Biometrics 000, 000–000 DOI: 000
000 0000

局部拟指数增长模型：核微分方程回归和稀疏数据

Chunlei Ge* and W. John Braun**

Department of Computer Science, Mathematics, Physics and Statistics,
The University of British Columbia Okanagan, Kelowna, BC V1V 1V7, Canada.

**email:* chunlei.ge@ubc.ca

***email:* john.braun@ubc.ca

Summary: 局部多项式回归在处理稀疏数据时面临多个挑战。难以捕捉基础函数的局部特征可能导致对真实关系的潜在误表示。此外，由于局部邻域中的数据点有限，估计量的方差可能会显著增加。局部多项式回归还需要大量的数据来生成良好的模型，使其对于稀疏数据集来说效率较低。本文采用由 Ding, A. A., and Wu, H. (2014) 引入的微分方程约束回归方法，用于局部准指数增长模型。通过引入一阶微分方程，该方法扩展了局部多项式回归处理稀疏设计的能力，同时减少了偏差和方差。我们使用一阶泰勒多项式讨论核估计量的渐近偏倚和方差。利用小鼠肿瘤生长数据进行模型比较，并在模拟不同噪声水平和增长率的各种场景下进行仿真研究。

Key words: 非参数回归；局部多项式回归；稀疏设计；微分方程。

This paper has been submitted for consideration for publication in *Biometrics*

1. 介绍

非参数回归避免了对函数形式或误差分布的严格假设，使其适用于参数模型可能失败的情景。局部多项式回归通常能够捕捉到复杂、非线性关系，这是参数方法可能会遗漏的。Fan, J. & Gijbels, I. (1996) 研究了局部多项式回归的效率和渐近性质，而 Ruppert, D., & Wand, M. P. (1994) 则为多变量局部回归发展了一种渐近分布理论。

微分方程在 Hirsch, M. W. et al. (1974) 中讨论了函数与其导数之间的关系，并且是建模动态系统的基础。泰勒展开或泰勒级数，由 Taylor, B. (1717) 引入，提供了一个函数及其在一个点上的导数的近似值。它可以用于数值求解微分方程。

传统上，局部多项式回归在稀疏数据区域（尤其是在边界处）遇到困难。通过引入微分方程约束，局部多项式回归可以更好地在外推稀疏区域时减少偏差和方差。本文将采用由 Ding, A. A., and Wu, H. (2014) 引入的带有微分方程约束的回归方法，应用于局部拟指数增长模型，这是一种相对简单但通用的增长模型。

本文组织如下。第 2 节概述了 DE 约束局部多项式回归方法。第 3 节讨论了 DE 约束估计的渐近性质。数值属性在第 4 节中进行了讨论，并应用于一个实际数据集。之后，有一个讨论部分来回顾我们的方法并概述我们未来的工作。

2. 方法论

2.1 微分方程约束回归模型

我们考虑通过一阶微分方程增强的回归模型。我们的目标是在一个简单但实用的环境中研究受微分方程约束的局部回归估计器。

给定 n 个解释变量 x 和响应变量 y 的独立观测值，我们考虑如下形式的模型

$$y_i = g(x_i) + \varepsilon_i, \quad g'(x) = F(x, g(x)), \quad i = 1, 2, \ldots, n,$$

其中 F 是某个 Lipschitz 连续函数，并且误差项 ε_i 互不相关且均值为零。我们假设设计点是从区间 $[a, b]$ 按照概率密度函数 f 随机采样的，或者是在该区间内根据固定抽样方案选取的。Ding, A. A., and Wu, H. (2014) 考虑了这种设置并开发了一个类似的程序，但他们的目标是估计微分方程中的参数；他们没有研究函数估计过程本身。

一个简单的受微分方程约束的回归模型是具有特定微分方程约束的局部指数增长模型：

$$y_i = g(x_i) + \varepsilon_i, \quad g'(x) = \lambda g(x), \quad i = 1, 2, \ldots, n,$$

