基于乘法季节模型的月航线客运量模型分析

黄泳瑜*，熊薪叶†，王晨凯‡

数学系，南方科技大学

2019 年 12 月

摘要

我们观察并分析了 1949 年到 1960 年月航线客运量这一数据，建立了相应的时间序列模型并进行了预测。这篇报告主要包括了模型识别、拟合、诊断和预测等部分。在模型识别部分，通过观察样本的自相关函数以及差分结果，初步选择建立 $ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型。接着我们对模型进行了拟合，得到了具体参数。在模型诊断部分我们分析了残差并且进行了过拟合，确认了 $ARIMA(0,1,1)\times(0,1,1)_{12}$ 为最优模型。最后我们用建立的时间序列模型进行了近两年的预测。

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

*邮箱地址: 11713034@mail.sustech.edu.cn;
†邮箱地址: 11713039@mail.sustech.edu.cn;
‡邮箱地址: 11710619@mail.sustech.edu.cn.

1 数据

1.1 背景介绍

George E. P. Box，英国统计学家，主要从事质量控制、时间序列分析、实验设计和贝叶斯推理等领域的工作。他被称为“20 世纪最伟大的统计学家之一”。Gwilym Meirion Jenkins，英国统计学家、系统工程师。他最出名的是与 George E. P. Box 在时间序列分析中的 Box- jenkins 模型。除此之外，最先由 Box 和 Jenkins 研究的月航线客运量时间序列也被视为典型的时间序列。因此我们选取了这一时间序列进行分析和研究。

1.2 数据来源

我们直接从 R 的一个名为”TSA” 的包里读取了”AirPassengers” 这个数据集进行数据分析。

1.3 数据截取

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

1.4 时间序列图

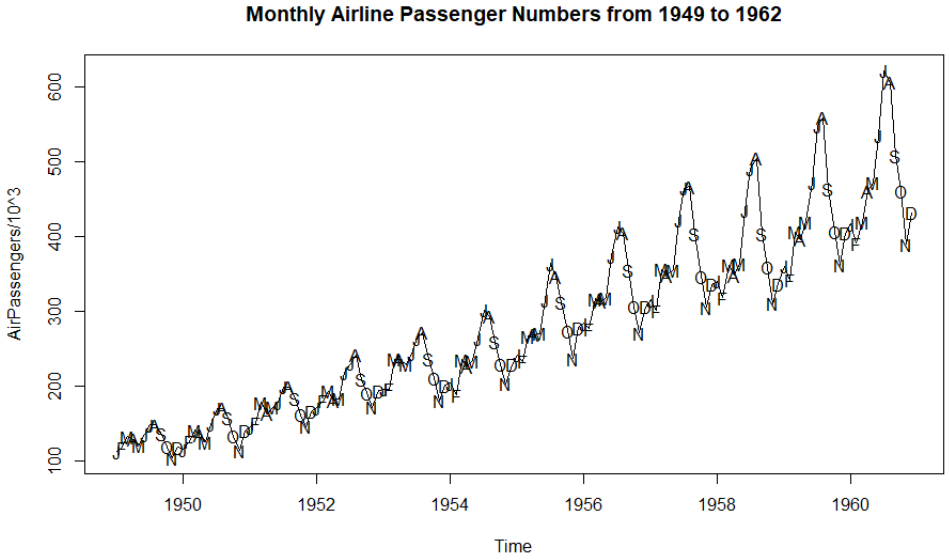

图 1: 1949-1962 年月航线客运量时间序列图

图 1为 1949 年 1 月到 1960 年 12 月每月国际航班乘客总人数（单位: 千）。通过上图发现：每年的乘客总人数逐年递增，乘客人数呈现季节趋势，且乘客数量的分散性与年份呈显著正相关。通过上述信息，我们将尝试建立一个非平稳模型。

