Statistics and Application 统计学与应用, 2018, 7(3), 359-365
Published Online June 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.73042

Hans汉斯

Statistical Diagnosis for Mean and Variance Models with Missing Data

Fangyi Zhu, Yu Zheng

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

Received: Jun. 2nd, 2018; accepted: Jun. 22nd, 2018; published: Jun. 29th, 2018

Abstract

A statistical diagnosis method based on the data deletion model is considered for the mean and variance models with response variables random missing. It is mainly based on the regression imputation and random regression imputation and the Gauss-Newton iterative algorithm to give the maximum likelihood estimation of the unknown parameters in the models, and then based on the likelihood distance to carry out the diagnosis and analysis of the abnormal values. Finally, through simulation analysis, the results show that the proposed model and statistical method are feasible and effective.

Keywords

Mean and Variance Models, Data Deletion Model, Imputation, Statistical Diagnosis

缺失数据下均值与方差模型的统计诊断

朱方怡，郑　玉

浙江农林大学统计系，浙江　杭州
Email: 175384319@qq.com

收稿日期：2018年6月2日；录用日期：2018年6月22日；发布日期：2018年6月29日

摘　要

针对响应变量随机缺失下均值与方差模型，考虑了基于数据删除模型的统计诊断方法。其中主要基于回归插补法和随机回归插补法以及结合Gauss-Newton迭代计算算法给出该模型中未知参数的极大似然估计，进而基于似然距离进行异常值诊断分析。最后通过模拟研究分析，结果表明所提出的模型和统计方法是可行有效的。

文章引用: 朱方怡, 郑玉. 缺失数据下均值与方差模型的统计诊断[J]. 统计学与应用, 2018, 7(3): 359-365.
DOI: 10.12677/sa.2018.73042

关键词

均值与方差模型，数据删除模型，插补，统计诊断

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。其中对方差进行建模分析是目前比较流行的处理异方差问题的统计方法，并且也已经有了大量的研究成果。例如，Aitkin [1]基于正态分布提出了联合均值与方差模型的极大似然估计；黄丽等[2]研究了对数正态分布的联合模型的极大似然估计；Taylor和Verbyla [3]在自由度未知的情况下提出了t分布的联合位置与尺度模型的参数估计和检验；Verbyla [4]针对方差建模模型研究了限制极大似然估计和考虑了模型的影响诊断分析；吴刘仓等[5]研究了基于正态分布的联合均值与方差模型的变量选择；具体的其它成果内容还可以参见[6] [7] [8] [9]。

另外，数据缺失是数据分析或研究工作中所遇到的非常普遍的现象。若采用完全数据的统计方法进行统计推断，则所获得的统计分析结果会出现较大的偏差。所以，缺失数据分析一直是统计学者们研究的热点问题。特别是缺失数据下均值回归模型的统计推断，国内外学者都已经有了深入的研究。但是，有关缺失数据下均值与方差联合建模的成果大多数都是集中研究参数估计、变量选择等问题，几乎没人研究过统计诊断问题。因此就有必要在缺失数据下研究均值与方差模型的统计诊断问题。

本文主要针对缺失数据下均值与方差联合模型，通过数据删除模型的参数估计和统计诊断，基于回归插补法和随机回归插补法以及结合Gauss-Newton迭代计算算法给出该模型中未知参数的极大似然估计，比较删除模型与未删除模型相应统计量之间的差异，进而有效的识别异常点。最后通过模拟研究分析，表明本文提出的理论和方法是有用和有效的。

2. 缺失数据下均值与方差模型

针对异方差数据，既对响应变量的均值建模，同时又对响应变量的方差进行建模，因此研究的均值与方差模型如下：

$$\begin{cases} y_i \sim N\left(\mu_i, \sigma_i^2\right) \\ \mu_i = x_i^{\mathrm{T}}\beta \\ \log\sigma_i^2 = z_i^{\mathrm{T}}\lambda \\ i = 1,2,\cdots,n \end{cases} \tag{1}$$

其中 $x_i = \left(x_{i1}, x_{i2}, \cdots, x_{ip}\right)^{\mathrm{T}}$ 和 $z_i = \left(z_{i1}, z_{i2}, \cdots, z_{iq}\right)^{\mathrm{T}}$ 分别为均值模型和方差模型中的解释变量(由于可能存在同时影响均值及方差的变量，因此对于 x_i 和 z_i 而言可完全相同、部分相同或完全不相同)，$\beta = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^{\mathrm{T}}$ 和 $\lambda = \left(\lambda_1, \lambda_2, \cdots, \lambda_q\right)^{\mathrm{T}}$ 是对应的均值及方差模型中的未知参数向量，$y = \left(y_1, y_2, \cdots, y_n\right)^{\mathrm{T}}$ 则是模型的被解释变量或响应变量。

