基于异质性数据的 Logit 变量选择模型研究*

斯介生 李 扬 谢邦昌

内容提要:在大数据时代,数据的异质性和变量的稀疏性是不可回避的两大问题。本文针对上述问题构建了异质性 Logit 变量选择模型。研究显示,在不同的异质性条件下,本文的方法可以明显区分有效变量和冗余变量。而且,通过 Gmeans 等评价指标可知该模型具有很好的预测效果。在对上市公司财务预警分析的应用研究中,本文方法得到了具有解释意义的结果,说明该方法具有一定的实证价值。

关键词:异质性;变量选择;财务预警

DOI: 10.19343/j.cnki.11-1302/c.2017.12.010

中图分类号: O212 文献标识码: A 文章编号: 1002-4565(2017)12-0110-09

The Study of Variable Selection in Logit Model Based on Heterogeneous Data

Si Jiesheng Li Yang Xie Bangchang

Abstract: The heterogeneity in data and the sparsity of variables are two important problems and can-not be ignored in big data era. In this paper, a new Logit model is proposed when the data is of heterogeneity, sparsity and the dependent variable is binary. The results show that the method can effectively distinguish the redundant variables in different groups. On the other hand, It shows that the model can predict well by Gmeans and other evaluation indicators. Finally, the method is applied to the research on financial early warning of listing corporation and some meaningful results are obtained, which shows the method in this paper has some practical value.

Key words: Heterogeneity; Variable Selection; Financial Early Warning

一、引言

随着研究的深入,越来越多的研究者认识到“大数据”不仅仅是“数据大”,而是指“大而复杂”的数据。随着数据量的增长,纳入研究的数据可能来自具有不同背景的子总体,构成了“分析单位之间在特征、属性和状态上的差别或不同”[1]的异质性。譬如在财务预警研究中,虽然存在影响企业未来经营状况的共同财务指标因素,但其影响程度可能在不同行业间存在差异。特别地,对于不同类型的企业,也会存在特异性的影响因素。如果在财务预警模型建立时忽略上述异质性影响,可能会导致模型估计不一致和推断错误[2]。另一方面,随着数据采集技术的发展,越来越多的

* 本文是中国人民大学科学研究基金(中央高校基本科研业务费专项资金资助)项目“生物医学大数据的统计方法基础研究”(15XNI011)的阶段性成果。

影响因素变量被纳入预测模型中,这就形成了高维数据。高维变量的稀疏性成为研究者关心的问题。如果不能有效地剔除模型中的冗余变量,会出现“模型解释性差、产生多重共线性、模型统计推断失效等”问题[3]。因此,如何在大数据时代发展经典统计,特别是对多源异质性大数据的整合分析是当前大数据研究的重要目标[4]。

关于异质性数据的研究,一种思路是假定变量的回归系数为常系数,数据的异质性通过引入可加形式的固定和随机效应进行刻画。其本质上是通过刻画误差的异方差性实现对异质性的建模,例如协方差分析模型和混合效应模型[5]。这类模型因为有很好的数学性质而被广泛研究。但一些研究表明,有些情况下回归系数是常系数的假定会显得过于严格。Su 和 Chen[6]、Browning 和 Carro[7]等研究都表明常系数的假定与现实不符。因此,有研究者提出假定回归系数为变系数的建模方法。Hsiao[2]对变系数模型做了详细总结。但也有研究者指出,这类模型因待估参数过多而损失过多自由度,在实际应用中存在局限性[8]。鉴于上述问题,有研究者提出了组异质性(Group Heterogeneity)回归模型(Lin 和 Ng[9]; Su 等[8])的方法。对于截面数据,模型定义为:

$$y_{ig} = x'_i\beta_g + u_{ig}, i \in \text{第 } g \text{ 个总体} \quad (1)$$

其中,$g = 1, 2, \cdots, G$ 是子总体个数。式(1)通过对不同子总体设定不同的回归系数刻画数据的异质性。这样的建模方式考虑了异质性,同时不必像变系数模型那样损失过多自由度,与前两者相比,方法更加灵活。

关于变量选择的研究,随着 Lasso 的提出,使惩罚函数方法在变量选择的研究中具有核心的地位。近年来,研究者针对同质性的截面数据提出了多种推广的惩罚函数方法,例如 Group Lasso、Adaptive Lasso、SCAD、MCP 等。关于异质性模型的变量选择研究,集中在基于回归系数为常系数的模型基础上讨论[3]。但是对于式(1)的变量选择研究,目前鲜有文献讨论。另外,式(1)的响应变量是连续变量。而在实际中,分类响应变量的异质性数据是大量存在的。对于这类数据,虽然可以按照不同的板块划分数据分别进行变量选择建模,但分开建模的办法会割裂数据的整体性,仅考虑了数据的异质性而忽略变量影响的共性。因此,需要在响应变量为二分类情形推广式(1),将数据纳入一个模型进行综合分析。

