中国科学 数学
SCIENTIA SINICA Mathematica

Nonlinear regression in COVID-19 forecasting

崔恒建 and 胡涛

Citation: 中国科学: 数学; doi: 10.1360/SSM-2020-0055

View online: http://engine.scichina.com/doi/10.1360/SSM-2020-0055

Published by the 《中国科学》杂志社

中国科学: 数学 2020 年 第 50 卷 第 ? 期: 1 ∼ 12
SCIENTIA SINICA Mathematica
论 文

新型冠状病毒肺炎疫情预测预报的非线性回归方法

崔恒建[1,2*], 胡涛[1,2]

1 首都师范大学数学科学学院, 北京 100048;
2 首都师范大学交叉科学研究院, 北京 100048
E-mail: hjcui@bnu.edu.cn, hutaomath@foxmail.com

收稿日期: 2020-02-27; 接受日期: 2020-03-05; 网络出版日期: 2020-03-12; * 通信作者
国家自然科学基金 (批准号: 11971324 和 11471223)、北京市科技创新平台建设 (批准号: 19530050181)、首都师范大学交叉科学研究院和生物统计交叉学科资助项目

摘要 本文介绍几种累计新型冠状病毒肺炎 (COVID-19) 疫情预测预报中的非线性增长曲线, 并说明 Richards 增长曲线在这次 COVID-19 疫情预测预报中的合理性和可行性; 在此基础上, 建立累计 COVID-19 疫情预测预报中的非线性回归点模型, 并给出参数估计方法; 对全国 COVID-19 疫情进行即时跟踪预测预报, 包括数据校准、整体和分时间段的预测预报, 同时获得全国 COVID-19 疫情随时间的预测预报结果, 为进一步的疫情防控打下良好基础.

关键词 新型冠状病毒肺炎 非线性回归模型 Richards 曲线 数据校准

MSC (2010) 主题分类 62J02, 62F10

1 引言

2019 年 12 月至 2020 年 1 月初, 我国湖北省武汉市陆续发现了多例新型冠状病毒感染的肺炎患者 (世界卫生组织将此新型冠状病毒命名为 SARS-CoV-2, 由此新型冠状病毒感染的肺炎命名为 COVID-19), 随着疫情的发展, 我国其他地区及境外也陆续发现了此类病例. 现已将该病纳入《中华人民共和国传染病防治法》规定的乙类传染病, 并采取甲类传染病的预防、控制措施. 引发 COVID-19 的新型冠状病毒 SARS-CoV-2 属于 SARS 的进化病毒, 以呼吸道飞沫和密切接触为主要的传播途径, 病毒进入人体后干扰免疫系统, 让免疫系统误认为肺细胞是外来物从而发起进攻, 使得人体的部分器官功能受损. 目前一般采用激素压制免疫系统, 但这会使人的身体变得很脆弱. SARS-CoV-2 在某些方面比 SARS 对人体的危害更大, 具体表现在: (1) 潜伏期更长; (2) 存在极少数患者从感染到发病, 再到死亡, 体温始终正常, 甚至无任何症状; (3) 具有人传人、传播速度更快、传播能力更强的特点, 目前只获得呼吸道飞沫和密切接触传播的证据, 不排除还有其他途径. COVID-19 是具有很强传染性的严重急性

英文引用格式: Cui H J, Hu T. Nonlinear regression in COVID-19 forecasting (in Chinese). Sci Sin Math, 2020, 50: 1–12, doi: 10.1360/SSM-2020-0055

呼吸道疾病, 目前还没有特效药物来治疗该传染病. 这一疫情向全国扩散, 加之在春节前后, 人群密集, 又进一步加大扩散的风险, 给人民的身体健康和生命安全带来严重威胁. 截至 2020 年 3 月 3 日, 国内已报告 8 万余例确诊病例和 2,900 多例死亡病例, 全球除中国外已报告超过 10,000 感染病例. 据百度 (https://voice.baidu.com/act/newpneumonia/newpneumonia/?from=osari_aladin_top1) 和丁香园 (http://ncov.dxy.cn/ncovh5/view/pneumonia) 网站上 COVID-19 疫情实时大数据报告, 全国以及湖北的疫情趋势如图 1 所示. 图中横轴有两个时间点非常重要, 这两个时间点分别是 2020 年 1 月 23 日和 2020 年 2 月 12 日. 从第一个关键日期 1 月 23 日起, 为遏制疫情蔓延, 中国政府采取了大规模检疫、限制出行和对可疑病例监控等前所未有的全国性防疫干预措施, 这些政策措施对疫情传播规模的控制起着重要作用. 为了进一步缓解人们的恐慌情绪, 采取客观冷静的态度, 科学地认识疫情的发展和有效地进行控制与防治, 我们有必要对全国 COVID-19 疫情数据进行科学分析和认识, 其中利用流行病建模方法来了解疫情发展规模是人们常用的方法之一.