本文主要考虑响应变量数据存在缺失，即假设 $\left(y_i, x_i, z_i, \delta_i\right), i = 1,2,\cdots,n$ 为来自模型(1)的不完全随机样

本，其中解释变量 x_i, z_i 能够完全观测，响应变量 y_i 部分缺失，并假定如果 y_i 缺失，则 $\delta_i = 0$；若 $\delta_i = 1$ 时，则表示 y_i 可以观测。在本文主要假定 y_i 为随机缺失(missing at random, MAR)，即 $P(\delta_i = 1 | y_i, x_i, z_i) = P(\delta_i = 1 | x_i, z_i)$。$y_i$ 在 MAR 机制下，模型(1)可以写成为

$$\begin{cases} \delta_i y_i \sim N\left(\mu_i, \sigma_i^2\right) \\ \mu_i = x_i^{\mathrm{T}} \beta \\ \log\left(\sigma_i^2\right) = z_i^{\mathrm{T}} \lambda \\ i = 1, 2, \cdots, n. \end{cases} \tag{2}$$

为了方便，令 $r = \sum_{i=1}^{n} \delta_i$ 为没有缺失数据(即能观测到)的个体数，$m = n - r$ 表示有缺失数据(即不能观测)的个体数，$s_r = \{i : \delta_i = 1, i = 1, 2, \cdots, n\}$ 表示能观测到的个体的集合，$s_m = \{i : \delta_i = 0, i = 1, 2, \cdots, n\}$ 表示不能观测的个体的集合。

3. 统计诊断分析

3.1. 数据删除模型

数据删除是统计诊断中最基本的方法，比较未删除模型和删除模型相应的参数估计量之间的差异，能得出一定的结论。缺失数据下均值方差模型的数据删除模型可以表示为：

$$\begin{cases} \delta_i y_i \sim N\left(\mu_i, \sigma_i^2\right) \\ \mu_i = x_i^{\mathrm{T}} \beta \\ \log\left(\sigma_i^2\right) = z_i^{\mathrm{T}} \lambda \\ i = 1, 2, \cdots, n, i \neq j \end{cases} \tag{3}$$

对于模型(2)和(3)，为了解第 j 个数据点 $\left(x_j, y_j\right)$ 在整个数据集中的作用和影响，可通过比较第 j 个数据点 $\left(x_j, y_j\right)$ 删除前后统计推断结果的变化，来判断这个点是否是异常点或强影响点，其中，统计推断结果的变化可以通过一些诊断统计量来表述，统计诊断量的具体计算见下文。删除第 j 个数据点之后的模型称之为数据删除模型，本文研究的数据删除模型为模型(3)。

3.2. 完全数据下的极大似然估计

对于模型(1)，假设 $\left(y_i, x_i, z_i\right)$ 为数据集中的第 i 个数据点。由模型(1)可知其似然函数为：

$$l(\beta, \lambda) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \exp\left(z_i^{\mathrm{T}} \lambda\right)}} \exp\left[-\frac{\left(y_i - x_i^{\mathrm{T}} \beta\right)}{2\exp\left(z_i^{\mathrm{T}} \lambda\right)}\right] \tag{4}$$

对式(4)取自然对数，得其对数似然函数为：

$$l(\beta, \lambda) = -\frac{1}{2}\sum_{i=1}^{n} \ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n} z_i^{\mathrm{T}} \lambda - \frac{1}{2}\sum_{i=1}^{n} \frac{\left(y_i - x_i^{\mathrm{T}} \beta\right)}{2\exp\left(z_i^{\mathrm{T}} \lambda\right)} \tag{5}$$

由于极大化该对数似然函数(5)无法获得未知参数的极大似然估计的显示表达式，因此主要采用 Gauss-Newton 迭代计算算法。

为了方便，令 $\theta = \left(\beta^{\mathrm{T}}, \lambda^{\mathrm{T}}\right)^{\mathrm{T}}$，则 $l(\beta, \lambda) = l(\theta)$，因此：

