Statistics and Application 统计学与应用, 2020, 9(6), 964-971
Published Online December 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.96101

Hans汉斯

混合效应模型在葡萄糖耐糖量测试中的应用研究

张银香，郭　靖，王　涛*

云南师范大学数学学院，云南　昆明
Email: *156289801@qq.com

收稿日期：2020年11月17日；录用日期：2020年12月2日；发布日期：2020年12月9日

摘　要

本文是对一组来自标准葡萄糖耐糖量实验的平衡纵向数据构建线性混合效应模型、*lasso*回归和广义线性混合模型，利用*AIC*准则以及*lasso*的变量选择选取最优的模型，并对所得模型进行模拟预测，然后计算模型的均方误差并进行比较，选出均方误差相对较小的模型。发现线性混合模型在这组数据中具有较好的预测效果。本文所有计算均用R软件完成。

关键词

线性混合效应模型，lasso回归，广义线性混合效应模型，模拟预测，均方误差

Application of Mixed Effect Model in Glucose Tolerance Test

Yinxiang Zhang, Jing Guo, Tao Wang*

College of Mathematics, Yunnan Normal University, Kunming Yunnan
Email: *156289801@qq.com

Received: Nov. 17th, 2020; accepted: Dec. 2nd, 2020; published: Dec. 9th, 2020

Abstract

This article is about a group of glucose sugar levels in experimental equilibrium longitudinal resistance from standard data to construct linear mixed effects models, lasso regression and generalized linear mixed models, using the AIC criterion as well as the lasso variable selection to select

*通讯作者。

文章引用: 张银香, 郭靖, 王涛. 混合效应模型在葡萄糖耐糖量测试中的应用研究[J]. 统计学与应用, 2020, 9(6): 964-971. DOI: 10.12677/sa.2020.96101

the optimal model and forecast the model simulated, then the mean square error of the model is calculated and compared, and the model with relatively small mean square error is selected. It is found that the linear mixed model has a good predictive effect in this set of data. All calculations in this paper are completed by R software.

Keywords

Linear Mixed Effect Model, Lasso Regression, Generalized Linear Mixed Effect Model, Simulation Prediction, Mean Square Error (mse)

1. 引言

在标准葡萄糖耐糖量实验中[1]，得到一组平衡纵向数据。我们的研究目的是：评估一下，在这个实验中对照组和肥胖组是否有显著差异。初步的探索表明构建合适的混合效应模型是分析此类数据的关键。

针对我们的数据，我们首先基于探索性数据分析，给出了模型的基本结构[2]，并根据 *AIC*，*BIC* 准则选择合适的协方差结构和固定效应[3] [4]，选出最佳模型。我们还用 *lasso* 方法进行变量选择，看这两种方法选择的变量是否一致。参数模型的缺点是对分布假设高度敏感，如果分布假设不能满足，那么结论会相差很多，甚至得到完全错误的结论，所以我们进一步用广义线性混合效应模型来拟合这组数据。最后，对三个模型进行 400 次模拟预测，比较一下哪种方法可以更好地拟合该数据。

2. 模型理论简介

2.1. 线性混合效应模型

线性混合效应模型是 *Laird* 和 *Ware* 于 1982 年提出的，它的一般形式为：

$$\begin{cases} Y_i = X_i\beta + Z_i b_i + \varepsilon_i \\ b_i \sim N_q(0, D) \\ \varepsilon_i \sim N_{n_i}(0, R_i) \\ b_1, \cdots, b_m, \varepsilon_1, \cdots, \varepsilon_m \text{独立} \end{cases} \tag{2.1}$$

这里的 $i = 1, 2, \cdots, m$，X_i 和 Z_i 是已知的设计矩阵，β 是 $p \times 1$ 维固定效应，b_i 是个体随机效应向量。ε_i 为随机误差向量。一般而言，Z_i 的列是 X_i 的子集。而 Y_i 是受试者响应向量，它服从以下这样一个多元正态分布：

$$\begin{cases} E[Y_i] = X_i\beta \\ Cov[Y_i] = R_i + Z_i G Z_i^{\mathrm{T}} = \Sigma_i \\ Y_i \sim N_{n_i}(X_i\beta, \Sigma_i) \end{cases} \tag{2.2}$$