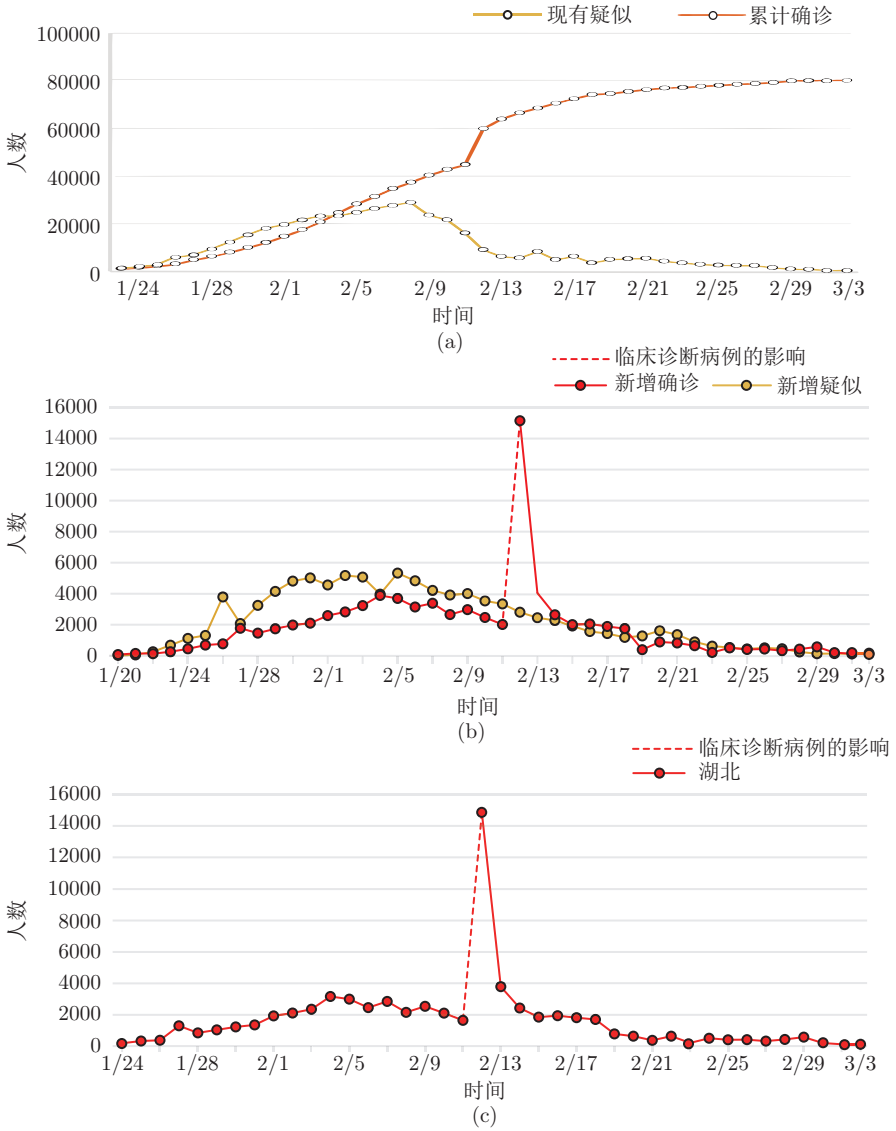

图 1 **(a) 全国累计确诊病例和疑似病例; (b) 全国新增确诊病例和疑似病例; (c) 湖北新增确诊病例**

在疫情蔓延早期, Wu 等[1] 预测武汉的疫情规模将在 1 月 25 日达到 75,185 例. 同时, Read 等[2] 预测, 在没有控制措施的情况下, COVID-19 将在 2 月 4 日达到高峰, 高达 19 万例. 这些关于 COVID-19 疫情规模的预测现在看来都与疫情进展差距很大. 近来, Yang 等[3] 通过优化的易感 - 暴露 - 感染 - 退出 (SEIR) 模型和人工智能 (AI) 方法预测国内 COVID-19 疫情的总规模将达到 73,180 例 (95% 置信区间为 51,308–85,839). 严阅等[4] 基于全国各级卫生健康委员会每日公布的 COVID-19 累计确诊数和治愈数, 提出了一类基于时滞动力学系统的传染病动力学模型并反演了模型参数, 有效地模拟了目前疫情的发展并预测了疫情未来的趋势. 在另一个关键日期 2 月 12 日, 湖北省卫生健康委员会根据影像学检查、中性粒细胞计数和流行病学关联等把确诊病例归为临床诊断确诊病例, 使得一夜之间增加了 1.6 万多病例, 致使全国 COVID-19 确诊病例数出现了大的波动, 同时也对疫情的统计建模提出了挑战. 另外, 许多学者对 COVID-19 流行病学的其他方面进行了研究, 例如, Li 等[5] 收集了 2020 年 1 月 22 日前武汉市 COVID-19 前 425 例确诊病例的数据, 估计了关键的流行病学延时分布、指数增长初期 COVID-19 的流行倍增时间和基本再生数. Backer 等[6] 根据疫情早期 (2020 年 1 月底) 在武汉以外地区发现的 88 例确诊病例研究平均潜伏期, 为确定合理适当的隔离期限提供了初步依据. Guan 等[7] 提取了截至 2020 年 1 月底全国 1,099 例实验室确诊的 COVID-19 患者的资料进行了描述性统计分析, 并指出 COVID-19 通过人际传播迅速传播, 中位潜伏期 3 天, 病亡率相对较低; 有些患者放射学检查结果正常, 可能不发热, 腹泻少见.

综上所述, 本文从统计非线性回归建模的角度来预测 COVID-19 的发展规模, 具体聚焦如下问题的研究: (1) 用非线性 Richards 增长曲线模型来研究 COVID-19 疫情的规模, 包括累计病例和累计病亡人数等的预测预报, 研究模型中的参数估计问题; (2) 通过统计非线性回归建模来说明 1 月 23 日起实施的武汉封城对疫情规模的影响; (3) 采用以核光滑为主要工具的数据校准方法来克服 2 月 12 日数据调整的影响, 这对疫情数据进行合理准确的预报十分有利. 截至 3 月 3 日, 通过分析 (本文采用的 COVID-19 疫情数据均来自国家卫生健康委员会网站 http://www.nhc.gov.cn/xcs/yqtb/202002/4a611bc7fa20411f8ba1c8084426c0d4.shtml), 我们获得如下结论.

(1) 全国疫情规模、累计确诊病例及其病亡率置信上限: 累计确诊病例总规模点估计为 81,127, 1σ (85%) 单边置信上限为 83,819, 2σ (95%) 单边置信上限为 84,915. 疫情整体态势趋于稳定, 期待 3 月中旬 (估计 3 月 7 日后) 新增病例降至两位数, 并有望在 3 月底 4 月初新增病例降至个位数. 病亡率点估计为 4.1%, 1σ (85%) 单边置信上限为 4.7%, 2σ (95%) 单边置信上限为 5.2%.

(2) 封城政策的影响: 武汉封城一周内累计病例呈指数级数增长, 对数累计病例的线性斜率稳定在 0.3652; 一周后, 指数增长趋势变缓, 说明封城政策有效地阻止了疫情的指数级传播蔓延.

(3) 数据的校准与调整能较充分地揭示疫情在改变检测方式前后的发展规律, 并使我们发现武汉封城两周后新增病例达到峰值.

