

基于主成分分析对新零售销售量灰色预测GM(1,1)模型

沈乔羽, 卫俊峰, 冯爱芬

河南科技大学数学与统计学院, 河南 洛阳
Email: 1114356618@qq.com

收稿日期: 2020年11月27日; 录用日期: 2020年12月22日; 发布日期: 2020年12月31日

摘 要

基于预先处理得到的前九个月销售量数据性质离散 ∞ , 通过建立灰色预测GM(1,1)模型, 预测后三个月的销售量并进行误差分析。再结合主成分分析, 通过降维来更好地突出模型的优点, 提出主成分回归-灰色预测模型, 减少了回归分析所需要考虑的变量个数, 并使预测结果具有较好的准确性, 具有较好应用前景。本文结合主成分分析、灰色预测的优点, 通过降维将粗糙但相关性高的成分组合, 预测销售量, 给出进货建议。

关键词

灰色预测, 主成分分析, 回归预测

GM(1,1) Model for Grey Prediction of New Retail Sales Based on Principal Component Analysis

Qiaoyu Shen, Junfeng Wei, Aifen Feng

Mathematics and Applied Mathematics, Henan University of Science and Technology, Luoyang Henan
Email: 1114356618@qq.com

Received: Nov. 27th, 2020; accepted: Dec. 22nd, 2020; published: Dec. 31st, 2020

Abstract

Based on the discretization of the first nine months sales data obtained by pre-processing, the grey forecast GM(1,1) model is established to predict the sales volume in the last three months

文章引用: 沈乔羽, 卫俊峰, 冯爱芬. 基于主成分分析对新零售销售量灰色预测 GM(1,1)模型[J]. 统计学与应用, 2020, 9(6): 1040-1047. DOI: 10.12677/sa.2020.96109

and to carry out error analysis. Combined with principal component analysis, the advantages of the model are better highlighted by dimensionality reduction, and the principal component regression-gray prediction model is proposed to reduce the number of variables to be considered in regression analysis, and to make the prediction results more accurate. It has a good application prospect. This paper combines the advantages of principal component analysis and grey prediction to predict sales volume and give purchase suggestions by reducing dimension to combine rough but highly correlated components.

Keywords

Grey Prediction, Principal Component Analysis, Regression Prediction

Copyright © 2020 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着我国经济发展,消费市场的逐步发展,市场上的消费模式不断变化,逐渐由“以物为主”转变成“以客为主”的销售模式[1]。消费者的需求日趋多样化、个性化,对应的市场供应也需调整。在此类销售模式和这类特殊需求的推动下,新零售企业的生产模式逐步向多品种、小批量迈进,这让商场内零售店铺里的饰品和玩具等种类变得更加琳琅满目,然而需求小众,单品的购买量小[2]。要求门店种类繁多,数量少。本文要解决的具体问题是以种类为层级给出按月精准需求预测,并给出商场进货建议。

2. 数据来源

本文用到的数据来自某大型商贸平台给定区域内 2018~2019 年不同销售小类的产品销售信息,本文预测给定区域内目标小类在 2019 年 10 月 1 日后 3 个月中每个月的销售量,与实际比较后,给出每个月预测值的平均绝对百分比误差(MAPE)。其中,目标小类为历史销售时间处于 2019 年 6 月 1 日至 2019 年 10 月 1 日内且累计销售额排名前 10 的小类。本文最后给出了部分数据样表。

3. 主成分分析回归模型的建立

3.1. 研究方法

主成分分析在物质科学领域对复杂现象的研究,这是通过将粗糙但相关性高的成分组合,并挑选其中贡献率大的指标来研究现实问题,得以更清晰看待问题。这里用来研究影响目标小类的销售量的因素,合适性比较好[3]。