2 模型识别

2.1 原数据的识别和分析

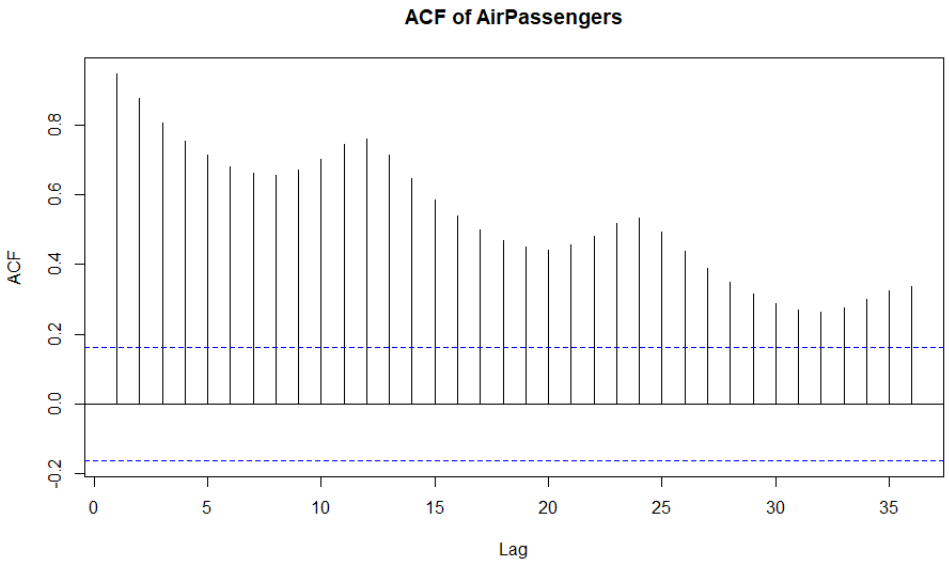

图 2: 样本自相关函数图

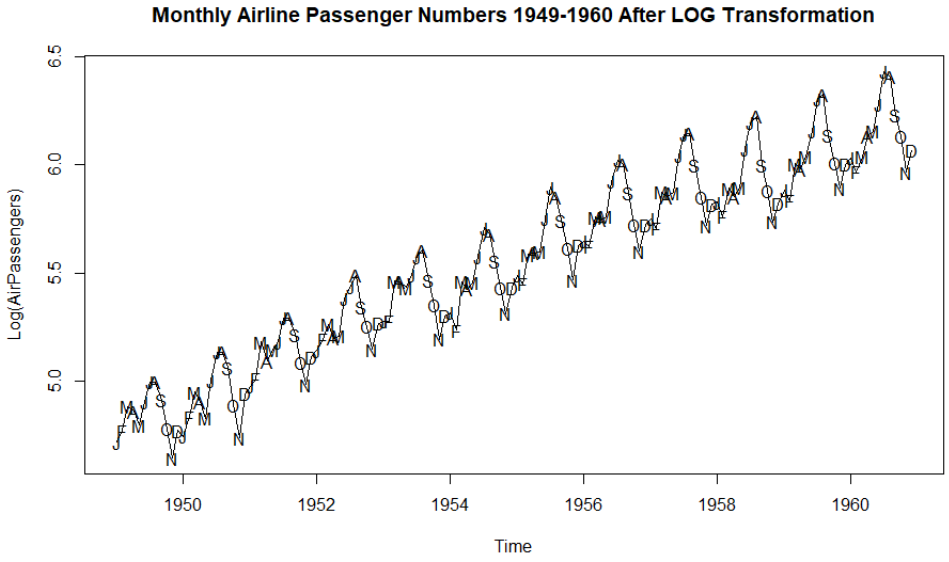

图 3: 对原数据取对数后的时间序列图

从图 1可以看出，这组时间序列数据同年中不同月份的乘客人数的分散性逐年递增。所以首先对这组数据取自然对数得到 $logY_t$, 绘制在图 3中。由图可以看出，分散性与年份的显著正相关性消失。绘制 $logY_t$ 的自相关函数图如图 4。

