应用概率统计　第三十卷
第三期　2014年6月

Chinese Journal of Applied Probability
and Statistics Vol.30 No.3 Jun. 2014

生长曲线模型的分位数回归 *

张　雨　　刘　倩　　曾林蕊*

(四川大学数学学院, 成都, 610065)　　(华东师范大学金融与统计学院, 上海, 200241)

摘　要

生长曲线模型有着广泛的应用, 在经济学、生物学、医学等各个领域的研究都起着重要的作用. 已有文献关于生长曲线模型参数矩阵的估计基本上是使用最小二乘方法或极大似然方法. 使用最小二乘方法, 当误差项服从偏峰分布、厚尾分布、或者存在异常点时, 得出的估计不是有效的; 使用极大似然方法, 要求分布已知, 实际使用时很难满足这一点. 分位数回归能弥补如上这些缺陷, 所得估计具有很好的稳健性. 本文使用分位数回归方法给出生长曲线模型参数矩阵的估计, 及其渐近正态性.

关键词:　生长曲线模型, 分位数回归, 渐近正态性.

学科分类号:　O212.1.

§1. 引　　言

生长曲线模型是生命科学中最重要的统计模型之一, 它主要用于短时间内生长问题的研究, 比如植物生长过程中, 以大小、重量、数量及这些特征在时间上的变化来描述物种群体、个体、器官的生长过程; 儿童成长发育过程中, 以身高、体重、胸围、坐高等随时间的变化为特征来描述儿童的生长发育状况. 随着研究者的不断探索, 该模型的理论内涵得到了不断的丰富, 应用背景得到了不断的推广, 在经济学、生物学、医学等领域的研究都起到了至关重要的作用.

生长曲线模型是Wishart (1938)在研究不同组间动植物的生长情况时引入的模型. Potthoff和Roy (1964)给出了该模型的一个非常详细的应用背景和一些研究. 假设n只小白鼠共有r个品种, 以品种把这n只小白鼠分为r个小组, 第i个小组有n_i只小白鼠, 记录在$t_1, t_2, \ldots, t_p$　p个时刻点这n只小白鼠的体重, 则第i个小组的小白鼠的体重有一条生长曲线$(i = 1, \cdots, r)$, 即

$$y_{ijl} = \beta_{i0} + \beta_{i1}t_l + \beta_{i2}t_l^2 + \cdots + \beta_{i,m-1}t_l^{m-1} + \varepsilon_{ijl}, \qquad l = 1, \ldots, p,\ j = 1, \ldots, n_i,$$

其中, y_{ijl}表示第i组中第j个小白鼠在t_l时刻的体重, $\beta_i = (\beta_{i0}, \ldots, \beta_{i,m-1})^T$为第$i$组小白鼠多项式生长曲线模型的回归系数, ε_{ijl}表示第i组中第j个小白鼠在t_l时刻的测量误差, m是多项式回归模型的阶数.

*2013年四川大学“大学生创新训练计划”项目资助.

*通讯作者, E-mail: lrzeng@stat.ecnu.edu.cn.

本文2013年11月7日收到, 2014年3月11日收到修改稿.
doi: 10.3969/j.issn.1001-4268.2014.03.007

记$Y_{p\times n}=(Y_1,Y_2,\ldots,Y_r)$, $Y_i=(y_{ijl})_{p\times n_i}$, $\varepsilon_{p\times n}=(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_r)$, $\varepsilon_i=(\varepsilon_{ijl})_{p\times n_i}$, $i=1,2,\ldots,r$, 1_n为元素全是1的$n\times 1$的列向量, $B=(\beta_1,\ldots,\beta_r)$,

$$X=\begin{pmatrix}1 & t_1 & \cdots & t_1^{m-1}\\ 1 & t_2 & \cdots & t_2^{m-1}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & t_p & \cdots & t_p^{m-1}\end{pmatrix},\qquad Z=\begin{pmatrix}1_{n_1}^T & 0 & \cdots & 0\\ 0 & 1_{n_2}^T & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1_{n_r}^T\end{pmatrix},$$

则生长曲线模型的矩阵形式为

$$\begin{cases}Y_{p\times n}=X_{p\times m}B_{m\times r}Z_{r\times n}+\varepsilon_{p\times n},\\ \mathsf{E}(\varepsilon)=0,\quad \mathsf{Cov}\left(\mathrm{vec}(\varepsilon)\right)=I_n\otimes\Sigma,\end{cases}$$

称Y是$p\times n$阶观察矩阵, X,Z分别是$p\times m$, $r\times n$阶的设计矩阵, B是$m\times r$阶的参数矩阵, ε是$p\times n$阶的误差矩阵, 且$\mathrm{Rank}(X)=m$, $\mathrm{Rank}(Z)=r$, Σ是$p\times p$的正定矩阵, $\mathrm{vec}(\varepsilon)$表示把矩阵$\varepsilon$按列拉直, $\otimes$表示Kronecker乘积.

