NBER WORKING PAPER SERIES

CORRECTING ENDOGENEITY VIA NONPARAMETRIC COPULA CONTROL FUNCTIONS

Xixi Hu
Yi Qian
Hui Xie

Working Paper 33607
http://www.nber.org/papers/w33607

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2025, Revised July 2026

The authors gratefully acknowledge financial support from Social Sciences and Humanities Research Council of Canada [grant 435-2018-0519 and 435-2023-0306], Natural Sciences and Engineering Research Council of Canada [grant RGPIN-2018-04313 and RGPIN-2024-06629] and US National Institute of Health [grant R01CA178061]. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2025 by Xixi Hu, Yi Qian, and Hui Xie. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Correcting Endogeneity via Nonparametric Copula Control Functions
Xixi Hu, Yi Qian, and Hui Xie
NBER Working Paper No. 33607
March 2025, Revised July 2026
JEL No. C0, C01, C10, C14, C51

ABSTRACT

We propose a new framework to address endogenous regressors using a novel conditional copula endogeneity model. Endogenous regressor models are nonparametric and agnostic to the functions that determine the values of endogenous regressors from exogenous regressors and unobservables. To capture the regressor-error dependence unexplained by exogenous regressors, conditional Gaussian copulas are used to link the structural error terms and the nonparametric models for endogenous regressors. Building on the model, we develop a two-stage nonparametric control function approach for endogeneity correction without relying on instrumental variables. Specifically, the approach constructs control functions using nonparametric estimates of the conditional cumulative distribution functions of endogenous regressors given exogenous regressors. The method relaxes the assumption of regressors and error jointly following Gaussian copula dependence structure and eliminates the need to model regressors. It unifies and generalizes existing copula-based endogeneity correction methods, while minimizing assumptions about how endogenous regressors are determined. Unlike existing copula control function methods, it can handle discrete endogenous regressors (e.g., binary or low-count) by leveraging variation in relevant exogenous control regressors. We demonstrate the robustness and broad applicability of the proposed method compared to existing copula-based endogeneity correction methods in simulation studies and empirical applications.

Xixi Hu
University of British Columbia
xixi.hu@sauder.ubc.ca

Yi Qian
University of British Columbia
Department of Marketing
and NBER
yi.qian@sauder.ubc.ca

Hui Xie
University of Illinois Chicago
School of Public Health
huixie@uic.edu

1 Introduction

A common challenge in empirical research is estimating structural models with endogenous regressors. Ignoring correlations between regressors and the error term can lead to significant estimation bias, known as endogeneity bias. While instrumental variable (IV) methods address this issue, they rely on valid and strong instruments that satisfy exclusion restrictions. Identifying and defending such instruments in practice is often difficult (Wooldridge 2010, Ebbes et al. 2009). These challenges spur the development and application of IV-free methods that do not rely on observed IVs or on full-information modeling (Lewbel 1997, Rigobon 2003, Ebbes et al. 2005, Lewbel et al. 2024, Papies et al. 2017, 2023, Rutz and Watson 2019, Eckert and Hohberger 2023).

Among IV-free methods, copula-based endogeneity corrections have gained widespread use across fields, such as management, economics, marketing, strategy, and information systems (Qian et al. 2026). First introduced by Park and Gupta (2012), these methods, as described in Qian et al. (2026), address endogeneity by modeling the regressor-error dependence using copulas, a widely used multivariate dependence model (Danaher and Smith 2011). Recent advances have expanded copula endogeneity correction to endogenous regressors with relaxed distributional requirements or correlated with exogenous regressors, noncontinuous endogenous regressors, and panel data. [1]

Despite their flexibility, all existing copula endogeneity correction methods require parametric or semiparametric auxiliary modeling of the endogenous regressor P given a vector of observed exogenous regressors W and an unobserved independent random term ξ_p. Existing methods use either linear auxiliary models with additive separability (Breitung et al. 2024, Mayer and Wied 2026) or semiparametric transformation models for P (Haschka 2022, Yang et al. 2024, 2025, Liengaard et al. 2025). In practice, the economic function that determines the values of P is often nonadditive and nonseparable in W and ξ_p, vio-

[1] See Qian et al. (2026) for a guide for the appropriate use of these methods.

lating the assumption of additive separability. For example, an exogenous regressor (e.g., weather or season) can affect not only the mean of the endogenous regressor (e.g., mean price of ice cream) but also its dispersion (e.g., variance of ice cream price). This rules out the possibility that P is an additive function of W and ξ_p, as occurred in our empirical application of price elasticity estimation. As shown subsequently, copula correction mistakenly assuming additive separability in the auxiliary model for P can yield biased model estimates. Semiparametric-transformation models relax the additive separability of W and ξ_p on the P scale, but impose different functional form assumptions on the model. To the best of our knowledge, no copula correction methods exist that employ a nonparametric, possibly nonadditive model for P as a function of W and ξ_p. Furthermore, no existing copula control function methods can handle noncontinuous endogenous regressors, which are common in marketing (Qian et al. 2026).

To address these important research gaps, this article introduces a two-stage endogeneity correction procedure (2sCOPE-np) based on a nonparametric, possibly nonadditive model for P as $P = G(W, \xi_p)$, where no restrictions are imposed on the functional form of G or on the distribution of ξ_p. It fully nests the existing methods, which use (additively separable) linear auxiliary models or semiparametric-transformation auxiliary models, as special cases. The proposed approach eliminates additive separability and functional-form assumptions in the auxiliary models for P, using data to nonparametrically determine the form of the function G that generates P from W and ξ_p, without imposing any restrictive assumptions on G. We then use conditional copulas to link this nonparametrically determined G function and the structural error (or its endogenous part) in the outcome model of interest *given* the exogenous regressors W. This is equivalent to using conditional copulas to link the random term ξ_p in the auxiliary model for P and the structural error in the outcome model, when the G function is strictly monotone in ξ_p. As a result, 2sCOPE-np can handle endogeneity due to omitted variables, simultaneity, and reverse causality (Web

Appendix C.1).[2]

Under the general conditional copula endogeneity framework, 2sCOPE-np consists of two steps. The first step derives the control variable, which is constructed using the conditional cumulative distribution function (CDF) of the endogenous regressor P given the exogenous regressors W (hereafter referred to as the conditional CDF). The conditional CDF is non-parametrically estimated, requiring neither specification of the first-stage model that determines the value of P, nor estimation of any joint models for the outcome and regressors. The control variable is shown to break the regressor-error dependence in that when the outcome model estimation conditions on the control variable, all regressors and the structural error become uncorrelated. Then, in the second step, the control variable is added in the outcome model as a generated regressor to address endogeneity. 2sCOPE-np nests all existing copula-based control function approaches as special cases and makes the following contributions to the literature.

First, 2sCOPE-np is agnostic to the functions determining the values of endogenous regressors (hereafter referred to as first-stage auxiliary models) and minimizes the estimation bias from misspecifying the auxiliary models. By introducing nonparametric, possibly nonadditive models for endogenous regressors, 2sCOPE-np eliminates the need to specify or justify first-stage auxiliary models, allowing for unknown function forms and potentially complex, interactive effects of the W variables and ξ_p on the distribution of P. By contrast, existing copula correction methods either assume independence between endogenous and exogenous regressors (hence P is determined by ξ_p only) or postulate specific first-stage models. Unlike linear additive auxiliary models which restricts W to only affect the mean of P, 2sCOPE-np allows W to affect all moments of P. Unlike semiparametric transfor-

[2] The copula model does not encompass all linear additive endogeneity models in which the endogenous regressor itself is linearly related to the error term in the outcome model. Such linear additive endogeneity models are not point identified without external IVs or additional distributional assumptions. We discuss the limited practical relevance of these models, given their strong additive separability assumptions, and present theoretical considerations and diagnostic tools for detecting or ruling them out (Section 3.2.3).

mation auxiliary models, 2sCOPE-np achieves this flexibility without restricting functional forms of the first-stage models. To our knowledge, 2sCOPE-np is the first copula method with a model-free first stage that requires no auxiliary model specification.

Second, 2sCOPE-np relaxes the key dependence assumption of existing copula-based endogeneity correction methods, which is that the regressors P, W and the structure error (or its endogenous part) jointly follow a Gaussian copula (GC) dependence structure. By using conditional GC copulas that link the nonparametric auxiliary model and the unobservable in the outcome model, 2sCOPE-np captures the regressor-unobservable dependence unexplained by exogenous regressors and nests existing copula models as special cases. As will be shown later, the conditional GC specification has increased flexibility and robustness and is more general than the joint GC model, which is widely regarded as flexible in capturing multivariate dependence (Danaher and Smith 2011). This is because the endogenous regressors and unobservables, unconditionally, are now not required to follow any pre-specified joint dependence structure, representing a substantial relaxation of existing approaches that impose a particular copula model to characterize their joint dependence. The structural error term is also not required to follow any specific distribution.

Third, 2sCOPE-np can handle endogenous regressors that are noncontinuous or have limited support[3]. Unlike existing copula control function methods, which can be inconsistent due to non-unique inverse mappings from the discrete marginal CDFs (Qian and Xie 2024), 2sCOPE-np avoids this problem by using inverse mappings from conditional CDFs that leverage variations in exogenous regressors. Our result aligns with Trivedi and Zimmer (2017)'s finding that identification of copula models with discrete margins can be achieved through variation in relevant exogenous variables. This result has broad practical implications, as endogenous regressors with insufficient support, such as binary treatment or low-count variables, are common in practice.

[3]For example, multiple observations may share the same value of the endogenous regressor. Common cases include binary or count endogenous regressors with few unique values.

Lastly, 2sCOPE-np offers a practical and doubly robust alternative in applications where valid IVs are unavailable or where the exclusion restriction is difficult to defend. Like exogenous regressors, instrumental variables must be independent of the unobserved determinants of the outcome. Additionally, they must satisfy the exclusion restriction, meaning they affect the outcome only through the endogenous regressors and have no direct effect on the outcome, making valid IVs considerably harder to identify and validate. 2sCOPE-np requires neither the condition that the endogenous regressor contain an additive exogenous component (observed or latent) satisfying the exclusion restriction, nor knowledge of functional forms of the first-stage model. Unlike the popular control function approach using IVs (Petrin and Train 2010), 2sCOPE-np employs model-free control functions and eliminates the need for IVs. Simulation studies and theoretical extensions also demonstrate that the assumptions used to derive 2sCOPE-np are sufficient but not necessary conditions, indicating the method's robustness. [4]

We apply our method in two empirical applications: estimating the price elasticity and assessing the return of education. In the pricing empirical application, the proposed 2sCOPE-np produces similar estimation results to those using a known strong price IV, thus cross-validating each other. In the estimation of the return to education, which is measured by years of education (a count regressor) and by postsecondary education (a binary regressor), 2sCOPE-np yields return-to-education estimates that are much more accurate and usable than the conventional estimates using the popular but weak IV.

The remainder of the paper is organized as follows. Section 2 introduces the proposed modeling framework, highlight the advantages of a nonparametric approach for modeling the conditional distribution of endogenous regressors, and develop flexible conditional copulas to capture endogeneity. We then derive the corresponding control function procedures and extend the framework to accommodate multiple endogenous regressors and hetero-

[4]See Section 3.2 for a discussion on the assumptions.

geneous samples. We also delineate how our approach differs from existing methods. In particular, the framework emphasizes the central role of the conditional CDF in addressing endogeneity. We show that our method unifies existing copula-based control function approaches and nests them as special cases.

Section 3 provides practical guidance for implementation, including strategies for handling higher-order and interaction terms, approaches to managing the dimensionality of conditional CDF estimation, and extensions to panel and clustered data. We also discuss the underlying assumptions, their practical implications and feasibility, methods for assessing their validity, as well as boundary conditions and diagnostics to mitigate potential pitfalls. Section 4 presents simulation results demonstrating that our approach improves accuracy, robustness, and applicability relative to existing copula-based control function methods across a range of settings. Section 5 illustrates the proposed method in two empirical applications,[5] and Section 6 concludes.

2 Method

2.1 An Overview of the Proposed Approach

For simplicity, we focus on the following linear structural model[6]

$$Y_i = \mu + \alpha P_i + W_i^{'}\beta + E_i, \tag{1}$$

where $i = 1, 2, \cdots, n$ indexes either time or cross-sectional units. Y_i is a (1×1) dependent variable; P_i is a (1×1) continuous endogenous regressor; W_i is a $(Q \times 1)$ vector of exogenous regressors that may include both continuous and discrete variables; E_i is the structural error term, and μ, α, β are the model parameters. Throughout the paper, we

[5]Sample code accompanying the paper has been uploaded to a GitHub repository and is available upon request. Due to confidentiality concerns during the review process, the repository is currently private, but access can be provided upon request.

[6]We follow the econometric convention that defines a structural model as one whose parameters have causal interpretations rather than a regression model merely capturing statistical associations (Wooldridge 2010, Ch. 4). A structural model may be derived from economic theory or may represent causal relationships that generate observed data. See Park and Gupta (2012) and Yang et al. (2025) for applications of copula methods to identifying and estimating structural models in the presence of endogenous regressors.

maintain the following assumptions.

Structural Model Assumptions: (1) E_i is independently distributed; (2) $\mathbb{E}\left[(1, P_i', W_i')'(1, P_i', W_i')\right]$ has full rank; and (3) E_i can be correlated with P_i, but E_i is uncorrelated with W_i.

Assumption (1) states that E_i is independent across $i = 1, \cdots, n$, with its distribution left unspecified. Because the intercept μ is included, we can assume $\mathbb{E}(E_i) = 0$ without loss of generality, where $\mathbb{E}(\cdot)$ is the expectation operator. Assumption (2) implies the full rank of the regressor design matrix, and no perfect collinearity among the regressors in Equation 1. Assumption (3) states that P_i can be correlated with E_i, generating the regressor endogeneity problem. In contrast, W_i, the exogenous control variables, are uncorrelated with E_i.[7] Dependence between W_i and P_i is allowed.

Formally, we express the joint distribution of (E, P, W)[8] as

$$f_{\theta,\phi,\psi}(E, P, W) \quad = \quad f_{\theta,\phi}((E, P)|W) f_{\psi}(W). \tag{2}$$

In the above, $f_{\theta,\phi}((E, P)|W)$ denotes the joint density for the structural error E and the endogenous regressor P given the exogenous regressors W, where $\theta = (\theta_1, \theta_2)$ with θ_1 denoting the parameters in the model of interest in Equation 1 and θ_2 denoting the dependence parameters between P and E given W; ϕ denotes the parameters governing the distribution of $P|W$. Since the marginal distribution of the exogenous regressors, $f_{\psi}(W)$, is ancillary to (θ, ϕ), inference about (θ, ϕ) depends only on $f_{\theta,\phi}((E, P)|W)$.

Regressor endogeneity arises in nonexperimental settings because units cannot be randomized to different levels of the endogenous variable P, allowing unmeasured confounders in the error term E_i to differ across levels of P and creating a correlation between P and E. Because the distribution of the endogenous regressor P is informative about θ_1 through its dependence on the structural error E, single equation methods such as ordinary least squares (OLS) applied to Equation 1 ignore this information and therefore can produce

[7] Note that W_i needs not to be randomly assigned (Wooldridge 2010).

[8] When possible, we suppress the subscript index i for notational simplicity.

biased estimates that reflect associations rather than causal effects. Regressor endogeneity is pervasive and a central challenge in empirical work using observational data.

To correct for endogeneity bias, our approach accounts for regressor-error dependence. It does so without specifying the joint conditional distribution $f_{\theta,\phi}((E,P)|W)$ or optimizing the corresponding joint likelihood function, thereby improving robustness and reducing computation. Specifically, we propose a computationally tractable bias correction method that augments Equation 1 with a control variable $\Phi^{-1}(F_\phi(P|W))$[9], derived from a conditional copula model capturing regressor-error dependence. As a result, the method can be readily implemented using standard software designed for models that assume all regressors are exogenous. In the next section, we introduce a nonparametric model for $P|W$, which is typically not of primary interest but an auxiliary model needed to correct endogeneity in the structural model; thus, robustness takes priority over simplicity. We then introduce the conditional copula model that links the outcome model and the nonparametric auxiliary regressor model, followed by deriving control variables under the model.

2.2 A Nonparametric Model for $P|W$ Without Additive Separability

For the endogenous variable P, our approach considers the following nonparametric model given the exogenous regressors in W and the unobserved term ξ_p, which can be nonadditive,

$$\begin{aligned} P &= G(W, \xi_p), & (3) \\ &= M(W) + \Lambda(W, \xi_p), & (4) \end{aligned}$$

where the random unobserved term ξ_p is distributed independently of W; the function G is unknown, and can be a structural equation or simply a function determining the values of P from exogenous variables W and unobserved disturbance ξ_p; $M(W)$ is the conditional mean of P given W; $\Lambda(W, \xi_p)$ is the difference between P and its conditional mean $M(W)$.

[9] $F_\phi(P|W)$ denotes conditional CDF of P given W; $\Phi^{-1}(\cdot)$ is the inverse standard normal CDF.

The nonparametric model in Equation 3 recognizes that the economic function that determines the values of the endogenous regressor P is often nonadditive, nonseparable in W and ξ_p, such that $\Lambda(W, \xi_p)$ generally depends on both W and ξ_p. Equation 3 includes the following two distinct models as special cases.

2.2.1 Special Case I: Semiparametric transformation models for $P|(W, \xi_p)$

$$G^{-1}(P) = H(W, \gamma) + \xi_p, \tag{5}$$

where $H(W, \gamma)$ are unknown transformations of W with the unknown parameters γ. The semiparametric transformation model accommodates bounded and skewed outcomes. By modeling the underlying latent transformed variable, it avoids restrictions imposed by additive separability on the observed scale. This flexibility makes it applicable to a wide range of empirical settings (Horowitz 1996). The model falls within the class of nonparametric models defined in Equation 3 because its restricts $G(W, \xi_p) = G(H(W, \gamma) + \xi_p)$.

Yang et al. (2025) develop 2sCOPE procedure under a GC model for (P, W, E)[10], hereafter referred to as 2sCOPE-GC, with the first-stage model for P following the above semiparametric transformation model with $\xi_p = \rho O + R$, where O denotes the omitted variable contained in the structural error E, R is the independent disturbance term and the parameter ρ captures the endogeneity. This semiparametric transformation model for P is motivated by the idea that the underlying unbounded latent variable (copula transformation of P) can be expressed as a linear combination of copula-transformed exogenous regressors, omitted variables contained in the structural error, and the noise R (Yang et al. 2025; Qian and Xie 2024 Web Appendix D.1). When P is normally distributed, their first-stage model reduces to an additive separable model directly on P as $P = H(W, \gamma) + \xi_p$.

By modeling the underlying unrestricted latent data, the semiparametric transformation endogeneity model in Equation 5 avoids the issue of logical inconsistency that can occur if an additive separable model is directly imposed on P, e.g., when P is bounded or when

[10]The joint GC model is also used in other works (Haschka 2022, Liengaard et al. 2025, Yang et al. 2024).

W affects not only the mean but also higher moments of P (Horowitz 1996, Qian et al. 2026). Notably, the GC model used in 2sCOPE captures dependence structures through the rank ordering of the data and is therefore invariant to strictly monotonic transformations of individual variables. Consequently, the validity of the GC model does not depend on specifying the correct marginal distributions of the regressors or identifying the appropriate transformations. As long as there exist strictly monotonic transformations of the regressors (e.g., logarithmic transformations) under which the GC model is valid, applying the GC model directly to the raw data will correctly recover the underlying dependence structure.

The nonparametric first-stage model in Equation 3 employed in the proposed 2sCOPE-np model continues to possess these merits of the semiparametric transformation models, while relaxing the additive separability assumption on $G^{-1}(P)$ in Equation 5 and avoiding the need to impose a joint GC assumption on the regressors and the structural unknowns. This improves the accuracy and broadens the applicability of copula correction, as shown in evaluations using simulations and real-life data subsequently.

2.2.2 Special Case II: Linear models with additive separability for $P|(W, \xi_p)$

$$P \quad = \quad H(W, \gamma) + \xi_p, \tag{6}$$

which is the first-stage mean regression model assumed in Breitung et al. (2024) and Mayer and Wied (2026). This is a special case of the nonparametric model in Equation 4 with the conditional mean $M(W) = H(W, \gamma)$ and $\Lambda(W, \xi_p) = \xi_p$. Thus, their approaches impose the implicit assumption that W and ξ_p are additive, separable on P. Their procedures require the copula transformation of ξ_p to be a linear deterministic function of the endogenous part of E, which achieves model identification.

The assumption of $\Lambda(W, \xi_p) = \xi_p$, implied by additive separability, means that $\Lambda(W, \xi_p)$, or the difference between P and its conditional mean given W, is unrelated to W. That is, W affects only the mean of P and cannot influence higher-order characteristics of the distribution of P, such as its variance or skewness. This restriction limits the applicability of these methods in certain empirical settings. For example, this assumption is known to be violated for bounded or truncated variables, but can also be questionable for any continuous endogenous regressors. [11] We name this assumption mean-dependence-only, and refer to this method as 2sCOPE-MD hereafter. We note that our proposed method (and 2sCOPE-GC) does not require the strong assumption of mean-dependence-only. In addition, although 2sCOPE-MD originated from copula correction, it generally does not permit a GC regressor-error dependence structure, as demonstrated in the simulation in Section 4.2. As shown later, copula endogeneity correction requires the conditional CDF of P given W rather than merely the conditional mean; the conditional mean is useful to predict P but generally insufficient for copula correction.

