Correcting Endogeneity via Nonparametric Copula Control Functions

Abstract

We propose a new framework that addresses endogenous regressors using a novel conditional copula endogeneity model to capture the regressor-error dependence unexplained by exogenous regressors. Building on the model, we develop a two-stage nonparametric copula control function approach for endogeneity correction without relying on instrumental variables. The method relaxes the restrictive assumption of the Gaussian copula regressor-error dependence structure and eliminates the need to model regressors. It unifies and generalizes existing copula-based endogeneity correction methods, while minimizing assumptions about the first-stage auxiliary dependence structures among regressors. Specifically, the approach constructs control functions using nonparametric estimates of the conditional cumulative distribution functions (CDFs) of endogenous regressors given exogenous variables, enhancing the accuracy and robustness of endogeneity correction. Unlike existing copula control function methods, it applies to broader dependence structures and 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.

Keywords: causal inference, endogeneity, generated regressor, instrumental variable, kernel estimation

Disclosure Statement: The authors report there are no competing interests to declare.

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. However, identifying and defending such instruments in practice is often difficult (Wooldridge 2010, Ebbes et al. 2009). These challenges spur development and application of instrument-free methods that do not rely on observed IVs or on full-information modeling (Lewbel 1997, Rigobon 2003, Ebbes et al. 2005, 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 economics, management, strategy, information systems, and marketing (Becerra and Markarian 2021, Blinded 2025, Ananthakrishnan et al. 2025). First introduced by Park and Gupta (2012), these methods as described in Blinded (2025) 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 insufficient nonnormality or correlated with exogenous regressors, discrete endogenous regressors, panel data, and spatial data (Christopoulos et al. 2021, Haschka 2022, Yang et al. 2024, 2025, Qian and Xie 2024, Breitung et al. 2024, Liengaard et al. 2024, Lin and Song 2025). Despite their flexibility, existing copula-based endogeneity correction methods rely upon one or more of the following assumptions or requirements: independence between endogenous and exogenous regressors, specific copulas governing regressor-error dependence, auxiliary modelling of endogenous regressors, continuity of endogenous regressors, or joint estimation of the outcome and regressor models (see Web Appendix 1 Table W1 for a comparison of these methods).

To relax these assumptions and data requirements, this article introduces a nonparametric two-stage endogeneity correction procedure (2sCOPE-np) based on the conditional copula theory. Specifically, we use conditional copulas to capture the dependence between the error (or its endogenous part) and endogenous regressors *given* exogenous control regressors, without restricting the conditional marginal distributions in the copula to be from the same family. Under the general conditional copula endogeneity modeling framework, 2sCOPE-np consists of two steps. In the first stage, 2sCOPE-np uses robust nonparametric estimates of the conditional CDFs of endogenous regressors given exogenous control variables to construct copula generated regressors. These generated regressors break the regressor-error dependence and are included as additional control variables in the second-stage augmented model to address endogeneity. As a control function approach, the nonparametric 2sCOPE-np is relatively straightforward to implement, requiring neither specification of the first-stage auxiliary models for regressors nor estimation of any joint models for the outcome and regressors. It can be viewed as a nonparametric generalization of the control functions, which nests the existing copula-based control function approaches and offers several substantial advantages over current methods.

First, 2sCOPE-np relaxes the key assumption, imposed by existing copula-based endogeneity correction methods, that the regressors and unobservable (the structural error or its endogenous part) follow a Gaussian copula (GC) dependence structure. 2sCOPE-np is based on a flexible *conditional* copula model that captures the regressor-unobservable dependence not explained by exogenous regressors and nests existing models as special cases. The conditional specification allows for increased flexibility and robustness; unconditionally, the endogenous regressors and unobservables need not follow any pre-specified joint dependence structure, which represents a major relaxation relative to existing approaches that assume particular joint copula models. In addition, the structural error term needs not to follow any specific distributions.

Second, 2sCOPE-np is agnostic to the dependence structure between endogenous and exogenous regressors, as it nonparametrically models the conditional distribution of endogenous regressors given exogenous regressors, eliminating the need to specify or justify first-stage auxiliary models. By contrast, existing copula correction methods either assume independence between endogenous and exogenous regressors or postulate specific first-stage models, thereby restricting functional form relationships (Web Appendix 1). Simulation studies confirm that 2sCOPE-np is the only method that can eliminate endogeneity bias across all regressor dependence structures.

Third, 2sCOPE-np can handle endogenous regressors that are noncontinuous or have limited support[1]. 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.

Lastly, 2sCOPE-np offers a practical and doubly robust alternative in applications where valid IVs are unavailable or where exclusion restrictions or auxiliary model assumptions are difficult to defend. The 2sCOPE-np requires neither knowledge of functional forms of the first-stage model nor the condition that the endogenous regressor contains an exogenous component (observed or latent) satisfying the exclusion restriction. Unlike the popular control function approach using IVs (Petrin and Train 2010), 2sCOPE-np employs model-free control functions, eliminating the need for IVs or assumptions (such as additive separability) about first-stage auxiliary models for endogenous regressors. Simulation studies and

[1]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.

theoretical extensions also demonstrate that the assumptions used to derive 2sCOPE-np are sufficient but not necessary conditions, indicating the robustness of the method.

Overall, the proposed 2sCOPE-np method can substantially increase the robustness, accuracy, and applicability of copula-based endogeneity correction methods. 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 conditional copula modeling framework and develops the 2sCOPE-np procedure. The modeling framework highlights the central role of the conditional CDF of endogenous regressors given exogenous regressors for controlling endogeneity. We show that 2sCOPE-np unifies all existing copula control function approaches and nests them as special cases. In Section 3, we conduct simulations to demonstrate that 2sCOPE-np improves accuracy, robustness, and applicability relative to existing copula control function methods. We then illustrate the usage of our proposed method in Section 4 and conclude with Section 5.

2 Method

2.1 The Regressor Endogeneity Problem

For simplicity, we focus on the following linear structural model with one endogenous regressor in this subsection, though our approach is applicable to nonlinear models (e.g.,

discrete choice models) and to multiple endogenous regressors:

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

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 impose 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 distributional form 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 the error term E_i. [2] Dependence between the exogenous regressors in W_i and the endogenous regressor P_i is allowed.

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. This creates correlation between P and E, causing methods that assume exogeneity—such as ordinary least squares (OLS)—to produce biased estimates that reflect associations rather than causal effects. Regressor endogeneity is pervasive and a central challenge in empirical work with observational data.

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

2.2 The Conditional Copula Endogeneity Model

Formally, we express the joint distribution of (E, P, W)[3] as

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

Since the marginal distribution of the exogenous regressors, $f_{\psi}(W)$, is ancillary to (θ, ϕ), inference about (θ, ϕ) depends only on $f_{\theta,\phi}((E, P)|W)$, the joint distribution of the error term and the endogenous regressor conditional on the exogenous regressors. Because the distribution of the endogenous regressor P carries information about θ via its dependence on the structural error E, single equation methods like OLS, which ignore this dependence, can yield biased estimates. As shown later, a computationally tractable bias correction method augments the model in Equation 1 with a control variable $\mathbb{E}(E|P, W)$, derived from a conditional copula model capturing this regressor-error dependence.