多年来, 许多统计学家对此模型做了大量的研究, 如: 潘建新(1988)研究了生长曲线模型中回归参数矩阵的最小二乘估计; 张日权(2000)研究了PC准则下生长曲线模型回归参数矩阵岭估计的优良性; 还有许多研究文献, 文献[4]、[5]、[6]、[7]和[18]研究了该模型的BLU估计、综合邻估计、协方差估计等等; 文献[3]讨论了该模型的非参数估计, 这里不一一列举. Pan和Fang (2002, 2007)对该模型的估计及其诊断做了比较全面的总结和介绍. 已有文献关于生长曲线模型参数矩阵的估计基本上是使用最小二乘方法或极大似然方法. 最小二乘方法刻画的仅仅是响应变量的均值问题, 很多时候不能全面地刻画自变量与响应变量之间的关系, 特别是, 当误差项服从偏峰分布, 厚尾分布, 或者存在异常点时, 最小二乘方法得出的估计就会失效. 极大似然方法要求分布已知, 实际使用时很难满足这一点.

Koenker和Bassett (1978)首次提出了分位数回归的思想. 近几十年来, 分位数回归的方法在各个领域中得到了迅猛的发展, 比如, 环境科学方面有关城市日死亡率与空气污染集中度的相互关系研究; 生态学方面有关不同河流对鲜鱼密度的影响研究; 金融方面有关风险价值和共同基金的投资类型研究; 在经济学中的应用研究包括教育回报, 财富分配不均, 失业持续时间, 酒精使用需求以及日间用电需求等等问题.

近年来, 有关分位数的理论研究很多学者都作出了非常大的贡献. 比如, Bassett 和 Koenker (1978)推导出分位数回归系数估计的渐近性质; Powell (1986)基于删失模型提出了非线性分位数回归; Koenker和Xiao (2002)解决了分位数回归过程中存在的特定推断问题; Koenker (2005)详细研究讨论总结了有关分位数回归的理论问题, 同时还提出了一些有待解决的问题; Koenker (2004)讨论了纵向数据的分位数回归理论及其渐近性质.

一般而言, 为了刻画自变量对响应变量某个分位数, 进而整个分布的影响, 同时也为了降低异常值对模型估计的影响, 统计学家常常会考虑使用分位数回归方法. 这是由于分位

数回归所得估计具有很好的稳健性, 不受异常值的影响, 同时还具有良好的大样本性质. 本文通过分位数回归方法, 研究生长曲线模型的估计, 给出该模型参数矩阵的估计, 及其渐近正态性.

§2. 生长曲线模型的分位数回归

由上节可知, 生长曲线的分位数回归模型为

$$Y = XB(\tau)Z + \varepsilon_\tau,$$

其中, Y, X, Z如前定义, $B(\tau)$是每个元素都与分位数τ $(0 < \tau < 1)$有关的参数矩阵, ε的τ分位数为0.

定义生长曲线模型的分位数回归估计

$$\widehat{\beta}(\tau) = \arg\min \sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r} \rho_\tau(y_{ijl} - t_l^T\beta_i(\tau)),$$

其中, 损失函数$\rho_\tau(u) = u(\tau - I(u < 0))$, 当$u < 0$时, $I(u < 0) = 1$, 否则, 为0; $t_l = (1, t_l, \cdots, t_l^{m-1})^T$, $\beta_i(\tau) = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{i,m-1})^T$.

不难看出, 这样的损失函数, 降低了异常值对模型估计的影响, 所得估计具有很好的稳健性, 不受异常值的影响. 对于不同的τ, 得到不同的分位数回归参数估计, 也就是说随着τ的变动, 能得到一簇分位数回归模型, 刻画了自变量对响应变量某个分位数, 进而整个分布的影响, 不像均值回归只能得到一个模型.

