ARIMA与ARIMAX模型在私人汽车拥有量预测中的应用

张淑娴

(安徽建筑大学数理学院，合肥 230601)

摘要： 为了提高私人汽车拥有量的预测精度，利用时间序列分析方法对全国2005—2020年的私人汽车拥有量数据进行研究，建立基于动态回归(ARIMAX)模型。运用Lasso模型和灰色关联分析得出影响私人汽车拥有量的主要因素，并将主要因素作为回归项引入差分自回归移动平均(ARIMA)模型。然后，在ARIMA模型的基础上建立ARIMAX模型。模型预测的对比结果揭示了ARIMAX的拟合效果更佳，适用于全国私人汽车拥有量的预测。

关键词： 动态回归(ARIMAX)模型；Lasso模型；私人汽车拥有量

中图分类号： F542 **文献标志码：** A **文章编号：** 1671－1807(2024)09－0189－06

随着国民经济的持续快速发展，全国私人汽车的拥有量急剧攀升。汽车的广泛普及便捷了人们的生活，但又在一定程度上对环境造成了危害。因此，研究全国私人汽车拥有量对于改善环境具有重要意义。

目前，一些学者对汽车拥有量的影响因素及预测进行了研究。张琪[1]综合考虑了私人汽车拥有量与经济、城市和交通这3种属性之间的关系，分别建立了随机效应模型、固定效应模型与混合回归模型，通过分析发现城镇居民家庭人均可支配收入为私人汽车拥有量的主导因素。周亚林等[2]首先借助机器学习中的极度梯度提升树法识别得到了影响新疆私人汽车保有量的因素，然后比较了极端梯度提升树(extreme gradient boosting，XGBoost)、随机森林和神经网络这3种方法的预测结果，结果表明神经网络预测效果最好。杨昆等[3]采用M-K(Mann-Kendall)检验、Theil指数、线性倾向率和面板数据模型，从全国、8大经济区域、各省3个尺度研究了中国民用汽车拥有量的时空变化特征及其与地区生产总值、公路里程和居民消费水平这3个影响因素的关系，结果表明在不同的时间阶段和空间尺度，各因素对民用汽车拥有量的作用方向以及强度上表现出了差异。Kai等[4]在经典指数曲线模型和修正指数曲线模型的基础上，提出了一种具有一阶多项式项的新型指数曲线模型，将这3种模型的预测结果进行比较，结果显示运用新型指数曲线模型预测中国私家车拥有量具有更高的精度。郭艳莉[5]采用灰色-广义回归神经网络预测模型分析私人汽车拥有量，预测结果表明该模型优于回归预测模型、灰色预测模型和反向传播(back propagation，BP)神经网络预测模型。李炳炎等[6]利用多元线性回归模型和向量自回归(vector autoregression，VAR)模型预测江苏省私人汽车拥有量，结果显示年末总人口数为私人汽车拥有量的主导因素。上述研究大多局限于单变量的时间序列分析。

本文选取2005—2020年全国私人汽车拥有量的相关数据进行实证分析，分别构建基于Lasso和灰色关联分析方法下的差分自回归移动平均模型(autoregressive integrated moving average，ARIMA)模型与动态回归(ARIMAX)模型，分析影响私人汽车拥有量的关键因素，借此进一步预测私人汽车拥有量的变化趋势，以期为汽车数量的有效控制提供依据。值得一提的是，在本文的研究中考虑了多因素影响的ARIMAX模型，其预测效果优于ARIMA模型，能更好地反映各变量之间在时间上的动态关系。

1 模型简介

1.1 Lasso模型

考虑一个具有标准化自变量和因变量的线性回归$\boldsymbol{Y}=\boldsymbol{\beta X}+\boldsymbol{\varepsilon}$，其中$\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$；$\boldsymbol{X}=(x_1,x_2,\cdots,x_p)$，$\boldsymbol{x}_p=(x_{1i},x_{2i},\cdots,x_{ni})^{\mathrm{T}}$，$i=1,2,\cdots,$

收稿日期： 2024-01-23

基金项目： 安徽省高等学校科学研究重点项目(2022AH050247)；安徽建筑大学科研项目(2016QD118)