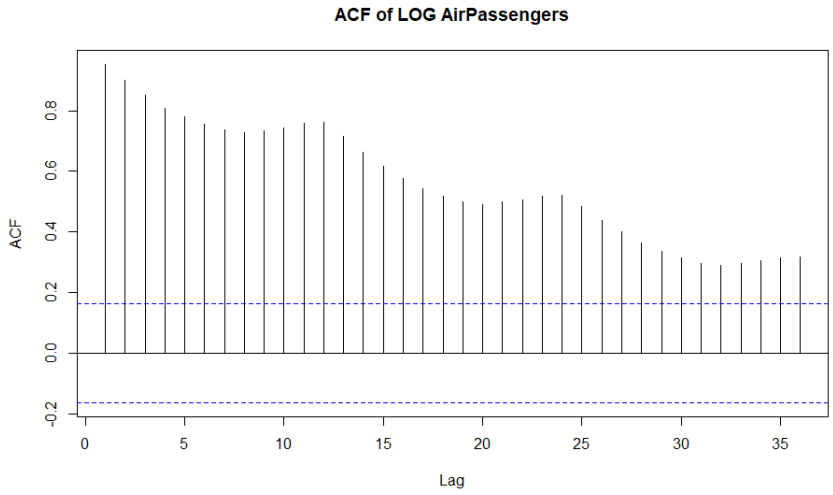

图 4: 取对数后的样本自相关函数图

2.2 一阶差分 $logY_t$ 和一阶季节差分 $\nabla logY_t$

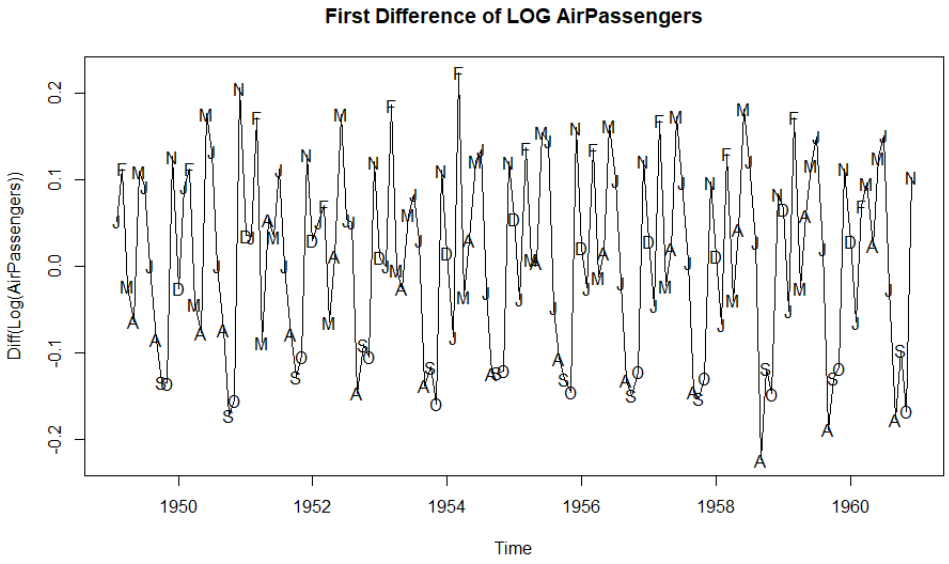

图 5: 样本取对数并进行一阶差分后的时间序列图

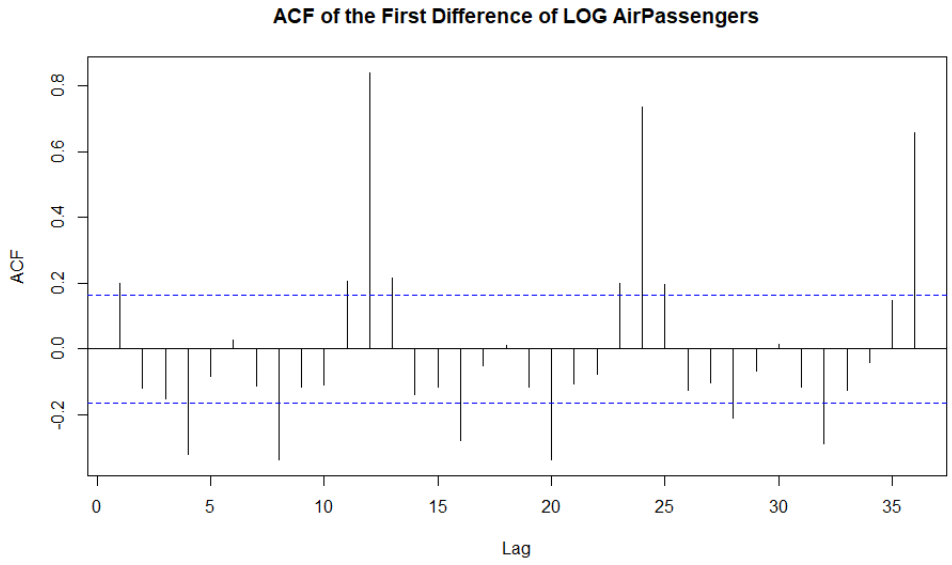

图 6: 样本取对数并进行一阶差分后的自相关函数图

图 1出现了一般性的上升趋势，所以我们将 $logY_t$ 进行一阶差分得到 $\nabla logY_t$, 绘制到图 5中。由图我们可以看出序列中一般性的上升趋势已经不存在，而强烈的季节性仍然存在。图 6为 $\nabla logY_t$ 的自相关函数。由图 6可知，$\nabla logY_t$ 在滞后 12，24，36 上具有强自相关性。这也验证了 $\nabla logY_t$ 的季节性。为了将这种趋势平稳化，我们接下来尝试使用季节差分法。

图 7显示了一次差分和季节差分后的 $logY_t$:$\nabla_{12}\nabla logY_t$ 的时间序列图。此时，明显的季节性已经消失了。绘制 $\nabla_{12}\nabla logY_t$ 的自相关函数图（如图 8）。图 8印证了经两次差分后的时间序列已经不再具有自相关性。此图也说明也许建立一个只在滞后 1 和 12 上具有自相关性的简单模型就够了。

结论：我们初步考虑为 $logY_t$ 建立乘法季节模型：$ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型。最终是否取这个模型取决于我们后面的模型诊断的结果。

