

基于半参数广义可加模型的社会消费品 零售总额影响因素研究

凌 兰 兰

(安徽建筑大学 数理学院, 合肥 230601)

摘要:基于半参数广义可加模型的思想,对全国社会消费品零售总额的影响因素进行了研究。首先运用灰色关联分析和单变量分析筛选出 2 个非线性因素和 3 个线性因素,发现两个非线性项之间存在交互效应,最终建立含有交互项的半参数广义可加模型,且模型的 R^2 达到 97.7%,说明模型具有极高的解释能力。结果表明,M1、GDP 和农村家庭居民人均可支配收入对全国社会消费品零售总额具有正线性影响;M2 和全国人均可支配收入具有非线性影响;人均可支配收入较低时,M2 越大会使得社会消费品零售总额越低。

关键词:半参数广义可加模型;灰色关联分析;社会消费品零售总额

中图分类号:F812.41 **文献标志码:**A **文章编号:**1671-1807(2023)16-0240-05

现阶段我国经济正处于高速增长阶段转向高质量发展阶段,主要矛盾也已发生重大改变。为解决人民日益增长的美好生活的需要和不平衡不充分发展之间的矛盾,必须坚定实施扩大内需战略。

作为反映全社会消费情况的主要指标,社会消费品零售总额涉及范围较广,不仅包括售给个人用于生活消费用的商品和建筑材料,也包括售给社会团体用作非生产、非经营的商品和餐饮服务所取得的收入金额等。它不仅直接反映了国内消费需求,还可以衡量社会经济景气程度。扩大国内投资、刺激国内消费和扩大外贸出口常被称为经济增长的“三驾马车”,只有保证经济发展才能扩大消费,积极促进国家或地区经济增长。随着国内城镇化扩大和人均可支配收入增加,消费在经济增长中的重要性日渐凸显。世界各国尤其是发达国家,为了发展经济,不断强调扩大国内消费、促进消费升级、培育消费热点。因此,研究社会消费品零售总额的影响因素对处在发展关键期的我国具有重要的经济意义。

1 文献综述

关于影响因素的选取,杨远裴^[1]从消费需求、商品供给等 4 个层面初步选取了 10 个变量,用随机森林回归模型从中选择出了 6 个主要影响变量,包括广义货币供应量(M2)、狭义货币供应量(M1)、国内

生产总值(GDP)固定资产投资额、邮电业务量和商品房销售额,使用这 6 个变量构建了含有自变量的动态回归模型进行预测;李小星和徐永利^[2]采用函数型数据分析方法对数据进行了预测,将数据分为长期趋势成分和季节性成分分别加以预测;梁红梅和赵宏宝^[3]研究发现人均可支配收入可以正向推动社会消费品零售总额的增长;马强等^[4]基于乘积季节模型对数据进行时间序列分析,确定最终模型为 ARIMA(1,1,1)(1,1,0);蒋翠清和乔晗^[5]在构建安徽省社会消费品零售总额累计增幅预测模型时,融合了股市数据;韩玉锦^[6]建立了灰色预测模型 GM(1,1)预测新冠病毒对武汉地区社会消费品零售总额的影响值大小。目前关于该问题的研究多采用多元线性回归和时间序列方法,但多元线性回归法往往忽略自变量之间非线性的因果关系。为克服这个缺陷,本文首次将半参数回归模型运用到对社会消费品零售总额的影响因素的研究。

由于半参数回归模型具有高度的灵活性和适应性,国内外将其应用到各个领域的研究越来越多。关于半参数回归方法的详细介绍及求解过程,梅长林和王宁^[7]、Ferraty 和 Vieu^[8]均在其著作中有详细说明;黄宏超等^[9]分别用半参数回归模型和

收稿日期:2023-04-13

基金项目:安徽省高等学校科学研究重点项目(2022AH050247);安徽建筑大学科研项目(2016QD118)。

作者简介:凌兰兰(1997—),安徽安庆人,安徽建筑大学数理学院,硕士研究生,研究方向为数据分析中的统计方法及应用。

传统参数模型来描述华北落叶松和白桦树高与胸径的关系,比较发现半参数回归模型的拟合效果明显更佳;马馨悦^[10]为研究余额宝的影响因素,分析变量的线性或非线性影响及交互作用,从而建立了半参数可加模型;李呈呈等^[11]将半参数分位数回归模型应用于老年人医疗费用的研究;鲁万波和陈映彤^[12]基于半参数误差修正模型对中国市场汇率进行预测研究。然而在已有半参数回归研究中,不少学者忽略了变量间的交互作用。

