



基于单纯形分布的比例数据的回归分析

张朋^{①*}, 邱振国^②

① 浙江大学数学系, 杭州 310027;

② CancerControl, Alberta Health Services, Edmonton, Alberta, T5J 3H1, Canada

E-mail: pengz@zju.edu.cn, zhenguo.qiu@albertahealthservices.ca

收稿日期: 2012-07-21; 接受日期: 2013-01-07; * 通信作者

摘要 很多应用领域中的实验结果都表达成连续比例型数据, 这类数据通常度量成为百分比、比率或比例, 并取值于单位区间. 为了采用弥散模型中的单纯形分布来模拟此类实验结果, 本文首先研究单纯形分布的部分重要性质, 在回归分析中参数估计和统计推断需要运用这些性质. 模拟研究表明, 当所研究情形不满足分布假设时, 单纯形回归模型比 Beta 回归和分对数 - 正态回归模型更为稳健. 通过对体外造血干细胞移植技术的真实数据分析, 本文阐释这种方法和它针对异常值的稳健性. 在 R 软件中, 单纯形回归可以由程序包 “simplexreg” 实现, 读者可以自行下载这个程序包, 地址为 <http://my.zju.edu.cn/share/2466293> (验证码: 7919).

关键词 Beta 回归 弥散模型 广义线性模型 分对数 - 正态回归 随机数生成

MSC (2010) 主题分类 62J12, 62E15

1 引言

广义线性模型 (GLM) [1] 理论表明, 在回归分析中需要合理识别响应变量的类型. 尽管在实际应用中人们常使用正态分布, 但 Jørgensen [2] 指出, 除非数据弥散度很小, 服从正态分布的数据通常不是主流, 而是特例. Fisher [3] 也强调过按类型自然描述数据的重要性. 如果有相关信息, 分析非正态形数据的时候需要考虑其数据类型.

Nelder 和 Wedderburn [4] 通过引入广义线性模型, 首次提出, 许多类型的非正态形数据可以通过一种统一的方法来分析. 广义线性模型起初用于指数分布族, 其核心思想后来推广到一类更广泛的模型, 即弥散模型 [2]. 可再生弥散模型 $DM(\mu, \sigma^2)$ (其中, μ 为位置参数, σ^2 为散度参数) 是一族分布函数, 它的概率密度函数在合适的测度下可以写成

$$a(y; \sigma^2) \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\}, \quad y \in C,$$

其中, $a \geq 0$ 是一个合适的标准化函数, d 是定义在 $C \times \Omega$ 上的单位偏差函数, $\mu \in \Omega$, $\sigma^2 > 0$. 弥散模型的主要贡献是, 均值和偏差的表示相应地推广为位置和偏差, 所以可以处理更多类型的数据. 同时, 方差分析中的残差平方和被推广的偏差函数, 使得偏差分析在模型拟合和模型选择中成为一个一般性的推断工具.

英文引用格式: Zhang P, Qiu Z G. Regression analysis of proportional data using simplex distribution (in Chinese). Sci Sin Math, 2014, 44: 89–104, doi: 10.1360/012013-200

这类弥散模型包含了一系列非正态分布, 包括描述正值数据、非负值数据、计数数据、二项数据、方向数据和比例数据的分布. 弥散模型中很重要的一类是指数弥散模型, 也就是现在熟悉的广义线性模型. 当散度参数固定时, 一个指数弥散模型正好是一个指数族分布. 单纯形分布^[2]也是一种弥散模型. 通过变动位置和散度参数, 单纯形分布包含了一大类定义在 $(0, 1)$ 区间上的分布. 如果散度较大则会出现两个极值点, 这是因为单纯形模型本质上是一种混合模型.

当单纯形分布用于广义线性模型中的随机机制时, 所得到的回归模型可以有效用于连续比例数据的分析, 而连续比例常出现在应用领域中. 由于单纯形模型实际上属于弥散模型, 它与指数弥散模型一样具有很多共同的解析性质. 所以, 针对连续比例数据的广义线性模型可以在经典广义线性模型^[6]的基础上得出.

由于分对数 - 正态分布 (LN)^[7] 的均值和方差没有解析解, 我们对比例作分对数变换^[8], 得到线性模型. Beta 分布是另一种在 Bayes 统计学中针对二项比例的常用先验分布. 通过变换 Beta 分布中的参数, Ferrari 和 Cribari-Neto^[9] 提出了针对比率或比例的 Beta 回归, 并用 R 语言程序包 “betareg”^[10] 实现了相关的参数估计和统计推断.

在以往文献中, 单纯形边际模型^[6, 11] 和单纯形混合效应模型^[12] 已被深入讨论. 本文讨论单纯形分布及其广义线性模型以及相关的计算问题. 第 2 节介绍单纯形分布及其基本性质, 如随机数生成法; 第 3 节介绍基于单纯形分布的广义线性模型及其统计推断和计算; 第 4 节报告模拟研究结果, 在违反分布假设情形下, 比较单纯形广义线性模型、Beta 回归模型和分对数变换线性模型的稳健性; 第 5 节分析研究干细胞低温保存后回复率的真实数据, 以此阐释上述方法; 第 6 节通过讨论得出结论. 技术细节详见附录. 我们通过 R 语言程序包 “simplexreg” 实现本文中讨论的单纯形广义线性回归, 读者可以自行下载这个程序包, 地址为 <http://my.zju.edu.cn/share/2466293> (验证码: 7919).

2 单纯形分布

单纯形分布可由广义逆 Gauss 分布^[5] 导出. 考虑由下式定义的重正化鞍点近似 (当 $\sigma^2 \rightarrow 0$ 时):

$$p(y; \alpha_1, \alpha_2, \mu, \sigma^2) = c(\alpha_1, \alpha_2, \mu, \sigma^2) y^{\alpha_1-1} (1-y)^{\alpha_2-1} \exp \left\{ -\frac{1}{2\sigma^2} d_{\alpha_1, \alpha_2}(y; \mu) \right\}, \quad y \in (0, 1),$$

其中, 单位偏差函数 $d_{\alpha_1, \alpha_2}(y; \mu)$ 定义为

$$d_{\alpha_1, \alpha_2}(y; \mu) = \mu^{2\alpha_1-1} (1-\mu)^{2\alpha_2-1} \frac{(y-\mu)^2}{y(1-y)},$$

$c(\alpha_1, \alpha_2, \mu, \sigma^2)$ 是满足

$$\int_0^1 p(y; \alpha_1, \alpha_2, \mu, \sigma^2) dy = 1$$

的规格化项, 分布中的参数 $(\alpha_1, \alpha_2)^T \in \mathbb{R}^2$, $\sigma^2 > 0$, $\mu \in (0, 1)$, 这类分布被称为广义单纯形分布, 记为 $Y \sim S(\alpha_1, \alpha_2, \mu, \sigma^2)$. 显然, 在 $\sigma^2 \rightarrow \infty$ 的极端情形下, 上述分布就是参数为 (α_1, α_2) 的 Beta 分布^[2]. 令 $(\alpha_1, \alpha_2) = (-1/2, -1/2)$, 我们得到标准单纯形分布, 本文中称为单纯形分布.

均值参数 $\mu \in (0, 1)$ 和散度参数 $\sigma^2 > 0$ 的单纯形分布具有类似于正态分布的密度函数:

$$f(y; \mu, \sigma^2) = [2\pi\sigma^2 \{y(1-y)\}^3]^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\}, \quad y \in (0, 1), \quad (2.1)$$

其中, 单位偏差函数为

$$d(y; \mu) = \frac{(y - \mu)^2}{y(1 - y)\mu^2(1 - \mu)^2},$$

单位方差函数为 $V(\mu) = \mu^3(1 - \mu)^3$.