The nonparametric first-stage model in Equation 3 employed in our proposed procedure relaxes the additive separability assumption and permits the first-stage mean regression error to depend on both ξ_p and W without imposing functional-form assumptions on the G function or imposing distributional assumptions on ξ_p. By fully embedding the two distinct models described above, our proposed method unifies and generalizes these two methods (2sCOPE-GC and 2sCOPE-MD), making copula correction agnostic to the specification of the auxiliary regressor models.

2.3 The Conditional Copula Endogeneity Model

To capture regressor endogeneity, we link Equations 1 and 3 via flexible conditional copulas. The regressor endogeneity often arises from correlation between P and unobserved confounders in the error E. To make the source of endogeneity explicit, write $E = O + \epsilon$,

[11] Bounded, censored, or truncated variables are common in marketing and economics (e.g., Danaher and Smith 2011). Atefi et al. (2018) considered a case in which the endogenous regressor is the proportion of the sales personnel being trained, which is a fractional variable bounded between 0 and 1. Count endogenous regressors (e.g., the number of promotions or the number of years in education) are bounded below from zero. The product pricing and promotion can be bounded by minimum and maximum values (Qian and Xie 2011, Kocherlakota 2021). Consumer purchase frequency often exhibits left truncation, as firms' databases only include customers who have made at least one purchase (Danaher and Smith 2011, Qian and Xie 2022). Our empirical application later also shows that the exogenous regressors affect more than just the mean of the category price distribution, violating the additive separability assumption.

where O denotes the mean-zero combined effect of omitted confounders with $\mathbb{E}(O|P,W) \neq 0$ due to endogeneity; ϵ is a mean-zero disturbance term that includes only exogenous shocks, such that $\mathbb{E}(\epsilon|P,W) = 0$, but otherwise the distribution of ϵ is unspecified (e.g., permitting error nonnormality or heteroscedasticity). Because O embodies endogeneity, it suffices to model the joint conditional distribution of (O,P) given W: $f_{\theta,\phi}((O,P)|W)$.

To obtain $f_{\theta,\phi}((O,P)|W)$, we use conditional copulas to link O and P given W. Let $F_{\theta,\phi}((O=o,P=p)|W=w) = \Pr((O \leq o, P \leq p)|W=w)$ be the joint CDF of (O,P) conditional on W, and let $F_{\theta_1}(O=o|W=w) = \Pr(O \leq o|W=w)$, $F_{\phi}(P=p|W=w) = \Pr(P \leq p|W=w)$ be the univariate CDFs of O and P conditional on W, respectively. Both $F_{\theta_1}(o|w)$ and $F_{\phi}(p|w)$ have a uniform distribution on $[0,1]$.[12] Then, according to Sklar's Theorem (Sklar 1959), there exists a unique copula function, $C(\cdot,\cdot)$, which links the two uniform marginal distributions on $[0,1]$, $F_{\theta_1}(o|w)$ and $F_{\phi}(p|w)$, to the two-dimensional joint conditional CDF, $F_{\theta,\phi}((o,p)|w)$ (Joe 2015):

$$F_{\theta,\phi}((o,p)|w) \quad = \quad C_{\theta_2}(F_{\theta_1}(o|w), F_{\phi}(p|w)). \tag{7}$$

Equation 7 expresses the joint distribution of (O,P) conditional on W as a copula function $C_{\theta_2}(\cdot,\cdot)$ of the marginal CDFs of O and P conditional on W. Copula models are widely used for modeling multivariate dependence, offering greater flexibility than traditional methods like linear correlation. They separate the estimation of marginal distributions from dependence structures, allowing nonparametric marginals that preserve key features for endogeneity correction and model identification. A popular and robust copula model is the Gaussian copula (Danaher and Smith 2011), which we use for this conditional copula. Gaussian Copula is also scalable to high-dimensional data, making it practical and computationally tractable.

[12]This is known to hold when O and P are continuous. For noncontinuous P, this requires relevant exogenous regressors in W generate sufficient variations for the values of $F_{\phi}(p|w)$ to be continuous and follow approximately a uniform distribution on [0,1]. This is akin to the requirement that relevant exogenous regressors in W generate sufficient variations in $\mathbb{E}(P|W)$, which aids identification in copula models with discrete marginals (Trivedi and Zimmer 2017).

Let $\Psi_\rho(\cdot,\cdot)$ be the CDF of the bivariate standard normal distribution with a conditional correlation coefficient, ρ, where ρ captures the regressor-error dependence unexplained by the exogenous regressors in W. A non-zero ρ indicates the presence of endogeneity. The conditional Gaussian copula states the following model:

$$F_{\theta,\phi}((o,p)|w) \quad = C_{\theta_2}(F_{\theta_1}(o|w), F_\phi(p|w)) = \Psi_\rho(\Phi^{-1}(u_{o|w}), \Phi^{-1}(u_{p|w})), \tag{8}$$

$$\text{where} \qquad u_{o|w} = F_{\theta_1}(o|w), u_{p|w} = F_\phi(p|w). \tag{9}$$

$\Phi(\cdot)$ is the CDF of standard normal distribution, and $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard normal distribution. The conditional copula endogeneity model has the following joint density function conditional on W.

$$\begin{aligned} f_{\theta,\phi}((o,p)|w) \quad &= \tfrac{f_{\theta_1}(o|w)f_\phi(p|w)}{(1-\rho^2)^{1/2}} \exp\Big\{-\tfrac{\rho^2}{2(1-\rho^2)}(\Phi^{-1}(u_{o|w})^2 + \Phi^{-1}(u_{p|w})^2) \\ &\qquad + \tfrac{\rho}{1-\rho^2}\Phi^{-1}(u_{o|w})\Phi^{-1}(u_{p|w})\Big\}, \end{aligned} \tag{10}$$

where $f_{\theta_1}(o|w)$ and $f_\phi(p|w)$ are univariate conditional density functions of O and P given W, respectively. Since the distribution of ϵ is left unspecified, the joint conditional distribution $f_{\theta,\phi}((E,P)|W)$ remains unspecified.

When the G function that determines P is strictly monotone in ξ_p, the conditional copula model in Equation 8 is equivalent to assuming a conditional copula model for the unobservables (O, ξ_p) given W.[13]

2.4 Derivation of Endogeneity Control Functions

To derive the control function estimation procedure under the conditional copula endogeneity model, we express the structural model in Equation 1 as:

$$Y \quad = \quad \mu + p\alpha + w'\beta + \underbrace{\mathbb{E}(o|p,w)}_{\text{Control Function}} + \underbrace{\overbrace{o+\epsilon}^{\text{Structural error } E=O+\epsilon} \qquad -\mathbb{E}(o|p,w)}_{\text{Exogenous Disturbance } \eta}, \tag{11}$$

where $\mathbb{E}(o|p,w)$ is the expectation of the omitted effect o given p and w, and the new error term $\eta = (o - \mathbb{E}(o|p,w)) + \epsilon$ is uncorrelated with all other right-hand side random

[13]This holds because when $P = G(W,\xi_p)$ is strictly monotone in ξ_p, $F_\phi(P|W) = F_\phi(G(W,\xi_p)|W) = F_\phi(\xi_p|W)$. Both the semiparametric transformation models (Equation 5) and the linear models with additive separability (Equation 6) are strictly monotone in ξ_p.

variables $p, w, \mathbb{E}(o|p,w)$ in Equation 11[14]. Thus, if $\mathbb{E}(o|p,w)$ were known and included in the regression, it would break regressor-error dependence and yield consistent structural model estimates. Hence, $\mathbb{E}(o|p,w)$ serves as the control function and is derived in Proposition 1.

Proposition 1. *Under a normal distribution $N(0,\sigma_o^2)$ for O and the conditional GC model in Equation 8 for (O,P) given W with correlation coefficient ρ, then*

$$\mathbb{E}(o|p,w) = \sigma_o \cdot \rho \cdot C_p,$$

where the copula-generated regressor $C_p = \Phi^{-1}(u_{p|w}) = \Phi^{-1}(F_\phi(p|w))$, and $(F_\phi(p|w))$ is the conditional CDF of P given W.

Proof. See Web Appendix D.1. □

Consequently, the correction term $C_p = \Phi^{-1}(F_\phi(p|w))$ can be added into the Y model as a generated regressor to control for endogeneity. When $\rho = 0$ (i.e., no endogeneity), the true coefficient for C_p is zero, so testing this coefficient provides a direct test for endogeneity. The copula term C_p depends on the conditional CDF, $F_\phi(p|w)$, which is unknown and must be estimated. Because $F_\phi(p|w)$ is typically not of primary interest but an auxiliary model needed to correct endogeneity in the structural model, robustness takes priority over simplicity. Thus, we choose not to impose functional form assumptions on $F_\phi(p|w)$, but instead estimate it nonparametrically using the kernel methods proposed by Li and Racine (2008). Details on the kernel estimation methods can be found in Web Appendix B.1. Consistency of the two-step estimation procedure requires consistency of the plug-in estimates obtained from the first-stage estimation (Wooldridge 2010). Li and Racine (2008) shows that, under standard regularity conditions, these kernel estimators provide consistent estimates of conditional CDFs.

[14] This holds because $\mathbb{E}(\eta|p,w,\mathbb{E}(o|p,w)) = 0$, which is a sufficient condition for the zero correlation of η with each of $(p,w,\mathbb{E}(o|p,w))$ (Wooldridge 2010). Note that $\mathbb{E}(\eta|p,w,\mathbb{E}(o|p,w)) = \mathbb{E}(O - \mathbb{E}(o|p,w) + \epsilon|p,w,\mathbb{E}(o|p,w)) = 0$ since (1) it is apparent that $\mathbb{E}(O - \mathbb{E}(o|p,w)|p,w,\mathbb{E}(o|p,w)) = 0$ and (2) $\mathbb{E}(\epsilon|p,w,\mathbb{E}(o|p,w)) = \mathbb{E}(\epsilon|p,w) = 0$ since $\mathbb{E}(o|p,w)$ is a function of (p,w) and ϵ is an exogenous disturbance term satisfying $\mathbb{E}(\epsilon|p,w) = 0$.

In practice, endogenous regressors are often conceptually continuous but measured or recorded coarsely, leading to discrete marginals. In the return to education example in Section 5, the endogenous regressor, education, is recorded in whole years. In this case, it is natural to smooth such variables, which is common in nonparametric regression analysis. In other cases where the endogenous regressors are truly discrete, such as low-count variables (e.g., the number of promotions), we treat them as coarsened forms of hypothetical and continuous variables and apply smoothing accordingly. The kernel estimator defined in Equation 43 in the Web Appendix shows that the resulting conditional CDF estimates take on continuous values with no ties, even if p is discrete, as long as the exogenous regressors in W are non-trivially related to the endogenous regressor p. This parallels how variation in exogenous variables aids identification in copula models with discrete marginals (Trivedi and Zimmer 2017). 2sCOPE-np differs from latent variable models in that the kernel conditional CDF estimator does not assume a specific distributional form for the underlying conceptually continuous variable.

2.5 Standard Errors

Additional care is needed for inference, as 2sCOPE-np is a two-step procedure. Rather than relying on analytic asymptotic variances—whose form can be complex and may perform poorly in finite samples—we obtain standard errors using a bootstrap procedure. Bootstrapping offers a flexible and reliable approach for two-step estimators, as it naturally accounts for the sampling variability introduced in both stages.

To apply the bootstrap procedure with a sample of n independent observations, first randomly draw n observations with replacement from the original sample. For each bootstrap sample, apply the two-step 2sCOPE-np procedure to estimate the parameters. This process is repeated many times to obtain a sample of parameter estimates. The empirical distribution of the resulting bootstrap estimates approximates the sampling distribution of the estimator, and the standard deviations of this sample of estimated parameters provide the bootstrapped standard error for the parameters of interest.

2.6 Multiple Endogenous Regressors

Let the structural model contain K continuous and possibly correlated endogenous regressors $(P_1, \cdots, P_K)$,

$$Y_i = \mu + \sum_{k=1}^{K} P_{i,k}\alpha_k + W_i^{'}\beta + E_i. \tag{12}$$

We extend the bivariate conditional GC model in Equation 8 to the following multivariate conditional GC model with correlation matrix Σ:

$$\begin{aligned} F_{\theta,\phi}((o, p_1, \cdots, p_K)|w) = & \quad C_{\theta_2}(F_{\theta_1}(o|w), F_{\phi_1}(p_1|w), \cdots, F_{\phi_K}(p_K|w)), & (13) \\ = & \quad \Psi_\Sigma(\Phi^{-1}(u_{o|w}), \Phi^{-1}(u_{p_1|w}), \cdots, \Phi^{-1}(u_{p_K|w})), & (14) \\ \text{where} & \quad u_{o|w} = F_{\theta_1}(o|w), u_{p_k|w} = F_{\phi_k}(p_k|w),\ k = 1, \cdots, K. & (15) \end{aligned}$$

Proposition 2 then extends the result of Proposition 1 to multiple endogenous regressors.

Proposition 2. *Under a normal distribution $N(0, \sigma_o^2)$ for the omitted effect O and the multivariate conditional GC model in Equation for $(O, P_1, \cdots, P_K)$ given W with correlation matrix $\Sigma = \begin{pmatrix} 1 & \Sigma_{OP} \\ \Sigma_{OP}^\top & \Sigma_{PP} \end{pmatrix}$, then*

$$\mathbb{E}(o|p_1, \cdots, p_K, w) = \sum_{k=1}^{K} \gamma_k C_{p_k},$$

where the k^{th} copula-generated regressor $C_{p_k} = \Phi^{-1}(F_{\phi_k}(p_k|w))$, $F_{\phi_k}(p_k|w)$ is the conditional CDF of P_k given W, and $(\gamma_1, \cdots, \gamma_k) = \sigma_o \left[\Sigma_{OP}\, \Sigma_{PP}^{-1}\right]$.

Proof. See Web Appendix D.2. □

2sCOPE-np approach estimates the following augmented regression model:

$$Y_i = \mu + \sum_{k=1}^{K} P_{i,k}\alpha_k + X_i^T\beta + \beta' W_i + \sum_{k=1}^{K} C_{i,p_k}\gamma_k + \eta_i, \tag{16}$$

Table 1 summarizes 2sCOPE-np estimation procedure.

Table 1: Summary of 2sCOPE-np Estimation Procedure

Stage 1:
• For the k-th endogenous regressor P_k, perform nonparametric kernel conditional CDF estimation for P_k given exogenous regressors W (Equation 41 or Equation 43) and obtain the copula correction term $C_{i,p_k} = \Phi^{-1}(\widehat{F}_{\phi_k}(P_{i,k}\|W_i))$.
• Repeat the above for $k = 1, \cdots, K$ and obtain $\{C_{i,p_k}\}$ for all K endogenous regressors.
Stage 2:
• Add $\{C_{i,p_k}\}, k = 1, \cdots, K$, to the outcome structural regression model as generated regressors to control for endogeneity of $\{P_k\}$. The augmented regression model takes the form of Equation 16.

Note: Bootstrap is used to obtain standard errors of parameter estimates. When W contains multiple variables and/or mixed types of continuous and discrete variables, a generalized product kernel is used in conditional CDF estimation (Li and Racine 2008).

2.7 Extensions to Heterogeneous Samples

Suppose that researchers have reasons to believe that the data consists of J groups with heterogeneous endogeneity within groups; Proposition 3 below shows that our model can accommodate heterogeneous groups. A benefit of using the heterogenous groups strategy is that it also allows non-normality in the marginal distribution of the omitted variable.

Proposition 3. *Suppose that the data can be partitioned into J heterogeneous samples. Assume that conditions in Table 3 are satisfied within each sample j; such that either the structural error E_j or the omitted variable $O_j \sim N(0, \sigma_j^2)$, either $(O_j, P_{j1}, \cdots, P_{jK})|W_j$ or $(E_j, P_{j1}, \cdots, P_{jK})|W_j$ follows the multivariate conditional GC model in subsample j with correlation matrix ,* $\Sigma_j = \begin{pmatrix} 1 & \Sigma_{O_jP_j} \\ \Sigma_{O_jP_j}^\top & \Sigma_{P_jP_j} \end{pmatrix}$*, and full column rank for identification. Then*

$$\mathbb{E}(o|p_1, \cdots, p_K, w) = \sum_{k=1}^{K} \sum_{j=1}^{J} I(w \in S_j)\gamma_{jk} C_{p_{jk}},$$

where $I(\cdot)$ is the indicator function, S_j denotes the subsample j, and the k^{th} copula-generated regressor for subsample j is $C_{p_{jk}} = \Phi^{-1}(F_{\phi_{jk}}(p_{jk}|w_j))$ where $F_{\phi_{jk}}(p_{jk}|w_j)$ is the conditional CDF of P_{jk} given W_j in subsample j, and

$$(\gamma_{j1}, \cdots, \gamma_{jk}) \;=\; \sigma_j \left[\Sigma_{O_jP_j} \Sigma_{P_jP_j}^{-1}\right]$$

Proof. See Web Appendix D.3. □

Proposition 3 allows the GC correlation structure (and therefore the endogeneity structure) to vary across subsamples and relaxes the assumption of O being normally distributed. In fact, the marginal distribution of O is a mixture of J normal distributions [15] and need not follow a normal distribution [16].

According to Proposition 3, the augmented regression adds J copula-generated terms, $I(w \in S_j)C_{p_{jk}}, j = 1, \cdots, J$, for each endogenous regressor P_k, where S_j denotes subsample j. Equivalently, one can also parameterize the J coefficients as $\gamma_1 + \sum_{j=2}^{J} \gamma_{j1} I(w \in S_j)$ for $j \in \{1, \ldots, J\}$. , with group 1 set as the reference group, such that the control function coefficients γ_{j1} for the group $j \neq 1$ test whether there are group differences between the group j and the reference group. We use this parameterization in the empirical example of return to education.

2.8 Summary and Comparison of Methods

In summary, 2sCOPE-np generalizes and unifies existing copula-based control functions approaches as summarized in Table 2. Notably, the copula correction term $C_p = \Phi^{-1}(\widehat{F}_\phi(P|W))$, derived in Proposition 1, has strong face validity and sheds light on earlier copula control function methods. It offers a statistically general and natural extension of the original P&G method (Park and Gupta 2012) that explicitly accounts for correlation between exogenous and endogenous regressors. When P and W are independent, the conditional CDF $F_\phi(P|W)$ reduces to the marginal CDF $F(P_i)$, and consequently the copula term simplifies to $P_i^* = \Phi^{-1}(\widehat{F}(P_i))$ as used in the P&G method.

The 2sCOPE-np procedure also generalizes and unifies recently developed residual-based two-stage methods (Yang et al. 2025, 2024, Breitung et al. 2024, Liengaard et al.

[15] Mechanically speaking, a mixture of normal distributions is highly flexible and can approximate a wide variety of complex (marginal) distributional shapes.

[16] The error E is allowed to be nonnormal or heteroscedastic regardless of the distribution of O, because $E = O + \epsilon$ and the exogenous component ϵ is left unspecified except for the condition $\mathbb{E}(\epsilon|P, W) = 0$.

Table 2: A Comparison of Copula Correction Methods

Methods	Estimation Method	Permit both GC & non-GC dependence		First-stage dependence model for (P, W)	Permit normal endogenous regressor	Permit noncontinuous endogenous regressor
		$P-E$	$P-O$			
Park and Gupta (2012)*	Likelihood CF	No Yes	No No	P and W assumed independent	No	No
Haschka (2022)	Likelihood	No	No	GC	No	No
Qian and Xie (2024) (SORE)	Likelihood	Yes	Yes	Semiparametric odds ratio model	No	Yes
Yang et al. (2025, 2024) Liengaard et al. (2025) (2sCOPE-GC)	CF	Yes	No	GC	Yes	No
Breitung et al. (2024) Mayer and Wied (2026) (2sCOPE-MD)	CF	No	No	Mean-dependence -only model	No	No
Lin and Song (2025)	Likelihood	TC	No	TC	Yes	No
This Work (2sCOPE-np)	**CF**	**Yes**	**Yes**	**Model free**	**Yes**	**Yes**

Note: *: Other earlier work (Christopoulos et al. 2021, Tran and Tsionas 2021, Becker et al. 2021, Eckert and Hohberger 2023) also assume independence between P and W and require nonnormal and continuous endogenous regressors. Ignoring correlations between P and W can lead to biases in copula correction (Haschka 2022, Yang et al. 2025). CF: control function that can use moment-based estimation. P and W denote the endogenous and exogenous regressors, respectively; E denotes the structural error; O denotes the omitted variable; GC: Gaussian copula; TC: t-copula.

2025, Mayer and Wied 2026). Under the auxiliary first-stage models assumed in these approaches, the general copula correction term in 2sCOPE-np $C_p = \Phi^{-1}(\widehat{F}_\phi(P|W))$ reduces to their corresponding residual-based copula terms. Importantly, 2sCOPE-np shows that the appropriate control function can be estimated nonparametrically without postulating specific first-stage auxiliary regressors models. Furthermore, Proposition 3 shows that, in heterogeneous samples, 2sCOPE-np nests several existing approaches (Christopoulos et al. 2021, Liengaard et al. 2025, Yang et al. 2024) as special cases.

