Addressing Endogeneity using a Two-stage Copula Generated Regressor Approach

Fan Yang
NEOMA Business School
Yi Qian
University of British Columbia
Hui Xie
Simon Fraser University

Cite as:
Yang Fan, Qian Yi, Xie Hui (2023), Addressing Endogeneity using a Two-stage Copula Generated Regressor Approach. *Proceedings of the European Marketing Academy*, 52nd, (114525)

Paper from the 52nd Annual EMAC Conference, Odense/Denmark, May 23-26, 2023

EMAC
EUROPEAN MARKETING ACADEMY

Addressing Endogeneity using a Two-stage Copula Generated Regressor Approach

Abstract

A prominent challenge when drawing causal inference using observational data is the ubiquitous presence of endogenous regressors. The classical econometric method to handle regressor endogeneity requires instrumental variables that must satisfy the stringent condition of exclusion restriction, making it infeasible to use in many settings. We propose a new instrument-free method using copula to address the endogeneity problem. Existing copula correction methods require sufficiently non-normal endogenous regressors. Furthermore, existing copula control function methods presume the independence of exogenous regressors and the endogenous regressor. Our proposed two-stage copula endogeneity correction (2sCOPE) method simultaneously relaxes the two key identification requirements, and we theoretically prove that 2sCOPE yields consistent causal-effect estimates with correlated endogenous and exogenous regressors as well as normally distributed endogenous regressors. Besides relaxing identification requirements, 2sCOPE has superior finite-sample performance and addresses the significant finite sample bias problem due to insufficient regressor non-normality. 2sCOPE employs generated regressors derived from existing regressors to control for endogeneity, and is straightforward to use and broadly applicable. Overall, 2sCOPE can greatly increase the ease and broaden the applicability of using instrument-free methods to handle regressor endogeneity. We further demonstrate the performance of 2sCOPE via simulation studies and an empirical application.

Our paper is intended for the 'Methods, Modelling & Marketing Analytics' track.

Keywords:
endogeneity, instrument-free method, correlated regressors

1 Methods

In this section, we develop a copula-based instrument-free method (2sCOPE) to handle endogenous regressors with insufficient non-normality and correlated with exogenous regressors. The 2sCOPE method jointly models the endogenous regressor, P_t, the correlated exogenous variable, W_t, and the structural error term, ξ_t, using the Gaussian copula model, which implies that $[P_t^*, W_t^*, \xi_t^*]$ follows the multivariate normal distribution:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix} \right), \tag{1}$$

where $P_t^* = \Phi^{-1}(H(P_t))$, $W_t^* = \Phi^{-1}(L(W_t))$, and $\xi_t^* = \Phi^{-1}(G(\xi_t))$, and $H(\cdot)$, $L(\cdot)$ and $G(\cdot)$ are marginal CDFs of P_t, W_t and ξ_t respectively. Under the above Gaussian copula model, we have the following system of equations that are similar to two-stage least-squares method using IVs. However, we do not require any variable that satisfies the exclusion restriction.

$$Y_t = \mu + P_t\alpha + W_t\beta + \xi_t \tag{2}$$

$$P_t^* = W_t^*\gamma + \epsilon_t, \tag{3}$$

where the two error terms ϵ_t and ξ_t are correlated because of the endogeneity of P_t. Under the assumption that ξ_t follows a normal distribution, ϵ_t and ξ_t follow a bivariate normal distribution, since they are a linear combination of tri-normal variate (ξ_t^*, P_t^*, W_t^*) under the Gaussian copula assumption. Equation (3) expresses the copula transformation of the endogenous regressor, determined by the rank-order of P_t, as a linear combination of observed and unobserved variables.