选取 2016 年 7 月至 2022 年 6 月的月度数据,经过灰色关联分析和单变量非线性效应分析,筛选确定 5 个影响因素,将其中的 2 个非线性影响因素作为半参数广义可加模型的非线性部分,并考虑了它们之间是否存在交互效应,发现考虑交互作用而建立的模型比不考虑交互作用的模型解释性更好,从而建立了含有交互项的半参数广义可加模型。

2 变量选取与指标设计

通过对相关文献的总结、比较,不难发现社会消费品零售总额受一些经济指标的影响,故初步筛选出 7 个经济自变量如表 1 所示。

表 1 变量选取与含义

变量	含义
x_1	M1,狭义货币供应量
x_2	M2,广义货币供应量
x_3	GDP,国内生产总值
x_4	全国居民人均可支配收入
x_5	农村家庭居民人均可支配收入
x_6	CPI,消费者物价指数
x_7	美元兑人民币汇率

选取的数据为 2016 年 7 月至 2022 年 6 月的月度数据,来自国家统计局和中国人民银行。鉴于 GDP、全国居民人均可支配收入和农村家庭居民人均可支配收入均是季度数据,故借用 Eviews 软件的插值法将其转化为月度数据。考虑到国家统计局不单独开展 1 月和 2 月的零售总额数据,故剔除每年 1~2 月的数据,对剩下 60 个月的数据进行建模研究。

3 理论模型

3.1 灰色关联分析

灰色关联分析方法被用于描述因素之间变化趋势的相似或相异程度,又称为“灰色关联度”,若变化趋势较为一致,则二者关联程度较大;反之,则较小。该方法的具体计算步骤如下。

第一步:确定母序列和子序列。

母序列:

$$X_0^* = [x_0(1) \quad x_0(2) \quad \cdots \quad x_0(m)] \tag{1}$$

子序列:

$$[X_1^* \quad X_2^* \quad \cdots \quad X_n^*] = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ x_1(2) & x_2(2) & \cdots & x_n(2) \\ \vdots & \vdots & & \vdots \\ x_1(m) & x_2(m) & \cdots & x_n(m) \end{bmatrix} \tag{2}$$

第二步:对母序列和子序列进行初值化处理。

$$X_0 = \left[1 \quad \frac{x_0(2)}{x_0(1)} \quad \cdots \quad \frac{x_0(m)}{x_0(1)} \right], \tag{3}$$

$$[X_1 \quad X_2 \quad \cdots \quad X_m] = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \frac{x_1(2)}{x_1(1)} & \frac{x_2(2)}{x_2(1)} & \cdots & \frac{x_n(2)}{x_n(1)} \\ \vdots & \vdots & & \vdots \\ \frac{x_1(m)}{x_1(1)} & \frac{x_2(m)}{x_2(1)} & \cdots & \frac{x_n(m)}{x_n(1)} \end{bmatrix} \tag{4}$$

第三步:计算子序列中各个指标与母序列相对应指标的关联系数。

$$\eta_j(k) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \rho \max_i \max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_j(k)| + \rho \max_i \max_k |x_0(k) - x_i(k)|} \tag{5}$$

式中: $i, j = 1, 2, \cdots, n, ; k = 1, 2, \cdots, m$; $\eta_j(k)$ 为 $X_0(k)$ 和 $X_j(k)$ 间的关联系数; $\rho \in [0, 1]$, 一般取 0.5。

