目 录

基于多项式预测中国人口规模

摘 要

人口规模是中国发展政策制定的重要指标，合理人口规模有助于平衡资源消耗，维护社会和谐。预测人口规模有助于更好的制定政策，减轻人口规模与环境、经济、社会、资源等之间的矛盾。常见的人口模型有马尔萨斯人口模型、Logistic 增长模型等，但中国人口增长受中国政策影响较大，指数不显然，故而使用多项式进行拟合。具体求解时，包括插值模型、回归模型，以及由以上两种方法优化的三次样条插值、取对数的回归模型。其中插值模型效果最差，会有龙格现象，其余几种对应相似，但在使用 2 次多项式作为指数的函数模型拟合效果非常不错，误差不到 1%，模型的稳定性也最高。最后指数为多项式的函数拟合实际上为马尔萨斯人口模型、Logistic 增长模型的推广，对应一次多项式与二次多项式作为指数的函数拟合模型，但其计算量更小，效果不弱于 Logistic 增长模型，且更加多样。

关键词：人口增长模型, 多项式, 回归模型, 插值模型, 指数函数

第 1 章　背景与初步假设

§1.1　选题依据

中国是世界第一人口大国，人口数量占比达到世界人口 20%。中国的人口数量变化对于世界人口稳定、经济发展、环境保护、资源利用等有重要影响。对中国自身而言，较为准确的预测未来中国人口变化，有助于中国人口政策的制定，减轻人口数量与环境、经济、社会、资源等之间的矛盾，提高人民在未来时间内的幸福感。同时让中国的发展进一步与国情相结合，实现发展道路的优化。

§1.2　研究背景

1949-2023 年，经过 70 十多年的发展，中国一跃成为世界第一人口大国。期间为了减轻人口数量与国家安全之间的矛盾，开国初期提倡多生；50 年代中期开始意识到人口过多的问题，70 年代开始逐步加强控制生育，做到计划生育，80 年代后期到今天，为了减轻老年化的进程，逐步倡导人们多生育。中国的人口数量一直是党和国家关心的重要事情。中国政策对于国家发展与人口变化起到了至关重要的作用，这也导致了中国的人口规模变化与美国等国家有较大的差别。常见的人口增长模型有马尔萨斯人口模型、Logistic 增长模型，或者说是指数模型。

表 1.1　中国人口[2]

年份	1949	1953	1965	1982	1990	2000	2010	2020
人口/亿	5.42	5.88	7.25	10.17	11.43	12.67	13.40	14.43

§1.3　模型初步假设及合理性分析

中国的人口受政策影响较大，致使其表面上并不符合指数模型。考虑到任意函数，均可以使用多项式进行拟合，故这里考虑使用一元多项式函数模型。

$$F(x)=\sum_{0}^{n}a_ix^i=a_0+a_1x^1+a_2x^2+\cdots+a_nx^n \tag{1.1}$$

其中 a_i 为每一项的系数，x 为年份，$F(x)$ 为多项式模型预测的人口数量。由多项式函数模型，我们需要确定最高次数与每一项的系数。其中，由多项式函数性质知，条

件的数量与所能求取的方程最高次数成正比。而条件的数量与系数的具体值与选取的求解方法有关。

为了更好的评估模型好坏，不仅要定性，更要定量。这里将 2020 年真实数据作为模型准确率依据，计算不同多项式模型下的相对误差

$$\text{相对误差} = \left| \frac{\underset{\text{(基于 1949-2010 年数据)}}{\text{2020 年预测人口数}} - \text{2020 年真实人口数量}}{\text{2020 年真实人口数量}} \right| * 100\%$$

高次模型的预测结果有时会不错，但加入新数据点后，预测值会变化很大，故而需要引入稳定性指标。这里将加入新数据点（2020 真实人口数据）对 2030 年预测变化作为指标，使用相对误差描述。

$$\text{相对误差} = \left| \frac{\underset{\text{(基于 1949-2020 年数据)}}{\text{2030 年预测人口数}} - \underset{\text{(基于 1949-2010 年数据)}}{\text{2030 年预测人口数}}}{\underset{\text{(基于 1949-2010 年数据)}}{\text{2030 年预测人口数}}} \right| * 100\%$$

