Statistical and Application 统计学与应用, 2014, 3, 68-75
Published Online June 2014 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2014.32010

Hans汉斯

Statistical Analysis for Nonlinear Joint Mean and Variance Models

Mengqi Zhou, Dengke Xu, Jiahong Yang, Mengying Wang

Department of Statistics, Zhejiang Agriculture and Forest University, Hangzhou
Email: 175384319@qq.com

Received: Apr. 25th, 2014; revised: May 23rd, 2014; accepted: Jun. 3rd, 2014

Abstract

We propose nonlinear joint mean and variance models in this paper and investigate the estimate for unknown parameters in the model based on Gauss-Newton iterative algorithm. Furthermore, we make some simulations to show that the proposed procedure works satisfactorily. Lastly, two real examples are presented to illustrate the proposed methodology.

Keywords

Nonlinear Joint Mean and Variance Models, Heteroscedasticity, Gauss-Newton, Maximum Likelihood Estimate

非线性联合均值方差模型的统计分析

周梦齐，徐登可，杨佳红，王梦滢

浙江农林大学统计系，杭州
Email: 175384319@qq.com

收稿日期：2014年4月25日；修回日期：2014年5月23日；录用日期：2014年6月3日

摘　要

在提出非线性均值方差模型的基础上，研究了该模型中未知参数的估计问题。主要是基于Gauss-Newton

迭代算法给出该模型中未知参数的极大似然估计。通过大量随机模拟实验验证了所提出方法的有效性。最后，结合实际问题数据验证了该模型与方法具有实用性和可行性。

关键词

非线性联合均值方差模型，异方差，Gauss-Newton，极大似然估计

1. 引言

在回归模型中，对误差项进行等方差假设是一个标准的假设。违反这个假设，估计量的有效性就可能得不到保证。因此很重要也很有必要去处理回归分析中的异方差情况。很多作者已经讨论了异方差情况下不同模型的估计和检验等统计推断问题。White[1]提出了在异方差线性回归模型下参数协方差阵的估计，这种估计不依赖于异方差的结构形式模型；Andrews[2]介绍了异方差和自相关函数形式未知的情形下协方差阵的估计；Smyth[3]基于异方差回归模型研究了模型中未知参数的限制极大似然估计。

事实上，随着人们对现实世界越来越深入的认识，很多现实生活的事件、现象、过程等也表现得越来越复杂，这也将导致我们研究的实际数据也是错综复杂的。如果只是用简单的统计模型来描述和研究，很多分析已经不能得到真实的接近实际的结果。因此我们很有必要针对这些复杂现象，采用比较复杂的模型来描述，联合均值方差模型就是其中一种。有关基于均值方差同时建模的双重回归异方差模型也已有了大量的研究成果。Park[4]在高斯模型中提出了方差参数的对数线性模型，采用两阶段过程来估计参数；Harvey[5]在一般条件下讨论了均值和方差效应的极大似然估计和子序列似然比检验；Aitkin[6]提供了联合均值和方差模型的极大似然估计，并且把它应用到了 Minitab tree 数据中；Verbyla[7]利用限制极大似然估计参数和在 MLE 和限制似然下考虑了模型的影响诊断分析；Wu 和 Li[8]提出了逆高斯分布的均值和方差联合建模模型的同时变量选择问题；Xu 等[9]基于惩罚伪似然研究了双重广义线性模型的变量选择问题；吴刘仓等[10]基于 Box-Cox 变换下研究了联合均值与方差模型的参数估计；马婷等[11]基于偏正态分布联合位置、尺度与偏度模型给出了该模型参数的估计方法；徐登可等[12]基于双重 logistic 回归模型对影响妊高病的危险因素进行变量选择和预测分析。

总之，有关异方差数据处理方法以及用均值方差模型处理异方差数据都已经有了很多的研究成果。虽然，对于均值方差同时建模双重回归异方差模型已经有了大量的研究成果，但是上述成果大多数都是基于线性模型，很少有推广到非线性模型。而在现实生活数据中的变量与变量之间的关系可能存在非线性关系，因此我们很有必要基于非线性模型发展一种非线性联合均值方差模型。

故本文主要目的是基于Gauss-Newton迭代算法研究提出的非线性联合均值方差模型的极大似然估计以及考虑其应用。我们的方法在非线性的基础上能同时对均值和方差建立模型，使得均值方差模型更加一般化，也更具有应用的广泛性。最后，通过随机模拟和实例研究分析表明所提出的模型与方法是有用和有效的。