$Y \sim S^-(\mu, \sigma^2)$ 表示随机变量 Y 服从均值 (位置参数) 为 μ 散度为 σ^2 的单纯形分布. Jørgensen [2] 给出了 Y 的方差 τ^2 :

$$\tau^2 = \mu(1 - \mu) - \frac{1}{\sqrt{2\sigma^2}} \exp \left\{ \frac{1}{2\sigma^2\mu^2(1 - \mu)^2} \right\} \Gamma \left\{ \frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1 - \mu)^2} \right\}, \quad (2.2)$$

其中,

$$\Gamma(a, b) = \int_b^\infty t^{a-1} e^{-t} dt$$

是不完备 Γ 函数.

图 1 展示了具有相同均值和方差的单纯形分布、Beta 分布、和分对数 - 正态分布的密度函数. 直观上看, 单纯形密度函数压缩正态密度的两端到单位区间的边界点 0 和 1. 当散度参数 σ^2 很大时, 单纯形分布具有两个极大值点. 在单位区间的两端, 单纯形分布和分对数 - 正态分布的密度趋向于 0, 但是 Beta 分布的密度值很大. 当均值 μ 接近 0 或 1 时, Beta 分布在单位区间两端的密度值是无穷大.

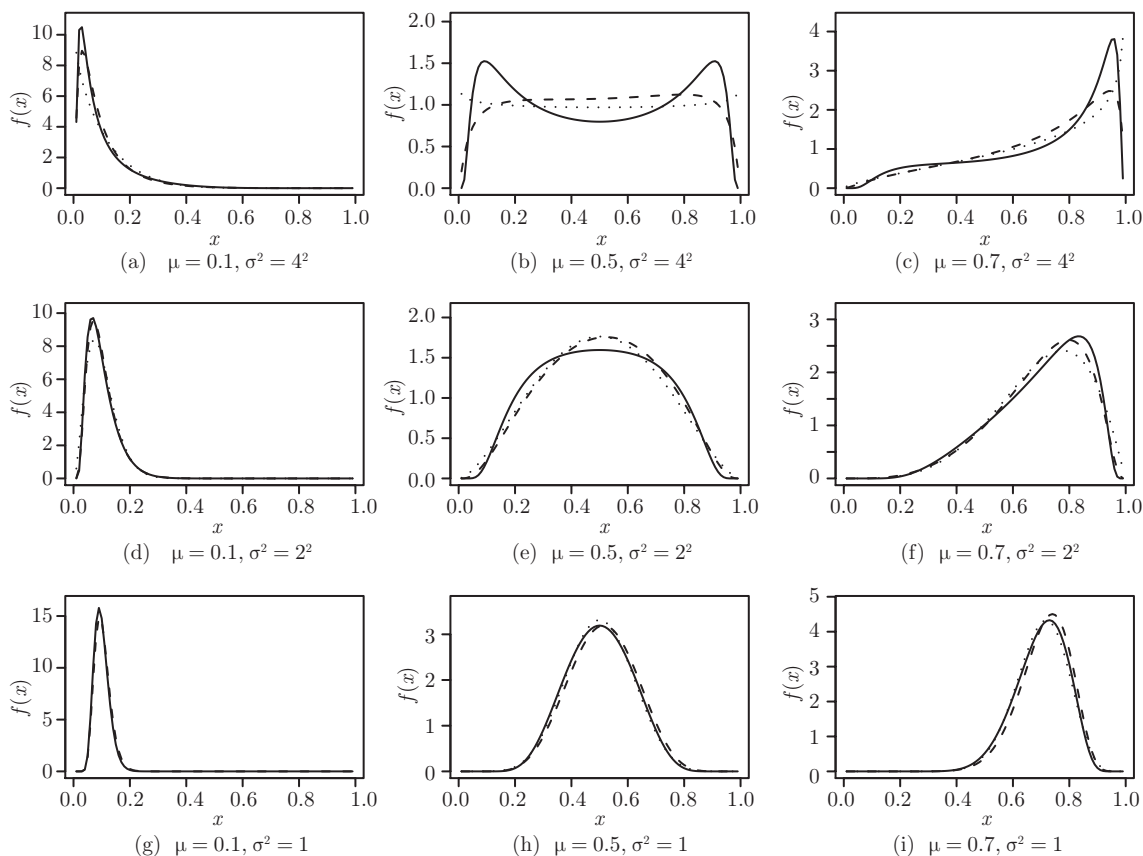


图 1 有相同期望和方差的单纯形密度、Beta 密度和分对数 - 正态密度, 其中, 实线: 单纯形密度, 点线: Beta 密度, 虚线: 分对数 - 正态密度

当 σ^2 较小时, 单纯形分布、Beta 分布和分对数 - 正态分布是类似的. 特别地, 当 $\sigma^2 \rightarrow 0$ 时, 由小散度渐进理论可得

$$\frac{Y - \mu}{\sigma\sqrt{V(\mu)}} \xrightarrow{d} N(0, 1),$$

其中, $Y \sim S^-(\mu, \sigma^2)$. 密度函数关于单位区间中点 (即 0.5) 对称, 即 $S^-(\mu, \sigma^2)$ 与 $S^-(1 - \mu, \sigma^2)$, $\text{Beta}(\alpha, \beta)$ 与 $\text{Beta}(\beta, \alpha)$ 和 $\text{LN}(\mu, \sigma^2)$ 与 $\text{LN}(-\mu, \sigma^2)$ 是关于 0.5 对称的.

2.1 单纯形随机变量生成

在很多方面我们都需要开发服从单纯形分布的随机数的算法, 其中一个例子就是模拟实验. 同时, 生成单纯形随机数的有效算法也是 Bayes 推断中 Markov chain Monte Carlo 方法的基础. 这种算法可以基于某种变换, 而这种变换的动因是单纯形分布是由逆 Gauss 混合分布的变换得到的^[13].

逆 Gauss 混合分布, 记为 $\text{M-IG}(\xi, \epsilon^2, p)$, 是逆 Gauss 分布 (以 $1 - p$ 的概率) 和其互补分布 (以概率 p) 的混合, 它具有如下密度函数:

$$f(x; \xi, \epsilon^2, p) = (2\pi\epsilon^2 x^3)^{-1/2} \left(1 - p + \frac{px}{\xi}\right) \exp\left\{-\frac{1}{2\epsilon^2} \frac{(x - \xi)^2}{\xi^2 x}\right\}, \quad x > 0.$$

以下命题表明, 单纯形随机变量可以通过变换服从逆 Gauss 混合分布的随机变量得到^[14].

命题 1 若 $X \sim \text{M-IG}(\xi, \epsilon^2, p)$, 令

$$y = \frac{x}{1+x}, \quad \mu = \frac{\xi}{1+\xi}, \quad \xi^2 = \frac{\epsilon^2}{(1-p)^2} = \epsilon^2(1-\xi)^2.$$

那么,

$$Y \sim S^-(\mu, \sigma^2).$$

注意到在这个变换中 Jacobian 式为 $(1 - y)^{-2}$, 证明是显然的. 所以, 为了生成服从单纯形分布 $S^-(\mu, \sigma^2)$ 的随机数, 我们可以先从逆 Gauss 分布 $\text{M-IG}(\xi, \epsilon^2, p)$ 中产生随机数, 其中,

$$p = \mu, \quad \xi = \frac{\mu}{1-\mu}, \quad \epsilon^2 = \sigma^2(1-\mu)^2,$$

再作一一变换

$$y = \frac{x}{1+x}.$$

Jørgensen^[13] 介绍了一种从逆 Gauss 分布和 χ_1^2 分布中产生服从逆 Gauss 混合分布的随机数的方法. 逆 Gauss 分布具有很多统计和概率性质, 其中, 包括抽样理论中的 χ^2 - 分布和 F - 分布^[5]. 本文采用 Michael 等人^[15] 提出的方法生成逆 Gauss 随机变量, 这种方法的基础是逆 Gauss 密度的核是 χ^2 - 分布. 这里提出的单纯形随机数生成的变换方法仅仅建立在 χ^2 随机数生成器和均匀分布随机数生成器之上. 这种产生单纯形分布随机数的方法可由 R 软件包 “simplexreg” 实现. 图 2 对比了从单纯形随机数生成器得到的 5000 个随机数的直方图和相应的单纯形分布的密度函数.