下面讨论估计的大样本性质. 如上目标函数为

$$\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r} \rho_\tau(y_{ijl} - t_l^T\beta_i(\tau)).$$

假设$B^0(\tau) = (\beta_1^0(\tau), \ldots, \beta_r^0(\tau))$为真值, 记$\delta = (\delta_1, \ldots, \delta_r) = \sqrt{np}(B(\tau) - B^0(\tau))$, 则

$$\frac{\delta}{\sqrt{np}} + B^0(\tau) = B(\tau).$$

代入上述目标函数, 有

$$\begin{aligned}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r} \rho_\tau(y_{ijl} - t_l^T\beta_i(\tau)) &= \sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r} \rho_\tau\Big(y_{ijl} - t_l^T\Big(\frac{\delta_i}{\sqrt{np}} + \beta_i^0(\tau)\Big)\Big)\\ &= \sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r} \rho_\tau\Big(y_{ijl} - t_l^T\beta_i^0(\tau) - t_l^T\frac{\delta_i}{\sqrt{np}}\Big).\end{aligned}$$

故而可设目标函数为

$$Q_{np}(\delta) = \sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r} \rho_\tau\Big(y_{ijl} - t_l^T\beta_i^0(\tau) - \frac{t_l^T\delta_i}{\sqrt{np}}\Big) - \rho_\tau(y_{ijl} - t_l^T\beta_i^0(\tau)).$$

则$\widehat{\delta}=\sqrt{np}(\widehat{B}(\tau)-B^0(\tau))$为使得目标函数$Q_{np}(\delta)$达到最小值的解, 即

$$\widehat{\delta}=\arg\min Q_{np}(\delta).$$

记$S=Z^T\otimes X=(t_{ij})_{np\times mr}$, β为系数矩阵B的按列拉直向量. 如下的假设, 在定理中是需要的. 假设:

条件 2.1 假设y_{ijl}的密度函数$0<f_{ijl}(\cdot)<\infty$, 且$f'_{ijl}(\cdot)$在$t_l^T\beta_i(\tau)$处有界, $i=1,2,\ldots,r$, $j=1,2,\ldots,n_i$, $l=1,2,\ldots,p$.

条件 2.2 记$\Phi=\mathrm{diag}(f_{ijl}(t_l^T\beta_i(\tau)))$,

$$D_0=\lim_{n,p\to\infty}\frac{\tau(1-\tau)}{np}S^TS,\qquad D_1=\lim_{n,p\to\infty}\frac{1}{np}S^T\Phi S,$$

且D_0和D_1是正定矩阵.

条件 2.3 $\max\|t_{ij}\|<M$.

定理 2.1 若满足条件2.1–2.3, 则

$$\sqrt{np}(\widehat{B}(\tau)-B^0(\tau))\xrightarrow{\mathrm{d}}N(0,D_1^{-1}D_0D_1^{-1}).$$

证明: 由Night等式

$$\rho_\tau(u-v)-\rho_\tau(u)=-v\psi_\tau(u)+\int_0^v[I(u\le s)-I(u\le 0)]\mathrm{d}s,$$

其中, $\psi_\tau(u)=\tau-I(u<0)$. 并记$v_{il}=t_l^T\delta_i/\sqrt{n}$, 可得

$$\begin{aligned}
Q_{np}(\delta) &= \sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}\rho_\tau\Big(y_{ijl}-t_l^T\beta_i^0(\tau)-\frac{t_l^T\delta_i}{\sqrt{np}}\Big)-\rho_\tau(y_{ijl}-t_l^T\beta_i^0(\tau))\\
&= \sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}-\frac{t_l^T\delta_i}{\sqrt{np}}\psi_\tau(y_{ijl}-t_l^T\beta_i^0(\tau))\\
&\quad+\int_0^{t_l^T\delta_i/\sqrt{np}}[I(y_{ijl}\le t_l^T\beta_i^0(\tau)+s)-I(y_{ijl}\le t_l^T\beta_i^0(\tau))]\mathrm{d}s\\
&= -\frac{1}{\sqrt{p}}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}\Big[\frac{t_l^T\delta_i}{\sqrt{n}}\psi_\tau(y_{ijl}-t_l^T\beta_i^0(\tau))\\
&\quad-\sqrt{p}\int_0^{t_l^T\delta_i/\sqrt{np}}[I(y_{ijl}\le t_l^T\beta_i^0(\tau)+s)-I(y_{ijl}\le t_l^T\beta_i^0(\tau))]\mathrm{d}s\Big]\\
&= -\frac{1}{\sqrt{p}}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}v_{il}\psi_\tau(y_{ijl}-t_l^T\beta_i^0(\tau))\\
&\quad+\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}\int_0^{t_l^T\delta_i/\sqrt{np}}[I(y_{ijl}\le t_l^T\beta_i^0(\tau)+s)-I(y_{ijl}\le t_l^T\beta_i^0(\tau))]\mathrm{d}s
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{\sqrt{p}}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r} v_{il}\psi_\tau(y_{ijl}-t_l^T\beta_i^0(\tau)) \\
&\quad +\frac{1}{\sqrt{p}}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}\int_0^{v_{il}}\Big[I\Big(y_{ijl}\le t_l^T\beta_i^0(\tau)+\frac{t}{\sqrt{p}}\Big)-I(y_{ijl}\le t_l^T\beta_i^0(\tau))\Big]\mathrm{d}t \\
&\widehat{=} Q_{np}^{(1)}(\delta)+Q_{np}^{(2)}(\delta),
\end{aligned}
$$

