Journal of Business Economics
https://doi.org/10.1007/s11573-023-01179-z

ORIGINAL PAPER

Endogeneity of marketing variables in multicategory choice models

Harald Hruschka[1]

Accepted: 4 September 2023

Abstract

A regressor is endogenous if it is correlated with the unobserved residual of a model. Ignoring endogeneity may lead to biased coefficients. We deal with the omitted variable bias that arises if firms set marketing variables considering factors (demand shocks) that researchers do not observe. Whereas publications on sales response or brand choice models frequently take the potential endogeneity of marketing variables into account, multicategory choice models provide a different picture. To consider endogeneity in multicategory choice models, we follow a two-step Gaussian copula approach. The first step corresponds to an individual-level random coefficient version of the multivariate logit model. We analyze yearly shopping data for one specific grocery store, referring to 29 product categories. If the assumption of a Gaussian correlation structure is met, the copula approach indicates the endogeneity of a category-specific marketing variable in about 31% of the categories. The majority of marketing variables rated as endogenous are positively correlated with the omitted variable, implying that ignoring endogeneity leads to an overestimation of the coefficients of the respective marketing variable. Finally, we investigate whether taking endogeneity into account by the copula approach leads to different managerial implications. In this regard, we demonstrate that for our data ignoring endogeneity often suggests a level of marketing activity that is too high.

Keywords Marketing · Retailing · Multicategory choice · Market basket analysis · Endogeneity

JEL Classification M31 · L81 · D1 · C35

✉ Harald Hruschka
harald.hruschka@ur.de

[1] Faculty of Business, Economics, and Management Information Systems, University of Regensburg, Universitätsstraße 31, 93040 Regensburg, Bavaria, Germany

Published online: 25 September 2023

1 Introduction

A regressor is endogenous if it is correlated with the unobserved residual of a model. Ignoring endogeneity may lead to a bias, according to which the estimated coefficient of the regressor differs from its true value. Wooldridge (2010) distinguishes four sources of endogeneity, namely variable omission, measurement error, selection, and simultaneity.

These sources generate bias under the following conditions:

- the omitted variable is related to both the regressor and the dependent variable.
- the dependent variable also affects the regressor (simultaneity.
- the error in measuring the regressor is correlated with the dependent variable;
- the selection into a sample is not random, or the assignment of a treatment (e.g., the assignment of an ad to households) is not random.

Here we deal with the omitted variable bias that arises if firms set marketing variables considering factors (demand shocks) that researchers do not observe. Such unobserved factors are not included in the model, they are incorporated into the structural error of the model. Biased estimates generated this way prevent the inference of valid causal effects, e.g., changes of sales or choice probabilities produced by changes of marketing variables (Papies et al. 2017). The magnitude and direction of the endogeneity bias depend on the way managers react to unobserved demand shocks. The sign of a bias corresponds to the sign of the correlation between the respective marketing variable and the excluded factor (Papies et al. 2017). Therefore, a positive (negative) bias of a marketing variable's coefficient implies that the marketing variable and the unobserved demand shock are positively (negatively) correlated, i.e., that managers raise (decrease) the marketing variable in the case of a positive shock in demand.

The most widespread approach correcting for endogeneity uses observed instrumental variables. Ideally, an instrumental variable should be correlated with the marketing variable and should be uncorrelated with the error term of the respective model. Another approach adds a supply-side model explaining observed realizations of the endogenous variables. Typically, the supply-side model is based on either prior knowledge or a theory of the decision-making process of a firm. Obviously, this approach is prone to misspecification bias due to erroneous assumptions on decision-making. Instrument-free alternatives such as the higher moments approach (Lewbel 1997), the heteroskedastic errors approach (Lewbel 2012), the latent instrument variable approach (Ebbes et al. 2005), and the Gaussian copula approach (Park and Gupta 2012) do not require observed instrumental variables or knowledge of the decision-making process.

For sales response and brand choice models, we can find several papers that deal with endogeneity [examples are Andrews and Ebbes 2014; Villas-Boas and Winer 1999; Petrin and Train 1999] by using instrumental variables or adding a supply-side model (Besanko et al. 1998). Publications using instrument-free

methods focus on sales response data (Ebbes et al. 2009; Hruschka and Gerhardt 2012; Park and Gupta 2012; Elshiewy and Boztuğ 2018; Atefi et al. 2018; Keller et al. 2019; Yang et al. 2022). Datta et al. (2015) consider endogeneity of marketing variables in a binomial choice model by an instrument-free approach. Whereas the publications mentioned so far are restricted to one brand or one product category, Datta et al. (2017) include an instrument-free approach to deal with endogeneity in a market share model for 25 categories.

Multicategory choice models allow pick-any choices, i.e., households may purchase multiple product categories at the same occasion. Multinomial choice models, on the other hand, are appropriate only if households choose exactly one category at each purchase occasion. The two dominant functional forms of multicategory choice models are the multivariate logit (Russell and Petersen 2000; Boztuğ and Hildebrandt 2008; Boztuğ and Reutterer 2008; Dippold and Hruschka 2013; Aurier and Mejia 2014; Hruschka 2021; Richards et al. 2018; Solnet et al. 2016) and the multivariate probit (Chib et al. 2002; Duvvuri et al. 2007; Manchanda et al. 1999; Hruschka 2017; Aurier and Mejia 2014; Xia et al. 2019).

In contrast to single category models, which may lead to biased conclusions on the effects on consumer behavior (Seetharaman et al. 2005), multicategory models like the multivariate logit and the multivariate probit model allow for interactions between purchases of different categories. A positive interaction exists if the purchase of category A increases the purchase probability of another category B. For example, the purchase of snacks could increase the purchase probability of beverages. In a negative interaction, on the other hand, the purchase of category A decreases the purchase probability of another category B (e.g., the purchase of cold cereal could decrease the purchase probability of beer).

