Web Appendix 1: Proof of proposition 1

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

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

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

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

$\Psi_\rho(\cdot,\cdot)$ 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 2: 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:

$$\begin{pmatrix} \Phi^{-1}(u_{o|w}) \\ \Phi^{-1}(u_{p_1|w}) \\ \cdot \\ \Phi^{-1}(u_{p_K|w}) \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & \rho_{op_1} & \cdot & \rho_{op_k} \\ \rho_{op_1} & 1 & \cdot & \rho_{p_1p_k} \\ \cdot & \cdot & 1 & \cdot \\ \rho_{op_k} & \rho_{p_1p_k} & \cdot & 1 \end{bmatrix}\right). \tag{W4}$$

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

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

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

so $O = \sigma_o Z_O$.

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

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

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

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

Therefore, using $O = \sigma_o Z_O$,

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

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

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

where each coefficient

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

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

Thus,

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

Web Appendix 3: Proof of Proposition 3

Suppose that the data can be divided into J heterogenous samples based on W. 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 $\Sigma_j = \begin{pmatrix} 1 & \Sigma_{O_jP_j} \\ \Sigma_{O_jP_j}^\top & \Sigma_{P_jP_j} \end{pmatrix}$.

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 $O_j = \sigma_j Z_{O_j}$,

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

Writing the $1 \times K$ vector $\sigma_j \, \Sigma_{O_j P_j} \, \Sigma_{P_j P_j}^{-1}$ 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 4: 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).

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

$$\tag{W6}$$

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

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

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

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

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

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 W1: 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 W1 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 [10] and higher computational cost), which is expected for any nonparametric estimator.

[10] The 2sCOPE-GC method also contains a nonparametric portion - the estimation of the marginal CDFs using

Web Appendix 5: Double Robustness Property of 2sCOPE-np

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 2) and that GC captures the joint distribution of (P, O) or (P, E) given W (Assumption 2 Table 2). 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. When there is reason to believe that the omitted effect (O) follows a normal distribution, we place no assumptions on ϵ, the exogenous part of the error term. Therefore, the structural error term, the sum of the two random variables $(E = O + \epsilon)$, can be nonnormal and left unspecified. In this case, since no assumption is placed on the joint distribution of ϵ and P, the regressor-error distribution (i.e., $P - E$) needs not to follow a GC or any specific copula dependence structure.

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

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

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

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

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

$$\epsilon_i \sim H \text{ as specificied in Table W2} \quad \text{(W16)}$$

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

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

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 4. The result for 2sCOPE-GC is similar regardless of estimation methods.

Table W2: 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.

table.

Simulation results reported in Table W2 show that although the error term E_i is nonnormally 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 6: 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 7: The Case of Count and Binary Regressors

In this section, 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.

7.A Simulation Case 4: Count Endogenous Regressor

In this simulation, we consider the following DGP.

$$W_i \sim N(0, 1), \qquad O_i \sim N(0, 1) \tag{W18}$$

$$P_i = F^{-1}_{\text{Poisson}(\exp(-2+W_i))}(\Phi(O_i)) \tag{W19}$$

$$E_i = O_i + \epsilon_i, \epsilon_i \sim N(0, 1) \tag{W20}$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 0 - 1 \cdot P_i + 1 \cdot W_i + O_i + \epsilon_i. \tag{W21}$$

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 W19, 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 (W19)) and also enters the regression equation for Y_i (Equation (W21)), 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 W3: 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

Results in Table W3 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 W3). As expected, 2sCOPE-GC and 2sCOPE-MD perform poorly for count endogenous regressors, resulting in estimation biases of 51.2% and 44.8%, respectively.

7.B Simulation 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.

$$\begin{pmatrix} P_i^* \\ W_i^* \\ E_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{pe} \\ \rho_{pw} & 1 & 0 \\ \rho_{pe} & 0 & 1 \end{bmatrix} \right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right), \tag{W22}$$

$$E_i = G^{-1}(U_{E,i}) = G^{-1}(\Phi(E_i^*)) = \Phi^{-1}(\Phi(E_i^*)) = 1 \cdot E_i^*, \tag{W23}$$

$$P_i = \mathbb{I}_{\{U_{P,i} > p_1\}} = \mathbb{I}_{\{\Phi(P_i^*) > p_1\}} = \mathbb{I}_{\{P_i^* > \Phi^{-1}(p_1)\}}, \tag{W24}$$

$$W_i = L^{-1}(U_{W,i}) = L^{-1}(\Phi(W_i^*)), \tag{W25}$$

$$Y_i = \mu + \alpha \cdot P_i + \beta \cdot W_i + E_i = 0 + 1 \cdot P_i + 2 \cdot W_i + E_i. \tag{W26}$$

where

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

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 W4) 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 W4). As expected, 2sCOPE-GC and 2sCOPE-MD show significant bias when applied to binary endogenous regressors.

Table W4: 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 4.

