美国轻型汽车零售市场销售趋势探究

统计与管理学院

辛柏赢 2020111753

一、 案例背景

轻型汽车通常是指相较于传统汽车而言车身质量更轻的一种车型，其中既包括载人用的客运车辆，也包括一些运输用途的轻型卡车等。其轻巧的车身提高了燃油效率，降低了能源成本，也因此受到了许多消费者的青睐。此外随着节能减排的观念的不断普及以及新能源的广泛利用，尤其在货物运输方面，也能看到越来越多的轻型电动卡车正逐渐走向市场。在美国，如雷克萨斯、奥迪、英菲尼迪等都是消费者非常信任的（轻型）汽车生产商。

在本报告中，着重收集了美国 1976 年至 2019 年共 44 年的年度轻型汽车零售销售额。考虑到美国本身具有的相当庞大规模的汽车保有量等，其轻型汽车的销售趋势与规律对于我国对应市场的研究也具有一定的指导意义。尤其是在当下低碳环保的发展趋势下，轻型汽车在日后亦会具有更大的市场潜力与发展空间[1]。

二、 数据介绍与描述性统计

（一） 数据获取与说明

数据由商用数据网站 statista.com 提供下载，其对应的数据整理自美国商业经济分析局（U.S. Bureau of Economic Analysis, www.bea.gov）。数据共 44 条记录，没有缺失值或异常值，具体说明如表 1 所示。

在本报告中，预留了 2017～2019 三年的销售数据作为测试集验证模型拟合效果，其余 41 条数据作为训练集对模型进行拟合。

表 1 变量说明表

变量名称	变量含义	变量示例
year	时间数据	1976
sales	轻型汽车销售量（单位：千台）	12969.8

（二） 描述性统计分析

从下图 1 的时序图中可以发现，整体的轻型车销售呈现显著的波动趋势。可能由于国际局势、能源供应、金融危机等外部因素影响，在约 1978、1986、2007 年左右都出现了比较显著的下降，而其余年份普遍上升。且总体销售水平普遍在触底反弹后略有增加。

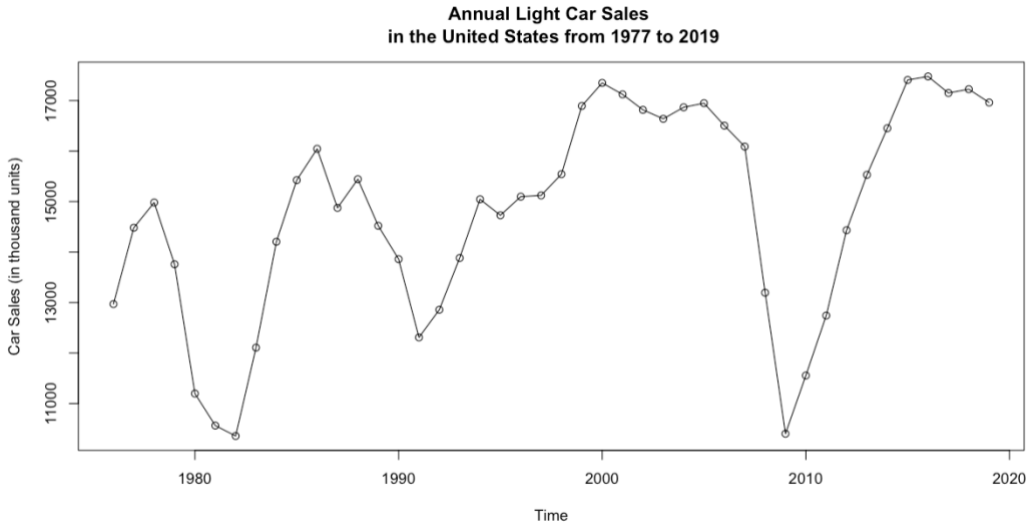

图 1 美国汽车 1977～2019 年轻型汽车销量时序图（单位：千台）

对数据另外绘制其分布 Q-Q 图。由图 2 可以发现轻型车销量呈现一定的左偏趋势，这也和上述数据中的几个特殊影响相对应：由于某些事件的冲击导致轻型车的销量出现了较为极端的下滑，在其余时刻普遍销量维持在较高水平。