The causal identification challenge often arises from correlation between P and unobserved confounders in the error E. To make the source of endogeneity explicit, write $E = O + \epsilon$, 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 we leave the distribution of ϵ unspecified (e.g., permitting error nonnormality or heteroscedasticity). Because O embodies regressor endogeneity, it suffices to model the joint conditional distribution of (O, P) given W: $f_{\theta,\phi}((O, P)|W)$.

Unlike prior copula endogeneity models, we address the problem using a conditional copula model for $f_{\theta,\phi}((O, P)|W)$ that reveals the central role of the conditional CDF of the endogenous regressor P given the exogenous regressors in W for constructing valid copula control functions. Specifically, 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

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

P conditional on W, respectively. Both $F_{\theta_1}(o|w)$ and $F_\phi(p|w)$ have a uniform distribution on $[0,1]$. 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):

$$F_{\theta,\phi}((o,p)|w) \;=\; C_{\theta_2}(F_{\theta_1}(o|w), F_\phi(p|w)). \tag{3}$$

Equation 3 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 (GC) is also scalable to high-dimensional data, making it practical and computationally tractable.

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 (GC) states the following model:

$$F_{\theta,\phi}((o,p)|w) \;= C_{\theta_2}(F_{\theta_1}(o|w), F_\phi(p|w)) = \Psi_\rho(\Phi^{-1}(u_{o|w}), \Phi^{-1}(u_{p|w})), \tag{4}$$

$$\text{where} \qquad u_{o|w} = F_{\theta_1}(o|w), u_{p|w} = F_\phi(p|w). \tag{5}$$

$\Phi(\cdot)$ 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.

$$\begin{aligned} f_{\theta,\phi}((o,p)|w) \;&= \tfrac{f_{\theta_1}(o|w) f_\phi(p|w)}{(1-\rho^2)^{1/2}} \exp\Big\{ -\tfrac{\rho^2}{2(1-\rho^2)} (\Phi^{-1}(u_{o|w})^2 + \Phi^{-1}(u_{p|w})^2) \\ &\qquad + \tfrac{\rho}{1-\rho^2} \Phi^{-1}(u_{o|w}) \Phi^{-1}(u_{p|w}) \Big\}, \end{aligned} \tag{6}$$

where $f_{\theta_1}(o|w)$ and $f_{\phi}(p|w)$ are univariate conditional density functions of O and P given W, respectively. Unlike prior likelihood-based approaches (Park and Gupta 2012, Haschka 2022), the joint distribution of (E, P) or (E, P, W) is left unspecified and modeling assumptions are relaxed.

The conditional copula framework above generalizes existing copula endogeneity models, while minimizing assumptions about endogenous regressors. Unlike prior copula-based correction methods, Equation 4 constructs the copula *conditional* on W. Thus, it generalizes Park and Gupta (2012) and Christopoulos et al. (2021), which model the *unconditional* joint distribution $F(e, p)$. By modeling the conditional joint distributions $F_{\theta,\phi}((o, p)|w)$, our approach captures endogeneity unexplained by W and relaxes the assumption that P is independent of W. Compared with unconditional joint copula models that jointly specify the dependence structure of (P, W, E) (Haschka 2022, Yang et al. 2025, Lin and Song 2025), conditioning on W improves robustness and reduces computational burden by lowering the number of variables being modeled. Notably, under our framework, neither (P, O) nor (P, E) must follow any specific copula dependence structure unconditionally, and no model for W is required. See Web Appendix Table W1 for a detailed comparison with existing copula correction methods.

2.3 Derivation of Control Functions for Endogeneity

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

$$Y = \mu + p\alpha + w'\beta + \mathbb{E}(o|p, w) + \underbrace{\overbrace{\epsilon + O}^{\text{Structural error } E=O+\epsilon} - \mathbb{E}(o|p, w)}_{\xi}, \tag{7}$$

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

variables $p, w, \mathbb{E}(o|p,w)$ in Equation 7[4]. 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 4 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 2. □

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 model $F_\phi(x|w)$, but instead estimate it nonparametrically. 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):

$$\widehat{F}_\phi^a(p|w) \;=\; \frac{\sum_{i=1}^n I(P_i \le p) K_h(W_i - w)}{\sum_{i=1}^n K_h(W_i - w)}. \tag{8}$$

where $I(P_i \le p)$ is the indicator function for the event $P_i \le p$; n denotes the number of

[4]This holds because $\mathbb{E}(\xi|p,w,\mathbb{E}(o|p,w)) = 0$, which is a sufficient condition for the zero correlation of ξ with each of $(p,w,\mathbb{E}(o|p,w))$ (Wooldridge 2010). Note that $\mathbb{E}(\xi|p,w,\mathbb{E}(o|p,w)) = \mathbb{E}(O - \mathbb{E}(o|p,w) + \epsilon|p,w,\mathbb{E}(o|p,w)) = 0$ 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 $\mathbb{E}(\epsilon|p,w) = 0$.

observations; $K_h(W_i - w)$ is a weight function defined as:

$$K_h(W_i - w) \quad = \quad \frac{1}{h} k\left(\frac{W_i - w}{h}\right), \tag{9}$$

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

$$\widehat{F}^b_\phi(p|w) \quad = \quad \frac{\sum_{i=1}^n G((p - P_i)/h_0) K_h(W_i - w)}{\sum_{i=1}^n K_h(W_i - w)} \tag{10}$$

(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.

In practice, endogenous regressors are often conceptually continuous but measured or recorded coarsely, leading to discrete marginals. In the return to education example in Web Appendix 10, 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. Equation 10 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.

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 5).

2.4 Model Identification

Table 1 lists the key assumptions to derive 2sCOPE-np and to ensure consistent estimation and model identification, noting that the 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 are sufficient conditions for deriving 2SCOPE-np, but they are not necessary (Web Appendix 6). Furthermore, although Assumptions 1 and 2 are used to derive 2sCOPE-np, they are not strictly required (Web Appendix 7).

Table 1: Assumptions to Derive 2sCOPE-np

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 regressors $(1, P, W)$.*

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 6).

Assumption 1 means that the structural error E need not be normally distributed (Web Appendix 6). The assumption holds as long as one of E and O is normal[5]. 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 due to 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. Furthermore, 2sCOPE-np is robust to nonnormality of O (Proposition 3 and Web Appendix 7).

Assumption 2 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. Furthermore, the conditional copula model permits both

[5]The normality requirement satisified by O can be further relaxed by Proposition 3.