2 COVID-19 预测预报非线性模型及其机理简介

自 COVID-19 从我国湖北省武汉市向全国及全世界蔓延时, 我们就对世界和我国的 COVID-19 疫情的数据, 特别是在 1 月 23 日武汉封城后疫情的发展数据, 进行了即时的统计分析, 并通过分别观察全国 (包括武汉、湖北等) 的累计病例、新增病例的走势, 提出了用累积增长的 Richards S 型曲线去拟合和预测 COVID-19 累计病例的发展趋势, 在此基础上, 我们自然用 Richards 增长曲线的差分拟合和预测新增病例. 之所以选用 Richards 增长曲线进行累计病例的拟合与预测, 是因为, 其一, 传染病的

累计病例通常遵循 S 型曲线增长, 而 Richards 增长曲线是被广泛采用的具有 4 参数的 S 型增长曲线 (也称广义逻辑 (logistic) 曲线); 其二, 在 2003 年预测预报北京和全国 SARS 累计病例数中, 我们就使用了 Richards 增长曲线模型, 拟合和预测效果很好, 参见文献 [8,9]; 其三, Richards 增长曲线模型用于描述传染病的传染增长机理较为清楚, 下面进行简单介绍.

由于人们已认识到 COVID-19 是一种传染性很强的传染病, 且具有聚集性传染的特点, 随着人们日常交往的频繁和城市人口的密集, 假设它的传染力大于 0, 即如果 $A = A(t_0)$ 为从时刻 t_0 开始的最终感染人数, $I(t_0) \approx I_0$ 和 $I(t)$ 分别为 t_0 和 t ($t \geqslant t_0$) 时刻的累计病例数, 令 $y = [I(t) - I(t_0)]/A$, 则 y 满足如下增长率方程:

$$\frac{d\log(y)}{dt} = f(y, t - t_0), \tag{2.1}$$

其中 $f(y,t)$ 是形式已知的关于 $0 < y < 1$ 和 t 的连续函数. 常见的 $f(y,t)$ 是如下 Terner- 型函数:

$$f(y,t;p,m,\lambda) = \lambda(1-y^m)^{1-p}(y^{-m}-1)^p, \tag{2.2}$$

其中 $\lambda, m > 0$ (参见文献 [10,11]). 当取 $p = 0$, $m = 1$ 时, 有 $\frac{d\log(y)}{dt} = \lambda(1-y)$, 即有

$$I(t) \approx I(t_0) + \frac{A}{1+\exp\{-K(t-t_0)+b\}},$$

即为人们熟知的逻辑 (logistic) 增长模型; 当 $p = 0$ (或 $p = 1$), $m > 0$ 时, 即为 Richards 增长方程. 而 $p = 0$ 时的 Richards 方程具有以下特点:

(1) 它是 logistic 增长模型 ($m = 1$) 的自然推广, 且当 $t \sim t_0$, $y \sim 0$ 时, $\frac{d\log(y)}{dt} \approx \lambda$, 因此, $y \approx \exp\{\lambda(t-t_0)+b\}$, 即为指数函数;

(2) 它有如下形式的显式解:

$$I(t) = I(t_0) + Ay = I(t_0) + A[1+\exp\{-K(t-t_0)+b\}]^{-B}, \tag{2.3}$$

其中参数 A 为增长量的极限值, B 为曲线形状参数, K 为增长速率, b 为初始值参数.

(3) 遵循 “负反馈” 微分方程:

$$\frac{dI}{dt} = \alpha I - \beta I^{m+1}, \tag{2.4}$$

其中 $\alpha > 0$, $\beta > 0$, 且满足 $A = (\alpha/\beta)^{1/m}$, $B = 1/m$, $K = \alpha m$. Richards 增长曲线 (2.3) 见图 2.

注意到这次武汉 COVID-19 疫情在开始时, 具有聚集性传染、来势凶猛的特点, 且累计病例 I 基本是以指数速度自然增长的, 这从武汉封城一周内累计暴发数据就可以看出. COVID-19 与其他通常传染病相比, 有不同的特点, 它传染力很强, 潜伏期长, 致死率相对较低. 人们起初对它认识不清楚, 因而认为它起初感染的增长遵循 “负反馈方程” 中的第一项 αI, 即对数线性的速度增长, 而随着疫情的发展, 人们对它有所认识, 病亡、病愈、抗体人群的产生以及采取的严格控制措施, 对感染的增长作用速率认为是第二项 $-\beta I^{m+1}$, 因而, 其累计病例增长数应近似遵循增长方程 (2.3). 由于 S 型曲线 (2.3) 含有 4 参数 A、B、K 和 b, 因此具有整体拟合与预测的灵活性, 其参数估计可借助线性回归计算方法快捷获得 (参见第 3 节).

鉴于以上分析, 对照 S 型曲线 (2.3) 的性质和这次 COVID-19 疫情的特点, 本文选择 Richards 非线性增长曲线 (2.3) 作为非线性回归方程来拟合全国 COVID-19 的累计病例 (或累计病亡人数) 数据.

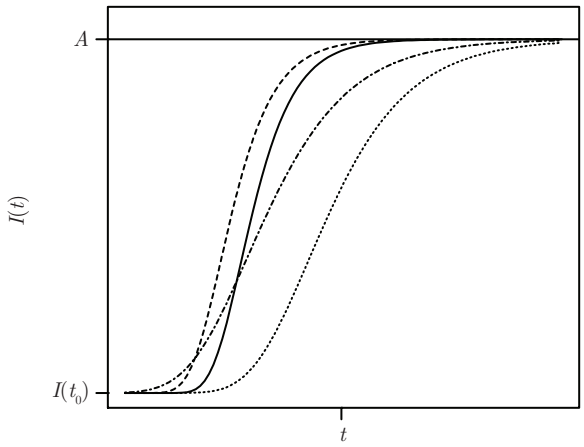

图 2 不同参数 $B>0$ 和 $K>0$ 下的 Richards 增长曲线

3 非线性 Richards 回归模型及其参数的估计方法

3.1 非线性 Richards 回归模型

对于 COVID-19 的累计确诊病例, 我们建立如下的非线性 Richards 回归模型:

$$\begin{aligned}\mathbf{I}(t) &= I(t)+\epsilon(t)\\ &= I(t_0)+A[1+\exp\{-K(t-t_0)+b\}]^{-B}+\epsilon(t),\end{aligned}\tag{3.1}$$

其中 $\epsilon(t)$ 为随机误差, 满足 $\mathrm{E}(\epsilon(t))=0$. $\mathbf{I}(t)$ 为 t 时刻可观测到的确诊病例累计量, $I(t_0)$ 为初始确定值.