由于影响目标小类销售的因素很多,各因素之间具有一定的相关性,进行回归预测之前需要采用主成分分析法选取贡献率高的主成分,构造出与主成分之间的回归预测模型[4]。

1. 首先样本自变量矩阵进行标准化处理:

$$\text{按列计算均值 } \bar{x}_j = \frac{1}{n} \sum_{i=1}^n X_{ij} \text{ 和标准差 } S_j = \sqrt{\sum_{i=1}^n (X_{ij} - \bar{x}_j)^2}$$

$$\text{标准化数据 } X_{ij} = \frac{(X_{ij} - \bar{x}_j)}{S_j}$$

- 2. 计算 R 的特征值和特征向量
 - 3. 计算主成分贡献率以及累计贡献率
- 本文通过 matlab 求解矩阵得到主成分分析，结果如表 1 所示。

Table 1. Summary of principal component analysis results
表 1. 主成分分析结果汇总表

特征向量	a_1	a_2	a_3		
	0.702044	0.035057	0.324482	-0.18087	-0.60655
	0.029542	-0.63515	0.188917	-0.68442	0.302636
	-0.70854	-0.05907	0.291035	-0.16851	-0.61756
销售价	-0.05695	0.529225	0.740037	-0.11475	0.394783
折扣	0.03145	-0.55838	0.476093	0.676234	0.057174
特征值	2.249432	1.59588	1.140183	0.014189	0.000315
贡献率	0.449886	0.319176	0.228037	0.002838	6.31E-05
累计贡献率	0.449886	0.769062	0.997099	0.999937	1

利用结果分析，前三个主成分分析已经达到 99.8%，包含了大部分信息部分，所以我们取前三个作为指标。

$$\begin{aligned} F_1 &= 0.7020X_1 + 0.0295X_2 - 0.7085X_3 - 0.0570X_4 + 0.0315X_5, \\ F_2 &= 0.0350X_1 - 0.6352X_2 - 0.0591X_3 + 0.5292X_4 - 0.5584X_5, \\ F_3 &= 0.3245X_1 - 0.1890X_2 - 0.2910X_3 + 0.7400X_4 - 0.4761X_5 \end{aligned}$$

对三主成分解释：

F_1 对 X_1 有着高度的正载荷，对 X_3 高度的负载荷，可以解释为库存与价格平衡度。
 F_2 对 X_2 ， X_5 有着高度的负载荷，对 X_4 有高度的正载荷，可以解释为折扣敏感度。
 F_3 对 X_4 有着中等的正载荷对 X_4 高度的负载荷，可以解释为销售价波动感知。

3.2. 基于主成分分析的回归模型建立

通过以上分析和计算，我们用 3 个新变量来替代 5 个可量化变量[5]。这两个新变量相互独立，且充分反映了原来 5 个影响因素包含的信息。这 3 个主成分体现了之前 5 个可量化指标 99.8%的信息，提取效果较好。

接下来我们用得出的三个新变量用最小二乘法来建立回归模型[6]。将 F 使用主成分得到主成分指标，并将因变量销量 y 标准化，整理得到主成分自变量分析表，如表 2 所示。

Table 2. Analysis table of independent variables
表 2. 自变量分析表

变量	参数估计	标准误差	II 型 SS	F 值	Pr > F
主成分	0.01442	0.06116	354.10450	5917.20	<0.0001
Z_1	0.019679	0.02600	47.00164	785.41	<0.0001
Z_2	0.0008919	0.04624	2.85128	47.65	<0.0001
Z_3	0.00993	0.07208	1.25237	20.93	0.0006

R 检验所得到的 $R^2 = 0.9801$ ，而 R^2 越接近于 1，则说明模型的拟合程度越好，因此本模型的拟合是比较好的。利用 F 检验判断回归方程显著性的时候，则关注目标小类的回归预测方程：