作者简介： 张淑娴(2000—)，女，江苏常州人，硕士研究生，研究方向为数据分析中的统计方法及应用。

p，n 为样本的个数；$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$，$p$ 为解释变量的个数；$\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^{\mathrm{T}}$，$\boldsymbol{\varepsilon}$ 为误差向量且满足 $E(\boldsymbol{\varepsilon})=0$，$\mathrm{Cov}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}$，$\boldsymbol{I}$ 为单位矩阵。

在线性模型的基础上产生的 Lasso 筛选变量公式为

$$\begin{cases}\hat{\beta}=\min\dfrac{1}{2}\sum\limits_{i=1}^{n}\left(y_i-\sum\limits_{j}\beta_j x_{ij}\right)^2\\ \text{s. t.}\sum\limits_{j=1}^{p}|\beta_j|\leqslant t\end{cases}\tag{1}$$

式中：$\hat{\beta}$ 为回归系数的估计值；t 为调和参数；σ^2 为方差。该函数又可以表示成拉格朗日形式：

$$\hat{\beta}=\mathrm{argmin}\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\sum_{j}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\tag{2}$$

式中：$\sum\limits_{j=1}^{p}|\beta_j|$ 即为 l_1 范数，$\lambda\sum\limits_{j=1}^{p}|\beta_j|$ 相当于在原回归方程的残差平方和（损失函数）基础上添加了一个惩罚，λ 为压缩系数，原理是通过选取合适的 λ 来压缩模型的系数，使得那些与被解释变量关系较弱的系数变小（甚至压缩为 0），从而提高了估计的准确度。

1.2　ARMA 模型

自回归移动平均（ARMA）模型[7]是通过自回归模型与移动平均模型相结合产生的，其定义为

$$\begin{cases}y_t=\phi_0+\sum\limits_{i=1}^{p}\phi_i y_{t-i}+\varepsilon_t-\sum\limits_{i=1}^{q}\theta_i\varepsilon_{t-i}\\ \phi_p\neq 0,\ \theta_q\neq 0\\ E(\varepsilon_t)=0,\ \mathrm{Var}(\varepsilon_t)=\sigma_\varepsilon^2,\ E(\varepsilon_t\varepsilon_s)=0,\ s\neq t\\ E(x_s\varepsilon_t)=0,\ \forall s<t\end{cases}\tag{3}$$

式中：y_t 为当前序列值；ϕ_0 为常数项；p 为 AR(p) 模型的偏自相关系数 p 阶截尾；ϕ_i 为自相关系数；q 为 MA(q) 模型的自相关系数 q 阶截尾；θ_i 为偏自相关系数；ε_t 为随机干扰项；y_{t-i} 为 $t-i$ 时刻的序列值；ε_{t-i} 为 $t-i$ 时刻的残差值；ϕ_p 为 p 时刻的自相关系数；θ_q 为 q 时刻的偏自相关系数；ε_s 为 s 时刻的残差；x_s 为 s 时刻（过去）的序列值；s 为 s 时刻（过去时刻）；t 为 t 时刻（当期）。

差分自回归移动平均（ARIMA）模型与 ARMA 模型的区别是 ARIMA 模型需要对时间序列进行 d 阶差分，从而得到平稳的时间序列。

1.3　多元时间序列 ARIMAX 动态回归模型

ARIMAX 模型构造之前必须满足响应序列 $\{y_t\}$ 和输入变量序列 $\{x_{1t},x_{2t},\cdots,x_{kt}\}$ 均为平稳序列；若不是平稳序列则需要采用差分或对数化的方法使其变平稳，随后便能够构造响应变量与输入变量之间的模型。

ARIMAX 模型构造的基本思想[7]为：考虑响应序列 $\{y_t\}$（即因变量序列）与输入变量序列（即自变量序列）$\{x_{1t},x_{2t},\cdots,x_{kt}\}$ 均平稳，构建因变量序列与自变量序列的回归模型为

$$y_t=\mu+\sum_{i=1}^{k}\frac{\Theta_i(B)}{\Phi_i(B)}B^{l_i}x_{it}+\varepsilon_t\tag{4}$$

式中：μ 为模型常数项均值；B 为移位算子；$\Phi_i(B)$ 为第 i 个输入变量的自回归系数多项式；$\Theta_i(B)$ 为第 i 个输入变量的移动平均系数多项式；l_i 为第 i 个输入变量的延迟阶数；$\{\varepsilon_t\}$ 为回归残差序列。

