·临床研究·

基于灰色关联与偏最小二乘回归的住院费分析*

吕亚兰[1]，王润华[2]，叶孟良[2△]

（重庆医科大学：1. 信息管理系；2. 公共卫生与管理学院，重庆 400016）

摘 要：目的 结合灰色系统理论中的关联分析法与偏最小二乘回归模型，建立人均住院费用的预测模型。**方法** 采用灰色关联分析筛选出人均住院费用的主要影响因子，对因子间进行共线性诊断，建立人均住院费用与主要影响因子间的偏最小二乘回归预测模型，通过实例证明该模型的预测精度。**结果** 经灰色关联分析筛选出人均住院费用最主要影响因素为西药费、中药费，其次为诊疗费，其他费用、检查费、床费和手术费对人均住院费用影响也较大；偏最小二乘回归模型对住院费用拟合和预测准确率较高，平均相对误差较低，分别为－0.000 2%、0.349 3%。**结论** 灰色关联分析与偏最小二乘回归适宜于住院费用影响因素与预测分析，可为样本量小、变量间存在严重共线性资料分析提供参考。

关键词：灰色关联分析；偏最小二乘回归；住院费

doi：10.3969/j.issn.1671-8348.2013.23.009 **文献标识码：**A **文章编号：**1671-8348(2013)23-2722-03

Grey relational and partial least squares regression analysis on the hospitalization expenses*

Lv Yalan[1]，*Wang Runhua*[2]，*Ye Mengliang*[2△]

（*Chongqing Medical University*：1. *Department of Information Management*；

2. *Public Health and Management Faculty*，*Chongqing*，400016，*China*）

Abstract：Objective To combine grey relation analysis and partial least squares regression model to establish the forecasting model of per-patient hospitalization expenses. **Methods** Gray relational analysis was used to filter out the main factors affecting per-patient hospitalization expenses，and then collinearity was examined between these factors. Partial least squares regression was used to establish prediction model of per-patient hospitalization expenses，and the prediction accuracy was proved. **Results** After filtered by gray relational analysis，the order of the importance of factors affecting per-patient hospitalization expenses was the western medicine fee，traditional Chinese medicine fees，diagnosis and treat fees，other fees，inspection fees，bed fees and operation fees. The established partial least squares regression model had a higher accuracy on fitting and prediction，with low average relative error，respectively，－0.000 2% and 0.349 3%. **Conclusion** The gray relational analysis and partial least squares regression are suitable for the influencing factors and prediction analysis of hospitalization costs. It provides a reference for data with the small sample size and high collinearity between the variables.

Key words：grey relational analysis；partial least squares regression；hospitalization expenses

医疗费用长期以来都是社会焦点问题，其变化受主观和客观因素的影响，资料呈现灰色特性[1]，且各项费用间多不独立，因此，一般的统计方法多不适用于分析医疗费用。灰色关联分析（grey relation analysis）是灰色系统理论的一个重要内容，它通过参考序列与比较序列各点之间的距离分析来确定各序列之间的差异性和相近性，从而找出各因子之间的影响关系及影响系统行为的主要因子[2]。其不必考虑样本量的多少以及样本有无典型分布规律，因而在系统信息匮乏、统计数据灰度大的情况下，尤为实用。偏最小二乘回归（partial least squares，PLS）由 Wold 等[3]提出，主要用来解决多元回归分析中的变量多重相关性或变量多于样本点等实际问题。它集多元线性回归分析、主成分分析和典型相关分析的基本功能为一体[4]，该方法目前已广泛应用于化学、工业、经济等领域。因此，针对住院费用内部的复杂性和相关特性，本文结合灰色关联分析与偏最小二乘回归，对人均住院费用的主要影响因素和住院费用预测进行分析，为有效控制住院费用提供方法和参考。

1 资料与方法

1.1 一般资料 资料来源于重庆市各医院病案科 2010 年1～12 月登记的住院患者数据，选取其中总住院费用、床费、护理费、西药费、中药费、化验费、诊疗费、手术费、检查费和其他费用等 10 个指标，用 1～10 月数据建立模型，11、12 月数据验证模型预测精确度。