$$y_1 = 0.01442 + 0.019679Z_1 + 0.0008919Z_2 + 0.00993Z_3$$

本文通过 Excel 对 2019 年目标小类 10 月到 12 月的三个月销量的预测(表 3~表 5)。

Table 3. October 2019 forecast table

表 3. 2019 年 10 月份预测表

排名	小标签类	第一主成分 F1	第二主成分 F2	第三主成分 F3	实际销售量	预测销售量	MAPE
1	27050401	32314519.71	-891723.7381	80350779.31	1655852	1433005.358	0.394184227
2	27060804	15139301.78	-382992.1361	36736466.78	664065	662377.8586	0.007441415
3	27112849	527483.776	-554018.624	15397838.44	187408	162786.7742	0.384801016
4	27071209	4298798.884	-58542.36975	9119559.168	163811	175101.0863	0.201868643
5	27164944	650694.7176	-364901.0798	10676765.3	140886	118499.8599	0.46539964
6	27196225	3539560.462	-4619.868083	6372358.732	50705	132928.4265	0.249626387
7	27092025	228160.6045	-221959.1506	6197052.35	110337	65828.75143	0.18150074
8	27123264	75163.03153	-185141.7741	4965868.766	89872	50625.09662	0.27907395
9	27217089	3294134.664	-1563.069709	5860354.844	35347	123017.22	0.264641728
10	27102436	136170.0267	-82627.27109	2396755.833	59757	26405.79453	0.634697559
主成分贡献率		0.449886	0.319176	0.228037	0.997099	0.025822387	

Table 4. November 2019 forecast table

表 4. 2019 年 11 月份预测表

排名	小标签类	第一主成分 F1	第二主成分 F2	第三主成分 F3	实际销售量	预测销售量	MAPE
1	27050401	21112130.2	-1455638.7	75291199.7	1302844	1161809	0.317066
2	27060804	9921830.03	-646670.79	34407199.2	521737	536338.43	0.081971
3	27112849	-2007735.1	-710698.33	15011459.7	175950	108919.72	0.115826
4	27071209	3080372.9	-120781.93	8593101.89	129792	145840.45	0.362159
5	27164944	-1093826.6	-474159.88	10448647.6	130367	81806.768	0.091008
6	27196225	1135959.18	-108416.98	4838154.59	30572	70300.733	0.806234
7	27092025	-725745.96	-279551.07	6016125.65	103520	45208.856	0.649841
8	27123264	-583023.86	-227833.54	4918187.06	84127	37161.081	0.635167
9	27217089	1122746.16	-94918.564	4463802.23	21297	66335.434	0.194119
10	27102436	-335041.99	-111721.72	2324218.82	54837	16386.571	0.053724
主成分贡献率		0.449886	0.319176	0.228037	0.997099		

Table 5. Sales volume in December 2019
表 5. 2019 年 12 月份销售量

小标签类	第一主成分 F1	第二主成分 F2	第三主成分 F3	预测销售量
27050401	33048588.37	-1240071.53	90743003.5	1550335.2
27060804	15073630.47	-564291.911	41354663.44	706782.5
27112849	-871852.125	-674335.795	16068145.89	141798.09
27071209	4108470.682	-113713.972	10224338.2	182276.87
27164944	-165900.315	-429961.164	10933413.65	104920.58
27196225	981196.7375	-129941.094	5126749.578	70101.714
27092025	-197186.539	-257172.664	6365361.586	59098.249
27123264	-69869.1839	-202120.443	5153121.481	49615.284
27217089	987745.4741	-114707.867	4742243.562	66426.028
27102436	-201810.822	-109096.667	2490997.033	20666.876
主成分贡献率	0.449886	0.319176	0.228037	

由此可见，主成分回归预测中，百分比误差不稳定，甚至个别达到了百分之八十，于是接下来我们在主成分回归的基础上建立灰色预测模型。

4. 灰色预测 GM(1,1)模型的建立

4.1. 研究思路方法

灰色预测模型是一种灰色系统理论提出了一种新的分析方法即关联度分析方法[7]，即根据因素之间发展态势的相似或相异程度来衡量因素间关联的程度，它揭示了事物动态关联的特征与程度[8]。目标小类商品销售信息如表 6。

Table 6. Sales information of target small categories of goods
表 6. 目标小类商品销售信息

小类编码	销量 y_1	小类销额 x_1	标价 x_2	库存 x_3	销售价 x_4	折扣 x_5
27050401	4558989	541409433.8	161.34	427314790	118.7565	0.736075
27060804	1866284	221500903.8	111.08	169840399	118.6855	1.068515
27112849	542730	63394592.5	85.53	58539158	116.8069	1.365606
27071209	466616	55344005	178.43	41758141	118.6072	0.664736
27164944	401234	47459953.75	129.29	43647794	118.285	0.914864
27196225	197030	40352573.75	206.50	20943155	204.8042	0.991805
27092025	319424	37214623.75	181.49	33984852	116.5054	0.641945
27123264	251468	30255705	185.77	27630862	120.3163	0.647666
27217089	139553	28742578.75	186.91	13790791	205.9617	1.101942