其中,

$$
\begin{aligned}
Q_{np}^{(1)}(\delta) &= -\frac{1}{\sqrt{p}}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r} v_{il}\psi_\tau(y_{ijl}-t_l^T\beta_i^0(\tau)), \\
Q_{np}^{(2)}(\delta) &= \frac{1}{\sqrt{p}}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}\int_0^{v_{il}}\Big[I\Big(y_{ijl}\le t_l^T\beta_i^0(\tau)+\frac{t}{\sqrt{p}}\Big)-I(y_{ijl}\le t_l^T\beta_i^0(\tau))\Big]\mathrm{d}t.
\end{aligned}
$$

记$\Omega=\mathrm{diag}(\psi_\tau(y_{ijl}-t_l^T\beta_i(\tau)))$, $C^T=1_{np}^T\Omega S/\sqrt{n}$, 则

$$
Q_{np}^{(1)}(\delta)=-\frac{1}{\sqrt{p}}\frac{1_{np}^T\Omega S}{\sqrt{n}}\delta=-\frac{1}{\sqrt{p}}C^T\delta.
$$

由于

$$
\begin{aligned}
&\mathsf{E}(Q_{np}^{(1)}(\delta))=-\frac{1}{\sqrt{p}}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r} v_{il}\mathsf{E}(\psi_\tau(y_{ijl}-t_l^T\beta_i^0(\tau)))=0, \\
&\mathsf{Var}\,(C)=\frac{1}{n}\mathsf{Var}\,(S^T\Omega 1_{np})=\frac{1}{n}S^T\mathsf{Var}\,(\Omega 1_{np})S=\frac{\tau(1-\tau)}{n}S^TS,
\end{aligned}
$$

所以,

$$
C_0\widehat{=}-\frac{1}{\sqrt{p}}C\xrightarrow{\mathrm{d}}N(0,D_0).
$$

下面讨论第二部分, 因为

$$
\begin{aligned}
\mathsf{E}(Q_{np}^{(2)}(\delta)) &= \frac{1}{\sqrt{p}}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}\int_0^{v_{il}}F_{ijl}\Big(t_l^T\beta_i^0(\tau)+\frac{t}{\sqrt{p}}\Big)-F_{ijl}(t_l^T\beta_i^0(\tau))\mathrm{d}t \\
&= \frac{1}{\sqrt{p}}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}\int_0^{v_{il}}f_{ijl}(t_l^T\beta_i^0(\tau))\frac{t}{\sqrt{p}}\mathrm{d}t \\
&= \frac{1}{p}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}f_{ijl}(t_l^T\beta_i^0(\tau))\frac{v_{il}^2}{2} \\
&= \frac{1}{2p}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}f_{ijl}(t_l^T\beta_i^0(\tau))\frac{v_{il}^2}{2}\Big(\frac{t_l^T\delta}{\sqrt{n}}\Big)^T\Big(\frac{t_l^T\delta}{\sqrt{n}}\Big) \\
&= \frac{1}{2np}\sum_{l=1}^{p}\sum_{j=1}^{n_i}\sum_{i=1}^{r}f_{ijl}(t_l^T\beta_i^0(\tau))\delta^Tt_lt_l^T\delta \\
&= \frac{1}{2np}\delta^TS^T\Phi S\delta.
\end{aligned}
$$

由条件2.3可知, $\mathsf{Var}\,(Q_{np}^{(2)}(\delta)) \to 0$, 故

$$Q_{np}^{(2)}(\delta) \xrightarrow{\mathrm{p}} \frac{1}{2}\delta^T D_1 \delta.$$