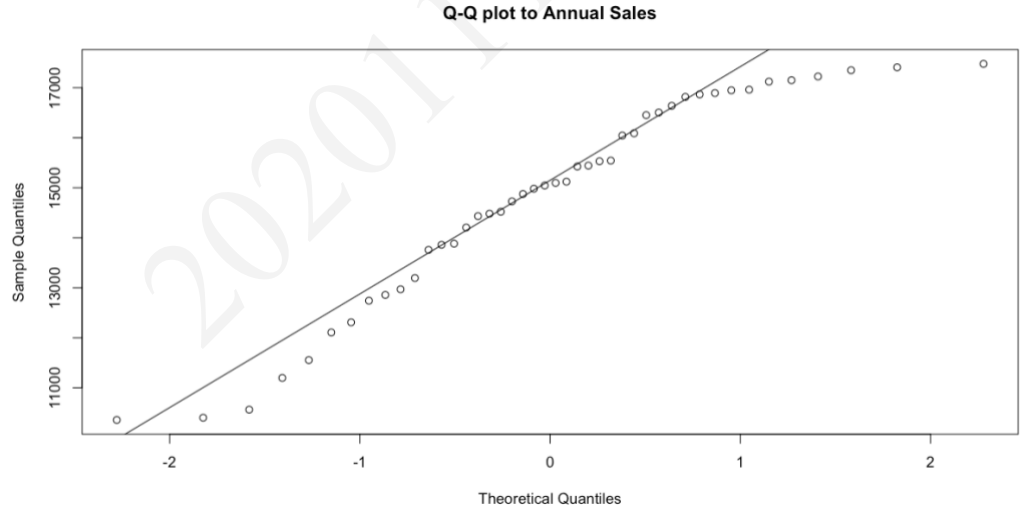

图 2 轻型车销量 Q-Q 图

三、 ARIMA 模型建模与预测

（一） 模型的平稳化及相应检验

由上述时序图可以看出原数据显然是非平稳的。为拟合 ARIMA 模型，需要尝试对模型进行平稳化处理。

首先对模型尝试进行一次差分，计算其轻型车年度销售增量。一次差分后的时序图如图 3 所示。由时序图可以发现增量数据的平稳性得到大幅改善。为进一步定量分析其平稳水

平，进一步对数据进行 ADF 单位根检验[2]。此处采用零均值假设，并由后续的 PACF 计算结果暂时对模型做出$AR(3)$的拟合假设。经计算可知其 Dickey- Fuller 统计量为-4.0044，对应的 p 值小于 0.01，故可以拒绝原假设。更进一步，参考 Schwert 的相关实证经验意见，其建议的 ADF 检验最大滞后阶数为：

$$k_{max} = [12(T/100)^{\frac{1}{4}}]$$

其中T为模型的序列长度，$[\cdot]$表示取整[3]。

代入本报告可以算得相应的推荐最大滞后阶数为 9，再次进行 ADF 检验，得到相应统计量为-2.5953，对应的 p 值为 0.01148，仍然拒绝原假设。综合上述分析可知，直到滞后 9 阶的模型假设下都可以认为该差分序列都是统计意义上显著平稳的。

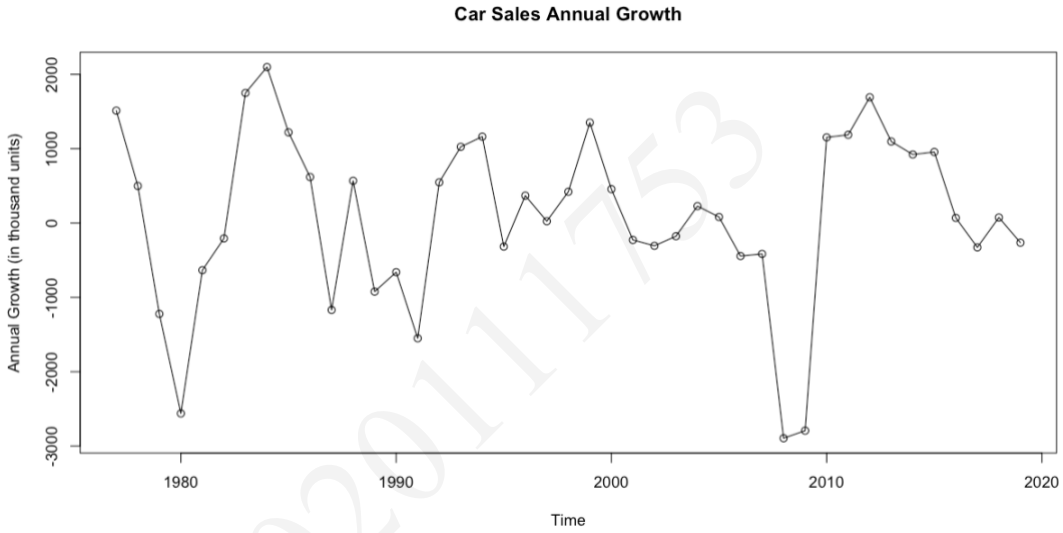

图 3 轻型车年度销售增量

进一步再次尝试对数差分变化，如图 4 所示。由时序图可见在经过对数差分后，其大致增长趋势与原差分序列基本一致，但取对数操作会额外增加模型的复杂性与解释难度。故在后续的报告中将主要针对差分数据进行研究。