The main idea of 2sCOPE is to make use of the fact that, by conditioning on ϵ_t, the structural error term ξ_t becomes independent of both P_t and W_t. That is, by conditioning on the component of P_t causing the endogeneity of P_t (i.e, ϵ_t here), the structural error is not correlated with both P_t and W_t, thereby ensuring the consistency of standard estimation methods. In this sense, ϵ_t serves as a (scaled) control function to address the endogeneity bias. To demonstrate this point, note that the Gaussian copula model in Equation (1) can be rewritten as follows:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{pw} & \sqrt{1-\rho_{pw}^2} & 0 \\ \rho_{p\xi} & \frac{-\rho_{pw}\rho p\xi}{\sqrt{1-\rho_{pw}^2}} & \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2\rho_{p\xi}^2}{1-\rho_{pw}^2}} \end{pmatrix} \cdot \begin{pmatrix} \omega_{1,t} \\ \omega_{2,t} \\ \omega_{3,t} \end{pmatrix},$$

where $[\omega_{1,t}, \omega_{2,t}, \omega_{3,t}]$ are standard normal. Given the above joint normal distribution for (P_t^*, W_t^*, ξ_t^*) and $\xi_t^* = \sigma_\xi \xi_t$, we have

$$P_t^* \quad = \quad \rho_{pw} W_t^* + \sqrt{(1-\rho_{pw}^2)} \cdot \omega_{2,t} = \rho_{pw} W_t^* + \epsilon_t, \tag{4}$$

which shows γ in Equation (3) is ρ_{pw} and $\epsilon_t = \sqrt{(1-\rho_{pw}^2)} \cdot \omega_{2,t}$, and

$$\begin{aligned} Y_t &= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} P_t^* + \frac{-\sigma_\xi \rho_{pw}\rho_{p\xi}}{1-\rho_{pw}^2} W_t^* + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t} \\ &= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} (P_t^* - \rho_{pw} W_t^*) + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}, \\ &= \mu + P_t\alpha + W_t\beta + \frac{\sigma_\xi \rho_{p\xi}}{1-\rho_{pw}^2} \epsilon_t + \sigma_\xi \sqrt{1-\rho_{p\xi}^2 - \frac{\rho_{pw}^2 \rho_{p\xi}^2}{1-\rho_{pw}^2}} \cdot \omega_{3,t}. \end{aligned} \tag{5}$$

Equation (5) suggests adding the estimate of the error term ϵ_t from the first stage regression as a generated regressor to the outcome regression instead of using P_t^* and W_t^*. The new error term $\omega_{3,t}$ is uncorrelated with all the regressors in Equation (5), ensuring the consistency of model estimates. This two-step procedure, named as 2sCOPE, adds the first-stage residual term $\widehat{\epsilon}_t$ to control for endogeneity and in this aspect is similar to the control function approach of Petrin and Train (2010). However, unlike Petrin and Train (2010), 2sCOPE requires no use of instrumental variables.

2 Simulation Study

In this section, we conduct Monte Carlo simulation studies for the following goals: (a) to assess the performance of the proposed method for correlated regressors, (b) to assess the performance of the proposed method under regressor normality and near normality, and (c) to assess the performance of the proposed method under various types of structural models. We compare the performance of 2sCOPE with existing methods (OLS and Copula$_{\text{Origin}}$ from Park and Gupta (2012)). Following Park and Gupta (2012), we measure the estimation bias using t_{bias} calculated as the ratio of the absolute difference between the mean of the sampling distribution and the true parameter value to the standard error of the parameter estimate. As defined above, t_{bias} represents the size of bias relative to the sampling error.

2.1 Case 1: Non-normal Regressors

We first examine the case when P and W are correlated. The data-generating process (DGP) is summarized below:

$$\begin{pmatrix} P_t^* \\ W_t^* \\ \xi_t^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{pw} & \rho_{p\xi} \\ \rho_{pw} & 1 & 0 \\ \rho_{p\xi} & 0 & 1 \end{bmatrix}\right) = N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}\right), \quad (6)$$

$$\xi_t = G^{-1}(U_{\xi,t}) = G^{-1}(\Phi(\xi_t^*)) = \Phi^{-1}(\Phi(\xi^*)) = 1 \cdot \xi_t^*, \quad (7)$$

$$P_t = H^{-1}(U_{P,t}) = H^{-1}(\Phi(P_t^*)), \quad W_t = L^{-1}(U_{W,t}) = L^{-1}(\Phi(W_t^*)), \quad (8)$$

$$Y_t = \mu + \alpha \cdot P_t + \beta \cdot W_t + \xi_t = 1 + 1 \cdot P_t + (-1) \cdot W_t + \xi_t, \quad (9)$$