在式(1)的框架下,目前尚无文献综合考虑分类响应变量情形的异质性数据建模及变量选择问题。根据实际应用的需要,在响应变量是分类变量的情形下推广式(1)需要考虑到三个方面的问题: 一是模型形式需要改变,需要建立相应的参数估计、变量选择、模型评价方法。二是对异质性问题进行变量选择时,需要考虑到异质性数据存在的组异质性,以及同一变量在不同子总体之间存在的共性。这就使变量选择过程变得复杂,在理论上属于异质性数据的双水平变量选择。马双鸽等[10]以及其中提到的相关文献研究了不同情形的双水平变量选择问题,但都没有针对异质性分类响应变量数据进行研究。三是在具体的建模过程中,需要考虑到分类数据的不平衡性[11]。对于异质性数据而言,每个子总体的不平衡率不同,这就使得问题变得更加复杂。在实际应用中,这三个方面的问题不可回避。如何建模,并进行参数估计和变量选择是研究需要面对的挑战。

本文考虑当响应变量为二分类变量、解释变量为连续变量时的异质性截面数据建模及其变量选择问题。具体研究内容包括: 针对响应变量为二分类变量的情形,构建了异质性数据的 Logit 模型; 对于新构建的模型,建立了基于极大似然的参数估计方法; 综合考虑异质性数据存在的组异质性,同一变量在不同子总体之间存在的共性,不同子总体间存在不同的不平衡性,引入稀疏组惩罚函数(Sparse Group Penalized Function),实现了对新构建模型的变量选择; 在实际建模中,采用交叉验证等方法选择衡量数据不平衡的阈值和惩罚函数的调节参数; 针对新构建的模型,通过引入 Gmeans 等评价指标,建立了相应的模型评价方法。

二、基于异质性数据的 Logit 变量选择模型

当响应变量为二分类变量且数据存在异质性时，下面建立基于异质性数据的 Logit 模型（以下简称为异质性 Logit 模型），并给出了惩罚似然方法，实现参数估计和变量选择。

（一）异质性 Logit 模型

假定响应变量 Y 是取值 0 和 1 的分类变量，解释变量 $X_1,X_2,\cdots,X_p$ 是与响应变量可能相关的确定性变量。n 组观测数据来自 G 个不同的子总体，其中每个子总体分别有 $n_1,n_2,\cdots,n_G$ 次观测，即满足 $\sum_{g=1}^{G} n_g = n$。对于来自第 g 个子总体的第 i 个样本的实现，记 y_{gi} 为响应变量的实现值，$x_{1gi},\cdots,x_{pgi}$ 为解释变量的实现值。

由于响应变量为二分类变量，记 $\pi_{gi} = P(y_{gi} = 1 \mid X_1 = x_{1gi},\cdots,X_p = x_{pgi})$，表示解释变量为 $X_1 = x_{1gi},\cdots,X_p = x_{gpi}$ 条件下，响应变量取 1 的条件概率。本文定义异质性 logit 模型为：

$$\text{logit}(\pi_{gi}) = \log\left(\frac{\pi_{gi}}{1-\pi_{gi}}\right) = x'_{gi}\beta_g, i \in \text{第 } g \text{ 个总体} \tag{2}$$

其中，$\beta_g = (\beta_{g1},\cdots,\beta_{gp})'$ 为第 g 个子总体对应的回归系数向量。$x_{gi} = (x_{1gi},\cdots,x_{pgi})'$ 是来自第 g 个子总体的第 i 个样本解释变量向量。

可见，在式（2）中，在每个子总体内回归系数为常系数。而在不同的子总体之间，回归系数不相同，这就刻画了总体的异质性。

（二）异质性 Logit 模型的变量选择

1. 模型的对数似然函数。

为了实现对式（2）的参数估计和变量选择，本文通过对其似然函数加入适当的惩罚函数，即建立惩罚似然，并最优化惩罚似然达到目标。为此，首先建立式（2）的似然函数。假定所有观测间相互独立，则得到式（2）的对数似然函数为：

$$l(\beta) = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left[y_{gi}\log\left(\frac{\pi_{gi}}{1-\pi_{gi}}\right) + \log(1-\pi_{gi})\right] \tag{3}$$

其中，$\beta = (\beta'_1,\cdots\beta'_G)'$ 是式（2）的未知参数。

2. 模型的惩罚似然函数。

为了对式（3）引入罚函数，需要根据式（2）的特点，构造适当的罚函数。首先将式（3）中的未知参数 β 改写为：

$\beta = (\beta^{(1)},\beta^{(2)},\cdots,\beta^{(p)})$，其中 $\beta^{(j)} = (\beta_{1j},\beta_{2j},\cdots,\beta_{Gj})'$，$j = 1,2,\cdots,p$