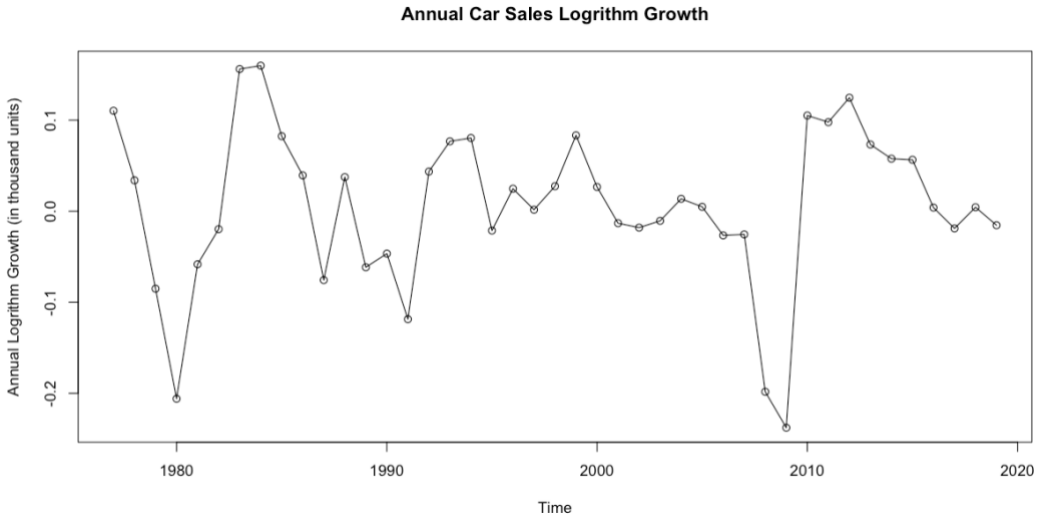

图 4 轻型车销量对数增长额度

（二） 模型定阶

对于一阶差分后的序列，在确认其平稳性后尝试对其进行初步的定阶工作。首先求解其样本 ACF 及 PACF 数据并绘图如图 5 所示。一方面由 ACF 结果可知可以尝试建立$MA(1)$模型；另一方面由 PACF 结果，其在滞后 3 阶处的 PACF 略超出接受区间，故可以尝试建立$AR(3)$模型。此外，考虑到模型的简洁性与解释性等，暂不考虑由 PACF 结果拟合$AR(12)$模型。

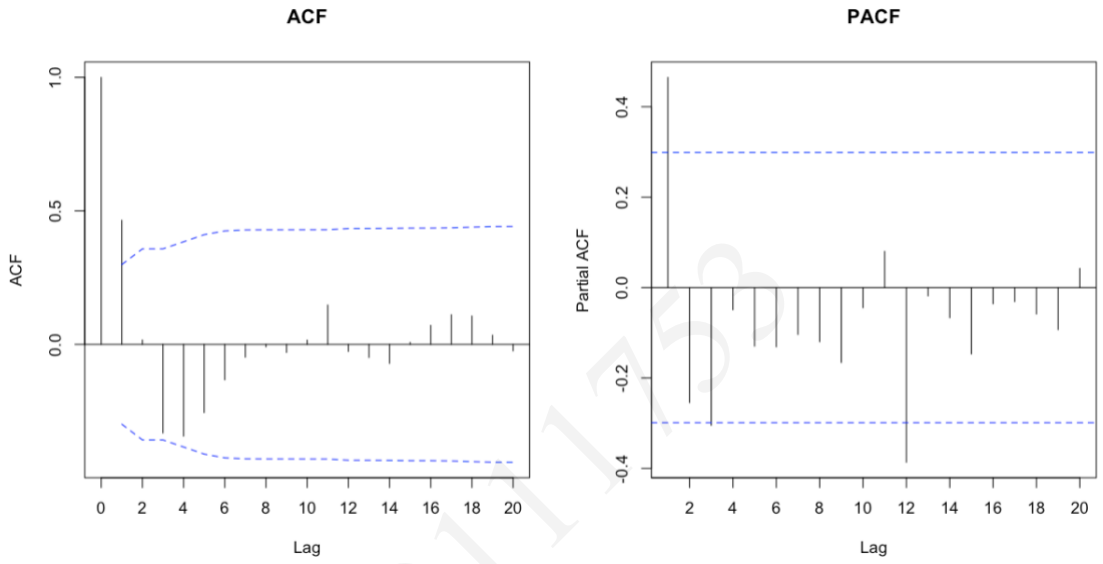

图 5 差分数据样本 ACF 及 PACF

（三） 模型拟合与修订

这里将尝试分别对这两种模型进行拟合，并通过信息准则及测试集对预测效果对模型进行评估。需要注意的是，由于上述定阶的工作是基于一阶差分的数据进行的，故在下列模型拟合过程中将相应转化为$ARIMA(0,1,1)$模型和$ARIMA(3,1,0)$模型的拟合与评估。

为叙述方便，下对模型记号进行统一说明：记S_t为第t年的美国轻型车销量情况，a_t为对应年份的无法由模型解释的随机扰动项，B为滞后算子。

1. $ARIMA(3,1,0)$模型拟合与修订

首先建立如下的$ARIMA(3,1,0)$模型：

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(S_t-S_{t-1})=a_t \quad \cdots(1)$$