本文的组织结构安排如下：第 2 节，首先介绍了非线性均值方差模型；然后给出了模型中未知参数的极大似然估计。第 3 节，详细介绍了 Gauss-Newton 迭代算法。第 4 节，通过随机模拟实验验证该方法的有效性。第 5 节，结合实际问题数据验证该模型与方法的实用性与可行性。最后是本文的小结与讨论。

2. 非线性联合均值方差模型

首先针对异方差数据和基于非线性回归，我们既对均值建模，同时又对方差进行建模，提出如下非线性联合均值方差模型：

$$\begin{cases} y_i \sim N\left(\mu_i, \sigma_i^2\right) \\ \mu_i = f\left(x_i, \beta\right) \\ \sigma_i^2 = g\left(h_i, \gamma\right) \\ i = 1, 2, \cdots, n. \end{cases} \tag{1}$$

其中 $y_i = \left(y_1, y_2, \cdots, y_n\right)^T$ 为响应变量，$x_i = \left(x_{i1}, x_{i2}, \cdots, x_{ip}\right)^T$ 和 $h_i = \left(h_{i1}, h_{i2}, \cdots, h_{iq}\right)^T$ 分别为影响均值部分和方差部分的解释变量，$\beta = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^T$ 是非线性均值模型中的未知参数，$\gamma = \left(\gamma_1, \gamma_2, \cdots, \gamma_q\right)^T$ 是非线性方差模型中的未知参数。x_i、h_i两个解释变量可能完全不相同，完全相同或者部分相同，即均值模型、方差模型可能包含不同的解释变量，包含相同的解释变量也可能包含部分相同的解释变量。$f\left(x_i, \beta\right)$是一个含有未知参数 β 的已知非线性函数，$g\left(h_i, \gamma\right)$ 是一个含有未知参数 γ 的已知非线性函数。若 f 和 g 变成了线性函数，那么 $f\left(x_i, \beta\right) = x_i^T\beta$，$g\left(h_i, \gamma\right) = h_i^T\gamma$，即变成了线性均值方差模型，也就是说线性均值方差模型是该模型的一种特例。

由模型(1)可以得到：

$$p\left(y_i\right) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{\left[y_i - f\left(x_i, \beta\right)\right]^2}{2\sigma_i^2}\right\} \tag{2}$$

对(2)式两边取自然对数，得到：

$$\log p\left(y_i\right) = -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log 2\sigma_i^2 - \frac{\left[y_i - f\left(x_i, \beta\right)\right]^2}{2\sigma_i^2} \tag{3}$$

由(3)式可得到对数似然函数：

$$l\left(\beta, \gamma\right) = \log p\left(y\right) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\log g\left(h_i, \gamma\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i - f\left(x_i, \beta\right)\right]^2}{g\left(h_i, \gamma\right)} \tag{4}$$

3. 参数的极大似然估计

3.1. Gauss-Newton 迭代算法

由于该模型无法通过普通的极大似然估计得到参数估计的显示表达式，所以我们采用了 Gauss-Newton 迭代算法。

为了方便，令 $\theta = \left(\beta^T, \gamma^T\right)^T$,则 $l\left(\beta, \gamma\right) = l\left(\theta\right)$，因此

$$U\left(\theta\right) = \frac{\partial l\left(\theta\right)}{\partial\theta} = \left(U_1^T\left(\beta\right), U_2^T\left(\gamma\right)\right)^T \tag{5}$$

其中

$$U_1\left(\beta\right) = \frac{\partial l\left(\beta, \gamma\right)}{\partial\beta},\quad U_2\left(\gamma\right) = \frac{\partial l\left(\beta, \gamma\right)}{\partial\gamma}, \tag{6}$$

$$\frac{\partial l\left(\beta, \gamma\right)}{\partial\beta} = -\frac{1}{2}\sum_{i=1}^{n}\frac{-2\left[y_i - f\left(x_i, \beta\right)\right]}{g\left(h_i, \gamma\right)}\frac{\partial f\left(x_i, \beta\right)}{\partial\beta} = \sum_{i=1}^{n}\frac{y_i - f\left(x_i, \beta\right)}{g\left(h_i, \gamma\right)}\frac{\partial f\left(x_i, \beta\right)}{\partial\beta} \tag{7}$$

$$\frac{\partial l\left(\beta, \gamma\right)}{\partial\gamma} = -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{g\left(h_i, \gamma\right)}\frac{\partial g\left(h_i, \gamma\right)}{\partial\gamma} + \frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i - f\left(x_i, \beta\right)\right]^2}{g\left(h_i, \gamma\right)^2}\frac{\partial g\left(h_i, \gamma\right)}{\partial\gamma} \tag{8}$$