第四步:计算第 j 个对象的灰色关联度。

$$r_j = \frac{1}{m} \sum_{k=1}^m \eta_j(k) \tag{6}$$

3.2 半参数广义可加模型

在许多实际问题中,某个变量 Y 往往与变量 $X_1, X_2, \cdots, X_p$ 有关,但这种相关关系经常不甚明确,因此只能说 Y 部分由 $X_1, X_2, \cdots, X_p$ 的取值决定。这时我们可以认为 Y 的取值由两部分构成:一部分由 $X_1, X_2, \cdots, X_p$ 通过某个函数关系 $m(X_1, X_2, \cdots, X_p)$ 确定;另一部分为随机因素,记为 ϵ , 于是 Y 与 $X_1, X_2, \cdots, X_p$ 的关系可表示为

$$Y = m(X_1, X_2, \cdots, X_p) + \epsilon \tag{7}$$

式中: Y 为因变量; $X_1, X_2, \cdots, X_p$ 称为自变量; $m(X_1, X_2, \cdots, X_p)$ 称为回归函数; Y 与 $X_1, X_2, \cdots, X_p$ 的关系称为回归关系; ϵ 称为误差项,并假定 $E(\epsilon) = 0$ 。如果函数形式是已知的,即

$$m(X_1, X_2, \cdots, X_p) =$$
$$f(X_1, X_2, \cdots, X_p; \beta_1, \beta_2, \cdots, \beta_r) \quad (8)$$

式中： f 是已知函数； $\beta_1, \beta_2, \cdots, \beta_r$ 是未知参数。

式(7)为参数回归模型。然而，如果不假定回归函数 $m(X_1, X_2, \cdots, X_p)$ 的具体形式，只要求 p 元回归函数是连续或光滑的，此时式(7)为非参数回归模型。这类模型具有高度的灵活性，但是当自变量的维数增加时，常常会出现难以控制的维数灾难问题，为了克服该缺陷，半参数广义可加模型应运而生。

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \sum_{i=1}^r f_i(X_{p+i}) + \epsilon \quad (9)$$

式中： $r \geq p + 1, p \geq 1; \beta_0$ 为截距项； ϵ 为误差项； $\sum_{j=1}^p \beta_j X_j$ 为模型的线性部分； $\sum_{i=1}^r f_i(X_{p+i})$ 为模型的非参数部分。

4 实证分析

根据上面对灰色关联度计算步骤的描述，借助 Matlab 软件编程求解，得到各相关因子与社会消费品零售总额的灰色关联度(表 2)。

表 2 灰色关联度

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y	0.82	0.77	0.78	0.79	0.71	0.63	0.62

由表 2 可知，影响社会消费品零售总额的相关因子程度为： $x_1 > x_4 > x_3 > x_2 > x_5 > x_6 > x_7$ ，且前 5 个相关因子的影响程度显著大于后 2 个。

4.1 单变量分析

当因变量不通过正态性检验时，即不能建立线性回归模型，此时建立半参数回归模型才具有意义，通过 QQ 图(图 1)来检验社会消费品零售总额的正态性。

鉴别样本数据是否近似于正态分布，只需看 QQ 图上的点是否近似地在 45° 的直线上。若散点图是直线则说明是正态分布，且直线的斜率和截距分别表示标准差和均值。显然图 1 中大部分数据偏

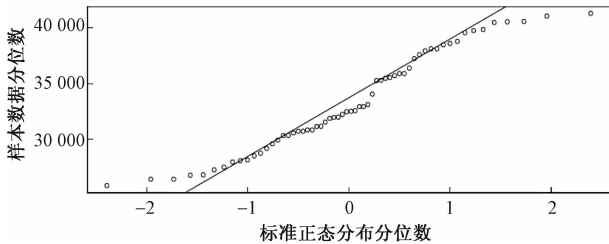


图 1 正态检验 QQ 图

离该直线，所以社会消费品零售总额不服从正态分布，可以采用半参数回归模型对其进行建模。

接下来计算各自变量与社会消费品零售总额的线性相关系数，由 R 软件的 $\text{cor}(\cdot)$ 函数实现(图 2)。

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y	1.00	0.70	0.60	0.90	0.57	0.64	-0.10	-0.16
x_1	0.70	1.00	0.97	0.85	0.82	0.70	-0.13	-0.27
x_2	0.60	0.97	1.00	0.82	0.83	0.69	-0.10	-0.26
x_3	0.90	0.85	0.82	1.00	0.67	0.65	-0.02	-0.15
x_4	0.57	0.82	0.83	0.67	1.00	0.95	-0.11	-0.27
x_5	0.64	0.70	0.69	0.65	0.95	1.00	-0.12	-0.25
x_6	-0.10	-0.13	-0.10	-0.02	-0.11	-0.12	1.00	0.60
x_7	-0.16	-0.27	-0.26	-0.15	-0.27	-0.25	0.60	1.00