通过训练集数据对模型进行估计，这里在极大似然估计准则下，得到如下估计模型：

$$(1-0.5266\,B+0.0615\,B^2+0.3301\,B^3)(S_t-S_{t-1})=a_t \quad \cdots(2)$$

相似地，通过条件最小二乘估计方法得到的模型估计如下：

$$(1-0.5096\,B+0.0264\,B^2+0.3303\,B^3)(S_t-S_{t-1})=a_t \quad \cdots(3)$$

可以看到对于该模型，两种估计方法得到的估计参数较为接近。

现对极大似然方法下的估计模型进行残差检验如下，检验结果如下图 6、图 7 所示。

由残差时序图可知，残差基本上围绕 0 均值上下波动，其波动方差亦基本上保持在一定水平中。说明$ARIMA(3,1,0)$模型大致可以提取大多数的有效信息。值得注意的是，在约 2008 年对应数据上仍然存在着较大波动，这里认为与当时次贷危机等外部冲击导致的购买力显著下降相关。

残差正态性方面，由残差直方图及 Q-Q 图可以看出，其分布大致呈现钟形曲线模式，大多数残差点也近似落在 Q-Q 图中的参考直线上。除少部分极端值点，大多数的数据均近似服从正态分布。进一步地，对残差序列进行 Shapiro-Wilk 正态性检验[4]。其检验的原假设为H_0: *数据是平稳的*。经计算可知相应统计量为$W = 0.98238$，其对应的 p 值为 0.7649，故在 5%的水平下不能拒绝原假设。因此可以认为残差在统计意义上是基本服从正态分布的。

残差相关性检验如图 7 所示。由残差 ACF 可见，在不同滞后阶上残差彼此都是不相关的。进一步通过 Ljung-Box 检验衡量其联合相关情况，亦可发现在最高直到 15 阶的情况下其相应统计量的 p 值均大于 0.3，可以较充分地印证上述残差不相关的结论。