GC and non-GC types of regressor-error dependence unconditionally, as will be further demonstrated in the *Simulation Studies* section. Assumption 2 is also plausible 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 [6]. Assumption 2 is also not strictly required; even conditional on W, (P, O) is not required to follow a GC dependence structure (Web Appendix 7) for the proposed method to work.

Prior studies (Haschka 2022, Qian and Xie 2024) show that grossly mis-specifying regressor-error dependence can lead to weakly or non-identifications with huge standard errors (often $>$ 8-10 times those of the uncorrected estimates). Thus, checking for large standard-error inflation is recommended when applying copula correction.

Assumption 3 ensures full column rank of the augmented regressor matrix $(1, P, W, C_p)$ for model identification, preventing perfect multicollinearity. Since C_p is generated from existing regressors, it may be nearly collinear with $(1, P, W)$, inflating standard errors. In the extreme 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).

Consistency of the two-step estimation procedure requires consistency of the plug-in estimates obtained from the first-stage estimation(Wooldridge 2010). For 2sCOPE-np, we require the consistency of nonparametric kernel estimates of the conditional CDF, $\widehat{F}_{\phi}(P|W)$,

[6] Section 3.3 provides one such example.

in Equations 8 and 10. Li and Racine (2008) shows that, under standard regularity conditions, these kernel estimators provide consistent estimates of conditional CDFs. An especially useful property of these kernel CDF estimators is their built-in ability to automatically perform variable selection and remove irrelevant W variables that do not affect the conditional distribution (Li and Racine 2008). This property results in impressive performance of nonparametric CDF estimation in high-dimensional settings, where in practical applications, models often include covariates that turn out to be irrelevant.

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 endogenous regressors $(P_1, \cdots, P_K)$,

$$Y_i = \mu + \sum_{k=1}^{K} P_{i,k}\alpha_k + W_i'\beta + E_i. \tag{11}$$

We extend the bivariate conditional GC model in Equation 4 to the following multivariate conditional GC model with correlation matrix Σ:

$$F_{\theta,\phi}((o, p_1, \cdots, p_K)|w) = C_{\theta_2}(F_{\theta_1}(o|w), F_{\phi_1}(p_1|w), \cdots, F_{\phi_K}(p_K|w)), \quad (12)$$

$$= \Psi_\Sigma(\Phi^{-1}(u_{o|w}), \Phi^{-1}(u_{p_1|w}), \cdots, \Phi^{-1}(u_{p_K|w})), \quad (13)$$

$$\text{where} \quad u_{o|w} = F_{\theta_1}(o|w), u_{p_k|w} = F_{\phi_k}(p_k|w), \ k = 1, \cdots, K. \quad (14)$$

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 3. □

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

$$Y_i = \mu + \sum_{k=1}^{K} P_{i,k}\alpha_k + X_i^T\beta + \beta' W_i + \sum_{k=1}^{K} C_{i,p_k}\gamma_k + \xi_i, \quad (15)$$

Table 2 summarizes the 2sCOPE-np estimation procedure. For multiple endogenous regressors, the above 2sCOPE-np only requires performing the kernel conditional CDF estimation for each endogenous regressor one by one, without the need to perform a joint conditional CDF estimation for all endogenous regressors simultaneously. Thus, the computational workload for 2sCOPE-np only increases linearly with the number of endogenous regressors and is scalable to multiple endogenous regressors.

Table 2: 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 8 or Equation 10) 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 15.

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 the data can be divided into J heterogeneous samples indexed by sample index j. Then Proposition 2 can be extended into the following form.

Proposition 3. *Suppose that the data can be partitioned into J heterogeneous samples. Assume that conditions in Table 1 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 $, \Sigma_j = \begin{pmatrix} 1 & \Sigma_{O_j P_j} \\ \Sigma_{O_j P_j}^\top & \Sigma_{P_j P_j} \end{pmatrix}$, and full column rank for identification. Then*

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

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

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

Proof. See Web Appendix 4. □

Proposition 3 allows the GC correlation 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 [7] and need not follow a normal distribution [8]. According to Proposition 3, the augmented regression adds J copula generated terms, $I(w \in j)C_{p_{jk}}, j = 1, \cdots, J$, for each endogenous regressor P_k.

2.8 Summary

In summary, 2sCOPE-np generalizes and unifies existing copula-based control functions approaches (see Web Appendix 1 for details). 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 way of extending the P&G method (Park and Gupta 2012) to account 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, Breitung et al. 2024, Mayer and Wied 2025, Liengaard et al. 2024, Yang et al. 2024). 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 noparametrically without postulating specific first-stage auxiliary regressors models. Furthermore, Proposition 3 shows that, in heterogeneous samples, 2sCOPE-np nests several existing approaches (Christopou-

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

[8] 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 $\mathbb{E}(\epsilon|P, W) = 0$.

los et al. 2021, Liengaard et al. 2024, Yang et al. 2024) as special cases.

Relative to likelihood-based alternatives (Haschka 2022, Qian and Xie 2024, Lin and Song 2025), the 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, computational 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 supports established likelihood-based tests and model comparisons (Qian and Xie 2024, Lin and Song 2025).

3 Simulation Studies

2sCOPE-np makes minimal assumptions regarding the first-stage model and accommodates broader types of dependence structures and endogenous regressors. This section presents simulation studies to demonstrate that the proposed 2sCOPE-np corrects for endogeneity under different data settings. We compare 2sCOPE-np with existing copula control function methods simulating data from five underlying models. The first three evaluate the performance of 2sCOPE-np for continuous endogenous regressors. The last two cases test whether 2sCOPE-np can correct for endogeneity when the endogenous regressor is a binary or count variable with limited support, a setting in which existing copula control function methods are known to fall short. We expect 2sCOPE-np to outperform existing copula control function methods, due to its generality and nonparametric first-stage estimation. 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.

3.1 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 using the following data generating process (DGP).

$$\begin{pmatrix} P_i^* \\ W_i^* \\ E_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix}\right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}\right), \quad (16)$$

$$E_i = G^{-1}(U_{E,i}) = G^{-1}(\Phi(E_i^*)) = \Phi^{-1}(\Phi(E_i^*)) = 1 \cdot E_i^*, \quad (17)$$

$$P_i = H^{-1}(U_{P,i}) = H^{-1}(\Phi(P_i^*)), \;\; W_i = L^{-1}(U_{W,i}) = L^{-1}(\Phi(W_i^*)). \quad (18)$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 1 + 1 \cdot P_i + (-1) \cdot W_i + E_i. \quad (19)$$

The interest is to estimate the structural model in Equation 19 using simulated data, where we know the true parameter values and can evaluate the accuracy of parameter estimates. The simulated data has an endogeneity problem 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 $H(P)$ 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. We set $L(W)$ to be the CDF of the standard normal distribution. Both regressor-error and regressor-regressor ($P - W$) follow GC dependence structure.