3.2 线性化与光滑化参数估计

下面给出这个参数的估计方法. 不妨假设 $t_0=0$, $I(t_0)=0$, 先改写 $I(t)$ 的形式:

$$\frac{dI(t)}{dt}=ABK[1+\exp\{-Kt+b\}]^{-(B+1)}\exp\{-Kt+b\}=(A^{-\frac{1}{B}}BK\mathrm{e}^b)I(t)^{1+\frac{1}{B}}\exp\{-Kt+b\}.$$

于是,

$$\log\left(\frac{1}{I(t)}\frac{dI(t)}{dt}\right)=\log(A^{-\frac{1}{B}}BK\mathrm{e}^b)+\frac{1}{B}\log(I(t))-Kt=:\beta_0+\beta_1\log(I(t))+\beta_2 t,\tag{3.2}$$

其中参数 β_0、β_1 和 β_2 满足关系

$$\beta_0=\log(A^{-\frac{1}{B}}BK\mathrm{e}^b),\quad \beta_1=\frac{1}{B},\quad \beta_2=-K.$$

下面的问题就转化成参数 β_0、β_1 和 β_2 的估计问题. 显然, 我们可以用 t_i 时刻的累计病例观测值 $\mathbf{I}(t_i)$ 来替代 $I(t_i)$, 用差分来逼近微分. 为了进一步降低数据的误差, 采用适当光滑化 (降噪方法, 参见文献 [12]) 的方法, 也就是说, 给定一个光滑权重函数 $w_n(t,t_j)$, $1\leqslant j\leqslant n$ (本文采用核权重光滑函数), 令

$$\boldsymbol{\Delta}(t):=\sum_{j=1}^{n-1}w_{n-1}(t,t_j)\frac{\mathbf{I}(t_{j+1})-\mathbf{I}(t_j)}{(t_{j+1}-t_j)\mathbf{I}(t_j)}\approx\frac{1}{I(t)}\frac{dI(t)}{dt}.\tag{3.3}$$

这样由非线性回归模型 (3.1) 出发, 就近似转化成如下数据型线性回归模型:

$$\boldsymbol{\Delta}(t_i) = \beta_0 + \beta_1 \log(\mathbf{I}(t_i)) + \beta_2 t_i + e_i, \quad 1 \leqslant i \leqslant n-1. \tag{3.4}$$

记 $\mathbf{1} = (1,1,\ldots,1)^{\mathrm{T}} \in \mathbb{R}^{n-1}$; $\boldsymbol{Y}_n = (\log(\boldsymbol{\Delta}(t_1)), \log(\boldsymbol{\Delta}(t_2)), \ldots, \log(\boldsymbol{\Delta}(t_{n-1})))^{\mathrm{T}}$; $\boldsymbol{X}_{1n} = (\log(\mathbf{I}(t_1)), \log(\mathbf{I}(t_2)), \ldots, \log(\mathbf{I}(t_{n-1})))^{\mathrm{T}}$; $\boldsymbol{X}_{2n} = (t_1, t_2, \ldots, t_{n-1})^{\mathrm{T}}$, $\boldsymbol{X}_n = (\mathbf{1}, \boldsymbol{X}_{1n}, \boldsymbol{X}_{2n})$; $\boldsymbol{e}_n = (e_1, e_2, \ldots, e_{n-1})^{\mathrm{T}}$; $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$, 因此, (3.4) 可写成

$$\boldsymbol{Y}_n = \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{e}_n. \tag{3.5}$$

借助于线性模型的最小二乘估计, 即 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_n^{\mathrm{T}} \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^{\mathrm{T}} \boldsymbol{Y}_n$ 来估计 $\boldsymbol{\beta}$. 由 3 个线性回归系数要解得参数 A、B、K 和 b, 还需加一约束, 于是, 约束曲线 $I(t)$ 经过当前时间点 $(t_n, \mathbf{I}(t_n))$, 这样可解得

$$\hat{A} = \frac{\mathbf{I}(t_n)}{1 - \frac{\mathbf{I}(t_n)^{1/\hat{B}}}{\hat{K}\hat{B}} \exp\{-\hat{K} t_n + \hat{\beta}_0\}},$$

其中 $\hat{K} = -\hat{\beta}_2$, $\hat{B} = 1/\hat{\beta}_1$. 一般情形下, β_0 和 K 的估计相对较稳定. 如果我们需要调整 A 和 B, 则可固定 $K = \hat{K}$, $\beta_0 = \hat{\beta}_0$, 以 A 和 B 的最小二乘估计为初值, 代回到方程 (3.1), 使得非线性回归残差平方和

$$\sum_{i=1}^{n} \left(\mathbf{I}(t_i) - I_0 - A \left[1 + \frac{A^{1/B}}{B\hat{K}} \exp\{-\hat{K} t_i + \hat{\beta}_0\} \right]^{-B} \right)^2$$

达到最小而获得参数估计 A 和 B 的调整估计.

注意到, 在 Richards 增长模型的前提下, 由于观测数据 (累计病例、新增病例等) 都带有误差, 处理这样的非线性数据拟合问题通常采用最小二乘方法, 即

$$\min \sum_{i=1}^{n} (\mathbf{I}(t_i) - I_0 - A[1 + \exp\{-K t_i + b\}]^{-B})^2.$$

但是, 这个非线性优化问题对初值比较敏感, 经常陷入局部最优化, 且计算速度较慢, 拟合效果有时也不尽人意, 其原因是数据有时误差较大 (尤其刚开始的数据), 规律不明显. 为此, 我们将非线性模型进行 “线性化” 处理使之变成线性模型 (3.2), 其目的是借助于线性模型快速稳定的算法获得回归参数 β 的估计. 同时, 为了降低数据误差, 对 (3.2) 左边因变量对数函数中对应的观测数据进行光滑化 (去噪) 处理, 因此形成了可计算的线性模型 (3.5). 这种参数估计方法稳定性好, 计算快速, 整体拟合效果好. 此外还有通过固定点法 (称为四点法) 和分段拟合法 (称为三段法) 等方法, 这些方法对点或段的选择要求较高, 对有误差的数据, 拟合误差难以控制. 比较几种方法, 以转化线性模型的方法参数估计效果最稳定, 整体拟合误差最小, 具体比较可参见文献 [13].