图 2 各变量间的相关系数

显然 x_6, x_7 与 y 的相关系数极小，因此 CPI、美元兑人民币汇率与社会消费品零售总额间不存在线性关系，于是考虑它们之间是否存在非线性关系。

由表 3 和表 4 的最后一列 P 可知，在显著性水平为 0.05 情况下， x_6 和 x_7 的非线性效应不显著，因此 x_6, x_7 与 y 也不存在非线性关系，这与表 1 灰色关联度计算结果相一致。 x_6, x_7 与 y 的关联度最小，于是将这 2 个变量剔除，对剩下 5 个自变量进行研究。

由图 2 可知， x_3 与 y 的相关系数达到 0.9，两者之间存在高度正相关，因此将 x_3 作为半参数广义可加模型的线性部分。接下来考虑其他 4 个自变量与因变量之间的关系。

表 3 x_6 单变量分析结果

参数	参数估计	标准误	t	$P(> t)$
截距项	33 282	469	70.97	$< 2 \times 10^{-16} ***$
变量	估计自由度	参考自由度	F	$P(> F)$
$f_6(x_6)$	2.386	2.993	2.535	0.06

注：*表示 $P < 0.1$ ，**表示 $P < 0.05$ ，***表示 $P < 0.01$ 。

表 4 x_7 单变量分析结果

参数	参数估计	标准误	t	$P(> t)$
截距项	33 281.9	478.3	69.58	$< 2 \times 10^{-16} ***$
变量	估计自由度	参考自由度	F	$P(> F)$
$f_7(x_7)$	3.585	4.432	1.483	0.225

注：*表示 $P < 0.1$ ，**表示 $P < 0.05$ ，***表示 $P < 0.01$ 。

图 3 和图 4 分别是考虑 x_1 、 x_4 作为单变量时对因变量的非线性影响。实线越弯曲表明非线性作用越显著。显然, x_1 非线性效果不显著, x_5 同样不显著, x_2 与 x_4 的非线性效果较显著, 恰好印证了它们与 y 的线性相关系数最小。因此, 选取 x_2 与 x_4 作为模型的非参数部分, x_1 、 x_3 与 x_5 作为线性部分, 得到部分回归结果如表 5 所示。

其中输出 R^2 结果为 0.96, 表明模型的解释能力为 96%, 可解释偏差输出结果为 97.1%, 表明模型的可解释偏差为 97.1%, 两项结果可以看出模型的解释能力很好。由表 5 和表 6 可知, M1, GDP 和农村家庭居民人均可支配收入对社会消费品零售总额均具有正线性影响, 且所有变量均通过显著性检验, 所以可初步得到模型为

$$y = -62\,900 + 0.109\,9x_1 + 0.178\,1x_3 + 15.33x_5 + f_1(x_2) + f_2(x_4) + \epsilon \tag{9}$$

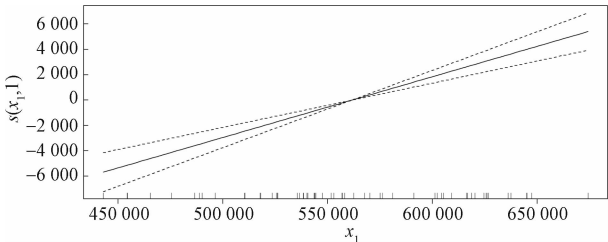


图 3 x_1 单变量分析效应

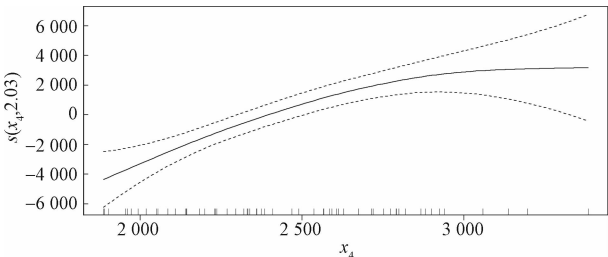


图 4 x_4 单变量分析效应

表 5 模型的线性部分回归结果

变量	参数估计	标准误	t	$P(> t)$
截距项	-6.290×10^4	7.928×10^3	-7.934	$5.50 \times 10^{-10}^{***}$
x_1	1.099×10^{-1}	1.645×10^{-2}	6.678	$3.62 \times 10^{-8}^{***}$
x_3	1.781×10^{-1}	3.668×10^{-2}	4.855	$1.60 \times 10^{-5}^{***}$
x_5	15.33	4.013	3.821	0.000 419 ***