Web Appendix 8: Box plot of Log Price and Kernel Estimates of CDFs

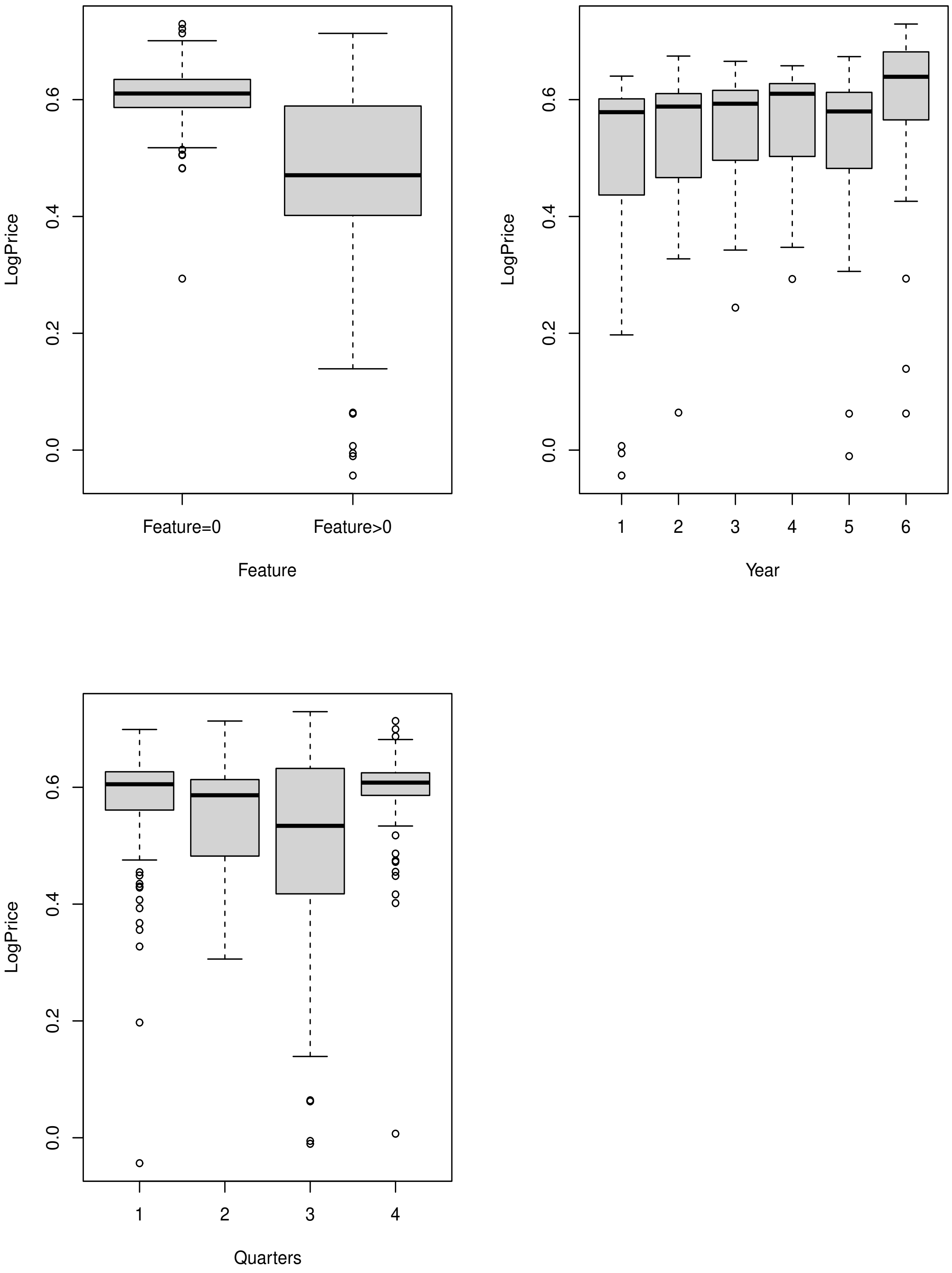
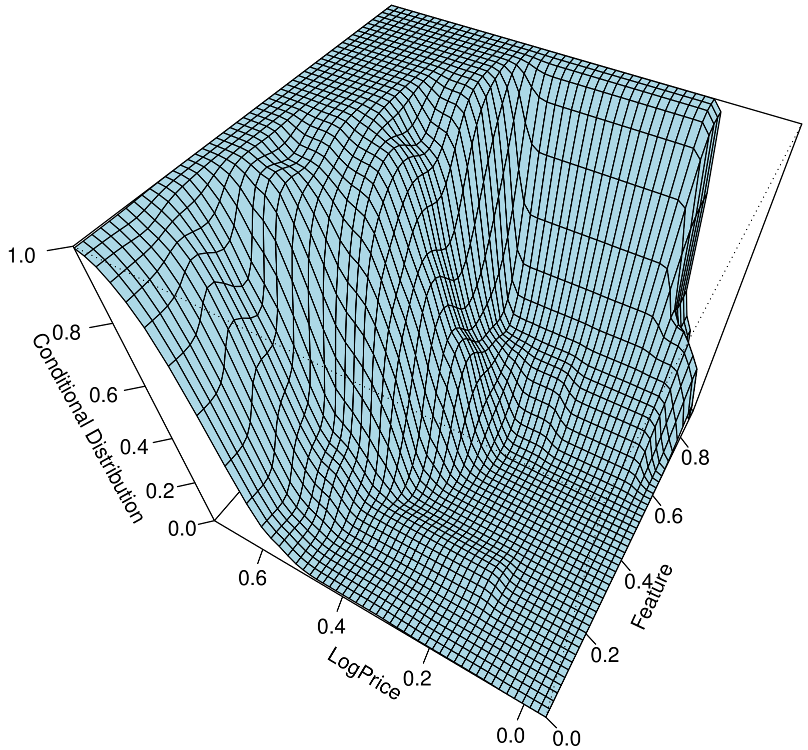