We consider four different estimation methods: (1) OLS regression of Equation (19); (2) 2sCOPE-GC; (3) 2sCOPE-MD; and (4) the proposed 2sCOPE-np. We expect the OLS regression to yield biased estimates because it ignores the regressor endogeneity problem. The three correction methods in (2)-(4) aim to correct for endogeneity bias. All three copula correction methods involve two steps, in which the second step estimates the following

augmented OLS regression

$$Y_i \quad = \quad \mu + \alpha P_i + \beta W_i + \gamma \mathtt{C}_{i,p} + \xi_i, \tag{20}$$

where $\mathtt{C}_{i,p}$ estimated in the first step denotes the control function. These three methods differ in how the control function $\mathtt{C}_{i,p}$ is estimated in the first step, as described below.

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

$$\mathtt{C}_{i,p} \quad = \quad P_i^* - \hat{\delta} W_i^*, \tag{21}$$

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 the 2sCOPE-GC procedure makes no assumptions about the marginal distributions of regressors (and thus can accommodate bounded regressors automatically) [9], and permits 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,

$$\mathtt{C}_{i,p} \quad = \quad \Phi^{-1}(\widehat{F}_v(\widehat{V}_i)), \tag{22}$$

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

[9] In the results presented below, 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.

residuals from this first stage regression. Following Mayer and Wied (2025), 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 9). 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, $F_v(P_i - g(W_i)) \neq F_\phi(P_i|W_i)$, breaking the mean-dependence-only assumption and introducing bias in the endogeneity correction. Although the violation of this assumption could occur for any types of variables, 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,

$$\mathtt{C}_{i,p} \;=\; \Phi^{-1}(\widehat{F}_\phi(P_i|W_i)), \tag{23}$$

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 10 that smooths both P and W.

We generate 1000 replicate datasets with sample size of 1000 each using the above DGP,

and apply the four estimation methods to all generated datasets producing the parameter estimates' empirical distributions.

Table 3 reports the results. The columns for "Bias" in Table 3 compute the differences between the true parameter values and the average parameter estimates of all 1000 simulated data samples for each method. The columns for "SE" compute the standard deviations of these parameter estimates resulting from all 1000 simulated datasets, and measure how much each estimator varies from sample to sample. The columns for "RMSE" provide comprehensive measures of the overall performance of an estimator and are computed as $\sqrt{Bias^2 + SE^2}$. RMSE informs the expected absolute difference between the estimator and the corresponding true parameter values, taking into account both the bias and the variability of the estimator.

Table 3: 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: The 2sCOPE-MD implements the method of Mayer and Wied (2025) and the 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 3). 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 com-

parable) RMSEs. Although 2sCOPE-MD reduces bias relative to OLS, substantial bias remains (e.g., bias=0.385 for α in Table 3). 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 the 2sCOPE methods. In comparison, 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.

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

In this simulation, we consider the following DGP.

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

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

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

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

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 the omitted variable O_i according to Equation 25, and consequently is an endogenous regressor. 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) jointly follow a GC dependence structure. We then apply the same four methods: (1) OLS; (2)

2sCOPE-GC; (3) 2sCOPE-MD; and (4) 2sCOPE-np, as described previously. We generate 1000 datasets, each with a sample size of 1000. We apply these four estimation methods to all generated datasets, producing the parameter estimates' empirical distributions.

Table 4: 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 3.

Results reported in Table 4 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 the 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.

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

This simulation examines a case, in which the DGP follows neither the 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 standard 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 the 2sCOPE-np assumption, but (P, W) does not follow a joint GC structure (see Web Appendix 8 for proof), thus violating the assumption of 2sCOPE-GC. Higher moments of P depend on W violating the mean-dependence-only assumption.

We apply the same four methods: (1) OLS; (2) 2sCOPE-GC; (3) 2sCOPE-MD; and (4) 2sCOPE-np, as described previously. We generate 1000 datasets as replicates, each with a sample size of 1000. We apply these four estimation methods to all generated datasets, to obtain the empirical distributions of the parameter estimates.

Table 5: 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 3.

As expected, without endogeneity correction, OLS results in a large bias of 49.2% for the coefficient α (Table 5). 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.

3.4 Case 4: Count Endogenous Regressor with few unique values

In this and the next subsections, we present the simulation studies where the focal endogenous regressor is count or binary with limited support, comparing 2sCOPE-np with existing copula control function methods. As noted in the *Introduction*, existing copula 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 regression equation for Y_i (Equation (37)), thereby creating endogeneity.

We generate 1000 replicate datasets of sample size of 2000 each using the DGP. We apply the four estimation methods:(1) OLS; (2) 2sCOPE-GC; (3) 2sCOPE-MD; and (4) 2sCOPE-np, as described previously, to all generated datasets, producing the parameter estimates' empirical distributions.

Table 6: 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 3.

Results in Table 6 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 6). 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.

3.5 Case 5: 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.

$$\begin{pmatrix} P_i^* \\ W_i^* \\ E_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right), \quad (38)$$

$$E_i = G^{-1}(U_{E,i}) = G^{-1}(\Phi(E_i^*)) = \Phi^{-1}(\Phi(E_i^*)) = 1 \cdot E_i^*, \quad (39)$$

$$P_i = \mathbb{I}_{\{U_{P,i} > p_1\}} = \mathbb{I}_{\{\Phi(P_i^*) > p_1\}} = \mathbb{I}_{\{P_i^* > \Phi^{-1}(p_1)\}}, \quad (40)$$

$$W_i = L^{-1}(U_{W,i}) = L^{-1}(\Phi(W_i^*)), \quad (41)$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 0 + 1 \cdot P_i + 2 \cdot W_i + E_i. \quad (42)$$

where

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

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: (1) OLS regression; (2) 2sCOPE-GC; (3) 2sCOPE-MD; and (4) 2sCOPE-np, as described previously, to all generated datasets, producing the estimates' empirical distributions.

Without endogeneity correction, OLS yields large biases (91.2%, Table 7) 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 7). As expected, 2sCOPE-GC and 2sCOPE-MD show significant bias when applied to binary endogenous regressors.

Table 7: 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 3.

3.6 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.

4 Empirical Applications

We illustrate the 2sCOPE-np procedure in two empirical applications: the marketing problems on estimating price elasticity, where the focal endogenous regressor price is continuous, and the labour economics problem of estimating return to education (Web Appendix 10),

where the focal endogenous regressor can be count or binary. We discuss the motivation for using 2sCOPE-np, steps to conduct estimation and inference, how to check verifiable data requirements, and how to interpret results in the empirical application. In addition to IV methods, we also benchmark the performance of 2sCOPE-np using 2sCOPE-GC and 2sCOPE-MD, consistent with the *Simulation Studies* section.

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:

$$\ln \texttt{sales}_t = \mu + \alpha P_t + \beta' W_t + E_t, \tag{43}$$

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. While 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 (IV) 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, after controlling for pre-planned feature activity and the quarterly trend, it appears reasonable to capture the remaining dependence between the joint effect of contemporaneous omitted variables and the price variable using conditional copula model.

