

基于 ARIMA 和 ARIMAX 模型的北京市 GDP 预测分析

张月

北京邮电大学, 北京 100089

摘 要： 为了更加准确地预测北京市 GDP 走势，采用时间序列分析方法对 1982—2022 年的北京市 GDP 数据展开研究，尝试构建基于动态回归（ARIMAX）模型。通过 Lasso 回归确定影响 GDP 变化的因素，并将这些因素作为回归项引入差分自回归移动平均（ARIMA）模型，构建 ARIMAX 模型。

关 键 词： GDP；ARIMA 模型；ARIMAX 模型

Forecast analysis of Beijing GDP based on ARIMA and ARIMAX models

Zhang Yue

Beijing University of Posts and Telecommunications, Beijing 100089

Abstract： In order to predict the GDP trend of Beijing more accurately, the time series analysis method was used to study the GDP data of Beijing from 1982 to 2022, and the dynamic regression (ARIMAX) model was tried to be built. The factors affecting GDP change were determined by Lasso regression, and these factors were introduced into the differential autoregressive moving average (ARIMA) model as regression terms to build the ARIMAX model.

Keywords： GDP; ARIMA model; ARIMAX model

国内生产总值（GDP）作为一个关键的经济指标，具有重大意义。它反映了一个国家或地区的经济规模、发展水平和综合实力。GDP 涵盖了所有最终商品和劳务的市场价值，体现了经济活动的活跃程度。在全球以及中国的核算体系中，国内生产总值（GDP）扮演着核心角色。截至2022 年，北京实现地区生产总值 41610.9 亿元，比上年增长 0.7%^[1]。准确预测北京市 GDP 有助于政府制定合理的宏观经济政策。

时间序列分析广泛应用于日常生活的诸多场景。陈聪聪采用 ARIMA 模型及 ARIMAX 模型，通过将 1975 年至 2013 年间山东省的 GDP 数据与第三产业产值作为分析对象，成功地进行了模型拟合^[2]。在 2019 年的研究中，徐向辉选择了 ARIMAX 模型作为工具，对财政收入的变化趋势进行了前瞻性的对比分析，聚焦于未来五年的预测^[3]。在 2019 年，Anggraei W 等研究者将水稻相关变量及季节因素视为影响大米价格未来走势的关键自变量，采用人工神经网络与 ARIMAX 模型相融合的预测方式，为定价决策的制定提供有力支持^[4]。2021 年，田庆等人基于山东省肺结核的季节变化及发展规律，构建了 ARIMA 乘积季节性模型，旨在预测肺结核病例数量^[5]。张淑嫻则采用时间序列分析方法，先借助 Lasso 模型和灰色关联分析确定影响私人汽车拥有量的主要因素，进而构建 ARIMA 和 ARIMAX 模型^[6]。

本文选取 1982-2022 年北京市 GDP 的相关数据进行实证分析，分别构建动态回归（ARIMAX）、差分自回归移动平均（ARIMA）模型，分析影响北京市 GDP 的关键因素并预测其变化趋势。

一、模型介绍

（一）ARIMA 模型

如果平稳时间序列 $\{X_t\} = \{X_t, t \in \mathbb{N}\}$ 满足

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + Z_t - \theta_1 Z_{t-1} - \theta_2 Z_{t-2} - \cdots - \theta_q Z_{t-q}$$

其中 $\phi_0, \phi_1, \dots, \phi_p$ 为 $p+1$ 个实数， $\phi_p \neq 0$ ， $\theta_1, \dots, \theta_q$ 为 q 个

实数， $\theta_q \neq 0$ ，且满足

$$\begin{cases} \Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \neq 0, & |z| \leq 1 \\ \Theta(z) = 1 - \theta_1 z - \theta_2 z^2 - \cdots - \theta_q z^q \neq 0, & |z| \leq 1 \end{cases}$$

$\Phi(z)$ 和 $\Theta(z)$ 互素

其中 $Z_t \sim WN(\mu, \sigma^2)$ ，则可将 $\{X_t\}$ 视为阶数为 (p, q) 阶的平稳自回归移动平均模型 ARMA(p, q)^[7]。应用延迟算子，上述模型可以表示为 $\Phi(B)x_t = \Theta(B)Z_t$ ，其中 $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ ，

$$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q.$$

ARIMA 模型的结构如下

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(\varepsilon_t x_s) = 0, \forall s < t \end{cases}$$