另外, 我们还可以用一种接受 - 拒绝算法^[16] 从单纯形分布中生成随机数. 当散度参数较大时, 将 Beta 分布作为包裹分布; 反之, 则用截断的正态密度^[14]. 值得注意的是, 尽管生成 Beta 随机数是很便利的, 但此接受 - 拒绝算法并不非常有效. 基于 Beta 分布的接受 - 拒绝算法对比了单纯形分布和 Beta 分布, 而 Beta 分布是广义单纯形分布的极限分布, 且二者具有相同的样本空间.

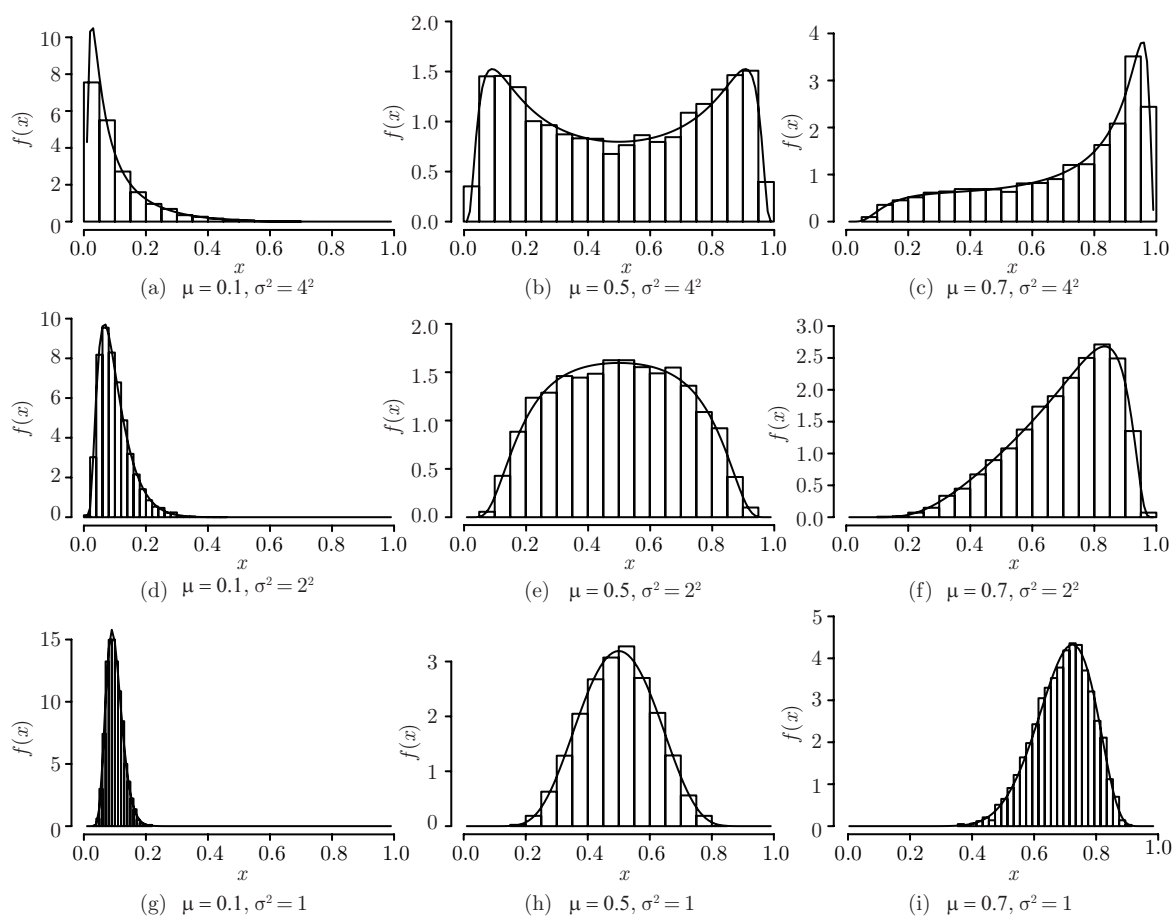


图 2 服从单纯形分布的随机数的直方图, 实线表示相应的单纯形分布密度函数的曲线

2.2 性质

易知单纯形模型满足 y 与 $1-y$, μ 与 $1-\mu$ 间微分与积分的可交换性. 若 $Y \sim S^-(\mu, \sigma^2)$, 则其得分函数如下:

$$u(y; \mu) = \frac{y - \mu}{\mu(1 - \mu)} \left\{ d(y; \mu) + \frac{1}{\mu^2(1 - \mu)^2} \right\},$$

于是 $d'(y; \mu) = -2u(y; \mu)$, 这里 ' 和 '' 分别表示一阶和二阶导数.

计算 Fisher 信息量时, 我们需要知道 $E[d(Y; \mu)]$, $E[(Y - \mu)d'(Y; \mu)]$ 和 $E[(Y - \mu)d(Y; \mu)]$.

命题 2 令 $Y \sim S^-(\mu; \sigma^2)$. 那么, $E[Y] = \mu$, 且

- (i) $E[d(Y; \mu)] = \sigma^2$;
- (ii) $E[(Y - \mu)d'(Y; \mu)] = -2\sigma^2$;
- (iii) $E[(Y - \mu)d(Y; \mu)] = 0$;
- (iv) $E[d'(Y; \mu)] = 0$;
- (v) $\frac{1}{2}E[d''(Y; \mu)] = \frac{3\sigma^2}{\mu(1-\mu)} + \frac{1}{\mu^3(1-\mu)^3}$;
- (vi) $\text{Var}[d(Y; \mu)] = 2(\sigma^2)^2$;
- (vii) $\text{Var}[u(Y; \mu)] = \frac{3\sigma^4}{\mu(1-\mu)} + \frac{\sigma^2}{\mu^3(1-\mu)^3}$.

(i)–(v) 的证明可参见文献 [6, 附录], 文献 [17] 给出了 (vi) 的证明, 本文附录给出了 (vii) 的证明.

命题 2 中描述的性质可用于统计建模和推断. 例如, 由 (i) 可以得出散度参数 σ^2 的矩法估计, 而 (iv)–(vii) 可用于推导单纯形广义线性模型^[14]、单纯形边际模型^[6, 17] 和单纯形混合效应模型^[12] 的统计推断.

3 单纯形广义线性模型

广义线性模型是一类从经典线性模型自然推广出的统计模型, 以线性回归作为其特例, 包括方差分析模型、针对二项数据的分对数和正态数模型、针对计数的分对数线性模型和多元离散选择模型, 以及一些针对存活数据的常用模型. 广义线性模型的主要思想可扩展到弥散模型, 如单纯形模型. 接下来将阐述单纯形广义线性模型, 简称为 SGLM, 它使用单纯形分布来描述比例数据. 为了避免与下面介绍的单纯形模型中的替代反应量发生混淆, 从现在起, 我们用 y 表示随机变量或者它的观察值, Y 将不再表示 y 的随机变量, 而将仅仅被用作表示替代反应量.