另外，令

$$H(\theta)=\begin{pmatrix}\dfrac{\partial^2 l(\theta)}{\partial\beta\partial\beta^T} & \dfrac{\partial^2 l(\theta)}{\partial\beta\partial\gamma^T}\\ \dfrac{\partial^2 l(\theta)}{\partial\beta^T\partial\gamma} & \dfrac{\partial^2 l(\theta)}{\partial\gamma\partial\gamma^T}\end{pmatrix}，\text{其中} \tag{9}$$

$$\frac{\partial^2 l(\beta,\gamma)}{\partial\beta\partial\beta^T}=-\sum_{i=1}^{n}\frac{1}{g(h_i,\gamma)}\frac{\partial f(x_i,\beta)}{\partial\beta}\frac{\partial f(x_i,\beta)}{\partial\beta^T}+\sum_{i=1}^{n}\frac{y_i-f(x_i,\beta)}{g(h_i,\gamma)}\frac{\partial^2 f(x_i,\beta)}{\partial\beta\partial\beta^T} \tag{10}$$

$$\begin{aligned}\frac{\partial^2 l(\beta,\gamma)}{\partial\gamma\partial\gamma^T}&=\frac{1}{2}\sum_{i=1}^{n}\frac{1}{g(h_i,\gamma)^2}\frac{\partial g(h_i,\gamma)}{\partial\gamma}\frac{\partial g(h_i,\gamma)}{\partial\gamma^T}-\sum_{i=1}^{n}\frac{\left[y_i-f(x_i,\beta)\right]^2}{g(h_i,\gamma)^3}\frac{\partial g(h_i,\gamma)}{\partial\gamma}\frac{\partial g(h_i,\gamma)}{\partial\gamma^T}\\&\quad-\frac{1}{2}\sum_{i=1}^{n}\frac{\left\{g(h_i,\gamma)-\left[y_i-f(x_i,\gamma)\right]^2\right\}}{g(h_i,\gamma)^2}\frac{\partial^2 g(h_i,\gamma)}{\partial\gamma\partial\gamma^T}\end{aligned} \tag{11}$$

$$\frac{\partial^2 l(\beta,\gamma)}{\partial\beta\partial\gamma^T}=\sum_{i=1}^{n}-\frac{y_i-f(x_i,\beta)}{g(h_i,\gamma)^2}\frac{\partial f(x_i,\beta)}{\partial\beta}\frac{\partial g(h_i,\gamma)}{\partial\gamma^T} \tag{12}$$

$$\frac{\partial^2 l(\beta,\gamma)}{\partial\beta^T\partial\gamma}=\sum_{i=1}^{n}-\frac{y_i-f(x_i,\beta)}{g(h_i,\gamma)^2}\frac{\partial f(x_i,\beta)}{\partial\beta^T}\frac{\partial g(h_i,\gamma)}{\partial\gamma} \tag{13}$$

最后，将(5)-(13)这 9 个式子带入下面的(14)式进行迭代计算，

$$\theta_1=\theta_0-\left(H(\theta)\right)^{-1}U(\theta)\Big|_{\theta=\theta_0} \tag{14}$$

直到$\left|\theta_1-\theta_0\right|<\delta$，即认为$\theta_1$为$\hat{\theta}$的极大似然估计的近似值，其中$\delta$为预先给定的充分正小数，如$\delta=10^{-6}$。

3.2. 迭代步骤

给出以下算法步骤对模型(1)中的参数进行极大似然估计迭代计算。

步骤 1：给定参数的迭代初值$\theta^{(0)}=\left(\left(\beta^{(0)}\right)^T,\left(\gamma^{(0)}\right)^T\right)^T$；

步骤 2：给定当前值$\theta^{(m)}=\left(\left(\beta^{(m)}\right)^T,\left(\gamma^{(m)}\right)^T\right)^T$，代入下式进行迭代更新

$$\theta^{(m+1)}=\theta^{(m)}-\left(H\left(\theta^{(m)}\right)\right)^{-1}U\left(\theta^{(m)}\right);$$