显然，$\beta^{(j)}$ 表示第 j 个变量对应的 G 个子总体的回归系数。对于同一个变量而言，这 G 个回归系数刻画了不同子总体间的异质性。但是对于不同的变量而言，$\beta^{(j)}$ 体现的效应都是由于第 j 个变量产生，这个变量是诸子总体的共同变量，所以存在共性。一方面考虑到共同变量刻画的共性，我们对 $\beta^{(j)}$ 整体施加二范数罚。另一方面考虑到不同子总体间的异质性，我们对每个回归系数施加一范数罚。于是得到式（2）的惩罚似然函数：

$$S_{\lambda,\alpha}(\beta) = -l(\beta) + \lambda\alpha\sum_{j=1}^{p}\|\beta^{(j)}\|_2 + \lambda(1-\alpha)\sum_{j=1}^{p}\sum_{g=1}^{G}|\beta_{gj}| \tag{4}$$

其中，$\lambda > 0, 0 < \alpha < 1$ 为调节参数。最小化式（4）就得到式（2）的参数估计。从式（4）中可以看出，本文对所有待估参数进行分组，不仅对每个组的变量整体加罚，而且对组内每个变量加罚。因而本质上可以归结为稀疏组 Lasso[12] 方法。Vincent 和 Hansen[13] 关于稀疏组 Lasso 提出了快速而有效的优化算法，本文基于该算法的思想实现对式（4）的最小化。

3. 调节参数的选择。

式(2)中存在 3 个调节参数。式(4)中含两个调节参数 α 和 λ,以及处理不平衡问题的阈值。其中,λ 的值决定了模型整体的稀疏程度,当其足够大时,所有参数估计值都将为 0; 当其为 0 时,相当于对模型不加罚。α 的值决定模型对成组参数和单个参数的惩罚力度。其值越大,对 $\beta^{(j)}$ 整体惩罚力度越大。其值越小,对单个参数的惩罚力度越大。在本文中选择 $\alpha = 0.5$,可以同时兼顾成组变量选择和单一变量选择。对于 λ,需要根据模型的预测效果进行选取。理论上可以在 $0 \sim \lambda_{\max}$ 之间选取多个值进行尝试,根据模型预测效果确定最优的 λ。其中,$\lambda_{\max} = \inf\{\lambda > 0 \mid \hat{\beta}(\lambda) = 0\}$,本文根据 Vincent 和 Hansen[13] 的方法数值计算得到。对于阈值,由于实际问题中真实的不平衡率是未知的。因此从 0 ~ 0.5 之间取多个值进行尝试,根据预测结果选择最优的阈值。

(三)模型评价

模型的评价需要从两个方面进行考虑。第一是预测效果,第二是变量选择效果。对于二分类问题,一般情况下可以用真阳性率(TPR)和真阴性率(TNR)进行判断,其定义为:

$$\mathrm{TPR} = \frac{\sum_{i=1}^{n} I(Y_i = 1 \text{ 且 } \hat{Y}_i = 1)}{\sum_{i=1}^{n} I(Y_i = 1)},\ \mathrm{TNR} = \frac{\sum_{i=1}^{n} I(Y_i = 0 \text{ 且 } \hat{Y}_i = 0)}{\sum_{i=1}^{n} I(Y_i = 0)} \tag{5}$$

其中,$Y_i(i = 1,2,\cdots,N)$ 是每个样本的结局,$\hat{Y}_i(i = 1,2,\cdots,N)$ 是每个样本的预测结果。$i(\cdot)$ 为示性函数。由于二分类数据往往存在不平衡性[11],需要综合考虑真阳性率和假阳性率。本文采用两者的几何平均数作为评价指标,定义为 $\mathrm{Gmeans} = \sqrt{\mathrm{TPR} \times \mathrm{TNR}}$,当 Gmeans 达到最大时的参数估计即为我们需要的估计值和变量选择结果。

三、模拟研究

(一)研究目标和仿真设定

1. 研究目标。

对于模型(2),研究者关心其变量选择效果和模型的预测效果。对变量选择的效果,应从两个角度进行考察: 一是对于同一变量在不同子总体上的表现; 二是对模型的预测效果,通过对响应变量的预测值与真实值进行比较来考察。具体采用式(5)以及 Gmeans。

2. 仿真设定。

第一,为了针对不同复杂程度的模型考察方法效果,本文设定不同数量的子总体(记子总体的个数为 G)比较模型的效果。本文对 $G=3$、$G=5$、$G=7$ 三种子总体数不同的情况进行讨论。第二,由于二分类变量可能存在不平衡性,本文设定模型的不平衡率为 $1:9(\#\{y=1\}:\#\{y=0\} = 1:9)$。第三,变量个数为 $p=21$,其中第一个变量为截距项。样本量分别为 450、750 和 1050。