3.1 模型建立

考虑独立的比例反应变量 y_i , $0 < y_i < 1$, $i = 1, 2, \dots, m$. 令 $\mathbf{x}_i$ 为 p 维自变量向量, $\mathbf{z}_i$ 为 $\mathbf{x}_i$ 的子集. 下面将反应变量的均值和散度模型化为这些协变量的函数.

令 $y_i | \mathbf{x}_i \sim S^-(\mu_i, \sigma_i^2)$, 其中, 均值 μ_i 服从广义线性模型,

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (3.1)$$

这里函数 $g: (0, 1) \rightarrow (-\infty, \infty)$ 是关联函数, $\boldsymbol{\beta}$ 是回归系数向量. 散度参数 σ_i^2 可以与一些自变量联系在一起,

$$h(\sigma_i^2) = \mathbf{z}_i^T \boldsymbol{\gamma}, \quad (3.2)$$

其中, 函数 $h: (0, 1) \rightarrow (-\infty, \infty)$ 是关联函数, $\boldsymbol{\gamma}$ 是回归系数向量. 这是一个广义线性模型的扩展. 我们将关注 g 的分对数关联函数和 h 的对数关联函数. 这个模型可以处理更复杂的数据, 如由自变量导致的多模式聚类.

SGLM 同样满足广义线性模型的基本原则. 首先, 给定 $\mathbf{x}_i$, y_i , $i = 1, \dots, m$ 是独立分布的, 并有均值 $E(y_i | \mathbf{x}_i) = \mu_i$; 其次, 参数 μ_i 可以根据观测对象改变而改变, 它是通过线性函数 $\mathbf{x}_i^T \boldsymbol{\beta}$ 连接密度函数和回归变量的参数; 最后, 分对数关联函数是单调可微的. 但是, 单纯形密度函数不是广义形式上的指数分布族的成员, y_i 的方差也并不只随着 $\mathbf{x}_i$ 通过 μ_i 变化, 实际上, 也依赖于散度参数.

3.2 推断

在散度参数的变化下, 为了估计 SGLM 的参数 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$, 依照广义线性模型的经典理论, 我们用重赋权最小二乘迭代法计算 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$ 的极大似然估计. 显然, 对数似然函数有如下形式:

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\frac{1}{2} \sum_{i=1}^m \left[\frac{d(y_i; \mu_i)}{\sigma_i^2} - \log \sigma_i^2 \right], \quad (3.3)$$

另附带有个常数项.

定义替代或作用向量 $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)^T$ 的第 i 个元素

$$Y_i = g(\mu_i) + \frac{u(y_i; \mu_i)}{\sigma_i^2 \omega_i g'(\mu_i)}, \quad (3.4)$$

其中, $u(y_i; \mu_i)$ 是 (2.3) 中定义的得分函数, 而权重由命题 2(v) 给出,

$$\omega_i = \frac{E d''(y_i; \mu_i)}{2 \sigma_i^2 [g'(\mu_i)]^2} = \frac{1}{\sigma_i^2 [g'(\mu_i)]^2} \left(\frac{3 \sigma_i^2}{\pi_i} + \frac{1}{\pi_i^3} \right), \quad (3.5)$$

其中, $\pi_i = \mu_i(1 - \mu_i)$. 根据命题 2(iv)–(vii) 容易看出 $E[Y_i] = \eta_i = g(\mu_i)$ 并且 $\text{Var}[Y_i] = \omega_i$.

替代反应变量的表达式 (3.4) 是弥散模型下的自然推广. 对于指数分布族模型, 替代反应变量可以写成 $Y_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$, 这是经典广义线性模型中的常见形式.

(3.3) 和 (3.4) 的导数详见附录. 通过这些导数, 我们可以用重赋权最小二乘迭代法估计 β 和 γ [14], 即迭代求解 β 和 γ 的得分方程

$$\begin{aligned} \beta^{(k+1)} &= \beta^{(k)} + \left(\sum_{i=1}^m \mathbf{x}_i^T \omega_i^{(k)} \mathbf{x}_i \right)^{-1} \sum_{i=1}^m \mathbf{x}_i^T \omega_i^{(k)} (Y_i^{(k)} - \eta_i^{(k)}), \\ \gamma^{(k+1)} &= \gamma^{(k)} + \left(\sum_{i=1}^m \mathbf{z}_i^T u_i^{(k)} \mathbf{z}_i \right)^{-1} \sum_{i=1}^m \mathbf{z}_i^T v_i^{(k)}, \end{aligned}$$

其中, 当 h 为对数关联函数时,

$$u_i = \frac{1}{2}, \quad v_i = \frac{d(y_i; \mu_i)}{2 \sigma_i^2} - \frac{1}{2}.$$

$\omega_i^{(k)}, Y_i^{(k)}, \eta_i^{(k)}, u_i^{(k)}$ 和 $v_i^{(k)}$ 可在计算 β 和 γ 的第 k 步迭代中得到. 采用分对数关联函数 g , 注意到 $g'(\mu_i) = \pi_i^{-1}$, 我们可以将 (3.5) 简化为

$$\omega_i = 3\pi_i + \frac{1}{\sigma_i^2 \pi_i}.$$

β 的初始值可以从线性模型分对数变换的反应变量得到, 而 γ 的初始值, 根据命题 2(i), 可以通过基于 Γ -分布的反应变量 $d(y_i; \mu_i)$ 的对数线性模型得到.

回归参数 β 和 γ 的估计基于极大似然估计. 关于回归参数的假设检验可以选择 Wald 检验, 似然性检验和得分检验. 回归参数及它们的标准差的估计可以在 “simplexreg” 软件包中得到. 例如, 如果我们检验 $H_0: \beta_i = 0$ 与 $H_a: \beta_i \neq 0$, 令 $\hat{\beta}_i$ 和 $se(\hat{\beta}_i)$ 为 β_i 的估计及其标准误差估计, Wald 统计量由下式给出:

$$t = \left| \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \right|,$$

其 p 值为 $2 \times Pr(Z > t)$, 这里 Z 为服从标准正态分布的随机变量. β_i 的 $A(1 - \alpha) \times 100\%$ 置信区间为

$$(\hat{\beta}_i - Z_{\alpha/2} se(\hat{\beta}_i), \hat{\beta}_i + Z_{\alpha/2} se(\hat{\beta}_i)),$$

其中, $Z_{\alpha/2}$ 是标准正态分布在 $\alpha/2$ 水平下的临界值.

单纯形广义线性模型的一个特殊例子是固定散度参数, 即对 $y_i \sim S^-(\mu_i, \sigma^2)$, 我们只建立广义线性模型 (3.1). β 的估计方法不变, 而通过解得分方程, σ^2 的估计有解析形式

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m d(y_i; \mu_i),$$

这与由矩法得到的 σ^2 的估计是一致的 (见命题 2(i)). 文献 [11] 称此为 Jørgensen 估计量. 为了得到更好的无偏估计, 可调整自由度, 并得

$$\sigma^2 = \frac{1}{m - (p + 1)} \sum_{i=1}^m d(y_i; \mu_i),$$

其中, p 为 (3.1) 中的协变量数量.

但是, 当固定散度参数假设存疑时, 模拟实验^[6] 更倾向于可变的散度参数模型 (3.1) 和 (3.2). 如果没有因素影响散度参数, 模型 (3.2) 退化为特殊情形 — 常数散度. 相反地, 如果某些因素是聚类结构的关键因素, 如呈现多模式分布情形, 模型 (3.2) 在散度水平上有助于聚类过程模型化.