注: * 表示 $P < 0.1$, ** 表示 $P < 0.05$, *** 表示 $P < 0.01$ 。

表 6 模型非线性部分回归结果

变量	估计自由度	参考自由度	F	$P(> F)$
$f_1(x_2)$	8.105	8.739	11.046	$< 2 \times 10^{-16}^{***}$
$f_2(x_4)$	4.481	5.537	2.526	0.037 1 *

注: * 表示 $P < 0.1$, ** 表示 $P < 0.05$, *** 表示 $P < 0.01$ 。

4.2 交互作用分析

值得注意的是, 非参数广义可加模型的非线性变量间有时会存在交互效应现象, 对两个自变量而言, 这种现象表现为一个自变量对因变量的效应会随着另一个自变量的变化而变化。于是对 x_2 和 x_4 进行交互作用分析, 得到交互作用效应如图 5 所示。

表 7 显示在考虑交互作用时, 所有自变量包括 x_2 和 x_4 的交互项均通过显著性检验, 且模型输出 R^2 结果为 0.977, 表明模型的解释能力为 97.7%, 可解释偏差输出结果为 98.4%, 表明模型的可解释偏差为 98.4%, 相较于模型(9)效果更好。图 5 纵轴表示因变量预测值, 可以看出 M2 和全国人均可支配收入交互效用明显。人均可支配收入较低时, M2 越大会使得社会消费品零售总额越低; 在 M2 接近 260 000 元/人、人均可支配收入略小于 2 000 元时, 社会消费品零售总额最低; M2 较大时, 人均可支配收入越接近于 2 500 元/人, 社会消费品零售总额越高; M2 和人均可支配收入同时较低时, 会使得社会消费品零售总额达到极大值。

综上, 关于全国社会消费品零售总额的影响因素分析, 最终建立模型。

$$y = -40\,000 + 0.069\,19x_1 + 0.352\,9x_3 + 4.61x_5 + f_1(x_2) + f_2(x_4) + f_3(x_1, x_4) + \epsilon \tag{10}$$

表 7 交互作用分析结果

变量	参数估计	标准误	t	$P(> t)$
截距项	-4.000×10^4	5.949×10^3	-6.725	$3.89 \times 10^{-8}^{***}$
x_1	6.919×10^{-2}	1.209×10^{-2}	5.721	$1.05 \times 10^{-6}^{***}$
x_3	3.529×10^{-1}	2.265×10^{-2}	15.584	$< 2 \times 10^{-16}^{***}$
x_5	4.610	2.333	1.976	0.005 49 **
变量	估计自由度	参考自由度	F	$P(> F)$
$f_1(x_2)$	1.460	1.611	27.647	$9.56 \times 10^{-7}^{***}$
$f_2(x_4)$	3.013	3.209	4.326	0.006 89 **
$f_3(x_2, x_4)$	10.201	11.435	9.043	$< 2 \times 10^{-16}^{***}$

注: ** 表示 $P < 0.05$, *** 表示 $P < 0.01$ 。

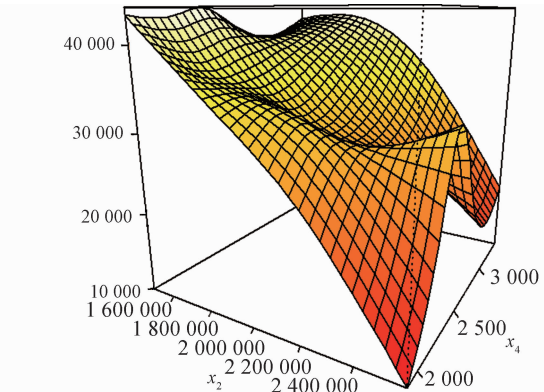


图 5 x_2 和 x_4 的交互作用效应

5 结论

基于半参数广义可加模型的思想,本文对全国社会消费品零售总额的影响因素进行了研究。首先对初步选取的 7 个可能影响因素进行了灰色关联分析和相关系数计算,并对每个影响因素都单独进行了非线性效应分析。结果表明,CPI 和美元兑人民币汇率 2 个影响因子与社会消费品零售总额之间不存在线性或非线性关系,而这也恰好印证了这 2 个影响因子与社会消费品零售总额之间的灰色关联度最低。