4 全国 COVID-19 疫情数据的预测预报

以上导出的非线性 Richards 增长曲线模型 (3.1) 和参数估计方法具有模型简单、机理清楚和算法简明快速的特点. 崔恒建等[8] 已成功地将 Richards 增长曲线模型用于 2003 年 SARS 疫情的预测预报, 所以, 我们继续利用 Richards 增长曲线模型对全国 COVID-19 疫情进行预测预报, 包括动态的即时预报 (5 天以内) 和中长期预报, 给出新增确诊病例数和累计确诊病例数等的预报及预报误差范围. 依据国家卫生健康委员会网站 http://www.nhc.gov.cn/xcs/yqtb/202002/4a611bc7fa20411f8ba1c8084426c0d4.shtml 上公布的疫情数据, 可以计算模型 (3.1) 中的参数 A、B 和 K, 以及这些参数随时间变化的规律.

4.1 武汉封城一周内的累计确诊病例数的对数线性增长态势预测预报

全国的累计确诊 (临床) 病例模型参数与预测见表 1, 总体上来看其预测预报结果在武汉封城 (1 月 23 日) 一周左右基本上遵循对数指数的暴发的发展态势. 我们把武汉封城前 3 天和后一周内 (1 月 20 日至 1 月 30 日, 共 11 天) 的累计确诊病例数据进行对数变换, 从而明显地看到随时间的变化规律, 即有

$$\log(\mathbf{I}(t_i)) = b + kt_i + \epsilon_i, \tag{4.1}$$

$j = 6, 7, \ldots, 16$, 其中 $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, 参见图 3. 1 月 20 日后 (1 月 14 日对应 $t_0 = 0$), 武汉封城一周内第 i 天的对数累计确诊病例数据集 $\{\mathbf{I}(t_j), 6 \leqslant t_j \leqslant 9 + i\}$ 与时间 $\{t_j, 6 \leqslant t_j \leqslant 9 + i\}$ 的相关系数 $(1 \leqslant i \leqslant 7)$ 见表 2.

表 1 全国 COVID-19 疫情拟合预测结果及参数 A 的变化

日期	A	次日预测 2σ 置信区间	次日预测 (实测)	5 天预测	1 周拟合标准差
1 月 31 日	212,859	(14,338, 16,113)	15,225 (14,380)	35,410	444
2 月 1 日	93,773	(16,946, 18,677)	17,812 (17,205)	34,609	433
2 月 2 日	69,187	(19,791, 21,463)	20,627 (20,428)	35,334	418
2 月 3 日	67,221	(23,105, 24,841)	23,973 (24,324)	38,267	434
2 月 4 日	79,792	(27,149, 29,362)	28,255 (28,018)	44,107	553
2 月 5 日	80,604	(31,007, 33,039)	32,023 (31,161)	47,549	507
2 月 6 日	71,594	(34,144, 35,588)	34,866 (34,546)	48,202	361
2 月 7 日	69,695	(37,460, 38,671)	38,066 (37,198)	50,191	303
2 月 8 日	63,583	(39,663, 40,863)	40,263 (40,171)	50,142	300
2 月 9 日	63,317	(42,285, 43,705)	42,995 (42,638)	51,836	355
2 月 10 日	61,803	(44,412, 45,850)	45,131 (44,653)	52,670	359
2 月 11 日	59,949	(46,139, 47,423)	46,781 (59,804*)	52,990	321

注: 2σ 置信区间为近似 95% 置信区间; $*$ 改变确诊 (临床) 方法后的数据

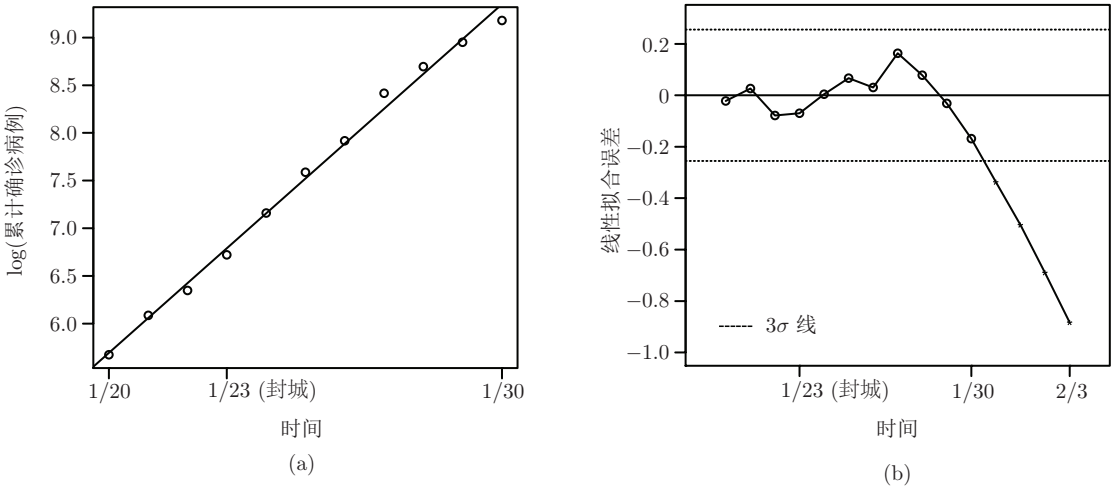

图 3 (a) log(累计确诊病例) 的线性拟合; (b) log(累计确诊病例) 线性拟合误差

表 2 封城一周内 log (累计确诊病例) 与时间的相关系数变化

封城后的天数	1	2	3	4	5	6	7
相关系数	0.9969	0.9973	0.9983	0.9982	0.9987	0.9984	0.9973

可以看出, 相关系数在武汉封城一周内比较稳定, 表明其累计病例确实是呈指数级数增长. 对数累计病例的线性斜率和误差标准差的估计分别为 $\hat{k} = 0.3652$ 和 $\hat{\sigma} = 0.0851$. 从图 3(b) 中可以看出, 1 月 31 日后误差已突破 3σ 下限, 表明 log(累计确诊病例) 将不再以 $k = 0.3652$ 的斜率增长, 斜率渐渐变小, 指数增长趋势变缓, 由于 SARS-CoV-2 的潜伏期一般认为平均为 1 周左右, 这也说明武汉封城一周后, 有效地阻止了疫情的指数级传播蔓延.