步骤 3：重复步骤 2，直到迭代收敛。

4. 模拟

接下来我们对上述参数估计方法的有限样本性质进行模拟研究，参数的估计精度使用均方误差(MSE)来评价和衡量，其定义如下：

$$MSE\left(\hat{\beta}_i\right)=E\left(\hat{\beta}_i-\beta_{0i}\right)^2\quad i=1,2,\cdots,p,$$

$$MSE\left(\hat{\gamma}_j\right)=E\left(\hat{\gamma}_j-\gamma_{0j}\right)^2\quad j=1,2,\cdots,q.$$

其中，β_0和γ_0分别是β和γ的真值，β_{0i}和γ_{0j}分别是β_0和γ_0的第i个分量和第j个分量。我们将通过下

面两个具体的非线性联合均值方差模型例子来说明所提出模型与方法的有效性。

4.1. 例子 1

根据模型(1)，建立如下的具体的非线性联合均值方差模型(15)，根据该模型并产生模拟数据

$$\begin{cases} y_i \sim N\left(\mu_i, \sigma_i^2\right) \\ \mu_i = \beta_1 \exp\left(x_{i1}\beta_2\right) \\ \sigma_i^2 = \prod_{j=1}^{q} h_{ij}^{\gamma_j} = \exp\left(\sum_{j=1}^{q} \gamma_j \log h_{ij}\right) \\ i = 1, 2, \cdots, n. \end{cases} \tag{15}$$

其中 $y_i\left(i=1,2,\cdots,n\right)$ 相互独立，且服从正态分布 $N\left(\mu_i,\sigma_i^2\right)$，$x_i$ 和 h_i 的分量分别相互独立，且 x_i 产生于均匀分布 $U\left(-1,1\right)$，h_i 产生于均匀分布 $U\left(0,1\right)$。β_1 的真值取1.0，β_2 的真值取1.0，γ 的真值为 $\left(0.8,0.8,0.8\right)^T$，取样本量 $n=80,120$和160，重复模拟2000次。具体模拟结果见表1。

表1结果显示，模型(15)中的参数的估计随着样本量的递增越来越接近真实值，参数极大似然估计的均方误差也越来越小，这说明模型(15)所使用的极大似然估计方法取得了较理想的效果。

4.2. 例子 2

根据模型(1)，建立如下的具体的非线性均值方差模型(16)，根据该模型并产生模拟数据

$$\begin{cases} y_i \sim N\left(\mu_i, \sigma_i^2\right) \\ \mu_i = \beta_1 \exp\left(x_{i1}\beta_2\right) \\ \sigma_i^2 = \exp\left(h_i^T \gamma\right) \\ i = 1, 2, \cdots, n. \end{cases} \tag{16}$$

其中 $y_i\left(i=1,2,\cdots,n\right)$ 相互独立，且服从正态分布 $N\left(\mu_i,\sigma_i^2\right)$，$x_i$ 和 h_i 的分量分别相互独立，且 x_i 产生于均匀分布 $U\left(-1,1\right)$，h_i 产生于均匀分布 $U\left(-1,1\right)$。β_1 的真值取1.0，β_2 的真值取1.0，γ 的真值取 $\left(0.8,0.8,0.8\right)^T$，取样本量 $n=80,120$和160，重复模拟2000次。具体模拟结果见表2。

Table 1. Maximum likelihood estimate of unknown parameters in nonlinear joint mean and variance models in Example 1
表 1. 例 1 中非线性联合均值方差模型中未知参数的极大似然估计结果

	β				γ					
n	$\hat{\beta}_1$	$\hat{\beta}_2$	$MSE\left(\hat{\beta}_1\right)$	$MSE\left(\hat{\beta}_2\right)$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$MSE\left(\hat{\gamma}_1\right)$	$MSE\left(\hat{\gamma}_2\right)$	$MSE\left(\hat{\gamma}_3\right)$
80	1.0001	1.0013	0.0011	0.0015	0.8352	0.7998	0.8288	0.0350	0.0357	0.0316
120	0.9994	1.0010	0.0003	0.0008	0.8209	0.8192	0.8206	0.0179	0.0179	0.0190
160	1.0000	1.0004	0.0002	0.0006	0.8166	0.8140	0.8156	0.0125	0.0136	0.0130

Table 2. Maximum likelihood estimate of unknown parameters in nonlinear joint mean and variance models in Example 2
表 2. 例 2 中非线性联合均值方差模型中未知参数的极大似然估计结果