由于 $\{y_t\}$，$\{x_{1t}\}$，$\{x_{2t}\}$，…，$\{x_{kt}\}$ 均平稳，那么平稳序列的线性组合仍然是平稳的，也就是说残差序列 $\{\varepsilon_t\}$ 是平稳序列，$\{\varepsilon_t\}$ 的表达式为

$$\varepsilon_t=y_t-\left[\mu+\sum_{i=1}^{k}\frac{\Theta_i(B)}{\Phi_i(B)}B^{l_i}x_{it}\right]\tag{5}$$

接着借助 ARMA 模型继续提取残差序列 $\{\varepsilon_t\}$ 中的相关信息，最终得到的模型称为动态回归模型，简记为 ARIMAX。该模型表达式为

$$\begin{cases}y_t=\mu+\sum\limits_{i=1}^{k}\dfrac{\Theta_i(B)}{\Phi_i(B)}B^{l_i}x_{it}+\varepsilon_t\\ \varepsilon_t=\dfrac{\Theta(B)}{\Phi(B)}a_t\end{cases}\tag{6}$$

式中：$\Phi(B)$ 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。

2　变量筛选

选取国家统计局提供的数据进行归纳整理，将全国私人汽车拥有量看作被解释变量，居民消费价格指数、公路里程等 8 个因素看作解释变量进行分析，见表 1，并基于 Lasso 和灰色关联分析方法筛选影响私人汽车拥有量的关键因素。

表 1　影响私人汽车拥有量的变量选取及其含义

变量	变量符号	单位
私人汽车拥有量	y	万辆
居民消费价格指数（1978 年为 100）	x_1	—
公路里程	x_2	万 km
城镇居民人均可支配收入	x_3	元
国内生产总值	x_4	亿元
公路营运汽车拥有量	x_5	万辆
钢材产量	x_6	万 t
年末总人口数	x_7	万人
就业人数	x_8	万人

2.1 Lasso 筛选关键因素

通过 R 语言中的 glmnet 函数对私人汽车拥有量和 8 个影响因素的数据构建 Lasso 回归模型。压缩系数 λ 的取值不同,模型的系数变化不同,图 1 中每一条曲线的变化代表每个变量的回归系数随 λ 变化的趋势。

在图 1 中,横坐标表示压缩系数 λ 的对数值 lgλ,纵坐标表示模型的回归系数值。图形顶部的数字表示在对应的 λ 下得到的非零系数的个数。结果显示模型系数随着压缩系数 λ 值而变化,变化越来越平稳,最终模型系数趋近于一个相同的值。说明只要找到相对合理的 λ 值,就能够筛选出有效准确的变量。因此运用交叉验证方法选取最优参数 λ 值,结果如图 2 所示。

在图 2 中,横轴表示压缩系数 λ 的对数值 lgλ,纵轴表示模型均方误差(MSE),结果显示压缩系数 lgλ 越小,均方误差越稳定,左侧虚线表示在均方误差最小时所对应的模型包含了 5 个变量,右侧虚线表示在一倍标准误(SE)内更简洁的模型,包含了 3 个变量。