1.2 灰色关联分析 灰色关联分析的基本思想是根据序列曲线几何形状的相似程度来判断其联系是否紧密。曲线越接近，相应序列之间关联度就越大，反之就越小[2]。其分析步骤如下：

（1）确定参考数列（Y_0）和比较数列（X_i）。

（2）对参考数列和比较数列进行无量纲化处理，本文采用数据初值化，即将数列的每一项数据均除以该数列的第 1 个数据，得到参考数列 $Y_0(t)=\{y_0(1),y_0(2),\cdots\cdots,y_0(n)\}$ 与比较数列 $X_i(t)=\{x_i(1),x_i(2),\cdots\cdots,x_i(n)\}$（i＝1，2，……，9；$t$＝1，2，……，$n$；$n$＝10）。

（3）计算关联系数。关联系数是指比较数列 $X_i(t)$ 与参考数列 $Y_0(t)$ 在各个观察点的关联程度值，其在观察点 k 的值为：

* **基金项目：**重庆市卫生局 2012 年医学科研项目资助（2012-2-516）。 **作者简介：**吕亚兰（1988～），硕士，主要从事多元统计及其在医学中的应用研究（工作）。 △ **通讯作者**，Tel：15310939053；E-mail：yemengliang12@gmail.com。

$$\xi_i(k)=\frac{\min_i\min_k|y_0(k)-x_i(k)|+\rho\max_i\max_k|y_0(k)-x_i(k)|}{|y_0(k)-x_i(k)|+\rho\max_i\max_k|y_0(k)-x_i(k)|}$$

式中，$|y_0(k)-x_i(k)|$称为观察 k 点处 $y_0(k)$与 $x_i(k)$的绝对差，记为 $\Delta_i(k)$；$\min_i\min_k|y_0(k)-x_i(k)|=\Delta_{min}$，成为最小绝对差；$\max_i\max_k|y_0(k)-x_i(k)|=\Delta_{max}$，成为最大绝对差；ρ 为分辨系数，ρ∈[0,1]，一般取 ρ=0.5。

(4)计算关联度及排列关联次序。通过式 $r_i=\frac{1}{n}\sum_{k=1}^{n}\xi_i(k)$ 计算关联度，并按照 r_i 的大小顺序予以排列，组成关联序，关联序越靠前，说明该比较数列对参考数列的影响程度越大；反之，则越小。将关联度大于 0.6 的因子作为对人均住院费用影响较大的因子。

(5)将灰色关联分析筛选出来的影响较大因素分别与人均住院费用做相关分析，将相关系数假设检验无显著性的因子进一步排除，筛选最终进入回归模型的因素。

1.3 偏最小二乘回归模型(partial least-squares regression model)建立　(1)共线性诊断。采用条件指数(conditional index)和方差比例作为共线性诊断指标。条件指数在 10～30 为弱相关；在 30～100 为中等相关；大于 100 为强相关。在大的条件指数中由方差比例超过 0.5 的自变量构成的变量子集就认为是相关变量集[5]。(2)若经诊断自变量间具有共线性，则以经过灰色关联分析和相关分析筛选的影响因素为自变量，形成自变量集合(X)，以人均住院费用为因变量(Y)构建偏最小二乘回归模型。偏最小二乘回归模型的原理是分别在自变量 X 和因变量 Y 中提取成分 t_1 和 u_1，二者必须尽可能多地携带它们各自数据中的变异信息，且相关程度能够达到最大[6]。在第一成分 t_1 和 u_1 被提取后，偏最小二乘回归模型分别实现 X 对 t_1 的回归及 Y 对 u_1 的回归，如果回归方程已经达到满意的精度，则算法终止；否则将利用 X 被 t_1 解释后和 Y 被 u_1 解释后的残差进行第二轮的成分提取，依次类推，直到达到回归方程满意的精度。各成分相互独立，建立这些成分与自变量的回归方程；然后转变成自变量与因变量的回归方程[7-8]。(3)模型评价。通过对模型提取因子的累计解释能力、变量投影重要性指标(VIP)和误差分析对模型进行评价。采用 1～10 月的数据进行模型拟合程度的评价，11、12 月份的数据进行模型预测能力验证。