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

其中 $U_1(\beta)=\dfrac{\partial l(\beta,\lambda)}{\partial\beta}$，$U_2(\beta)=\dfrac{\partial l(\beta,\lambda)}{\partial\lambda}$，

$$\frac{\partial l(\beta,\lambda)}{\partial\beta}=\sum_{i=1}^{n}\frac{\left(y_i-x_i^{\mathrm{T}}\beta\right)x_i}{\exp\left(z_i^{\mathrm{T}}\lambda\right)}$$

$$\frac{\partial l(\beta,\lambda)}{\partial\lambda}=-\frac{1}{2}\sum_{i=1}^{n}z_i+\frac{1}{2}\sum_{i=1}^{n}\frac{\left(y_i-x_i^{\mathrm{T}}\beta\right)^2 z_i}{\exp\left(z_i^{\mathrm{T}}\lambda\right)}$$

另外，令 $H(\theta)=\begin{pmatrix}\dfrac{\partial^2 l(\theta)}{\partial\beta\partial\beta^{\mathrm{T}}} & \dfrac{\partial^2 l(\theta)}{\partial\beta\partial\lambda^{\mathrm{T}}}\\ \dfrac{\partial^2 l(\theta)}{\partial\beta^{\mathrm{T}}\partial\lambda} & \dfrac{\partial^2 l(\theta)}{\partial\lambda\partial\lambda^{\mathrm{T}}}\end{pmatrix}$，其中：

$$\frac{\partial^2 l(\beta,\lambda)}{\partial\beta\partial\beta^{\mathrm{T}}}=-\sum_{i=1}^{n}\frac{x_i x_i^{\mathrm{T}}}{\exp\left(z_i^{\mathrm{T}}\lambda\right)},\quad \frac{\partial^2 l(\beta,\lambda)}{\partial\lambda\partial\lambda^{\mathrm{T}}}=-\frac{1}{2}\sum_{i=1}^{n}\frac{\left(y_i-x_i^{\mathrm{T}}\right)^2 z_i z_i^{\mathrm{T}}}{\exp\left(z_i^{\mathrm{T}}\lambda\right)},$$

$$\frac{\partial^2 l(\beta,\lambda)}{\partial\beta\partial\lambda^{\mathrm{T}}}=-\sum_{i=1}^{n}\frac{\left(y_i-x_i^{\mathrm{T}}\right)^2 z_i x_i^{\mathrm{T}}}{\exp\left(z_i^{\mathrm{T}}\lambda\right)},\quad \frac{\partial^2 l(\beta,\lambda)}{\partial\beta^{\mathrm{T}}\partial\lambda}=-\sum_{i=1}^{n}\frac{\left(y_i-x_i^{\mathrm{T}}\right)^2 x_i z_i^{\mathrm{T}}}{\exp\left(z_i^{\mathrm{T}}\lambda\right)}。$$

最后，将以上几个式子带入式(6)进行迭代计算：

$$\theta_1=\theta_0-\left(H(\theta)\right)^{-1}U(\theta)\big|_{\theta=\theta_0}\tag{6}$$

直到 $|\theta_1-\theta_0|<\delta$，即认为 θ_1 为极大似然估计 $\hat{\theta}$ 的近似值，其中 δ 为预先给定的充分小的正数，如 $\delta=10^{-6}$。

最后给出以下获得模型(1)中未知参数的极大似然估计的具体迭代计算算法与步骤。

步骤 1：首先给定未知参数的初始值 $\theta^{(0)}=\left(\left(\beta^{(0)}\right)^{\mathrm{T}},\left(\lambda^{(0)}\right)^{\mathrm{T}}\right)^{\mathrm{T}}$。

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

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

步骤 3：重复步骤 2，直到迭代收敛。

3.3. 基于回归插补法的缺失数据下参数估计

回归插补法的主要思想是依据响应变量 y 和与之对应的解释变量 x 之间的关系建立回归模型，即 $\hat{y}_i=x_i^{\mathrm{T}}\hat{\beta}$。然后根据已知的解释变量 x 的数据信息，对相应的响应变量 y 进行插补。

在缺失数据下均值方差模型(2)中，对 $(y_i,x_i,z_i),i\in s_r$ 根据完全数据下极大似然估计的方法得到均值模型中未知参数的估计 $\hat{\beta}_0$，然后令缺失值 $y_i^*=x_i^{\mathrm{T}}\hat{\beta}_0,i\in s_m$，记补全后的“完全数据”为 $y_{i1}=\delta_i y_i+(1-\delta_i)y_i^*$。接着使用Gauss-Newton迭代计算算法就可获得基于回归插补法缺失数据下均值方差模型中未知参数的极大似然估计值。