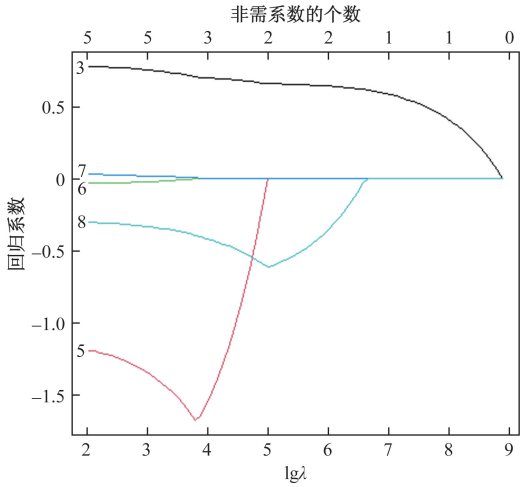


图 1 模型的回归系数值随着压缩系数的变化趋势

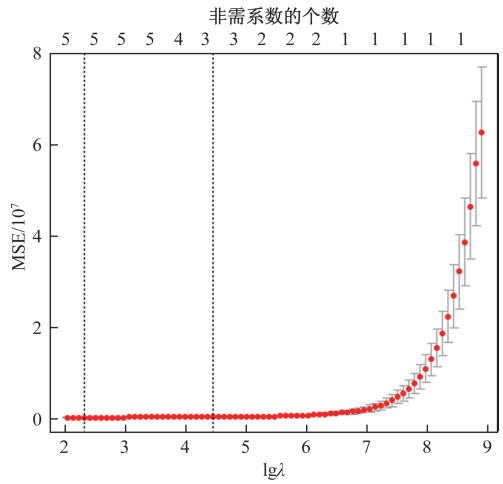


图 2 模型的压缩系数与均方误差 MSE 的变化

最终,Lasso 从所有变量中筛选出 x_3(城镇居民人均可支配收入)、x_5(公路营运汽车拥有量)、x_8(就业人数)3 个变量,其他的变量则被压缩至 0。这就说明城镇居民人均可支配收入、公路营运汽车拥有量、就业人数主要影响着私人汽车拥有量的变化。

借助 R 软件作参数估计,由表 2 可知,得到变量 x_8 的检验 P 值为 0.298,大于显著性水平 $\alpha=0.05$,故没有通过检验,所以在后续研究中剔除变量 x_8,只保留变量 x_3、x_5。

表 2 Lasso 回归模型参数估计

变量	参数估计	标准误	t 统计量	P
截距项	1.828×10^{4}	2.092×10^{4}	0.874	0.399 5
x_3	7.213×10^{-1}	2.456×10^{-2}	29.372	1.51×10^{-12}
x_5	−2.481	9.390×10^{-1}	−2.642	0.021 5
x_8	-3.018×10^{-1}	2.774×10^{-1}	−1.088	0.298 0

2.2 灰色关联分析法筛选关键因素

灰色关联分析法可以用来衡量变量之间发展趋势的相近或相异程度,故用此分析方法有利于筛选出影响私人汽车拥有量的因素。此外,灰色关联度的大小代表着各个序列影响主序列的程度大小,有利于分析变量的动态历程。

(1)将私人汽车拥有量作为主序列 $X_0=\{x_0(1),\cdots,x_0(k)\}$,$k=1,2,\cdots,16$,将居民消费价格指数等 8 个因素作为影响序列 $X_i=\{x_i(1),x_i(2),\cdots,x_i(k)\}$,$i=1,2,\cdots,8$;$k=16$。

(2)对数据进行无量纲化处理 $x_i(k)'=\dfrac{x_i(k)}{x_i(1)}$,$k=1,2,\cdots,16$;$i=0,1,2,\cdots,8$,目的是让数据的增长趋势更为明显。

(3)计算出无量纲化后的指标序列 $X'_i=\{x'_i(1),\cdots,x'_i(k)\}$ 与全国私人汽车拥有量序列 $X'_0=\{x'_0(1),\cdots,x'_0(k)\}$ 的差序列 $\Delta_i(k)=|x'_0(k)-x'_i(k)|$,随后找出差序列中的最大值 $\Delta(\max)$ 与最小值 $\Delta(\min)$。这时候,就可以得到关联系数 γ^*_{oi} 的表达式

$$\gamma^*_{oi}(k)=\frac{\Delta(\min)+\rho\Delta(\max)}{\Delta_i(k)+\rho\Delta(\max)},\quad k=1,2,\cdots,16;\ i=1,\cdots,8 \tag{7}$$

式中:ρ 为分辨系数,$\rho\in(0,1)$,一般取 0.5。

(4)计算全国私人汽车拥有量与各影响因素之间的灰色关联度 γ_{oi}。

$$\gamma_{oi}=\frac{1}{16}\sum_{k=1}^{16}\gamma_{oi}^{*}(k),\ i=1,2,\cdots,8 \qquad (8)$$

根据计算步骤,借助 MATLAB 软件编程求解,得到各相关因子与私人汽车拥有量的灰色关联度,并从大到小排序,结果见表 3。

根据灰色关联度排序结果可知,在 0.5 的分辨系数下,这 8 个因素对私人汽车拥有量的影响程度为:国内生产总值>城镇居民人均可支配收入>居民人均消费价格指数>钢材产量>公路里程>公路营运汽车拥有量>年末总人口数>就业人员。