In addition, 2sCOPE-np relaxes the dependence structure assumed by existing copula endogeneity models. Unlike prior copula correction methods, 2sCOPE-np imposes dependence assumptions after conditioning on the exogenous regressors W, thereby capturing the endogeneity unexplained by W and relaxing the assumption needed on the dependence structure among P, W, and E. It generalizes Park and Gupta (2012), which models the *unconditional* joint distribution between structural error and endogenous regressor. Compared with unconditional joint copula models that jointly specify the dependence structure of (P, W, E) (Haschka 2022, Yang et al. 2025, Liengaard et al. 2025, Lin and Song 2025), conditioning on W improves robustness and reduces computational burden by lowering the number of variables being modeled. More importantly, under our framework, neither (P, O) nor (P, E) must follow GC or any specific copula dependence structure unconditionally, making the conditional copula model much more flexible and applicable.

Relative to likelihood-based alternatives (Haschka 2022, Qian and Xie 2024, Lin and Song 2025) in Table 2, 2sCOPE-np does not require specifying joint or marginal distributions for regressors and error terms, nor does it rely on a likelihood function. Consequently, 2sCOPE-np is more robust, computationally simpler, and readily implementable using standard software designed for models without endogenous regressors. In contrast, one-step joint likelihood estimation may offer greater estimation efficiency (smaller standard errors) and support established likelihood-based tests and model comparisons (Qian and Xie 2024). See Web Appendix A for a more detailed comparison of these methods.

3 Practical Guidance for Applying 2sCOPE-np

3.1 Implementation Considerations for 2sCOPE-np

Practical applications can encounter multiple endogenous regressors and functions of endogenous regressors (e.g., polynomial or interaction terms). Furthermore, our method relies on nonparametric conditional distribution estimation for each P_k given W. As with any nonparametric approach, one potential concern is the curse of dimensionality, which may create sparse-data issues in the local neighborhoods used for estimation. Several features of our approach and practical implementation considerations help mitigate these concerns.

3.1.1 Multiple endogenous regressors

For multiple and possibly correlated endogenous regressors $\{P_k, k = 1, \cdots, K\}$, 2sCOPE-np only requires performing the univariate conditional CDF estimation $(\widehat{F}_{\phi_k}(P_k|W))$ for each P_k one by one and separately from the other $K - 1$ endogenous regressors, without the need to perform a joint conditional CDF estimation on all endogenous regressors $(P_1, \cdots, P_k)$ simultaneously (Proposition 2 and Table 1). Thus, the computational workload for 2sCOPE-np only increases linearly with the number of endogenous regressors and is scalable to multiple endogenous regressors.

3.1.2 Endogenous higher-order or interaction terms

Adding copula control functions for any endogenous higher-order terms (e.g., P_k^2) or interaction terms (e.g., $P_k * P_{k-1}$ or $P_k * W$) appearing in the outcome model is neither necessary nor desirable. In fact, doing so is suboptimal and substantially degrades performance (Qian et al. 2026). Only copula terms for the first-order endogenous terms $(P_1, \cdots, P_k)$ are needed. This considerably simplifies implementation and represents an important advantage of copula correction methods.

3.1.3 Excluding Covariates in W that are irrelevant for P

Including irrelevant covariates can be particularly costly because the effective sample size within a local neighborhood decreases rapidly as the number of conditioning variables grows, making the estimation of the conditional CDF increasingly difficult. As a practical measure to mitigate the curse of dimensionality, we recommend excluding from the conditional CDF estimation variables that are known *a priori* to be irrelevant, such as prognostic variables in W that affect Y but not P_k.

A powerful feature of the kernel-based conditional CDF estimators for $\widehat{F}_{\phi_k}(P_k|W)$ is its automatic variable-selection and dimension-reduction property, such that covariates in W

that are irrelevant for P are automatically removed during the conditional CDF estimation process (Li and Racine 2008). This useful feature substantially reduces the curse of dimensionality issue in the presence of irrelevant covariates, a situation that arises quite often in empirical work (Li and Racine 2008).

3.1.4 Exogenous higher-order and interaction terms in W

As a nonparametric conditional CDF estimator, $\widehat{F}_{\phi_k}(P_k|W)$ does not require conditioning on polynomials terms (e.g., W_1^2) or interaction terms (e.g., $W_1 * W_2$): any such terms should be removed from the conditional CDF estimation because the nonparametric estimator automatically accommodates such nonlinearities (Hayfield and Racine 2008). This effectively reduces the burden of functional-form specification when estimating $\widehat{F}_{\phi_k}(P_k|W)$, as conditioning on higher-order transformations of W is unnecessary.

3.1.5 Categorical covariates in W for P and Panel/Clustered Data

Another common cause of high-dimensionality is that W contains categorical variables with possibly many levels (e.g., panel or clustered data). The nonparametric conditional CDF can handle and smooth categorical Ws via specialized kernels (Web Appendix B.1). An alternative approach is to do nonparametric conditional CDF estimation by groups, which estimates the nonparametric conditional CDF of P given only continuous W variables separately within each group defined by the combination of categorical W variables. This group-wise approach reduces the number of W variables to be smoothed and may also reduce computation time while yielding similar conditional CDF estimates as those obtained from smoothing all W variables, when each group has a sufficient sample size for reliable nonparametric conditional CDF estimation by groups.

When the levels of the combination of categorical Ws are large and cannot be collapsed to a small number relative to sample sizes, smoothing categorical W variables may become computationally burdensome, and the conditional CDF estimation by groups may encounter the issue of too few observations to estimate within-group conditional CDFs. In this case, to balance robustness and feasibility, a recommended approach is to treat the categorical levels as fixed effects and remove these effects through within-group demeaning of P and the continuous W regressors in a similar way to the within-group transformation in panel data to remove panel fixed-effects (Qian et al. 2026). Then apply the conditional CDF estimation and Stage 1 of 2sCOPE-np in Table 1 to the group-demeaned continuous regressors to obtain control functions. Variations of this approach can be considered in practice. For example, in data clustered by zipcode, the within-group demeaning could be done at the zipcode level, while the conditional CDF estimation using group demeaned regressors could be done by states or provinces to allow for sufficient sample size for conditional CDF estimation within each group. Similarly, demeaning could be performed for each panel unit (e.g., firms), while the nonparametric conditional CDF estimation could be performed by groups of panel units (e.g., industrial sectors of firms).

3.2 Assessing Assumptions and Identification: Practicality, Diagnostics, Robustness, and Recommendations

Table 3 lists the assumptions to derive 2sCOPE-np and to ensure consistent estimation and model identification. We note that 2sCOPE-np can also be derived under normal E and a conditional GC model for $(E, P)|W$ without decomposing E into $O + \epsilon$. Importantly, Assumptions 1 and 2 mean that the normal error distribution and GC regressor-error dependence conditions are sufficient but not necessary conditions for deriving 2sCOPE-np (Web Appendix C.3).[17] Furthermore, although Assumptions 1 and 2 are used to derive 2sCOPE-np, they are not strictly required (Web Appendix C.4). [18]

[17]Web Appendix C.3 shows that 2sCOPE-np works under weaker conditions than the normal error distribution and the conditional GC regressor-error dependence condtions.

[18]Web Appendix C.4 shows that 2sCOPE-np has reasonable robustness to violations of the assumptions.

Table 3: Assumptions for 2sCOPE-np Derivation and Identification

Assumption 1. *The error E or its endogenous part O is normally distributed.*
Assumption 2. *Either $(P, O)|W$ or $(P, E)|W$ follows a conditional GC.*
Assumption 3. *The copula correction term C_p is not a linear combination of $(1, P, W)$.*

Note: Assumption 1 means that the error term E can be nonnormally distributed. Assumption 2 states that GC captures the regressor endogeneity unexplained by exogenous control covariates in W, and that regressor-error needs not to follow a specific copula structure (Web Appendix C.3). Although Assumptions 1 and 2 are used to derive 2sCOPE-np, they are not strictly required (Web Appendix C.4).

3.2.1 Assumption 1

Assumption 1 means that the structural error E need not be normally distributed (Web Appendix C.3). The assumption holds as long as one of E and O is normal. [19] As the exogenous shock ϵ is not required to follow any specific distribution, structural error E, or the sum of ϵ and O, are allowed to be non-normal or heteroscedastic without affecting the consistency of the estimation of the augmented regression.

Assumption 1 is often reasonable in practice for two reasons. Firstly, a normal distribution assumption for E is commonly invoked in existing literature (e.g., Villas-Boas and Winer 1999, Yang et al. 2003). Secondly, the omitted variable O often represents a sum of multiple unmeasured confounders' combined effects and approaches a normal distribution according to the central limit theorem, even when the individual unmeasured confounders are nonnormally distributed. For instance, the omitted variable, an individual's ability in the return-to-education empirical example, reflects the aggregate influence of numerous genetic traits and environmental factors and is often deemed to be normally distributed. In addition, 2sCOPE-np is robust to nonnormality of O as shown by Proposition 3 and Web Appendix C.4.

[19]The normality assumption imposed on the omitted variable O can be further relaxed through Proposition 3, which allows O to follow a finite mixture of normal distributions rather than a single normal distribution

3.2.2 Assumption 2

Assumption 2 links the nonparametric model for endogenous regressors with the structural unknowns. It implies that, marginally, regressor-error dependence need not follow GC or any specific copula: as shown previously, 2sCOPE-np can be derived assuming a conditional GC dependence structure for (P, O) given W, while leaving the dependence structure for (P, E) given W unrestricted. Therefore, the conditional copula model permits both GC and non-GC types of regressor-error dependence unconditionally, as will be further demonstrated in the *Simulation Studies* section.

Assumption 2 is plausible in practice because, as noted previously, the joint GC model is widely regarded as flexible in capturing multivariate dependence (Danaher and Smith 2011), and the conditional GC model is more general than the joint GC model [20]. Assumption 2 is also not strictly required; even conditional on W, (P, O) is not required to follow a GC dependence structure for the proposed method to work (Web Appendix C.4). [21]

In practice, we recommend assessing the plausibility of Assumption 2 by (1) considering the source and nature of endogeneity (e.g., knowledge about the omitted variables, the distribution of P, and the links between them), and (2) applying diagnostic tools to detect potential model misspecification (e.g., checking for large standard-error inflation as a warning sign for misspecified copula models).

3.2.3 Comparison of 2sCOPE-np and linear additive endogeneity model

As noted previously, the conditional copulas flexibly model endogeneity independent of endogenous regressors' distributions, making them broadly applicable. In contrast, linear endogeneity models assume additive separability and a linear dependence between an endogenous regressor and the structural error. Such a representation can be internally in-

[20] Section 4.4 includes an example where the data follows joint GC but not conditional GC.

[21] Appendix C.4 shows that the model is often robust to symmetric departures in the copula structure, such as following a t-copula instead of GC.

consistent and is generally unable to capture complex multivariate dependence structures (Danaher and Smith 2011).

To be more specific, consider the linear endogeneity model studied in the prior literature (Haschka 2022, Lewbel et al. 2024, Qian and Xie 2024, Dost and Haschka 2025):

$$P = \rho O + V, \tag{17}$$

where the omitted variable O affects both the structural error and the endogenous regressor additively and generally does not follow a GC model with P. This model assumes (1) a linear relationship between O and P (linear endogeneity), and (2) V is independent of O (additive separability).

Condition (1) can be violated when O affects P nonlinearly. For instance, the linear relationship is known to not hold when P is bounded while O is unbounded (e.g., normally distributed). In contrast, the conditional copula model is flexible to capture nonlinear relationships and model bounded P.

Condition (2) is notably much stronger than no correlations between V and O (Lewbel et al. 2024). The additive separability implies that O can only affect the mean of P, but not higher moments of P, which can be a restrictive assumption. For example, this excludes the possibility that the ability (O) affects the variance of years of education (P). Our proposed model in Equation 3 relaxes the additive separability assumption.

Lewbel et al. (2024) shows that the linear endogeneity model is not point identified when O is normally distributed. [22] Consequently, when the underlying structural model itself is not point identified, copula-based methods should not be expected to correct for endogeneity bias in this case. As shown in Haschka (2022), Qian and Xie (2024), and Web Appendix C.2, applying copula correction to data arising from such linear endogeneity models is found to yield diffuse estimates with huge standard errors, reflecting the fact

[22] The linear endogeneity model is generally not identifiable without external IVs or meeting certain higher-moment conditions on unobservables O and V (Lewbel et al. 2024) that can be difficult to verify.

that the underlying linear endogeneity model is not point identified.

In practice, we recommend checking the distribution of P. In many empirical applications, linear endogeneity models can be ruled out when the endogenous regressor P is bounded, truncated, or discrete (Note 11) while O is unbounded. Dependence of higher-order moments of P on O likewise rules out the linear endogeneity model discussed above. Furthermore, checking for large standard-error inflation is a recommended diagnostic when applying copula correction, because grossly mis-specifying regressor-error dependence can lead to weak or non-identifications with huge standard errors (often $>$ 8-10 times those of the uncorrected estimates). An inflation of 6 is suggested as a cutoff value for dependence model misspecification when other model assumptions are met (Qian et al. 2026).

3.2.4 Assumption 3

Assumption 3 ensures full column rank of the augmented regressor matrix $(1, P, W, C_p)$ for model identification, preventing perfect multicollinearity. In the case when C_p is exactly a linear combination of $(1, P, W)$, the augmented regression fails to be identifiable. Because such perfect or severe collinearity causes significantly inflated standard errors (typically $>$ 6 times the OLS standard errors), we recommend comparing the standard errors of the copula-corrected estimates to those of the uncorrected estimates: an inflation of >6 times in standard error is a red flag for potential non-identification. Conceptually, this parallels Shea's measure for detecting multicollinearity with weak instruments, which compares the variances of IV estimates with those of OLS estimates (Wooldridge 2010, p.110).

4 Simulation Studies

This section presents main simulation studies to demonstrate the performance of the proposed method under different data endogeneity settings. We compare the proposed method with existing copula control function methods simulating data from different data generating processes (DGP) as outlined in Table 4.

The first three simulation studies assess the performance of 2sCOPE-np under DGP considered in the existing copula control function literature and under a more general setting involving continuous endogenous regressors where current copula methods fail. Two additional studies investigate the ability of 2sCOPE-np to correct for endogeneity when the endogenous regressor is count-valued or binary with limited support, a class of problems for which existing copula control function methods are known to be inadequate.

Due to space limitations, the binary example is included in the Web Appendix, where we further conduct a series of simulation studies to assess the robustness of the proposed method under model misspecification. Specifically, we consider settings generated from a linear additive endogeneity model and scenarios that violate Assumptions 1 and 2, thereby examining the performance of the proposed method when its key identifying assumptions fail to hold. We also discuss a diagnostic tool for detecting violations of the assumptions in practice and simulations assessing the tradeoff between flexibility and efficiency of the method. Table 4 presents an overview of all simulation studies.

Table 4: Overview of Simulation Studies

Section	Data Generating Process
4.2	Joint gaussian copula data
4.3	Exogenous variable W affects the mean of endogenous P only
4.4	Neither joint gaussian copula nor mean-dependence
4.5	Count endogenous regressor with few unique values
Web Appendix	
C.1	Simultaneity, omitted variable, and reverse causality
C.2	Linear additive endogeneity model
C.3	Non-normal structure error and failing joint gaussian copula (*Double robustness property*)
C.4	Robustness to departures from Assumptions 1 and 2
C.5	Multiple endogenous regressors
C.6	Binary endogenous regressor

4.1 Benchmark Methods and Performance Metrics

In the simulation studies, we estimate the structural model in Equation 18 using simulated data and known parameter values to evaluate the accuracy of the estimation.

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i \tag{18}$$

We consider four different estimation methods: (1) OLS regression; (2) 2sCOPE-GC; (3) 2sCOPE-MD; and (4) the proposed 2sCOPE-np. We expect the OLS to yield biased estimates because it ignores the regressor endogeneity problem. The three copula-based endogeneity correction methods in (2)-(4) aim to correct for endogeneity bias; they involve two steps, in which the second step estimates the following augmented OLS regression

$$Y_i = \mu + \alpha P_i + \beta W_i + \gamma \mathtt{C}_{i,p} + \eta_i, \tag{19}$$

where the copula correction term $\mathtt{C}_{i,p}$ is estimated in the first step. These three methods differ in how $\mathtt{C}_{i,p}$ is estimated in the first step, as described in Web Appendix B.2.

We evaluate the performance of these methods using bias, empirical standard errors (SE), and root mean squared error (RMSE) of the structural model parameter estimates. Bias computes the differences between the true parameter values and the average parameter estimates of all simulated data samples. SE computes the standard deviations of these parameter estimates resulting from all simulated datasets. RMSE provides comprehensive measures of the overall performance of an estimator and is computed as $\sqrt{Bias^2 + SE^2}$. RMSE informs the expected absolute difference between the estimator and the corresponding true parameter values, considering both the bias and the variability of the estimator.

4.2 Case 1: The Joint Gaussian Copula Data

To show that 2sCOPE-np embeds Gaussian copula as a special case, we first generate data from a joint Gaussian copula as described below.

$$\begin{pmatrix} P_i^* \\ W_i^* \\ E_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix}\right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}\right), \quad (20)$$

$$E_i = G^{-1}(U_{E,i}) = G^{-1}(\Phi(E_i^*)) = \Phi^{-1}(\Phi(E_i^*)) = 1 \cdot E_i^*, \quad (21)$$

$$P_i = H^{-1}(U_{P,i}) = H^{-1}(\Phi(P_i^*)), \;\; W_i = L^{-1}(U_{W,i}) = L^{-1}(\Phi(W_i^*)). \quad (22)$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 1 + 1 \cdot P_i + (-1) \cdot W_i + E_i. \quad (23)$$

Endogeneity problem arises due to the correlation between E_i^* and P_i^* ($\rho_{pe} = 0.5$). The control variable W_i is exogenous (i.e., uncorrelated with E_i), but is correlated with the endogenous regressor P_i ($\rho_{pw} = 0.5$) and also affects the outcome Y ($\beta \neq 0$). Thus, W is an observed confounder to be controlled for in the outcome regression model. In the simulation, we set $H(P)$ as the CDF of a truncated standard normal with a lower bound of 0 and an upper bound of 2. This emulates a scenario of an endogenous regressor with a bounded range. We set $L(W)$ to be the CDF of the standard normal distribution.

We generate 1000 replicate datasets with a sample size of 1000 each using the above DGP, and apply the methods as described previously. Table 5 reports the results.

Table 5: Simulation Study 1: The Case of Joint Copula Data

		OLS			2sCOPE-MD			2sCOPE-GC			2sCOPE-np		
Parameters	True	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
μ	1	-0.908	0.052	0.909	-0.265	0.331	0.424	-0.013	0.197	0.197	0.006	0.220	0.220
α	1	**1.259**	**0.057**	**1.260**	**0.385**	**0.446**	**0.589**	**0.014**	**0.262**	**0.262**	**-0.010**	**0.296**	**0.296**
β	-1	-0.308	0.028	0.309	-0.068	0.121	0.139	-0.006	0.071	0.071	-0.010	0.074	0.074
σ_E	1	-0.170	0.018	0.171	-0.066	0.088	0.110	-0.005	0.068	0.068	0.020	0.088	0.090

Note: 2sCOPE-MD implements the method of Mayer and Wied (2026) and 2sCOPE-GC implements the method of Yang et al. (2025). Bias, SE and RMSE denote the mean, standard deviation and root mean squared error of parameter estimates across all 1,000 simulated samples.

As expected, OLS yields large biases for all parameter estimates (Table 5). The coefficient α for the endogenous regressor P has a large bias of 1.259, a magnitude exceeding 100%. Because of such large parameter estimation bias, the OLS estimates have the largest RMSE, even though they have the smallest SE values among all methods.

In comparison, all three copula correction methods outperform OLS. They eliminate or reduce the endogeneity bias of OLS, and have substantially smaller RMSEs. 2sCOPE-GC and 2sCOPE-np perform the best with no estimation bias and the smallest (and comparable) RMSEs. Although 2sCOPE-MD reduces bias relative to OLS, substantial bias remains (e.g., bias=0.385 for α in Table 5). This bias occurs because the DGP violates the mean-dependence-only assumption: W affects only the mean, not the higher moments of $P|W$. In addition, 2sCOPE-MD cannot accommodate the GC regressor-error dependence. 2sCOPE-MD has substantially greater SEs than 2sCOPE-GC and 2sCOPE-np, leading to the highest overall RMSE among 2sCOPE methods.

2sCOPE-np performs on par with the correct model 2sCOPE-GC without assuming the joint GC dependence structure on regressors. 2sCOPE-np shows negligible bias, with SEs and RMSEs closely matching those of the correctly specified 2sCOPE-GC model.