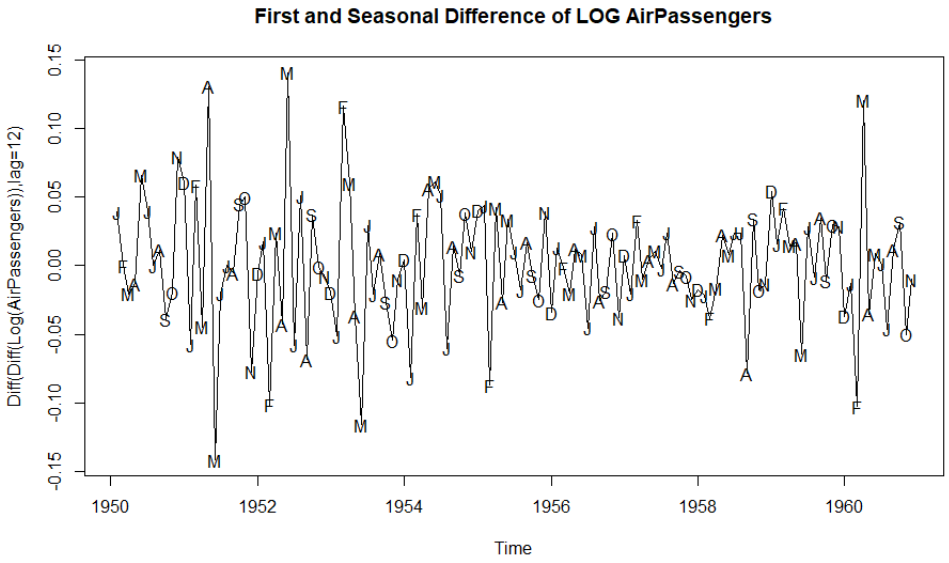

图 7: 样本取对数并进行一阶季节差分后的时间序列图

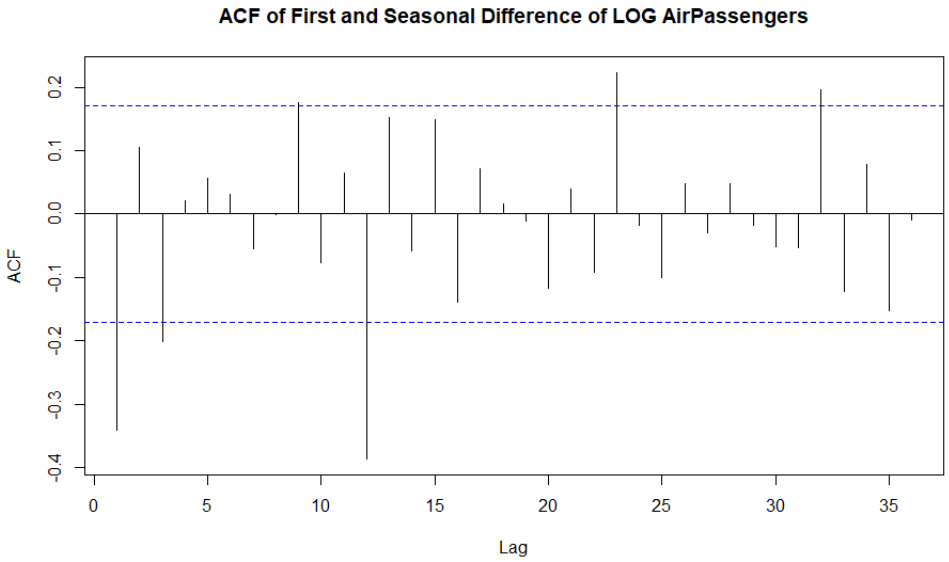

图 8: 样本取对数并进行一阶季节差分后的自相关函数图

3 模型拟合

接下来我们将尽可能有效地估计试探性模型：$ARIMA(0,1,1)\times(0,1,1)_{12}$ 中的参数。

用直接运行 R 软件内置的 *arima* 函数我们得到如下数据:

```
Call:
arima(x = LogAirPassengers, order = c(0, 1, 1), seasonal = list(order = c(0,
1, 1), period = 12))

Coefficients:
ma1      sma1
−0.4018  −0.5569
s.e.  0.0896  0.0731

sigma^2 estimated as 0.001348:  log likelihood = 244.7,  aic = −485.4
```

系数	θ	Θ
标准值	-0.4018	-0.5569
标准误差	0.0896	0.0731
$\sigma_e^2 = 0.001348$	对数似然值 = 244.7	AIC = -485.4

上表给出了最小二乘估计及其标准误差。然后我们用极大似然估计的方法再次评估数据，得到的参数及其标准误差与上表完全相同。上述两种估计的所有的系数估计值都是完全相同的，所以取 $\theta = -0.4018, \Theta = -0.5569$。接下来将对此模型加以检验。

4 模型诊断

4.1 残差图

为了对估计后的 $ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型进行诊断，我们先获取残差的时间序列图。

