.

Davidson, J. (1994): *Stochastic limit theory: An introduction for econometricians*. OUP Oxford.

Davidson, R., J. G. MacKinnon, et al. (2004): *Econometric theory and methods*, vol. 5. Oxford University Press New York.

Delgado, M. A., and J. Vainora (2023): "Conditional Distribution Model Specification Testing Using Chi-Square Goodness-of-Fit Tests," .

Domínguez, M. A., and I. N. Lobato (2004): "Consistent estimation of models defined by conditional moment restrictions," *Econometrica*, 72(5), 1601–1615.

Durbin, J., and M. Knott (1972): "Components of Cramér–von Mises statistics. I," *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 290–307.

Fryer, R. G., and M. Greenstone (2010): "The changing consequences of attending historically black colleges and universities," *American economic journal: Applied economics*, 2(1), 116–148.

Gessaman, M. (1970): "A consistent nonparametric multivariate density estimator based on statistically equivalent blocks," *The Annals of Mathematical Statistics*, 41(4), 1344–1346.

Giné, E., and J. Zinn (1984): "Some limit theorems for empirical processes," *The Annals of Probability*, pp. 929–989.

Greenwood, P. E., and M. S. Nikulin (1996): *A guide to chi-squared testing*, vol. 280. John Wiley & Sons.

Han, Y., P. Ma, H. Ren, and Z. Wang (2023): "Model checking in large-scale data set via structure-adaptive-sampling," *Statistica Sinica*, 33, 1–27.

Hardle, W., and E. Mammen (1993): "Comparing nonparametric versus parametric regression fits," *The Annals of Statistics*, pp. 1926–1947.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009): *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71(4), 1161–1189.

Kallenberg, W. C. M., J. Oosterhoff, and B. Schriever (1985): "The number of classes in chi-squared goodness-of-fit tests," *Journal of the American Statistical Association*, 80(392), 959–968.

Khmaladze, E. V. (1982): "Martingale approach in the theory of goodness-of-fit tests," *Theory of Probability & Its Applications*, 26(2), 240–257.

Khmaladze, E. V., and H. L. Koul (2004): "Martingale transforms goodness-of-fit tests in regression models," .

Kiefer, N. M., T. J. Vogelsang, and H. Bunzel (2000): "Simple robust testing of regression hypotheses," *Econometrica*, 68(3), 695–714.

Koul, H. L., and P. Ni (2004): "Minimum distance regression model checking," *Journal of Statistical Planning and Inference*, 119(1), 109–141.

Koutrouvelis, I. A., and J. Kellermeier (1981): "A goodness-of-fit test based on the empirical characteristic function when parameters must be estimated," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 43(2), 173–176.

Kuan, C.-M., and W.-M. Lee (2006): "Robust M tests without consistent estimation of the asymptotic covariance matrix," *Journal of the American Statistical Association*, 101(475), 1264–1275.

Lee, W.-M., C.-M. Kuan, and Y.-C. Hsu (2014): "Testing over-identifying restrictions without consistent estimation of the asymptotic covariance matrix," *Journal of Econometrics*, 181(2), 181–193.

Lütkepohl, H., and M. M. Burda (1997): "Modified Wald tests under nonregular conditions," *Journal of Econometrics*, 78(2), 315–332.

Moore, D. S. (1977): "Generalized inverses, Wald's method, and the construction of chi-squared tests of fit," *Journal of the American Statistical Association*, 72(357), 131–137.

Moore, D. S., and M. C. Spruill (1975): "Unified large-sample theory of general chi-squared statistics for tests of fit," *The Annals of Statistics*, pp. 599–616.

Newey, W. K. (1985): "Generalized method of moments specification testing," *Journal of econometrics*, 29(3), 229–256.

——— (1990): "Efficient instrumental variables estimation of nonlinear models," *Econometrica: Journal of the Econometric Society*, pp. 809–837.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Nikulin, M. S. (1973): "Chi-square test for continuous distributions with location and scale parameters," *Teoriya Veroyatnostei i ee Primeneniya*, 18(3), 583–591.

Pearson, K. (1900): "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157–175.

Polachek, S. W., et al. (2008): "Earnings over the life cycle: The mincer earnings function and its applications," *Foundations and Trends® in Microeconomics*, 4(3), 165–272.

Pollard, D. (1979): "General chi-square goodness-of-fit tests with data-dependent cells," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 50(3), 317–331.

——— (1980): "The minimum distance method of testing," *Metrika*, 27(1), 43–70.

——— (1984): *Convergence of stochastic processes*. Springer Science & Business Media.

Price, G. N., and A. C. Viceisza (2023): "What Can Historically Black Colleges and Universities Teach about Improving Higher Education Outcomes for Black Students?," *Journal of Economic Perspectives*, 37(3), 213–232.

Ramsey, J. B. (1969): "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 31(2), 350–371.

Rao, C. R., and S. K. Mitra (1972): "Generalized inverse of a matrix and its applications," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, vol. 6, pp. 601–621. University of California Press.

Rao, K. C., and B. Robson (1974): "A chi-squabe statistic for goodies-of-fit tests within the exponential family," *Communications in Statistics-Theory and Methods*, 3(12), 1139–1153.

Rosenbaum, P. R. (1987): "Model-based direct adjustment," *Journal of the American statistical Association*, 82(398), 387–394.

Rosenbaum, P. R., and D. B. Rubin (1983): "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2), 212–218.

Sant'Anna, P. H., and X. Song (2019): "Specification tests for the propensity score," *Journal of Econometrics*, 210(2), 379–404.

Schott, J. R. (2016): *Matrix analysis for statistics*. John Wiley & Sons.

Seetharaman, P., and P. K. Chintagunta (2003): "The proportional hazard model for purchase timing: A comparison of alternative specifications," *Journal of Business & Economic Statistics*, 21(3), 368–382.

Stute, W. (1997): "Nonparametric model checks for regression," *The Annals of Statistics*, pp. 613–641.

Stute, W., W. G. Manteiga, and M. P. Quindimil (1998): "Bootstrap approximations in model checks for regression," *Journal of the American Statistical Association*, 93(441), 141–149.

Stute, W., and L.-X. Zhu (2002): "Model checks for generalized linear models," *Scandinavian Journal of Statistics*, 29(3), 535–545.

Tauchen, G. (1985): "Diagnostic testing and evaluation of maximum likelihood models," *Journal of Econometrics*, 30(1-2), 415–443.

Van Buuren, S., and K. Groothuis-Oudshoorn (2011): "mice: Multivariate imputation by chained equations in R," *Journal of statistical software*, 45, 1–67.

Van der Vaart, A. W. (2000): *Asymptotic statistics*, vol. 3. Cambridge university press.

Van Der Vaart, J. W. (1996): *Weak convergence*. Springer.

Watson, G. S. (1959): "Some recent results in chi-square goodness-of-fit tests," *Biometrics*, pp. 440–468.

Wooldridge, J. M. (1990): "A unified approach to robust, regression-based specification tests," *Econometric Theory*, 6(1), 17–43.