Such interactions turn out to be especially important for promotion, assortment, and store layout decisions. Considering interactions is a prerequisite evaluating the effects of promotions across a retailer's assortment (Russell and Petersen 2000). By using information on interactions, management avoids the error of eliminating low-profit categories though the latter have strong positive interactions with other, high-profit categories (Boztuğ and Silberhorn 2006). Placing categories with positive interactions close to each other in store layouts increases total sales (Boztuğ and Silberhorn 2006).

Research on multicategory choice models, as a rule, leaves out endogeneity concerns. We are aware of only one exception, Richards et al. (2018) who tackle endogeneity by adding a supply-side model to a multivariate logit model for choices in four categories. Let us put forward two possible reasons for this research gap with respect to multicategory choice models.

Maybe some researchers think that the endogeneity of marketing variables is a lesser problem at the category level. Such a consideration can be justified if the following two conditions are both valid. One condition requires that a firm's managers, who respond to a shock at the category level, do not set marketing variables of most brands that belong to the category in the same way. According to the other condition, endogeneity only occurs for low share brands or becomes low because random shocks affect marketing variables of brands in different

directions with appropriate positive and negative correlations. We doubt that these two conditions apply in general.

Empirical evidence for the endogeneity of marketing variables at the category level is rare, with the exception of Park and Gupta (2012) and Richards et al. (2018), who analyze category sales and purchase incidence data, respectively. The demand surge of toilet paper and pasta during the first weeks of the COVID-19 crisis in Germany offers an example, as retail managers reduced sales promotion activities in these categories (Consumer Index 2020). Such circumstances lead to negative omitted variable biases, because positive demand shocks are negatively correlated with sales promotion activities.

Another reason for this research gap may be the fact that the number of endogenous marketing variables is much higher compared to sales response or brand choice models. As a rule, about seven endogenous marketing variables are investigated in brand choice models. The usual number of endogenous marketing variables in sales response models is even lower. Quite contrary, we analyze 29 endogenous marketing variables in our empirical study. A high number of marketing variables makes it difficult to either find appropriate instrumental variables or develop a supply-side model. That is why we turn to the instrument-free Gaussian copula-correction method of Park and Gupta (2012) that under certain assumptions reproduces the correlation between marketing variables and the error term.

For models without intercept Park and Gupta (2012) demonstrate their method's robustness if the distribution of structural errors is mispecified. On the other hand, it is well known that the majority of models in marketing encompass intercepts. For such models the recent simulations of Becker et al. (2022) show that the Gaussian copula-correction method is sensitive to deviations of the structural error terms from normality and the Gaussian correlation structure. According to the results of Becker et al. (2022), normality of the error terms can be reliably assessed based on regression residuals. Nonetheless, we have to emphasize that the Gaussian correlation structure of the error term with any endogeneous variable remains as untestable assumption.

In the next section, we present the homogeneous multivariate logit model. We also show how the endogeneity of marketing variables can be taken into account, based on a two-step Gaussian copula approach developed for the individual-level mixed logit model. Subsequently, we explain our estimation approach. We also describe how we investigate the difference of the effects of marketing variables between a model with coefficients corrected by the copula approach and another model with uncorrected coefficients. The empirical part of our paper starts by characterizing the data set. The following section compares to the performance of a related model that excludes interactions between purchases of product categories. We then discuss coefficients of interactions, of category-specific marketing variables and of their copula correction terms. Assuming a Gaussian correlation structure the significance of the latter group of coefficients indicates that category-specific marketing variables are endogenous. In this case, the sign of a copula correction coefficient shows whether marketing variables are subject to positive or negative omitted variable biases. We also examine to what extent the testable requirements of the Gaussian copula correction method are fulfilled. Afterwards, we discuss managerial

implications. In the concluding section, we summarize the main results, mention further application areas, and discuss the limitations of our approach together with corresponding model extensions.

2 Models

In this section, we present the homogeneous multivariate (MVL) model. Then we show how endogeneity of marketing variables can be considered by the Gaussian copula approach and explain our estimation approach.

J column vector y_{mt} denotes market basket t of household m and consists of binary purchase indicators (J symbolizes the number of product categories). If household m purchases category j at purchase occasion t, the respective element y_{jmt} equals one. Vector x_{mt} consists of regressors relevant for the market basket t of household m. In our study, these regressors consist of category loyalties and one category-specific marketing variable.

We compute the loyalty of household m for category j in market basket t in analogy to exponentially smoothed brand loyalties (Guadagni and Little 1983):

$$loy_{jmt} = \alpha\, y_{jmt-1} + (1-\alpha)\, loy_{jmt-1} \tag{1}$$

$0 \leq \alpha \leq 1$ denotes the smoothing constant. The binary purchase incidence y_{jmt-1} equals one, if household m purchases category j at the previous purchase occasion $t-1$. The current category loyalty depends on the previous purchase incidence y_{jmt-1} and the previous loyalty loy_{jmt-1}. In a manner similar to the brand loyalty of Guadagni and Little (1983) we set initial values loy_{jm0} equal to the relative purchase frequency of the respective category j across all households and shopping visits ($t = 1$ denotes the first shopping visit). The lower the smoothing constant α is, the less the loyalty variable reflects fluctuating purchases.

2.1 Homogeneous multivariate logit model

In the homogeneous MVL model, each coefficient is constant across households. Extending the expression for the homogeneous MVL model without regressors (also known as auto-logistic model) given in Besag (1972) we define the probability of market basket y_{mt} conditional on regressors x_{mt} as follows:

$$\begin{aligned} &\exp(y'_{mt}a + x'_{mt}B\,y_{mt} + 1/2\,y'_{tm}Vy_{mt})/C \\ &\text{with} \quad C = \sum_{v\in\{0,1\}^J} \exp(v'a + x'_{mt}B\,v + 1/2\,v'Vv) \end{aligned} \tag{2}$$

Expression (2) shows that computation of this probability requires division by the so-called normalization constant C that is obtained by summing over all possible market baskets defined by different binary vectors v. Coefficients contained in the (J, J) matrix V measure pairwise interactions between categories. As a pairwise interaction of a category with itself does not make sense, all diagonal elements of

V are zero. Off-diagonal elements are symmetric, i.e., $V_{j1,j2} = V_{j2,j1}$. Column vector a consists of J constants. The (K, J) matrix B holds the effect of K regressors on purchase probabilities. The homogeneous MVL model has been applied to market basket data by Russell and Petersen (2000) building upon earlier publications in statistics (Cox 1972; Besag 1974).