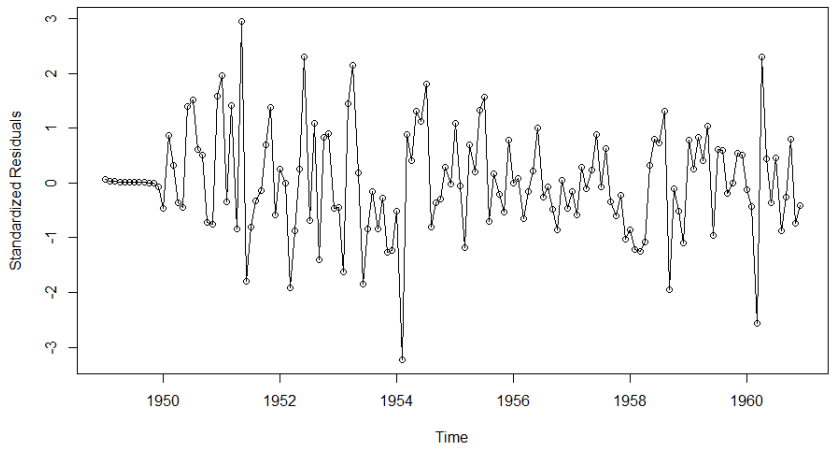

图 9: $ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型的残差

4.2 残差 ACF

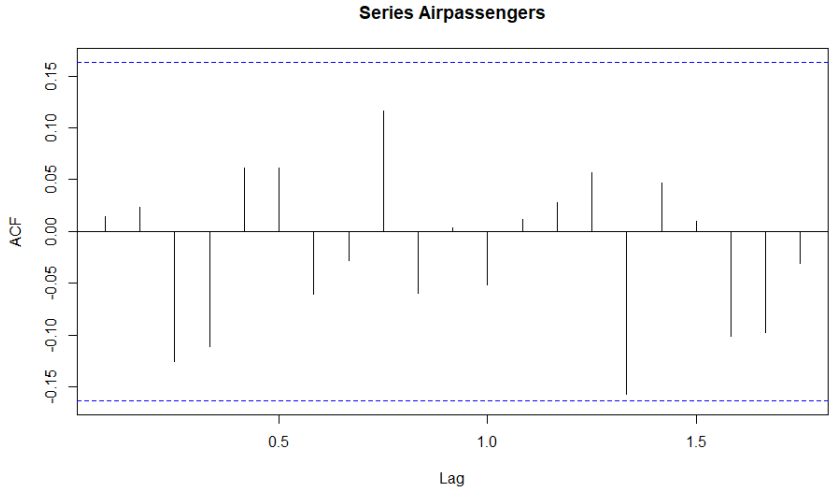

图 10: $ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型的残差的样本自相关函数

图 9给出了标准残差图，除去序列中间的少许异常，此残差图并未表现明显的不规则性。

图 10绘出了残差的 ACF 以便进行进一步的检验。现在我们对模型进行 Ljung-Box 检验：

```
> signif(acf(residuals(ap),plot = F)$acf[1:6],2)
[1]  0.014  0.024 -0.130 -0.110  0.061  0.061
```

$K = 6$ 时，$Q_* = 144(144+2)\left(\frac{0.014^2}{144-1} + \frac{0.024^2}{144-2} + \frac{(-0.130)^2}{144-3} + \frac{(-0.110)^2}{144-4} + \frac{(0.061)^2}{144-5} + \frac{(0.061)^2}{144-6}\right)$
≈ 5.58>0.41 (自由度为 5，显著水平为 5% 的卡方 p-value)

故不能拒绝误差项是不相关的原假设。

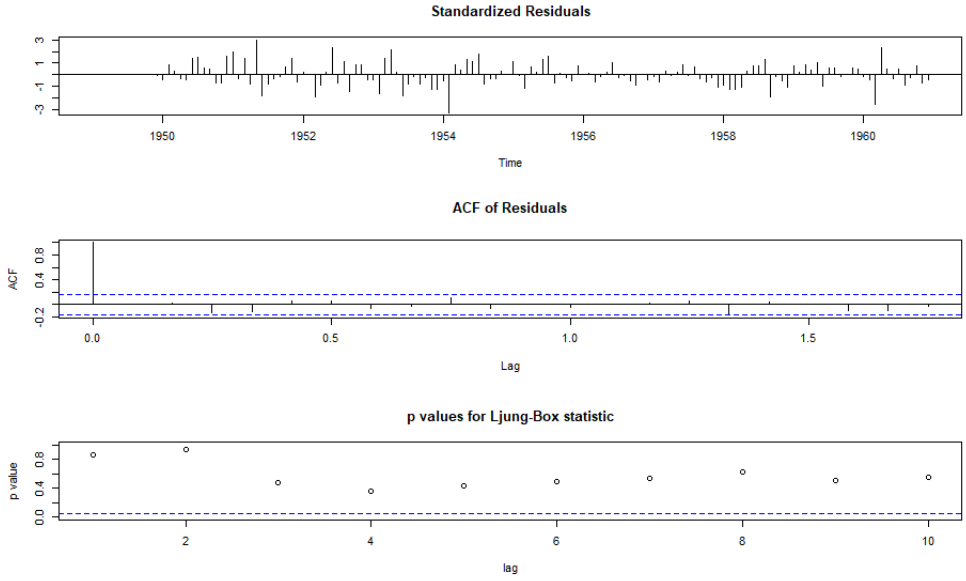

图 11: $ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型的诊断演示

图 11展示了三种诊断——标准残差序列、残差样本 ACF、K 从 0 至 10 的 Ljung-Box 检验统计量的 p 值。在该图中，所有 p 值都在 5% 水平虚线之上，故估计的 $ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型能够较好的解释 Airpassengers 的时间序列结构。