The 2sCOPE estimates the following augmented regression model to correct price endogeneity

$$\ln \texttt{sales}_t \quad = \quad \mu + \alpha P_t + \beta' W_t + \gamma C_{price,t} + \xi_t, \tag{44}$$

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

Figure W1 in Web Appendix 9 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 can 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) using 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 W2 in Web Appendix 9 presents the CDF estimates of logPrice conditional on feature intensity only (the top one) 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

44 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 8 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 8: Estimation Results for Price Elasticity in Store Sales Data.

	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	6.87	0.08	92.04	7.06	0.09	74.89	7.15	0.16	43.62	7.23	0.17	42.61	7.50	0.13	55.79
Price	**-1.82**	**0.12**	**-15.57**	**-2.13**	**0.15**	**-14.13**	**-2.30**	**0.27**	**-8.51**	**-2.43**	**0.28**	**-8.78**	**-2.86**	**0.22**	**-12.98**
Feature	0.26	0.06	4.05	0.13	0.08	1.76	0.07	0.10	0.68	0.02	0.12	0.20	-0.18	0.09	-1.90
Week	0.00	0.00	1.76	0.00	0.00	2.01	0.00	0.00	2.24	0.00	0.00	2.53	0.00	0.00	2.53
Q_2	-0.06	0.03	-2.26	-0.06	0.03	-2.32	-0.06	0.03	-2.03	-0.07	0.03	-2.31	-0.07	0.03	-2.39
Q_3	-0.08	0.03	-2.80	-0.09	0.03	-3.18	-0.09	0.03	-2.93	-0.10	0.03	-3.20	-0.14	0.02	-4.80
Q_4	-0.05	0.03	-1.70	-0.05	0.03	-1.71	-0.04	0.02	-1.56	-0.04	0.02	-1.79	-0.04	0.03	-1.35
C_{price}							**0.07**	**0.03**	**3.01**	**0.10**	**0.02**	**4.12**	**0.10**	**0.01**	**10.91**
Wu-Hausman test				**H=11.45, p=0.0004**											

Note: Table presents coefficient estimates, bootstrapped standard errors (SE), and t-values.

In comparison, the 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. The 2sCOPE-GC price estimate deviates more, differing from 2SLS by roughly the same margin as OLS does. The 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 W1. 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 control function term, agreeing on the presence of endogeneity bias. In particular, testing for the

copula correction 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.

2SLS price estimate has a smaller standard error (0.15) compared with that of the 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 the 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.

Example 2: Return to Education (the case of weak IV). In Web Appendix 10, we apply 2sCOPE-np to estimate the return to education, operationalized either as a count variable (years of education) or a binary variable (received post-secondary education or not).

5 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. By replacing the marginal CDFs of regressors with the 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 to handle 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. Only 2sCOPE-np can eliminate endogeneity bias under all considered data settings. This shows that 2sCOPE-np greatly increases the robustness and accuracy of endogeneity correction, broadening the applicability of copula correction without restricting to GC regressor-error dependence or needing to specify a first-stage model for endogenous regressors.

We apply 2sCOPE-np to estimate price elasticity in the peanut butter category using historical store sales data, and to estimate the return to education. The analysis shows 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 and the practical importance of its ability to handle noncontinuous endogenous regressors and to guard against bias from auxiliary model mis-specifications.

The 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. Future research could extend 2sCOPE-np to other types of models and data, such as discrete choice models for multinomial outcomes. Numerous opportunities exist for empirical applications of this method.

References

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

Ananthakrishnan UM, Xu C, Mitkina M, Tan Y (2025) Political consumerism and the emergence of rare information on user-generated content platforms. *Management Science, Forthcoming, Available at SSRN: https://ssrn.com/abstract=5237716* .

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.

Becerra M, Markarian G (2021) Why are firms with lower performance more volatile and unpredictable? a vulnerability explanation of the bowman paradox. *Organization Science* 32(5):1327–1345.

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.

Blinded (2025) *Journal of Marketing* .

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.

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.

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

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.

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.

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 (2024) Dealing with regression models endogeneity by means of an adjusted estimator for the gaussian copula approach. *Journal of the Academy of Marketing Science* 1–21.

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

Mayer A, Wied D (2025) Endogeneity corrections in binary outcome models with nonlinear transformations: Identification and inference. *Oxford Bulletin of Economics and Statistics* .

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, 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 G (2003) Bayesian analysis of simultaneous demand and supply. *Quantitative Marketing and Economics* 1:251–275.

Web Appendix 1: Related Literature

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 W1. The original copula-based endogeneity correction method was introduced by Park and Gupta (2012). Like Park and Gupta (2012), other earlier work (Tran and Tsionas 2021, Christopoulos et al. 2021, Becker et al. 2021, Eckert and Hohberger 2023) also assume independence between P and W and require nonnormal and continuous endogenous regressors. 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. As reviewed below, our proposed approach falls under the control function form extension, unifying and generalizing all existing copula control function methods.

1.A 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. [10]

[10]Recent extensions of 2sCOPE-GC (Liengaard et al. 2024, 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 W4) to vary across heterogenous subsamples (Proposition 3).

Mayer and Wied (2025) and Breitung et al. (2024) proposed using a mean regression auxiliary model for the endogenous regressor to estimate the conditional distribution of P 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 [11] 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 W1). This assumption is known to be violated for bounded or truncated variables, but may also be questionable for any continuous endogenous regressors.[12] 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 GC regressor-error dependence structure (Table W1). Our proposed method unifies and generalizes these two methods (2sCOPE-GC and 2sCOPE-MD), making copula correction more robust.

1.B 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)

[11] This correlation coefficient is defined right above Equation 4.

[12] 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. The product pricing and promotion can be bounded by minimum and maximum values (Qian and Xie 2011). 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). Although category price does not have explicit bounds, our empirical application later also shows that the exogenous regressors affect more than just the mean of the category price distribution.

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 W1). 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.

1.C Comparison among methods

The proposed 2sCOPE-np method unifies and nests all other methods, except for SORE in Table W1, 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

Table W1: 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
Christopoulos et al. (2021)	CF	GC TC	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. (2024) (2sCOPE-GC)	CF	Yes	No	GC	Yes	No
Breitung et al. (2024) Mayer and Wied (2025) (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 (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 (endogenous regressors) and W (exogenous regressors) 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.

regressor-error dependence (GC or non-GC) and eliminates first-stage modeling of endogenous regressors (Table W1).

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[13]. Consequently, 2sCOPE-np nests all existing copula-based endogeneity correction methods as special cases, and has increased robustness and broader applicability

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

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.