For the homogeneous MVL model we can write the purchase probability of category j in market basket t of household m conditional on purchases of the other categories collected in vector y_{-jtm}, the category-specific marketing variable $mvar_{jt}$, and the category-specific loyalty loy_{jmt} as:

$$P(y_{jmt} = 1|y_{-jmt}, x_{mt}) = \varphi(Z_{jmt})$$
$$\text{with} \quad Z_{jmt} = a_j + b_j mvar_{jt} + c_j loy_{jmt} + \sum_{l \neq j} V_{j,l}\, y_{lmt} \tag{3}$$

$\varphi(Z)$ denotes the binomial logistic function $1/(1 + exp(-Z))$. Z_{jmt} can be interpreted as latent variable referring to category j and market basket t of household m.

Maximum likelihood estimation of the MVL model requires computation of the so-called normalization constant obtained by summing over all possible market baskets (see expression (2)) in each iteration. For the 29 categories in our study, we would have to deal with more than 5.36×10^8 possible market baskets. Maximum pseudo-likelihood (MPL) estimation (Bel et al. 2018) offers a viable alternative maximizing the log pseudo-likelihood LPL across households, market baskets and categories:

$$LPL = \sum_{m=1}^{M} \sum_{t=1}^{T_m} \sum_{j=1}^{J} \log(\tilde{P}_{jmt}) \tag{4}$$

T_m symbolizes the number of market baskets of household m, $\tilde{P}_{jmt}$ the pseudo-probability of a (non) purchase of category j in market basket t of household m. Summing logarithmic pseudo-probabilities across product categories makes MPL estimation feasible as it replaces summing across all possible baskets, which would be necessary in maximum likelihood estimation. The pseudo-probability $\tilde{P}_{jmt}$ can be written as:

$$\tilde{P}_{jmt} = P(y_{jmt} = 1|y_{-jmt}, x_{mt})^{y_{jmt}} \, (1 - P(y_{jmt} = 1|y_{-jmt}, x_{mt}))^{1-y_{jmt}} \tag{5}$$

Expression (3) shows how to compute the conditional probability $P(y_{jmt} = 1|y_{-jmt}, x_{mt})$ for the homogeneous MVL model. y_{jmt} denotes the binary purchase indicator, which is set to one if basket t of household m contains category j. One can see from Eq. (5) that its first part is relevant if category j is purchased and its second part if category j is not purchased. In a nutshell, LPL estimation looks at J different binomial logit models representing conditional probabilities.

2.2 Taking endogeneity into account

We treat the category-specific marketing variables as potentially endogenous regressors and consider the category loyalties to be exogenous. Our approach is based on

the two-step estimation method that (Park and Gupta 2012) develop for the random coefficient multinomial logit model for individual level data. In a simulation generating data for the random coefficient multinomial logit model, Park and Gupta (2012) demonstrate that under certain assumptions this two-step estimation method provides unbiased coefficients. In contrast to other models dealt with in their simulations, Park and Gupta (2012) include brand constants (which are analogous to intercepts) in the random coefficient multinomial logit model. On the other hand, Park and Gupta (2012) do not investigate whether their method is robust with respect to violations of the assumed Gaussian correlation structure.

In accordance with Sect. 2.1 we replace the random coefficient multinomial logit model by J different random coefficient binomial logit models, one for each product category. To simplify matters, we speak of the random coefficient multivariate logit model (RMVL) in the following to denote this set of J random coefficient binomial logit models.

We specify the latent variable for category j as follows:

$$\begin{aligned} Z_{jmt} &= (\bar{a}_j + \bar{b}_j mvar_{jt} + \xi_{jt}) + a_{jm} + b_{jm} mvar_{jt} + c_j loy_{jt} \\ &\quad + \sum_{l \neq j} V_{j,l}\, y_{lmt} + \epsilon_{mjt} \\ &= \delta_{jt} + a_{jm} + b_{jm} mvar_{jt} + c_j loy_{jt} + \sum_{l \neq j} V_{j,l}\, y_{lmt} + \epsilon_{mjt} \end{aligned} \tag{6}$$

$\bar{a}_j$ and $\bar{b}_j$ denote average coefficients, a_{jm} and b_{jm} their random deviations. Correlation of the structural error ξ_{jt} with the endogenous marketing variable leads to the endogeneity problem. ϵ_{mjt} are Gumbel distributed errors.

In the first step we estimate random deviations a_{jm} and b_{jm} together with fixed loyalty coefficients c_j, fixed interaction coefficients $V_{j,l}$, and fixed coefficients δ_{jt}. The latter are specific to the category j and week t. First step estimates of δ_{jt} can be written as $\hat{\delta}_{jt} = \delta_{jt} + \varsigma_{jt}$ with ς_{jt} being asymptotically normal.

In the second step we regress these estimates on the marketing variable $mvar_{jt}$ of category j and its Gaussian copula correction term $mvar^*_{jt}$:

$$\hat{\delta}_{jt} = \bar{a}_j + \bar{b}_j mvar_{jt} + d_j mvar^*_{jt} \tag{7}$$

The Gaussian copula directly correlates the error term of a structural equation with a non-normally distributed endogenous regressor. This way, the endogenous regressor is treated as a random variable from any (non-normal) marginal population distribution.