where ξ_t^* and P_t^* are correlated ($\rho_{p\xi} = 0.5$), generating the endogeneity problem; W_t^* is exogenous and uncorrelated with ξ_t^*; W_t^* and P_t^* are correlated ($\rho_{pw} = 0.5$), and thus W_t and P_t are correlated. We consider four different estimation methods: (1) OLS, (2) $\text{Copula}_{\text{Origin}}$ and (3) the proposed 2sCOPE in the form of Equation (5). We set the sample size $T = 1000$, and generate 1000 data sets as replicates using the DGP above. In the simulation, we use the gamma distribution $Gamma(1, 1)$ with shape and rate equal to 1 for P_t and the exponential distribution $Exp(1)$ with rate 1 for W_t. Models are estimated on all generated data sets, providing the empirical distributions of parameter estimates.

			OLS			$\text{Copula}_{\text{Origin}}$			2sCOPE		
ρ_{pw}	Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
0.5	μ	1	0.689	0.045	6.964	1.231	0.081	2.849	1.009	0.059	0.157
	α	1	1.571	0.036	15.75	1.055	0.069	0.791	0.986	0.070	0.197
	β	-1	-1.259	0.031	8.236	-1.289	0.031	9.169	-0.995	0.042	0.123
	$\rho_{p\xi}$	0.5	-	-	-	0.570	0.047	1.504	0.504	0.038	0.097
	σ_ξ	1	0.862	0.020	6.902	1.011	0.043	0.244	1.006	0.040	0.143
0.7	μ	1	0.730	0.041	6.629	1.307	0.076	4.037	1.005	0.053	0.088
	α	1	1.800	0.041	19.67	1.260	0.068	3.838	0.991	0.075	0.118
	β	-1	-1.529	0.037	14.21	-1.567	0.037	15.36	-0.994	0.056	0.110
	$\rho_{p\xi}$	0.5	-	-	-	0.633	0.043	3.130	0.500	0.026	0.000
	σ_ξ	1	0.799	0.018	11.18	0.980	0.044	0.468	1.003	0.040	0.084

Table 1: Results of the Simulation Study Case 1: Non-normal Regressors
Note: Mean and SE denote the average and standard deviation of parameter estimates over all the 1,000 simulated samples.

Table 1 reports estimation results. As expected, OLS estimates of both α and β are biased ($t_{bias} = 15.75/8.24$) due to the regressor endogeneity. $\text{Copula}_{\text{Origin}}$ reduces the bias, but still shows significant bias for the coefficient estimates of P_t and W_t. The bias of $\text{Copula}_{\text{Origin}}$ depends on the strength of the correlation between W and P. Stronger correlations between P^* and W^* can cause a larger bias of $\text{Copula}_{\text{Origin}}$ estimates. For example, when the correlation between W^* and P^* increases from 0.5 to 0.7, the bias of estimated α increases by around five times (from 0.055 to 0.260 in Table 1 under the column "$\text{Copula}_{\text{Origin}}$"). The bias confirms our derivation in the model section, demonstrating that using the existing copula method may not solve the endogeneity problem completely with correlated regressors. The proposed 2sCOPE method provides consistent estimates without using instruments. The average estimates of $\rho_{p\xi}$ is close to the true value 0.5 and is significantly different from 0, implying regressor endogeneity detected correctly using 2sCOPE.

2.2 Case 2: Normal Regressors

Next, we examine the case when the endogenous regressor and (or) the correlated exogenous regressor are normally distributed. We pay special attention to this case because normality is not allowed for endogenous regressors in Park and Gupta (2012). We use the same DG as described in Equations (6) to (9) to generate the data, except that the marginal CDFs for regressors, $H(\cdot)$ and $L(\cdot)$, are chosen according to the distributions listed in the first two columns in Table 2.

Table 2 summarizes the estimation results. As expected, OLS estimates are biased. $\text{Copula}_{\text{Origin}}$ produces biased estimates whenever the endogenous regressor P follows a normal distribution. The estimates of $\text{Copula}_{\text{Origin}}$ are biased when P follows a gamma distribution (first row of Table 2) for a different reason: P and W are correlated. By contrast, the proposed 2sCOPE method provides consistent estimates as long as P_t and W_t are not both normally distributed. Both α and β are tightly distributed near the true value whenever P_t or W_t is nonnormally distributed. Unlike $\text{Copula}_{\text{Origin}}$, 2sCOPE adds the residual term obtained from regressing P_t^* on W_t^* as the generated regressor. Thus, as long as P_t and W_t are not both normally distributed, the residual term is not perfectly co-linear with the original regressors, permitting model identification. Only when both P_t and W_t are normally distributed (the last scenario in Table 2), the residual term added into the structural regression model becomes a linear combination of P_t and W_t, causing perfect co-linearity and model non-identification.