4.2 封城两周改变检测方法后的数据校准及其预测预报

封城一周后虽然累计病例的指数级增长有所减缓, 但累计病例数据却开始艰难地 “爬坡”, 即近似线性增长趋势, 由于此时基数增大, 线性斜率 “爬坡” 较大, 给防控带来了巨大压力. 因此, 及时预测预报对人们消除紧张恐惧心理具有重要作用. 这期间采用非线性回归模型 (3.1) 对 COVID-19 全国累计病例进行预测预报, 2 月 1 日得到的预测上限是 94,000 左右, 见表 1. 当时的这一预测对消除当时人们的恐惧心理, 为政府部门制定进一步的相关决策是至关重要的, 即使从现在来看, 也是合理可行的. 从模型预测预报的整体看, 模型对数据的跟踪效果是很好的, 一周内的拟合误差均控制在 5% 内. 次日预测效果较好, 但从 5 天的预测数据来看, 实测数据明显偏低于预测值, 例如, 在 2 月 5 日预测 2 月 10 日的值相差近 5,000.

这次 SARS-CoV-2 的检测受到试剂盒数量的很大限制, 特别是武汉封城一周后这个问题尤为突出, 有许多患者因得不到及时检测而不能有效确诊, 患者救治和隔离均受到一定程度的影响, 确诊病例滞后于真正发病的实际人数, 因此在数据上就出现了比较大的低估, 从表 1 中 “5 日预测” 的数据也看到这一现象. 到了封城 3 周时 (2 月 12 日), 把用试剂盒测核酸的方法改成用临床方法 (包括 CT 方法) 确诊, 使得累计确诊病例突然增多, 这对病人的及时救治和隔离带来了好处, 但另一方面, 这可能会导致 “假阳性” 现象的出现, 从后面模型的预测上可看到统计数据的变化.

武汉封城 1 周后的模型 “5 日预测” 的高估是合理的, 因为疫情指数暴发, 而试剂盒受限, 导致许多患者没能及时确诊, 因此, 这部分实际数据是没有的, 造成数据低估. 对此我们有必要对原来的数据, 特别是改变确诊方法时前一周 (封城第 3 周) 的数据使用我们的模型进行校准. 我们选择采用 2 月 4 日的拟合模型 (见图 4) 进行校准, 此模型比较接近 2 月 12 日变化的数据. 以此模型为依据, 并调整使得该增长曲线模型经过点 (29, 59,804). 这样可用这个曲线对相应时间 (2 月 5 日至 2 月 11 日) 的数据 (共 7 天) 进行插值, 见表 3.

4.3 数据校准后的模型拟合与预测预报

有了经过模型校准的数据, 就可以对接下来的疫情继续使用我们的模型进行拟合和预测预报, 表 4 给出校准数据后全国整体 COVID-19 疫情的预测结果. 注意到 2 月 18 日的次日预测结果, 实测值没有在预测值的 2σ 置信区间内, 其原因是 2 月 19 日实行了订正核减数据, 就是对确诊病例中来源于原 “临床诊断病例” 者进行核酸检测, 通过综合分析将核酸检测结果为阴性的病例从确诊病例中核减, 共核减 279 例. 统计模型就是对统计数据规律的探索与跟踪, 如果数据打破了通常规律, 则会在模型的预测或拟合过程中有所体现. 截至 3 月 3 日, 3 周内的拟合相对误差控制在 2% 内, 经过计算获得全

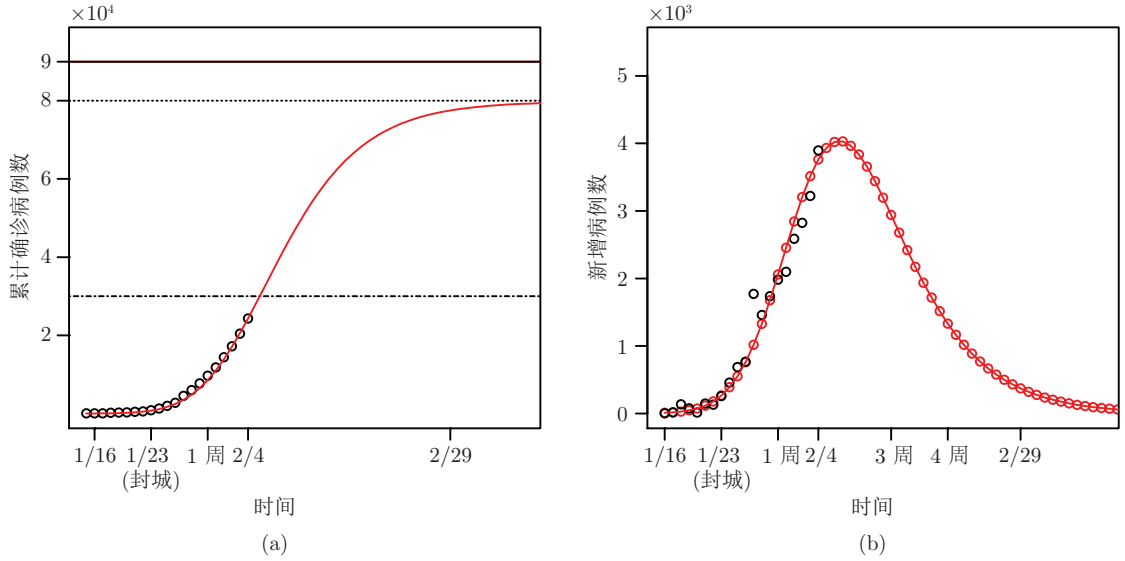

图 4 **(a) 2 月 4 日全国累计确诊病例的拟合与预测; (b) 2 月 4 日全国新增确诊病例的拟合与预测**

表 3 **累计确诊病例的校准数据**

日期	实测数据	校准数据	校准差
2 月 5 日	28,018	31,364	3,346
2 月 6 日	31,161	35,827	4,666
2 月 7 日	34,546	40,299	5,753
2 月 8 日	37,198	44,701	7,503
2 月 9 日	40,171	48,960	8,789
2 月 10 日	42,638	53,020	10,382
2 月 11 日	44,653	56,839	12,186

国累计确诊病例总规模 (极限参数 A) 的点估计为 81,127, 1σ (85%) 单边置信上限为 83,819, 2σ (95%) 单边置信上限为 84,915. 其整体发展态势见图 5. 其中过去 1 周内和将来 1 周内的拟合相对误差和累计确诊病例预测见表 5.