记系数矩阵为 $\beta^{[4]} = (\beta_{gi})_{G\times p}$,则数据由模型 $\mathrm{logit}(\pi_g) = \beta_{g0} + \beta_{g1}x_{g1} + \cdots + \beta_{g20}x_{g20}$ 产生。其中 $g = 1,\cdots,G$。$x_{gj} \sim N(0,1)$ 独立同分布。当 $G = 3$、5、7 时,有:

$$\beta^{[3]} = \begin{pmatrix} 0 & 0.8 & 0 & 0 & 0.6 & 0 & 0 & 0.8 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0.8 & 0 & 0 & 0.8 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 \end{pmatrix}$$

$$\beta^{\text{固}}=\begin{pmatrix}1&0&0&0.6&0&1&0&0.8&0&0&0&1&0&0&0&0.8&0&0.8&0&0&0\\1&0&0&0.8&0&0.8&0&0&0&0&0&1&0&0.8&0&0.8&0&0.8&0&0&0\\1&0&0&1&0&0&0&1&0&0&0&0&0&0.8&0&1&0&0.6&0&0&0\\1&0&0&0.8&0&0.8&0&0&0&1&0&0&0&1&0&1&0&0.6&0&0&0\\1&0&0&0.6&0&1&0.8&0&1&0&0&0&0&0&0&0&0&1&0&0&0\end{pmatrix}$$

$$\beta^{\text{团}}=\begin{pmatrix}1&0&0&0.6&0&1&0&0&0&1&0&0&0&0.8&0&0&0&1&0&0&0\\1&0&0&0.8&0&0.8&0&1&0&0&0&0&0&0&0&1&0&0.6&0&0&0\\1&0&0&1&0&0&0&0.8&0&0&0&1&0&0.8&0&0&0&0.6&0&0&0\\1&0&0&1&0&0.6&0&0&0&0&0&1&0&1&0&0.8&0&0.8&0&0&0\\1&0&0&1&0&0&0&0.8&0&0&0&0&0&1&0&0.6&0&1&0&0&0\\1&0&0&0.8&0&0.8&0&1&0&0&0&0&0&0&0&1&0&0.8&0&0&0\\1&0&0&0.6&0&1&0&0&0&1&0&0&0&0&0&0.8&0&1&0&0&0\end{pmatrix}$$

其中，系数矩阵的第 1 列是截距项。其余各列为各变量在子总体上的回归系数。冗余变量对应的列元素全部为 0，其余不全为 0 的列对应的变量为有效变量。容易看出当 $G=3$ 时，第 3、4、6、7、9、10、12、13、15、16、18、19 列对应的变量全为冗余变量，不同的子总体冗余变量不全相同。当 $G=5$ 以及 $G=7$ 时意义相同。

（二）模拟结果及解释

为了对异质性 logit 模型的变量选择效果进行比较，对人工数据采用三种建模方法：利用异质性 logit 模型对全部样本进行建模（记为 M1）；对不同的子总体样本分别建立 logit-lasso 模型（记为 M2），即对不同子总体分开建模；全部样本建立 logit-lasso 模型。其中，冗余变量用▽表示，有效变量用▼表示。按照上述设定，模拟 100 次，统计各变量选出的频率，$G=3$、4、5 的结果见表 1～3。

表 1　　**G=3 时各变量选出的频率**　　（%）

g=1	▼	▽	▽	▼	▽	▽	▼	▽	▽	▼	▽	▽	▽	▽	▽	▼	▽	▽	▽	▽
M1	100	21	21	97	22	28	100	19	19	98	20	21	43	19	16	95	26	18	37	36
M2	100	32	53	100	40	41	100	43	45	100	47	48	48	49	49	100	50	46	45	48
M3	100	64	71	100	67	64	100	65	54	62	64	64	100	52	53	62	66	64	59	100
g=2	▼	▽	▽	▼	▽	▽	▽	▽	▽	▼	▽	▽	▼	▽	▽	▽	▽	▽	▼	▽
M1	100	18	26	100	27	34	50	22	20	95	29	26	100	25	32	44	35	22	100	42
M2	100	49	47	100	50	38	52	51	49	100	42	39	100	48	50	46	41	48	100	45
M3	100	64	71	100	67	64	100	65	54	62	64	64	100	52	53	62	66	64	59	100
g=3	▼	▽	▽	▼	▽	▽	▼	▽	▽	▽	▽	▽	▼	▽	▽	▽	▽	▽	▽	▼
M1	100	22	27	100	26	25	100	16	27	49	28	19	100	23	17	42	23	19	40	89
M2	100	47	41	100	38	45	100	43	45	44	43	46	100	49	41	44	38	38	46	100
M3	100	64	71	100	67	64	100	65	54	62	64	64	100	52	53	62	66	64	59	100