其中，$\lambda \neq 0$ 和 ε_i 是不相关的均值为零的误差。

然而，指数模型常常导致对后期增长的高估。我们考虑一个一般的生长模型，即局部拟指数增长模型，如下所示。

2.2 局部拟指数增长模型

假设我们有一个模型，其形式一般为拟指数型

$$y_i = g(x_i) + \varepsilon_i, \quad g'(x) = \lambda g^{\alpha}(x), \quad i = 1, 2, \ldots, n, \tag{1}$$

其中 $0 < \alpha \leqslant 1$、$\lambda \neq 0$ 和 ε_i 是不相关的均值为零的误差。ε 的方差是 σ_{ε}^2。

由于增长大致可以用指数模型来近似，因此对响应变量进行对数变换，并假设误差在对数尺度上是加性的更有意义。在正态分布的假设下，这意味着在原始尺度上，误差是乘性的，并且服从对数正态分布。因此，在对数尺度上的模型为

$$y_i = g(x_i)e^{\epsilon_i}, \quad g'(x) = \lambda g^{\alpha}(x), \quad i = 1, 2, \ldots, n,$$

经过对数变换后，该模型变为

$$\log(y_i) = G(x_i) + \epsilon_i, \quad G'(x) = \lambda e^{(\alpha-1)G(x)}, \quad i = 1, 2, \ldots, n, \tag{2}$$

其中 $G(x) = \log(g(x))$,$0 < \alpha \leqslant 1$,$\lambda \neq 0$ 和 ϵ_i 是均值为零、方差为 σ_ϵ^2 的正态分布。

对数变换简化了模型，使其在参数方面呈线性关系，这通常从统计处理的角度来看更容易应对。

在本文的以下部分中，我们将重点研究微分方程约束回归模型的这些特定案例，并探讨微分方程约束估计的理论性质。

2.3 受微分方程约束的估计

我们描述了模型 (1) 中 $g(x)$ 的估计程序。模型 (2) 中 $G(x)$ 的估计程序是类似的。

2.3.1 *线性尺度估计.* 对于给定的评估点 x，其中 $x \in (a, b)$，通过最小化局部最小二乘目标函数

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 K_h(x_i - x),$$

获得约束于微分方程的估计器 $g(x)$，其中 x_i 是设计点，核函数 $K_h(x)$ 是一个按带宽 h 缩放的对称概率密度函数，并且核函数 K 满足正则性条件：非负性、归一化和有限二阶矩。带宽 h 满足以下条件：$h \to 0$ 和 $nh \to \infty$ 作为 $n \to \infty$.

使用模型 (1) 中的微分方程 $g'(x) = \lambda g^{\alpha}(x)$，我们得到 $g(x)$ 的 p^{th} 阶导数函数：

$$g^{(p)}(x) = \lambda^p \left(\prod_{l=1}^{p} (l-1)\alpha - (l-2) \right) g^{p\alpha - p + 1}(x).$$

我们将 $\prod_{l=1}^{p}(l-1)\alpha-(l-2)$ 表示为后续的 $\pi_{\alpha,p}$。用这种表示方法，我们写出 $g^{(p)}(x) = \lambda^p \pi_{\alpha,p} g^{p\alpha-p+1}(x)$。

应用 k^{th} 阶泰勒展开式在 x 充分小的邻域内对 $g(x_i)$ 进行展开，我们得到：

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - g(x_i))^2 K_h(x_i - x) \\
&\doteq \sum_{i=1}^{n} \left\{ y_i - g(x) - \sum_{p=1}^{k} \frac{1}{p!}(x_i - x)^p g^{(p)}(x) \right\}^2 K_h(x_i - x) \\
&= \sum_{i=1}^{n} \left\{ y_i - g(x) - \sum_{p=1}^{k} \frac{1}{p!}(x_i - x)^p \lambda^p \pi_{\alpha,p} g^{p\alpha-p+1}(x) \right\}^2 K_h(x_i - x)
\end{aligned} \tag{3}$$