综合上述几个方面，可以认为该$ARI(3,1)$模型基本通过了残差检验。

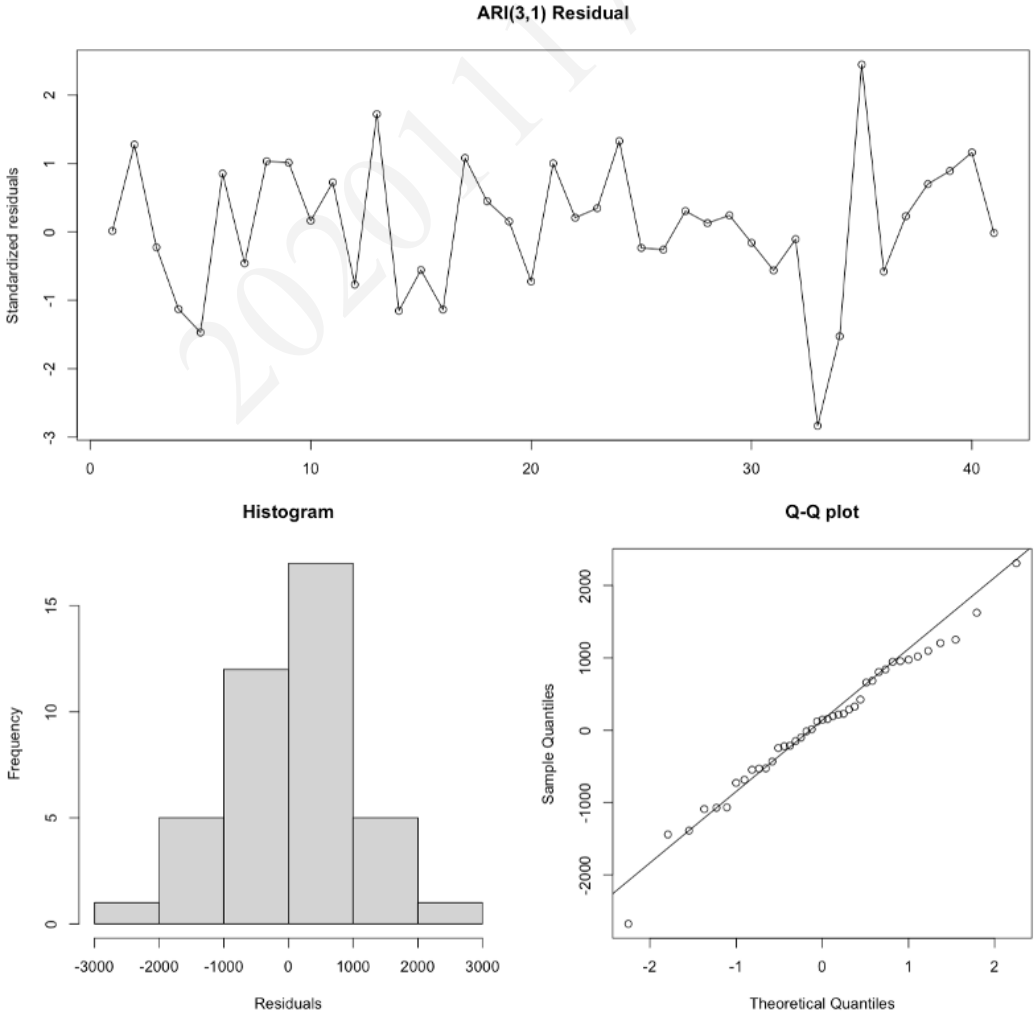

图 6 ARI(3,1)模型残差时序图、直方图与 Q-Q 图

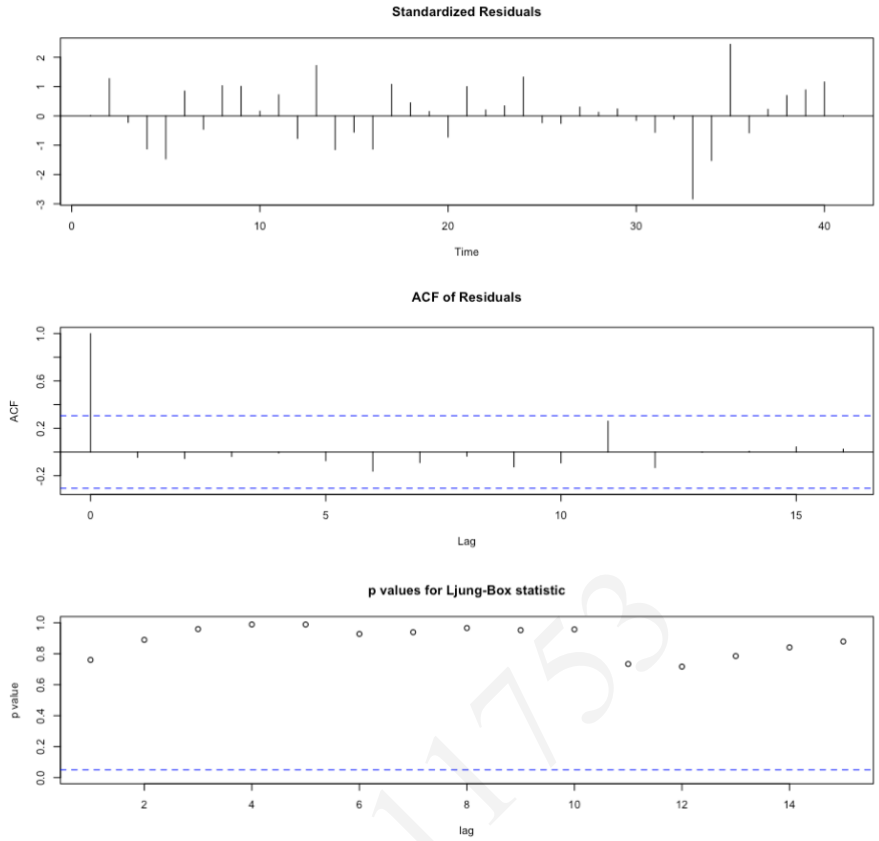

图 7 ARI(3,1)模型残差相关性检验

2. $ARIMA(0,1,1)$模型拟合与修订

与上述过程类似，首先给出欲建立的$ARIMA(0,1,1)$模型：

$$S_t - S_{t-1} = a_t + \theta a_{t-1} \quad \cdots(4)$$

类似地，通过极大似然估计方法得到如下拟合模型：

$$S_t - S_{t-1} = a_t + 0.4398 a_{t-1} \quad \cdots(5)$$

通过条件最小二乘得到如下拟合模型：

$$S_t - S_{t-1} = a_t + 0.4348 a_{t-1} \quad \cdots(6)$$

亦不难发现两估计方法得到的估计量较为相近。

下对极大似然估计得到的模型进行残差分析，结果如图 8、图 9 所示。由残差相关图象可知，整体残差依然围绕 0 均值进行波动，但相较于前述$ARIMA(3,1,0)$模型而言，其正态性有所偏离。再观察其 ACF 图，可以发现在第 3 阶时出的残差 ACF 略超出接受区间。这提示残差中存在着一定的关联性。进一步从 Ljung-Box 检验图中可以看出，在 3～5 阶时滞处其相关统计量值均接近临界值区间，这也与 ACF 检验的结果相互印证。

但综合考虑 ACF 中残差样本的超出规定范围并不明显，Ljung-Box 检验结果在 5%区间在联合效应上并无显著相关残差，以及模型的解释性、简洁性，暂不考虑将此处的

$ARIMA(0,1,1)$进一步修改为更高阶的$ARIMA(0,1,4)$模型。不过由于在残差分析中出现的相关性、正态性问题，在后续的分析中将更倾向于上一小节所展示的$ARIMA(3,1,0)$模型。