4.3 Case 2: W affects the Mean of P Only (Mean-Dependence).

In this simulation, we consider the following DGP based on the mean dependence model.

$$O_i \sim N(0,1), \qquad W_i \sim Gamma(3,2), \tag{24}$$

$$P_i = \gamma_0 + \gamma_1 W_i + V_i = W_i + F^{-1}_{Gamma(1,1)}(\Phi(O_i)), \tag{25}$$

$$E_i = O_i + \epsilon_i, \quad \epsilon_i \sim N(0,1) \tag{26}$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 1 + 1 \cdot P_i + (-1) \cdot W_i + O_i + \epsilon_i. \tag{27}$$

The error term $E_i = O_i + \epsilon_i$, where O_i and ϵ_i are the omitted variable and the exogenous shock, respectively. P_i is related to O_i according to Equation 25, and consequently is endogenous. After generating P_i from Equation 25, we divide P_i by its standard deviation; the standardized P_i (with unit standard deviation) then enters model in Equation 27.

The DGP corresponds to a mean-dependence-only model for $P_i|W_i$, because W only affects the mean of the conditional distribution $f(P|W)$ and is unrelated to the first-stage model error term V_i, according to Equation 25. Moreover, neither (P, E) nor (P, O) follow a GC dependence structure. We generate 1000 datasets, each with a sample size of 1000.

Table 6: Simulation Study 2: $P|W$ follows a Mean-Dependence-Only Model.

		OLS			2sCOPE-GC			2sCOPE-MD			2sCOPE-np		
Parameters	True	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
μ	1	-0.902	0.085	0.906	-0.166	0.159	0.229	-0.012	0.159	0.159	0.015	0.138	0.139
α	1	**1.194**	**0.054**	**1.195**	**0.221**	**0.170**	**0.279**	**0.031**	**0.110**	**0.114**	**0.033**	**0.122**	**0.126**
β	-1	-0.907	0.063	0.909	-0.219	0.137	0.258	-0.030	0.121	0.125	-0.042	0.101	0.109
σ_E	1.44	-0.350	0.027	0.351	-0.140	0.077	0.160	-0.038	0.062	0.073	0.017	0.074	0.076

See the note under Table 5.

Results reported in Table 6 show large endogeneity bias for the OLS estimates. All three 2sCOPE methods outperform OLS by reducing bias and achieving substantially smaller RMSE. Although 2sCOPE-GC reduces the bias, noticeable bias still remains. As data are generated from 2sCOPE-MD model (satisfying the mean-dependence-only condition), 2sCOPE-MD performs well, recovering the true parameter values and eliminating bias entirely. 2sCOPE-np performs comparably to 2sCOPE-MD, despite not making the mean-dependence-only assumption of 2sCOPE-MD.

4.4 Case 3: Neither Joint Gaussian Copula Nor Mean-Dependence

This simulation examines a DGP that follows neither GC nor the mean-dependence model.

$$W_i \sim \text{student-t}(3), \tag{28}$$

$$P_i|W_i \sim \text{truncated standard normal}(a = \min(0, -2W_i + 2), b = \max(2, -2W_i + 2)) \tag{29}$$

$$P_i^* = \Phi^{-1}(H_{TN(a,b)}(P|W)) \text{ where } H_{TN(a,b)} \text{ is the (truncated normal) CDF of } P|W \tag{30}$$

$$\begin{pmatrix} P_i^* \\ E_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right) \tag{31}$$

$$E_i = G^{-1}(\Phi(E_i^*)) = \Phi^{-1}(\Phi(E_i^*)) = 1 \cdot E_i^* \tag{32}$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 1 + 1 \cdot P_i + 2 \cdot W_i + E_i. \tag{33}$$

Conditional on W, P follows a truncated normal distribution, whose truncation points depend on W. This emulates a scenario in which exogenous conditions (e.g., weather or cost fluctuations) affect firms' price-setting bounds. (P_i^*, E_i^*) follow a bivariate normal in Equation 31; the non-zero correlation between them generates the endogeneity problem. The dependence structure of $(P, E)|W$ follows the conditional GC dependence structure satisfying 2sCOPE-np assumption, but neither (P, W) nor (P, E) follow a joint GC structure (see Web Appendix D.4 for proof), thus violating the assumption of 2sCOPE-GC. Higher moments of P depend on W, violating the mean-dependence-only assumption. We generate 1000 datasets as replicates, each with a sample size of 1000.

Table 7: Simulation Study 3: neither GC nor MD

		OLS			2sCOPE-GC			2sCOPE-MD			2sCOPE-np		
Parameters	True	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
μ	1	0.486	0.046	0.488	-0.136	0.133	0.190	-0.276	0.279	0.392	-0.022	0.097	0.099
α	1	**-0.492**	**0.036**	**0.493**	**0.143**	**0.131**	**0.194**	**0.286**	**0.283**	**0.403**	**0.030**	**0.093**	**0.098**
β	2	-0.122	0.026	0.125	0.041	0.040	0.057	0.071	0.076	0.104	-0.002	0.031	0.031
σ_E	1	-0.104	0.021	0.106	0.153	0.159	0.221	0.066	0.068	0.095	0.013	0.048	0.049

See the note under Table 5.

As expected, without endogeneity correction, OLS results in a large bias of 49.2% for the coefficient α (Table 7). With endogeneity correction, the other three methods reduce the bias, but only 2sCOPE-np fully eliminates the bias. 2sCOPE-GC is mis-specified for this DGP, resulting in a bias of 14.3%. The DGP also does not satisfy the mean-dependence assumption. Thus, 2sCOPE-MD has a bias of 28.6%. In comparison, 2sCOPE-np, which makes the fewest dependence assumptions, successfully reduces the bias to a negligible 3% and has the smallest RMSE of all four methods. 2sCOPE-np is also much more efficient than 2sCOPE-GC and 2sCOPE-MD, as shown by its smaller SEs of all coefficients.

4.5 Case 4: Count Endogenous Regressor with few unique values

In this and Web Appendix C.6, we present simulation studies where the focal endogenous regressor is count or binary with limited support. As noted in the *Introduction*, existing copula control function methods are known to suffer from inconsistent estimation for such endogenous regressors. The goal here is not to re-establish this known problem, but rather to examine how effective 2sCOPE-np is in addressing this known problem, relative to the other methods developed within the same realm of copula-based control functions. In this simulation, we consider the following DGP.

$$W_i \sim N(0,1), \qquad O_i \sim N(0,1) \tag{34}$$

$$P_i = F^{-1}_{\text{Poisson}(\exp(-2+W_i))}(\Phi(O_i)) \tag{35}$$

$$E_i = O_i + \epsilon_i, \epsilon_i \sim N(0,1) \tag{36}$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 0 - 1 \cdot P_i + 1 \cdot W_i + O_i + \epsilon_i. \tag{37}$$

The error term is $E_i = O_i + \epsilon_i$, where O_i is the omitted variable and ϵ_i is the exogenous shock. The endogenous variable P_i is a count variable following a Poisson distribution with mean $\exp(-2 + W_i)$ and assumes only a few unique values (for most simulated datasets, P_i does not exceed 5). In Equation 35, P_i is generated as the minimum count value whose conditional CDF given W_i exceeds $\Phi(O_i)$. The omitted variable O_i is correlated with P_i (Equation (35)) and also enters the equation for Y_i (Equation (37)), creating endogeneity. We generate 1000 replicate datasets of sample size of 2000 each using the DGP.

Table 8: Simulation Study 4: Count Endogenous Regressors with Limited Support

		OLS			2sCOPE-GC			2sCOPE-MD			2sCOPE-np		
Parameters	True	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
μ	1	-0.229	0.034	0.231	0.122	0.028	0.125	-0.087	0.044	0.097	-0.047	0.026	0.054
α	-1	**1.095**	**0.104**	**1.100**	**-0.512**	**0.131**	**0.528**	**0.448**	**0.216**	**0.497**	**-0.056**	**0.047**	**0.073**
β	1	-0.257	0.026	0.258	0.093	0.024	0.096	0.038	0.072	0.082	-0.054	0.025	0.060
σ_E	1.44	-0.116	0.016	0.117	0.125	0.055	0.137	-0.064	0.036	0.073	0.013	0.017	0.022

See the note under Table 5.

Results in Table 8 show large endogeneity bias for OLS, exceeding 100%. All three copula correction methods outperform OLS, reducing the bias and achieving substantially smaller RMSE compared to that of OLS. However, only 2sCOPE-np effectively eliminates the bias (Bias: -5.6%, RMSE: 7.3% Table 8). As expected, 2sCOPE-GC and 2sCOPE-MD perform poorly for count endogenous regressors with few unique values, resulting in estimation biases of 51.2% and 44.8%, respectively.

4.6 Conclusion from simulation studies

The simulation studies examine cases in which regressor-error dependence and regressor-regressor dependence follow either GC, MD, neither GC nor MD, as well as cases in which the endogenous regressors are noncontinuous with limited support such as count and binary.

Unlike existing copula control function methods, only 2sCOPE-np is capable of correcting endogeneity bias across all cases. These simulation studies demonstrate that 2sCOPE-np offers greater robustness, higher accuracy, and broader applicability compared to existing copula correction methods. The reason is that the proposed 2sCOPE-np is nonparametric in the first-stage model for endogenous regressors given exogenous regressors and permits broader types of regressor-error dependence. Moreover, the nonparametric modeling of the conditional CDF in the first step leverages the variation in the exogenous variables to bypass inverse mapping from marginal CDFs of the regressors, overcoming estimation bias caused by the nonuniqueness of such mappings and enabling more effective handling of noncontinuous endogenous regressors.

5 Empirical Applications

We illustrate 2sCOPE-np procedure in two empirical applications: estimating price elasticity, where the focal endogenous regressor price is continuous, and estimating return to education, where the focal endogenous regressor can be count or binary (Web Appendix

E.2). We discuss the motivation for using 2sCOPE-np, steps to conduct estimation and inference, how to check verifiable data requirements, how to interpret results in the empirical applications, and also how to use heterogeneous estimation. In addition to IV methods, we also benchmark the performance of 2sCOPE-np using OLS, 2sCOPE-GC and 2sCOPE-MD.

5.1 Example 1: Price Elasticity (cross-validate IV and copula correction).

Our first empirical application is to estimate price elasticity using weekly sales for the peanut butter category at a store in New York City, over 311 weeks from 2001 to 2006, using store sales data provided in the IRI Academic data set (Bronnenberg et al. 2008). The category price elasticity estimates are often of interest for packaged goods categories (Nijs et al. 2001, Park and Gupta 2012), for gasoline consumption (Li et al. 2014), etc. These category price elasticity estimates are key information for marketers and policymakers to design effective price strategies and policy interventions. The IRI data sets also contain information on which UPCs were featured in the store's weekly flyer (Bronnenberg et al. 2008). Following prior literature (Nijs et al. 2001, Park and Gupta 2012), we compute the category price (defined per pound) and feature intensity as market-share-weighted averages of UPC-level price and feature, respectively.

We are interested in estimating the following category sales demand model:

$$\ln \texttt{sales}_t = \mu + \alpha P_t + \beta' W_t + E_t, \tag{38}$$

where $t = 1, \cdots, 312$ is the week index, $\ln \texttt{sales}_t$ and P_t denote the logarithms of category sales volume and category price, respectively. The parameter α represents the category price elasticity. The vector W_t contains a set of control variables: two continuous variables, feature intensity and week, and three dummy variables, `Q2`, `Q3`, and `Q4` for the second, third, and fourth quarters.

Price is customarily considered endogenous in category demand models (Nijs et al. 2001, Li et al. 2014, Park and Gupta 2012) due to unobserved factors, such as retailer pricing strategies and number of shelf facings. If omitted from the model, these unobservables become part of the structural error and lead to regressor-error dependence. The feature variable is often treated as exogenous because decisions to feature products in store flyers are made well in advance of implementation. Thus, feature is unlikely to correlate with weekly unobserved factors (Sriram et al. 2007, Chintagunta 2002). Week is a control covariate to account for price trends over time.

The above considerations suggest caution in using OLS to estimate the sales demand model. Nonetheless, we conduct OLS analysis as the benchmark for comparison. A common approach to addressing price endogeneity is the instrumental variables approach. The challenge with the IV approach is that valid and good IVs are difficult to find in practical studies. In our context, the price of peanut butter at another store within the same market may serve as an IV (e.g., Park and Gupta 2012). Since prices across stores are typically correlated, likely due to similar wholesale prices (ensuring relevance), but unobserved product characteristics (such as retailer decisions on shelf facings and placement) are unlikely to influence wholesale prices (ensuring the exclusion restriction, ER), the IV approach is justified. However, the ER assumption is inherently untestable, and the strength of the IV may be insufficient. This highlights a key use case for copula correction: employing multiple methods, such as IV estimation and copula-based endogeneity correction, to cross-validate findings and enhance the robustness of causal inference.

Thus, we conduct both IV estimation via two-stage least squares (2SLS) and copula-based endogeneity methods to control for potential price endogeneity. It appears plausible to assume that the joint effect of omitted variables (retailer pricing strategies, shelf facings, and locations, etc.) is roughly normally distributed. Furthermore, because price is often bounded (Kocherlakota 2021), the linear endogeneity model is unlikely to be appropriate. After controlling for pre-planned feature activity and quarterly trends, the remaining dependence between price and the joint effect of contemporaneous omitted variables can plausibly be captured using a conditional copula model. [23]

2sCOPE estimates the following augmented regression model

$$\ln \texttt{sales}_t \quad = \quad \mu + \alpha P_t + \beta' W_t + \gamma C_{price,t} + \eta_t, \tag{39}$$

where $C_{price,t} = \Phi^{-1}(\widehat{F}_\phi(P_t|W_t))$ is the copula term to correct price endogeneity and $\widehat{F}_\phi(P_t|W_t)$ is the estimate of the conditional CDF of the price given exogenous regressors.

Figure 1 in Web Appendix E.1 shows the box plots of the price variables at different levels of exogenous regressors: Feature (feature intensity at two levels: Feature=0 vs. Feature>0), week (grouped into years) and quarters (Q1, Q2, Q3 and Q4). We observe that the exogenous regressors affect not only the mean, but also the variances and higher-order moments of the price distributions (e.g., skewness, kurtosis). To obtain accurate control functions, it is prudent to use nonparametric estimation of $F_\phi(P_t|W_t)$ for the conditional CDF to account for the complex data features occurring in real data.

To apply 2sCOPE-np, we first estimate the conditional CDF $\widehat{F}_\phi(P_t|W_t)$ using the nonparametric kernel estimates as implemented in the `npcdist()` in the R package `np`. The kernel conditional CDF estimator permits a mixture of continuous and non-continuous variables in regressors (P_t, W_t). We used the default settings that employ a generalized product kernel function (Hayfield and Racine 2008) with second-order Gaussian kernel functions for continuous variables and Aitchison and Aitken kernel functions for categorical variables (Q2, Q3, and Q4). The kernel conditional CDF estimation uses the following cross-validation selected bandwidth: 0.012 for logPrice, 0.084 for feature intensity, 33.734 for week, 0.128 for Q2, 0.202 for Q3, and 0.499 for Q4. Figure 2 in Web Appendix E.1 presents the CDF estimates of logPrice conditional on feature intensity only (the top one)

[23] An example of pricing equation from our model is $G^{-1}(Price) = \ln(Price - L_p) = H(W, \gamma) + \xi_p$, where L_p is the lower bound of price, $\xi_p = \rho O + R$ with O denoting omitted variables (e.g., unobserved product characteristics or demand shocks), and R denoting supply shocks. Cases 1 to 3 in the Simulation Studies section provide other examples of pricing equations from our model. 2sCOPE-np does not require knowing the exact form of the G function.

and conditional on week only (the bottom one). We observe that as the feature intensity increases, there is a marked shift to lower price levels. To obtain the copula term, we perform copula transformation of the conditional CDF estimates; no copula transformations of individual regressors are needed. To verify that the augmented regression model in Equation 39 is of full rank to achieve model identification, we will inspect the inflation of standard errors of copula corrected estimates relative to those for the uncorrected estimates.

Table 9 reports estimation results from OLS, 2SLS with IV, 2sCOPE-np, 2sCOPE-GC and 2sCOPE-MD. The OLS estimation yields the price elasticity estimate with the smallest absolute effect size (Est. = -1.82, SE=0.12, $p = 0.000$), which is substantially smaller than both the 2SLS estimate (Est. = -2.13, SE=0.15, $p = 0.000$), 2sCOPE-np estimate (Est. = -2.30, SE=0.27, $p = 0.000$), 2sCOPE-GC estimate (Est. = -2.43, SE= 0.28, $p = 0.000$), and 2sCOPE-MD estimate (Est. = -2.86, SE=0.22, $p = 0.000$). We note that the OLS estimate is more than two standard errors away from 2SLS and the three 2sCOPE estimates.

Table 9: Estimation Results for Price Elasticity in Store Sales Data

	OLS	2SLS	2sCOPE-np	2sCOPE-GC	2sCOPE-MD
Intercept	6.87	7.06	7.15	7.23	7.50
	(0.08)	(0.09)	(0.16)	(0.17)	(0.13)
Price	**-1.82**	**-2.13**	**-2.30**	**-2.43**	**-2.86**
	(0.12)	**(0.15)**	**(0.27)**	**(0.28)**	**(0.22)**
Feature	0.26	0.13	0.07	0.02	-0.18
	(0.06)	(0.08)	(0.10)	(0.12)	(0.09)
Week	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Q_2	-0.06	-0.06	-0.06	-0.07	-0.07
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Q_3	-0.08	-0.09	-0.09	-0.10	-0.14
	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)
Q_4	-0.05	-0.05	-0.04	-0.04	-0.04
	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
C_{price}			**0.07**	**0.10**	**0.10**
			(0.03)	**(0.02)**	**(0.01)**
Wu–Hausman test	**H = 11.45, p = 0.0004**				

Note: Table presents coefficient estimates with bootstrapped standard errors reported in parentheses. The 2sCOPE-np price elasticity estimate is within about one standard error (0.15) of the 2SLS estimate, while those from the other methods are all more than two standard errors away from the 2SLS price estimate.

In comparison, 2sCOPE-np's price estimate is very close to 2SLS's estimate, cross-validating each other. Among the three copula methods, 2sCOPE-np makes the fewest assumptions, and yields the price estimate closest to that from 2SLS. 2sCOPE-GC price estimate deviates more, differing from 2SLS by roughly the same margin as OLS does. 2sCOPE-MD shows the largest deviation—over twice the OLS–2SLS gap—likely due to the violation of its mean-dependence assumption illustrated in Figure 1. This highlights the value of robustness of 2sCOPE-np and the importance of its ability to guard against bias from misspecified dependence models.

All three 2sCOPE methods have significantly positive coefficients for copula terms, agreeing on the presence of endogeneity bias. In particular, testing for the copula term C_{Price} in 2sCOPE-np gives a statistically significant result (Est. $= 0.074$, SE $= 0.025$, $p = 0.003$), indicating the presence of endogeneity. This is consistent with the result of the Wu-Hausman test ($H = 11.45$, $p = 0.0004$) from 2SLS, which also suggests endogeneity. We note that the endogeneity tests in copula methods hinge on the adequacy of copula to capture regressor-error dependence, while the Wu-Hausman test relies on the IV validity.

2SLS price estimate has a smaller standard error (0.15) compared with that of 2sCOPE-np (0.27), 2sCOPE-GC (0.28), and 2sCOPE-MD (0.22), because the IV is strongly correlated with the endogenous regressor (correlation between logPrice in stores is 0.90). In datasets with weak IVs, 2SLS might have much greater SEs than copula methods. The standard error of 2sCOPE-np corrected price estimate is approximately 2.3 (=0.270/0.117) times that of the OLS price estimate, suggesting no weak identification issues.

5.2 Example 2: Return to Education (the case of weak IV).

Our second empirical application is concerned with estimating the return to education on earning - a central question in labor economics. Firms, educational institutions, and governments are often interested in whether investments in labor force education can substantially increase earnings. Quantifying the return to education is important for educational institutes to attract prospective students and for policymakers to design appropriate education funding policies. The interest might be in estimating the return to one more year of education, in which case the education variable is a count variable. Alternatively, one may be interested in whether educational milestones, such as postsecondary education, would increase earnings (Web Appendix E.2); here, postsecondary education is a binary variable.

The decision to acquire additional education or pursue postsecondary studies is widely recognized as endogenous. Educational choices are often correlated with unobserved individual characteristics that also influence wages, such as motivation and innate ability.[24] These latent factors may affect educational attainment in complex and potentially nonlinear ways.[25] Because such variables are typically unobserved by researchers, their omission can induce endogeneity through omitted variable bias.

If educational attainment serves as a selection mechanism that disproportionately admits or retains individuals with higher levels of motivation or ability, these individuals may earn higher wages regardless of their level of schooling. In such cases, OLS estimates may overstate the return to education. Conversely, measurement error in educational attainment can introduce attenuation bias. As discussed by Card (1999), measurement error may generate a negative correlation between schooling and the structural error in the earnings equation, leading to an underestimation of the return to education. Because endogeneity may arise from multiple sources that operate in different directions, failing to account for it can result in substantially biased estimates. Correcting for endogeneity is therefore essential for obtaining credible estimates of the causal effect of education on earnings.

We use the well-known dataset of the 3010 men from the US National Longitudinal Survey of Young Men 1976 analyzed in Card (1999), Ebbes et al. (2005), etc., to study the

[24]Individual ability is shaped by a combination of genetic and environmental factors. When measured across a large representative population, it is reasonable to approximate the distribution of ability as normal due to the aggregation of many independent influences.