k^{th} 阶 DE 约束估计量在 x 处是通过最小化权重和 (3) 关于单一参数 $g(x)$ 来获得的。该最小化问题通过对加权非线性最小二乘问题求解来实现，使用高斯-牛顿算法迭代求解很容易完成，前提是给出了合适的初始猜测值。例如，2^{nd} 阶 DE 约束估计量可以通过使用局部常数回归估计作为迭代的起始值来获得。对于 $g(x)$，局部常数估计在数值上处理稀疏设计比高阶局部多项式回归更好。因为它在相当一般的条件下以速率 $O_p(n^{-2/5})$ 渐近收敛到真实值，所以它可以为该迭代提供一个好的初始值。

2.3.2 对数尺度估计. 对于估计 $G(x)$，我们使用 $G(x)$ 的 p^{th} 阶导数函数：

$$G^{(p)}(x) = (p-1)!\lambda^p(\alpha-1)^{p-1}e^{p(\alpha-1)G(x)}$$

并在 x 足够小的邻域内对 $G(x_i)$ 应用 k^{th} 阶泰勒展开，我们得到：

$$\begin{aligned}
&\sum_{i=1}^{n}(\log(y_i) - G(x_i))^2 K_h(x_i - x) \\
&\doteq \sum_{i=1}^{n}\left\{\log(y_i) - G(x) - \sum_{p=1}^{k}\frac{1}{p!}(x_i-x)^p G^{(p)}(x)\right\}^2 K_h(x_i-x) \\
&= \sum_{i=1}^{n}\left\{\log(y_i) - G(x) - \sum_{p=1}^{k}\frac{1}{p}(x_i-x)^p \lambda^p(\alpha-1)^{p-1}e^{p(\alpha-1)G(x)}\right\}^2 K_h(x_i-x) \qquad (4)
\end{aligned}$$

通过最小化加权和 (4)，我们获得 $G(x)$ 的 k^{th} 阶 DE 约束估计器。

3. 渐近性质

当我们评估估计量的行为随样本容量无限增长时，渐近性质提供了有用的指导。在本节中，我们将讨论 k^{th} 阶 DE 约束估计量 $\hat{g}_k(x)$ 的条件渐近分析。对于模型 (1)，我们做以下假设：$g(x)$，均值函数，在 x 附近具有有界且连续的 $(k+1)^{th}$ 阶导数。设计密度 $f(x)$ 是二阶连续可微且为正的。并且核函数 $K(\cdot)$ 是非负的、对称的，并且在区间 $[a,b]$ 上具有紧支集的有界 PDF。核函数满足 $\int_{-\infty}^{\infty}K(w)dw = 1$，$R(K) = \int K^2(w)dw < \infty$，并且其矩有限到六阶。缩放后的核函数定义为 $K_h(\cdot) = h^{-1}K(\cdot/h)$，继承了原核函数的性质。

在上述假设下，我们有以下关于模型（1）的估计量 $\hat{g}_k(x)$ 的渐近条件偏差和方差的定理。

定理 1（渐近条件偏差） k^{th} 阶 DE 约束估计量 $\hat{g}_k(x)$ 在区间 $[a,b]$ 的内部具有渐近条件偏差

$$\text{Bias}(\widehat{g}_k(x)|x_1,...,x_n) = \frac{1}{(k+1)!}g^{(k+1)}(x)h^{k+1}\mu_{k+1} + o_p(h^{k+1}), \quad k \quad \text{odd}, \qquad (5)$$

其中 $\mu_{k+1} = \int w^{k+1}K(w)dw < \infty$，

并且当 k 是偶数时，

$$\text{Bias}(\widehat{g}_k(x)|x_1,...,x_n) = \frac{1}{(k+1)!}g^{(k+1)}(x)\left(\frac{\lambda[(k+1)\alpha-k]g^{\alpha-1}(x)}{k+2} + \frac{f'(x)}{f(x)}\right)h^{k+2}\mu_{k+2} + o_p(h^{k+2}), \qquad (6)$$