向量 Y 代表完全观察到的数据集合，即 $Y = (Y_1, \cdots, Y_m)^{\mathrm{T}}$。因为 Y 是一个多元正态随机向量的线性组合，故 Y 的分布如下：

$$Y \sim N_n(X\beta, \Sigma)$$

上式中，$N = \sum_{i=1}^{m} n_i$，$X = (X_1, X_2, \cdots, X_m)^{\mathrm{T}}$，$\Sigma = diag(\Sigma_1, \Sigma_2, \cdots, \Sigma_m)$，$\Sigma$ 的这种形式假定了不同个

体之间的所有观察结果都是独立的。

2.2. lasso 方法

lasso (*Least absolute shrinkage and selection operator*)是由 *Tibshirant* 于 1996 年提出的[5]，该方法可以压缩模型的系数，并将一些系数置为零。该方法的主要特点是可以同时进行参数估计和变量选择。

假定数据为$\left(x^i, y_i\right), i=1,\cdots,N$，*其中*，$x^i=\left(x_{i1},\cdots,x_{ip}\right)^{\mathrm{T}}$是预测变量，$y_i$是响应。在通常的回归中，我们假设观测值是独立的，或者给定x_{ij}时，y_i是有条件独立的。我们假设x_{ij}是标准化的，所以$\sum_i x_{ij}/N=0$，$\sum_i x_{ij}^2/N=1$。

令$\hat{\beta}=\left(\hat{\beta}_1,\cdots,\hat{\beta}_p\right)^{\mathrm{T}}$，*lasso* 估计$\left(\hat{\alpha},\hat{\beta}\right)$定义为：

$$\left(\hat{\alpha},\hat{\beta}\right)=\arg\min\left\{\sum_{i=1}^{N}\left(y_i-\alpha-\sum_j\beta_j x_{ij}\right)^2\right\}，\text{约束为}\sum_j\left|\beta_j\right|\le t$$

其中，$t\ge 0$，是调整参数，对于所有的 t，α 的所有解$\hat{\alpha}=\bar{y}$，不失一般性，假设$\bar{y}=0$，因此可以省略α。

2.3. 广义线性混合效应模型

广义线性混合模型(*GLMM*)的模型结合了广义线性模型和线性混合模型的优点，进一步扩展了广义线性混合模型，包括随机变量，以说明纵向或聚类数据的变化和相关性。另一方面，*GLMM* 也是对线性混合模型的扩展，允许非正态分布的响应数据。使用 *GLMM* 的一个优点是响应数据不必是正态分布的数据。*GLMM* 利用数据的真实本质，而不是依赖于数据的正态分布或转换假设。

广义线性混合效应模型由三部分组成：

(1) 随机组成：

1) 响应；$Y_{ij}\mid b_i\sim\exp\left\{\dfrac{y\theta-b(\theta)}{a(\phi)}+c(y,\theta)\right\}$

2) 随机效应：$b_i\sim N_q(0,G)$

(2) 系统组成：$\eta_{ij}=X_{ij}^{\mathrm{T}}\beta+Z_{ij}^{\mathrm{T}}b_i$

(3) 链接函数：$g\left\{E\left[y_{ij}\mid b_i\right]\right\}=X_{ij}^{\mathrm{T}}\beta+Z_{ij}^{\mathrm{T}}b_i\ (i=1,\cdots,m;\ j=1,\cdots,n_i)$

其中，y_{ij}是第 i 个人在第 j 个时间点的响应，而 x_{ij} 和 z_{ij} 分别是与组合和随机变量相关的协变量向量。β 和 b_i 分别是固定效应和随机效应。