对比以上两种方法,筛选出更加完善的影响汽车拥有量的关键因素,其中包括国内生产总值、城镇居民可支配收入和公路营运汽车拥有量。

表 3　灰色关联度排序

变量	灰色关联度
国内生产总值	0.792 07
城镇居民人均可支配收入	0.756 00
居民人均消费价格指数	0.707 07
钢材产量	0.653 82
公路里程	0.611 50
公路营运汽车拥有量	0.610 27
年末总人口数	0.576 01
就业人员	0.567 29

3　模型建立

3.1　用 ARIMA 模型预测私人汽车拥有量

借助 Eviews 软件进行 ADF 检验(augmented Dickey-Fuller test)可知 $\{y_t\}$ 是非平稳序列,但对数化后的序列能够通过 ADF 检验,说明 $\{\ln y_t\}$ 是平稳序列。考察对数化后序列的自相关图与偏自相关图的性质并结合 AIC(Akaike information criterion)准则进行定阶,构建模型 ARIMA(1,0,2),该模型的 AIC 值为−26.45,表达式为

$$\ln y_t=8.757\ 5+0.987\ 2\ln y_{t-1}+\varepsilon_t+1.638\ 2\varepsilon_{t-1}+\varepsilon_{t-2} \qquad (9)$$

下面利用 LB(Ljung-Box)检验对残差序列进行检验,当显著性水平 $\alpha=0.05$ 时,由图 3 的结果分析发现,有 95% 以上的标准化残差都是在区间[−2,2]以内的,此外,ARIMA(1,0,2)模型的残差的自相关函数在 0 阶后迅速下降至上下两条虚线之中,总体上 Ljung-Box 统计量的 P 值都大于显著性水平,表明该模型已充分提取信息。参数检验的结果亦显示模型的参数具有统计学意义。因此,可以判定建立的 ARIMA(1,0,2)模型是合理的。

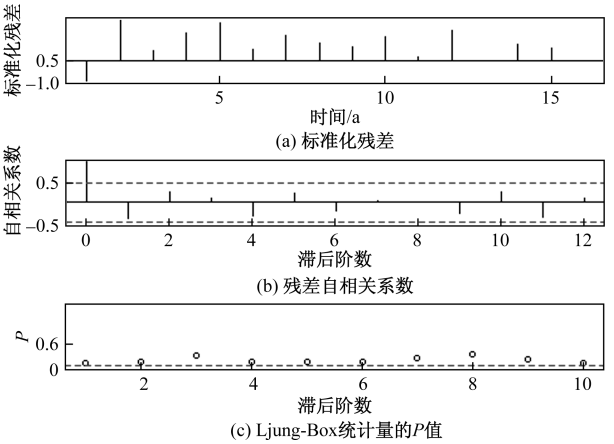


图 3　ARIMA(1,0,2)模型残差诊断检验

可以利用该模型对原始时间序列作预测,预测 2021 年、2022 年的私人汽车拥有量数据。

3.2　用 ARMAX 模型预测私人汽车拥有量

由于序列 $\{y_t\}$、$\{x_{3t}\}$、$\{x_{4t}\}$、$\{x_{5t}\}$ 均未通过 ADF 检验,所以它们是非平稳序列,考虑将这 4 个序列对数化可以使它们变平稳,然后借助 Eviews 检验对数化之后的序列是否平稳。

将城镇居民人均可支配收入、国内生产总值和私人汽车拥有量这 3 个因素对数化后可表示为 $\{\ln x_{3t}\}$、$\{\ln x_{4t}\}$、$\{\ln y_t\}$。采用 Eviews 中的 ADF 检验可以得到检验 P 值分别为 0.014 5,0.021 8,0.045 7,它们均小于显著性水平 $\alpha=0.05$。由此可知 $\{\ln x_{3t}\}$、$\{\ln x_{4t}\}$、$\{\ln y_t\}$ 这 3 个序列具有稳定性。然而 $\{\ln x_{5t}\}$ 与 $\{\ln y_t\}$ 非同阶单整,所以剔除序列 $\{\ln x_{5t}\}$。