In the Gaussian copula approach computation of the correction term of an endogenous regressor starts with its empirical cumulative distribution function. The value of the empirical cumulative distribution function $F(mvar_{j't'})$ of marketing variable for category j and basket t' is (Papadopoulos 2022):

$$F(mvar_{j't'}) = 1/\left(1 + \sum_{m=1}^{M} T_m\right) \sum_{m=1}^{M} \sum_{t=1}^{T_m} I\{mvar_{jt} \leq mvar_{j't'}\} \tag{8}$$

The indicator function I returns one if the condition inside the parentheses is fulfilled, zero otherwise. Therefore $F(mvar_{jm't'})$ equals the number of baskets in which the marketing variable for category j is less equal to $mvar_{jm't'}$ divided by the total number of baskets $\sum_{m=1}^{M} T_m$ plus one. Finally, the correction term $mvar^*_{jm't'}$ results from inserting the value of the empirical cumulative distribution function into the inverse normal cumulative distribution function Φ^{-1}:

$$mvar^*_{j't'} = \Phi^{-1}(F(mvar_{j't'})) \tag{9}$$

Park and Gupta (2012) demonstrate that multiple correction terms have to be included in the case of multiple endogenous regressors, i.e., one term for each regressor. Each of these terms is computed separately based on its empirical cumulative distribution function and the inverse normal cumulative distribution function according to Eqs. (8) and (9).

The copula approach requires that endogenous regressors are not (too) normal to make sure that variation due to the endogenous regressor and variation due to the structural error can be separated. Consistent estimates of average coefficients $\bar{a}_j$ and $\bar{b}_j$ by the second step regression (7) require non-normality of the endogenous variable as well as normality of error terms (Becker et al. 2022; Eckert and Hohberger 2022).

Based on extensive simulations, Becker et al. (2022) propose to assume nonnormality of a variable if its absolute skewness is greater than 0.8 and a nonnormality test provides a conservative (i.e., very low) p-value. Becker et al. (2022) also state that the error term's normality can be checked using the residuals of the second stage regression without the copula correction term.

According to the results of Eckert and Hohberger (2022) significance of the copula correction term in the second stage regression is indicative of endogeneity only when the assumption a Gaussian correlation structure is met. In this case, a positive (negative) significant coefficient shows that the correlation of the respective marketing variable and the excluded variable is positive (negative). Consequently, a positive (negative) omitted variable bias results, as explained in Sect. 1.

The copula approach also implicitly assumes that all exogenous regressors are uncorrelated with the linear combination of copula transformations of endogenous regressors (Yang et al. 2022). In our case, these linear combinations correspond to $d_j mvar^*_{jt}$. To tackle such correlations Yang et al. (2022) develop a two-stage copula endogeneity correction method that consists of adding residuals from regressing copula data for each endogenous regressor on copula data for the exogenous regressors. In a situation with noticeable correlations, the two-stage copula correction method should be used.

2.3 Estimation

We exclude the null basket for which all purchase indicators y_j equal zero in accordance with previous related publications (Russell and Petersen 2000; Boztuğ and Reutterer 2008; Kwak et al. 2015). This way, we model purchases conditional on the

purchase of at least one category. Therefore, the number of possible market baskets is $2^J - 1$.

We estimate the RMVL model by maximum simulated pseudo-likelihood using Halton draws and normal mixing distributions for the random coefficients (Train 2003). The coefficients of the second state regression models are estimated by least squares. We do not base coefficients' tests at this stage on conventional standard errors, which are incorrect for models with copula correction terms because the latter are estimated quantities. Instead, we compute the standard error of any coefficient as standard deviation of its estimates across 500 bootstrap samples (Papies et al. 2017). Each bootstrap sample has $\sum_{m=1}^{M} T_m$ observations, which are drawn with replacement from the original market baskets.

3 Derivation of managerial implications

We investigate whether taking the endogeneity of marketing variables into account leads to different managerial implications. We consider the decision problem of setting a marketing variable. This decision depends on the effect a marketing variable has on purchases of the corresponding category itself, the so-called own effect, as well as on cross effects, i.e., the effects on purchases of other categories.

As our estimation approach provides pseudo-probabilities, we cannot directly determine purchase probabilities and have to resort to simulation. We generate simulated purchases by iterated Gibbs-sampling from the conditional distributions (Besag 2004). The computation times of Gibbs sampling for each observed market basket would be prohibitively high. Therefore, we cluster market baskets by K-means with the category loyalties of each basket $loy_{jmt}, j = 1, \ldots, J$ as clustering variables.

K-means assigns each market basket to exactly one of C clusters. These assignments can be indicated by binary variables u_{cmt}. u_{cmt} is one, if basket t of household m belongs to cluster c. The size s_c of each cluster c is the sum of its assignments, $s_c = \sum_m^M \sum_t^{T_m} u_{cmt}$. The average loyalty in each cluster equals $1/s_c \sum_m^M \sum_t^{T_m} u_{cmt}\, loy_{jmt}$. We obtain cluster-specific purchase probabilities as averages of simulated purchases, which we generate by iterated Gibbs-sampling from the following cluster-specific conditional distribution:

$$P(y_{jc} = 1 | y_{-jc}, x_c) = \varphi(Z_{jc})$$

Average cluster-specific loyalties are part of the regressors x_c. Z_{jc} denotes the latent variable that is specified in analogy to expression (6). Finally, we compute the total purchase probability $P(y_j = 1)$ of each category j as weighted average of cluster-specific purchase probabilities $1/(\sum_m^M T_m) \sum_c^C s_c P(y_{jc} = 1 | y_{-jc}, x_c)$.

In the first run of our simulation approach, we set the marketing variable of a category k to a low value of and estimate the total purchase probability of category j. Then we estimate the total purchase probability of category j by setting the marketing variable of category k to a high value. In both cases, we fix the marketing variables for categories other than category k at their observed average values. The

difference of probability of the category for the marketing variable at the high value and the probability for the marketing variable at the low value gives the change of probability of category j that occurs with the given change of the marketing variable of category k:

$$\Delta P(y_j = 1) = P(y_j = 1|mvar_k = high) - P(y_j = 1|mvar_k = low) \tag{10}$$

Note that the purchase probabilities in Eq. (10) are only conditional on the marketing variable of category k, but not conditional on purchases of categories other than j.

In addition, we compute arc elasticities by dividing the relative change of purchase probability by the relative change of the marketing variable (Allen 1934):

$$\frac{\Delta P(y_j = 1)}{(P(y_j = 1||mvar_k = high) + P(y_j = 1|mvar_k = low))/2} \frac{(high + low)/2}{high - low} \tag{11}$$

Probability changes and elasticities measure own effects if $j = k$, they measure cross effects if $j \neq k$.