Web Appendix 2: Proof of Proposition 1

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

$$F_{\theta,\phi}((o,p)|w) = C_{\theta_2}(F_{\theta_1}(o|w), F_\phi(p|w)), \tag{W1}$$

$$= \Psi_\rho(\Phi^{-1}(u_{o|w}), \Phi^{-1}(u_{p|w})) \tag{W2}$$

$$\text{where} \quad u_{o|w} = F_{\theta_1}(o|w), u_{p|w} = F_\phi(p|w), \tag{W3}$$

$\Psi_\rho(\cdot,\cdot)$ 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,

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

Web Appendix 3: Proof of Propositions 2

Under the conditional GC model, $(\Phi^{-1}(u_{o|w}), \Phi^{-1}(u_{p_1|w}), \cdots, \Phi^{-1}(u_{p_K|w}))$ follow the standard multivariate normal distribution:

$$\begin{pmatrix} \Phi^{-1}(u_{o|w}) \\ \Phi^{-1}(u_{p_1|w}) \\ \cdot \\ \Phi^{-1}(u_{p_K|w}) \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & \rho_{op_1} & \cdot & \rho_{op_k} \\ \rho_{op_1} & 1 & \cdot & \rho_{p_1 p_k} \\ \cdot & \cdot & 1 & \cdot \\ \rho_{op_k} & \rho_{p_1 p_k} & \cdot & 1 \end{bmatrix} \right). \tag{W4}$$

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 (W4), 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 $\Sigma_{OP} = (\rho_{op_1}, \ldots, \rho_{op_K})$ 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 $c := Z_P\big|_{P=p} = \big(C_{p_1}, \ldots, C_{p_K}\big)^\top$ denote the observed values of the latent Z_P components where,

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

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 $O = \sigma_o Z_O$,

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

Writing the $1 \times K$ vector $\sigma_o\, \Sigma_{OP}\, \Sigma_{PP}^{-1}$ componentwise as $(\gamma_1, \ldots, \gamma_K)$, 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

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

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

Thus,

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

Web Appendix 4: Proof of 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

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

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

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

Web Appendix 5: Multiple regressors: tradeoff of 2sCOPE-np between flexibility and efficiency

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

$$\begin{pmatrix} E_i^* \\ P_{1i}^* \\ W_{1i}^* \\ P_{2i}^* \\ W_{2i}^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pe} & 0 & \rho_{pe} & 0 \\ \rho_{pe} & 1 & \rho_{pw} & \rho_{p_1p_2} & \rho_{pw} \\ 0 & \rho_{pw} & 1 & 0 & \rho_{w_1w_2} \\ \rho_{pe} & \rho_{p_1p_2} & 0 & 1 & \rho_{pw} \\ 0 & \rho_{pw} & \rho_{w_1w_2} & \rho_{pw} & 1 \end{bmatrix} \right), \tag{W5}$$

(W6)

$$\rho_{pe} = 0.2, \rho_{pw} = 0.5, \rho_{w_1w_2} = 0.2, \rho_{p_1p_2} = 0.2 \tag{W7}$$

$$E_i = G^{-1}(U_{E,i}) = G^{-1}(\Phi(E_i^*)) = \Phi^{-1}(\Phi(E_i^*)) = 1 \cdot E_i^*, \tag{W8}$$

$$P_{ji} = H_j^{-1}(\Phi(P_{ji}^*)), \quad W_{ji} = L_j^{-1}, (\Phi(W_j i^*))' \quad j = 1, 2. \tag{W9}$$

$$Y_i \quad = \mu + \alpha_1 \cdot P_{1i} + \alpha_2 \cdot P_{2i} + \beta_1 \cdot W_{1i} + \beta_2 \cdot W_{2i} + E_i \tag{W10}$$

$$= 1 + 1 \cdot P_{1i} + 1 \cdot P_{2i} + (-1) \cdot W_{1i} + (-1) \cdot W_{2i} + E_i. \tag{W11}$$

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 W2: 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: The 2sCOPE-MD implements the method of Mayer and Wied (2025) and the 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 W2 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. The 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 [14] and higher computational cost), which is expected for any nonparametric estimator.

Web Appendix 6: 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 1) and that GC captures the joint distribution of (P, O) or (P, E) given W (Assumption 2 Table 1). 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 Gaus-

[14] The 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 3. The result for 2sCOPE-GC is similar regardless of estimation methods.

sian copula model and E_i is not normally distributed. The correlation between P_i and O_i is non-zero, generating the endogeneity problem.

$$\begin{pmatrix} P_i^* \\ W_i^* \\ O_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix}\right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}\right), \tag{W12}$$

$$O_i = (\Phi(O_i^*)) = \Phi^{-1}(\Phi(O_i^*)) = O_i^*, \tag{W13}$$

$$P_i = F^{-1}(\Phi(P_i^*)) \tag{W14}$$

$$W_i = G^{-1}(\Phi(W_i^*)). \tag{W15}$$

$$\epsilon_i \sim H \text{ as specificied in Table W3} \tag{W16}$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + O_i + \epsilon_i = 0 + 1 \cdot P_i + 2 \cdot W_i + O_i + \epsilon_i. \tag{W17}$$

Table W3: 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.

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.

Simulation results reported in Table W3 show that although the error term E_i is non-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.

Web Appendix 7: Robustness to Departure from Assumption 1 and 2

In our proposed model, we impose a normality assumption and a Gaussian copula assumption (Assumptions 1 and 2 in Table 1) 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.

$$\begin{pmatrix} P_i^* \\ W_i^* \\ E_i^* \end{pmatrix} \sim t_2 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix} \right), \tag{W18}$$

$$P_i = H^{-1}(F(P_i^*)) \text{ where } H \text{ is the cdf of a Gamma (1,1) distribution ,} \tag{W19}$$

$$W_i = L^{-1}(F(W_i^*)) \text{ where } L \text{ is the cdf of a Exponential (1) distribution,} \tag{W20}$$

$$E_i = G^{-1}(F(E_i^*)) \text{ where } G \text{ is the cdf of a Uniform[-0.5, 0.5] distribution} \tag{W21}$$

$$F \text{ is the cdf of the student-t(2) distribution.} \tag{W22}$$

Simulation results presented in Table W4 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 W4: 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.

Web Appendix 8: Proof that (P, W) does not follow GC in simulation case 3

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.