令$Q_0(\delta) = -\delta^T C_0 + (1/2)\delta^T D_1 \delta$, $\widehat{\delta}_0 = \arg\min Q_0(\delta)$, 由于

$$Q_{np}(\delta) - Q_0(\delta) \xrightarrow{\mathrm{p}} 0,$$

故

$$\widehat{\delta} - \widehat{\delta}_0 \xrightarrow{\mathrm{p}} 0.$$

由于

$$\widehat{\delta}_0 = D_1^{-1} C_0 \xrightarrow{\mathrm{d}} N(0, D_1^{-1} D_0 D_1^{-1}),$$

故

$$\widehat{\delta} \xrightarrow{\mathrm{d}} N(0, D_1^{-1} D_0 D_1^{-1}). \qquad \square$$

参 考 文 献

[1] 潘建新, 增长曲线模型中回归系数的最小二乘估计及Gauss-Markov定理, 数理统计与应用概率, **3(2)** (1988), 169–185.

[2] 张日权, PC准则下生长曲线模型回归参数阵岭估计的优良性, 工程数学学报, **17(1)**(2000), 113–116.

[3] 高采文, 甘华来, 增长曲线模型的非参数估计, 应用概率统计, **29(6)**(2013), 655–665.

[4] 徐承彝, 杨文礼, 蒋文江, 增长曲线模型中的参数估计, 北京师范大学出版社, 1996.

[5] 梁小林, 生长曲线模型的条件BLU估计, 数学理论与应用, **22(3)**(2002), 47–50.

[6] 严国义, 刘贤龙, 桂咏新, 含有随机效应的增长曲线模型协差阵的最小二乘估计, 华中师范大学学报(自然科学版),**38(4)**(2004), 401–414.

[7] 刘乐平, 生长曲线模型的综合岭估计, 华东师范大学学报(自然科学版), **3**(1999), 27–32.

[8] Wishart, J., Growth-rate determinations in nutrition studies with the bacon pig, and their analysis, *Biometrika*, **30(1-2)**(1938), 16–28.

[9] Potthoff, R.F. and Roy, S.N., A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika*, **51(3-4)**(1964), 313–326.

[10] Pan, J.X. and Fang, K.T., *Growth Curve Models and Statistical Diagnostics*, New York: Springer, 2002.

[11] Pan, J.X. and Fang, K.T., *Growth Curve Models and Statistical Diagnostics – Mathematics Monograph Series 8*, Beijing: Science Press, 2007.

[12] Koenker, R. and Bassett, G., Regression quantiles, *Econometrica*, **46(1)**(1978), 33–50.

[13] Bassett, G. and Koenker, R., Asymptotic theory of least absolute error regression, *Journal of the American Statistical Association*, **73(363)**(1978), 618–622.

[14] Koenker, R., *Quantile Regression*, Cambridge University Press, 2005.

[15] Powell, J.L., Censored regression quantiles, *Journal of Econometrics*, **32(1)**(1986), 143–155.

[16] Koenker, R. and Xiao, Z., Inference on the quantile regression process, *Econometrica*,**70(4)**(2002), 1583–1612.

[17] Koenker, R., Quantile regression for longitudinal data, *Journal of Multivariate Analysis*, **91(1)**(2004), 74–89.

[18] Wong, C.S. and Cheng, H., Estimation in a growth curve model with singular covariance, *Journal of Statistical Planning and Inference*, **97(2)**(2001), 323–342.

Quantile Regression for Growth Curve Model

Zhang Yu

(*College of Mathematics, Sichuan University, Chengdu, 610065*)

Liu Qian Zeng Linrui

(*School of Finance and Statistics, East China Normal University, Shanghai, 200241*)

Growth curve model has broad application background, and plays an important role in some fields such as economics, biology, medical research. Many of existing estimation of its parameter matrix have been obtained based on the least squares method or maximum likelihood method. When distribution of the error term is partial peak, or heavy tail, or there exist outliers, estimation obtained by least square method will be invalid. The distribution of the error must be known in maximum likelihood estimation, which is often not satisfied. Quantile regression method can compensate for these defects and the estimation has good robustness. In this paper, quantile regression is used to give the estimation of growth curve model, and its asymptotic normality.

Keywords: Growth curve model, quantile regression, asymptotic normality.

AMS Subject Classification: 62J99.