4 Empirical study

4.1 Data

Our data refer to 24,047 shopping visits to one specific grocery store over a one-year period made by a random sample of 1500 households. For each shopping visit, we compose a market basket from the IRI data set (Bronnenberg et al. 2008). We represent a market basket by a binary vector whose elements indicate whether a household purchases each of 31 product categories (see Table 1). The average number of shopping visits per household amounts to 16.031, its standard deviation is 13.464. The average basket size (i.e., the number of purchased categories) is 3.852, and its standard deviation 2.654.

Table 2 shows relative marginal purchase frequencies for the 31 categories, and Table 3 the highest 20 pairwise relative frequencies. Milk is the category most frequently purchased. Carbonated beverages and milk are the two categories most frequently purchased together.

The data include the category-specific marketing variables displays, features, and price reductions. These marketing variables Displays are set up in stores. Frequently found variants of displays are front-walls, end-caps, wings, or in-aisle gondolas (Neslin 2002). Products are typically featured by newspaper ads or inserts. These marketing variables are defined as weekly market share-weighted averages of UPC level variables in the respective category and are shares taking values between zero and one.

Displays, features, and price reductions are promotional activities intended to increase sales of a store (Little 1998). Therefore, we expect that average coefficients for these marketing variables are positive. However, estimation of the

Table 1 Product categories and abbreviations

Beer and ale	beer	Blades	blades
Carbonated beverages	carbbev	Cigarettes	cigets
Coffee	coffee	Cold Cereal	coldcer
Deodorant	deod	Diapers	diapers
Facial tissue	factiss	Frozen dinners	fzdin
Frozen pizza	fzpizza	Household cleaners	hhclean
Frankfurters and hotdog	hotdog	Laundry detergent	laundet
Margarine and butter	margbutr	Mayonnaise	mayo
Milk	milk	Mustard and ketchup	mustketc
Paper towels	paptowl	Peanut butter	peanbutr
Photographic supplies	photo	Razors	razors
Salty snacks	saltsnck	Shampoo	shamp
Soup	soup	Spaghetti sauce	spagsauc
Sugar substitutes	sugarsub	Toilet tissue	toitisu
Tooth brush	toothbr	Toothpaste	toothpa
Yogurt	yogurt		

Table 2 Relative marginal frequencies

milk	0.476	carbbev	0.400	saltsnck	0.351	coldcer	0.280
yogurt	0.202	soup	0.197	spagsauc	0.184	toitisu	0.171
margbutr	0.158	paptowl	0.140	coffee	0.136	laundet	0.118
fzpizza	0.110	mayo	0.109	hotdog	0.103	mustketc	0.102
fzdin	0.090	factiss	0.084	peanbutr	0.080	beer	0.076
toothpa	0.059	shamp	0.053	deod	0.039	cigets	0.032
hhclean	0.030	diapers	0.020	blades	0.019	toothbr	0.014
sugarsub	0.011	photo	0.007	razors	0.002		

Table 3 Relative pairwise frequencies

carbbev	milk	0.199	carbbev	saltsnck	0.189	milk	saltsnck	0.176
coldcer	milk	0.154	coldcer	saltsnck	0.128	carbbev	coldcer	0.127
milk	yogurt	0.115	milk	soup	0.107	milk	spagsauc	0.094
carbbev	soup	0.092	milk	toitisu	0.089	carbbev	yogurt	0.089
carbbev	spagsauc	0.088	saltsnck	yogurt	0.088	saltsnck	soup	0.087
coldcer	yogurt	0.087	margbutr	milk	0.086	saltsnck	spagsauc	0.085
carbbev	toitisu	0.084	saltsnck	toitisu	0.080			

Shows the 20 highest relative pairwise frequencies

RMVL model with these marketing variables produces several negative average coefficients. Multicollinearity, reflected by high variance inflation factors for two marketing variables, is responsible for these negative coefficients. Specifically,

we obtain variance inflation factors greater than 5.0 in 23 and 22 categories for features and price reductions, respectively [a threshold of 5.0 corresponding to a R-squared value of 0.80 is recommended by Hocking and Pendelton (1983)]. On the other hand, for displays such a high variance inflation factor occurs in one category only. These results are congruent with the statement of Blattberg and Neslin (1990) that "multicollinearity is endemic in promotion regression models".

Negative average coefficients of features and price reductions run counter to theoretical expectations. They prevent both the comparison of uncorrected to copula-corrected coefficients and the derivation of plausible managerial implications. Therefore, we exclude features and price reductions from the model and only consider the marketing variable display as is least affected by multicollinearity. Table 4 shows the share of displays averaged across all market baskets. We see that carbonated beverages have the highest number of displays, whereas no displays occur for the two categories cigarettes and sugar substitutes. For this reason, we do not consider these two categories in the following analyses.

Table 4 also contains the average loyalty across all market baskets for each category for a smoothing constant $\alpha = 0.2$, which puts more weight on the loyalty of the previous shopping visit. This value of the smoothing constant leads to the best performing RMVL model with category loyalty according to a grid search over $[0.1, 0.2, 0.3, \ldots, 0.9]$. Given such a value, previous purchases are strongly smoothed. Milk attains the highest category loyalty.