这里都是相对误差越小，准确率越高，稳定性越高。

第 2 章　两种多项式模型

§2.1　插值模型

§2.1.1　模型假设及合理性分析[1]

假设中国人口数量调查为真值，没有误差。即寻求的多项式函数经过每一个已知数据点

$$F(x_i) = y_i, i = 0, 1, \cdots, n$$

由已知的 $n+1$ 个数据点，可以构建线性方程组

$$\begin{cases} a_0 + a_1 x_0 + \cdots + a_n x_0^n = y_0 \\ a_0 + a_1 x_1 + \cdots + a_n x_1^n = y_1 \\ \vdots \\ a_0 + a_1 x_n + \cdots + a_n x_n^n = y_n \end{cases} \tag{2.1}$$

由线性代数的知识，可以得到该线性方程组的系数矩阵

$$\mathbf{A} = \begin{Bmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^n \end{Bmatrix} \tag{2.2}$$

恰为范德蒙德矩阵，由于 $x_i (i = 0, 1, \cdots, n)$ 互异，故

$$det\mathbf{A} = \prod_{i,j=0, i>j}^{n-1} (x_i - x_j) \neq 0 \tag{2.3}$$

由行列式不为 0，故该线性方程组有唯一解，即多项式插值有唯一解。为了简化编程，这里选取与之等价的 $Lagrange$ 插值。

§2.1.2 拉格朗日插值

§2.1.2.1 n 次插值基函数

在插值节点处的值为 1，但在其它插值节点处的值为 0 的 n 次多项式函数，即

$$l_k(x) = \frac{(x-x_0)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}, k = 0, 1, \cdots, n$$

$$s.t. \quad l_k(x_j) = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases} \qquad j, k = 0, 1, \cdots, n \tag{2.4}$$

§2.1.2.2 Lagrange 插值多项式

由线性插值的两点式知，可以构造插值多项式

$$L_n(x) = \sum_{k=0}^{n} y_k l_k(x) \tag{2.5}$$

由基函数性质知，

$$L_n(x_j) = \sum_{k=0}^{n} y_k l_k(x_j) = y_j, \quad j = 0, 1, \cdots, n \tag{2.6}$$

可见 $L_n(x)$ 满足插值多项式等价。

§2.1.3 结果

将 1949-2010 年的数据与 1949-2020 数据分别代入拉格朗日插值多项式可以得到插值函数模型

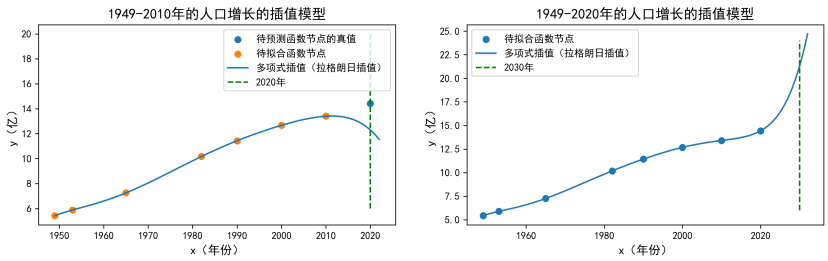

图 2.1 拉格朗日插值模型

基于 1949-2010 年的数据，其预测 2020 年人口为 12.30 亿人，相对误差为 14.78%，并不理想，但其预测 2030 年人口数量为 4.64 亿人，这大概率难以实现，在加入 2020 年的数据点，重新预测结果改观许多为 21.34 亿人，但仍不理想。从图像上分析得知，

在使用全数据点时，产生了龙格现象，即高次插值的病态性质，为此我们可以使用分段插值来避免高次插值。

§2.1.4 插值模型的优化

分段插值是一个降低次数不错的方法。然分段线性插值对插值区间内，即内插点效果不错，对于外插点则不尽人如意。若要同时兼顾连续性（考虑斜率与曲率）与插值节点的通过性，由条件数分析（插值节点本身，插值节点处的左右连续性，一阶导数的左右连续性，二阶导数左右连续性，再补上两个边值条件，刚好 $4n$ 个条件），则可构建出 3 次的函数，故而选择三次样条插值，将导数与分段插值结合在一起，并且最高次数只有 3 次的插值模型