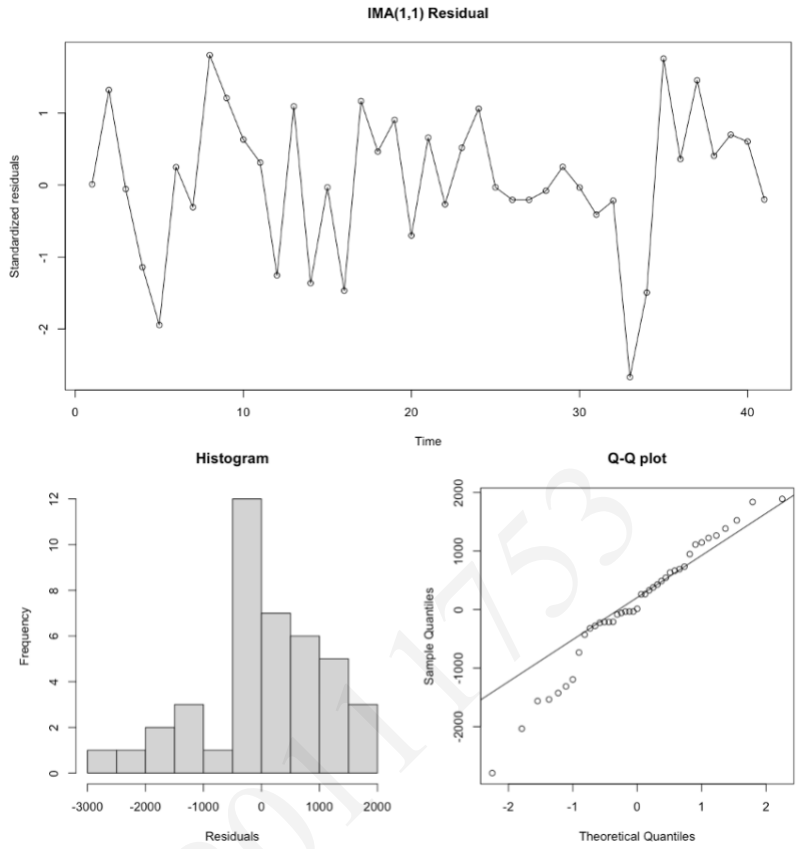

图 8 IMA(1,1) 模型残差时序图、直方图与 Q-Q 图

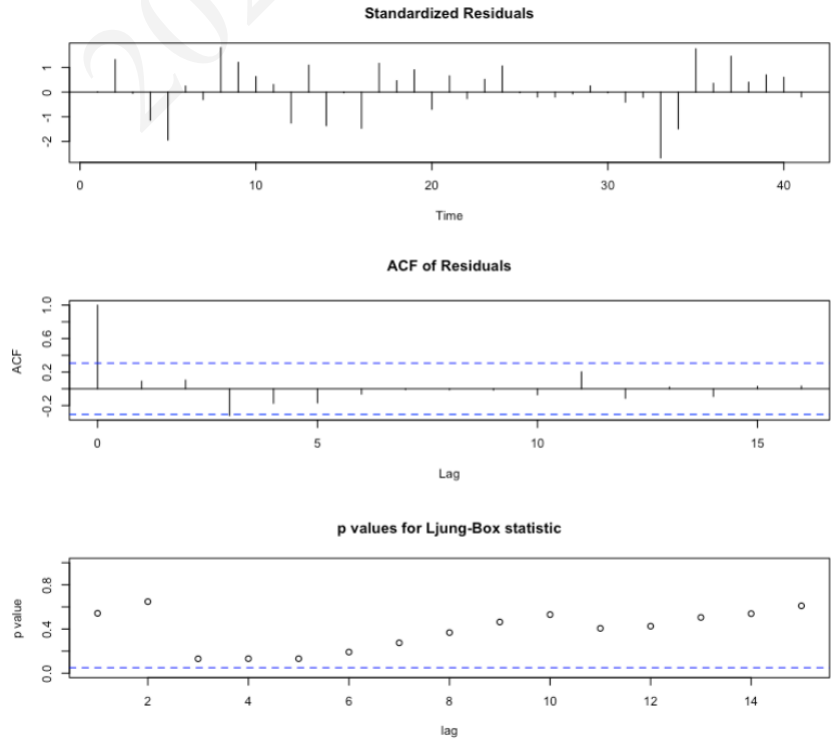

图 9　IMA(1,1) 模型残差相关性检验

（四） 模型拟合效果评估与预测

下对其估计效果进行评估。

可以分别计算两种模型的 AIC 与 BIC，以及其在训练集上的 RMSE、MAE，结果如下表 2 所示：

表 2 模型拟合效果比较

模型	AIC	BIC	RMSE	MAE
ARIMA(3,1,0)	664.30	665.99	932.61	721.97
ARIMA(0,1,1)	671.94	673.63	1033.06	792.37

从上表的各信息量及拟合优度指标来看，$ARIMA(3,1,0)$具有更小的信息量及更小的拟合误差。因此综合模型诊断建议以及拟合优度评估结果，最终选择对原始数据拟合$ARIMA(3,1,0)$进行预测。