Table 4 Average values of displays and category Loyalties

Displays							
beer	0.080	blades	0.090	carbbev	0.283	cigets	0.000
coffee	0.080	coldcer	0.114	deod	0.034	diapers	0.010
factiss	0.048	fzdin	0.007	fzpizza	0.121	hhclean	0.016
hotdog	0.034	laundet	0.081	margbutr	0.026	mayo	0.054
milk	0.009	mustketc	0.054	paptowl	0.071	peanbutr	0.053
photo	0.196	razors	0.206	saltsnck	0.267	shamp	0.077
soup	0.061	spagsauc	0.072	sugarsub	0.000	toitisu	0.081
toothbr	0.031	toothpa	0.045	yogurt	0.020		
Category Loyalties							
beer	0.058	blades	0.014	carbbev	0.307	cigets	0.026
coffee	0.103	coldcer	0.218	deod	0.032	diapers	0.015
factiss	0.065	fzdin	0.070	fzpizza	0.084	hhclean	0.023
hotdog	0.081	laundet	0.092	margbutr	0.119	mayo	0.084
milk	0.359	mustketc	0.081	paptowl	0.109	peanbutr	0.062
photo	0.004	razors	0.001	saltsnck	0.274	shamp	0.041
soup	0.149	spagsauc	0.142	sugarsub	0.009	toitisu	0.133
toothbr	0.010	toothpa	0.046	yogurt	0.161		

Table 5 Evaluation of random coefficient logit models

Model	Log pseudo-likelihood	AIC	BIC
Multivariate logit	− 175,544	355,090	371,274
Independent logit	− 183,353	369,896	382,796

Values are rounded to the nearest integer

Table 6 Pairwise interaction coefficients

paptowl	toitisu	1.122	(21.375)	carbbev	saltsnck	0.750	(20.510)
mayo	mustketc	1.009	(15.936)	toothbr	toothpa	1.882	(13.832)
factiss	toitisu	0.795	(11.960)	coldcer	yogurt	0.507	(11.567)
fzdin	fzpizza	0.812	(11.174)	coldcer	spagsauc	0.431	(9.545)
factiss	paptowl	0.651	(9.334)	coldcer	saltsnck	0.372	(9.305)
blades	deod	1.373	(9.137)	deod	toothpa	0.915	(8.907)
soup	spagsauc	0.416	(8.639)	shamp	toothpa	0.785	(8.125)
laundet	toitisu	0.486	(7.969)	coldcer	milk	0.285	(7.735)
deod	shamp	0.840	(7.718)	laundet	paptowl	0.472	(7.353)
laundet	toothpa	0.602	(7.348)	coldcer	peanbutr	0.453	(7.281)

20 interactions with highest absolute t-values shown in parentheses

4.2 Estimation results

We compare the RMVL model defined in Sect. 2.2 to the related, less complex random coefficient independent logit model that excludes category interactions. We compute the log pseudo-likelihoods of these two models and determine AIC and BIC values to take their different number of parameters into account. Interactions turn out to be important because the RMVL is clearly superior with respect to both AIC and BIC values (see Table 5). That is why we only consider the RMVL in subsequent analyses.

Out of a total of 406 interaction terms, 190, i.e., about 47%, are significant. Of these 190 interaction terms, 188 are positive, and only two are negative (photo and yogurt, beer and cold cereal). Table 6 shows the 20 interaction coefficients with the highest t-values. In these interactions, both categories belong either to the non-food assortment (e.g., paper towels and toilet tissue, tooth brush and toothpaste, facial tissue and toilet tissue) or to the food assortment (e.g., carbonated beverages and salted snacks, mayonnaise and mustard/ketchup, cold cereal and yogurt).

In the following we present the significant average display coefficients of the RMVL, which arise in all but the two categories cigarettes and sugar substitutes. These significant coefficients are all positive, i.e., more displays increase the pseudo-probability of a purchase in the respective category (see Table 7).

The second stage regressions provide nine significant coefficients of the copula correction term for displays (see Table 8). We also assess to what extent the testable requirements of the Gaussian copula correction mentioned in Sect. 2.2 are fulfilled. We obtain evidence for the non-normality of the display variables. In

Table 7 Significant average coefficients for displays

beer	4.927	blades	5.972	carbbev	3.143	coffee	6.126
coldcer	4.680	deod	4.002	diapers	2.953	factiss	5.944
fzdin	4.717	fzpizza	5.715	hhclean	5.775	hotdog	5.835
laundet	5.738	margbutr	4.598	mayo	5.788	musketc	5.960
paptowl	5.236	peanbutr	5.435	photo	8.061	razors	7.208
saltsnck	7.900	shamp	6.063	soup	4.912	spagsauc	4.612
toitisu	4.933	toothpa	5.064	yogurt	3.031		

Lowest t-value amounts to 2.685

Table 8 Significant copula correction coefficients and display coefficients

Category	Coefficient of copula	Display coefficients	
	Correction term	Copula-corrected	Uncorrected
coldcer	− 0.178	5.180	4.680
fzdin	0.030	4.520	4.717
hhclean	0.135	4.518	5.775
mayo	0.053	5.740	5.882
paptowl	0.147	4.956	5.236
saltsnck	0.096	7.660	7.900
shamp	− 0.253	7.824	6.063
spagsauc	0.065	4.377	4.612
toitisu	− 0.058	5.205	4.933

Contains correction term coefficients with absolute t-values ≥ 2.147

each of the 29 categories, displays attain absolute skewnesses greater than 0.8. For 27 categories, the Shapiro–Wilks tests of normality provide a p-value less than 0.10. We especially see very low p-values not greater than 0.0078, for the nine categories with significant copula correction terms.

We also investigate the residuals of the second stage regression without copula correction terms for these nine categories to assess the normality of error terms. Across these nine categories, the highest absolute skewness amounts to 0.373, and the lowest p-value of the Shapiro-Wilks test is 0.141. Therefore, we conclude that their error terms are normal. In one accepts the assumption of a Gaussian correlation structure, these results for the non-normality of displays and for the normality of error terms allow to infer endogeneity from significant coefficients of the correction terms.

Summing up we obtain evidence for the endogeneity of displays in about 31% of the 29 categories investigated. The majority of the nine significant coefficients of copula correction terms are positive. In other words, most of the nine display variables are positively correlated with the omitted variable. These correlations imply positive omitted variable biases, i.e., overestimation of display coefficients,

if they are not corrected. Corrected display coefficients in Table 8) that are lower than their uncorrected counterparts reflect this fact.

One anonymous reviewer asked whether correlations of the endogenous variable display with one of the omitted marketing variables (feature, display) might explain the sign of the omitted variable bias. As a rule, these correlations are positive in the investigated categories. For only two categories we obtain negative correlations with features, which are very low in absolute size. Therefore correlations with the two omitted marketing variables do not explain the signs of biases. Of course, the biases may be due to unknown factors that differ from the two omitted marketing variables.