其中 $\mu_{k+2} = \int w^{k+2}K(w)dw < \infty$。

在方程 (5) 和 (6) 中，$g^{(k+1)}(x)$ 是 $g(x)$ 的 $(k+1)^{th}$ 阶导函数，

$$g^{(k+1)}(x) = \lambda^{k+1}\pi_{\alpha,k+1}g^{(k+1)\alpha-k}(x).$$

定理 2（渐近条件方差） 在区间 $[a,b]$ 内部的 k^{th} 阶 DE-约束估计量 $\hat{g}_k(x)$ 具有渐近条件方差

$$\text{Var}(\widehat{g}_k(x)|x_1,...,x_n) \approx \frac{\sigma^2 R(K)}{nhf(x)} + o_p\left(\frac{1}{nh}\right). \qquad (7)$$

上述定理提供了一种通过最小化渐近均方误差（AMSE）来选择 $\hat{g}_k(x)$ 的渐近最优带宽的工具。

定理 3（渐近最优带宽） 在模型 (1) 的假设下，k^{th} 阶估计量 $\hat{g}_k(x)$ 的渐近最优带宽由以下给出：

$$h_{o,k}^{2k+3} = \frac{\sigma^2 R(K)((k+1)!)^2}{nf(x)\lambda^{2k+2}\pi_{\alpha,k+1}^2 g^{2(k+1)\alpha-2k}(x)(2k+2)\mu_{k+1}^2}, \qquad (8)$$

当 k 为奇数时，

和

$$h_{o,k}^{2k+5}=\frac{\sigma^2R(K)((k+1)!)^2}{nf(x)\lambda^{2k+2}\pi_{\alpha,k+1}^2g^{2(k+1)\alpha-2k}(x)\left(\frac{\lambda[(k+1)\alpha-k]g^{\alpha-1}(x)}{k+2}+\frac{f'(x)}{f(x)}\right)^2(2k+4)\mu_{k+2}^2}, \tag{9}$$

当 k 为偶数时，其中

$$g(x)=\{(1-\alpha)(\lambda x+g(0)\}^{1/(1-\alpha)}, \tag{10}$$

这是模型 (1) 中的微分方程的显式解。

备注: 定理 3 提供了渐近最优带宽的选择方法。公式（8）和（9）表示了对 k^{th} 阶 DE 约束回归的带宽评估。在实际应用中，如果我们获得了开始带宽当 $k=0$ 或 $k=1$ 时，以下公式（11）和（12）更为有用：

$$h_{o,k+2}=\left(\frac{(k+3)(k+1)}{\lambda^4[(k+2)\alpha-(k+1)]^2[(k+1)\alpha-k]^2g^{4\alpha-4}(x)}h_{o,k}^{2k+3}\right)^{1/(2k+7)}, \tag{11}$$

当 k 为奇数时，

当 k 是偶数时，

$$h_{o,k+2}=\left(\frac{(k+2)^3}{(k+4)\lambda^4[(k+1)\alpha-k]^2[(k+2)\alpha-(k+1)]^2g^{4\alpha-4}(x)}\frac{\left(\frac{\lambda[(k+1)\alpha-k]g^{\alpha-1}(x)}{k+2}+\frac{f'(x)}{f(x)}\right)^2}{\left(\frac{\lambda[(k+3)\alpha-(k+2)]g^{\alpha-1}(x)}{k+4}+\frac{f'(x)}{f(x)}\right)^2}h_{o,k}^{2k+5}\right)^{1/(2k+9)}, \tag{12}$$