结合图 5 可以看到, 武汉封城两周后新增病例达到峰值, 这恰恰是对数据进行校准后看出的结果, 它较充分地揭示了疫情在改变检测方式前后的发展规律. 再结合表 5 可得出, 疫情整体态势趋于稳定, 期待 3 月中旬 (估计 3 月 7 日后) 新增病例降至两位数, 并有望在 3 月底或 4 月初新增病例降至个位数.

4.4 累计病亡人数模型拟合与预测及病亡率估计

累计病亡人数是指到当前时间在 COVID-19 确诊累计病例中死亡的人数, 病亡率则是在 COVID-19 确诊总病例数中, 总死亡人数所占的比率. 截至 3 月 3 日, 全国 COVID-19 累计病亡人数见图 6. 同样用 Richards 增长曲线模型进行拟合和同样的参数估计方法, 经计算获得结果如下: 全国确诊病例累计病亡人数 (极限参数 A_d) 的点估计 $\hat{A}_d$ 为 3,358, 1σ (85%) 单边置信上限为 3,848, 2σ (95%) 单边置信上限为 4,203. 我们采用 $\hat{r}_d = \frac{\hat{A}_d}{\hat{A}}$ 作为病亡率 r_d 的点估计. 经过计算, 获得结果如下: 点估计 $\hat{r}_d = 3358/81127 = 4.1\%$, r_d 的 1σ (85%) 单边置信上限为 4.7%, 2σ (95%) 单边置信上限为 5.2%.

表 4 全国 COVID-19 疫情累计病例拟合预测结果及参数 A 变化

日期	A	次日预测 2σ 置信区间	次日预测 (实测)	5 天预测	1 周拟合标准差
2 月 12 日	91,832	(62,108, 64,623)	63,366 (63,851)	74,735	628
2 月 13 日	94,467	(65,573, 68,940)	67,257 (66,492)	78,056	841
2 月 14 日	91,289	(68,127, 70,762)	69,445 (68,500)	78,524	658
2 月 15 日	87,753	(69,959, 71,980)	70,969 (70,548)	78,313	505
2 月 16 日	86,325	(71,652, 73,669)	72,660 (72,436)	78,810	504
2 月 17 日	85,638	(73,210, 75,300)	74,255 (74,185)	79,471	522
2 月 18 日	85,478	(74,676, 76,858)	75,767 (74,576)	80,258	545
2 月 19 日	82,495	(74,849, 76,696)	75,773 (75,465)	79,054	461
2 月 20 日	81,702	(75,464, 77,412)	76,438 (76,288)	79,062	487
2 月 21 日	81,318	(76,098, 78,080)	77,089 (76,939)	79,226	496
2 月 22 日	80,985	(76,638, 78,552)	77,595 (77,150)	79,328	479
2 月 23 日	80,111	(76,776, 78,523)	77,649 (77,658)	78,943	437
2 月 24 日	80,105	(77,114, 79,031)	78,073 (78,064)	79,144	479
2 月 25 日	80,090	(77,485, 79,333)	78,409 (78,497)	79,297	462
2 月 26 日	80,253	(78,230, 79,357)	78,794 (78,824)	79,560	282
2 月 27 日	80,325	(78,570, 79,583)	79,077 (79,251)	79,730	253
2 月 28 日	80,630	(78,973, 79,984)	79,479 (79,824)	80,072	252
2 月 29 日	81,202	(79,538, 80,548)	80,043 (80,026)	80,624	252
3 月 1 日	81,222	(79,891, 80,539)	80,215 (80,151)	80,717	162
3 月 2 日	81,163	(79,977, 80,644)	80,311 (80,270)	80,736	167
3 月 3 日	81,127	(80,070, 80,741)	80,405 (-)	80,766	168