Across all categories, the maximum correlation in absolute size of the linear combinations of the copula correction term with the exogenous regressor loyalty amounts to 0.061. In view of these low correlations, we do not apply the two-stage copula endogeneity correction method of Yang et al. (2022).

4.3 Managerial implications

We now deal with the question of whether models $M0$ with uncorrected display coefficients and $M1$ with corrected display coefficients entail different managerial implications. We consider the number of purchases as managerial objective for displays in any category. As the number of purchases equals the sum of purchase probabilities across market baskets, we investigate how much probability changes and elasticities due to higher display activities differ between these models. Section 3 explains the computation of purchase probabilities, probability changes, and elasticities. Based on K-means for category loyalties, we choose six clusters. We alternatively set each category display variable to a high value of 0.25 and a low value of 0.05.

Table 9 shows the differences of probability changes and elasticities for purchases of the respective category of at least 0.01 in absolute size. We do not show differences for cross effects, i.e., effects on other categories, as these are all lower than 0.005 in absolute size. The signs of six significant copula correction term coefficients in Table 8 agree with the sign of the differences with respect to probability changes and elasticities in Table 9. Signs of the remaining three significant copula correction term coefficients agree with one of the two differences. To sum up, the differences between the two models are linked to the respective positive or negative omitted variable biases.

If the difference between models $M0$ and $M1$ is positive (negative) and managers rely on uncorrected display coefficients, they would choose a level of display activities in the respective category, which is too high (low). The majority of the differences in Table 9 are positive, i.e., display effects are often overestimated if uncorrected coefficients are used.

To assess the economic relevance of these differences we assume that a supermarket and a smaller convenience store attain 7000 and 13,000 weekly shopping visits, respectively. These numbers of weekly shopping visits are in accordance with published values (EHI 2017; Spar International 2020). In the case of a positive difference of the probability changes of 0.02, management would overestimate the

Table 9 Probability changes and elasticities of displays

Category	M1	M0	M0–M1	Category	M1	M0	M0–M1
Probability changes							
coldcer	0.340	0.294	− 0.046	hhclean	0.037	0.063	0.026
mayo	0.257	0.268	0.011	paptowl	0.203	0.228	0.025
saltsnck	0.531	0.541	0.010	shamp	0.230	0.121	− 0.109
spagsauc	0.251	0.274	0.023	toitisu	0.223	0.201	− 0.022
Elasticities							
coldcer	0.206	0.159	− 0.047	fzdin	0.426	0.439	0.013
fzpizza	0.194	0.259	0.065	hhclean	0.359	0.469	0.110
paptowl	0.295	0.397	0.102	peanbutr	0.463	0.453	− 0.010
saltsnck	0.356	0.383	0.027	shamp	0.473	0.490	− 0.017
spagsauc	0.384	0.413	0.029	toitisu	0.223	0.183	− 0.040
toothpa	0.333	0.237	− 0.096				

Probability changes with a minimum absolute difference between models of 0.01

Elasticities with a minimum absolute difference between models of 0.01

number of weekly purchases in the displayed category by 140 and 260, respectively. Consequently, management would set display activities at a level that is too high. For a negative difference of − 0.02, management would underestimate the number of weekly purchases by the same amounts (140 and 260) and opt for a level of display activities that is too low.

5 Conclusion

Under the assumption of a Gaussian correlation structure explained in Sect. 1, the two-step Gaussian copula approach indicates endogeneity of the marketing variable display in about 31% of the 29 categories investigated. The majority of display variables rated as endogenous are positively correlated with the omitted variable. Such positive correlations imply that ignoring endogeneity leads to overestimation of the coefficients of the respective display variables.

The own effects of increasing a promotion activity on purchases of the same category frequently differ between a model with display coefficients corrected by the Gaussian copula approach and the related model with uncorrected coefficients. On the other hand, both models agree on the size of the cross effects that display activities exert on purchases of other categories. In a similar manner to the results for model coefficients, the own effects are often overestimated if they rely on uncorrected display coefficients. Such overestimation tempts managers to set a level of display activities that is too high.

Future research efforts may investigate the endogeneity of marketing variables in multicategory choice models in other contexts than food retailing (e.g., consumer electronics, apparel). The fact that we only look at purchases constitutes a limitation of our study. One could also consider response variables such as store

choice, purchase quantity, and brand choice. Of course, adding response variables would lead to more encompassing multicategory models, but also to higher model complexity.

Author Contributions The authors contributed equally to the research and preparation of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability statement The dataset generated during the current study is not publicly available as it contains proprietary information that the authors acquired through a license. Information on how to obtain it and reproduce the analysis is available from the corresponding author on request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Allen RGD (1934) The concept of arc elasticity of demand. J Econ Stud 1:226–229

Andrews RL, Ebbes P (2014) Properties of instrumental variables estimation in logit-based demand models: finite sample results. J Model Manag 9:261–289

Atefi Y, Ahearne M, Maxham JG III et al (2018) Does selective sales force training work? J Mark Res 55:722–737

Aurier P, Mejia V (2014) Multivariate logit and probit models for simultaneous purchases: presentation, uses, appeal and limitations. Rech Appl Mark 29:79–98

Becker JM, Proksch D, Ringle CM (2022) Revisiting Gaussian copulas to handle endogenous regressors. J Acad Mark Sci 50:46–66

Bel K, Fok D, Paap R (2018) Parameter estimation in multivariate logit models with many binary choices. Econ Rev 37:534–550

Besag J (1972) Nearest-neighbour systems and the auto-logistic model for binary data. J R Stat Soc B 34:75–83

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc B 35:192–236

Besag J (2004) An introduction to Markov chain Monte Carlo methods. In: Johnson ME, Khudanpur SP, Ostendorf M et al (eds) Mathematical foundations of speech and language processing. Springer, New York, pp 247–270

Besanko D, Gupta S, Jain D (1998) Logit demand estimation under competitive pricing behavior: An equilibrium framework. Manag Sci 44:1533–1547