。

例如，在一个模拟研究中，我们可以使用公式 (9) 或局部常数回归来找到 $h_{o,0}$，即当 $k=0$ 时的渐近最优带宽。然后可以通过公式 (12) 获得当 $k=2$ 时的渐近最优带宽。类似地，我们使用公式 (8) 或局部线性回归来找到 $h_{o,1}$，即当 $k=1$ 时渐近最优带宽。然后可以通过公式 (11) 获得当 $k=3$ 时的渐近最优带宽。接着，我们可以逐步得到更高 k^{th} 阶回归的渐近最优带宽。

对于对数尺度模型 (2)，我们可以通过应用上述定理和函数 $G(x)$ 的 p^{th} 导数，获得关于估计量 $\hat{G}(x)$ 的渐近性质的类似定理，

$$G^{(p)}(x)=(p-1)!\lambda^p(\alpha-1)^{p-1}e^{p(\alpha-1)G(x)}.$$

4. 稀疏肿瘤生长数据的应用

在本节中，我们将考虑对数比例模型（2）的以下示例，该示例涉及一项动物实验中的化疗试验的一组控制数据。小鼠肿瘤数据（Plume, C. A. et al.（1993））是在一段时间内从小鼠身上收集的肿瘤体积数据。从单只小鼠处进行了肿瘤体积测量。时间以天为单位记录，体积以立方厘米为单位。

为了说明各种局部多项式模型在稀疏数据上的表现，我们人为移除了一些数据点。

使用模型 (2) 中的微分方程，$G(x_i)$ 在 x 充分小的邻域内的一阶泰勒展开给出

$$\log(y_i)\doteq G(x)+\lambda\mathrm{e}^{(\alpha-1)G(x)}(x_i-x)+\varepsilon_i. \tag{13}$$

此外，二阶泰勒展开给出

$$\log(y_i)\doteq G(x)+\lambda\mathrm{e}^{(\alpha-1)G(x)}(x_i-x)+\frac{1}{2}\lambda^2(\alpha-1)\mathrm{e}^{2(\alpha-1)G(x)}(x_i-x)^2+\varepsilon_i \tag{14}$$

当在 DE 约束回归方法中实现时，我们将应用展开（13）和（14）的模型分别称为一阶和二阶局部拟增长模型，它们采用了一阶和二阶 DE 约束回归。

由于我们不知道该数据集的真实模型，我们任意选择整个数据集的局部线性估计作为“真实值”。带宽是通过 R 中的核平滑包（Wand, M., & Ripley, B. (2015)）中的`dpill`函数（Ruppert, D., Sheather, S. J., & Wand, M. P. (1995)）获得：$h = 2.38$。残差的标准差为 0.089。局部线性拟合 $\widehat{G}_{\ell\ell}(x)$ 以及标准差成为另一项模拟研究的基础，在该研究中，新的观测值根据均值为 $\widehat{G}_{ll}(x_i)$（$i \in 1, 2, \ldots, 10$）和经验标准差的正态分布生成于原始设计点 x_i。

我们在模拟数据集上训练竞争模型时，每次都移除第 4 到第 8 个观测值。测试的模型包括局部常数（NW）、局部线性（LL）、局部二次（LQ）、一阶局部拟增生（DE1）、二阶局部拟增生（DE2）以及模型中微分方程解的非线性最小二乘估计（2）。

需要估计 α 和 λ 以用于两个局部增长模型。

从模型的微分方程（1）和对数变换方程的显式解（10）中，我们可以看出

$$G(x) \doteq \frac{1}{1-\alpha}(\log(1-\alpha) + \log(\lambda) + \log(x)), \quad \alpha \neq 1$$