4.3 残差的正态性

下面我们借助于残差来研究误差项的正态性程度。

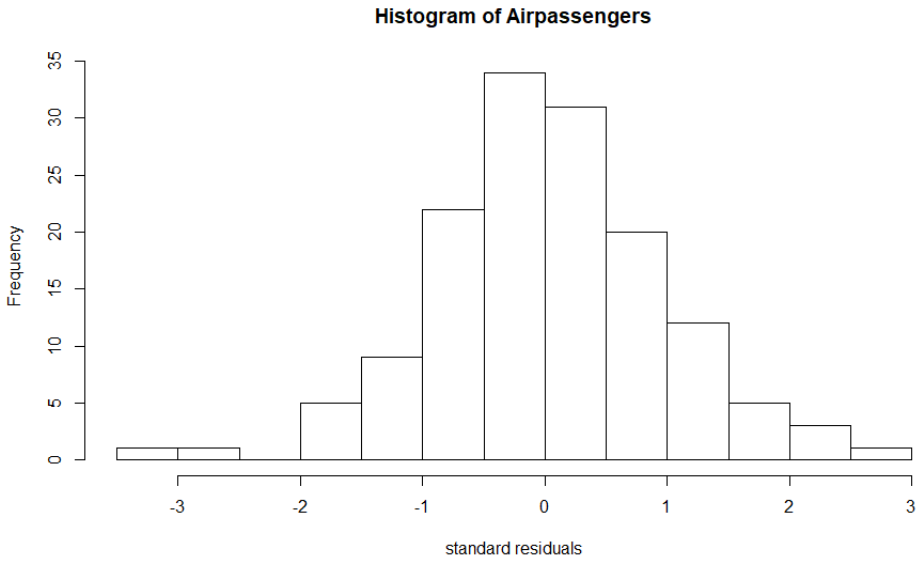

图 12: $ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型的残差直方图

图 12为残差直方图，形状与“钟形“十分相似，但不难看出似乎有些右偏，我们借助 QQ-plot 来获取更多的信息。

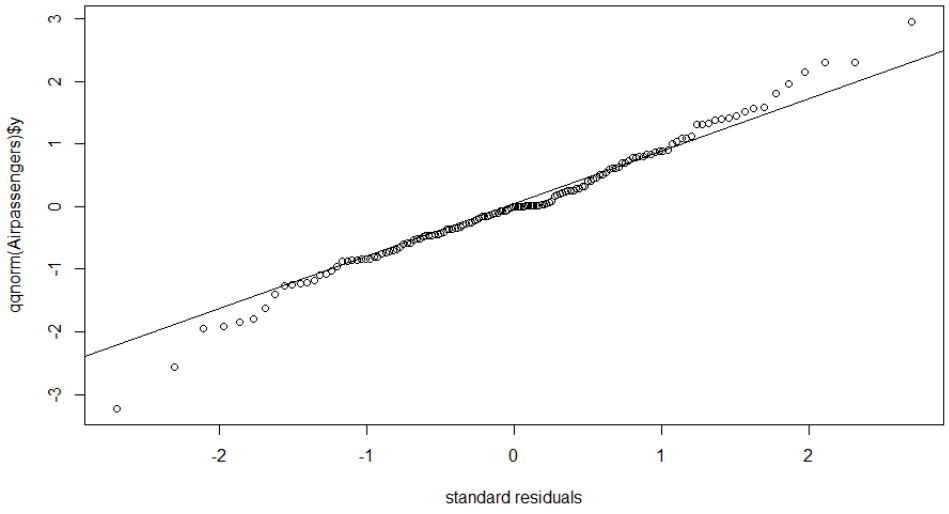

图 13: $ARIMA(0,1,1)\times(0,1,1)_{12}$ 的 Q-Q 图

从图 13我们可以看出，在大部分情况下，残差与直线的拟合程度非常高。

4.4 过度拟合

在这一部分我们将进行过度拟合，以此来确认 $ARIMA(0,1,1)\times(0,1,1)_{12}$ 是原时间序列的最优拟合。

1. 首先我们尝试 $ARIMA(0,1,2)\times(0,1,1)_{12}$ 模型。模型系数评估如下:

```
Call:
arima(x = ap1, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 1),
period = 12))

Coefficients:
ma1      ma2      sma1
−0.3961  −0.0397  −0.5590
s.e.  0.0859   0.0851   0.0732

sigma^2 estimated as 0.001345:  log likelihood = 244.81,  aic = −483.62
```