1.4 变量设置　人均住院费＝y、床费＝x_1、护理费＝x_2、西药费＝x_3、中药费＝x_4、化验费＝x_5、诊疗费＝x_6、手术费＝x_7、检查费＝x_8、其他费用＝x_9。

1.5 统计学处理　本文采用 Excel 2010 和 SAS9.13 进行数据整理和分析。

2 结　果

2.1 灰色关联分析　利用 1～12 月住院患者数据进行灰色关联分析。确定人均住院费为参考数列，床费、护理费、西药费、中药费、化验费、诊疗费、手术费、检查费和其他费用为比较数列。经灰色关联分析得到灰色关联系数值如表 1，最终计算得关联度由大到小的因素分别为：西药费(r_3＝0.89)、中药费(r_4＝0.89)、诊疗费(r_6＝0.79)、其他费用(r_9＝0.77)、检查费(r_8＝0.71)、床费(r_1＝0.70)、手术费(r_7＝0.63)、化验费(r_5＝0.54)、护理费(r_2＝0.53)。选取 $r>0.6$ 的因素作为主要影响因素，即西药费、中药费、诊疗费、其他费用、检查费、床费、手术费。

2.2 相关分析结果　灰色关联分析所获得主要影响因素分别与人均住院费用进行 Spearman 相关分析，进一步进行变量筛选，结果如表 2。结果表明中药费与人均住院费用不相关，相关系数比较差异无统计学意义($P>0.05$)。因此排除该因素，最终选择床费、西药费、诊疗费、手术费、检查费、其他费用作为因变量进行回归模型构建。

表 1　灰色关联系数

月份	ξ1(k)	ξ2(k)	ξ3(k)	ξ4(k)	ξ5(k)	ξ6(k)	ξ7(k)	ξ8(k)	ξ9(k)
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.64	0.50	0.84	0.84	0.60	0.78	0.38	0.69	0.78
3	0.69	0.40	0.90	0.90	0.56	0.72	0.46	0.67	0.68
4	0.70	0.39	0.89	0.89	0.67	0.88	0.59	0.52	0.74
5	0.65	0.37	0.88	0.88	0.39	0.71	0.55	0.60	0.66
6	0.73	0.43	0.85	0.85	0.63	0.71	0.53	0.71	0.74
7	0.67	0.45	0.88	0.88	0.50	0.84	0.63	0.80	0.81
8	0.70	0.47	0.93	0.93	0.52	0.73	0.85	0.65	0.82
9	0.65	0.49	0.91	0.91	0.52	0.94	0.67	0.65	0.75
10	0.59	0.44	0.86	0.86	0.41	0.66	0.56	0.76	0.67
11	0.57	0.57	0.86	0.86	0.34	0.72	0.70	0.69	0.71
12	0.86	0.80	0.93	0.93	0.36	0.78	0.59	0.80	0.84

表 2　因变量与各自变量 Spearman 相关系数

项目	床费	西药费	中药费	诊疗费	手术费	检查费	其他费用
r_s	0.890 9	0.981 8	0.400 0	0.836 4	0.863 6	0.945 5	0.790 9
P	0.000 2	<0.000 1	0.222 9	0.001 3	0.000 6	<0.000 1	0.003 7

2.3 共线性诊断　对最终选入模型的因素进行共线性诊断，结果如表 3。条件指数(φ)为 546.98，远大于 100，且西药费、检查费和其他费用的方差比例均大于 0.5，可以认为这些变量存在严重的多重共线性，故不能直接进行一般多元线性回归，可采用偏最小二乘回归。

表 3　共线性诊断

模型维数	条件指数(φ)	方差比例(υp)					
		x_1	x_2	x_6	x_7	x_8	x_9
1	1.000 00	0.000 00	0.000 00	0.000 01	0.000 02	0.000 01	0.000 00
2	61.987 20	0.009 54	0.001 58	0.010 38	0.221 98	0.000 24	0.003 81
3	124.131 51	0.038 89	0.000 35	0.004 81	0.055 49	0.393 35	0.010 19
4	156.155 18	0.023 65	0.000 64	0.710 57	0.425 75	0.018 89	0.003 08
5	304.750 33	0.926 00	0.053 19	0.056 53	0.052 70	0.062 43	0.250 66
6	546.984 28	0.001 91	0.944 24	0.217 70	0.244 06	0.525 09	0.732 26