4 模拟实验

为了说明恰当地运用单纯形广义线性模型分析数据的重要性, 并比较单纯形模型、Beta 模型和分对数 - 正态模型对应的三种数据分布假设下的稳健性, 本文作了一系列模拟研究, 所用数据是以上相应的三种数据分布假设下随机产生的, 并且每组数据均由三种模型进行拟合.

在单纯形模型和 Beta 回归模型下, 应变量的均值 $\mu_i, i = 1, \dots, 500$ 由以下模型得到:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 S_i,$$

其中, T 为实验分组变量, 取值于 0 或 1, S 为环境因素, 且为 (0, 5) 均匀分布中取值的连续变量. 为了使得本节中的模拟结果能与第 5 节数据分析的结果在不同散度值下进行比较和讨论, 本节通过调整单纯形分布和 Beta 分布的参数, 使其产生的数据与各自不同散度值下的回收率具有相同的形状. 根据 Ferrari 和 Cribari-Neto^[9] 的研究, 记精度参数 $\phi = a + b$, 那么, 对于分布 Beta(a, b), 我们有 $a = \mu\phi$, $b = (1 - \mu)\phi$, 其中, ϕ 设为 16. 而对于单纯形回归模型, 散度参数 $\sigma^2 = 16$.

当数据来自于分对数 - 正态模型时, 我们使用

$$\text{logit}(Y_i) = \beta_0 + \beta_1 T_i + \beta_2 S_i + \varepsilon_i,$$

其中, $\varepsilon_i \sim N(0, 1)$.

简化起见, 我们只考虑 Y 的分布对参数 $\beta_j, j = 1, 2, 3$ 估计值的影响. 其真值不妨设为

$$(\beta_0, \beta_1, \beta_2) = (1.5, 0.5, -0.5).$$

重复 200 次所得的参数估计结果综述于表 1.

表 1 所示为三个模型参数的估计均值及其均方误差. 当 Beta 模型中 $\phi = 16$ 时, 散度参数估计的平均值为 $\sigma^2 = 2.2$. 然而, 大概 13.7% 的产生于 $\sigma^2 = 16$ 的单纯形模型的数据不能用 Beta 回归拟合. 图 3 中展示了各种情况下的核密度估计. 从中我们发现:

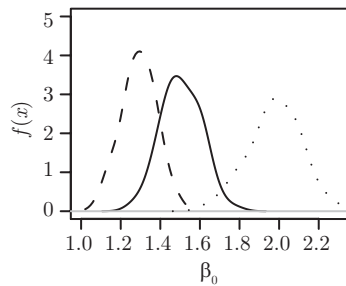
(i) 当分布假设不成立时, 单纯形回归模型比 Beta 模型和分对数 - 正态模型更有效. 这可以通过这一事实进行理解, 当数据并不是由单纯形分布产生时, 单纯形模型的偏差最小且均方误差增量最小.

(ii) 当数据来自单纯形分布或 Beta 分布时, 分对数 - 正态模型表现最差. 而 Beta 模型参数估计的标准差相较单纯形模型和分对数 - 正态模型更小.

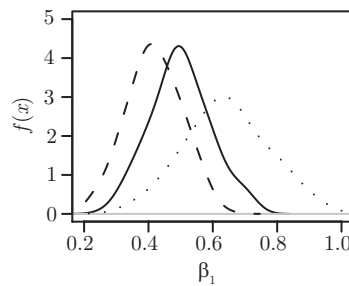
(iii) 对于比例数据来说, 单纯形模型的稳健性最好. 从图 3 中看出, 单纯形广义线性回归参数的密度估计总位于三种密度估计之间, 这说明当数据分布假设不成立时, 单纯形广义线性回归模型相较其他模型总有更小的估计误差.

表 1 基于 200 次重复实验的参数估计的汇总统计表

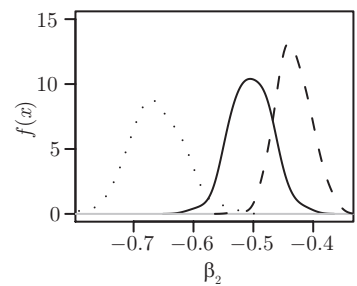
随机数生成模型	真值	拟合模型					
		单纯形		Beta		分对数 - 正态	
		均值	均方误差	均值	均方误差	均值	均方误差
单纯形	$\beta_0(1.5)$	1.5055	0.0205	1.2681	0.3081	1.9794	0.2653
	$\beta_1(0.5)$	0.5040	0.0170	0.4225	0.0199	0.6469	0.0539
	$\beta_2(-0.5)$	-0.5036	0.0023	-0.4279	0.0181	-0.6612	0.0297
Beta	$\beta_0(1.5)$	1.5258	0.0089	1.5029	0.0065	1.6373	0.0270
	$\beta_1(0.5)$	0.5093	0.0063	0.4989	0.0043	0.5480	0.0082
	$\beta_2(-0.5)$	-0.5086	0.0009	-0.5005	0.0006	-0.5457	0.0029
分对数 - 正态	$\beta_0(1.5)$	1.3229	0.1774	1.2343	0.1910	1.5225	0.1722
	$\beta_1(0.5)$	0.4263	0.1391	0.3990	0.1156	0.4971	0.1570
	$\beta_2(-0.5)$	-0.4374	0.0175	-0.4088	0.0201	-0.5046	0.0164



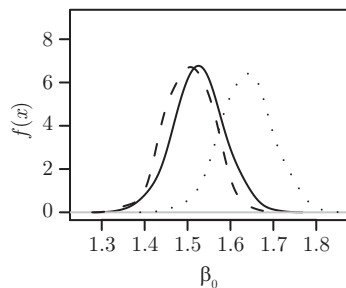
(a) 单纯形模型产生的数据



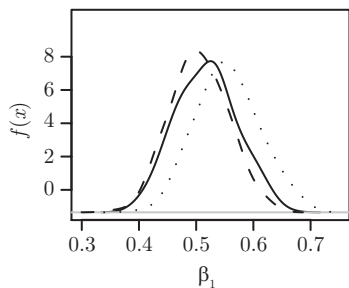
(b) 单纯形模型产生的数据



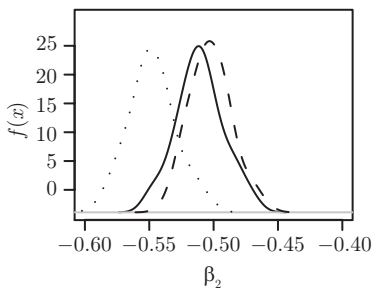
(c) 单纯形模型产生的数据



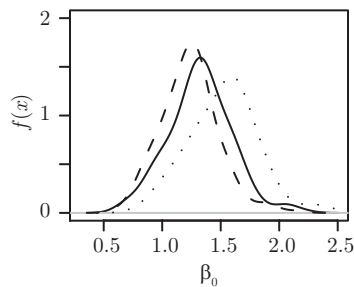
(d) Beta 模型产生的数据



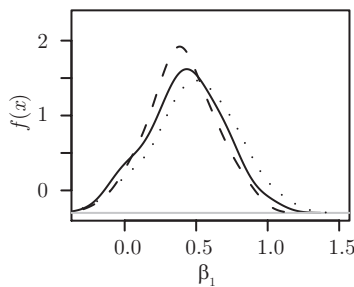
(d) Beta 模型产生的数据



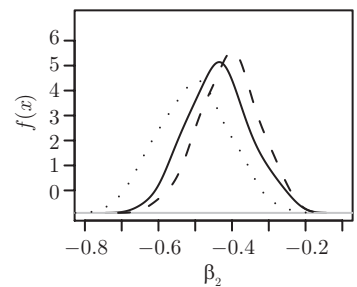
(d) Beta 模型产生的数据



(d) 分对数-正态模型产生的数据



(d) 分对数-正态模型产生的数据



(d) 分对数-正态模型产生的数据

图 3 基于 200 次重复实验的回归模型参数密度估计, 实线: 单纯形; 虚线: Beta; 点线: 分对数 - 正态

5 数据分析

自体外造血干细胞 (PBSC) 移植技术已广泛地用于多种恶性血液系统疾病的骨髓性治疗后的血液恢复. 造血功能重建主要取决于是否有足量的干细胞输注进入骨髓微环境^[19]. 其中活性 CD34+ 造血干细胞的剂量视为是否有充足 PBSC 可供采集的重要标志, 同时也是血液移植中的预测指标. 研究表明, 干细胞输注前的冷冻、低温保存和解冻过程将不可避免地损坏 PBSC, 并显著降低活性 CD34+ 细胞的数量^[20]. 活性 CD34+ 细胞的损耗程度一般通过比较低温保存后和冷冻前的 CD34+ 细胞的回收率变化来测算. 目前影响低温保存后细胞回收率的机制已引起了学术界的关注.