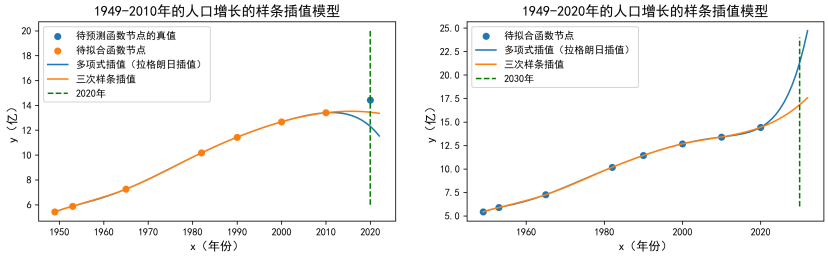

图 2.2　拉格朗日插值模型与三次样条插值模型

基于 1949-2010 年的数据，其预测 2020 年人口为 13.44 亿人，相对误差为 6.84%，不算糟糕。2030 年人口数量为 12.62 亿人，还能接受。在加入 2020 年的数据点，2030 年预测结果有较大改变为 16.84 亿人。虽然三次样条插值仍有许多不尽人如意，但其对于拉格朗日插值，龙格现象现象被大幅度削弱，无论是在模型的预测准确度，还是模型的稳定性都要远高于拉格朗日插值模型。

§2.2 回归模型

假设多项式经过每一个数据点并不能得到较好的多项式模型，一部分原因在于人口数据并不是十分准确，有一定的误差，故考虑函数拟合，找到一个多项式函数，让其在二范数下与人口数据点的误差最小。

$$\|\delta\|_2^2 = \sum_{i=0}^{n} \delta_i^2 = \sum_{i=0}^{n} [F^*(x_i) - y_i]^2 = \min \sum_{i=0}^{n} [F(x_i) - y_i]^2 \tag{2.7}$$