[25]For example, advancement in education is often determined by whether an individual's ability, aptitude, or test scores exceed certain thresholds required for admission, graduation, or advancement; thus, the relationship between education and the underlying (omitted) variables is non-linear.

return to education and estimate the following model:

$$\ln \texttt{wage}_t = \mu + \alpha Education_t + \beta' W_t + E_t, \tag{40}$$

where $t = 1, \cdots, 3010$ is the index of the individual worker; $\ln \texttt{wage}_t$ denotes the logarithms of the wage. $Education_t$ denotes the education variable. It can be a count variable denoting the number of years of education as discussed in this section or a binary indicator of postsecondary education as in Web Appendix E.2. The parameter α represents the return to education, which is of primary interest here. The vector W_t contains a rich set of control variables: five continuous variables: experience, experience squared, mother's education in years (motheduc), father's education in years (fatheduc), and the interaction between father and mother's education, and several categorical variables: whether the respondent is African American, whether father/mother's education is missing, family structure at 14 years old (momdad14, sinmom14, step14), and regional dummies for nine regions, and regional variables such as whether the region is in the South in 1966 and at the time of the survey. The full encoding of variable names can be found in Web Appendix E.3.

To correct for potential endogeneity bias, we consider both IV and IV-free methods: 2sCOPE-GC, 2sCOPE-MD, and 2sCOPE-np (including 2sCOPE-np with heterogeneous samples). For the IV approach, we apply the commonly used IV, proximity to college, for education (Card 1999). Being close to college affects one's chance of attending college (relevance condition) but does not directly affect one's earnings (exclusion restriction). For the IV-free methods, we follow the same process illustrated in the price elasticity example, including bootstrapping the error for all 2sCOPE methods and checking data requirements and model identification for 2sCOPE-np. We verify the data requirement of no collinearity between the copula terms and the original regressors. The inflation of standard errors of 2sCOPE-np corrected estimates relative to the corresponding OLS estimates is well below the threshold value of 6 (Tables 10 and 17), indicating no signs of weak model identification.

For 2sCOPE-np implementation with heterogeneous samples, we consider a setting where the effect of the omitted variable on educational attainment may vary across regions due to differences in local educational policies, demographic differences, etc. [26] Thus, the dependence between the omitted variable and educational attainment may vary across regions, possibly violating the assumption of a common endogeneity mechanism in the pooled sample. This motivates a heterogeneous sample implementation of 2sCOPE-np that allows the endogeneity to differ across regions. We operationalize our control functions in a way such that the coefficient of the first control function denotes the coefficient for the reference region, and all other coefficients of the control functions represent the regional differences between the underlying endogeneity and the reference region.

As reported in Table 10, among all methods, 2SLS estimates yield the highest return to education (Est.=0.1347, SE=0.05), compared with OLS (Est.= 0.0723, SE= 0.003), 2sCOPE-np (Est.=0.0697, SE=0.01), 2sCOPE-np-HET (EST.=0.0786, SE=0.00), 2sCOPE-GC (Est.=0.05, SE=0.01), and 2sCOPE-MD (Est.=0.04, SE=0.01)[27]. At first glance, the 2SLS and OLS results seem to suggest that OLS underestimates the return to education. However, the standard error of the 2SLS estimate is so large that, in fact, its confidence interval ([0.03,0.23]) covers that of the OLS ([0.06,0.08]). The Wu-Hausman test fails to reject the null hypothesis that the education variable is exogenous (H = 1.482, p-value= 0.224) and finds no evidence supporting the presence of endogeneity bias for the OLS estimate. Although the partial F test on the strength of IV rejects the null hypothesis (partial F= 15.383, p-value= 0.000), the partial correlation between the IV and education is only 0.066, and the IV is likely weak. The weak IV may explain why the confidence interval of the 2SLS estimate is too wide to be informative.

In contrast, 2SCOPE-np estimates that the return to log-earnings is 0.065 (95% CI: 0.05, 0.08), only slightly lower than the OLS estimate, and is much more precise than

[26] For example, regional differences in teacher quality and educational resources, admission requirements, and demographics may affect how unknowns such as ability and motivation influence schooling decisions.

[27] The 2SLS and OLS estimation results are almost the same as those reported in Card (1999), validating our results.

Table 10: Estimation Results for Count Education Variable

	OLS	2SLS	2sCOPE-np	2sCOPE-np-HET	2sCOPE-GC	2sCOPE-MD
(Intercept)	4.64	3.79	4.68	4.56	5.02	5.04
	(0.08)	(0.74)	(0.13)	(0.08)	(0.14)	(0.19)
educ	**0.07**	**0.13**	**0.07**	**0.08**	**0.05**	**0.04**
	(0.00)	**(0.05)**	**(0.01)**	**(0.00)**	**(0.01)**	**(0.01)**
exper	0.09	0.11	0.08	0.08	0.07	0.08
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
expersq	0.00	0.00	0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
black	-0.19	-0.16	-0.19	-0.19	-0.20	-0.20
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
fatheduc	0.00	0.00	0.00	0.00	0.01	0.01
	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)
motheduc	0.01	0.00	0.01	0.01	0.01	0.01
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
fatheduc × motheduc	0.00	0.00	0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Missing: fatheduc	0.01	-0.07	0.01	0.00	0.04	0.05
	(0.04)	(0.08)	(0.03)	(0.03)	(0.04)	(0.04)
Missing: motheduc	0.11	0.05	0.11	0.10	0.14	0.14
	(0.04)	(0.07)	(0.04)	(0.04)	(0.04)	(0.04)
momdad14	0.07	0.02	0.07	0.06	0.08	0.09
	(0.04)	(0.05)	(0.03)	(0.03)	(0.03)	(0.03)
sinmom14	0.04	0.04	0.04	0.04	0.04	0.04
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
step14	0.00	0.00	0.00	-0.00	-0.00	0.00
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
C_{common}			0.01		0.06	0.06
			(0.02)		(0.02)	(0.03)
C_{ref}				-0.01		
				(0.08)		
C_{21}				0.06		
				(0.09)		
C_{31}				-0.12		
				(0.08)		
C_{41}				-0.17		
				(0.10)		
C_{51}				0.01		
				(0.08)		
C_{61}				-0.05		
				(0.09)		
C_{71}				0.03		
				(0.09)		
C_{81}				-0.04		
				(0.13)		
C_{91}				0.01		
				(0.10)		
Wu–Hausman test		**H = 1.48, p = 0.22**				

Note: Dummy variables related to location (South, SMSA, region indicators, and SMSA66) are omitted from display. Standard errors are reported in parentheses and are estimated using bootstrap methods. In the heterogeneous specification (2sCOPE-np-HET), C_{ref} estimates the control-function coefficient for the reference region (Region 1), while coefficients $C_{21}, \ldots, C_{91}$ represent deviations from the reference region and therefore test whether the endogeneity differs across regions.

the 2SLS estimate (95% CI: 0.03, 0.23). The coefficient of the control function is positive but insignificant (Est.=0.01, SE=0.02, Table 10). Testing the coefficient of the control function yields a statistically insignificant result, consistent with the null result from the Wu-Hausman endogeneity test using 2SLS. This suggests that the biases from different sources may have offset each other. Alternatively, the set of control variables included in the model may have already significantly reduced endogeneity. In particular, parental education can be important confounders (especially mother's education, as shown by the statistically significant coefficient for the control variable motheduc in Table 10)[28].

The result from the heterogeneous-sample implementation of 2sCOPE-np with the nine geographic regions of the dataset (2sCOPE-np-HET) suggests that there is no statistically significant regional difference, as can be seen from the statistically insignificant control function coefficients ($C_{2ref}, \cdots, C_{9ref}$) at the 5% level. The estimated return to education remains close to that obtained from the homogeneous implementation, with overlapping confidence intervals (2sCOPE-np: Est. = 0.0697, SE = 0.01; 2sCOPE-np-HET: Est. = 0.0786, SE = 0.003).The coefficient of the control function for the reference region yields a statistically insignificant result, consistent with the homogeneous 2sCOPE-np specification.

The other two IV-free methods, 2sCOPE-GC and 2sCOPE-MD, theoretically cannot handle count regressors with limited support. They produce smaller estimates of the return to education than 2sCOPE-np by roughly two and three standard errors for 2sCOPE-GC and 2sCOPE-MD, respectively. While both 2SLS and 2sCOPE-np conclude no endogeneity, 2sCOPE-GC and 2sCOPE-MD yield positive and significant control function coefficients, suggesting a positive endogeneity bias in the OLS estimate, which may result from the inappropriateness of using the two methods for count regressors with limited support.

[28] We also considered the control covariate specification that removes the family education-related covariates, similar to Ebbes et al. (2005) which uses the latent IV (LIV) method. In this specification, 2sCOPE-np gives virtually the same estimates as the LIV estimates, further validating our proposed method.

6 Conclusion

Observational studies using historical data on markets, organizations, and individuals are invaluable for studying real-world causal relationships when experiments are either impractical or impossible to conduct. They offer insights that are generalizable, ethical, and cost-effective, making them essential tools in management and many other fields of research. However, their limitations for causal inference must be accounted for. One ubiquitous challenge in many nonexperimental casual studies is handling endogenous regressors.

To overcome the endogeneity problem, we propose a nonparametric two-stage copula-based endogeneity correction method (2sCOPE-np). 2sCOPE-np does not require IVs and can address endogeneity in various situations, including when good IVs are not available, when available IVs are weak or have questionable validity; in these cases, 2sCOPE-np can serve as a primary method or as a secondary method for robustness checking. One can also include 2sCOPE-np when multiple endogenous regressors in a structural model require combining different methods to address endogeneity.

We develop 2sCOPE-np under a new conditional copula endogeneity model that permits non-normal error distributions, both GC and non-GC types of regressor-error (or regressor-confounder) dependence, and nonparametric modeling of the first-stage auxiliary model for endogenous regressors. Thus, it generalizes and unifies all existing copula control function methods, while minimizing the assumptions in the first-stage model. We show that copula correction can be made agnostic to the functions determining the values of endogenous regressors. Operationally, this is achieved by using nonparametric conditional CDF estimates to construct control functions.

By replacing marginal CDFs of regressors with conditional CDFs of endogenous regressors given exogenous regressors in the copula term, 2sCOPE-np can overcome plateaus in discrete CDFs and achieve identification of noncontinuous endogenous regressors through relevant exogenous variables. This extends copula control function methods to a wide range of applications with noncontinuous endogenous regressors, an important feature that existing copula control function methods lack. Evaluation using simulation studies demonstrates significant improvement in endogeneity correction compared with existing methods.

We apply 2sCOPE-np to estimate price elasticity in the peanut butter category, and to estimate the return to education. The analyses show that IV-free 2sCOPE-np and 2SLS with IV largely agree on the test for endogeneity bias in the OLS estimate, but 2sCOPE-np produces more precise estimates of the return to education than 2SLS with weak IV. Compared with existing copula correction methods, 2sCOPE-np also demonstrates the value of its robustness, the practical importance of its ability to handle noncontinuous endogenous regressors, and to guard against bias from auxiliary model mis-specifications.

2sCOPE-np employs nonparametric kernel conditional CDF estimation in the first stage for endogenous regressors, making it computationally more demanding than existing copula-based control function methods. However, the available software programs, including our codes, and packages for nonparametric kernel conditional CDF estimation should mitigate the computational concerns. We also suggest simple strategies to facilitate its practical implementation (Section 3.1). While 2sCOPE-np uses a nonparametric auxiliary model for P and consequently is agnostic to the function determining the values of the endogenous regressor, model estimation and tests of endogeneity rely on the sufficiency of the conditional GC copula that links the model to structural unknowns. However, 2sCOPE-np demonstrates reasonable robustness to departures from the assumption. We also provide ways to assess the plausibility of the conditional copula model and diagnostics to detect severely misspecified dependence structures (Section 3.2).

References

Atefi Y, Ahearne M, Maxham III JG, Donavan DT, Carlson BD (2018) Does selective sales force training work? *Journal of Marketing Research* 55(5):722–737.

Becker JM, Proksch D, Ringle CM (2021) Revisiting gaussian copulas to handle endogenous regressors. *Journal of the Academy of Marketing Science* 50(1):1–21.

Breitung J, Mayer A, Wied D (2024) Asymptotic properties of endogeneity corrections using nonlinear transformations. *The Econometrics Journal* 27(3):362–383.

Bronnenberg BJ, Kruger MW, Mela CF (2008) Database paper - the iri marketing data set. *Marketing Science* 27(4):745–748.

Card D (1999) The causal effect of education on earnings. *Handbook of labor economics* 3:1801–1863.

Chintagunta PK (2002) Investigating category pricing behavior at a retail chain. *Journal of Marketing Research* 39(2):141–154.

Christopoulos D, McAdam P, Tzavalis E (2021) Dealing with endogeneity in threshold models using copulas. *Journal of Business & Economic Statistics* 39(1):166–178.

Danaher PJ, Smith M (2011) Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Science* 30(1):4–21.

Dost F, Haschka RE (2025) The gaussian copula control function method does not help against traditional omitted variable bias. *Available at SSRN 5285127* .

Ebbes P, Wedel M, Böckenholt U (2009) Frugal iv alternatives to identify the parameter for an endogenous regressor. *Journal of Applied Econometrics* 24(3):446–468.

Ebbes P, Wedel M, Böckenholt U, Steerneman T (2005) Solving and testing for regressor-error (in)dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics* 3(4):365–392.

Eckert C, Hohberger J (2023) Addressing endogeneity without instrumental variables: An evaluation of the gaussian copula approach for management research. *Journal of Management* 49(4):1460–1495.

Haschka RE (2022) Handling endogenous regressors using copulas: A generalization to linear panel models with fixed effects and correlated regressors. *Journal of Marketing Research* 59(4):860–881.

Hayfield T, Racine JS (2008) Nonparametric econometrics: The np package. *Journal of Statistical Software* 27:1–32.

Horowitz JL (1996) Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica* 103–137.

Joe H (2015) *Dependence Modeling with Copulas* (Boca Raton, FL: CRC Press).

Kocherlakota NR (2021) Bounds on price-setting. *Theoretical Economics* 16(3):979–1015, URL `http://dx.doi.org/https://doi.org/10.3982/TE4367`.

Lewbel A (1997) Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and r&d. *Econometrica* 65:1201–1214.

Lewbel A, Schennach SM, Zhang L (2024) Identification of a triangular two equation system without instruments. *Journal of Business & Economic Statistics* 42(1):14–25.

Li Q, Racine JS (2008) Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics* 26(4):423–434.

Li S, Linn J, Muehlegger E (2014) Gasoline taxes and consumer behavior. *American Economic Journal: Economic Policy* 6(4):302–342.

Liengaard BD, Becker JM, Bennedsen M, Heiler P, Taylor LN, Ringle CM (2025) Dealing with regression models' endogeneity by means of an adjusted estimator for the gaussian copula approach. *Journal of the Academy of Marketing Science* 53(1):279–299.

Lin Y, Song Y (2025) Addressing endogeneity issues in a spatial autoregressive model using copulas. *Journal of Econometrics* 252:106106.

Mayer A, Wied D (2026) Endogeneity corrections in binary outcome models with nonlinear transformations: Identification and inference. *Oxford Bulletin of Economics and Statistics* 88(1):45–56.

Nijs VR, Dekimpe MG, Steenkamps JBE, Hanssens DM (2001) The category-demand effects of price promotions. *Marketing Science* 20(1):1–22.

Papies D, Ebbes P, Feit EM (2023) *Endogeneity and Causal Inference in Marketing*, 253–300, (World Scientific).

Papies D, Ebbes P, Van Heerde HJ (2017) *Addressing Endogeneity in Marketing Models*, 581–627, (Cham: Springer).

Park S, Gupta S (2012) Handling endogenous regressors by joint estimation using copulas. *Marketing Science* 31(4):567–586.

Petrin A, Train K (2010) A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research* 47(1):3–13.

Qian Y, Koschmann A, Xie H (2026) A practical guide to endogeneity correction using copulas. *Journal of Marketing* 90(4):27–51.

Qian Y, Xie H (2011) No customer left behind: A distribution-free bayesian approach to accounting for missing xs in marketing models. *Marketing Science* 30(4):717–736.

Qian Y, Xie H (2022) Simplifying bias correction for selective sampling: A unified distribution-free approach to handling endogenously selected samples. *Marketing Science* 41(2):336–360.

Qian Y, Xie H (2024) Correcting regressor-endogeneity bias via instrument-free joint estimation using semiparametric odds ratio models. *Journal of Marketing Research* 61(5):914–936.

Rigobon R (2003) Identification through heteroskedasticity. *Review of Economics and Statistics* 85:777–792.

Rutz OJ, Watson GF (2019) Endogeneity and marketing strategy research: An overview. *Journal of the Academy of Marketing Science* 47(3):479–498.

Sklar M (1959) Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris* 8:229–231.

Sriram S, Balachander S, Kalwani MU (2007) Monitoring the dynamics of brand equity using store-level data. *Journal of Marketing* 71(2):61–78.

Tran KC, Tsionas MG (2021) Efficient Semiparametric Copula Estimation of Regression Models with Endogeneity. *Econometric Reviews* 41(5):1–28.

Trivedi P, Zimmer D (2017) A note on identification of bivariate copulas for discrete count data. *Econometrics* 5(1):10.

Villas-Boas JM, Winer RS (1999) Endogeneity in brand choice models. *Management Science* 45(10):1324–1338.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT Press).

Yang F, Qian Y, Xie H (2025) Addressing endogeneity using a two-stage copula generated regressor approach. *Journal of Marketing Research* 62(4):601–623.

Yang L, Qian Y, Xie H (2024) Handling endogenous marketing mix regressors in correlated heterogeneous panels with copula augmented mean group estimation. *NBER Working Paper w33265, https://www.nber.org/papers/w33265* .

Yang S, Chen Y, Allenby GM (2003) Bayesian analysis of simultaneous demand and supply. *Quantitative marketing and economics* 1(3):251–275.

Web Appendix

Table of Contents

A Comparison between Methods

This section reviews the origins and recent advances in copula-based endogeneity correction methods and compares our proposed method with existing approaches, as summarized in Table 2 in the main text. The original copula-based endogeneity correction method was introduced by (Park and Gupta 2012). Since then, there have been numerous extensions that can be categorized into two groups: (1) two-stage control function methods and (2) likelihood-based methods. Our proposed approach falls under the control function form extension, unifying and generalizing all existing copula control function methods.

Two-stage control function methods

Yang et al. (2025) introduced the two-stage copula endogeneity (2sCOPE) correction procedure using copula control functions. They show that properly constructed residuals from a first-stage model for endogenous regressors given exogenous regressors are copula control functions. Their method can handle normally distributed endogenous regressors (not allowed in (Park and Gupta 2012)), provided at least one related exogenous regressor is continuous and nonnormally distributed. They also assume that (P, W, E) follow a joint Gaussian copula dependence model (we refer to the method as 2sCOPE-GC hereafter to emphasize the GC assumption), where P, W, E denote endogenous regressors, exogenous regressors and the error term in the structural model, respectively. Our proposed method relaxes the joint Gaussian copula assumption[29].

Mayer and Wied (2026) and Breitung et al. (2024) proposed using a mean regression auxiliary model for the endogenous regressor to estimate the conditional distribution of P

[29]Recent extensions of 2sCOPE-GC (Liengaard et al. 2025, Yang et al. 2024) further permit heterogeneous joint GC model for (P, W, E) and the copula correction terms to vary by exogenous discrete regressors. Our proposed 2sCOPE-np method is more general and can permit the conditional GC dependence (to be precisely defined in Equation 80) to vary across heterogenous subsamples (Proposition 3).

given the exogenous regressors W, formulated as $P = g(W) + V$. They use the copula transformation of the residual V as the control function. This procedure is fully embedded in our framework and imposes the following extra assumptions: (1) the correlation coefficient in our conditional copula endogeneity model is set at ± 1 [30] and (2) the conditional CDF of P given W, $F_\phi(P|W) = F_v(V)$, where $F_v(V)$ is the marginal CDF of V. The latter assumption implicitly requires V to be unrelated to W. We name this assumption mean-dependence-only, and refer to this method as 2sCOPE-MD hereafter. That is, W affects only the mean of $P|W$, not its higher moments (Table 2). This assumption is known to be violated for bounded or truncated variables, but may also be questionable for any continuous endogenous regressors (see Note 11). In comparison, our proposed method (and 2sCOPE-GC) does not require such a strong assumption. Although 2sCOPE-MD originated from copula correction, it generally does not permit a GC regressor-error dependence structure (Table 2). Our proposed method unifies and generalizes these two methods (2sCOPE-GC and 2sCOPE-MD), making copula correction more robust.

Likelihood-based methods

Unlike control function approaches, these methods specify a full likelihood function for endogeneity correction. Haschka (2022) proposed maximizing the likelihood of a joint GC model for all regressors and the structural error, and generalized Park and Gupta (2012) to panel models with fixed-effects. Lin and Song (2025) consider maximizing the joint likelihood of a t-copula model for (P, W, E) (all regressors and the structural error) in spatial models with endogenous regressors and endogenous spatial weights. Qian and Xie (2024) considered the joint conditional distribution $f_{\theta,\phi}(E, P|W)$, which was modeled using a flexible distribution-free semiparametric odds ratio endogeneity model (SORE, Table

[30]This correlation coefficient is defined right above Equation 8.