2003–2008 年, 加拿大阿尔伯塔省医疗总部下属的埃德蒙顿造血干细胞癌症研究所 (the Edmonton Hematopoietic Stem Cell Lab in Cross Cancer Institute-Alberta Health Services) 在患者同意的条件下进行了总计 242 例骨髓性剂量化疗后进行的自体 PBSC 移植的临床实验. 患者的年龄、性别和临床数据, 如癌症类型和化疗方案等, 都被记录下来. 患者年龄从 18 岁至 71 岁不等, 其中大部分为男性 (170), 并被诊断为多发性骨髓瘤、非 Hodgkin 淋巴瘤、急性白血病、实体瘤和淀粉样变等病症类型. 病人接受了初期的化疗, 有一日方案和三日方案, 方式包括粒细胞集落刺激因子等方法, 用于活化 CD34+ 细胞. 在外周血 CD34+ 循环计数达到或超过 15 个细胞 / μL 的时候开始采集 PBSC. 之后 PBSC 产物被冷冻, 并存储在液氮蒸气中直到再次输注. PBSC 样本在收集当日 (预冻处理) 和低温保存后 (解冻后) 分别评估样本中绝对活性 CD34+ 细胞的数量. 低温保存后的细胞活性以冷冻后活性细胞或菌落的数量所占冷冻前其数量的百分比来计算.

5.1 单纯形回归分析

CD34+ 细胞回收率的范围在 40% 至 100%. 回收率的直方图显示了某种左偏态分布. 数据中出现的两处极值 (100%), 用 99% 替换以避免比率数据在端点的定义问题. 通过对患者不同化疗类型的分类我们构造出类型变量 Chemo 来表述, 包括 G-CSF 和另外一种化疗类型在内的两种化疗类型, 并根据化疗实施天数区分为为一天方案和者三天方案. 横行依据 Chemo 分组, 纵列依据年龄分组 (< 40 或 ≥ 40) 建立列联表

$$\begin{pmatrix} 4 & 40 \\ 129 & 69 \end{pmatrix}.$$

几乎所有小于 40 岁的年轻患者都被安排了三日化疗方案, 而对于大多数老年患者则安排了一日化疗方案. 初始年龄和化疗方案这两个因素都没有在分别进行的初始拟合中显示统计显著性, 说明模型中这两个因素的效果是相互抵消的. 在接下来的模型拟合过程中, 样本中年龄小于 40 岁的个体设定为基准年龄, 以调整其他个体的年龄 (模型中表述为 Age).

用齐性或非齐性的散度结构对数据进行单纯形回归分析, 假定应变量期望的模型为

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Chemo}.$$

在非齐性散度结构的假设下, 散度参数对年龄作回归如下:

$$\log(\sigma_i^2) = \gamma_0 + \gamma_1 \text{Age}. \quad (5.1)$$

由此, 可将参数的估计与同样使用分对数连接函数的 Beta 回归模型和应变量经过分对数函数变换的分对数 - 正态线性回归模型的估计进行比较. 表 2 中给出了三种模型回归系数的估计和标准差.

表 2 PBSC 研究中回归系数的估计

模型	系数	估计	标准差	t -值	p -值
分对数 - 正态	β_0	1.170	0.142	8.253	< 0.0001
	β_1	0.017	0.007	2.498	0.0132
	β_2	0.243	0.127	1.907	0.0578
齐性 - 单纯形	β_0	1.100	0.141	7.801	< 0.0001
	β_1	0.014	0.007	2.092	0.0360
	β_2	0.266	0.125	2.128	0.0330
非齐性 - 单纯形	β_0	1.116	0.141	7.915	< 0.0001
	β_1	0.013	0.006	2.016	0.0440
	β_2	0.252	0.122	2.066	0.0390
非齐性 - 单纯形的散度模型	γ_0	2.337	0.249	9.386	< 0.0001
	γ_1	-0.015	0.007	-2.143	0.0320
Beta	β_0	1.042	0.112	9.349	< 0.0001
	β_1	0.014	0.005	2.711	0.0070
	β_2	0.214	0.102	2.108	0.0350

每个单纯形模型对应的信息量准则用“simplexreg”程序包计算得出, 包括 Akaike 信息量准则 (AIC) 和 Bayes 信息量准则 (BIC) [21], 计算公式分别为

$$\text{AIC} = -2\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) + 2(p + q),$$

$$\text{BIC} = -2\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) + (p + q) \log N,$$

其中, p 为 (3.1) 中参数的数量, q 为 (3.2) 中参数的数量. 数据分析结果表明, 齐性和非齐性单纯形回归模型的 AIC 分别为 246.04 和 245.00, BIC 分别为 253.02 和 253.47.

在齐性单纯形模型的假设下, 数据的散度估计是 6.397. 化疗类型变量经过年龄变量修正, 在单纯形模型中显著, 而在分对数 - 正态模型中不显著. 另外, 非齐性单纯形模型的结果表明, 低温保存后细胞回收率的散度与患者年龄相关.

Beta 回归模型中回归系数的估计一般具有较小的标准差, 因此, 与单纯形模型和分对数 - 正态模型相比, p -值更小. 与数据模拟的结果类似, 除了在 β_2 的估计上单纯形模型比分对数 - 正态模型大, 单纯形模型中回归系数的估计介于分对数 - 正态模型与 Beta 回归模型之间, 这说明单纯形模型相较另外两种模型更为可靠.

5.2 稳健性分析

为了研究异常数据是否会影响这些回归模型的分析结果, 我们在分对数 - 正态模型中用标准线性回归准则识别出异常数据.

结果五个影响最大, 并且学生化残差大于 2.5 的病人数据被视为异常数据. 删除了五个异常数据后, 新数据集分别用分对数 - 正态模型、Beta 模型和单纯形模型各拟合一次, 结果表明, 删除异常数据前后单纯形模型有相同的结论; 但是在分对数 - 正态模型中, Chemo 变量在删除异常数据后变得具有统计显著性, 如表 3 所示. 此外, 所有回归系数的估计值经过标准差调整后, 分对数 - 正态模型的参

表 3 删除 5 个异常数据后回归系数的估计

模型	系数	参数估计	标准差	t- 值	p- 值
分对数 - 正态	β_0	1.153	0.142	8.105	< 0.0001
	β_1	0.016	0.007	2.471	0.0142
	β_2	0.257	0.128	2.006	0.0461
齐性 - 单纯形	β_0	1.087	0.141	7.698	< 0.0001
	β_1	0.013	0.007	2.059	0.039
	β_2	0.277	0.126	2.203	0.028
非齐性 - 单纯形	β_0	1.101	0.142	7.756	< 0.0001
	β_1	0.013	0.006	2.006	0.045
	β_2	0.263	0.122	2.153	0.031
非齐性 - 单纯形的散度回归模型	γ_0	2.350	0.250	9.404	< 0.0001
	γ_1	-0.015	0.007	-2.235	0.025
Beta	β_0	1.028	0.112	9.206	< 0.0001
	β_1	0.014	0.005	2.677	0.007
	β_2	0.225	0.102	2.206	0.027

数估计值相对变化最大, 其中, 变量 Chemo 的相对变化率为 5.19%; 而 Beta 模型和单纯形模型中这个值分别为 4.74% 和 3.47%.