3.4. 基于随机回归插补法的缺失数据下参数估计

回归插补是从确定性的角度出发，然后对缺失值进行插补的一种常用的缺失数据处理方法。即插补

是依据某种确定的函数关系获得一个确定的插补值，其表达式可写成 $y_i = f(x_i)$。然而随机回归插补法是指在确定性插补方法的基础上给插补值增加一个随机项，即 $y_i = f(x_i) + e_i$。进而可以把回归插补写成为 $y_i = x_i^{\mathrm{T}}\hat{\beta} + e_i$，这样的回归插补就被称作随机回归插补法。该随机项反应所预测的值的不确定性影响。在正态回归模型中，该随机项是取服从均值为零和方差为回归中剩余方差的正态分布。

在缺失数据下均值方差模型(2)中，对 $(y_i, x_i, z_i), i \in s_r$ 根据完全数据下极大似然估计的方法得到均值模型中未知参数的估计 $\hat{\beta}_0$ 及方差模型中未知参数的估计 $\hat{\gamma}_0$，然后令缺失值 $y_i^* = x_i^{\mathrm{T}}\hat{\beta}_0 + e_i, i \in s_m$，其中 $e_i \sim N\left(0, \exp\left(z_i^{\mathrm{T}}\hat{\gamma}_0\right)\right), i \in s_m$。记补全后的“完全数据”为 $y_{i2} = \delta_i y_i + (1-\delta_i) y_i^*$。接着使用 Gauss-Newton 迭代计算算法就可获得基于随机回归插补法缺失数据下均值方差模型中未知参数的极大似然估计值。

3.5. 基于数据删除模型的统计诊断量

在数据删除模型的基础上，这里将主要介绍诊断统计量似然距离的计算方法和表达式，以及诊断结果的解释说明。

由韦博成等[10]的《统计诊断》可知，对于本文中的未删除模型及其删除模型，第 i 个数据点的似然距离定义为：

$$LD(i) = 2\left\{L(\hat{\theta}) - L(\hat{\theta}(i))\right\} \tag{7}$$

由于 $L(\hat{\theta})$ 为全局最大值，因此恒有 $LD(i) \geq 0$。似然距离 $LD(i)$ 反映了第 i 个数据点 (x_i, y_i) 对参数 θ 的极大似然估计的影响。对于远大于其极大似然距离的点，说明删除该点对参数估计值影响较大，即该点为强影响点或异常点。

针对本文中的模型，似然距离没有显式解，因此需要用近似计算得出其数值解。

根据韦博成等[10]的《统计诊断》，对 $LD(i) = 2\left\{L(\hat{\theta}) - L(\hat{\theta}(i))\right\}$ 在 $\hat{\theta}$ 处进行 Taylor 展开可得：

$$\begin{aligned} LD(i) &= 2\left\{L(\hat{\theta}) - L(\hat{\theta}(i))\right\} = -2\left\{L(\hat{\theta}(i)) - L(\hat{\theta})\right\} \\ &\approx -2\left\{\dot{L}(\hat{\theta})(\hat{\theta}(i) - \hat{\theta}) + \frac{1}{2}(\hat{\theta}(i) - \hat{\theta})^T \ddot{L}(\hat{\theta})(\hat{\theta}(i) - \hat{\theta})\right\} \end{aligned}$$

又因为 $\dot{L}(\hat{\theta}) = 0$，从而可以得到似然距离 $LD(i)$ 的近似表达式为：

$$LD^I(i) = (\hat{\theta} - \hat{\theta}(i))^{\mathrm{T}}\left[-\ddot{L}(\hat{\theta})\right](\hat{\theta} - \hat{\theta}(i))$$

或 $LD^I(i) = (\hat{\theta} - \hat{\theta}(i))^{\mathrm{T}}\left[I(\hat{\theta})\right](\hat{\theta} - \hat{\theta}(i))$

其中 $I(\hat{\theta})$ 为 Fisher 信息阵，应用 Gauss-Newton 迭代法可得到参数的估计值 $\hat{\theta}$ 和 $\hat{\theta}(i)$。为了计算简便，本文主要使用 Fisher 信息阵计算似然距离。