Blattberg R, Neslin SA (1990) Sales Promotion. Prentice-Hall, Englewood Cliffs

Boztuğ Y, Hildebrandt L (2008) Modeling joint purchases with a multivariate MNL approach. Schmalenbach Bus Rev 60:400–422

Boztuğ Y, Reutterer T (2008) A combined approach for segment-specific market basket analysis. Eur J Oper Res 187:294–312

Boztuğ Y, Silberhorn N (2006) Modellierungsansätze in der Warenkorbanalyse im Überblick. J Betr Wirtsch 56:105–128
Bronnenberg BJ, Kruger MW, Mela CF (2008) Database paper: the IRI marketing data set. Mark Sci 27:745–748
Chib S, Seetharaman PB, Strijnev A (2002) Analysis of multi-category purchase incidence decisions using IRI market basket data. In: Franses PH, Montgomery AL (eds) Econometric models in marketing. JAI, Amsterdam, pp 57–92
Consumer Index (2020) Consumer Index der GfK: Bezahlte Preise ziehen im März deutlich an. URL advance.lexis.com/api/document?collection=news &id=urn:contentItem:5YRK-JFV1-F06W-T35V-00000-00 &context=1516831. Accessed 23 Aug 2023
Cox DR (1972) The analysis of multivariate binary data. J R Stat Soc C 21:113–120
Datta H, Foubert B, van Heerde HJ (2015) The challenge of retaining customers acquired with free trials. J Mark Res 52:217–224
Datta H, Ailawadi KL, van Heerde HJ (2017) How well does consumer-based brand equity align with sales-based brand equity and marketing mix response? J Mark 81:1–20
Dippold K, Hruschka H (2013) A model of heterogeneous multicategory choice for market basket analysis. Rev Mark Sci 11:1–31
Duvvuri SD, Ansari V, Gupta S (2007) Consumers' price sensitivities across complementary categories. Manag Sci 53:1933–1945
Ebbes P, Wedel M, Böckenholt U et al (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. Quant Market Econ 3:365–392
Ebbes P, Wedel M, Böckenholt U (2009) Frugal IV alternatives to identify the parameter for an endogenous regressor. J Appl Econ 24:446–468
Eckert C, Hohberger J (2022) Addressing endogeneity without instrumental variables: an evaluation of the Gaussian copula approach for management research. J Manag 49:1460–1495
EHI (2017) EHI handelsdaten aktuell 2017. EHI Retail Institute, Köln
Elshiewy O, Boztuğ Y (2018) When back of pack meets front of pack: how salient and simplified nutrition labels affect food sales in supermarkets. J Public Policy Mark 37:55–67
Guadagni PM, Little JDC (1983) A logit model of brand choice calibrated on scanner data. Mark Sci 2:203–238
Hocking RR, Pendelton OJ (1983) The regression dilemma. Commun Stat 12:497–527
Hruschka H (2017) Analyzing the dependences of multicategory purchases on interactions of marketing variables. J Bus Econ 87:295–313
Hruschka H (2021) Comparing unsupervised probabilistic machine learning methods for market basket analysis. Rev Manag Sci 15:497–527
Hruschka H, Gerhardt RG (2012) Endogeneity of store attributes in heterogeneous store-level sales response models. OR Spectr 31:199–214
Keller WIY, Deleersnyder B, Gedenk K (2019) Price promotions and popular events. J Mark 83:73–88
Kwak K, Duvvuri SD, Russell GJ (2015) An analysis of assortment choice in grocery retailing. J Retail 91:19–33
Lewbel A (1997) Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R and D. Econometrica 65:1201–1213
Lewbel A (2012) Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. J Bus Econ Stat 30:67–80
Little JDC (1998) Integrated measures of sales, merchandising and distributions. Int J Res Mark 15:475–485
Manchanda P, Ansari A, Gupta S (1999) The shopping basket: a model for multi-category purchase incidence decisions. Mark Sci 18:95–114
Neslin S (2002) Sales promotion. In: Weitz BA, Wensley R (eds) Handbook of marketing. Sage, Thousand Oaks, pp 310–338
Papadopoulos A (2022) Accounting for endogeneity in regression models using copulas: a step-by-step guide for empirical studies. J Econ Methods 11:127–154
Papies D, Ebbes P, van Heerde H (2017) Addressing endogeneity in marketing models. In: Leeflang PSH, Wieringa JE, Bijmolt THA et al (eds) Advanced methods for modeling markets. Springer International Publishing, New York, pp 581–627
Park S, Gupta S (2012) Handling endogenous regressors by joint estimation using copulas. Mark Sci 31:567–586

Petrin A, Train K (1999) A control function approach to endogeneity in consumer choice models. J Mark Res 47:3–13
Richards TJ, Hamilton SF, Yonezkawa K (2018) Retail market power in a shopping basket model of supermarket competition. J Retail 94:328–342
Russell GJ, Petersen A (2000) Analysis of cross category dependence in market basket selection. J Retail 76:69–392
Seetharaman PB, Chib S, Ainslie A et al (2005) Models of multi-category choice behavior. Mark Lett 16:239–254
Solnet D, Boztuğ Y, Dolnicar S (2016) An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue. Int J Hosp Manag 56:119–125
Spar International (2020) Spar International Geschäftsbericht. https://spar-international.com/wp-content/uploads/2021/07/German-SPAR-Annual-Review-2020-1.pdf. Accessed 9 June 2023
Train KE (2003) Discrete choice methods with simulation. Cambridge University Press, Cambridge
Villas-Boas JM, Winer RS (1999) Endogeneity in brand choice models. Manag Sci 45:1324–1338
Wooldridge JM (2010) Econometric analysis of cross section and panel data. MIT Press, Cambridge
Xia F, Chatterjee R, May JH (2019) Using conditional restricted Boltzmann machines to model complex consumer shopping patterns. Mark Sci 38:711–727
Yang F, Qian Y, Xie H (2022) Addressing endogeneity using a two-stage copula generated regressor approach. Technical report 29708, NBER working paper