Overall, this simulation study demonstrates the capability of the proposed 2sCOPE to relax the non-normality assumption in $\text{Copula}_{\text{Origin}}$ as long as one of P_t and W_t is nonnormally distributed.

2.3 Case 3: Random Coefficient Linear Panel Model

We investigate the performance of 2sCOPE in random coefficient linear panel model. We use the copula function and marginal distributions of $[P_{it}, W_{it}, \xi_{it}]$ as specified in Case 1 (Equations 6-8). We assign $\rho_{pw} = 0.7$ as an example. We then generate the outcome Y_{it} using the following standard random coefficient linear panel model:

$$Y_{it} = \bar{\mu} + \mu_i + P_{it}(\bar{\alpha} + a_i) + W_{it}(\bar{\beta} + b_i) + \xi_{it} = 1 + \mu_i + P_{it}(1 + a_i) + W_{it}(-1 + b_i) + \xi_{it},$$

where $[\mu_i, a_i, b_i] \sim N(0, I_3)$, $t = 1, ..., 50$ indexes occasions for repeated measurements, and $i = 1, ..., 500$ indexes the individual units. The above random coefficients model permits individual units to have heterogeneous baseline preferences (μ_i) and heterogeneous

Distribution				OLS			Copula$_{\text{Origin}}$			2sCOPE		
P	W	Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
Gamma	Normal	μ	1	0.431	0.045	12.63	1.018	0.078	0.227	1.015	0.077	0.190
		α	1	1.569	0.037	15.40	0.979	0.070	0.302	0.985	0.070	0.212
		β	-1	-1.259	0.030	8.619	-1.333	0.028	11.78	-0.997	0.045	0.067
		$\rho_{p\xi}$	0.5	-	-	-	0.640	0.039	3.556	0.506	0.036	0.151
		σ_ξ	1	0.861	0.019	7.240	1.064	0.046	1.394	1.005	0.038	0.134
Normal	Exp	μ	1	1.286	0.042	6.777	1.286	0.045	6.374	1.023	0.070	0.334
		α	1	1.628	0.031	20.36	1.532	0.462	1.152	1.048	0.126	0.381
		β	-1	-1.286	0.032	8.956	-1.287	0.032	8.960	-1.024	0.062	0.383
		$\rho_{p\xi}$	0.5	-	-	-	0.089	0.419	0.980	0.465	0.074	0.473
		σ_ξ	1	0.829	0.018	9.492	0.940	0.151	0.394	0.980	0.063	0.318
Normal	Normal	μ	1	1.001	0.026	0.046	1.002	0.030	0.052	1.002	0.028	0.057
		α	1	1.668	0.030	22.38	1.663	0.450	1.474	1.655	0.395	1.657
		β	-1	-1.335	0.029	11.44	-1.335	0.029	11.42	-1.328	0.197	1.668
		$\rho_{p\xi}$	0.5	-	-	-	0.006	0.412	1.198	0.010	0.303	1.616
		σ_ξ	1	0.816	0.019	9.687	0.917	0.155	0.534	0.879	0.092	1.317

Table 2: Results of Case 2: Normal Regressors

responses to regressors (a_i, b_i). Such random coefficients models are frequently used in marketing studies to capture individual heterogeneity and to profile and target individuals. The correlation between ξ_{it} and P_{it} creates the regressor endogeneity problem, which can cause biased estimates for standard linear random coefficient estimation methods ignoring the regressor-error correlation. We generate individual-level panel data as described above for 1000 times and use the data for estimation. Estimation results are in Table 3. LME is the standard estimation method for linear mixed models assuming all regressors are exogenous, as implemented in the R function `lme()`. LME and Copula$_{\text{Origin}}$ are biased because of endogeneity and correlated exogenous regressors, respectively. Our proposed method 2sCOPE provides unbiased estimates that are tightly distributed around the true values for all parameters.