因为在这个应用中，$g(0)$ 必然是一个非常小的值。这意味着模型 $\log(y)$ 与 $\log(x)$ 的简单线性回归斜率估计量是 $1/(1-\alpha)$ 的一个估计量。我们使用此来估计 α。根据这一估计，然后通过将非线性最小二乘法应用于模型

$$y = \{(1-\widehat{\alpha})(\lambda x)\}^{1/(1-\widehat{\alpha})}.$$

来进行对 λ 的估计。这再次基于对在 (10) 处给出的显式 DE 解的近似。

我们设计了两组人为稀疏的数据：一种情况是去除数据点 5、6、7 和 8，另一种情况是去除数据点 4、5、6、7 和 8。通过 200 次模拟，计算了 $\widehat{G}(x_i)$ 与 $\widehat{G}_{\ell\ell}(x_i)$ 之间的平方差对于 $i = 5, \ldots, 8$ 或 $i = 4, \ldots, 8$，并对六种估计方法进行了平均。这些平均平方差的平均值列在表 1中。平方差是在原始尺度和对数尺度上计算的。表 1中的数值表明，两个局部拟指数增长模型与局部多项式回归方法具有更大的兼容性，特别是在数据高度稀疏且人为去除额外的数据点时更是如此。局部线性模型也享有相对较小的平均平方误差，但由于基础数据遵循线性模型，因此存在轻微偏向局部线性模型的偏差。局部二次回归表现较差，而非线性增长模型的误差比较低阶的局部近似要大得多。这表明建议的局部增长模型可能并不真正适合这些数据。然而，总体信息是 DE 模型可以指导核回归方法达到满意的估计。

[表 1 about here.]

图 1比较了移除数据点 5、6、7 和 8 后来自各种回归模型的拟合曲线。一阶和二阶局部准指数增长模型的拟合曲线比局部多项式回归模型的曲线更为平滑，特别是在数据稀疏区域，这表明在有限观测下捕捉潜在趋势的表现有所提升。

[图 1 about here.]

5. 讨论

关于微分方程的信息可以改善局部多项式回归估计。所提出的方法简单且计算负载低，无需求解微分方程。一阶和二阶 DE 约束回归优于局部常数和局部二次回归。当处理稀疏区域时，DE 约束方法也与局部线性回归具有竞争力。

本文重点研究了一阶 DE 约束回归模型，该模型涉及一阶微分方程。此处考虑的模型是一阶非线性微分方程，并具有显式解。本文采用的方法不需要知道这个解；它可以推广到许多其他情况。在未来，我们将探讨更高阶的模型，例如二阶 DE 约束回归模型。

致谢

此项研究部分得到了加拿大自然科学与工程研究委员会（NSERC）的资助。

参考文献

Ding, A. A., & Wu, H. (2014). Estimation of ordinary differential equation parameters using constrained local polynomial regression. Statistica Sinica, 24(4), 1613.

Fan J. and Gijbels I. (1996). Local Polynomial Modelling and Its Applications, volume 66. CRC.

Hirsch, M. W., Devaney, R. L., and Smale, S. (1974). Differential equations, dynamical systems, and linear algebra (Vol. 60). Academic press.

Plume, C. A., Daly, S. E., Porter, A. T., Barnett, R. B., & Battista, J. J. (1993). The relative biological effectiveness of ytterbium-169 for low dose rate irradiation of cultured mammalian cells. International Journal of Radiation Oncology* Biology* Physics, 25(5), 835-840.

Wand, M., & Ripley, B. (2015). KernSmooth: Functions for kernel smoothing supporting Wand & Jones (1995). R package version 2.23-15. MR1319818.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The annals of statistics*, 1346-1370.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association, 90(432), 1257-1270.

Taylor, B. (1717). Methodus incrementorum directa and inversa. Inny.