Figure W1: Boxplots of LogPrice in Store 1.



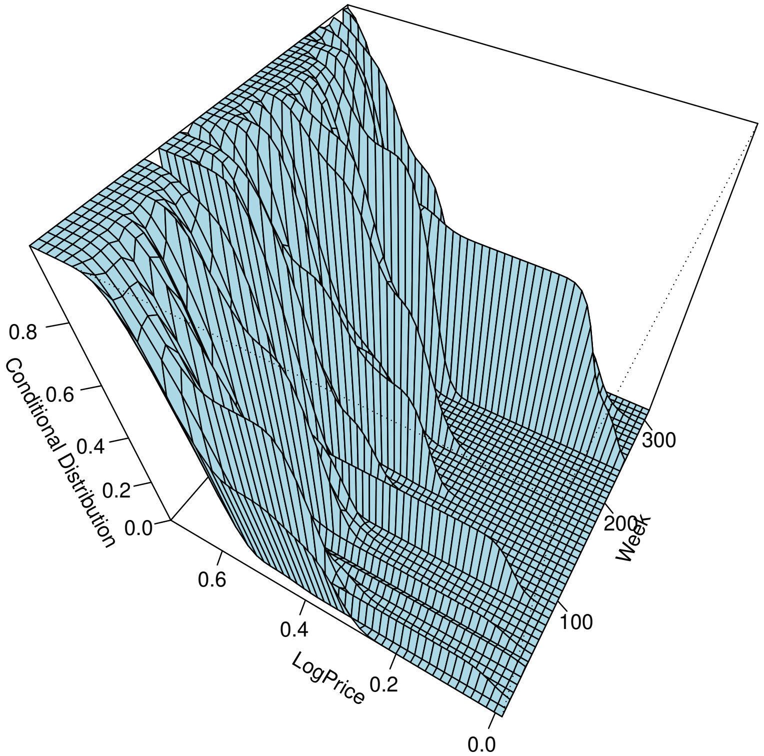


Figure W2: Kernel Estimates of CDFs of logPrice Conditional on Feature (top) and Week (Bottom) in Store 1.

Web Appendix 9: 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 earning; 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 decision correlate with unobserved traits, such as the individual's ability and motivation for success, which also affect wages. 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, Heckman et al. 2006). Due to the multiple potential sources and directions of bias, it is important to correct for endogeneity and account for potential bias.

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:

$$\ln \texttt{wage}_t = \mu + \alpha Education_t + \beta' W_t + E_t, \tag{W27}$$

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 10).

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.

9.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)[11]. 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

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

partial F test on the strength of IV rejects the null hypothesis (partial F= 15.383, p-value= 0.000), the partial correlation between the IV and education is only 0.066, and the IV is likely weak. The weak IV may explain why the confidence interval of the 2SLS estimate is too wide to be informative.

In contrast, 2SCOPE-np estimates that the return to log-earnings is 0.065 (95% CI: 0.05, 0.08), only slightly lower than the OLS estimate, and is much more precise than 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)[12].

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
$C_{\text{education}}$							**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 OLS estimate, which may result from the

[12] 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.

inappropriateness of using the two methods for count regressors with limited support.

9.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 ($\rho = 0.067$, partial F= 12.921, p-value= 0.0003). The Wu-Hausman test rejects the null hypothesis (H=3.005, p-value= 0.083) and suggests the presence of endogeneity at the 10% significance level. We conclude that the IV is likely weak [13]. This explains why the 2SLS result is statistically no different from that of OLS, despite the evidence of endogeneity.

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[14] 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-np and 2sCOPE-GC are close to each other. However, the standard error of 2sCOPE-GC estimate is so large that the postsecondary education

[13] 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.

[14] 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.

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
$C_{\text{education}}$							**-0.12**	**0.04**	**-2.84**	**-0.04**	**0.06**	**-0.69**	**-0.04**	**0.03**	**-1.21**
Wu-Hausman test				**H=3.01, p=0.083**											

See the note under Table W5

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 problematic, generating counterintuitive results.

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

Web Appendix 10: 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
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.