Web Appendix 9: Box plot of Log Price and Kernel Estimates of CDFs

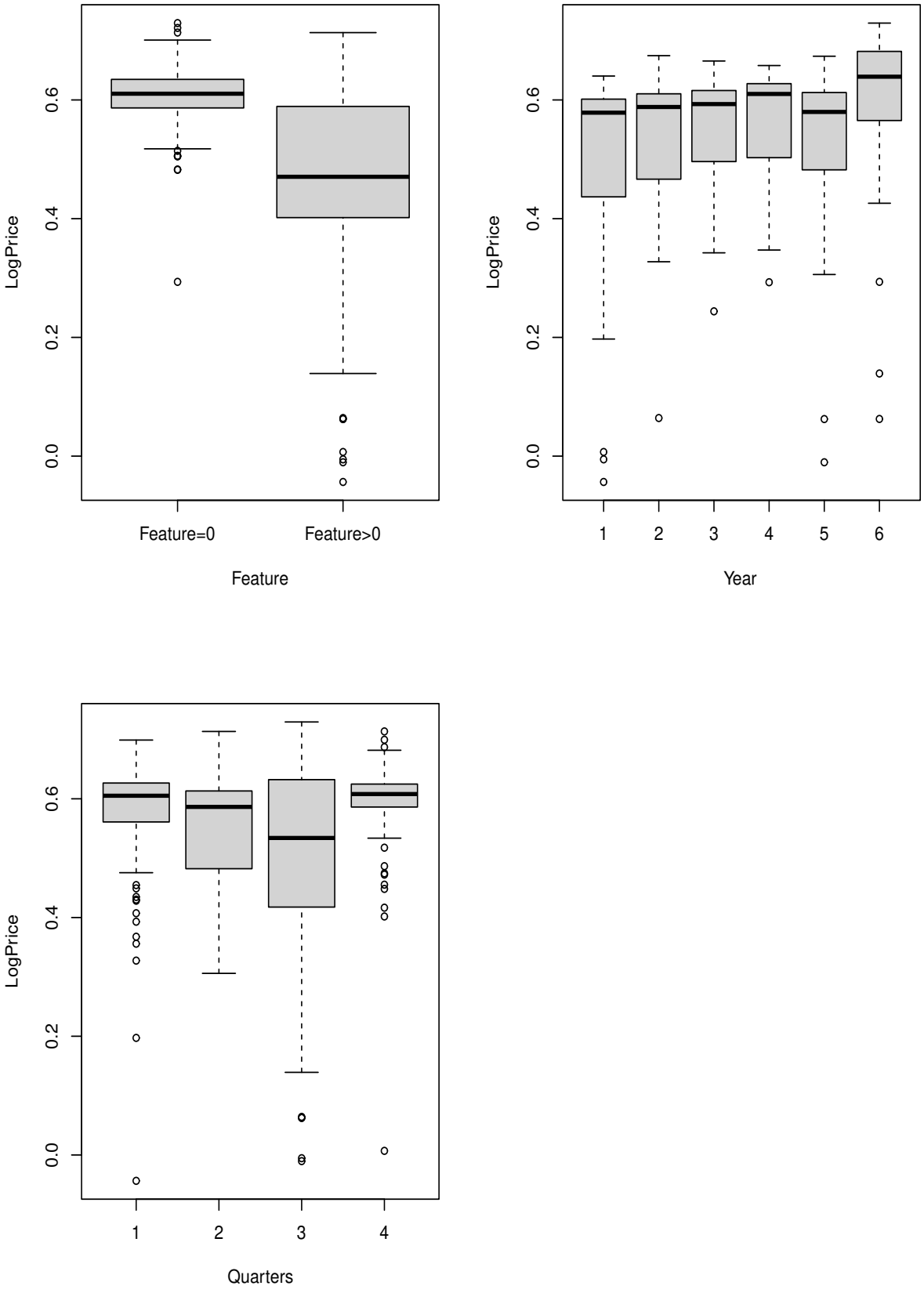


Figure W1: Boxplots of LogPrice in Store 1.

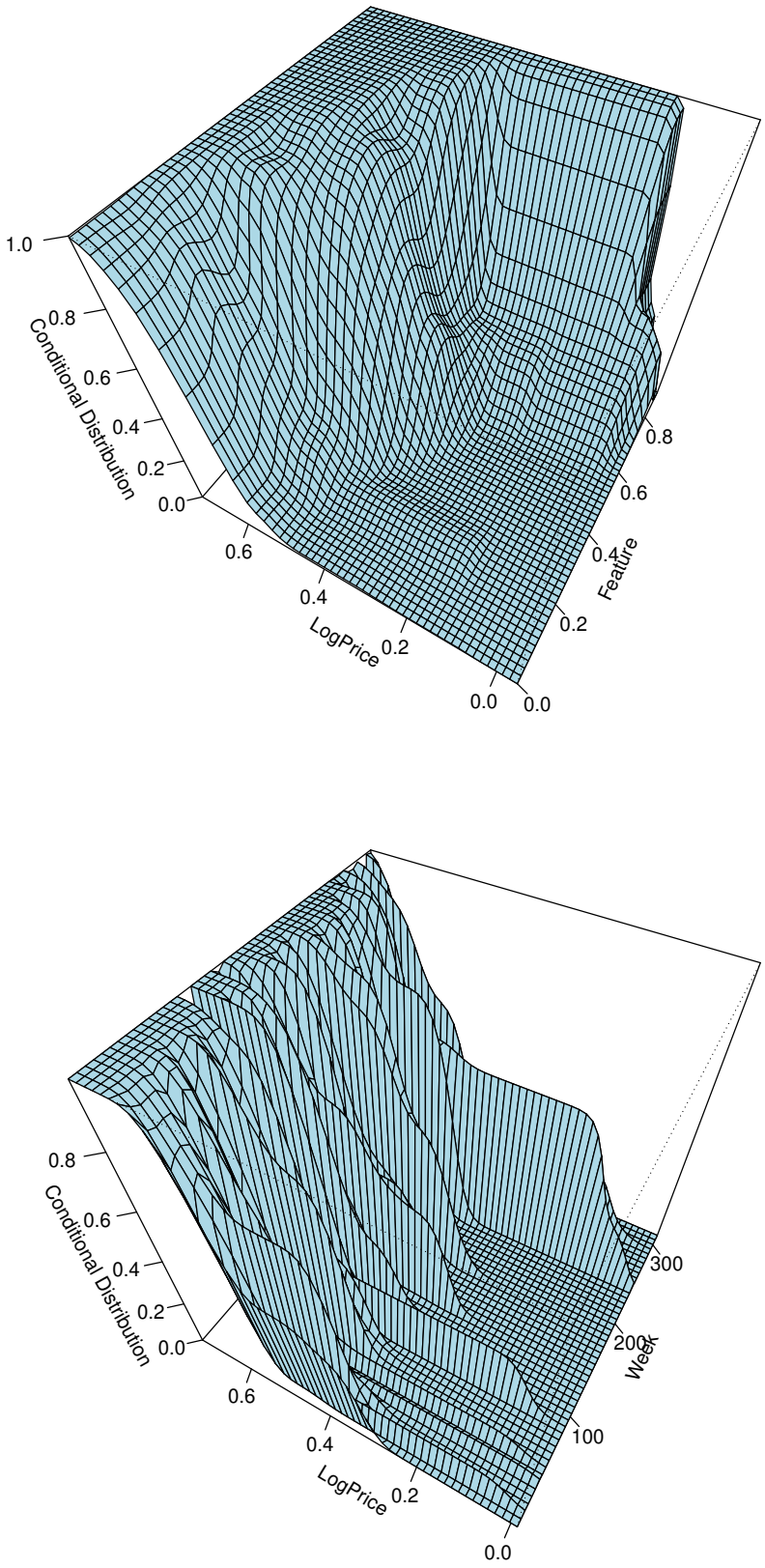


Figure W2: Kernel Estimates of CDFs of logPrice Conditional on Feature (top) and Week (Bottom) in Store 1.