3. 数据说明

本文数据来源于《Randomization Analysis of the Completely Randomized Design Extended to Growth and Response Curves》这篇文献。在一项有关高血糖和相对高胰岛素血症的研究中，对科罗拉多大学医学中心的儿科临床研究病房的 13 名对照组和 20 名肥胖组患者进行了标准的葡萄糖耐糖量试验。下表记录了在他们口服标准剂量葡萄糖后，分别在 0，0.5，1，1.5，2，3，4，5 小时取其血液样本，测其血浆无机磷值得到的，每个病人都测量了 8 次[6]。其中数据包括了编号(subject)、组别(treatment)、测量时间(time)、及每个时刻的血浆无机磷测量值(plasma inorganic phosphate measurement)。表 1 给出了部分病人的数据。

4. 数据分析

4.1. 线性混合效应模型

本文首先运用线性混合效应模型来拟合数据，由于线性混合效应模型假定数据是服从正态分布的，

所以分析数据时先进行正态性检验，原始测量值数据直方图和 QQ 图如图 1 所示。

Table 1. Data of some patients
表 1. 部分病人数据

subject	treatment	Plasma Inorganic Phosphate (mg/dl)							
		hours after glucose challenge							
		0.0	0.5	1.0	1.5	2.0	3.0	4.0	5.0
1	0	4.3	3.3	3.0	2.6	2.2	2.5	3.4	4.4[a]
2	0	3.7	2.6	2.6	1.9	2.9	3.2	3.1	3.9
3	0	4.0	4.1	3.1	2.3	2.9	3.1	3.9	4.0
4	0	3.6	3.0	2.2	2.8	2.9	3.9	3.8	4.0
5	0	4.1	3.8	2.1	3.0	3.6	3.4	3.6	3.7
14	1	4.3	3.3	3.0	2.6	2.2	2.5	2.4	3.4a
15	1	5.0	4.9	4.1	3.7	3.7	4.1	4.7	4.9
16	1	4.6	4.4	3.9	3.9	3.7	4.2	4.8	5.0
17	1	4.3	3.9	3.1	3.1	3.1	3.1	3.6	4.0
18	1	3.1	3.1	3.3	2.6	2.6	1.9	2.3	2.7

注：treatment 为 0 表示对照组，1 表示肥胖组。

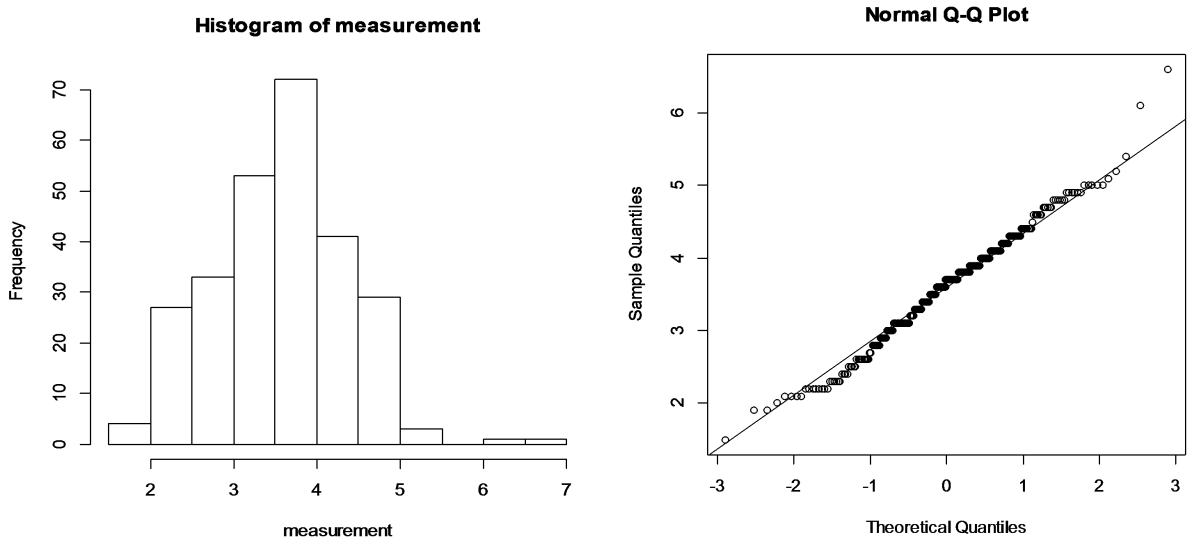

Figure 1. Data histogram and QQ chart of plasma inorganic phosphorus
图 1. 血浆无机磷测量值数据直方图和 QQ 图