其中 x_i, y_i 分别与真实人口数据中的年份，人口数量对应。

由条件数知道，基于 $1949 - 2010$ 年这 7 个数据最多可以构建 6 次多项式，那么不妨一个一个试验。

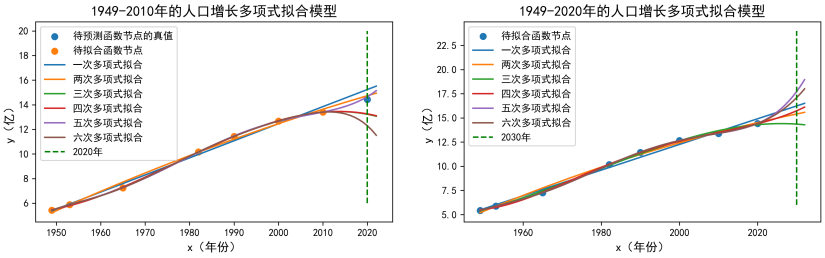

图 2.3　回归模型

将其两两进行比较

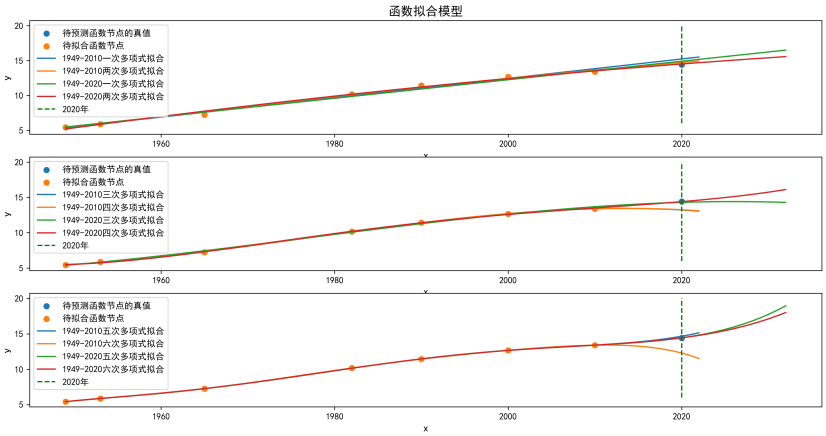

图 2.4　回归模型

分析得知，其在 2, 3, 5 次下的拟合效果还不错，再将这三个比较

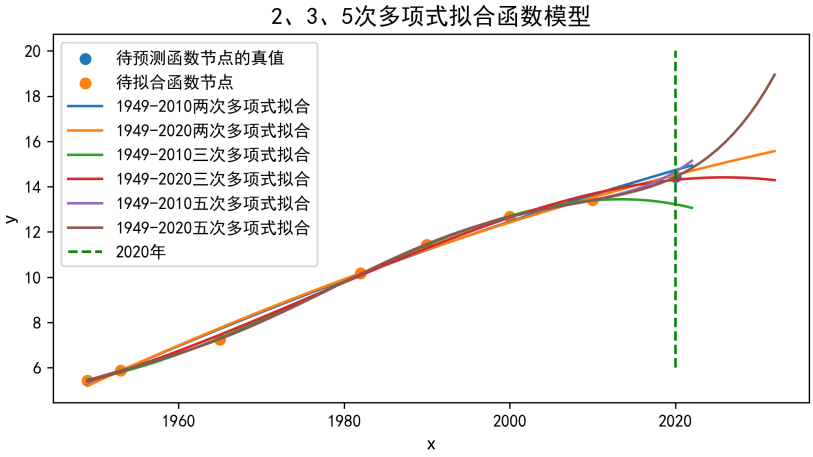

图 2.5　2、3、5 次回归模型

但 5 次下回归模型的的稳定性较次，且次数过高，可能会有龙格现象，故而 2 次多项式拟合的准确度高于 3 次，稳定性两者相近，故二次多项式拟合模型最优。

§2.2.1 回归模型的优化

马尔萨斯人口模型、Logistic 增长模型本质上仍为指数模型，那么使用指数函数结构的回归模型效果如何了。先从最简单的

$$F(x) = ae^{bx}$$

求解该模型，可以将两边同时取对数

$$\ln F(x) = \ln a + bx$$

，即线性回归模型

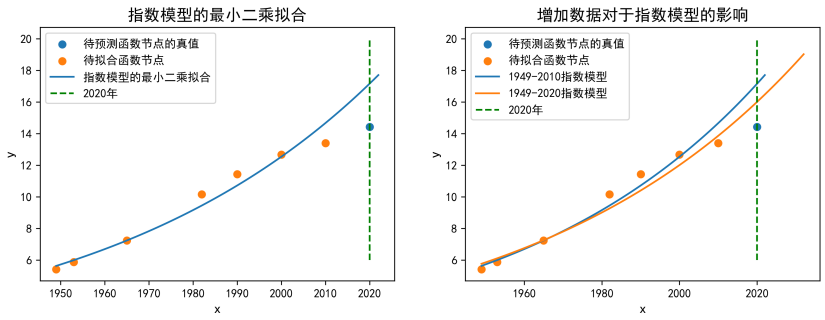

图 2.6 简单指数回归模型

结果比预想的好许多，但其在对未来几年的增长判断上不令人满意，那么不妨将指数模型与多项式模型结合，即对 $F(x)$ 取对数后，再建立其与时间的多项式关系，考虑到高次插值的龙格现象，这里只考虑多项式的回归模型，

$$\ln F(x) = \sum_{0}^{n} a_i x^i = a_0 + a_1 x^1 + a_2 x^2 + \cdots + a_n x^n \tag{2.8}$$