经计算可得最终点预测及区间预测结果，计算测试集的预测误差，其数值如表 3 所示，与原始数据合并在时序图中如图 10 所示。

表 3 真实值与模型预测结果

年份	真实值	点预测值	95%区间预测值		RMSE	MAE
			区间下界	区间上界		
2017	17150.1	17150.68	15300.1	19001.26		
2018	17224.9	16658.94	13281.76	20036.12	461.2604	376.8091
2019	16961.1	16397.21	11727.97	21066.46		

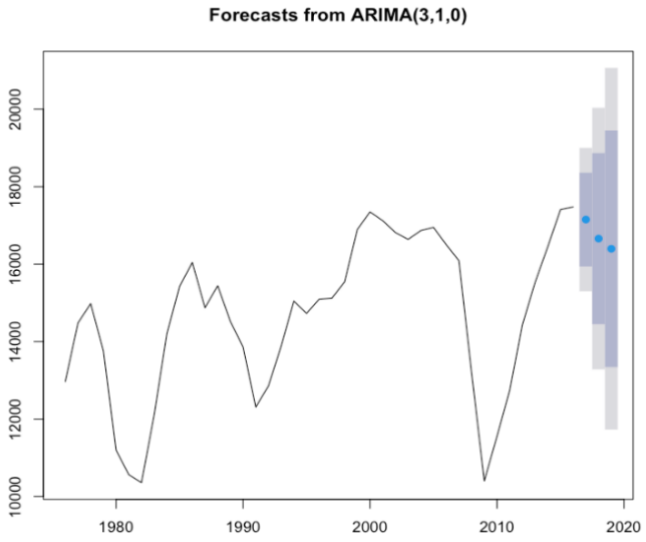

图 10 模型预测结果及预测区间

四、 总结与讨论

在本报告中主要就美国轻型汽车市场消费数据进行了时间序列分析。首先通过时序图就数据进行了平稳性分析与讨论，确定了一阶差分的平稳化方案。在模型建模方面，通过ACF、PACF 等一系列辅助工具，得到了$ARIMA(3,1,0)$ 与$ARIMA(0,1,1)$ 两个备选的 ARIMA 模型。通过对于拟合模型的残差检验以及拟合优度检验，最终确定了$ARIMA(3,1,0)$作为在当前框架下的最优模型。并通过对于划分的测试集对数据的预测效果进行了评估。可以看到模型的拟合效果较好，全部测试集的真实值均落在了 95%的预测区间内。

但是需要指出的是，由于 2020 年开始的新冠疫情的外部因素冲击，事实上轻型车销售市场乃至全球的销售市场都出现了大幅的下降。这导致现有模型对于近年的销售数据的预测能力有所下降。

但从已有的数据趋势可以看到，在美国地区，尽管有节能减排的倡议支持，轻型汽车的销售量已逐渐呈现下降的趋势。这亦可能是由于汽车市场过于饱和，对于该类型汽车的有效需求不足导致的。

通过本报告的讨论也可以进一步作为参考对我国的轻型车市场的销售状态提供一定的分析思路。

五、 参考文献

[1] 李茜, 吕力, 刘辰. 欧盟轻型车 CO_2 排放法规提案目标解读及对中国的启示 [J]. 中国汽车, 2022, (09): 28-32.

[2] DICKEY D A, FULLER W A. Distribution of the Estimators for Autoregressive Time Series with a Unit Root [J]. Journal of the American Statistical Association, 1979, 74(366a): 427-31.

[3] SCHWERT G W. Tests for Unit Roots [J]. Journal of Business & Economic Statistics, 2002, 20(1): 5-17.

[4] SHAPIRO S S, WILK M B. An analysis of variance test for normality (complete samples)† [J]. Biometrika, 1965, 52(3-4): 591-611.