为了研究不同组别对血浆无机磷测量值的结果是否有所不同，文章中画出了 0，0.5，1，1.5，2，3，4，5 小时这 8 个时间点对应的响应变量(measurement)均值折线图如图 2 所示[7]。

从图象中可以看出来测量值在 2 小时前后变化很大，不同组别有显著的差异，并且测量值对于时间的变化规律不是线性的，在模型中可能需要加入时间的高次项。基于以上分析，我们建立如下线性混合效应模型：

$$y_{ij} = \beta_0 + \beta_1 \cdot \text{time}_{ij} + \beta_2 \cdot \text{time}_{ij}^2 + \beta_3 \cdot \text{time}_{ij}^3 + \beta_4 \cdot \text{treatment} + \beta_5 \cdot \text{treatment} \cdot \text{time} + \varepsilon_{ij}$$

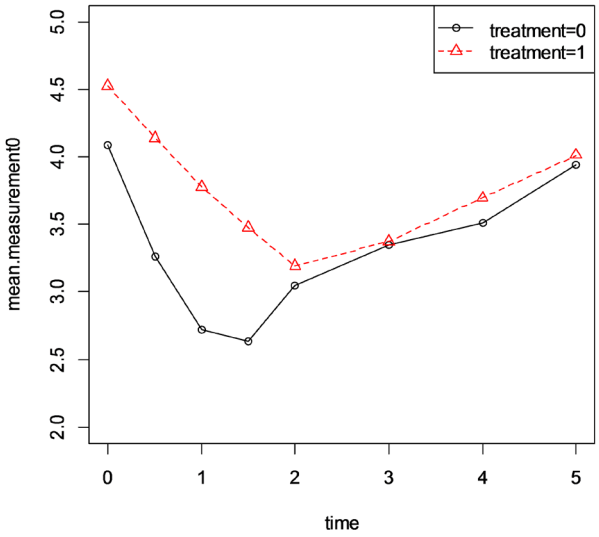

Figure 2. Mean line graph under different groups
图 2. 不同组别下的均值折线图

首先，对上述模型选取不同的协方差结构，基于 AIC，BIC 准则，我们选择个体内部的协方差结构为 AR(1)。这与数据的探索性分析中，与散点图得出的结果一致。然后，我们选取最佳的随机效应，根据 AIC，BIC 准则，当选择时间(time)为随机效应时模型更优。最后，基于我们选择的个体内部的协方差结构和随机效应的结构可以选择固定效应，固定效应包括：treatment，时间的三次多项式，时间与组别的交互项。根据模型的基本选取准则，接着我们对模型进行参数估计，由于经典最大似然估计方法估计出来的方差是有偏估计，因此我们这里使用限制最大似然估计(REML)方法来获得方差分量的最大似然估计(见表 2)。限制最大似然估计的想法是将数据分为两部分，一部分用来估计固定效应，另一部分用来估计协方差参数[8] [9]。可得模型及其限制极大似然估计为：

$$y = 3.947508 - 1.313492 * \text{time} + 0.467245 * \text{time2} - 0.041090 * \text{time3}$$

$$+ 0.667951 * \text{treatment} - 0.114951 * \text{time} * \text{treatment}$$

Table 2. Restricted maximum likelihood (REML) estimation for linear mixed models
表 2. 线性混合模型的限制最大似然(REML)估计

Fixed effect	Coefficient		T-statistic	P-value
	REML	ML	REML	REML
Intercept	3.947508	3.944325	22.325188	0.0000
time	−1.313492	−1.313265	−9.638033	0.0000
time2	0.467245	0.467722	7.427107	0.0000
time3	−0.041090	−0.041153	−5.080926	0.0000
treatment	0.667951	0.673348	3.056347	0.0046
time:treatment	−0.114951	−0.116712	−2.478084	0.0139