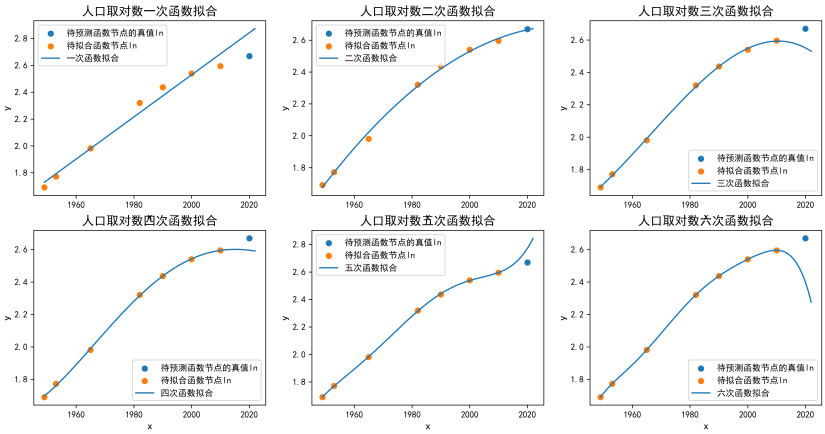

图 2.7 对数下的多项式回归模型

效果令人意外，可以说是非常好。还原函数后

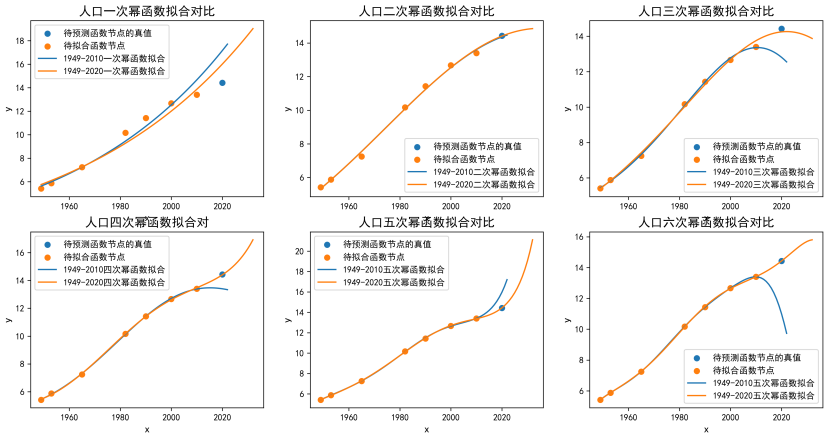

图 2.8　多项式为指数的回归模型

发现多项式幂次对于多项式回归本身的趋势影响并不大，但有一定的偏移，虽然大部分结果显示多项式幂次回归模型要弱于回归模型本身，但二次多项式作为指数的函数拟合模型是一个例外，不论误差，还是稳定性，均非常好。二次多项式作为指数的函数模型的半对数坐标系下的结果。