		LME			Copula$_{\text{Origin}}$			2sCOPE		
Parameters	True	Mean	SE	t_{bias}	Mean	SE	t_{bias}	Mean	SE	t_{bias}
$\bar{\mu}$	1	0.722	0.046	6.052	1.314	0.049	6.399	1.004	0.048	0.091
$\bar{\alpha}$	1	1.853	0.045	18.83	1.293	0.045	6.469	1.000	0.046	0.008
$\bar{\beta}$	-1	-1.557	0.045	12.39	-1.598	0.044	13.56	-1.000	0.044	0.005
σ_μ	1	0.985	0.033	0.459	0.982	0.033	0.547	0.984	0.031	0.522
σ_α	1	0.988	0.036	0.326	0.987	0.034	0.397	0.989	0.035	0.316
σ_β	1	0.993	0.031	0.235	0.992	0.033	0.249	0.992	0.033	0.248
$\rho_{p\xi}$	0.5	-	-	-	0.646	0.009	16.33	0.507	0.005	1.365
σ_ξ	1	0.794	0.004	57.71	0.957	0.010	4.439	0.985	0.009	1.640

Table 3: Results of Simulation Study Case 3: Random Coefficient Linear Panel Model
Note: $\sigma_\mu, \sigma_\alpha, \sigma_\beta$ are standard deviations of μ_i, a_i, b_i.

3 Empirical Application

In this section, we apply our method to a real marketing application. We illustrate the proposed method to address the price endogeneity issue using store-level sales data of toothpaste category in Chicago over 373 weeks from 1989 to 1997 [1]. To control for product size, we select toothpaste with the most common size, which is 6.4 oz. Retail price is usually considered endogenous. The endogeneity of retail price can come from unmeasured product characteristics or demand shocks that can influence both consumers' and retailers' decisions. Since these variables are unobserved by researchers, they are absorbed into the structural error, leading to the endogeneity problem. Prices of different stores are correlated and often used as an IV for each other. This allows us to test the performance of the proposed 2sCOPE method in an empirical setting where a good IV exists. Besides the endogenous price, two promotion-related variables, bonus promotion and direct price reduction, would also affect demand. Following Park and Gupta (2012), we treat the promotion variables as exogenous regressors. We focus on category sales in two large stores in Chicago (referred to as Stores 1 and 2). We convert retail price, in-store promotion and sales from UPC level to aggregate category level, weighted by weekly market share. The correlation between log retail price and bonus promotion in Store 1 (Store 2) is -0.30 (-0.15), and the correlation between log retail price and price reduction promotion in Store 1 (Store 2) is -0.23 (-0.35). Both the correlations are significantly different from zero. The appreciable correlations between price and promotion variables actually provide a good setting for testing our method. The moderate sample size (T=373) also provides an opportunity to compare finite sample performance of different copula correction methods in the presence of potentially insufficient regressor non-normality in real data. Summary statistics of key variables are summarized in Table 4.

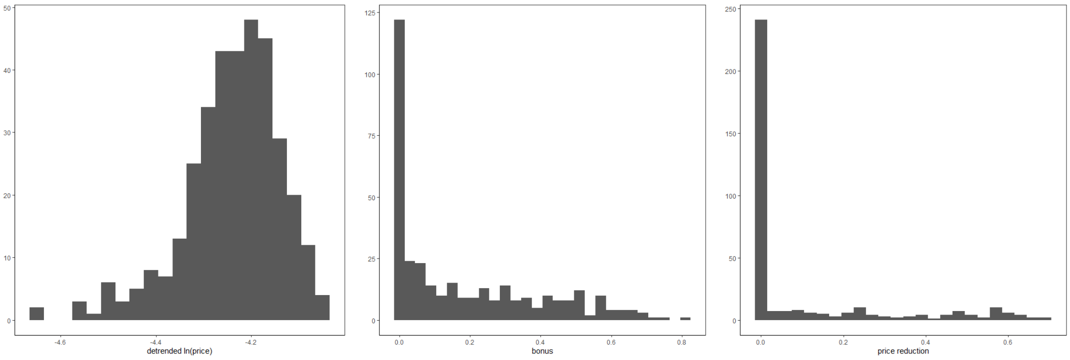
	Store 1				Store 2			
Variables	Mean	SD	Max	Min	Mean	SD	Max	Min
Sales (Unit)	115	52.8	720	35	165.7	93.7	1334	26
Price ($)	2.06	0.20	2.48	1.46	2.10	0.21	2.48	1.47
Bonus	0.18	0.20	0.80	0.00	0.16	0.19	0.79	0.00
PriceRedu	0.10	0.19	0.72	0.00	0.10	0.19	0.73	0.00