4.2. lasso 方法

用 R 软件里的 *lars* 程序包实现 *lasso* 变量选择。函数 *lars*()，提供了通过回归变量 *x* 和因变量 *y* 求解

其回归模型。

在对数据用线性混合效应建立模型的基础上，我们用 *lasso* 进行模型回归及其变量选择[10]。如图3表示在进行 *lasso* 回归时，自变量被选入的顺序。

可以看到图3中的竖线对应的是 *lasso* 中迭代的次数，对应的系数值不为0的自变量即为选入的，竖线的标号与 *lasso* 回归中的 *step* 相对应。

根据 MallowsCp 统计量，即 Cp 值选择 Cp 值最小的模型及模型对应系数，见表3。

4.3. 广义线性混合效应模型

取链接函数为密度函数，建立如下模型并采用极大似然(ML)估计方法来进行参数估计(见表4)。

$$\mu_{ij} = \beta_0 + \beta_1 \cdot \text{time}_{ij} + \beta_2 \cdot \text{treatment} + \beta_3 \text{time}_{ij} \cdot \text{treatment} + b_i$$

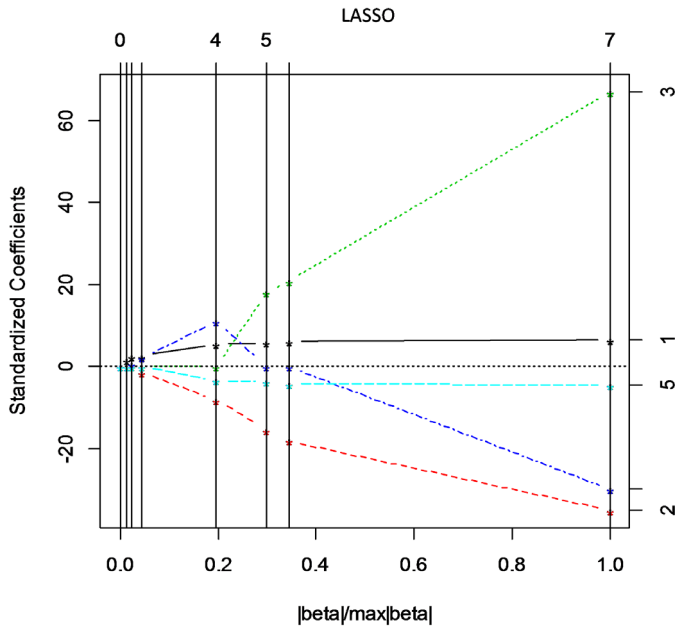

Figure 3. Lasso regression
图 3. lasso 回归

Table 3. The corresponding coefficient of lasso regression
表 3. lasso 回归对应系数

	Intercept	time	time2	time3	treatment	time: treatment
Coefficient	3.87211	−1.321535	0.488999	−0.044089	0.806052	−0.163956

Table 4. Maximum likelihood estimation of generalized linear mixed effects model
表 4. 广义线性混合效应模型的最大似然估计

	Value	Std.Error	DF	t-value	p-value
Intercept	3.1423740	0.16674730	229	18.845127	0.0000
time	0.0828566	0.03647356	229	2.271691	0.0240
Treatment	0.8081669	0.21411132	31	3.774517	0.0007
time:treatment	−0.1645818	0.04681937	229	−3.515250	0.0005

通过以上三种方法对该组数据进行拟合，发现每个系数都比较显著，没有剔除的项，三种方法都表明组别(treatment)的系数较大，说明肥胖对血浆的无机磷酸盐含量有较大影响。就线性混合效应模型与lasso 方法来看，两者之间的变量选择结果一致，而且参数估计的结果也比较接近。

5. 模型预测与比较

接下来我们用所建立的模型进行模拟预测，首先生成与原数据同分布的数据，然后再用所得模型进行预测，我们用均方误差来评价模型的预测能力。均方误差的定义为：

$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{observed}_i-\mathrm{predicted}_i\right)^2$$