5.3 模型诊断

使用某一特定的模型拟合数据也即对其预测子形式、连接函数和分布函数进行选择. 一般来说, Pearson- χ^2 统计量和偏差统计量是通用的拟合优度统计量.

Pearson 残差形式为

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{Var}}(y_i)}} = \frac{y_i - \hat{\mu}_i}{\hat{\tau}_i},$$

其中, $\hat{\tau}_i$ 因为包含不完全 Γ 函数, 所以没有解析表达式, 见 (2.2). 对散度过大的数据, 其散度参数 σ^2 比较大, 因此, (2.2) 中的应变量方差趋近于 $\mu(1 - \mu)$. 这就推导出 Pearson 残差的一个近似形式

$$r_i^a = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\mu}_i(1 - \hat{\mu}_i)}}.$$

将参数用其对应的估计值替代, 我们可以绘制 $\hat{r}_i^P$ 或 $\hat{r}_i^a$ 对 $\hat{\mu}_i$ 的散点图, 观察其均值 - 方差关系, 并进一步检验模型中关于单纯形分布的假设. 在 PBSC 研究中, 我们发现上述方法所绘图形不具有明显形态, 比较随机, 且 99% 的数据点都在 -1.96 到 1.96 的水平区域中.

检验连接函数假设是否合理的一种有效的非正式方法是 McCullagh 和 Nelder^[1] 提出的用调整后应变量 s_i 对线性预测子 $\hat{\eta}_i$ 作图的方法. 本文中, 定义

$$s_i = g(\mu_i) + \left\{ \frac{3\sigma^4}{\mu_i(1 - \mu_i)} + \frac{\sigma^2}{V(\mu_i)} \right\}^{-1/2} u(y_i; \mu_i),$$

其中, $u(y_i; \mu_i)$ 和 $V(\mu_i) = \mu_i^3(1 - \mu_i)^3$ 分别是得分函数和单纯形模型的单位方差函数. 由命题 2 可知, 因为 $E(u_i) = 0$, 所以 $E(s_i) = g(\mu_i)$, 且 $\text{Var}(s_i) = E\{s_i - g(\mu_i)\}^2 = 1$. 若连接函数的选择是恰当的, 则 $\hat{s}_i$ 对 $\hat{\eta}_i$ (见 (3.1)) 作图应是一条直线, 其中, 95% 的点落在上下边界为 $\hat{\eta}_i \pm 1.96$ 的区间中. 对于 PBSC 的研究数据, 包括分对数函数、probit、互补双对数函数 (complementary log-log) 和双对数函数 (log-log) 四种连接函数作出的散点图形状十分相似, 均有 98% 的点落入这个区间. 另一方面, 单纯形回归模型中不同连接函数对应的对数似然函数间的差异可以通过在 (3.3) 中插入参数估计值得到. 结果表明, 在对 PBSC 的研究中, 对数似然函数间的差异很小. 函数值最大的为双对数连接函数, 其对数似然函数值为 54.60, 稍大于分对数连接函数的 54.39.

作为广义线性模型的一个推广, 单纯形广义线性模型离差的测度或拟合优度可基于偏差统计量来构建. 当单位偏差函数 $d(y; \mu) = d(\mu; \mu) = 0$, 在完全拟合 $\hat{\mu}_i = y_i$ 假设下即得到饱和模型的对数似然函数. 因此, 由 (3.3), 偏差函数是

$$D = \sum_{i=1}^m D(y_i; \hat{\mu}_i) = 2 \sum_{i=1}^m \{\ell_i(y_i; y_i) - \ell_i(\hat{\mu}_i; y_i)\} = \sum_{i=1}^m \frac{d(y_i; \hat{\mu}_i)}{\sigma_i^2},$$

其服从 $\chi^2_{(m-p)}$ 分布. 对于 PBSC 的研究, 其齐性单纯形模型和非齐性单纯形模型 χ^2 检验的 p -值分别是 0.543 和 0.451, 这说明单纯形模型无失拟现象.

6 讨论

本文开发了基于一种弥散模型, 即单纯形分布, 对横断面研究中的比例型数据建模的方法. 用单纯形分布的性质, 在单纯形广义线性模型下建立了完善的统计推断和残差分析方法. 数据模拟和实际数据样例的计算结果充分说明, 与 Beta 回归模型和分对数 - 正态线性回归模型相比, 单纯形模型在分布假设和对异常数据的处理上具有显著的稳健性.

单纯形模型可以提供相当多的具有多种形态的分布以供选择, 从平稳对称分布到偏态分布, 从单一极值分布到多极值分布, 并且可以通过散度参数来延展. 特别地, 当散度很大时, 混合分布的特性允许我们对来自非齐性总体的数据建模. 只要合理地描述散度参数, 即能充分拟合随机误差, 这样就能在识别系统性关联时提高检验的功效. 本文的数据模拟和实证分析的结果已经说明了这点. 作为散度很大时一般单纯形分布的极限分布, Beta 分布可以良好的拟合散度较大的比例数据, 但是当数据分布的散度较小时, Beta 分布有可能过度估计回归分析中参数的影响.

本文完成了单纯形分布的密度函数, 累积概率分布及其逆函数的推导, 以及相应随机数生成器的构建. 基于函数变换的随机数生成器十分高效准确. 它或可作为基于模拟的推断方法, 如作为 Markov chain Monte Carlo 模拟的基础工具和用于拟合层次化单纯形广义线性模型下的多层比例数据.

多维单纯形模型是本文讨论的单纯形广义线性模型的一种重要推广. Jørgensen 和 Lauritzen^[22] 提出了一类处理多元非正态应变量的多元弥散模型. 多元分对数 - 正态分布已经用作对多元比例数据, 即成分数据^[8], 作回归分析. 而成分数据散度较大时, 多元单纯形分布^[22] 可作为多元回归模型的随机机制. 这将是今后进一步研究的课题之一.