2.4 偏最小二乘回归模型构建 以人均住院费用为因变量，以床位费、西药费、诊疗费、手术费、检查费、其他费用为自变量构建偏最小二乘回归方程。经交叉有效性检验，最终取成分 t_1 模型的预测能力较好，得到人均住院费用的偏最小二乘回归标准化变量模型：

$y=0.160x_1+0.195x_2+0.182x_6+0.182x_7+0.185x_8+0.176x_9$

转化为原始变量回归方程为：

$y=115.749+3.967x_1+0.525x_2+0.894x_6+0.897x_7+2.124x_8+1.659x_9$

2.5 模型精度分析

2.5.1 累计解释能力分析 在该模型构建中，从自变量中提取了 1 个成分 t_1，并计算出它对自变量和因变量的累计解释能力，该成分对自变量和因变量的累计解释能力均较高，分别为 83.470 5%、97.305 3%，可以认为这个成分的综合信息能代替 6 个自变量的信息。具体结果如表 4 所示。

表 4 成分 t_1 累计解释能力

成分	x_1	x_2	x_6	x_7	x_8	x_9	x	y
t_1	76.729 6	91.543 5	80.065 4	85.029 0	82.541 8	84.913 9	83.470 5	97.305 3

2.5.2 变量投影重要性指标分析 变量投影重要性指标（VIP_j）用于测度各个自变量对因变量的作用，如图 1 所示。由图可见，床费、西药费、诊疗费、手术费以及其他费用对人均住院费用影响均较大。

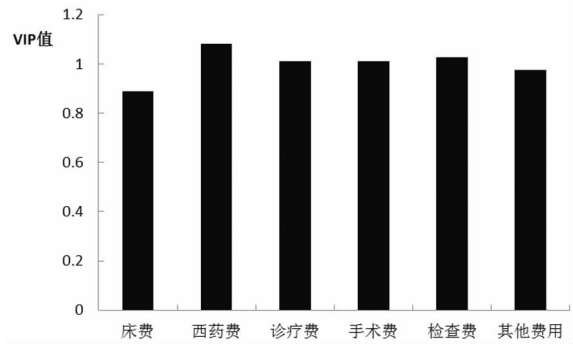

图 1 自变量的 VIP 直条图

2.5.3 误差分析 利用偏最小二乘回归的拟合值与实测值的误差，进行正态性检验和单样本 t 检（与均数 0 比较），差异均无统计学意义（正态性检验 $P>0.2$，t 检验 $P>0.05$），因此可认为模型绝对误差服从 $N(0,\sigma^2)$ 的正态分布，即可认为该误差为随机误差。且平均相对误差小，证明模型拟合较好。结果如表 5、图 2。

表 5 偏最小二乘回归模型拟合结果

月份	实测值	拟合值	绝对误差	相对误差(%)	平均相对误差(%)
1	5 312.454 0	5 354.542 6	−42.088 6	−0.786 0	−0.000 2
2	5 732.209 1	5 752.582 0	−20.372 9	−0.354 2	
3	5 819.604 8	5 827.403 4	−7.798 6	−0.133 8	
4	5 738.022 2	5 755.265 5	−17.243 3	−0.299 6	
5	5 738.202 4	5 728.905 2	9.297 2	0.162 3	
6	5 708.019 5	5 717.630 3	−9.610 8	−0.168 1	
7	5 663.175 8	5 652.136 4	11.039 4	0.195 3	
8	5 509.948 6	5 492.282 3	17.666 3	0.321 7	
9	5 620.087 7	5 601.526 0	18.561 7	0.331 4	
10	5 602.557 8	5 562.008 2	40.549 6	0.729 0	
11	5 703.079 0	5 653.809 0	49.270 0	0.871 4	0.349 3
12	5 892.395 9	5 902.595 2	−10.199 3	−0.172 8	