1. 三类建模方式变量选择情况。

根据不同子总体变量是否有效，我们将变量分为全部有效、全部冗余、部分有效、部分冗余四种情况，统计变量选出的平均频率。由表 1、表 2、表 3 可知：当 $G=3$ 时，本文方法（M1）对四类不同变量平均选出频率为：99.5%、23.3%、97.4%、42.6%；分开建模（M2）为 100%、44.7%、100%、46.4%；统一建模（M3）为 100%、62.3%、82.8%、78.2%。当 $G=5$ 时，M1 为 96.9%、21.9%、94.8%、41.6%；M2 为 100%、46.6%、100%、45.9%；M3 为 100%、69.9%、85.7%、86.0%。当 $G=7$ 时，M1 为 96.1%、11.4%、98.7%、29.3%；M2 为 100%、46.6%、100%、43.8%；M3 为 100%、74.8%、88.5%、86.1%。由此可以看出，三种情况中，对冗余变量的误选次数都是 M1 最小，对有

表 2　　G =5 时各变量选出的频率　　(%)

g = 1	▽	▽	▼	▽	▼	▽	▼	▽	▽	▽	▼	▽	▽	▽	▼	▽	▼	▽	▽	▽
M1	16	22	92	17	100	16	99	21	49	21	100	21	47	21	99	24	100	23	27	27
M2	54	41	100	41	100	38	100	51	36	48	100	48	52	50	100	56	100	41	47	46
M3	65	72	100	70	100	67	100	66	100	67	75	72	67	70	73	70	100	74	72	74
g = 2	▽	▽	▼	▽	▼	▽	▽	▽	▽	▽	▼	▽	▼	▽	▼	▽	▼	▽	▽	▽
M1	20	27	99	22	98	18	39	27	44	22	100	20	99	16	98	29	100	18	26	27
M2	45	50	100	46	100	42	47	47	46	46	100	51	100	47	100	52	100	49	44	49
M3	65	72	100	70	100	67	100	66	100	67	75	72	67	70	73	70	100	74	72	74
g = 3	▽	▽	▼	▽	▽	▽	▼	▽	▽	▽	▽	▽	▼	▽	▼	▽	▼	▽	▽	▽
M1	16	25	99	25	36	15	99	21	39	19	36	20	99	19	100	20	94	13	15	23
M2	48	53	100	42	46	53	100	49	47	44	54	47	100	44	100	46	100	47	44	45
M3	65	72	100	70	100	67	100	66	100	67	75	72	67	70	73	70	100	74	72	74
g = 4	▽	▽	▼	▽	▼	▽	▽	▽	▼	▽	▽	▽	▼	▽	▼	▽	▼	▽	▽	▽
M1	18	27	99	21	18	20	48	30	100	26	50	25	100	15	100	25	93	26	25	29
M2	44	47	100	44	100	51	41	48	100	44	49	46	100	44	100	57	100	53	54	50
M3	65	72	100	70	100	67	100	66	100	67	75	72	67	70	73	70	100	74	72	74
g = 5	▽	▽	▼	▽	▼	▽	▼	▽	▼	▽	▽	▽	▽	▽	▽	▽	▼	▽	▽	▽
M1	21	25	93	19	100	17	98	28	100	18	44	19	33	17	34	31	100	23	28	22
M2	40	44	100	43	100	39	100	51	100	43	41	45	45	40	47	57	100	41	41	40
M3	65	72	100	70	100	67	100	66	100	67	75	72	67	70	73	70	100	74	72	74

表 3　　G =7 时各变量选出的频率　　(%)