根据表所示结果，我们与最初拟合的 $ARIMA(0,1,1)\times(0,1,1)_{12}$ 进行对比：

```
Call:
arima(x = ap1, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1),
period = 12))

Coefficients:
ma1      sma1
−0.4018  −0.5569
s.e. 0.0896   0.0731

sigma^2 estimated as 0.001348:  log likelihood = 244.7,  aic = −485.4
```

与原结果进行比较可得 θ_1 与 Θ 的估计值无明显变化，尤其是在考虑进标准误差大小时。此外，新参数 θ_2 的估计值在统计上并不能显著区别于 0。并且，在 AIC 已经增加的前提下，σ^2 和对数似然估计值都没有显著变化。

2. 接着我们尝试用 $ARIMA(0,1,1)\times(0,1,2)_{12}$ 拟合原时间序列。模型系数评估如下：

```
Call:
arima(x = ap1, order = c(0,1,1), seasonal = list(order = c(0,1,2), period = 12))

Coefficients:
ma1      sma1     sma2
−0.4154  −0.5979  0.0685
```

```
s.e.    0.0900    0.0946    0.0910

sigma^2 estimated as 0.00134:  log likelihood = 244.98,  aic = -483.96
```

由表中的数据可知，θ_1 与 Θ 的估计值都与初始模型相比并无明显变化，且新参数在统计意义上为 0。还应注意，在 AIC 增加的情况下，σ^2 和对数似然估计值都没有显著变化。因此，我们确定数据适用于 $ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型。接下来我们将用该模型来预测。

5 模型的预测和解释

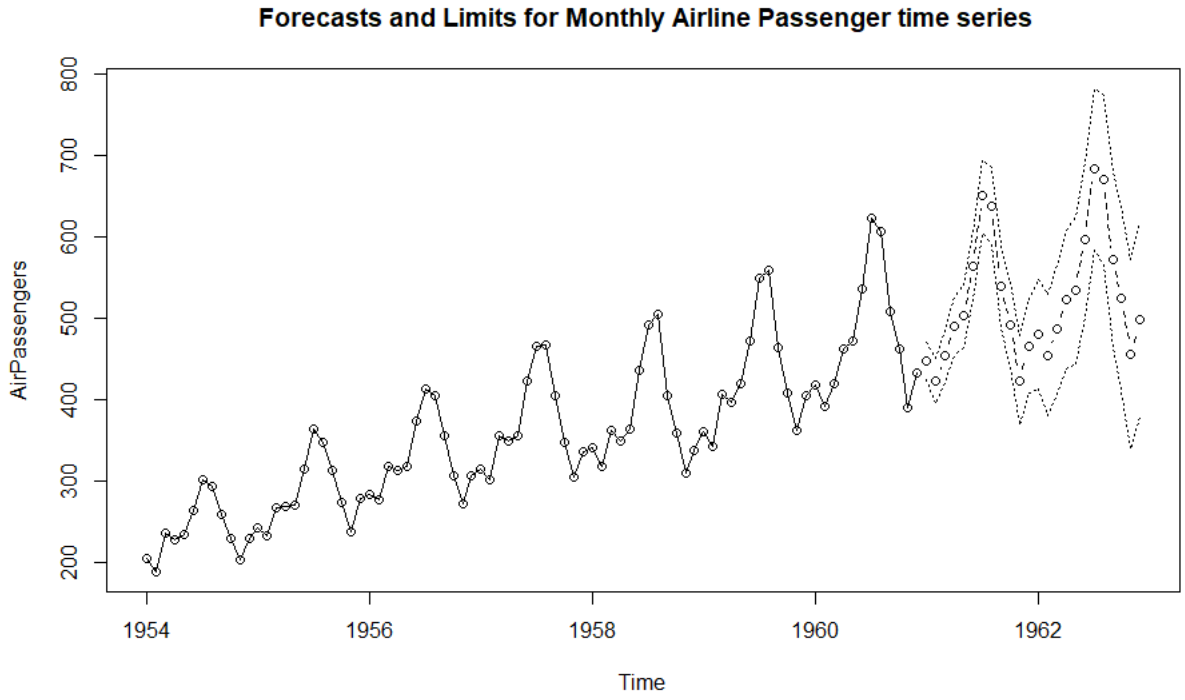

图 14: 客运量趋势预测及其极限

图 14显示了月平均航班人数最近七年和后续两年的预测及 95% 的预测极限。因为模型拟合相对不错，预测极限较为接近拟合趋势的预测。

6 总结

我们对 1949 年 1 月到 1960 年 12 月月航线客运量这一数据进行了分析和研究，并尝试利用时间序列分析的知识来解释数据并建立模型。通过对原数据取对数、差分等操作，发现乘法季节模型较为合适。确定具体模型后我们进行了模型的诊断和过拟合等一系列操作，最终确定 $ARIMA(0,1,1)\times(0,1,1)_{12}$ 模型较为合适。接着我们通过建立的模型进行预测分析，结果相对令人满意。由于预测的带有较高的不确定性，我们发现随着预测时间边长，预测区间就越宽，预测的精确性也越低。