上述分析发现 M1、GDP 和农村家庭居民人均可支配收入 3 个自变量对因变量均具有正线性影响,M2 和全国人均可支配收入 2 个自变量对因变量具有非线性影响,其中 GDP 和社会消费品零售总额高度正相关,因此可以初步建立模型(9)。

进一步考虑到两个非线性项之间可能存在交互效应,于是对它们进行交互效应分析。结果显示,M2 和全国人均可支配收入交互效用显著:人均可支配收入较低时,M2 越大会使得社会消费品零售总额越低;M2 较大时,人均可支配收入越接近于 2 500 元/人,社会消费品零售总额越高。因此,加入 M2 和全国人均可支配收入的交互项重新建模得到模型(10)。通过比较,得到模型(10)的判定系数 R^2 和可解释偏差结果都明显优于模型(9),说明模型(10)的解释能力更强,因此最终选定模型(10)。

参考文献

[1] 杨远裴. 社会消费品零售总额影响因素分析及预测研究[D]. 北京:北京工业大学,2021.

[2] 李小星,徐永利. 基于函数性数据分析的中国社会消费品零售总额数据的预测[J]. 北京化工大学学报(自然科学版),2018,45(3):107-112.

[3] 梁红梅,赵宏宝. 网络购物、人均可支配收入对社会消费品零售总额的动态关系研究——基于 Johansen 检验与 VEC 模型的分析[J]. 重庆理工大学学报(自然科学),2020,34(8):203-209.

[4] 马强,张琛婕,陈雪平,等. 基于乘积季节模型的社会消费品零售总额分析[J]. 数学的实践与认识,2021,51(6):87-94.

[5] 蒋翠清,乔晗. 融合股市数据的社会消费品零售总额预测方法[J]. 统计与决策,2022,38(2):151-154.

[6] 韩玉锦. 新冠肺炎疫情对经济会有多大影响——基于社会消费品零售总额的角度[J]. 产业经济评论,2020(2):13-23.

[7] 梅长林,王宁. 近代回归分析方法[M]. 北京:科学出版社,2012.

[8] FERRATY F, VIEU P. Nonparametric functional data analysis[M]. Berlin:Springer,2006.

[9] 黄宏超,谢栋博,段光爽,等. 华北落叶松和白桦半参数树高曲线模型[J]. 林业科学,2022(10):101-110.

[10] 马馨悦. 互联网金融产品收益率分布及影响因素的统计建模研究[D]. 南京:南京财经大学,2019.

[11] 李呈呈,刘红慧,陈曦来,等. 基于半参数分位数回归模型的老年人医疗费用影响因素研究[J]. 中国卫生统计,2022,39(4):513-517.

[12] 鲁万波,陈映彤. 人民币汇率的混频分析与预测——基于半参数误差修正模型[J]. 统计与信息论坛,2020,35(12):3-14.

Research on Influencing Factors of Total Retail Sales of Consumer Goods Based on Semi-Parametric Generalized Additive Model

LING Lanlan

(School of Mathematics and Physics, Anhui Jianzhu University, Hefei 230601, China)

Abstract: Based on the idea of a semi-parametric generalized additive model, the factors influencing the national total retail sales of consumer goods were examined. Firstly, grey relational analysis and univariate analysis were used to screen out two nonlinear factors and three linear factors, and interaction effects were found between the two nonlinear terms. Finally, a semi-parametric generalized additive model containing interaction terms was established, and R-squared of the model reaches 97.7%, indicating that the model has extremely high explanatory ability. The results are as follows. M1, GDP and per capita disposable income of rural households have positive linear effects on the total retail sales of consumer goods in China. The influence of M2 and national per capita disposable income is non-linear. When per capita disposable income is low, the larger M2 is, the lower the total retail sales of consumer goods are.

Keywords: semi-parametric generalized additive model; grey correlation analysis; total retail sales of consumer goods