2). Their method permits copula and non-copula types of endogeneity, as well as endogenous regressors with insufficient support (e.g., binary endogenous regressors). Estimation was done in one step via a profile-likelihood estimation that concentrates the likelihood on the parameters of interest, and eliminates nuisance nonparametric parameters for the baseline distribution from the likelihood. Similar to SORE, 2sCOPE-np considers the joint conditional distribution $f_{\theta,\phi}(E, P|W)$, which avoids modeling W, increasing robustness and reducing the number of variables to be modeled. Furthermore, (P, E) is not required to follow any specific copula dependence structure unconditionally. Unlike SORE, the two-step 2sCOPE-np requires no joint estimation of the outcome and regressor models.

Comparison among methods

The proposed 2sCOPE-np method unifies and nests all other methods, except for SORE in Table 2, as special cases. As a control function approach, 2sCOPE-np imposes fewer distributional assumptions (e.g., on the structural error distribution), is less sensitive to model mis-specifications, and is computationally simpler and faster than the likelihood-based SORE method. Meanwhile, the SORE method can be more efficient (smaller standard errors) and permits capturing endogeneity via flexible odds ratio functions and using likelihood-based tests and model comparisons. Overall, SORE and 2sCOPE-np are complementary to each other and non-hierarchical (i.e., one is not nesting the other).

Unlike existing copula methods, the proposed 2sCOPE-np is nonparametric regarding the first-stage auxiliary model for endogenous regressors; thus it permits more flexible regressor-error dependence (GC or non-GC) and eliminates first-stage modeling of endogenous regressors (Table 2).

The proposed framework highlights that proper copula control functions rely solely on the conditional CDF of the endogenous regressor given the exogenous regressors. This conditional CDF fully captures the influence of exogenous variables and can be estimated directly using nonparametric kernel methods, eliminating the need to impose a parametric first-stage model[31]. Consequently, 2sCOPE-np nests all existing copula-based endogeneity correction methods as special cases, and has increased robustness and broader applicability compared to existing methods. This generalization and increased robustness are important, because the first-stage models for endogenous regressors are not of primary interest and are preferred to be left unspecified.

B Implementation Details

B.1 Nonparametric kernel estimation of $F_\theta(P|W)$ in 2sCOPE-np

Specifically, we use the Nadaraya-Watson (NW) kernel regression estimator to estimate the conditional CDF via the following locally weighted average (Li and Racine 2008):

$$\widehat{F}^a_\phi(p|w) \quad = \quad \frac{\sum_{i=1}^n I(P_i \leq p) K_h(W_i - w)}{\sum_{i=1}^n K_h(W_i - w)}. \tag{41}$$

where $I(P_i \leq p)$ is the indicator function for the event $P_i \leq p$; n denotes the number of observations; $K_h(W_i - w)$ is a weight function defined as:

$$K_h(W_i - w) \quad = \quad \frac{1}{h} k\left(\frac{W_i - w}{h}\right), \tag{42}$$

[31] As a result, there exists no first-stage residuals to compute in 2sCOPE-np.

where $k(\cdot)$ is a user-supplied smooth and symmetric kernel function and h is the bandwidth parameter. Popular kernel functions include the second-order Gaussian kernel: $k(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$, and the Epanechnikov kernel $k(u) = 0.75(1-u^2)I(|u| \leq 1)$. The kernel conditional CDF estimation can also handle discrete W (Li and Racine 2008) using the Aitchison and Aitken kernel function (Aitchison and Aitken 1976). When W contains multiple variables and/or mixed types of continuous and discrete variables, a generalized product kernel is used (Li and Racine 2008). Intuitively, the NW estimator uses the weighted sample averages of the event $I(P_i \leq p)$ as the CDF estimate at the point w; the kernel function measures the distance between each sample observation and the point w at which the CDF is being estimated, assigning greater weight to observations that are closer to w when computing the weighted average. The NW estimator estimates the conditional CDF without imposing distributional assumptions and without assuming a specific functional form for the relationship between W and P.

Another estimator that smooths the continuous endogenous regressor P is defined as

$$\widehat{F}^b_\phi(p|w) = \frac{\sum_{i=1}^n G((p-P_i)/h_0)K_h(W_i - w)}{\sum_{i=1}^n K_h(W_i - w)} \tag{43}$$

(Li and Racine 2008), where $G(\cdot)$ is the CDF function defined by $G(v) = \int_{-\infty}^{v} k(u)du$ from the density function $k(u)$, and h_0 is the bandwidth for smoothing the outcome Y.

The performance of these kernel conditional CDF estimates depends on the judicious choice of the bandwidth tuning parameter h. If the bandwidth is too small, the resulting conditional CDF estimate can be too noisy (overfitting) while a bandwidth that is too large can introduce bias into conditional CDF estimation (underfitting). Unlike the unconditional kernel density/CDF estimation, however, there exists no known rule-of-thumb choice of bandwidth. In practice, the optimal bandwidth value can be estimated using an automatic, data-driven cross-validation that minimizes the out-of-sample prediction error of conditional CDF estimation (Li et al. 2013). As such, the method builds upon the advances in nonparametric kernel CDF estimation methods, and similarly trades off flexibility with a higher computational burden and potentially lower efficiency (Web Appendix C.5).

B.2 2sCOPE-GC, 2sCOPE-MD and 2sCOPE-np details

In this section, we describe in detail the calculation of the copula-based correction term during simulation studies.

- 2sCOPE-GC. Under the joint GC model for (P, W, E), Yang et al. (2025) shows that

$$\mathtt{C}_{i,p} \quad = \quad P_i^* - \hat{\delta} W_i^*, \tag{44}$$

  where $P_i^* = \Phi^{-1}(\widehat{H}(P_i))$ and $W_i^* = \Phi^{-1}(\widehat{L}(W_i))$ are copula transformations of regressors P_i and W_i, respectively; $\widehat{H}(\cdot)$ and $\widehat{L}(\cdot)$ are the empirical CDF estimates of the marginal distributions of P and W, respectively; $\hat{\delta}$ is the coefficient estimated from the first-stage model that regresses P_i^* on W_i^*. Consequently, the control function $\mathtt{C}_{i,p}$ is the residual of the first-stage regression and captures P's endogenous part that is orthogonal to the exogenous regressor W. Under the flexible joint GC model, the first-stage regression in 2sCOPE-GC procedure makes no assumptions about the marginal distributions of regressors (and thus can accommodate bounded regressors automatically) [32], and permit

[32] In the results presented in the main text, we use the empirical CDFs to estimate the marginal CDFs of regressors in 2sCOPE-GC, and to estimate $F_v(V)$ in 2sCOPE-MD. We also estimated these marginal CDFs in 2sCOPE-GC and 2sCOPE-MD using kernel methods, and the conclusions remain the same. Therefore, the conclusions we draw from simulation studies are unaffected by the methods used to estimate marginal CDFs.

possibilities of both mean and high-order moments of $P_i|W_i$ to depend on W_i. We expect 2sCOPE-GC to work well, since it captures GC regressor-error and regressor-regressor dependence in the data.

- 2sCOPE-MD. Under mean dependence only for $f(P|W)$: $P = g(W) + V$, where $g(W) = \mathbb{E}(P|W)$ can be a nonparametric function,

$$\mathtt{C}_{i,p} = \Phi^{-1}(\widehat{F}_v(\widehat{V}_i)), \tag{45}$$

where the residual $\widehat{V}_i = P_i - \widehat{g}(W_i))$ and $F_v(V)$ can be estimated by a univariate empirical CDF estimate: $\widehat{F}_v(t) = \frac{1}{n+1}\sum_{i=1}^{n} I(V_i \leq t)$. $\widehat{F}_v(\cdot)$ is the empirical CDF of the residuals from this first stage regression. FollowingMayer and Wied (2026), we use the `gam()` function in the R package `mgcv` to estimate the mean function $g(\cdot)$ nonparametrically and then estimate the marginal distribution of the first-stage residual using the empirical CDF estimate $\widehat{F}_v(\cdot)$ (see Footnote 32). Consequently, the approach permits nonlinear mean functional relationships between P and W. Note that the approach implicitly assumes that the higher moments of the first-stage residual do not depend on W. When this assumption is violated, the distribution of the residuals will depend on W, and consequently, $F_v(P_i - g(W_i)) \neq F_\phi(P_i|W_i)$, break the mean-dependence-only assumption and introduce bias in the endogeneity correction. Although the violation of this assumption could occur for any type of variable, this assumption is known to not hold for variables with bounded ranges such as censoring or truncation (e.g., truncated normal here). In addition, 2sCOPE-MD cannot capture GC regressor-error dependence. We use this in simulation to evaluate the potential bias resulting from a misspecified

2sCOPE-MD model.

- 2sCOPE-np. This is the proposed approach that uses nonparametric kernel CDF estimation for $F_\phi(P|W)$. It makes no assumptions about the distributions for regressors, the functional relationships among regressors, or the condition that $P|W$ depends on W only through its mean function. We expect 2sCOPE-np to work well because it is flexible enough to capture the GC regressor-regressor and regressor-error dependence in the data. With 2sCOPE-np,

$$\mathtt{C}_{i,p} \quad = \quad \Phi^{-1}(\widehat{F}_\phi(P_i|W_i)), \tag{46}$$

where we use the function `npcdist()` in the R package `np` (Hayfield and Racine 2008) to obtain the kernel estimate of the conditional CDF: $\widehat{F}_\phi(P_i|W_i)$. The function implements the estimator in Equation 43 that smooths both P and W.

C Additional Simulations

C.1 Simultaneity, omitted variable, and reverse causality

Simultaneity arises when the endogenous regressor and the outcome are jointly determined through a system of structural relationships. For example, in the classical demand–supply model, price is jointly determined by consumer demand and firms' supply decisions. The simultaneous relationships will induce a correlation between the error term in the demand-side equation and the error term in the supply-side equation determining the endogenous regressor, thereby creating endogeneity (Yang et al. 2003), such as in the following commonly considered model:

$$Y = \beta P + \epsilon_y, \tag{47}$$

$$P = G(\xi_p), \tag{48}$$

$$\begin{pmatrix} \epsilon_y \\ \xi_p \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{12} & \sigma_2 \end{pmatrix} \right), \tag{49}$$

where simultaneity occurs as the non-zero covariance σ_{12} between demand-side error ϵ_y and supply side error ξ_p. For simplicity, we omit exogenous covariates but the idea extends to the case with W (i.e., Equations 1 and 3). The above model is conceptually similar to the simultaneous demand and supply model considered in Section 2.3 of Yang et al. (2003), which assumes a linear supply equation with instrumental variables. In contrast, copula correction does not require instrumental variables in the supply equation.

The above model implies a GC regressor-error dependence structure, which is general in that the regressor-error dependence structure for (P, ϵ_y) depends on the rank-order of raw data only and is invariant to strictly monotonic transformation of P. As long as there exists a strictly monotonic transformation of P (e.g, $\ln P$, $\sqrt{P}$, or $\ln P + P$) such that the above simultaneity model holds, then the GC model applied to the raw data will recover the regressor-error dependence without the prior knowledge of the correct transformation (i.e., the G function).

To illustrate, we simulate data from the following system of simultaneous equations[33], where the endogenous regressor P (e.g., price) affects the outcome variable Y (e.g., quantity) in one structural equation, while the outcome variable simultaneously affects the

[33]For simplicity, we omit the index i and assume that the intercept is 0 without loss of generality. One can always demean first.

endogenous regressor in another equation.

$$Y = \beta P + \epsilon_y, \tag{50}$$

$$\ln P = \gamma Y + \epsilon_p, \tag{51}$$

where $\beta = 1$ and $\gamma = -1$, (ϵ_y, ϵ_p) follow a bivariate normal distribution with variance of 1 and a correlation of 0.2. Equation 51 corresponds to a log-linear model for P to account for skewness or boundedness. Substituting Equation 50 into Equation 51 yields $\ln(P) - \gamma\beta P = \ln(P) + P = \gamma\epsilon_y + \epsilon_p$, which can be rewritten to be in the form of Equation 48 as $P = G(\xi_p)$ with $\xi_p = \gamma\epsilon_y + \epsilon_p$. The endogeneity is caused by the appearance of ϵ_y in both Y and P models, similar to the endogeneity due to omitted variables. An important difference is that the equation for P is not of a known functional form here, demanding flexible nonparametric procedures to handle it. That is, given ϵ_y and ϵ_p, the value of P is

$$P = G(\xi_p) = \{P : \ln(P) - \gamma\beta P - \gamma\epsilon_y - \epsilon_p = 0\} \tag{52}$$

whose functional form is unknown. After generating P from Equation 52, we then generate Y given P and ϵ_y according to Equation 50.

The results of copula correction and OLS estimation are reported below in Table 11. The simulation results indicate that copula correction can recover the coefficients of the endogenous variable with little bias despite the presence of simultaneity. In contrast, the OLS model exhibits severe endogeneity bias, yielding an estimate of 0.231 (S.E=0.041) for β (true value -1), even reversing the sign of the effect. OLS also produces a biased intercept estimate. The standard errors of the copula-based model estimates are within the recommended threshold.

Table 11: Estimation result from a simultaneity model

		OLS		Copula		
Parameters	True Value	Mean	SE	Mean	SE	ICON
intercept	0	-0.890	0.035	0.003	0.094	2.69
β	-1	0.231	0.041	-1.010	0.124	3.02

Mean and SE denote the mean and standard deviation of the parameter estimates across 1,000 simulated datasets, each of size 1,000. The ICON statistic calculates the inflation of standard errors of the copula corrected estimate relative to the corresponding OLS estimates.

Omitted variable. Consider the following system of equations, where O denotes the omitted variable that affects both the outcome Y and the endogenous variable P, causing regressor-error dependence (i.e., endogeneity due to omitting variable O from the model).

$$Y = \alpha_0 + \alpha_1 P + \alpha_2 W + O + \epsilon_y \qquad \text{(outcome equation)} \tag{53}$$

$$P = g(W, O, \epsilon_p) \qquad \text{(endogenous regressor equation)} \tag{54}$$

Note that Equation 54 is embedded in Equation 3 in our main text in the discussion of P models. In Equation 3, the ξ_p term denotes the endogenous part in P that is independent of W, but correlated with the outcome Y. In Equation 54, ξ_p is decomposed into two parts, the omitted variable O and the independent random disturbance term ϵ_p that is uncorrelated with O. Specifically, under the conditional copula model for $(P, O)|W$ and a strictly monotone G function, (ξ_p, O) follows a GC distribution, such that $\xi_p^* = \rho * O + \epsilon_p$ where ξ_p^* is the normal score transformation of ξ_p. Notably, the GC model on (ξ_p, O) captures dependence structures through the rank ordering of the data and is therefore invariant to strictly monotonic transformations of ξ_p. Consequently, the validity of the GC model does not depend on specifying the correct marginal distributions of ξ_p or identifying the appropriate transformations. As long as there exist strictly monotonic transformations of the regressors (e.g., logarithmic transformations) under which the GC model is valid, applying the GC model directly to (ξ_p, O) will correctly recover the underlying dependence structure. When ξ_p is normal, $\xi_p = \rho O + \epsilon_p$.

Thus Equation 54 defines the source of endogeneity as omitted variable bias due to the omitted (unobserved) O affecting both the outcome and the endogenous regressor. Our derivation of 2sCOPE-np in Section 2.3 focuses explicitly on the notion of bias due to the omitted variable O, for which the method is designed to handle.

Reverse causality refers to a feedback mechanism in which the outcome directly influences the endogenous regressor. It can be viewed as a special case of simultaneity, where the outcome and the endogenous regressor are jointly determined through a system of structural equations. The simultaneity model in Equations 50 and 51 captures this setting, with the second equation explicitly allowing the outcome to feed back into the determination of the endogenous regressor.

C.2 Linear additive endogeneity model

In this section, we consider the case of a linear additive endogeneity model, where the omitted variable has a linear relationship with the endogenous regressor. In the simulation below, we simulate data from a linear additive model where the endogenous regressor P_i has a linear relationship with the structural error E_i, generating the endogeneity problem. P_i is equal to the sum of two independent variables: E_i from a standard normal distribution and S_i from a skwed normal distribution [34], thus P_i is non-normally distributed.

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 1 + 1 \cdot P_i + 1 \cdot W_i + E_i, \tag{55}$$

$$P_i = E_i + S_i \text{ where} \tag{56}$$

$$E_i \sim N(0,1) \text{ and} \tag{57}$$

$$S_i \sim \text{ skew normal distribution}(\xi = 0, \omega = 2, \alpha_1 = 3). \tag{58}$$

$$W_i \sim \text{student-t}(3). \tag{59}$$

As noted in the main text, the linear endogeneity model is not point identified with a normally distributed E (Lewbel et al. 2024). In Table 12, we present the mean and standard deviations of the parameter estimates across 1000 simulations of size 1000. Consistent with theoretical predictions, copula-based models, including 2sCOPE-np, encounter model mis-identification issue and yield estimates with bias in the opposite direction to OLS and huge standard errors.

Because grossly mis-specifying regressor-error dependence can lead to weak or non-identifications with huge standard errors (often $>$ 8-10 times those of the uncorrected estimates), checking for large standard-error inflation via the ICON statistic that compares the standard error of the copula corrected estimate to that of the uncorrected estimate is a recommended diagnostic when applying copula correction (Qian et al. 2026), An inflation of 6 is suggested as a cutoff value for dependence model misspecification when other model assumptions are met (Qian et al. 2026). The ICON statistics for the parameter α show a

[34] The notation for skewed normal distribution are as follows: ξ is the location parameter, $\omega > 0$ is the scale parameter, and α_1 is a shape parameter capturing the degree of skewness. The pdf of the skew normal distribution is $f(S = s) = \frac{2}{\omega}\phi\left(\frac{s-\xi}{\omega}\right)\Phi\left(\alpha_1\left(\frac{s-\xi}{\omega}\right)\right)$.

large inflation of the standard error with the ICON value far exceeding the threshold of 6.

Table 12: Linear additive model simluation

		OLS			2sCOPE-GC			2sCOPE-MD			2sCOPE-np		
Parameters	True	Mean	SE	ICON	Mean	SE	ICON	Mean	SE	ICON	Mean	SE	ICON
μ	0	0.44	0.032	-	1.816	0.388	12	1.832	0.392	12	1.829	0.529	17
α	1	1.37	0.016	-	0.461	0.254	16	0.451	0.256	16	0.451	0.349	22
β	1	1	0.015	-	1.002	0.033	2	1.001	0.03	2	1.001	0.017	1
σ_E	1	0.793	0.018	-	1.703	0.369	21	1.718	0.372	21	1.731	0.499	28

Mean and SD are the mean and SD of the estimated parameters across 1000 simulations of size 1000. The ICON statistic calculates the inflation of standard errors of the correpsonding estimate relative to the corresponding OLS estimates.

We note here that the linear endogeneity model above cannot hold when P is bounded while E is unbounded. In many marketing and economics applications, P is bounded (see Note 11). For these reasons, despite the more familiarity of linear additive models, models with more flexible dependence structures, such as copula models, are developed and used frequently in empirical studies.

C.3 Double robustness property

Our proposed 2sCOPE-np method is derived under the assumptions that the structural error or the omitted effect follows a normal distribution (Assumption 1 Table 3) and that GC captures the joint distribution of (P, O) or (P, E) given W (Assumption 2 Table 3). The double robustness property means that for 2sCOPE-np to work as intended, the error term can be nonnormal and regressor-error needs not to follow a specific copula structure. We place no assumptions on ϵ, the exogenous part of the error term. Therefore, the structural error term, the sum of the two random variables ($E = O + \epsilon$), can be nonnormal and left unspecified. In this case, since no assumption is placed on the joint distribution of ϵ and

P, the regressor-error distribution (i.e., $P - E$) needs not to follow a GC or any specific copula dependence structure.

In the simulations below, we demonstrate the double robustness property of the proposed 2sCOPE-np method. We simulate data from the following data generation process, where (P_i, W_i, O_i) follows a Gaussian copula model, but (P_i, W_i, E_i) does not follow a Gaussian copula model and E_i is not normally distributed. The correlation between P_i and O_i is non-zero, generating the endogeneity problem.

In the simulation, P_i follows the Gamma (1,1) distribution and W_i follows the exponential distribution with rate parameter 1. O_i follows the standard normal distribution, but ϵ_i follows either a uniform [-0.5, 0.5] distribution or a lognormal(0, 1)-$e^{0.5}$ distribution as specified in the table.

$$\begin{pmatrix} P_i^* \\ W_i^* \\ O_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right), \quad (60)$$

$$O_i = (\Phi(O_i^*)) = \Phi^{-1}(\Phi(O_i^*)) = O_i^*, \quad (61)$$

$$P_i = F^{-1}(\Phi(P_i^*)) \quad (62)$$

$$W_i = G^{-1}(\Phi(W_i^*)). \quad (63)$$

$$\epsilon_i \sim H \text{ as specificied in Table 13} \quad (64)$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + O_i + \epsilon_i = 0 + 1 \cdot P_i + 2 \cdot W_i + O_i + \epsilon_i. \quad (65)$$

Simulation results reported in Table 13 show that although the error term E_i is non-

Table 13: Simulation results for OLS and 2sCOPE-np under different distributions of ϵ_i.