1. 为了保证建模方法的可行性，需要对已知数据进行必要的检验处理，

设参数数据为: $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n))$

计算数列的级比 $\lambda(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, k = 2, 3, \dots, n$

按照 GM (1,1)模型可以得到预测值:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a} \right) e^{-ak} + \frac{b}{a}, k = 1, 2, \dots, n-1$$

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k), k = 1, 2, \dots, n-1$$

2. 检验预测值

(1) 残差检验:

令残差, 计算 $\varepsilon(k) = \frac{x^{(0)}(k) - \hat{x}^{(1)}(k+1)}{x^{(0)}(k)}, k = 1, 2, \dots, n$

如果残差 < 0.2 , 则可以认为达到一般要求, 如果则认为达到比较高的要求。

(2) 级比偏差检验:

首先由参考数据计算出级比偏差, 再用发展系数 a 求出相应的级比偏差。

$$p(k) = 1 - \left(\frac{1 - 0.5a}{1 + 0.5a} \right) \lambda(k)$$

如果 $p(k) < 0.2$, 则可认为达到一般要求; 如果 $p(k) < 0.1$, 则认为达到较高的要求

4.2. 灰色预测模型求解

选取连续标签号 27092025 作为例子, 对齐 1~9 月份的值进行建立灰色预测模型, 如表 7 所示, 用 matlab 计算解出 10~12 月份的预测值, 并与实际值比较[7], 得出 MAPE。

2019 年 10 到 12 月的销售量预测值分别为 110337, 103520, 106440, 如表 8 所示。

销量的预测以及误差分析如表 9。

Table 7. Small label sales for the first 9 months

表 7. 小标签前九个月销售量

小标签类	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月
27092025	166050	161139	135330	97249	117451	108140	106051	109812	101194

Table 8. Predicted values of small labels from October to December 2019

表 8. 小标签 2019 年 10 月到 12 月所求预测值

10 月总销量	11 月总销量	12 月总销量
110337	103520	106440

Table 9. Sales forecast and error analysis

表 9. 销量的预测以及误差分析

月份	原始	模型	残差	相对误差	级比偏差
1	166050	166050	0	0	0
2	327189	354072	26883	0.1668	0.0035
3	462519	468024	5505	0.0407	-0.1514
4	559768	531475	-28293	0.2909	-0.3457

Continued

5	677219	673270	-3949	0.0336	0.1993
6	785359	776105	-9254	0.0856	-0.0503
7	891410	883940	-7470	0.0704	0.014
8	1001222	1001256	34	0.0003	0.0661
9	1102416	1097454	-4962	0.049	-0.0494
10	1212753	1220435	7682	0.0696	0.1131
11	1316273	1320537	4264	0.0412	-0.0307
12	1422713	1433164	10451	0.0982	0.0595

回归曲线: $y = 4234999.0 - 4068944.0 * \exp(-0.0335519 * t)$

分析依然可以得知,除了个别值有较大的偏差,其余值都在误差 10%以内,说明灰色预测模型较好的符合本模型,个别值可能考虑是特殊因素或者季节因素,为了验证对于 2019 年缺失 1~9 月份的数据的类似离散值进行灰色预测的可行性,我们取值第一组的 27050401 的数值进行验证,数据如表 10~11:

Table 10. Monthly sales in 2018

表 10. 2018 年各月销售量

小类标签	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月
27050401	1840397	2484127	1933451	1836189	2650917	1997669	2103730	2053699	1841812