g = 1	▽	▽	▼	▽	▼	▽	▽	▽	▼	▽	▽	▽	▼	▽	▽	▽	▼	▽	▽	▽
M1	4	11	93	4	100	12	30	10	100	15	33	13	100	15	28	9	100	15	11	8
M2	42	42	100	46	100	41	43	45	100	49	47	40	100	51	45	45	100	38	56	53
M3	75	75	100	71	100	80	75	76	100	79	75	69	74	75	100	74	100	73	78	73
g = 2	▽	▽	▼	▽	▼	▽	▼	▽	▽	▽	▽	▽	▽	▽	▼	▽	▼	▽	▽	▽
M1	12	9	95	6	97	11	100	11	25	8	26	6	31	9	100	10	84	12	11	10
M2	42	48	100	44	100	47	100	55	52	38	46	49	32	47	100	38	100	53	49	53
M3	75	75	100	71	100	80	75	76	100	79	75	69	74	75	100	74	100	73	78	73
g = 3	▽	▽	▼	▽	▽	▽	▼	▽	▽	▽	▼	▽	▼	▽	▽	▽	▼	▽	▽	▽
M1	9	11	99	5	29	9	98	9	28	13	100	6	98	11	27	11	88	11	10	10
M2	51	43	100	46	44	44	100	54	32	42	100	36	100	50	35	41	100	41	45	44
M3	75	75	100	71	100	80	75	76	100	79	75	69	74	75	100	74	100	73	78	73
g = 4	▽	▽	▼	▽	▼	▽	▽	▽	▽	▽	▼	▽	▼	▽	▼	▽	▼	▽	▽	▽
M1	6	13	100	8	95	12	37	21	30	14	100	12	99	16	98	12	100	10	16	12
M2	49	46	100	51	100	40	48	46	42	48	100	49	100	51	100	48	100	63	51	43
M3	75	75	100	71	100	80	75	76	100	79	75	69	74	75	100	74	100	73	78	73
g = 5	▽	▽	▼	▽	▽	▽	▼	▽	▽	▽	▽	▽	▼	▽	▼	▽	▼	▽	▽	▽
M1	3	13	100	8	28	12	99	13	29	15	32	10	100	18	91	12	100	15	12	10
M2	46	42	100	45	52	51	100	44	41	39	47	44	100	47	100	57	100	44	46	45
M3	75	75	100	71	100	80	75	76	100	79	75	69	74	75	100	74	100	73	78	73
g = 6	▽	▽	▼	▽	▼	▽	▼	▽	▽	▽	▽	▽	▽	▽	▼	▽	▼	▽	▽	▽
M1	10	12	99	5	98	14	99	18	30	16	26	12	25	12	99	13	99	10	15	15
M2	43	45	100	48	100	41	100	44	44	44	47	54	41	51	100	47	100	47	54	53
M3	75	75	100	71	100	80	75	76	100	79	75	69	74	75	100	74	100	73	78	73
g = 7	▽	▽	▼	▽	▼	▽	▽	▽	▼	▽	▽	▽	▽	▽	▼	▽	▼	▽	▽	▽
M1	7	16	88	9	100	13	37	17	100	17	29	10	26	13	100	12	100	16	13	9
M2	52	42	100	50	100	42	46	47	100	47	45	53	46	41	100	49	100	42	45	49
M3	75	75	100	71	100	80	75	76	100	79	75	69	74	75	100	74	100	73	78	73

效变量的正确选择次数，M1 与 M2、M3 相差很小。这说明，本文方法(M1) 可以更好地区分冗余变量和有效变量。

2. 变量误选原因分析。

由表 1 ~ 3 可知，如果某变量在各子总体的冗余情况不一致，误选变量的频数会增加。M3 效果最差，M2 次之，M1 效果最好。并且随着样本量的增加，M2 和 M3 对冗余变量的误选频

数并没有变小，而 M1 的误选频数变小。我们认为原因如下：对于 M3 而言，直接对所有样本不加区分地建模，并没有因为样本量的增加而使变量误选频率降低。这是因为变量在不同子总体之间冗余情况不一，样本信息相互干扰，直接利用所有样本，无法排除这种干扰。对于 M2 而言，分开建模可以排除不同子总体之间的干扰信息，但是分开建模本质上只利用了一个子总体的信息，导致了大量信息损失。对于 M1 而言，本文方法综合利用了不同子总体的全部信息，所以效果最好。

3. 预测效果。

$G=3$、$G=5$ 和 $G=7$ 时本文方法（M1）的 Gmeans 值分别为 88.4%、89.5% 和 88.3%，这表明：首先，在三次模拟中，预测效果都较好；其次，在三次模拟中 Gmeans 值变化不大，可认为预测效果在三次模拟之间是比较稳定的。所以可知，当同一变量对应的不同参数值个数，即每组内待估参数个数在一定范围内变化时，模型的预测效果较好且保持相对稳定。

四、应用研究

为验证本文方法的实证意义，将其应用于 A 股上市公司财务预警研究。财务预警模型研究的重点在于预警指标的选择，并针对模型的学习和泛化能力展开讨论[14]。此外，财务预警研究中有三个问题值得注意：第一，如果对纳入模型的企业仅区分 ST 和非 ST 企业两类而不考察企业的背景，会忽略数据的异质性结构。第二，由于 ST 企业只占少部分，数据中存在不平衡性。但不考虑数据的实际而强行 1∶1配比[15]，会破坏原始数据的数据结构[11]。第三，预警指标变量的主观选取会因研究者的不同而得到不同的结果。而且，主观选取可能会遗漏关键变量。鉴于此，本文在强调模型的预测效果基础上，同时考虑企业数据的异质性，ST 企业和非 ST 企业的不平衡性，研究了基于数据实际情况下的变量选择问题。

本文所用数据包含 557 家上市公司 2010 年末的 84 个财务指标，及 2013 年初是否被评为 ST。557 家上市公司中，有 161 家来自金属非金属板块，106 家来自生物医学板块，290 家来自机械设备仪表板块。同时，不同板块被评为 ST 的企业比例也不同。金属非金属板块中，12 家为 ST，占 7.5%；生物医药板块中，7 家为 ST，占 6.6%；机械设备仪表板块中，14 家为 ST，占 4.8%。数据存在很强的不平衡性，并且不同板块的不平衡率是不同的。