由于涉及的自变量个数比较多,变量之间也可能会产生多重共线性,如果仅仅使用线性回归来分析输入变量与响应变量之间的关系,就会影响到参数估计的精确度。因此,利用转移函数的结构形式来构建模型能够避免发生以上问题。下面借助 R 软件的 forecast 程序包中的 arima 函数来对输入变量定阶:得到 $\{\ln x_{3t}\}$ 的拟合模型是 AR(1),AIC 值为−18.4;$\{\ln x_{4t}\}$ 的拟合模型是 AR(1),AIC 值为−12.07;对残差序列进行 LB 检验的结果为 P 值均显著大于显著性水平 $\alpha=0.05$,这表明模型拟合效果好。

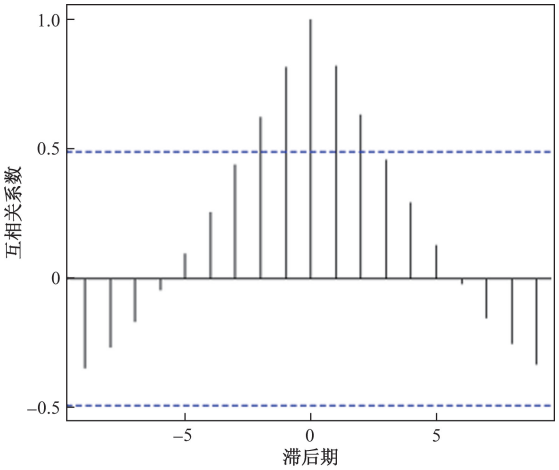
对城镇居民人均可支配收入对数化后建立如下拟合模型:

$$\ln x_{3t}=9.981\ 5+0.989\ 2\ln x_{3t-1}+\varepsilon_t \qquad (10)$$

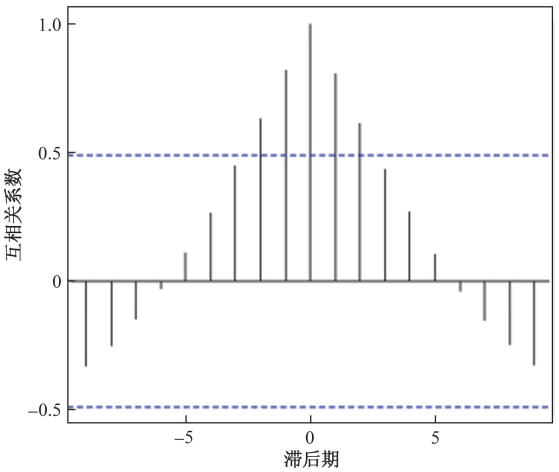
对国内生产总值对数化后建立如下拟合模型:

$$\ln x_{4t}=12.999\ 2+0.988\ 46\ln x_{4t-1}+\varepsilon_t \qquad (11)$$

根据图 4 绘制的 $\{\ln y_t\}$ 与 $\{\ln x_{3t}\}$、$\{\ln x_{4t}\}$ 的互相关图可知,序列在滞后阶数为 0 时相关系数最大,说明延迟 0 阶时最相关,可以同期建模。



(a) $\ln y_t$与$\ln x_{3t}$的互相关图



(b) $\ln y_t$与$\ln x_{4t}$的互相关图

图 4　互相关图

将上述关于城镇居民人均可支配收入与国内生产总值的两个模型作为输入变量模型运用到ARIMAX 模型中，借助 R 软件中的 TSA 程序包里的 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，那么 ARIMAX 模型的拟合精度高于没有考虑到影响因素的ARIMA模型。通过 LB 检验得到的残差序列的 P 值大于显著性水平 $\alpha = 0.05$。说明模型信息已提取充分。此外，利用条件最小二乘法估计模型参数，检验结果显示各参数亦均有统计学意义，即模型显著有效。该模型表达式为

$$\begin{cases} \ln y_t = -10.365\,6 + 1.331\,1\ln x_{3t} + 0.454\,7\ln x_{4t} + \varepsilon_t \\ \varepsilon_t = \dfrac{1}{1 - 0.607\,6B + 0.684\,1B^2} a_t \end{cases} \tag{12}$$