4. 模拟研究

下面用随机模拟方法来验证文中提出的基于似然距离对缺失数据下均值方差模型的统计诊断方法的可行性和有效性。

根据模型(2)，我们产生随机数据。其中 $x_i \sim U(-1,1)$，$z_i \sim U(-1,1), i = 1,2,\cdots,100$。取参数 $\beta = (3,2,1.5)^{\mathrm{T}}$，$\lambda = (1,1,1)^{\mathrm{T}}$，缺失比例大约 20%。然后将被解释变量 Y 的第 20 行的值改为 5，第 40 行的值改为 5，即人为的制造两个异常点，然后根据上述的诊断方法得出结果，验证上述统计量是否有效。模拟结果如图 1 和图 2。

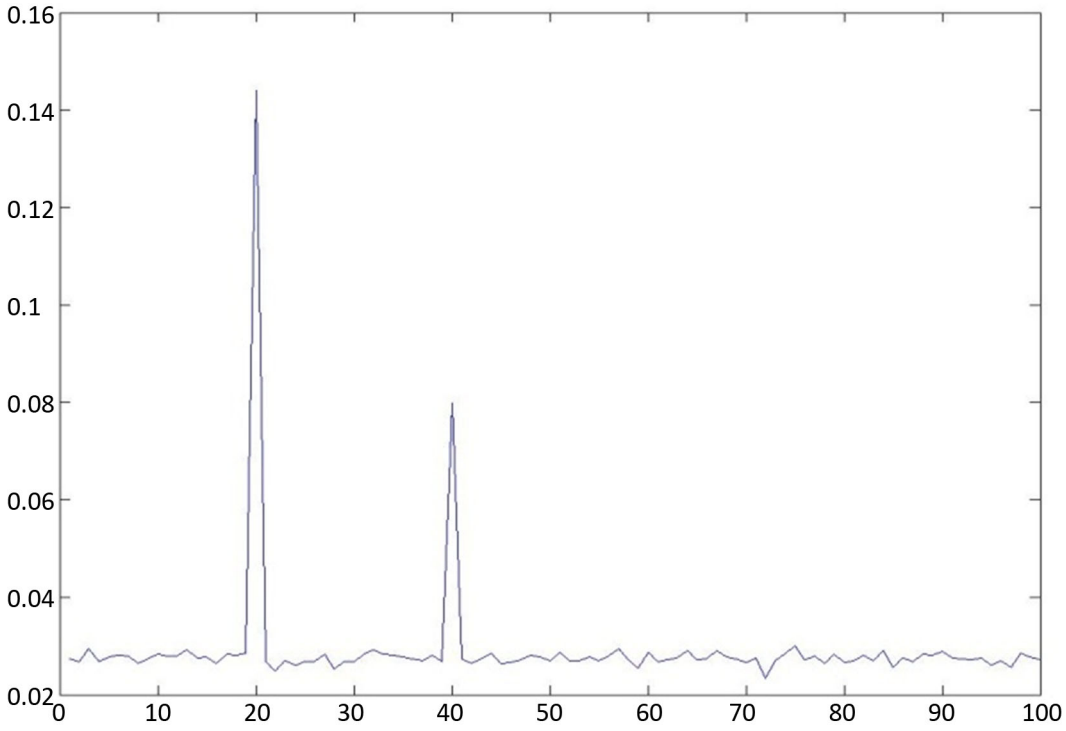

Figure 1. Scatter plot of likelihood distance based on the regression imputation method
图 1. 基于回归插补方法得到的似然距离的散点图