	β				γ					
n	$\hat{\beta}_1$	$\hat{\beta}_2$	$MSE\left(\hat{\beta}_1\right)$	$MSE\left(\hat{\beta}_2\right)$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$MSE\left(\hat{\gamma}_1\right)$	$MSE\left(\hat{\gamma}_2\right)$	$MSE\left(\hat{\gamma}_3\right)$
80	0.9971	1.0038	0.0144	0.0342	0.8115	0.8138	0.8249	0.0947	0.0958	0.0999
120	0.9989	1.0036	0.0097	0.0230	0.8047	0.8057	0.8111	0.0601	0.0575	0.0606
160	0.9989	1.0004	0.0071	0.0162	0.8088	0.8088	0.8100	0.0425	0.0455	0.0407

表 2 结果显示，模型(16)中的参数的估计随样本量的递增越来越接近真实值，参数极大似然估计的均方误差也越来越小，这说明模型(16)所使用的极大似然估计方法取得了较理想的效果。

5. 实例分析

5.1. 伦福德冷却实验数据

1978 年，Count Rumford 得到一组摩擦生热的数据[13]。首先在一个固定的炮管内插入一只钝管，应用螺丝固定在炮管的底部。让一对马连续转动达 30 分钟，然后再设置一只温度计。在将近 45 分钟内，每隔一段时间观察温度的变化，并记录温度的大小。

利用 SPSS 软件对伦福德数据进行正态性检验，得到图 1 为伦福德数据正态检验的 P-P 图，我们可以从图中发现，伦福德数据基本服从或近似服从正态分布。

因此利用模型(15)，建立如下模型：

$$\begin{cases} y_i \sim N\left(\mu_i, \sigma_i^2\right) \\ \mu_i = \beta_1 \exp\left(x_{1i}\beta_2\right) \\ \sigma_i^2 = \exp\left(\gamma_1 \log h_{1i}\right) \\ i = 1, 2, \cdots, n. \end{cases} \tag{17}$$

其中，y_i 为不同时间炮管的温度，x_1 为时间(在该模型中令 x_1 与 h_1 相同)，通过计算可得：$\hat{\beta}_1 = 125.8717$，$\hat{\beta}_2 = -0.0035$，$\hat{\gamma}_1 = -0.1302$。这也表明该数据中变量间存在一定的非线性关系。

5.2. 氟哌啶醇血浆浓度数据

1975 年，Wagner 记录了氟哌啶醇血浆浓度的数据[13]。

利用 SPSS 软件对氟哌啶醇血浆浓度数据进行正态性检验，得到图 2 为氟哌啶醇血浆浓度数据正态检验的 P-P 图，我们可以从图中发现，氟哌啶醇血浆浓度数据基本服从或近似服从正态分布。