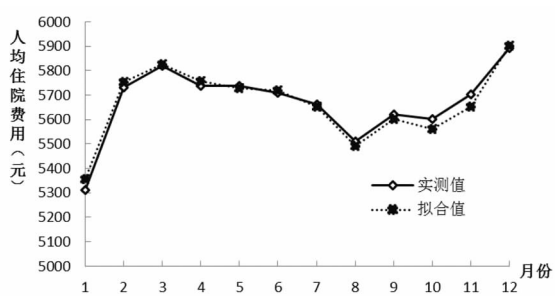

图 2 模型拟合值与实测值

3 讨 论

3.1 灰色关联分析适用于人均住院费用的影响因素筛选 灰色关联分析从系统行为特征数据与其相关因素数据入手，采用灰色关联度来描述主行为因子与相关行为因子之间关系亲密程度，其已广泛应用于医学卫生领域，已有多项研究[9-11]对住院费采用该方法进行分析。鉴于该方法对于数据分布及样本大小均无要求的优点，本文采用其筛选人均住院费用的主要影响因素，结果发现药品费是住院费用的第一影响因素，其次是诊疗费，这与郭朝伟[12]的研究结果相同。

3.2 偏最小二乘回归模型拟合较好 偏最小二乘回归被密西根大学的弗耐尔（Fornell）教授誉为第二代回归分析方法，是工业应用中用于软建模的流行方法，在医学领域也逐渐得到应用，对于存在共线性和样本量小的数据无非是构架预测模型的一种有效方法。对于住院费用各影响因素间相关性较强，存在严重的共线性，不适宜用一般多元线性回归分析。本文根据住院费用的特点，研究采用偏最小二乘回归对其建模的可行性和优越性。回归系数均大于 0，即各影响因素与人均住院费用呈正相关关系，符合一般规律。可见偏最小二乘回归应用于样本量小、自变量间共线性强的资料进行回归建模较好，能够较好地对人均住院费用进行合理预测。且经模型精度分析，定性辨析因变量与各自变量之间的关系，模型拟合和预测平均相对误差均较小，且误差属于随机误差，证明该模型拟合较好，预测较精确。

综上所述，灰色关联分析与偏最小二乘回归相结合，对于住院费用影响因素和预测分析取得较理想的结果，证明结合该方法能够很好地解决样本含量小、变量间关（下转第 2727 页）

平均时间比后者早约 40 d。对于那些发生转移的生存期很短的食管癌患者，这样明显的时间差异不仅有统计学上的意义，同时有积极的临床意义。但是，食管金属支架置入患者吞咽困难缓解效果持续时间很短，有很高的复发率，而且达到完全缓解（吞咽困难等级 0）的患者很少。

食管金属支架置入治疗能较早地缓解患者吞咽困难，但是放射治疗能达到较长的持续缓解时间和更有效的缓解效果。因此，可考虑将食管金属支架置入和放射治疗相联合，能更为有效地解决食管癌患者的吞咽困难症状。在治疗初期，食管金属支架置入可及时缓解患者的吞咽困难。支架置入后，通过放射治疗（后装或三维适形）保持缓解效果和避免复发。本研究中，只有 10 例患者接受了联合疗法，数量太少，因此，对于该疗法的安全性和有效性还无法评估。作者计划设计一个随机性的前瞻性试验，通过三维适形放疗联合金属支架置入，研究解决不能手术食管癌患者吞咽困难的最佳途径。

参考文献：

[1] Pisani P, Parkin DM, Bray F, et al. Estimates of the world wide mortality from 25 cancers in 1990[J]. Int J Cancer, 1999, 83(1): 18-29.

[2] Cwikiel M, Cwikiel W, Albertsson M. Palliation of dysphagia in patients with malignant esophageal strictures: comparison of results of radiotherapy, chemotherapy and esophageal stent treatment[J]. Acta Oncol, 1996, 35(1): 75-79.

[3] Frenken M. Best palliation in esophageal: surgery, stenting, radiation, or what? [J]. Dis Esophagus, 2001, 14(2): 120-123.