Table 4: Summary Statistics

We estimate the following linear regression model:

$$\log(\text{Sales}_t) = \beta_0 + \log(\text{Retail Price}_t) \cdot \beta_1 + W_t'\beta_2 + \xi_t,$$

[1] We obtained the data from https://www.chicagobooth.edu/research/kilts/datasets/dominicks.



(a) detrended log price **(b)** bonus **(c)** price reduction

Figure 1: Histogram of Log Retail Price, Bonus and Price Reduction in Store 1

where $t = 1, 2, ..., T$ indexes week. The vector W_t includes all exogenous regressors, which are two promotion variables, bonus promotion and price reduction, in this application. Figure 1 shows the histograms of detrended log retail prices and the two promotion variables. All the three variables are continuous variables. Moreover, except log retail price, which is a bit close to normal distribution, the other two regressors, bonus and price reduction, are both nonnormally distributed. Therefore, we expect that the proposed 2sCOPE method can exploit these additional features of exogenous regressors correlated with the endogenous regressor for model identification and estimation even if the endogenous regressor has a close-to-normal distribution. We estimate the model using the OLS, two-stage least-squares (TSLS), $\text{Copula}_{\text{Origin}}$, and our proposed 2sCOPE method.

We use the IV-based TSLS estimator as a benchmark to test the validity of our proposed method. Following Park and Gupta (2012), we use retail price at the other store as an instrument for price. This variable can be a valid instrument as it satisfies the two key requirements. First, retail prices across stores in a same market can be highly correlated because wholesale prices are usually offered the same (or very close). The Pearson correlation between the detrended log retail prices at Stores 1 and 2 is 0.79, providing strong explanatory power on the endogenous price. The correlation is comparable to that in Park and Gupta (2012). Second, some unmeasured product characteristics such as shelf-space allocation, shelf location and category location are determined by retailers and are usually not systematically related to wholesale prices (exclusion restriction). For the three copula-based methods, we make use of information from the existing endogenous and exogenous regressors and no extra IVs are needed. In $\text{Copula}_{\text{Origin}}$, we add the copula transformation of the detrended log price, $\text{logP}^* = \Phi^{-1}(\widehat{H}(\text{logP}))$, as a "generated regressor" to the outcome regression. For the 2sCOPE method, we first regress logP^* on Bonus^* $(= \Phi^{-1}(\widehat{L}_1(\text{Bonus})))$ and $\text{PriceRedu}^*(= \Phi^{-1}(\widehat{L}_2(\text{PriceRedu})))$, and then add the residual as the only "generated regressor" to the outcome regression. $\widehat{H}(\cdot), \widehat{L}_1(\cdot), \widehat{L}_2(\cdot)$ are all estimated CDFs using the univariate empirical distribution for each regressor. Standard errors of parameter estimates are obtained using bootstrap.

Table 5 reports the estimation results. Beginning with the results from Store 1, OLS estimates are significantly different from TSLS estimates, indicating that the price endogeneity issue occurs. Instrumenting for retail price changes the price coefficient estimate from -0.767 to -1.797, implying that there is a positive correlation between unobserved product characteristics and the price. The estimates of ρ in the three IV-free copula-based methods, representing the correlation between the endogenous regressor P_t and the error term, are all significantly positive, further confirming our previous conclusion. This direction of correlation is consistent with previous empirical findings (e.g., Villas-Boas and Winer 1999, Chintagunta, Dubé, and Goh 2005). The price elasticity estimates from