其中 $\nabla^d = (1-B)^d$, ε_t 为白噪声序列。特别的, 当 $d=0$ 时, ARIMA(p, d, q) 模型实际上就是 ARMA(p, q) 模型。

(二) Lasso 回归模型

考虑线性回归模型 $Y = X\beta + \varepsilon$, 其中 $Y = (y_1, y_2, \dots, y_N)^T$; $X = (x_1, x_2, \dots, x_N)^T$, $x_j = (x_{1j}, x_{2j}, \dots, x_{Nj})^T$, $i = 1, 2, \dots, p$, N 为样本的个数; $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$; $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N)^T$, $\varepsilon \sim N(0, \sigma^2)s$ 。

在这个线性模型基础上的 Lasso 筛选变量公式为

$$\begin{cases} \hat{\beta} = \min \frac{1}{2} \sum_{i=1}^n (y_i - \sum_j \beta_j x_{ij})^2 \\ s.t. \sum_{j=1}^p |\beta_j| \leq t \end{cases}$$

又可以表示为拉格朗日形式:

$$\hat{\beta} = \arg \min \frac{1}{2} \sum_{i=1}^n (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Lasso 方法通过构造 L_1 惩罚, 使模型精简。它强制系数绝对值之和小于固定值, 当该值较小时, 部分变量系数会被压缩为 0, 实现变量筛选, 同时完成未知参数估计^[8]。

(三) ARIMAX 模型

ARIMAX 模型^[9]的构建前提为: 输入序列 $\{x_{1t}\}, \{x_{2t}\}, \dots, \{x_{it}\}$ 为平稳序列, 响应序列 $\{y_t\}$ 也是平稳的, 并且两者线性相关, 响应序列会因输入序列的变化而变化。构建如下结构的回归模型:

$$y_t = \beta_0 + \sum_{i=1}^k \frac{\Theta_i(B)}{\Phi_i(B)} B^{l_i} x_{it} + \varepsilon_t$$

其中 $\Phi_i(B) = 1 - \phi_{i1}B - \dots - \phi_{ip_i}B^{p_i}$ 为第 i 个输入变量的自回归系数多项式, $\Theta_i(B) = 1 - \theta_{i1}B - \dots - \theta_{ip_i}B^{p_i}$ 为第 i 个输入变量的自回归系数多项式, l_i 为第 i 个输入变量的延迟阶数, $\{\varepsilon_t\}$ 为回归残差序列。

鉴于输入序列与响应序列均呈现出平稳特性, 由此可推知二者线性组合亦具备平稳属性, 换言之, 残差序列亦处于平稳状态, 进而残差序列可表示为

$$\varepsilon_t = y_t - \beta_0 - \sum_{i=1}^k \frac{\Theta_i(B)}{\Phi_i(B)} B^{l_i} x_{it}$$

对残差序列进一步运用 ARMA 模型进行拟合, 以提取其中信息, 从而得出:

$$\varepsilon_t = \frac{\Theta(B)}{\Phi(B)} a_t$$

其中 $\Phi(B)$ 为残差序列的自回归系数多项式, $\Theta(B)$ 为残差序列的移动平均系数多项式, $\{a_t\}$ 为零均值白噪声序列。最终所获得的动态回归 ARIMAX 模型的具体形式呈现如下

$$\begin{cases} y_t = \beta_0 + \sum_{i=1}^k \frac{\Theta_i(B)}{\Phi_i(B)} B^{l_i} x_{it} + \varepsilon_t \\ \varepsilon_t = \frac{\Theta(B)}{\Phi(B)} a_t \end{cases}$$

二、模型建立

(一) ARIMA 模型

首先选取北京市 1982 到 2019 年的 GDP 数据进行 ARIMA 模型拟合。观察序列的趋势可以看出序列是呈指数型增长的, 因此为了数据更容易平稳化且避免异方差现象的出现, 首先对要拟合的时间序列进行对数化处理, 对数化后的数据不是平稳的, 所以还需要对对数化后的序列进行差分处理。