84 个财务指标来自于：盈利能力、股东获利能力、营运能力、发展能力、现金流量能力、短期偿债能力、长期偿债能力、股权结构、成本水平等九方面。对这些变量进行删选，需要考虑到两个问题：一是同一指标对于不同板块企业的影响程度可能不同。二是对于所有企业而言，相同的财务指标存在共同的影响。对所有企业不区分板块，将忽略指标影响在不同板块上的差异。对不同板块分开建模，存在忽略指标影响在不同板块上共性的问题。

本文采用式（2）对上市公司财务数据进行分析，根据数据的实际情况选择财务预警指标。其中样本容量为 557，解释变量为 84 个，全部为连续变量。响应变量为取值 0 和 1（ST 类取值为 1）的二分类变量。

为使模型具有更好的泛化能力，同时避免过拟合。本文以式（5）的几何平均（即 Gmeans）作为评价指标，结合 5 折交叉验证确定模型中的调节参数，同时选择财务预警指标。由于三个板块的不平衡率不同，为处理不平衡问题，确定金属非金属、生物医药和机械设备仪表三个板块的阈值分别为 $\theta_1=0.069$、$\theta_2=0.07$、$\theta_3=0.058$。指标（变量）选择和参数估计的结果见表 4①。

① 为了节省篇幅，略去了没有被选出的指标。

表 4　　财务预警指标选择结果和参数估计结果

指标名称	金属非金属板块	生物医药板块	机械设备仪表板块
每股未分配利润(元/股)	-0. 177	-0. 789	-0. 916
净资产收益率(平均) (%)	-0. 075	0	0. 041
资产净利率(%)	-0. 404	-0. 179	-0. 422
投入资本回报率(%)	-0. 016	-0. 073	-0. 003
营业利润/营业总收入	0	-0. 036	0
营业总成本/营业总收入	0	-0. 001	0
营业利润率(%)	0	-0. 036	0
营业利润增长率(%)	-0. 021	-0. 045	0
每股经营现金流增长率(%)	0	0	-0. 068
净资产收益率增长率(%)	0	0	-0. 001
应收账款周转率(次)	-0. 001	0	0
股东权益周转率(次)	0	0	-0. 001
销售现金比率	0	-0. 026	0
资产负债率(%)	0	0. 001	0. 003
股东权益比率(%)	0	-0. 001	-0. 003
市净率	-0. 001	0	-0. 001
市现率	0	0	-0. 001
流动资产净利润率	-0. 114	-0. 050	0
营运资金比率	0	-0. 053	0
营运资金对资产总额比率	0	-0. 067	-0. 188
营运资金对净资产总额比率	-0. 032	0	0

表 4 的结果表明,非 0 参数组为 21 组。对于不同的板块而言,同一指标对于响应变量的影响不同。即同一指标对于上市公司是否被纳入 ST 类的影响程度不同。每股未分配利润、资产净利率、投入资本回报率属于一般性的,对三个板块都有影响的财务指标。净资产收益率、营业利润增长率、资产负债率等对三个板块中的两个板块有影响; 应收账款周转率、销售现金比率、营运资金比率等属于特异性的,仅对其中某一板块有影响。已有的文献选取的财务预警指标与这里的指标有很多相同之处,例如孔宁宁和魏韶巍[15]。本文基于数据实际情况进行指标选择的研究表明,不同板块的财务预警指标选取需要具体情况具体分析。目前鲜有文献这样讨论,值得进一步研究。

模型的预测能力对于财务预警有着重要意义。通过式(5) 以及 Gmeans 对预测效果进行模型评价。其中,式(5) 中的 TPR、TNR 以及 Gmeans 越大,说明模型的预测效果越好。预测结果见表 5,预测正确率绝大多数在 80% 以上,表明本文方法对财务预警有很好的应用价值。

表 5　　财务预警模型的预测结果　　(%)

板块	TPR	TNR	Gmeans
金属非金属	70. 5	91. 7	80. 4
生物医药	95. 0	100	97. 4
机械设备仪表	84. 1	78. 6	81. 3
三板块合计	82. 3	87. 9	85. 0

五、结论与讨论

本文以实际问题中数据存在异质性的问题为起点,针对二分类响应变量提出了异质性 Logit 变量选择模型。通过模拟仿真研究了模型的性能,并将此方法应用于财务预警研究中。在数值模拟中,所有情况下模型预测结果的 Gmeans 值都在 88. 0% 以上。在财务预警的应用研究中,预测的 Gmeans 值达到 85. 0%。说明模型具有较好的预测能力。同时,数值模拟显示: 不同异质性程度下,此模型不仅能比较明显地区分有效变量组和冗余变量组,也能比较明显地区分单个有效变量和单个冗余变量。说明此模型具有较好的变量选择能力。在应用研究中,利用该模型选出的变量与已有实证研究有很多相同之处。对于不同板块而言,相同的指标有着不同的影响。这在以往研究