3.3　模型预测比较

利用上述建立的 ARIMAX 模型对测试集(2021 年、2022 年)的全国私人汽车拥有量进行预测，并与 ARIMA 模型进行对比分析，结果见表 4。

由表 4 可知，ARIMAX 模型的预测值更接近于真实值。ARIMAX 模型的 AIC 值为－54.59，ARIMA模型的 AIC 值为－26.45，通过相对误差对预测效果进行定量评估可知，ARIMAX 模型的相对误差更小，因此，ARIMAX 模型的预测精度要高于 ARIMA 模型。

表 4　2021 年、2022 年全国私人汽车拥有量预测结果

年份	拥有量/万辆		
	ARIMA 预测值	ARIMAX 预测值	真实值
2021	24 551.05	25 657.47	26 152.02
2022	27 094.77	28 290.40	27 873

表 5　2021 年、2022 年不同方法的相对误差

年份	相对误差/%	
	ARIMA 模型	ARIMAX 模型
2021	－6.12	－1.89
2022	－2.79	1.50

4　结论

以 2015—2020 年全国的私人汽车拥有量数据为基础，利用 Lasso 回归和灰色关联分析方法综合筛选了变量，结果表明公路营运汽车拥有量、城镇居民人均可支配收入和国内生产总值是影响全国私人汽车拥有量的 3 个关键因素；随后将它们作为输入变量引入到模型中，分别构建了 ARIMA 模型与多因素影响的 ARIMAX 模型来对私人汽车拥有量进行预测。借助 R 软件对 ARIMA 模型进行参数估计与残差检验，结果显示建立的 ARIMA(1，0，2)模型是合理的。由于私人汽车拥有量与城镇居民人均可支配收入、国内生产总值均在滞后阶数为 0 时相关系数最大，可以同期建立 ARIMAX 模型，利用 LB 检验对模型的残差序列检验通过，说明建立的 ARIMAX 模型有效。

从实证分析结果来看，ARIMAX 模型具有更小的 AIC 值，预测结果的相对误差也更小，即ARIMAX模型的预测精度优于 ARIMA 模型，更适用于私人汽车拥有量的预测。通过该模型对未来私人汽车拥有量进行预测，能够为汽车产业的未来经营和发展提供依据。在未来的研究中，还可以进一步考虑将基于定量分析与政策定性分析相结合，从而实现更科学的预测。

参考文献

[1]　张琪. 中国城市私人汽车拥有量的影响因素分析[D]. 大连：大连理工大学，2016.

[2]　周亚林，叶琴，郭杰，等. 基于机器学习的私人汽车保有

量影响因素分析及预测：以新疆为例[J]. 交通运输研究，2022，8(4)：74-82.

[3] 杨昆，时燕，罗毅，等. 经济快速发展背景下中国民用汽车拥有量变化的时空特征[J]. 地理科学，2019，39(4)：654-662.

[4] KAI Z，HANG Z，XUEGE Z. An exponential curve model and its application in forecasting private car ownership of China [J]. Journal of Mathematics，2022，2022：6850263.

[5] 郭艳莉. 基于灰色-广义回归神经网络的私人汽车拥有量分析及预测[D]. 北京：华北电力大学(北京)，2023.

[6] 李炳炎，李世龙，廖月彬，等. 江苏省私人汽车拥有量预测及其影响因素分析：基于 VAR 模型[J]. 时代汽车，2023(4)：7-10.

[7] 王燕. 时间序列分析：基于 R[M]. 北京：中国人民大学出版社，2015.

Application of ARIMA and ARIMAX Models in Predicting Private Car Ownership

ZHANG Shuxian

(School of Mathematics and Physics，Anhui Jianzhu University，Hefei 230601，China)

Abstract：In order to improve the prediction accuracy of private car ownership，the data of private car ownership in China from 2005 to 2020 is analyzed by using time series analysis method，and a prediction model based on dynamic regression(ARIMAX) model is established. Lasso model and grey correlation analysis are used to get the main factors affecting private car ownership，and the main factors are introduced into autoregressive integrated moving average(ARIMA) model as regression terms. The ARIMAX model is established on the basis of the ARIMA model. Through the comparison of model prediction，it is found that ARIMAX model has better fitting effect，which is suitable for the prediction of private car ownership in China.

Keywords：dynamic regression(ARIMAX) model；Lasso model；private car ownership