二阶差分后认为处理后的序列是平稳的, 为了客观起见, 对二阶差分后的序列进行 ADF 单位根检验^[10], 检验结果显示 p 值小于 0.05, 所以可以判定经过二阶差分的对数序列是平稳的。然后对序列进行 Ljung-Box 检验, p 值同样小于 0.05, 说明序列不是白噪声序列, 能够进行进一步的建模。考察序列的自相关系数与偏自相关系数并结合 AIC 准则进行模型定阶, 使 aic 达到最小的模型被认为是相对最优模型, 于是构建模型 ARIMA(2, 2, 0), 该模型的 AIC 值为 -123.2, 表达式为

$$(1 + 0.2525B + 0.4512B^2)(1 - B)^2 y_t = \varepsilon_t$$

对残差序列进行随机性检验, p 值为 0.7826 远大于 5% 的显著性水平, 所以认为残差是白噪声序列。因此, 模型的残差通过了随机性检验, 认为模型的拟合效果较好, 可以用于进一步的预测。

(二) 变量筛选

以国内生产总值 (GDP) 总值作为响应变量 Y , 挑选了十个紧密关联经济社会活动的变量作为解释变量 X , 分别为: 城镇登记失业率 X_1 , 货运量 X_2 , 财政收入 X_3 财政支出 X_4 , 居民消费价格指数 X_5 , 农林牧渔业总产值 X_6 , 粮食产量 X_7 , 建筑业总产值 X_8 , 社会消费品零售总额 X_9 , 全社会固定资产投资 X_{10} 。

首先对变量 X 和 Y 做标准化处理, 去除单位的限制, 以此来提高预测的精准度。标准化的公式如下

$$\bar{x}_i = \frac{x_i - \mu}{\sigma}$$

x_i 表示第 i 个样本输入值, μ 表示样本均值, σ 表示样本标准差。

标准化处理 X 和 Y 后, 开展 Lasso 回归分析。当 λ 值有所变动时, 模型系数的变动幅度起初较为显著, 但随后逐渐趋于平缓, 这表明选取一定量 λ 值后, 在之后的 λ 值的变化对于模型的调整影响微弱。由此可获取较为可信的 λ 值, 进而确定变量选择的结果。并且鉴于当 λ 值上升至特定阈值后, 进一步扩充模型中自变量的数量, 也就是降低 λ 值, 无法对模型效能产生显著的提升效果, 综合考量 arimax 模型的可操作性与复杂性, 在此环节确定选用的 λ 值为 0.04159, 筛选出的变量为 $x_1, x_4, x_8, x_9, x_{10}$ 。

(三) ARIMAX 模型

同样首先对筛选出来的变量分别做对数化处理, 记为 $\ln x_1, \ln x_4, \ln x_8, \ln x_9, \ln x_{10}$, 发现对数化后的序列呈现出较为一致的变化趋势, 但不是平稳序列。于是需要对各个序列分别做差分处理来实现平稳化, $\ln y, \ln x_1, \ln x_4, \ln x_8, \ln x_9, \ln x_{10}$ 都是在二阶差分后达到平稳状态, 即为二阶单整的, 分别记为

$d \ln y, d \ln x_1, d \ln x_2, d \ln x_3, d \ln x_4, d \ln x_5, d \ln x_6, d \ln x_7, d \ln x_8, d \ln x_9, d \ln x_{10}$ 。对这些处理后的序列进行 ADF 检验，可以得到检验的 p 值分别为 0.0222, 0.0100, 0.0240, 0.0100, 0.0203, 0.0405, 均小于显著性水平 $\alpha = 0.05$ ，由此可判定 $d \ln y, d \ln x_1, d \ln x_2, d \ln x_3, d \ln x_4, d \ln x_5, d \ln x_6, d \ln x_7, d \ln x_8, d \ln x_9, d \ln x_{10}$ 这些序列是平稳的。对经过二次差分处理的序列执行白噪声测试，通过 Ljung-Box 检验来评估序列中是否存在未被提取的信息，检验的序列的 p 值均远小于 0.05，即这些差分序列均不能视为白噪声序列，可以用于进一步建立 ARIMA 模型。

对这些序列分别拟合 ARIMA 模型可得： $\ln x_1$ 的拟合模型是 ARIMA(0,2,1)，AIC 值为 -0.61； $\ln x_4$ 的拟合模型是 ARIMA(1,2,0)，AIC 值为 -65.56；得到 $\ln x_8$ 的拟合模型是 ARIMA(0,2,0)，这个模型没有什么意义，所以剔除 $\ln x_8$ 这个变量； $\ln x_9$ 的拟合模型是 ARIMA(0,2,1)，AIC 值为 -96.41； $\ln x_{10}$ 的拟合模型是 ARIMA(0,2,1)，AIC 值为 -42.58。通过随机性检验剩余变量的残差序列，结果表明所有得到的 p 值均显著地高于显著性水平 $\alpha = 5\%$ 。这一发现支持了模型具有良好的拟合效果，从而使其具备了应用于预测这些变量的可能性。