7 致谢

首先我们想感谢蒋老师。老师渊博的专业知识，严肃的科学态度，严谨的治学精神对我们有深远的影响。在蒋老师的指导下，我们习得了时间序列里一系列重要的概念，比如随机过程的平稳性，自回归-求和-滑动平均模型，参数估计与预测，模型诊断，异方差性及建模，更重要的是我们能够能够对一些时间序列数据进行建模，分析和预测。这无疑令人十分激动。

其次我们想感谢 TA 李卓航和康国钰。他们总能热心及时解答学习中遇到的一些问题，这对我们帮助很大。

最后希望这个项目不会是时间序列分析的终点，也希望前面这句话不只是希望。

8 附录

```
#Prepare the AirPassengers data and the window for graph
library(TSA)
data(AirPassengers)
LogAirPassengers = log(AirPassengers)
month = as.vector(season(AirPassengers))
win.graph(width = 9.7,height = 6)

#plot Time Series
plot(AirPassengers,ylab = "AirPassengers",
main = "Monthly Airline Passenger Numbers from 1949 to 1962")
points(AirPassengers,pch = month)
acf(as.vector(AirPassengers),lag.max = 36,main = "ACF of AirPassengers")

#plot Time Series been Logged
plot(LogAirPassengers,ylab = "Log(AirPassengers)",
main = "Monthly Airline Passenger Numbers 1949-1960 After LOG Transformation")
points(LogAirPassengers,pch = month)
acf(as.vector(LogAirPassengers),lag.max = 36,main = "ACF of LOG AirPassengers")

#deal with trend
DLAirPassengers = diff(LogAirPassengers)
plot(DLAirPassengers,ylab = "Diff(Log(AirPassengers))",
main = "First Difference of LOG AirPassengers")
points(DLAirPassengers,pch = month)
acf(as.vector(DLAirPassengers),lag.max = 36,
main = "ACF of the First Difference of LOG AirPassengers")

#deal with atrong seaonality
SDLAirPassengers = diff(DLAirPassengers,lag = 12)
plot(SDLAirPassengers,ylab = "Diff(Diff(Log(AirPassengers)),lag=12)",
main = "First and Seasonal Difference of LOG AirPassengers")
points(SDLAirPassengers,pch = month)
acf(as.vector(SDLAirPassengers),lag.max = 36,
main = "ACF of First and Seasonal Difference of LOG AirPassengers")
m1.LogAirPassengers = arima(LogAirPassengers,order = c(0,1,1),
seasonal = list(order = c(0,1,1),period = 12))
```

```
m2.LogAirPassengers = arima(LogAirPassengers,order = c(0,1,1),
seasonal = list(order = c(0,1,1),period = 12),method = 'ML')
plot(window(rstandard(m1.LogAirPassengers)),main = "The residuals")
points(window(rstandard(m1.LogAirPassengers)),pch = month)

#model diagnose
ap1=log(AirPassengers)
ap=arima(ap1,order=c(0,1,1),seasonal=list(order=c(0,1,1),period=12))
Airpassengers=rstandard(ap)

#plot residual of the model
plot(Airpassengers,xlab = 'Time',ylab='Standardized Residuals', type='o')

#plot ACF
acf(Airpassengers)

#plot  the histogram
plot(hist(Airpassengers),xlab='standard residuals')

#plot qq-plot
plot(qqnorm(Airpassengers),xlab = 'standard residuals')
qqline(Airpassengers)

#model diagnose
tsdiag(ap)

#get specific acf value of residuals
acf(residuals(ap),plot = F)$acf

#Ljung-Box
signif(acf(residuals(ap),plot = F)$acf[1:6],2)

#Analyze the over fitting model ARIMA(0, 1, 2) * (0, 1, 1)_12
ap2=arima(ap1,order=c(0,1,2),seasonal=list(order=c(0,1,1),period=12))
ap2

#Analyze the over fitting model ARIMA(0, 1, 1) * (0, 1, 2)_12
ap3=arima(ap1,order=c(0,1,1),seasonal = list(order=c(0,1,2),period=12))
```

ap3

```
#model forcasting
model = arima(AirPassengers, order = c(0, 1, 1),
seasonal = list(order = c(0, 1, 1), period = 12))
model
plot(model, n1 = c(1954,1), n.ahead = 24, ylab = 'AirPassengers',
main = 'Forecasts and Limits for Monthly Airline Passenger time series')
```