Appendix

表格和图形

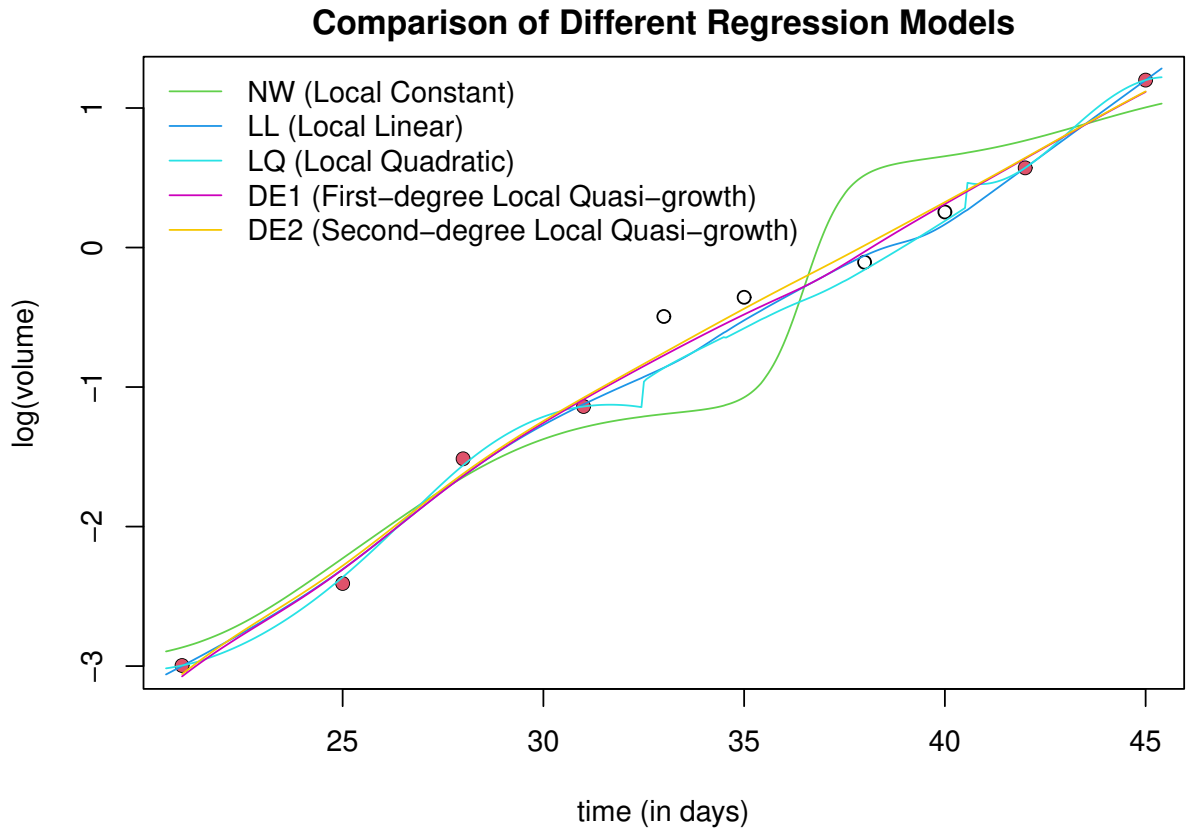

Figure 1. 不同回归模型在稀疏肿瘤数据中去除点 5、6、7 和 8 后的拟合曲线。

	log scale	original scale
NW	0.265	0.300
LL	0.016	0.014
LQ	0.024	0.021
DE1	0.017	0.014
DE2	0.019	0.023
NLS	0.181	0.045

	log scale	original scale
NW	0.443	0.352
LL	0.037	0.023
LQ	0.091	0.161
DE1	0.027	0.016
DE2	0.018	0.020
NLS	0.508	0.068

Table 1

两种不同稀疏设计的平均平方误差汇总：移除数据点 *5*、*6*、*7* 和 *8*（左）以及移除数据点 *4*、*5*、*6*、*7* 和 *8*（右）每种建模方法的结果：*NW*（局部常数）、*LL*（局部线性）、*LQ*（局部二次）、*DE1*（一阶局部准增长）、*DE2*（二阶局部准增长）以及 *NLS*（非线性最小二乘）。”对数尺度”列中的误差基于拟合值与观测值在对数尺度上的差异，而”原始尺度”列中的误差则基于指数化后的拟合值与原始观测值之间的差异。