参考文献

- 1 McCullagh P, Nelder J A. Generalized Linear Models, 2nd ed. London: Chapman and Hall, 1989

- 2 Jørgensen B. The Theory of Dispersion Models. London: Chapman and Hall, 1997
- 3 Fisher R A. Dispersion on a sphere. Proc R Soc Ser A Math Phys Eng Sci, 1953, 217: 295–305
- 4 Nelder J A, Wedderburn R W M. Generalized linear models. J Roy Statist Soc Ser A, 1972, 135: 370–384
- 5 Barndorff-Nielsen O E, Jørgensen B. Some parametric models on the simplex. J Multivariate Anal, 1991, 39: 106–116
- 6 Song P X-K, Qiu Z, Tan M. Modeling heterogeneous dispersion in marginal simplex models for continuous longitudinal proportional data. Biom J, 2004, 46: 540–553
- 7 Mead R. A Generalised logit-normal distribution. Biometrics, 1965, 21: 721–732
- 8 Aitchison J. The Statistical Analysis of Compositional Data. London: Chapman and Hall, 1986
- 9 Ferrari S L P, Cribari-Neto F. Beta regression for modelling rates and proportions. J Appl Stat, 2004, 31: 799–815
- 10 Cribari-Neto F, Zeileis A. Beta regression in R. J Stat Softw, 2010, 34: 1–24
- 11 Song P X-K. Correlated Data Analysis-Modeling, Analytics and Applications. New York: Springer-Verlag, 2007
- 12 Qiu Z, Song P X-K, Tan M. Simplex mixed-effects models for longitudinal proportional data. Scand J Stat, 2008, 35: 577–596
- 13 Jørgensen B. On the mixture of the inverse Gaussian distribution with its complementary reciprocal. Scand J Stat, 1991, 18: 77–89
- 14 Qiu Z. Simplex mixed models for longitudinal proportional data. PhD Thesis. Toronto: York University, 2001
- 15 Michael J R, Schucany W, Hass R W. Generating random variates using transformations with multiple roots. Stat Comput, 1976, 30: 88–90
- 16 Robert C P, Casella G. Monte Carlo Statistical Methods. New York: Springer-Verlag, 1999
- 17 Song P X-K, Tan M. Marginal models for longitudinal continuous proportional data. Biometrics, 2000, 56: 496–502
- 18 Green P J. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. J R Stat Soc Ser B Stat Methodol, 1984, 46: 149–192
- 19 Allan D S, Keeney M, Howson-Jan K, et al. Number of viable CD34+ cells reinfused predicts engraftment in autologous hematopoietic stem cell transplantation. Bone Marrow Transpl, 2002, 20: 967–72
- 20 Yang H, Acker J P, Cabuhat M, et al. Association of post-thaw viable CD34+ cells and CFU-GM with time to hematopoietic engraftment. Bone Marrow Transpl, 2005, 35: 881–887
- 21 Schwarz G. Estimating the dimension of a model. Ann Statist, 1978, 6: 461–464
- 22 Jørgensen B, Lauritzen S L. Multivariate dispersion models. J Multivariate Anal, 2000, 74: 267–281

附录

命题 2(vii) 的证明 从 (iv) 的证明出发可以得到 (vii). 因为

$$\int_0^1 [2\pi\sigma^2\{y(1-y)\}^3]^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\} dy = 1,$$

等式两边对 μ 求导, 得到

$$\int_0^1 d'(y;\mu)[2\pi\sigma^2\{y(1-y)\}^3]^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\} dy = 0,$$

因此, $E[d'(Y;\mu)] = 0$. 再次对等式两边对 μ 求导, 得到

$$\begin{aligned} & \int_0^1 d''(y;\mu)[2\pi\sigma^2\{y(1-y)\}^3]^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\} dy \\ &= \frac{1}{2\sigma^2} \int_0^1 (d'(y;\mu))^2 [2\pi\sigma^2\{y(1-y)\}^3]^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\} dy, \end{aligned}$$

即

$$E[d''(Y;\mu)] = \frac{1}{2\sigma^2} E[\{d'(Y;\mu)\}^2],$$

故

$$\text{Var}[u(Y;\mu)] = \frac{1}{4} \text{Var}[d'(Y;\mu)] = \frac{\sigma^2}{2} E[d''(Y;\mu)],$$

由此得 (vii). □

SGLM 的推导 将似然函数 (3.3) 对 β 求导, 得到

$$\frac{\partial \ell(\beta, \gamma)}{\partial \beta} = - \sum_{i=1}^m \mathbf{x}_i^T \frac{d'(y_i; \mu_i)}{2\sigma_i^2 g'(\mu_i)},$$

以及

$$\frac{\partial^2 \ell(\beta, \gamma)}{\partial \beta \partial \beta^T} = - \sum_{i=1}^m \mathbf{x}_i^T \mathbf{x}_i \frac{d''(y_i; \mu_i)}{2\sigma_i^2 \{g'(\mu_i)\}^2} - \frac{\partial \{g'(\mu_i)\}^{-1}}{\partial \mu_i} \mathbf{x}_i^T \mathbf{x}_i \frac{d'(y_i; \mu_i)}{2\sigma_i^2 g'(\mu_i)}.$$

由此可知, 似然函数对 β 的得分向量为

$$\frac{\partial \ell(\beta, \gamma)}{\partial \beta} = \sum_{i=1}^m \mathbf{x}_i^T w_i (Y_i - \eta_i),$$

而 Fisher 信息量为

$$I(\beta) = E \left(- \frac{\partial^2 \ell(\beta, \gamma)}{\partial \beta \partial \beta^T} \right) = \sum_{i=1}^m \mathbf{x}_i^T w_i \mathbf{x}_i,$$

因为由命题 2(iv) 可知, $Ed'(y_i; \mu_i) = 0$.

将似然函数 (3.3) 对 γ 求导, 得到

$$\frac{\partial \ell(\beta, \gamma)}{\partial \gamma} = - \sum_{i=1}^m \frac{\mathbf{z}_i^T}{2h'(\sigma_i^2)} \left(- \frac{d(y_i; \mu_i)}{2(\sigma_i^2)^2} + \frac{1}{\sigma_i^2} \right),$$

以及

$$\frac{\partial^2 \ell(\beta, \gamma)}{\partial \gamma \partial \gamma^T} = - \sum_{i=1}^m \frac{\mathbf{z}_i^T \mathbf{z}_i}{2h'^2(\sigma_i^2)} \left(\frac{2d(y_i; \mu_i)}{(\sigma_i^2)^3} - \frac{1}{(\sigma_i^2)^2} \right) + \sum_{i=1}^m \frac{h''(\sigma_i^2) \mathbf{z}_i^T \mathbf{z}_i}{2h'^3(\sigma_i^2)} \left(- \frac{d(y_i; \mu_i)}{(\sigma_i^2)^2} + \frac{1}{(\sigma_i^2)} \right).$$

由此可知, 似然函数对于 γ 的得分向量为

$$\frac{\partial \ell(\beta, \gamma)}{\partial \gamma} = \sum_{i=1}^m \mathbf{z}_i^T v_i,$$

而 Fisher 信息量为

$$I(\gamma) = E \left(- \frac{\partial^2 \ell(\beta, \gamma)}{\partial \gamma \partial \gamma^T} \right) = \sum_{i=1}^m \mathbf{z}_i^T u_i \mathbf{z}_i,$$

因为由命题 2(i), 有 $Ed(y_i; \mu_i) = \sigma_i^2$, 这里,

$$u_i = \frac{1}{2(\sigma_i^2)^2 h'^2(\sigma_i^2)}, \quad v_i = \frac{d(y_i; \mu_i) - \sigma_i^2}{2(\sigma_i^2)^2 h'(\sigma_i^2)}.$$

Regression analysis of proportional data using simplex distribution

ZHANG Peng & QIU ZhenGuo

Abstract Outcomes of continuous proportions arise in many applied areas. Such data are typically measured as percentages, rates or proportions confined in the unitary interval. In this paper, we propose to use a disper-

sion model of the simplex distribution to model such proportional outcomes. We study some key properties of the simplex distribution, which will be applied to establish estimation and inference in the regression analysis. Simulation studies show better robustness of the proposed simplex regression model against violation in some distributional assumptions than the beta distribution based regression model and the logit-normal linear regression model. We illustrate the proposed methodology by a real data analysis on peripheral blood stem cell transplants and robustness against outliers. The proposed framework is implemented on the R platform in a package named “simplexreg”, which can be downloaded from <http://my.zju.edu.cn/share/2466293> (verification code: 7919).

Keywords Beta regression, dispersion models, generalized linear models, logit-normal regression, random variable generation

MSC(2010) 62J12, 62E15

doi: 10.1360/012013-200