		OLS			TSLS			$\text{Copula}_{\text{Origin}}$			2sCOPE		
Store	Parameters	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value	Est	SE	t-value
Store 1	Constant	1.301	1.197	0.25	-2.993	1.646	1.82	-8.526	2.619	3.26	-3.908	2.314	1.69
	Price	-0.767	0.288	2.66	-1.797	0.396	4.54	-3.082	0.620	4.97	-2.014	0.555	3.63
	Bonus	0.371	0.122	3.31	0.104	0.141	0.74	0.415	0.115	3.61	0.064	0.171	0.37
	PriceRedu	0.498	0.115	4.33	0.285	0.125	2.28	0.544	0.111	4.90	0.275	0.143	1.92
	ρ	-	-	-	-	-	-	0.521	0.098	5.32	0.297	0.089	3.34
Store 2	Constant	-3.898	1.246	3.13	0.763	1.943	0.39	1.107	3.404	0.33	0.001	2.702	0.00
	Price	-1.982	0.300	6.61	-0.864	0.467	1.85	-0.799	0.807	0.99	-1.048	0.648	1.62
	Bonus	0.062	0.116	0.53	0.286	0.148	1.93	0.032	0.117	0.27	0.239	0.151	1.58
	PriceRedu	0.283	0.111	2.55	0.540	0.137	3.94	0.275	0.110	2.5	0.467	0.152	3.07
	ρ	-	-	-	-	-	-	-0.319	0.177	1.80	-0.188	0.109	1.72

Table 5: Estimation Results: Toothpaste Sales

the $\text{Copula}_{\text{Origin}}$ and the proposed method 2sCOPE are -3.082 and -2.014, respectively. Among the two estimates, the estimate of -2.014 from the proposed 2sCOPE is close to the estimate of -1.797 from the TSLS method, whereas the existing copula yields substantially greater-sized price elasticity estimates. We confirm in the literature that the TSLS and 2sCOPE estimates are reasonable because the price elasticity of toothpaste category is around -2.0 (Hoch et al. 1995, Mackiewicz and Falkowski 2015). Comparing the estimates of ρ from the three IV-free copula-based methods, our proposed 2sCOPE provides a much smaller estimate of ρ (0.297 for 2sCOPE vs 0.521 for $\text{Copula}_{\text{Origin}}$ in Table 5), consistent with the over-correction in $\text{Copula}_{\text{Origin}}$. Reasons for the substantial differences in the 2sCOPE estimates from the $\text{Copula}_{\text{Origin}}$ include (1) correlated endogenous and exogenous regressors and (2) the unimodal close-to-normality distribution for the logarithm of price variable, which can lead to poor finite sample performance for $\text{Copula}_{\text{Origin}}$.

Unlike Store 1, the results from Store 2 indicate that the retail price is not endogenous. First, the estimates of ρ (the correlation between price and the error term) are not significantly different from 0 for both $\text{Copula}_{\text{Origin}}$ and 2sCOPE (t-value ≤ 1.96 under columns "$\text{Copula}_{\text{Origin}}$" and "2sCOPE" for Store 2 in Table 5). Second, the estimated price coefficient of OLS is -1.982, which is very close to the estimates of TSLS and 2sCOPE in Store 1 and further confirming no endogeneity of price in Store 2. Overall, the price elasticity estimates from TSLS and the three IV-free copulas-based methods are close to each other for Store 2, and the observed differences between them and the OLS estimate can be attributed to estimation variability incurred from using more complicated models instead of the presence of endogeneity.

References

Chintagunta, Pradeep, Jean-Pierre Dubé, and Khim Yong Goh (2005), "Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household-level brand choice models," *Management Science*, 51 (5), 832–849.

Hoch, Stephen J, Byung-Do Kim, Alan L Montgomery, and Peter E Rossi (1995), "Determinants of store-level price elasticity," *Journal of Marketing Research*, 32 (1), 17–29.

Mackiewicz, Robert and Andrzej Falkowski (2015), "The Use of Weber Fraction as a Tool to Measure Price Sensitivity: a Gain and Loss Perspective.," *Advances in Consumer Research*, 43.

Park, Sungho and Sachin Gupta (2012), "Handling endogenous regressors by joint estimation using copulas," *Marketing Science*, 31 (4), 567–586.

Petrin, Amil and Kenneth Train (2010), "A control function approach to endogeneity in consumer choice models," *Journal of marketing research*, 47 (1), 3–13.

Villas-Boas, J Miguel and Russell S Winer (1999), "Endogeneity in brand choice models," *Management science*, 45 (10), 1324–1338.