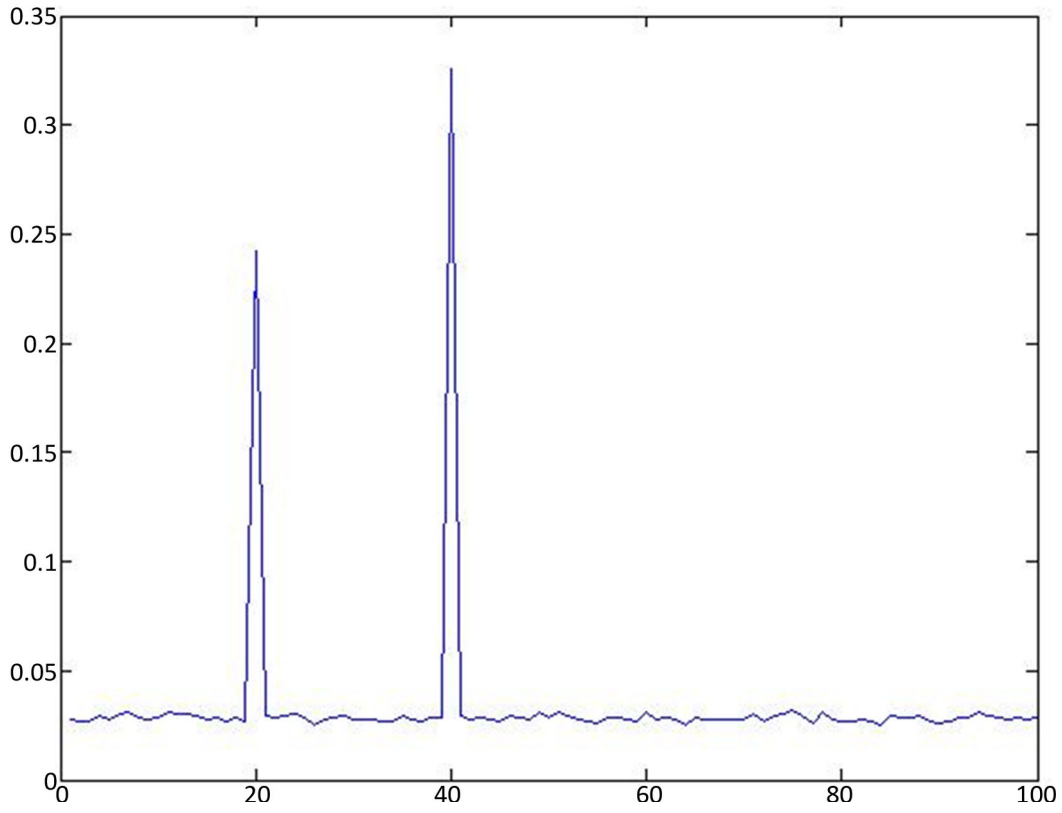

Figure 2. Scatter plot of likelihood distance based on the random regression imputation method
图 2. 基于随机回归插补方法得到的似然距离的散点图

从图 1 和图 2 可以观察到，第 20 和 40 号点为异常点或强影响点。我们人为制造的两个异常点，第 20 和第 40 号点均被诊断出来，说明本文提出的方法是行之有效的。

5. 结论

本文基于数据缺失情况下，运用回归插补和随机回归插补两种缺失插补方法，对均值和方差联合模型进行了统计诊断的研究。并介绍了有关获取未知参数极大似然估计中所使用的 Gauss-Newton 迭代算法以及两种缺失插补方法的主要思想。模拟研究结果表明，我们的方法能对数据缺失情况下的联合均值与方差模型进行统计诊断，且该研究模型与诊断方法具有可行性及有效性。

基金项目

浙江农林大学创新创业训练计划(110-2013200040)。

参考文献

[1] Aitkin, M. (1987) Modelling Variance Heterogeneity in Normal Regression Using GLIM. *Applied Statistics*, **36**, 332-339. https://doi.org/10.2307/2347792

[2] 黄丽, 吴刘仓. 基于对数正态分布下联合均值与散度广义线性模型的极大似估计[J]. 高校应用数学学报, 2011, 26(4): 379-389.

[3] Taylor, J.T. and Verbyla, A.P. (2004) Joint Modelling of Location and Scale Parameters of the t Distribution. *Statistical Modelling*, **4**, 91-112. https://doi.org/10.1191/1471082X04st068oa

[4] Verbyla, A.P. (1993) Modelling Variance Heterogeneity: Residual Maximum Likelihood and Diagnostics. *Journal of the Royal Statistical Society*: *Series B*, **52**, 493-508.

[5] 吴刘仓, 张忠占, 徐登可. 联合均值与方差模型的变量选择[J]. 系统工程理论实践, 2012(8): 1754-1760.

[6] 戴琳, 陶冶, 吴刘仓. 联合均值与方差模型的统计诊断[J]. 统计与信息论坛, 2017, 32(1): 14-19.

[7] 宋红凤, 汤杨冰, 徐登可. 缺失数据下非线性均值方差模型的参数估计[J]. 统计与决策, 2017(19): 10-14.

[8] 王子豪, 吴刘仓, 詹金龙. 联合均值与方差模型的经验似然推断[J]. 统计与决策, 2015(20): 8-10.

[9] 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. https://doi.org/10.1007/s00184-011-0352-x

[10] 韦博成, 林金官, 解锋昌. 统计诊断[M]. 北京: 高等教育出版社, 2009.