Table 11. Sales forecast for October-December 2019

表 11. 2019 年 10~12 月份预测销售量

10 月总销量	11 月总销量	12 月总销量
1655852	1302844	1480258

Table 12. Error analysis of 2019 sales volume

表 12. 2019 年销售量误差分析

月份	原始	模型	残差	相对误差	级比偏差
1	1840397	1840397	0	0	0
2	2484127	2270477	213650	0.086	0.2745
3	1933451	2223489	-290038	0.15	-0.2582
4	1836189	2177478	-341289	0.1859	-0.0312
5	2650917	2132456	518461	0.1956	0.3217
6	1997669	2088300	-90631	0.0454	-0.2995
7	2103730	2045100	58630	0.0279	0.0701
8	2053699	2002700	50999	0.0248	-0.0032
9	1841812	1961400	-119588	0.0649	-0.092

回归方程 $y = 1.11569e8 - 1.09729e8 * \exp(-0.0209087 * t)$

由表 12 可以看出,灰色预测的其误差值均在 8%左右,拟合程度较好。

5. 结果分析

发现经过主成分分析回归预测, 和灰色预测 GM(1,1)模型比较, 发现基于主成分分析的回归预测后的灰色预测能够校正主成分分析中的个别偏大误差, 得出的模型结合两种模型的优点可以用于预测销售量, 在进货上根据预测值进货、备货。

基金项目

河南省高校省级大学生创新创业训练计划项目(S202010464052), 河南科技大学 SRTP 项目(2020169) 2019 年河南省高等教育教学改革研究与实践项目(2019SJGLX262), 河南科技大学教育教学改革研究与实践项目(2019YBZD-009)。

参考文献

- [1] 滕树军, 郑惠文, 刘柏森. 基于多元回归分析的超市商品销售影响因素的研究[J]. 全国流通经济, 2018(14): 6-9.
- [2] 黄晓梅. 一元线性回归分析法在超市产品销售中的应用[J]. 科技信息, 2013(11): 77-78.
- [3] Lu, Q., Liu, S.L., Li, W. and Jin, X.J. (2020) Combination of Thermodynamic Knowledge and Multilayer Feed forward Neural Networks for Accurate Prediction of M_s Temperature in Steels. *Materials & Design*, **192**, 108696. <https://doi.org/10.1016/j.matdes.2020.108696>
- [4] 孟昌, 庞燕岭. 北京市出租车需求量测算与数量规制——基于课题组调查问卷和官方数据的测算[J]. 北京社会科学, 2020(5): 4-16.
- [5] Deng, B., Luo, X., Zhang, M., Ye, L. and Chen, Y. (2020) Corrigendum to “Quantitative Detection of Acyclovir by Surface Enhanced Raman Spectroscopy Using a Portable Raman Spectrometer Coupled with Multivariate Data Analysis. *Colloids Surf. B: Biointerfaces*, **173**, 286-294. <https://doi.org/10.1016/j.colsurfb.2018.09.058>
- [6] 任婕. 基于向量自回归模型的旅游热门景点预测方法研究[J]. 现代电子技术, 2020, 43(3): 158-161.
- [7] 吴越, 杨烨军, 黄钰婷. 基于多元回归和灰色预测的浙江省文化产业发展影响因素的计量分析[J]. 东莞理工学院学报, 2020, 27(3): 5-10.
- [8] 孙建华, 张志立, 石茜, 赵阳, 魏春荣. 基于主成分-逐步回归分析法的瓦斯涌出量预测研究[J]. 煤炭工程, 2020, 52(1): 89-94.

数据样表

排名	小标签类	求和项:skc 金额	求和项:skc 销量	平均值项:标签价	库存	销售价	折扣
1	27050401	541409433.8	4558989	161.34	427314790	118.76	0.74
2	27060804	221500903.8	1866284	111.08	169840399	118.69	1.07
3	27112849	63394592.5	542730	85.53	58539158	116.81	1.37
4	27071209	55344005	466616	178.43	41758141	118.61	0.66
5	27164944	47459953.75	401234	129.29	43647794	118.28	0.91
6	27196225	40352573.75	197030	206.50	20943155	204.80	0.99
7	27092025	37214623.75	319424	181.49	33984852	116.51	0.64
8	27123264	30255705	251468	185.77	27630862	120.32	0.65
9	27217089	28742578.75	139553	186.91	13790791	205.96	1.10
10	27102436	20777561.25	178224	145.00	18825015	116.58	0.80