$d \ln y$ 与 $d \ln x_1, d \ln x_2, d \ln x_3, d \ln x_4, d \ln x_5, d \ln x_6, d \ln x_7, d \ln x_9, d \ln x_{10}$ 在滞后阶数为 0 时序列相关系数达到最大值，表明在延迟 0 阶存在最强的相关性，此类序列适合进行同时期的模型构建。借助 R 中的 arimax 函数便于拟合 ARIMAX 模型，结果显示拟合后模型的 AIC 值为 -54.59。残差 p 值大于 0.05，表明模型信息已充分提取，通过条件最小二乘法获取模型参数，引入特定参数之后，能够得出该模型的数学表述为

$$\ln y = 0.9477 + 0.8190 \ln x_9 + 0.1672 \ln x_{10} + \frac{1}{(1 - 0.9358B)} a_t$$

三、预测比较

上述两个模型的建立是基于北京市 1982 到 2019 年的 GDP 数据和相关的经济指标，在这里我们使用上述模型对 2020-2022 年的测试集数据进行预测，来检验模型的拟合效果，预测结果如下：ARIMA 模型这三年的预测值分别为 40032.38、43796.32、47560.26；ARIMAX 模型这三年的预测值分别为 38474.26、41602.01、44729.75。而这三年的真实值为 35943.3、41045.6、41540.9。可以看出 ARIMAX 模型预测值比 ARIMA 模型预测值的表现更好，具体表现在相对误差更小，因此也可以说 ARIMAX 模型的预测精度更高。

四、结论

本文基于 1982 年到 2022 年的北京市 GDP 数据和相关影响 GDP 的经济社会指标，采用常用的时间序列分析方法，运用 ARIMA 模型和 ARIMAX 模型对北京市宏观 GDP 序列波动情况进行拟合。本文使用 Lasso 回归进行变量筛选，选取城镇登记失业率、财政支出、社会消费品零售总额、全社会固定资产投资四变量预测 GDP 变化。先证其与 GDP 序列的协整性，再依照互相关关系建动态回归模型（ARIMAX）。依据 AIC 准则的理论分析，ARIMAX 模型相较于 ARIMA 优势显著，实证结果明确显示 ARIMAX 模型的预测性能优于 ARIMA 模型，更加适合用于预测未来的 GDP 变化趋势。

参考文献

[1]. 李美娟. 基于 ARIMAX 和 BP 神经网络的安徽省 GDP 预测研究 [D]. 北方民族大学, 2022.DOI:10.27754/d.cnki.gbfmz.2022.000194.
[2]. 陈聪聪. 基于 ARIMA 模型和 ARIMAX 模型的山东省 GDP 的预测与分析 [D]. 山东大学, 2016.
[3]. 徐向辉. 基于 ARIMAX 模型的全国财政收入的预测与分析 [D]. 大连理工大学, 2019.DOI:10.26991/d.cnki.gdllu.2019.002701.
[4]. Anggraeni W, Mahananto F, Sari A Q, et al.Forecasting the Price of Indonesia's Rice Using Hybrid Artificial Neural Network and Autoregressive Integrated Moving Average (Hybrid NNs-ARIMAX) with Exogenous Variables [J]. Procedia Computer Science, 2019, 161:677-686.DOI:10.1016/j.procs.2019.11.171.
[5]. 田庆, 刘永鹏, 张晶晶, 等. ARIMA 乘积季节模型在山东省肺结核发病预测中的应用 [J]. 山东大学学报 (医学版), 2021, 59(07):112-118.
[6]. 李威. 基于非参数组合预测模型的江西省 GDP 的实证分析 [D]. 江西理工大学, 2022.
[7]. 王燕. 时间序列分析: 基于 R [M]. 中国人民大学出版社, 2015.
[8]. 司圃旭. 基于变量选择和 ARIMAX 模型的山东省 GDP 预测 [D]. 山东大学, 2019.DOI:10.27272/d.cnki.gshdu.2019.000983.
[9]. 汪远征, 徐雅静. 多元平稳时间序列 ARIMAX 模型的应用 [J]. 统计与决策, 2007, (18): 132-135.
[10]. 衡红军, 任鹏. 基于时间序列的机场短时段值机客流量预测 [J]. 计算机仿真, 2020, 37(02):26-32.