[4] Cunningham D, Allum WH, Stenning SP, et al. Perioperative chemotherapy versus surgery alone for resectable gastroesopha-geal cancer[J]. N Engl J Med, 2006, 355: 11-20.

[5] Weigel TL, Frumiento C, Gaumintz E. Endoluminal palliation for dysphagia secondary to esophageal carcinoma[J]. Surg Clin North Am, 2002, 82(4): 747-761.

[6] Siersema PD, Dees J, van Blankenstein M. Palliation of malignant dysphagia from oesophageal cancer. rotterdam oesophageal tumor study group[J]. Scand J Gastroenterol Suppl, 1998, 225: 75-84.

[7] Siersema PD, Hop WC, van Blankenstein M, et al. A comparison of 3 types of covered metal stents for the palliation of patients with dysphagia caused by esophagogastric carcinoma; a prospective, randomized study[J]. Gastrointest Endosc, 2001, 54: 145-153.

[8] Ferlay J, Shin HR, Bray F, et al. Estimates of world-wide burden of cancer in 2008: GLOBOCAN 2008[J]. Int J Cancer, 2010, 127(12): 2893-2917.

[9] Cunningham D, Allum WH, Stenning SP, et al. Perioperative chemotherapy versus surgery alone for resectable gastro-esophageal cancer[J]. N Engl J Med, 2006, 355(1): 11-20.

[10] Ychou M, Boige V, Pignon JP, et al. Perioperative chemotherapy compared with surgery alone for resectable gastro-esophageal adenocarcinoma: an FNCLCC and FFCD multicenter phase Ⅲ trial[J]. J Clin Oncol, 2011, 29(13): 1715-1721.

[11] Stahl M, Walz MK, Stuschke M, et al. Phase Ⅲ comparison of preoperative chemotherapy compared with chemoradio-therapy in patients with locally advanced adenocarcinoma of the esophagogastric junction[J]. J Clin Oncol, 2009, 27(6): 851-856.

（收稿日期：2013-01-08 修回日期：2013-04-22）

（上接第 2724 页）

系灰度大且具有较强共线性的资料分析问题。在医疗卫生领域常常存在类似资料，本文具有参考意义。

参考文献：

[1] 宋春华，马骏，崔壮，等. 参保精神分裂症患者住院费用构成分析[J]. 中国卫生统计，2011，28(5)：533-536.

[2] 邓聚龙. 灰色理论基础[N]. 上海：华中科技大学出版社，2002：12-19.

[3] Wold S, Albano C, Dunn M, et al. Pattern regression: Finding and using regularities in multivariate data[M]. London: Analysis Applied Science Publication, 1983.

[4] 丁磊. 偏最小二乘回归法改进及应用[D]. 乌鲁木齐：新疆大学应用数学系，2007.

[5] 马雄威. 线性回归方程中多重共线性诊断方法及其实证分析[J]. 华中农业大学学报：社会科学版，2008，74(2)：78-85.

[6] 秦浩，林志娟，陈景武. 偏最小二乘回归原理、分析步骤及程序[J]. 数理医药学杂志，2007，2(4)：450-451.

[7] 魏迎奇，张申. 基于偏最小二乘回归法的大坝渗透分析与预测[J]. 中国水利水电科学研究院学报，2011，9(3)：205-208.

[8] 汪春辉，罗飞，舒红平. 偏最小二乘回归在气温预测中的研究与应用[J]. 网络与通信，2012，28(5)：142-144.

[9] 徐萍. 腹腔胆囊切除术住院费用的灰色关联分析[J]. 中国卫生统计，2011，28(5)：596-598.

[10] 黄宝真，汤先钊，高侨，等. 基于灰色关联法的住院费用影响因素分析[J]. 解放军医院管理杂志，2011，18(10)：922-923.

[11] 刘娟，李系人. 7 种外科疾病住院费用的灰色关联分析[J]. 中国病案，2012，13(3)：49-50.

[12] 郭朝伟. 运用新灰色关联法对住院费用影响的因素分析[J]. 医学教育探索，2009，8(7)：886-888.

（收稿日期：2013-02-08 修回日期：2013-06-15）