中鲜有讨论，本文方法为相关研究提供了新的思路。

需要指出，本文方法依据 TNR、TPR 和 Gmeans 选择调节参数。本质上是根据模型的泛化能力最大确定调节参数。在实际问题中，变量的选择和模型的泛化能力都是需要重视的方面。例如，宋彪等[14]针对财务预警研究明确阐述了这样的观点。但是，实际中往往难以同时做到两者最优。事实上，模型的变量选择能力和泛化能力存在一种权衡关系，研究者要根据具体问题在这两者之间寻求平衡(李扬等)[16]。

本文研究的一个重要前提是已知数据的异质性情况。这样的假定在实际中有应用价值。例如，本文应用研究中已知公司来自不同的板块。而在工业统计中，企业具体来自哪些门类也是已知的。但是，存在异质性情况未知的实际问题。Lin 和 Ng[9]、Su 等[8]针对响应变量为连续的面板数据讨论了这种情形。当响应变量为分类变量，且异质性情况未知，目前还鲜有文献讨论，未来需要进一步研究。此外，异质性广泛存在于空间数据，这类数据是当前研究的重要方面(叶倩婷等)[17]，对于此类数据的变量选择方法是未来值得研究的内容。

参考文献

[1] 谢宇. 回归分析[M]. 第一版. 北京:社会科学文献出版社，2010.

[2] Hsiao C. Analysis of Panel Data[M]. Cambridge: Cambridge University Press，2003.

[3] 李扬，曾宪斌. 面板数据模型的惩罚似然变量选择方法研究[J]. 统计研究，2014(3):83-89.

[4] 大数据中的统计方法"课题组. 大数据时代统计学发展的若干问题[J]. 统计研究，2017，34(1):5-11.

[5] 王松桂. 线性模型引论[M]. 第一版. 北京:科学出版社，2004.

[6] Su L，Chen Q. Testing homogeneity in panel data models with interactive fixed effects[J]. Econometric Theory，2013，29(6):1079-1135.

[7] Browning M，Carro J M. Dynamic binary outcome models with maximal heterogeneity[J]. Journal of Econometrics，2014，178(2):805-823.

[8] Su L，Shi Z，Phillips P C B. Identifying Latent Structures in Panel Data[J]. Econometrica，2016，84(6):2215-2264.

[9] Lin C C，Ng S. Estimation of panel data models with parameter heterogeneity when group membership is unknown[J]. Journal of Econometric Methods，2012，1(1):42-55.

[10] 马双鸽，王小燕，方匡南. 大数据的整合分析方法[J]. 统计研究，2015，32(11):3-11.

[11] 李扬，李竟翔，王园萍. 基于 AUC 回归的不平衡数据特征选择模型研究[J]. 统计与信息论坛，2015(5):10-16.

[12] Noah Simon，Jerome Friedman，Trevor Hastie，et al. A Sparse-Group Lasso[J]. Journal of Computational & Graphical Statistics，2013，22(2):231-245.

[13] Vincent M，Hansen N R. Sparse group lasso and high dimensional multinomial classification[J]. Computational Statistics & Data Analysis，2014(71):771-786.

[14] 宋彪，朱建明，李煦. 基于大数据的企业财务预警研究[J]. 中央财经大学学报，2015(6):55-64.

[15] 孔宁宁，魏韶巍. 基于主成分分析和 Logistic 回归方法的财务预警模型比较——来自我国制造业上市公司的经验证据[J]. 经济问题，2010(6):112-116.

[16] 李扬，朱建锋，谢邦昌. 变量选择方法及其在健康食品市场研究中的应用探究[J]. 统计与信息论坛，2013(10):17-24.

[17] 叶倩婷，龙志和. 层级数据空间误差自回归模型的估计方法研究[J]. 数量经济技术经济研究，2016，33(5):143-161.

作者简介

斯介生，男，2015 年毕业于中国人民大学统计学院，获经济学博士学位，现为杭州电子科技大学经济学院讲师，中国人民大学统计咨询研究中心兼职研究员。研究方向为决策与预测。

李扬，男，2010 年毕业于中国人民大学统计学院，获经济学博士学位，现为中国人民大学统计学院副教授，博士生导师，国际统计学会(ISI)推选会员，国际生物统计学会中国分会(IBS-China)青年理事委员，中国人民大学统计咨询研究中心主任。研究方向为决策与预测。

谢邦昌，男，1991 年毕业于台湾大学，获生物统计学博士学位，现为台北医学大学大数据研究中心及管理学院院长兼主任，中央财经大学统计学院博士生导师，对外经济贸易大学统计学院硕士生导师，对外经济贸易大学大数据与风险管理研究中心副主任。研究方向为数据挖掘。

(责任编辑:青 青)