Distribution				OLS			2sCOPE-np		
ϵ_i	Skewness of E_i	Parameters	True	Bias	SE	RMSE	Bias	SE	RMSE
lognormal(0, 1)-$e^{0.5}$	4.62	μ	0	0.313	0.058	0.318	0.014	0.071	0.073
		α	1	-0.569	0.042	0.570	-0.020	0.090	0.092
		β	2	0.257	0.040	0.260	0.011	0.055	0.056
		σ_E	2.38	-0.069	0.138	0.155	-0.018	0.136	0.138
U(-0.5, 0.5)	0	μ	0	0.311	0.02	0.311	0.011	0.027	0.029
		α	1	-0.569	0.016	0.569	-0.019	0.032	0.038
		β	2	0.258	0.015	0.258	0.011	0.020	0.023
		σ_E	1.04	-0.132	0.009	0.132	-0.008	0.017	0.019

Note: Bias, SE and RMSE denote the mean, standard deviation, and root mean squared error of parameter estimates across all 1,000 simulated samples of size 5000.

normally distributed and does not follow GC or any specific copula structured jointly with P_i, 2sCOPE-np can recover the true parameter values and eliminate the endogeneity bias of OLS estimations, demonstrating the double robustness property of 2sCOPE-np.

C.4 Robustness to departure from Assumption 1 and 2

In our proposed model, we impose a normality assumption and a conditional Gaussian copula assumption (Assumptions 1 and 2 in Table 3) to derive 2sCOPE-np estimator. In practice, however, the dependence structure may deviate from the Gaussian copula, and the omitted variable may not follow a normal distribution. In this section, we examine the robustness of our approach in reducing bias when the normality assumption is violated and when the data are not generated from a Gaussian copula. Our results demonstrate that Assumptions 1 and 2 are not strictly required for satisfactory performance.

To relax Assumption 1 and Assumption 2 at the same time, we simluate data from the a t-copula and assume that the structural error follows the a uniform distribution and as defined below.

$$\begin{pmatrix} P_i^* \\ W_i^* \\ E_i^* \end{pmatrix} \sim t_2 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix} \right), \tag{66}$$

$$P_i = H^{-1}(F(P_i^*)) \text{ where } H \text{ is the cdf of a Gamma (1,1) distribution },\tag{67}$$

$$W_i = L^{-1}(F(W_i^*)) \text{ where } L \text{ is the cdf of a Exponential (1) distribution,}\tag{68}$$

$$E_i = G^{-1}(F(E_i^*)) \text{ where } G \text{ is the cdf of a Uniform[-0.5, 0.5] distribution}\tag{69}$$

$$F \text{ is the cdf of the student-t(2) distribution.}\tag{70}$$

Simulation results presented in Table 14 show that despite the mis-specification of GC model and the error distribution, 2sCOPE-np can successfully correct for bias. Therefore, the proposed method demonstrates reasonable robustness to the mis-specifications of the assumed copula dependence structure and normality.

Table 14: Simulation results for OLS and 2sCOPE-np under $E_i \sim Unif[-0.5, 0.5]$ and t-copula.

		OLS			2sCOPE-np		
Parameters	True	Bias	SE	RMSE	Bias	SE	RMSE
μ	0	0.080	0.010	0.080	0.009	0.014	0.016
α	1	-0.158	0.009	0.158	-0.023	0.017	0.029
β	2	0.079	0.009	0.079	0.015	0.011	0.019
σ_E	0.29	-0.034	0.004	0.035	-0.009	0.007	0.011

Note: Bias, SE, and RMSE denote the mean, standard deviation, and root mean squared error of parameter estimates across all 1,000 simulated samples of size 2000.

C.5 Multiple regressors: tradeoff between flexibility and efficiency

To show that the tradeoff between flexibility and efficiency when using 2sCOPE-np method, we generate data from a joint Gaussian copula using the following data generating process (DGP).

$$\begin{pmatrix} E_i^* \\ P_{1i}^* \\ W_{1i}^* \\ P_{2i}^* \\ W_{2i}^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho_{pe} & 0 & \rho_{pe} & 0 \\ \rho_{pe} & 1 & \rho_{pw} & \rho_{p_1p_2} & \rho_{pw} \\ 0 & \rho_{pw} & 1 & 0 & \rho_{w_1w_2} \\ \rho_{pe} & \rho_{p_1p_2} & 0 & 1 & \rho_{pw} \\ 0 & \rho_{pw} & \rho_{w_1w_2} & \rho_{pw} & 1 \end{bmatrix}\right),$$

$$\rho_{pe} = 0.2, \quad \rho_{pw} = 0.5, \quad \rho_{w_1w_2} = 0.2, \quad \rho_{p_1p_2} = 0.2,$$

$$E_i = G^{-1}(U_{E,i}) = G^{-1}(\Phi(E_i^*)) = \Phi^{-1}(\Phi(E_i^*)) = E_i^*,$$

$$P_{ji} = H_j^{-1}\big(\Phi(P_{ji}^*)\big), \qquad W_{ji} = L_j^{-1}\big(\Phi(W_{ji}^*)\big), \qquad j = 1, 2,$$

$$\begin{aligned} Y_i &= \mu + \alpha_1 P_{1i} + \alpha_2 P_{2i} + \beta_1 W_{1i} + \beta_2 W_{2i} + E_i \\ &= 1 + 1 \cdot P_{1i} + 1 \cdot P_{2i} + (-1) \cdot W_{1i} + (-1) \cdot W_{2i} + E_i. \end{aligned} \tag{71}$$

The simulated data has an endogeneity problem because of the correlation between E_i^*, P_{1i}^* and P_{2i}^* ($\rho_{pe} = 0.5$). The control variable W_{1i} and W_{2i} are exogenous (i.e., uncorrelated with E_i), but is correlated with the endogenous regressor ($\rho_{pw} = 0.5$) and also affects the outcome Y ($\beta_1, \beta_2 \neq 0$). Thus, W_1 and W_2 are observed confounders to be controlled for in the outcome regression model. In the simulation, we set H_1 to be the CDF of a truncated standard normal with a lower bound of 0 and an upper bound of 2. This emulates a scenario of an endogenous regressor with bounded range. H_2 is the CDF of a chi-squared distribution with 2 degrees of freedom. We set L_1 to be the CDF of the standard normal distribution, and L_2 to be the DCF of a Gamma distribution with rate parameter of 9, and scale parameter of 0.5. All regressors and error term follow GC dependence structure.

Table 15: Simulation Study W1: The Case of Joint Copula Data with multiple covariates

		OLS			2sCOPE-GC			2sCOPE-MD			2sCOPE-np		
Parameters	True	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
μ	1	0.036	0.069	0.077	0.002	0.082	0.082	-0.110	0.101	0.150	0.045	0.089	0.100
α_1	1	0.652	0.054	0.654	-0.054	0.219	0.231	0.201	0.453	0.496	-0.026	0.336	0.337
α_2	1	0.118	0.012	0.119	-0.014	0.034	0.037	0.003	0.036	0.036	-0.009	0.033	0.034
β_1	-1	-0.110	0.025	0.113	0.013	0.052	0.053	-0.043	0.100	0.108	-0.001	0.068	0.068
β_2	-1	-0.041	0.004	0.041	0.004	0.010	0.011	-0.002	0.018	0.018	-0.001	0.012	0.012
σ_E	1	-0.055	0.015	0.057	0.016	0.032	0.036	-0.002	0.042	0.042	0.065	0.065	0.092

Note: 2sCOPE-MD implements the method of Mayer and Wied (2026) and 2sCOPE-GC implements the method of Yang et al. (2025). Bias, SE and RMSE denote the mean, standard deviation and root mean squared error of parameter estimates across all 1,000 simulated samples of size 2000.

Results reported in Table 15 show that OLS estimates and 2sCOPE-MD exhibit large endogeneity bias and RMSE, in comparison to 2sCOPE-GC and 2sCOPE-np estimates. This is expected because, when the underlying data are generated from the GC model, 2sCOPE-GC and 2sCOPE-np should eliminate the bias. 2sCOPE-GC method imposes more assumptions about the dependence structure than 2sCOPE-np. However, since these assumptions are correctly specified in this simulation, 2sCOPE-GC is more efficient than 2sCOPE-np, yielding smaller standard errors and RMSE for the endogenous parameters.

Our proposed 2sCOPE-np's flexibility, due to the lack of modelling assumptions, comes at the expense of some efficiency loss compared with 2sCOPE-GC, which is evident in the larger standard errors observed across simulations. Estimating the conditional CDF

$F(P|W)$ is more challenging than estimating the marginal CDF, leading to slower convergence rates (requiring larger sampler size [35] and higher computational cost), which is expected for any nonparametric estimator.

C.6 Binary endogenous regressor

Teasing out endogeneity in treatment status is an important step in management research to identify the true effect of marketing interventions, such as advertising, price promotions, etc. The binary treatment status may be correlated with an unobserved confounder absorbed in the structural error term, creating an endogeneity problem. If left unaddressed, this can lead to biased estimates of the treatment effect. Thus, binary endogenous regressors play a crucial role in endogeneity correction.

This simulation study considers the following DGP with a binary endogenous regressor.

$$\begin{pmatrix} P_i^* \\ W_i^* \\ E_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right), \tag{72}$$

$$E_i = G^{-1}(U_{E,i}) = G^{-1}(\Phi(E_i^*)) = \Phi^{-1}(\Phi(E_i^*)) = 1 \cdot E_i^*, \tag{73}$$

$$P_i = \mathbb{I}_{\{U_{P,i} > p_1\}} = \mathbb{I}_{\{\Phi(P_i^*) > p_1\}} = \mathbb{I}_{\{P_i^* > \Phi^{-1}(p_1)\}}, \tag{74}$$

$$W_i = L^{-1}(U_{W,i}) = L^{-1}(\Phi(W_i^*)), \tag{75}$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 0 + 1 \cdot P_i + 2 \cdot W_i + E_i. \tag{76}$$

where

[35] 2sCOPE-GC method also contains a nonparametric portion - the estimation of the marginal CDFs using empirical CDFs. When comparing 2sCOPE-np and 2sCOPE-GC, we also tried using kernel-method rather than empirical CDFs to estimate these marginal CDFs in 2sCOPE-GC in Table 5. The result for 2sCOPE-GC is similar regardless of estimation methods.

$$p_1 = 0.5, \text{ and } \mathbb{I}_{\{x>0\}} = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases} \text{ is the indicator function.}$$

P_i^* can be considered as a latent variable that determines the binary treatment status P_i. An endogeneity problem arises because structural error E^* and latent variable P^* are correlated. The control variable W_i is an observed exogenous confounder, uncorrelated with the structural error E_i. In the simulation, we set $L(W)$ as the student-t distribution (df=3).

We generate 1000 datasets as replicates, each with a sample size of 2000, using the DGP above. Then, we apply the same four estimation methods as described previously, to all generated datasets, producing the estimates' empirical distributions.

Without endogeneity correction, OLS yields large biases (91.2%, Table 16) for the endogenous regressor. Due to such a large bias, OLS estimates have the largest RMSE, despite having the smallest SEs. With endogeneity correction, all three 2sCOPE methods outperform OLS; however, only 2sCOPE-np successfully corrects for endogeneity bias (Bias: 5.7% Table 16). As expected, 2sCOPE-GC and 2sCOPE-MD show significant bias when applied to binary endogenous regressors.

Table 16: Simulation Study 5: Binary Endogenous Regressors

		OLS			2sCOPE-GC			2sCOPE-MD			2sCOPE-np		
Parameters	True	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
μ	0	-0.456	0.032	0.457	0.110	0.163	0.197	-0.213	0.070	0.225	-0.029	0.137	0.140
α	1	**0.912**	**0.047**	**0.913**	**-0.219**	**0.320**	**0.388**	**0.428**	**0.132**	**0.447**	**0.057**	**0.267**	**0.273**
β	2	0.096	0.019	0.097	-0.027	0.040	0.048	0.021	0.025	0.032	0.005	0.032	0.032
σ_E	1	-0.096	0.015	0.097	0.054	0.080	0.096	-0.066	0.022	0.070	-0.006	0.056	0.056

See the note under Table 5.

D Proof

D.1 Proposition 1

Let $(O,P)|W$ follow a Gaussian Copula with correlation coefficient ρ. According to Sklar's theorem,

$$F_{\theta,\phi}((o,p)|w) = \quad C_{\theta_2}(F_{\theta_1}(o|w), F_\phi(p|w)), \tag{77}$$

$$= \quad \Psi_\rho(\Phi^{-1}(u_{o|w}), \Phi^{-1}(u_{p|w})) \tag{78}$$

$$\text{where} \quad u_{o|w} = F_{\theta_1}(o|w), u_{p|w} = F_\phi(p|w), \tag{79}$$

$\Psi_\rho(\cdot,\cdot)$ is the CDF of the bivariate standard normal distribution with a conditional correlation coefficient ρ; $\Phi(\cdot)$ denotes the standard normal CDF.

Let $Z_O := \Phi^{-1}(u_{o|w}), Z_P := \Phi^{-1}(u_{p|w})$. Because $(O,P) \mid W$ follows a Gaussian copula with parameter ρ, (Z_O, Z_P) is bivariate standard normal with correlation ρ. That is,

$$(Z_O, Z_P) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

For the bivariate standard normal vector, the conditional expectation of Z_O given Z_P is

$$\mathbb{E}[Z_O \mid Z_P = z] = \rho z.$$

Define

$$C_p := \Phi^{-1}(F_\phi(p \mid w)),$$

where $P = p$ which is the realized value of P, and C_p is the realized value of Z_P.

Since $O \perp W$, then $O|W \sim N(0, \sigma_o^2)$, and $\frac{O}{\sigma_o}|W \sim N(0,1)$ and $F_{\theta_1}(o \mid w) = \Phi(o/\sigma_o)$. Thus

$$Z_O = \Phi^{-1}(\Phi(O/\sigma_o)) = \frac{O}{\sigma_o}, \qquad \text{so that} \qquad O = \sigma_o Z_O.$$

Therefore,

$$\mathbb{E}(O \mid P = p, w) = \sigma_o \, \mathbb{E}[Z_O \mid Z_P = C_p] = \sigma_o \rho \, C_p.$$

Thus,

$$\mathbb{E}(O \mid P = p, w) = \sigma_o \rho \, \Phi^{-1}(F_\phi(p \mid w)) \, .$$

D.2 Proposition 2

Under the conditional GC model, $(\Phi^{-1}(u_{o|w}), \Phi^{-1}(u_{p1|w}), \cdots, \Phi^{-1}(u_{p_K|w}))$ follow the standard multivariate normal distribution:

$$\begin{pmatrix} \Phi^{-1}(u_{o|w}) \\ \Phi^{-1}(u_{p_1|w}) \\ \cdot \\ \Phi^{-1}(u_{p_K|w}) \end{pmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & \rho_{op_1} & \cdot & \rho_{op_k} \\ \rho_{op_1} & 1 & \cdot & \rho_{p_1 p_k} \\ \cdot & \cdot & 1 & \cdot \\ \rho_{op_k} & \rho_{p_1 p_k} & \cdot & 1 \end{bmatrix} \right). \tag{80}$$

Define the latent Gaussian probability integral transforms

$$Z_O := \Phi^{-1}(u_{o|w}), \qquad Z_P := \begin{pmatrix} \Phi^{-1}(u_{p_1|w}) \\ \vdots \\ \Phi^{-1}(u_{p_K|w}) \end{pmatrix}.$$

By the conditional Gaussian copula model (80), the $(K+1)$-vector $(Z_O, Z_P^\top)^\top$ is multivariate standard normal with mean zero and covariance

$$\Sigma = \begin{pmatrix} 1 & \Sigma_{OP} \\ \Sigma_{OP}^\top & \Sigma_{PP} \end{pmatrix},$$

where $\Sigma_{OP} = (\rho_{op_1}, \dots, \rho_{op_K})$ is the $1 \times K$ vector of correlations between Z_O and each Z_{P_k}, and Σ_{PP} is the $K \times K$ correlation matrix among the Z_{P_k}'s (entries $\rho_{p_i p_j}$).

Because marginally $O \sim N(0, \sigma_o^2)$ and $O \perp W$, then $O|W \sim N(0, \sigma_o^2)$, and we have

$$Z_O = \Phi^{-1}\big(F_{\theta_1}(O \mid w)\big) = \Phi^{-1}\big(\Phi(O/\sigma_o)\big) = \frac{O}{\sigma_o},$$

so $O = \sigma_o Z_O$.

Let $c := Z_P\big|_{P=p} = \big(C_{p_1}, \dots, C_{p_K}\big)^\top$ denote the observed values of the latent Z_P components where,

$$C_{p_k} := \Phi^{-1}\big(F_{\phi_k}(p_k \mid w)\big), \qquad k = 1, \dots, K.$$

For a multivariate standard normal vector, the conditional expectation of the scalar Z_O given $Z_P = c$ is

$$\mathbb{E}[Z_O \mid Z_P = c] \;=\; \Sigma_{OP}\, \Sigma_{PP}^{-1}\, c.$$

Therefore, using $O = \sigma_o Z_O$,

$$\mathbb{E}[O \mid P = p_1, \dots, p_K, w] = \sigma_o\, \mathbb{E}[Z_O \mid Z_P = c] = \sigma_o\, \Sigma_{OP}\, \Sigma_{PP}^{-1}\, c.$$

Writing the $1 \times K$ vector $\sigma_o\, \Sigma_{OP}\, \Sigma_{PP}^{-1}$ componentwise as $(\gamma_1, \dots, \gamma_K)$, we obtain the linear form

$$\mathbb{E}[O \mid P = p_1, \ldots, p_K, w] = \sum_{k=1}^{K} \gamma_k \, C_{p_k},$$

where each coefficient

$$\gamma_k \;=\; \sigma_o \left[\Sigma_{OP} \, \Sigma_{PP}^{-1}\right]_k$$

is the k-th element of the row vector $\sigma_o \, \Sigma_{OP} \, \Sigma_{PP}^{-1}$.

Thus,

$$\mathbb{E}(O \mid P = p_1, \cdots, p_k, w) = \sum_{k=1}^{K} C_{p_k} \gamma_k.$$

D.3 Proposition 3

Suppose that the data can be divided into J heterogenous samples. For subsample j, assume that the omitted variable $O_j \sim N(0, \sigma_j^2)$ and the multivariate conditional GC model for $(O_j, P_{j1}, \cdots, P_{jK)}$ given W_j with correlation matrix $\Sigma_j = \begin{pmatrix} 1 & \Sigma_{O_j P_j} \\ \Sigma_{O_j P_j}^{\top} & \Sigma_{P_j P_j} \end{pmatrix}$.

For each subsample $j = 1, \cdots, J$, define

$$u_{o_j|w_j} := F_{\theta_j}(o_j|w_j), u_{p_{jk}|w_j} =: F_{\phi_{jk}}(p_{jk}|w_j) \text{ for } k = 1, \cdots, K$$

and the latent Gaussian probability integral transforms as

$$Z_{O_j} := \Phi^{-1}(u_{o_j|w_j}), \qquad Z_{P_j} := \begin{pmatrix} \Phi^{-1}(u_{p_{j1}|w_j}) \\ \vdots \\ \Phi^{-1}(u_{p_{jK}|w_j}) \end{pmatrix}.$$

In subsample j, $O_j|W_j \sim N(0, \sigma_j^2)$, and $\frac{O_j}{\sigma_j}|W_j \sim N(0,1)$ and $F_j(o_j \mid w_j) = \Phi(o_j/\sigma_j)$.

Thus

$$Z_{O_j} = \Phi^{-1}(\Phi(O_j/\sigma_j)) = \frac{O_j}{\sigma_j}, \qquad \text{and} \qquad O_j|W_j = \sigma_j Z_{O_j}.$$

By the conditional Gaussian copula model, the $(K+1)$-vector $(Z_{O_j}, Z_{P_j}^\top)^\top$ is multivariate standard normal with mean zero and covariance Σ_j.

Let $c_j := Z_{P_j}\big|_{P_j=p_j} = \big(C_{p_j1}, \ldots, C_{p_jK}\big)^\top$ denote the observed values of the latent Z_{P_j} components in the subsample j, where

$$C_{p_jk} := \Phi^{-1}\big(F_{\phi_{jk}}(p_{jk} \mid w_j)\big), \qquad k = 1, \ldots, K.$$

For a multivariate standard normal vector, the conditional expectation of the scalar Z_{O_j} given $Z_{P_j} = c_j$ is

$$\mathbb{E}[Z_{O_j} \mid Z_{P_j} = c_j] \; = \; \Sigma_{O_jP_j}\, \Sigma_{P_jP_j}^{-1}\, c_j.$$

Therefore, using $O_j = \sigma_j Z_{O_j}$,

$$\mathbb{E}[O_j \mid P_j = p_{j1}, \ldots, p_{jK}, w_j] = \sigma_j\, \mathbb{E}[Z_{O_j} \mid Z_{P_j} = c_j] = \sigma_j\, \Sigma_{O_jP_j}\, \Sigma_{P_jP_j}^{-1}\, c_j.$$

Writing the $1 \times K$ vector $\sigma_j\, \Sigma_{O_jP_j}\, \Sigma_{P_jP_j}^{-1}$ componentwise as $(\gamma_{j1}, \ldots, \gamma_{jK})$, we obtain the linear form

$$\mathbb{E}[O_j \mid P = p_{j1}, \ldots, p_{jK}, w_j] = \sum_{k=1}^{K} \gamma_{jk}\, C_{p_{jk}},$$

in subsample j and

$$\mathbb{E}(o|p_1, \cdots, p_K, w) = \sum_{k=1}^{K} \sum_{j=1}^{J} I(w \in S_j) \gamma_{jk} C_{p_{jk}},$$

where the (j,k)th copula-generated regressor for subsample j and endogenous regressor k is $C_{p_jk} = \Phi^{-1}(F_{\phi_{jk}}(p_{jk}|w_j))$, and $F_{\phi_k}(p_{jk}|w_j)$ is the conditional CDF of P_{jk} given W_j in subsample j, and

$$(\gamma_{j1}, \cdots, \gamma_{jk}) \ = \ \sigma_j \left[\Sigma_{O_j P_j} \, \Sigma_{P_j P_j}^{-1} \right]$$

S_j denotes the space of subsample j,

D.4 (P, W) does not follow GC in Section 4.4

Let W follow a continuous marginal distribution (i.e. Student-t). Conditional on $W = W_i$, $P_i | W_i \sim$ truncated standard normal($a = \min(0, -2W_i + 2), b = \max(2, -2W_i + 2)$). This means, for $w \in (0, 1)$ we have $(P \mid W = w) \sim N(0, 1)$ truncated to $[0, 2]$.