Web Appendix 10: Empirical 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; here, postsecondary education is a binary variable.

The decision to have more years of education or to pursue postsecondary education is known to be endogenous. Education decisions correlate with unobserved traits that may also affect wages, such as motivation and the individual's ability [15]. These variables are often unobserved by the researcher, resulting in endogeneity due to omitted variables. If more education acts as a selection mechanism that successfully selects individuals with higher motivation or abilities, then these individuals might be able to earn a higher wage at any educational level. Without correcting for endogeneity, OLS can overestimate the return to education. Another source of bias may be the measurement errors in the education variable. Measurement error can lead to a negative correlation between schooling and the structural error of the earnings equation, resulting in an underestimation of the return to education (Card 1999). Due to the multiple potential sources and directions of bias, it is important to correct for endogeneity and account for potential bias.

[15] An individual's ability is often the result of numerous genetic traits and environmental factors. When measured across a large representative population, it is reasonable to assume that the population's abilities follow a normal distribution according to the Central Limit Theorem.

To study the return to education, 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. We are interested in estimating the following model:

$$\ln \texttt{wage}_t = \mu + \alpha Education_t + \beta' W_t + E_t, \tag{W23}$$

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 or a binary indicator of postsecondary education. 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 such as whether the region is in the South in 1966 and at the time of the survey, etc. Encoding of the variable names can be found in Table W7 (Web Appendix 11).

To correct for potential endogeneity bias, we consider both IV and IV-free methods: 2sCOPE-GC, 2sCOPE-MD and 2sCOPE-np. For the IV approach, we apply the commonly used instrumental variable, 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. In particular, we bootstrap the error for all 2sCOPE methods and check data requirements and model identification for 2sCOPE-np. We verify the data requirement of no collinearity between the copula terms and the original regressors. For the 2sCOPE-np analyses reported in Tables W5 and W6, the R^2 for the regressions that regress the generated copula terms on the other regressors are 0.77

and 0.44, respectively. This indicates that copula terms in the two 2sCOPE augmented models are not linearly dependent of existing regressors. The inflation of standard errors of the 2sCOPE-np corrected estimates relative to the corresponding OLS estimates is well below the threshold value of 6 (Tables W5 and W6), indicating no signs of weak model identification.

10.A Results: a count regressor for education

To examine the effect of an additional year of education, we use the education variable as a count variable (years of education), and estimate the return to education using OLS, 2SLS with IV, and the three 2sCOPE methods.

As reported in Table W5, 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-GC (Est.=0.05, SE=0.01), and 2sCOPE-MD (Est.=0.04, SE=0.01)[16]. 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:

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

0.05, 0.08), only slightly lower than the OLS estimate, and is much more precise than 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 W5). 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 statistical significant coefficient for the control variable motheduc in Table W5)[17].

Table W5: Estimation Results for Count Education Variable

	OLS			2SLS			2sCOPE-np			2sCOPE-GC			2sCOPE-MD		
Parameters	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value
(Intercept)	4.64	0.08	55.32	3.79	0.74	5.13	4.68	0.13	36.92	5.02	0.14	36.93	5.04	0.19	26.53
educ	**0.07**	**0.00**	**19.67**	**0.13**	**0.05**	**2.51**	**0.07**	**0.01**	**9.15**	**0.05**	**0.01**	**5.08**	**0.04**	**0.01**	**3.05**
exper	0.09	0.01	12.71	0.11	0.02	5.36	0.08	0.01	12.09	0.07	0.01	10.14	0.08	0.01	9.78
expersq	0.00	0.00	-7.23	0.00	0.00	-6.86	0.00	0.00	-6.45	-0.00	0.00	-7.20	-0.00	0.00	-6.39
black	-0.19	0.02	-9.78	-0.16	0.03	-4.88	-0.19	0.02	-9.58	-0.20	0.02	-10.02	-0.20	0.02	-10.19
fatheduc	0.00	0.00	0.61	0.00	0.01	-0.40	0.00	0.00	0.64	0.01	0.01	1.00	0.01	0.00	1.00
motheduc	0.01	0.00	2.03	0.00	0.01	-0.06	0.01	0.00	2.09	0.01	0.00	2.64	0.01	0.00	2.54
fatheduc × motheduc	0.00	0.00	-0.92	0.00	0.00	-0.99	0.00	0.00	0.00	-0.00	0.00	-0.87	-0.00	0.00	-0.88
Missing: fatheduc	0.01	0.04	0.25	-0.07	0.08	-0.91	0.01	0.03	0.26	0.04	0.04	1.07	0.05	0.04	1.12
Missing: motheduc	0.11	0.04	2.78	0.05	0.07	0.65	0.11	0.04	2.73	0.14	0.04	3.39	0.14	0.04	3.61
momdad14	0.07	0.04	1.91	0.02	0.05	0.48	0.07	0.03	1.94	0.08	0.03	2.44	0.09	0.03	2.57
sinmom14	0.04	0.04	0.90	0.04	0.04	0.91	0.04	0.04	0.90	0.04	0.04	0.84	0.04	0.04	0.85
step14	0.00	0.05	-0.03	0.00	0.05	0.07	0.00	0.05	-0.04	-0.00	0.05	-0.10	0.00	0.05	-0.08
$C_{education}$							**0.01**	**0.02**	**0.40**	**0.06**	**0.02**	**3.35**	**0.06**	**0.03**	**2.22**
Wu-Hausman test				**H=1.48, p=0.22**											

Note: dummy variables related to location omitted from display. Standard errors of parameter estimates are estimated using bootstrap.

The other two IV-free methods, 2sCOPE-GC and 2sCOPE-MD, theoretically cannot handle count regressors with limited support. In this dataset, 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 that there is no endogeneity bias, 2sCOPE-GC and 2sCOPE-MD yield positive and significant control function coefficients, suggesting a positive endogeneity bias in the

[17]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.

OLS estimate, which may result from the inappropriateness of using the two methods for count regressors with limited support.

10.B Results: 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 W6, 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 (ρ = 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 [18]. This explains why the 2SLS result is statistically no different from that of OLS, despite the evidence of endogeneity.

[18]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, the 2sCOPE-np estimation result has the advantage of being more accurate. The estimated return to log wage using 2sCOPE-np is 0.36[19] 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 the 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-

Table W6: 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
$C_{education}$							**-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 W5

np and 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). In contrast, both 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 problem-

[19]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.

atic, 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.

Web Appendix 11: Variable definition for the education example

Table W7: 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
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

Variable name	Description
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.