六、 程序代码

```
library(TSA)
library(fUnitRoots)
library(forecast)
library(modelr)

set.seed(123)

#########################################################################

# 1. Load Data
rm(list=ls())
setwd("/Users/xinby/Desktop/Sufe/TSA/ts_proj1/TS_proj1")
dat <- readxl::read_excel("car.xlsx")

CarSale <- ts(dat, start = 1976, end = 2019)
CarSale.train <- ts(CarSale[1:41],start=1976)
CarSale.test <- ts(CarSale[42:44],start=2017)

#########################################################################
# 2. Description

plot(CarSale,type='o',ylab="Car Sales (in thousand units)",
     main="Annual Light Car Sales \n in the United States from 1977 to
2019")
qqnorm(CarSale, main="Q-Q plot to Annual Sales"); qqline(CarSale)


#########################################################################

# 3. Stationary Test

## diff(car)

diff.CarSale <- diff(CarSale) # get diff data

plot(diff.CarSale,type = 'o',main="Car Sales Annual Growth",
     ylab = "Annual Growth (in thousand units)") # plot

adfTest(diff.CarSale, lags = 3, type='nc')
adfTest(diff.CarSale,lags=9,type='nc') # adf test with lag=9
```

```r
par(mfrow=c(1,2))
acf(diff.CarSale,xaxp=c(0,20,10),lag.max=20,main="ACF",ci.type='ma') #
acf: suggest MA(1)
pacf(diff.CarSale,xaxp=c(0,20,10),lag.max=20,main="PACF") # pacf: suggest
AR(3)

eacf(diff.CarSale) # eacf: poor performance


## diff(log(car))

diff.log.CarSale <- diff(log(CarSale)) # get log diff data

par(mfrow=c(1,1))
plot(diff.log.CarSale,type = 'o',
     main="Annual Car Sales Logrithm Growth",
     ylab = "Annual Logrithm Growth (in thousand units)") # plot



adfTest(diff.log.CarSale,lags = 9,type='nc') #adf test

AIC(demo)

#########################################################################

# 4. Parameter Estimation

## ARI(3,1)

ari31.ml <- TSA::arima(x = (CarSale.train), order = c(3,1,0), method =
"ML")
ari31.css <- TSA::arima(x = (CarSale.train), order = c(3,1,0), method =
"CSS")

ari31.ml
ari31.css

BIC(ari31.ml)

## IMA(1,1)
```

```r
ima11.ml <- TSA::arima(x = (CarSale.train), order = c(0,1,1), method = "ML")
ima11.css <- TSA::arima(x = (CarSale.train), order = c(0,1,1), method = "CSS")

ima11.ml
ima11.css

#####################################################################

# 5. Model Diagnosis

## IMA(1,1)
par(mfrow=c(1,1))
plot(rstandard(ima11.ml),ylab='Standardized residuals',type='o', main = "IMA(1,1) Residual") # residual plot

par(mfrow=c(1,2))
hist(residuals(ima11.ml), xlab='Residuals', main='Histogram') # residual histogram
qqnorm(residuals(ima11.ml), main='Q-Q plot'); qqline(residuals(ima11.ml)) #residual qqplot

par(mfrow=c(1,1))
shapiro.test(rstandard(ima11.ml)) # shapiro Normality test
acf(as.numeric(rstandard(ima11.ml)), xaxp = c(0,24,12), main = "")
Box.test(rstandard(ima11.ml), lag = 6, type = "Ljung-Box", fitdf = 2)
Box.test(rstandard(ima11.ml), lag = 12, type = "Ljung-Box", fitdf = 2)
Box.test(rstandard(ima11.ml), lag = 18, type = "Ljung-Box", fitdf = 2)
Box.test(rstandard(ima11.ml), lag = 24, type = "Ljung-Box", fitdf = 2)
tsdiag(ima11.ml,gof=15,omit.initial=F, dof = 20)

## ARI(3,1)
plot(rstandard(ari31.ml),ylab='Standardized residuals',type='o', main = "ARI(3,1) Residual")

par(mfrow=c(1,2))
hist(residuals(ari31.ml), xlab='Residuals', main='Histogram') # residual histogram
qqnorm(residuals(ari31.ml), main='Q-Q plot'); qqline(residuals(ari31.ml)) #residual qqplot
par(mfrow=c(1,1))
```

```r
shapiro.test(rstandard(ari31.ml))

acf(as.numeric(rstandard(ari31.ml)), xaxp = c(0,24,12), main = "")
Box.test(rstandard(ari31.ml), lag = 6, type = "Ljung-Box", fitdf = 4)
Box.test(rstandard(ari31.ml), lag = 12, type = "Ljung-Box", fitdf = 4)
Box.test(rstandard(ari31.ml), lag = 18, type = "Ljung-Box", fitdf = 4)
Box.test(rstandard(ari31.ml), lag = 24, type = "Ljung-Box", fitdf = 4)
tsdiag(ari31.ml,gof=15,omit.initial=F)


#########################################################################

# 6. Model Comparison & Prediction

## Comparison
  ## IMA(1,1)
  ima11.ml
  ima <- forecast(CarSale.train, h=3, model =ima11.ml)
  summary(ima)
  ## ARI(1,3)
  ari31.ml
  ari <- forecast(CarSale.train, h=3, model =ari31.ml)
  summary(ari)

## ARI prediction

sale_pred <- forecast::forecast(CarSale.train,h=3, model=ari31.ml,
                                xlab='Year',ylab='Car Sale (in thousand
units)')
sale_pred
predict(ari31.ml,n.ahead=3)
CarSale.test
plot(sale_pred)

accuracy(sale_pred,CarSale.test)
```