Suppose that (P, W) follow Gaussian copula. Let F_P, F_W denote the marginal CDFs and f_P, f_W their densities. If (P, W) has a Gaussian copula with density $c(u, v)$, then

$$f_{P|W}(p \mid w) = c(F_P(p), F_W(w)) \; f_P(p).$$

Since the Gaussian copula density $c(u, v) > 0$ for all $(u, v) \in (0, 1)^2$, it follows that

$$f_P(p) > 0 \quad \Longrightarrow \quad f_{P|W}(p \mid w) > 0 \quad \text{for all } w.$$

This means that under a Gaussian copula, whenever p lies in the support of P, it must also lie in the conditional support of $P \mid W = w$ for every w.

Take $p = -1$. Observe that $-2w + 2 \leq -1$ iff $w \geq 3/2$, hence for every $w \geq 3/2$ the truncation interval is $[a(w), b(w)]$ contains -1.

Since W has positive probability on $[3/2, \infty)$, we obtain $f_{P|W}(-1 \mid w) > 0$ for a set of w of positive probability, thus the marginal satisfies $f_P(-1) > 0$.

However, for every $w \in (0, 1)$ the conditional support is $[0, 2]$, so

$$f_{P|W}(-1 \mid w) = 0 \quad \text{for all } w \in (0, 1).$$

Thus there exists p with $f_P(p) > 0$ but $f_{P|W}(p \mid w) = 0$ on a set of w with positive probability, contradicting the positivity implication above. (P, W) cannot be generated by a GC.

E Additional Results of Empirical Applications

E.1 Box plot of log price and kernel estimates of CDFs

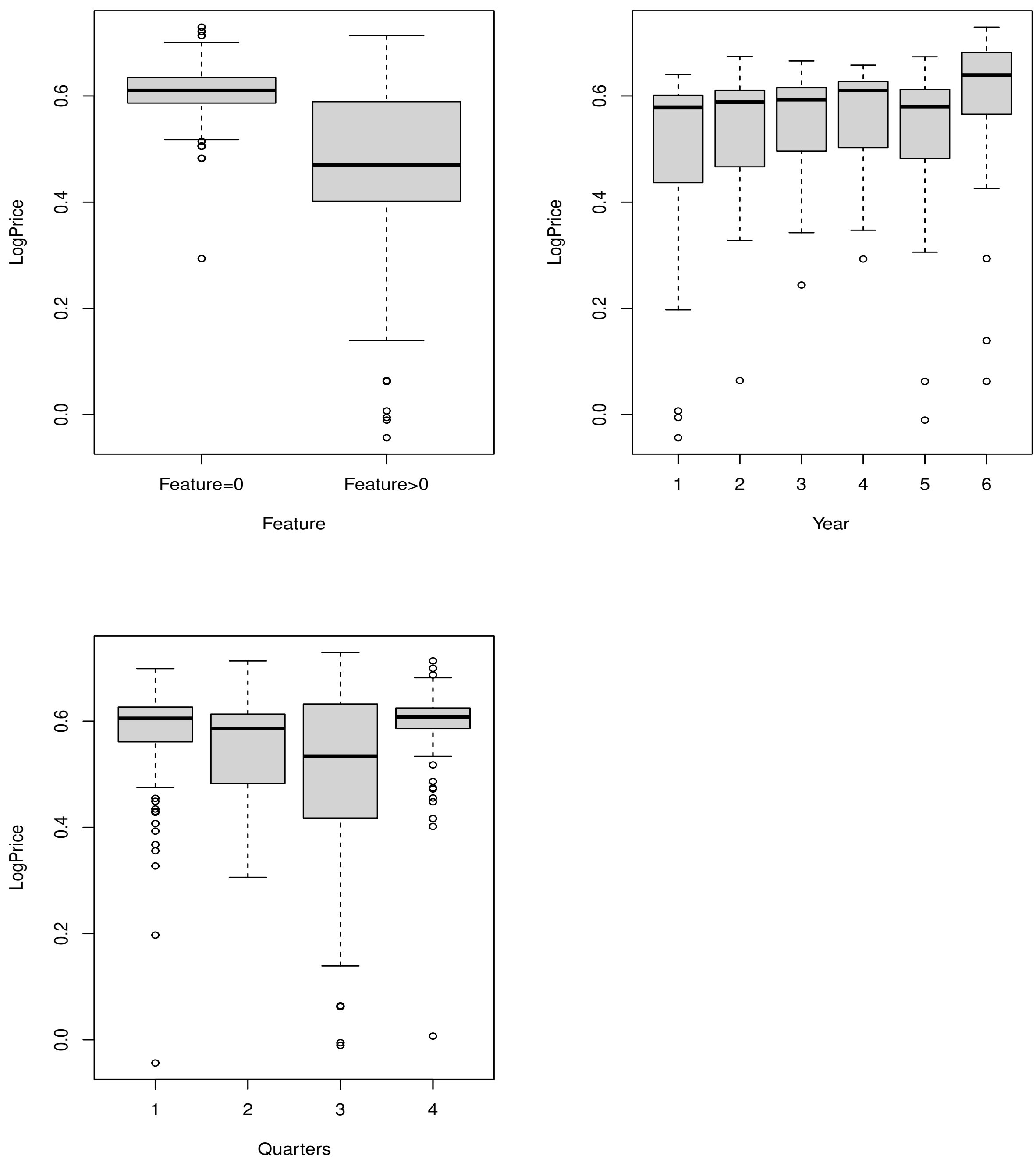

Figure 1: Boxplots of LogPrice in Store 1.

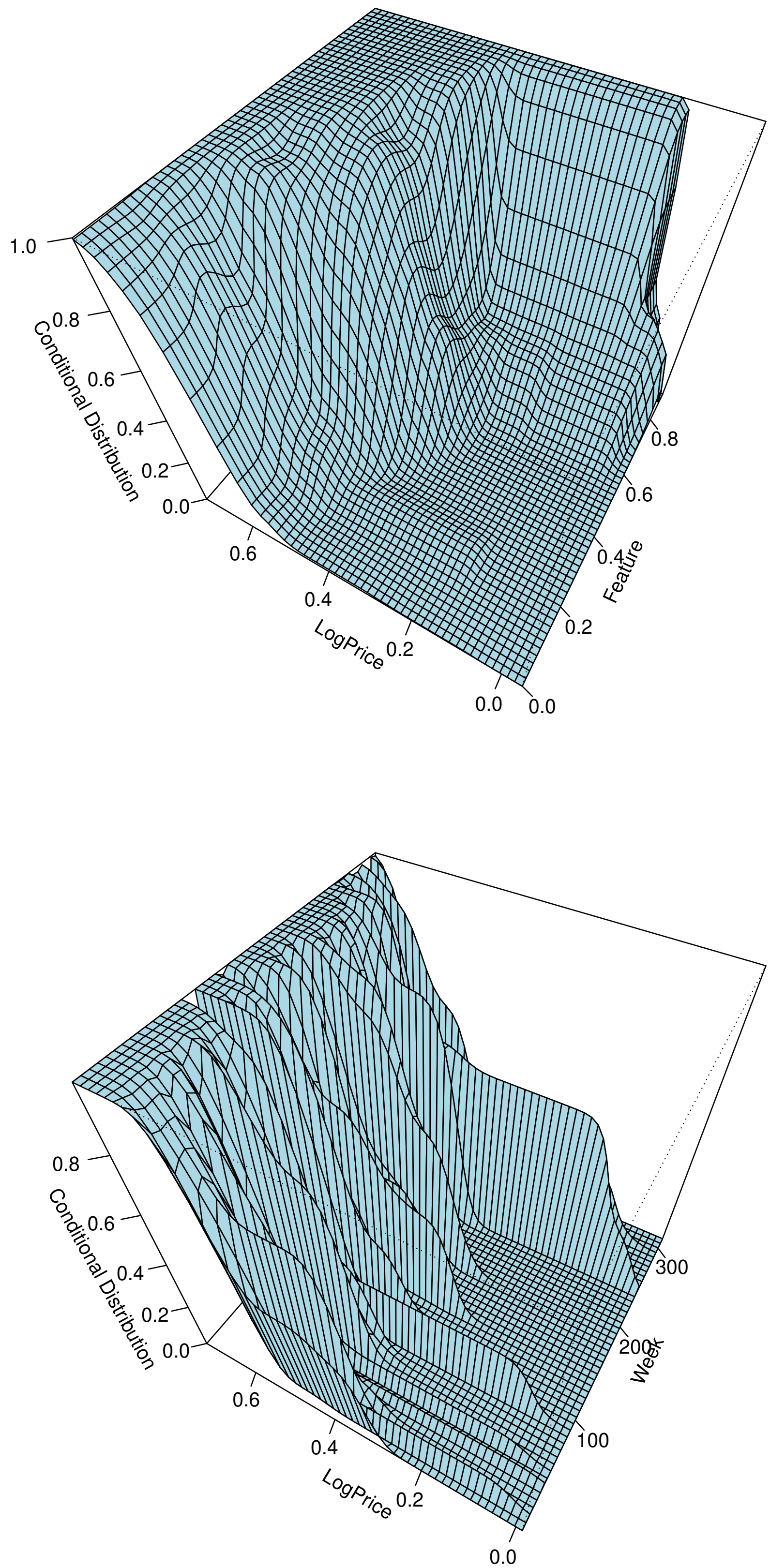

Figure 2: Kernel Estimates of CDFs of logPrice Conditional on Feature (top) and Week (Bottom) in Store 1.

E.2 Empirical example 2: return to education (the case of weak IV)- a binary regressor for postsecondary education

To examine the return to postsecondary education, we operationalize education as a binary variable of postsecondary education and estimate education return using OLS, 2SLS with IV, and the three 2sCOPE methods. We define postsecondary education as more than 12 years of education (end of high school).

As reported in Table 17, both 2SLS (Est.=0.69, SE=0.30) and 2sCOPE-np (Est.=0.36, SE=0.06) yield higher estimates of return to postsecondary education than the OLS estimate (Est.=0.22, SE= 0.02). The 2SLS estimate of the return to postsecondary education corresponds to a point estimate of 0.69 increase in log-wage, but has a large standard error of 0.30. This makes it difficult to interpret the 2SLS estimation result - the 95% confidence interval on the log-wage translates to a wage increase between 11% to 260%, which differ drastically in magnitude. In addition, the 2SLS confidence interval is so large that, in fact, it again covers the 95% confidence interval of the OLS estimate - making the interval estimation interpretation of 2SLS no different from that of OLS. Testing for weak instruments using the first-stage F-statistic rejects the null, but the partial correlation between the IV and postsecondary education is only 0.067 and the F statistic is borderline bigger than 10 ($\rho = 0.067$, partial F= 12.921, p-value= 0.0003). The Wu-Hausman test rejects the null hypothesis (H=3.005, p-value= 0.083) and suggests the presence of endogeneity at the 10% significance level. We conclude that the IV is likely weak [36]. This explains why the 2SLS result is statistically no different from that of OLS, despite the evidence of endogeneity.

[36] Card (1999) argues that other IVs such as family background variables are potentially endogenous, whereas the proximity to college variable has a clear exclusion restriction. Ebbes et al. (2005) also comments that proximity to college appears to be a weak instrumental variable. This further highlights the empirical challenge of finding a good IV that is highly relevant and satisfies the exclusion restriction.

In this case, where 2SLS uses a weak IV, 2sCOPE-np estimation result has the advantage of being more accurate. The estimated return to log wage using 2sCOPE-np is 0.36[37] with 95% confidence interval of (0.26, 0.47), compared to OLS results of 0.22 and 95% confidence interval (0.18, 0.25). The two confidence intervals do not overlap, and 2sCOPE-np estimate is strictly larger than that of OLS, indicating the presence of potential negative bias in the OLS estimate. At first look, the point estimation results of 2sCOPE-np and

Table 17: Estimation Results for Binary Education Variable

	OLS			2SLS			2sCOPE-np			2sCOPE-GC			2sCOPE-MD		
Parameters	Est	SE	t-val	Est	SE	t-val	Est	SE	t-val	Est	SE	t-val	Est	SE	t-val
(Intercept)	5.46	0.07	76.96	5.10	0.24	20.94	5.33	0.09	62.16	5.37	0.15	36.84	5.40	0.09	62.90
postsecondary education	**0.22**	**0.02**	**12.17**	**0.69**	**0.30**	**2.28**	**0.36**	**0.06**	**6.54**	**0.35**	**0.20**	**1.77**	**0.31**	**0.08**	**4.00**
exper	0.08	0.01	11.64	0.13	0.03	4.10	0.10	0.01	10.15	0.09	0.02	5.73	0.09	0.01	9.33
expersq	0.00	0.00	-8.58	0.00	0.00	-4.82	0.00	0.00	-7.85	-0.00	0.00	-7.94	-0.00	0.00	-8.71
black	-0.21	0.02	-10.42	-0.18	0.03	-5.74	-0.21	0.02	-10.11	-0.20	0.03	-7.82	-0.20	0.02	-9.60
feduc_imp	0.01	0.00	1.33	0.00	0.01	0.00	0.01	0.00	1.26	0.00	0.01	0.91	0.00	0.01	0.95
meduc_imp	0.01	0.00	3.25	0.00	0.01	0.42	0.01	0.00	3.09	0.01	0.01	1.67	0.01	0.00	2.30
feduc × meduc	0.00	0.00	-0.94	0.00	0.00	-1.18	0.00	0.00	-0.00	-0.00	0.00	-1.03	-0.00	0.00	-0.89
miss_feduc	0.06	0.04	1.54	-0.04	0.07	-0.56	0.05	0.04	1.33	0.03	0.05	0.52	0.03	0.04	0.86
miss_meduc	0.16	0.04	3.82	0.10	0.06	1.55	0.15	0.04	3.89	0.14	0.05	2.63	0.15	0.04	3.34
momdad14	0.09	0.04	2.57	0.04	0.05	0.89	0.08	0.03	2.32	0.08	0.04	1.81	0.08	0.04	2.25
sinmom14	0.04	0.04	0.89	0.04	0.05	0.94	0.03	0.04	0.79	0.04	0.04	0.87	0.04	0.04	0.89
step14	0.01	0.05	0.14	0.04	0.06	0.66	0.01	0.05	0.32	0.02	0.05	0.29	0.01	0.05	0.27
$C_{\text{education}}$							**-0.12**	**0.04**	**-2.84**	**-0.04**	**0.06**	**-0.69**	**-0.04**	**0.03**	**-1.21**
Wu-Hausman test				**H=3.01, p=0.083**											

See the note under Table 10

2sCOPE-GC are close to each other. However, the standard error of 2sCOPE-GC estimate is so large that the postsecondary education coefficient is statistically insignificant, contrary to theoretical predictions. The point estimation results of 2sCOPE-MD are close to one standard deviation away from 2sCOPE-np, and generate a confidence interval which overlaps with that of OLS.

Testing for endogeneity with the copula correction term in 2sCOPE-np gives a statistically significant result (Est.= -0.12, SE=0.04, p-value=0.003). [38] In contrast, both

[37] The estimation result of 0.36 with a standard error of 0.06 is reasonable, considering past research on the annual return to education (Ebbes et al. 2005), and the years of education during postsecondary.

[38] Running 2sCOPE-np with heterogeneous samples yields insignificant region-difference control function coefficients, suggesting that there is no heterogeneity on the region level. Thus, we omit the results here and only include the main specification.

2sCOPE-GC and 2sCOPE-MD yield statistically insignificant copula control function term coefficients, indicating no endogeneity bias. In this case, using methods (2sCOPE-MD, 2sCOPE-GC) that theoretically cannot handle binary regressors in a binary setting can be problematic, generating counterintuitive results.

The negative and significant coefficient of the copula correction term in 2sCOPE-np suggests the presence of a downward bias in the OLS estimates, which could be the result of measurement error. When discretizing the education variable from a count measure into a binary indicator, it is possible that the measurement bias in the education variable is amplified, resulting in overall downward bias. Overall, the estimation result and the test of endogeneity under 2sCOPE-np are consistent with theoretical predictions of potential bias in OLS, as well as more accurate and usable than those of 2SLS using IVs, 2sCOPE-GC and 2sCOPE-MD.

E.3 Variable definition for the education example

Table 18: Variable Definition

Variable name	Description
nearc4	indicator for whether a subject grew up near a four-year college
educ	subject's years of education
fatheduc	subject's father's years of education
motheduc	subject's mother's years of education
momdad14	indicator for whether subject lived with both mother and father at age 14
sinmom14	indicator for whether subject lived with a single mom at age 14

Variable name	Description
step14	indicator for whether subject lived with stepparent at age 14
reg661	indicator for whether subject lived in region 1 (New England) in 1966
reg662	indicator for whether subject lived in region 2 (Middle Atlantic) in 1966
reg663	indicator for whether subject lived in region 3 (East North Central) in 1966
reg664	indicator for whether subject lived in region 4 (West North Central) in 1966
reg665	indicator for whether subject lived in region 5 (South Atlantic) in 1966
reg666	indicator for whether subject lived in region 6 (East South Central) in 1966
reg667	indicator for whether subject lived in region 7 (West South Central) in 1966
reg668	indicator for whether subject lived in region 8 (Mountain) in 1966
reg669	indicator for whether subject lived in region 9 (Pacific) in 1966
south66	indicator for whether subject lived in South in 1966
black	indicator for whether subject's race is black
smsa	indicator for whether subject lived in SMSA in 1976
south	indicator for whether subject lived in the South in 1976
smsa66	indicator for whether subject lived in SMSA in 1966
wage	subject's wage in cents per hour in 1976
exper	subject's years of labor force experience in 1976
lwage	subject's log wage in 1976
expersq	square of subject's years of labor force experience in 1976

Note: SMSA: Standard Metropolitan Statistical Area, now referred to as Metropolitan Statistical Areas.

Appendix References

Aitchison J, Aitken CG (1976) Multivariate binary discrimination by the kernel method. *Biometrika* 63(3):413–420.

Breitung J, Mayer A, Wied D (2024) Asymptotic properties of endogeneity corrections using nonlinear transformations. *The Econometrics Journal* 27(3):362–383.

Card D (1999) The causal effect of education on earnings. *Handbook of labor economics* 3:1801–1863.

Ebbes P, Wedel M, Böckenholt U, Steerneman T (2005) Solving and testing for regressor-error (in)dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics* 3(4):365–392.

Haschka RE (2022) Handling endogenous regressors using copulas: A generalization to linear panel models with fixed effects and correlated regressors. *Journal of Marketing Research* 59(4):860–881.

Hayfield T, Racine JS (2008) Nonparametric econometrics: The np package. *Journal of Statistical Software* 27:1–32.

Li Q, Lin J, Racine JS (2013) Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics* 31(1):57–65.

Li Q, Racine JS (2008) Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics* 26(4):423–434.

Liengaard BD, Becker JM, Bennedsen M, Heiler P, Taylor LN, Ringle CM (2025) Dealing with regression models' endogeneity by means of an adjusted estimator for the gaussian copula approach. *Journal of the Academy of Marketing Science* 53(1):279–299.

Lin Y, Song Y (2025) Addressing endogeneity issues in a spatial autoregressive model using copulas. *Journal of Econometrics* 252:106106.

Mayer A, Wied D (2026) Endogeneity corrections in binary outcome models with nonlinear transformations: Identification and inference. *Oxford Bulletin of Economics and Statistics* 88(1):45–56.

Park S, Gupta S (2012) Handling endogenous regressors by joint estimation using copulas. *Marketing Science* 31(4):567–586.

Qian Y, Xie H (2024) Correcting regressor-endogeneity bias via instrument-free joint estimation using semiparametric odds ratio models. *Journal of Marketing Research* 61(5):914–936.

Yang F, Qian Y, Xie H (2025) Addressing endogeneity using a two-stage copula generated regressor approach. *Journal of Marketing Research* 62(4):601–623.

Yang L, Qian Y, Xie H (2024) Handling endogenous marketing mix regressors in correlated heterogeneous panels with copula augmented mean group estimation. *NBER Working Paper w33265, https://www.nber.org/papers/w33265* .