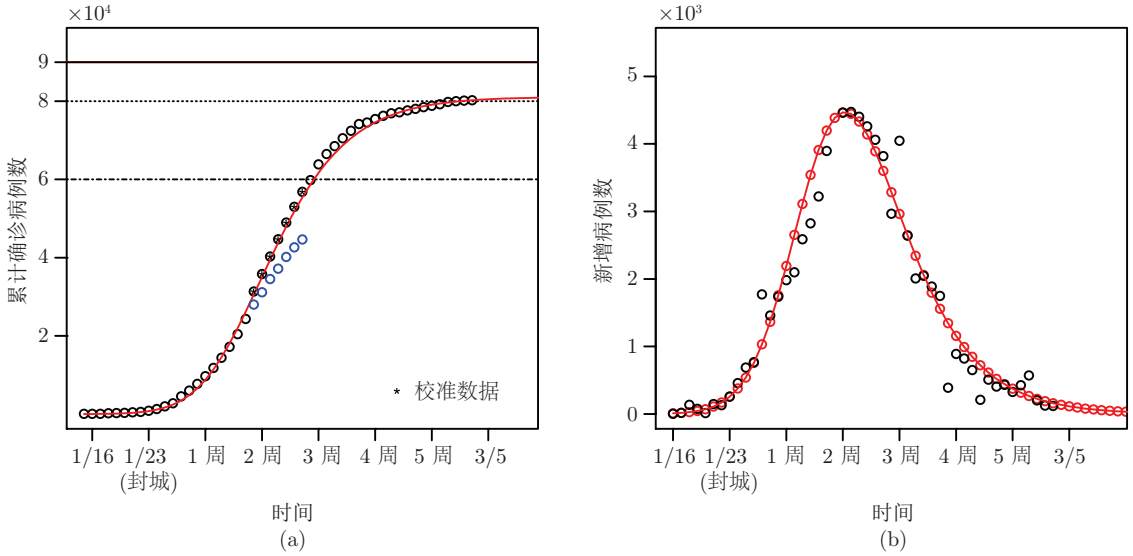

图 5 (a) 全国累计 (临床) 确诊病例拟合与预测; (b) 全国新增 (临床) 确诊病例拟合与预测

表 5 累计病例 1 周内模型预测预报与拟合结果

间隔天数	第 1 天	第 2 天	第 3 天	第 4 天	第 5 天	第 6 天	第 7 天
	3 月 4 日	3 月 5 日	3 月 6 日	3 月 7 日	3 月 8 日	3 月 9 日	3 月 10 日
未来 1 周预测预报	80,405	80,520	80,616	80,697	80,766	80,823	80,871
	3 月 2 日	3 月 1 日	2 月 29 日	2 月 28 日	2 月 27 日	2 月 26 日	2 月 25 日
过去 1 周拟合误差 (%)	−0.052	−0.134	−0.165	0.218	0.360	0.300	0.290

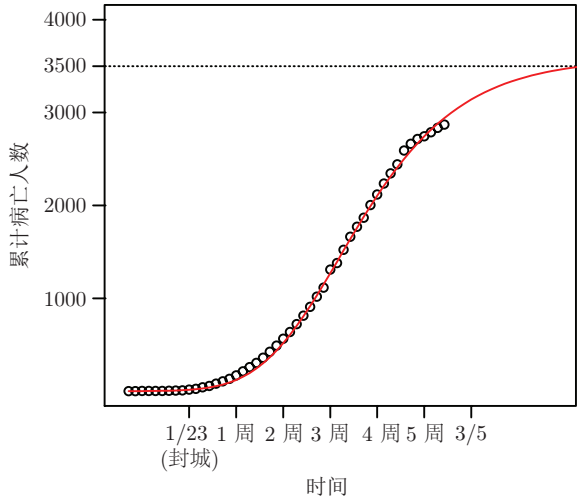

图 6 全国 COVID-19 累计病亡人数拟合与预测

虽然疫情整体趋势趋于平稳, 但疫情还在持续, 新增病例仍在以 3 位数增加, 疫情阻击战到了关键中的关键. 鉴于 SARS-CoV-2 的潜伏期长、传染力强、存在无症状患者和假阴性、聚集性传染等特点, 我们一刻也不能松懈, 必须保持高度警惕, 严格隔离措施, 保持良好心态和个人卫生, 为战胜 COVID-19 疫情贡献我们的力量.

注 4.1 本文是针对全国整体的预测预报, 但其机理和方法完全适用于其他地区或省市的预测预报, 包括武汉市、湖北省 (包括地区)、广东省、山东省、浙江省、安徽省、黑龙江省等省市的预测预报. 另外, 湖北省特别是武汉市疫情是这次 COVID-19 疫情的重中之重, 所以, 全国的疫情实际上取决于湖北省、特别是武汉市的疫情, 对全国疫情的分析, 很大程度上是对湖北省 (武汉市) 的疫情分析. 关于湖北省以外省市或地区疫情趋势的研究可能有所不同 (至少没有检测方式的改变, 没有封城带来的影响等), 这将成为我们下一步的研究任务.

致谢 本文在写作过程中得到西安交通大学徐宗本院士和北京大学耿直教授的大力支持与鼓励, 在此向他们表示衷心感谢. 向编委会及审稿人高效、快速、认真的审稿以及富有价值的审稿意见和建议表示谢意.

参考文献

1 Wu J T, Leung K, Leung G M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. Lancet, 2020, 395: 689–697
2 Read J M, Bridgen J R E, Cummings D A T, et al. Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions. BMJ, 2020, doi: 10.1101/2020.01.23.20018549
3 Yang Z, Zeng Z, Wang K, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thoracic Disease, 2020, in press
4 严阅, 陈瑜, 刘可伋, 等. 基于一类时滞动力学系统对新型冠状病毒肺炎疫情的建模和预测. 中国科学: 数学, 2020,

50: 385–392
5 Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med, 2020, doi: 10.1056/NEJMoa2001316
6 Backer J A, Klinkenberg D, Wallinga J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28, January 2020. Eurosurveillance, 2020, 25: 2000062
7 Guan W J, Ni Z Y, Hu Y, et al. Clinical characteristics of 2019 novel coronavirus infection in China. MedRxiv, 2020, doi: 10.1101/2020.02.06.20020974
8 崔恒建, 李仲来, 杨华, 等. SARS 疫情预测预报中的分段非线性回归方法. 遥感学报, 2003, 7: 245–250
9 李仲来, 崔恒建, 杨华, 等. SARS 预测的 SI 模型和分段 SI 模型. 遥感学报, 2003, 7: 345–349
10 Zeger S L, Harlow S D. Mathematical models from laws of growth to tools for biological analysis: Fifty years of "Growth". Growth, 1987, 51: 1–21
11 France J, Dijkstra J, Thornley J H M, et al. A simple but flexible growth function. Growth Development Aging, 1996, 60: 71–83
12 Cui H, Hu T. On nonlinear regression estimator with denoised variables. Comput Statist Data Anal, 2011, 55: 1137–1149
13 程毛林. Richards 模型参数估计及其模型应用. 数学的实践与认识, 2010, 40: 139–143

Nonlinear regression in COVID-19 forecasting

Hengjian Cui & Tao Hu

Abstract This paper introduces some kinds of nonlinear growth curves for forecasting cumulative COVID-19 patients. It is shown that the Richards curve is reasonable and flexible in this COVID-19 forecasting. The nonlinear growth curve regression model is established for forecasting cumulative COVID-19 patients and the parameter estimation approach for the model is also given. Specifically, the COVID-19 situation forecasting in China is made well which includes forecasting based on consecutive and piecewise time fitting. It provides a good basis for the future work.

Keywords **COVID-19, nonlinear regression model, Richards growth curve, data calibration**

MSC(2010) **62J02, 62F10**

doi: 10.1360/SSM-2020-0055