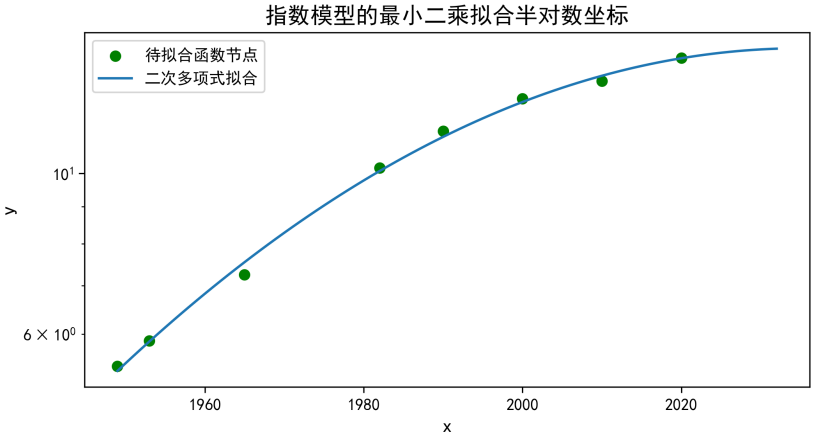

图 2.9　半对数坐标下的二次多项式指数回归曲线

第 3 章　模型分析与总结

§3.1　模型验证/分析/对比

这里将文中涉及的每一个模型制成表格 其中, 预测值 (2020): 基于 1949-2010 年

表 3.1　中国人口

模型	**真值** (2020) (亿人)	**预测值** (2020) (0) (亿人)	**相对误差** (1)	**预测值** (2030) (1) (亿人)	**预测值** (2030) (2) (亿人)	**相对** 误差 (2)
插值	14.43	12.30	14.78%	4.64	21.34	360%
样条插值	14.43	13.44	6.84%	12.62	16.84	33.36%
1 次拟合	14.43	15.23	5.57%	16.62	16.24	2.30%
2 次拟合	14.43	14.73	2.08%	15.75	15.41	2.17%
3 次拟合	14.43	13.22	8.37%	11.94	14.36	20.25%
4 次拟合	14.43	13.25	8.16%	12.05	15.74	30.67%
5 次拟合	14.43	14.65	1.55%	18.73	17.74	5.30%
6 次拟合	14.43	12.29	14.86%	4.58	17.10	273.22%
指数拟合	14.43	17.16	18.92%	20.07	18.49	7.91%
1ln 拟合	14.43	17.16	18.92%	20.07	18.49	7.91%
2ln 拟合	14.43	14.36	0.47%	14.74	14.82	0.54%
3ln 拟合	14.43	12.81	11.22%	11.04	14.01	26.91%
4ln 拟合	14.43	13.40	7.12%	12.92	16.30	26.09%
5ln 拟合	14.43	15.97	10.65%	29.80	19.04	36.11%
6ln 拟合	14.43	11.00	23.75%	3.22	15.70	387.30%

数据预测 2020 年人口数量。相对误差 (1): 基于 1949-2010 年数据预测的值与 2020 年真值的相对误差。值越小越模型越准确。预测值 1(2030): 基于 1949-2010 年数据预测 2030 年人口数量。预测值 2(2030): 基于 1949-2020 年数据预测 2030 年人口数量。相对误差 (2): 基于 1949-2010 年数据预测 2030 年人口数量与基于 1949-2020 年数据预测 2030 年人口数量的相对误差, 或者说增加数据对于模型预测的相对改变量。值越小模型越稳定。

其中指数函数拟合模型与一次多项式作为指数的模型的结果一致的原因在于

$$F(x) = ae^{bx+c} = e^{\ln a+bx+c} = e^{bx+c}$$

分析表格知道，二次多项式作为指数的拟合模型，其准确率与稳定性都是最高的，其次是 2 次多项式与 1 次多项式拟合。5 次多项式拟合模型数据上结果虽不错，但次数过高，容易出现龙格现象。整体上，次数低的模型，其稳定性更好，6 次仿若一个临界点，大于等于 6 次，稳定性会急速下降。进一步分析知道，虽然马尔萨斯人口模型、Logistic 增长模型是微分方程的形式，但其符号解可以视作多项式作为指数的函数模型。然由符号解的结构知道，即使是同一次数的多项式，微分方程下的符号解也是不全面的，未考虑低于最高次数的每一项。更不必说微分方程符号解最多只考虑了两次多项式。从节的结构上看微分方程是指数为多项式的幂函数模型的特例。实际计算中微分方程模型难度远高于取对数后多项式回归模型。

§3.2 总结

多项式插值会因次数过高，而出现龙格现象，分段虽然可以削弱，但对于外插节点的预测差强人意，使用多项式拟合，整体上比插值模型好，但高次的拟合结果不稳定，最后是将多项式作为指数的函数进行拟合，效果整体与多项式拟合接近，但会有一个二次多项式作为指数的函数模型效果非常好，无论是误差，还是稳定性都是所有方法中最好的。其实多项式作为指数的模型可以视作马尔萨斯人口模型、Logistic 增长模型的推广，增长率不变的视作在一次多项式作为指数的模型，增长率变化的 Logistic 增长模型则可以视作二次多项式作为指数的增长模型，而在多项式作为指数的函数拟合，其多项式的次数讨论范围更广，计算更加简便。在某种程度上得到的结果回比使用 Logistic 增长模型得到结果更好。

参考文献

[1] 李庆扬, 王能超, 易大义. 数值分析 [M]. 北京: 清华大学出版社, 2012.

[2] 姜启源, 谢金星, 叶俊. 数学模型 [M]. 北京: 高等教育出版社, 2019.