其中，N 为样本容量。当预测值与真实值完全相同时为 0，误差越大，MSE 值越大。如图 4 给出了三个模型的均方误差

从图 4 可以看出，黑色的水平线代表了线性混合模型的均方误差的平均值，要低于其他两种模型的均方误差的平均值，而且线性混合模型整体的均方误差也低于其他两种的均方误差，所以在该组数据下我们采用线性混合模型来拟合会更好，这是因为这组数据本身接近正态分布，有着很好的性质。在实际应用中，我们要基于数据本身选择更优的模型以便更好的解决问题。

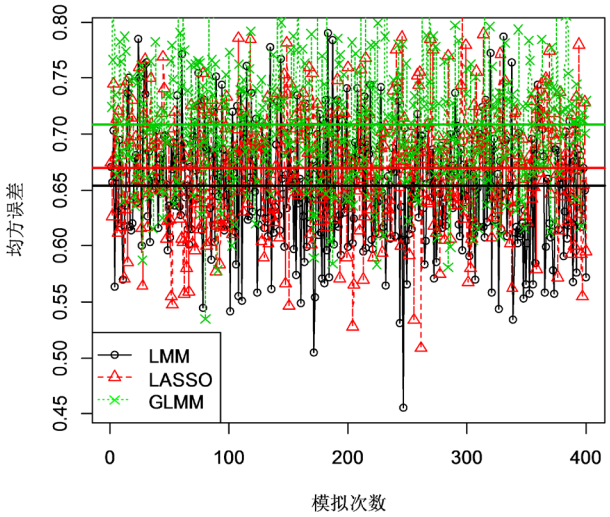

Figure 4. Mean square error chart
图 4. 均方误差图

6. 结论

基于对数据的探索性分析，如果数据接近正态分布，我们可以通过线性混合效应模型来对数据进行分析和预测。如果数据不是正态分布，可以对数据进行转换，然后用线性混合效应模型分析数据。在本文中，线性混合效应模型的拟合精度要比 *lasso* 回归和广义线性混合模型的拟合精度要高。我们发现肥胖对血浆中的无机磷酸盐含量有显著的影响，进而会导致高血压、糖尿病等多种疾病，所以，对于在儿童中由肥胖引起的血浆中的无机磷酸盐含量变化研究是有必要的。

参考文献

[1] Zerbe, G.O. (1979) Randomization Analysis of the Completely Randomized Design Extended to Growth and Response

Curves. *Journal of the American Statistical Association*, **74**, 215-221. https://doi.org/10.1080/01621459.1979.10481640

[2] Verbeke, G., Molenberghs, G. and Rizopoulos, D. (2010) Random Effects Models for Longitudinal Data. In: van Montfort, K. and Oud J., Satorra A., Eds., *Longitudinal Research with Latent Variables*, Springer, Berlin, Heidelberg, 37-96. https://doi.org/10.1007/978-3-642-11760-2_2

[3] 王云. 线性混合效应模型协方差阵的估计问题[D]: [硕士学位论文]. 北京: 北京工业大学, 2006.

[4] Wolfinger, R.D. (1995) Heterogeneous Variance-Covariance Structures for Repeated Measures. *Journal of Agricultural, Biological, and Enviromental Statistics*, **1**, 205-230. https://doi.org/10.2307/1400366

[5] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statal Society, Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[6] 赵晋芳. 重复测量线性混合模型在医学研究中的应用[D]: [硕士学位论文]. 太原: 山西医科大学, 2002.

[7] 吴喜之. 复杂数据统计方法:基于 R 的应用[M]. 北京: 中国人民大学出版社, 2012.

[8] Pan, J. and Fang, K. (2002) Growth Curve Models and Statistical Diagnostics. Springer Series in Statistics, New York. https://doi.org/10.1007/978-0-387-21812-0

[9] Davidian, M. (2005) Applied Longitudinal Data Analysis. Department of Statistics, North Carolina State University, Raleigh, North Carolina.

[10] 曲婷, 王静. 基于 Lasso 方法的平衡纵向数据模型变量选择[J]. 黑龙江大学自然科学学报, 2012, 29(6): 715-722.