因此利用模型(16)，建立如下模型：

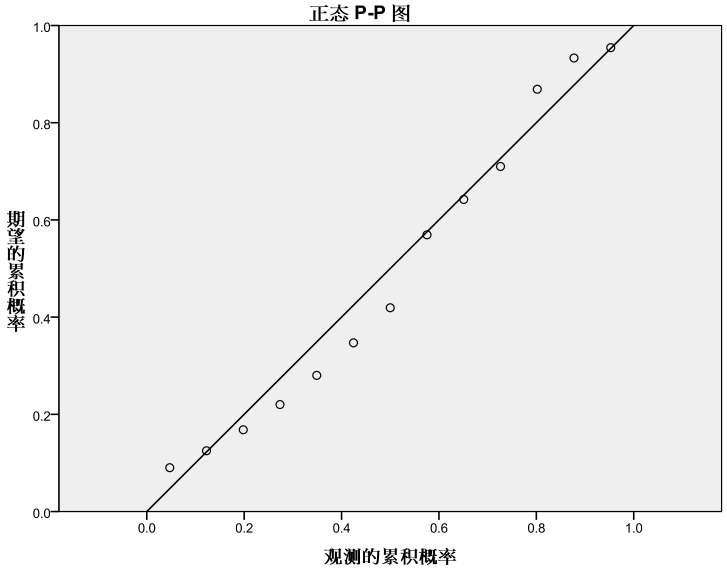

Figure 1. Normal P-P plot for Rumford data
图 1. 伦福德数据的正态 P-P 图

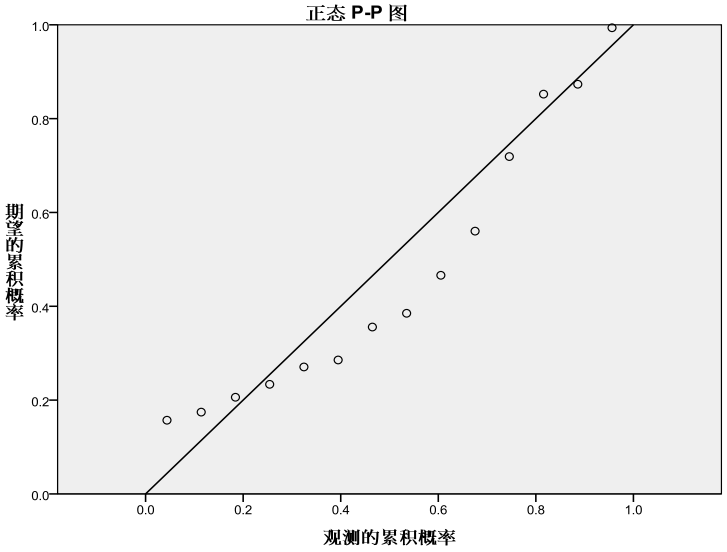

Figure 2. Normal P-P plot for concentration data of haloperidol plasma
图 2. 氟哌啶醇血浆浓度数据的正态 P-P 图

$$\begin{cases} y_i \sim N\left(\mu_i, \sigma_i^2\right) \\ \mu_i = \beta_1 \exp\left(x_{1i}\beta_2\right) \\ \sigma_i^2 = \exp\left(\gamma_1 h_{1i}\right) \\ i = 1, 2, \cdots, n. \end{cases} \tag{18}$$

其中， y_i 为氟哌啶醇血浆浓度， x_1 为时间(在该模型中令 x_1 与 h_1 相同)，通过计算可得： $\hat{\beta}_1 = 3.9037$ ， $\hat{\beta}_2 = -0.4164$ ， $\hat{\gamma}_2 = -0.0737$ 。这也表明该数据中变量间也存在一定的非线性关系。

6. 结论

本文建立了非线性联合均值方差模型，主要研究了该模型中未知参数的极大似然估计，并介绍了极大似然估计中常用的迭代算法——Gauss-Newton 迭代算法的具体实现步骤。模拟结果显示，通过 Gauss-Newton 迭代算法得到了较为满意的联合模型的参数估计结果，并且在实例分析中，对伦福德冷却实验数据和氟哌啶醇血浆浓度数据的应用也表明了该模型和所运用的方法是有用和有效的。另外在现实生活中，缺失数据也是经常碰到的复杂数据类型之一，以后可以运用合适的借补方法来研究分析缺失数据下非线性联合均值方差模型的统计推断问题。

基金项目

国家自然青年科学基金项目(11301485)；浙江农林大学校科研发展基金人才启动项目(2013FR079)；浙江农林大学创新创业训练计划(201311006)。

参考文献 (References)

[1] White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Journal of the Econometric Society*, **48**, 817-838.

[2] Andrews Donald, W.K. (1991) Heteroscedasticity and autocorrelation consistent covariance matrix estimation. *Journal of the Econometric Society*, **59**, 817-858.

[3] Smyth, G.K. (2002) An efficient algorithm for REML in heteroscedastic Regression. *Journal of Graphical Statistics*, **11**, 836-847.

[4] Park, R.E. (1966) Estimation with heteroscedastic error terms. *Econometrica*, **34**, 888.

[5] Harvey, A.C. (1976) Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, **44**, 460-465.

[6] Aitkin, M. (1987) Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics*, **36**, 332-339.

[7] Verbyla, A.P. (1993) Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society*: *Series B*, **52**, 493-508.

[8] Wu, L.C. and Li, H.Q. (2012) Variable selection for joint mean and dispersion models of the inverse gaussian distribution. *Metrika*, **75**, 795-808.

[9] Xu, D.K., Zhang, Z.Z. and Wu, L.C. (2014) Variable selection in high-dimensional double generalized linear models. *Statistical Papers*, **55**, 327-347.

[10] 吴刘仓, 黄丽, 戴琳(2012) Box-Cox 变换下联合均值与方差模型的极大似然估计. *统计与信息论坛*, **5**, 3-8.

[11] 马婷, 吴刘仓, 黄丽(2013) 基于偏正态分布联合位置, 尺度与偏度模型的极大似然估计. *数理统计与管理*, **3**, 433-439.

[12] 徐登可, 张忠占, 张松, 张蕾 (2012) 妊娠期高血压疾病危险因素的统计分析. *应用概率统计*, **2**, 134-142.

[13] 贝茨, 沃茨, 著, 韦博成, 等, 译 (1997) 非线性回归分析及其应用. 中国统计出版社, 北京.