Covid-19 全国疫情预测分析

姓名：于沁涵 学号：1910227

一、背景介绍

2020 年的上半年对于我们来说是不平凡的。2019 年年底，新型冠状病毒首次出现，到二月三月扑面而来的疫情席卷武汉乃至全国。于是全国人民开始了一场疫情防控的自卫反击战，防护医疗人员奔赴一线，全国人民放弃外出游玩春节聚餐的时间，居家隔离。“江城津门同在，日新月异抗疫。”这是我在疫情防控期间为武汉加油写下的两行字，在那段艰难的时期，我们都在贡献自己绵薄的力量，也正是那段时期，让我感受到了团结就是力量的深深贯彻。在人民共同的努力下，疫情很快得到了有效的控制，短短几个月的时间武汉的新增病例实现了清零的胜利。

但是幸运并没有持续很长的时间，全球疫情的爆发以及各种因素让国内疫情无法实现全方面清零的胜利，境外输入、无症状感染者、病毒的变异.....2022 年 2 月，疫情再次出现了反弹，人们不得不再次减少出行、囤货，做好再次抗击疫情的准备。

不可否认的是，人们的生活在此之后开始逐渐发生了变化。口罩成为人们的必需品，减少聚众、保持距离，人们开始出示健康码行走于各个地方，而这些已经成为生活的常态。虽然我们已经把这些习惯当成了生活中的一部分，但反复爆发的疫情仍然给我们带来了不小的困扰。高考的推迟、大学生的封校、航班的取消.....人们每天的计划都可能被疫情左右，为此，寻找一个更加了解疫情走势的途径仍是当务之急。

疫情实时图是当前人们了解疫情的直接途径，通过疫情地图，人们可以了解到每天全国及各个地方新增病例、累计确诊病例以及死亡和治愈的人数等，但是对于预测方面，却没有给出具体显示。

设想如果能科学客观地预测新冠病毒的新增确诊数量变化，对人们实施必要的防控措施、准确评估新冠疫情的影响具有重要的现实意义，同时也可以让人们每日对下一日的计划更加合理。

为了解决此问题，我们考虑利用已有的疫情相关数据为 Covid-19 全国疫情进行预测，探究疫情的影响。利用箱线图、时间序列图、线性回归模型对数据进行分析，并得到结论。

二、数据来源和说明

本文数据选取于国家卫健委官网，其中包含新增确诊人数、累计确诊人数、累计治愈人数、正在医学观察人数，新增接受医学观察人数、新增密切接触人数6个主要指标以及湖北省对应的以上6个指标和日期，一个13个变量，234条数据。本文选用其中的2020年1月25日到2020年2月28日的35条数据以及前7个变量进行分析，具体变量说明见表1。

表1 原始数据变量说明

变量名称	详细说明	取值范围	备注
全国新增确诊人数	连续变量	327～3887	t日新增确诊人数
日期	连续变量	2020年1月25日～2020年2月28日	
全国累计确诊人数	连续变量	1975～79251	t日累计确诊人数
全国累计死亡人数	连续变量	56～2835	t日累计死亡人数
全国累计治愈人数	连续变量	49～39002	t日累计治愈人数
全国正在接受医学观察人数	连续变量	21556～189660	t日正在接受医学观察人数
全国新增接受医学观察人数	连续变量	2533～31470	t日新增接受医学观察人数

三、描述性分析

首先我们对原始数据进行描述性分析。将数据读入后，绘制 6 个主要指标的时间序列变化图。利用 R 语言进行编程，得到结果图如下：

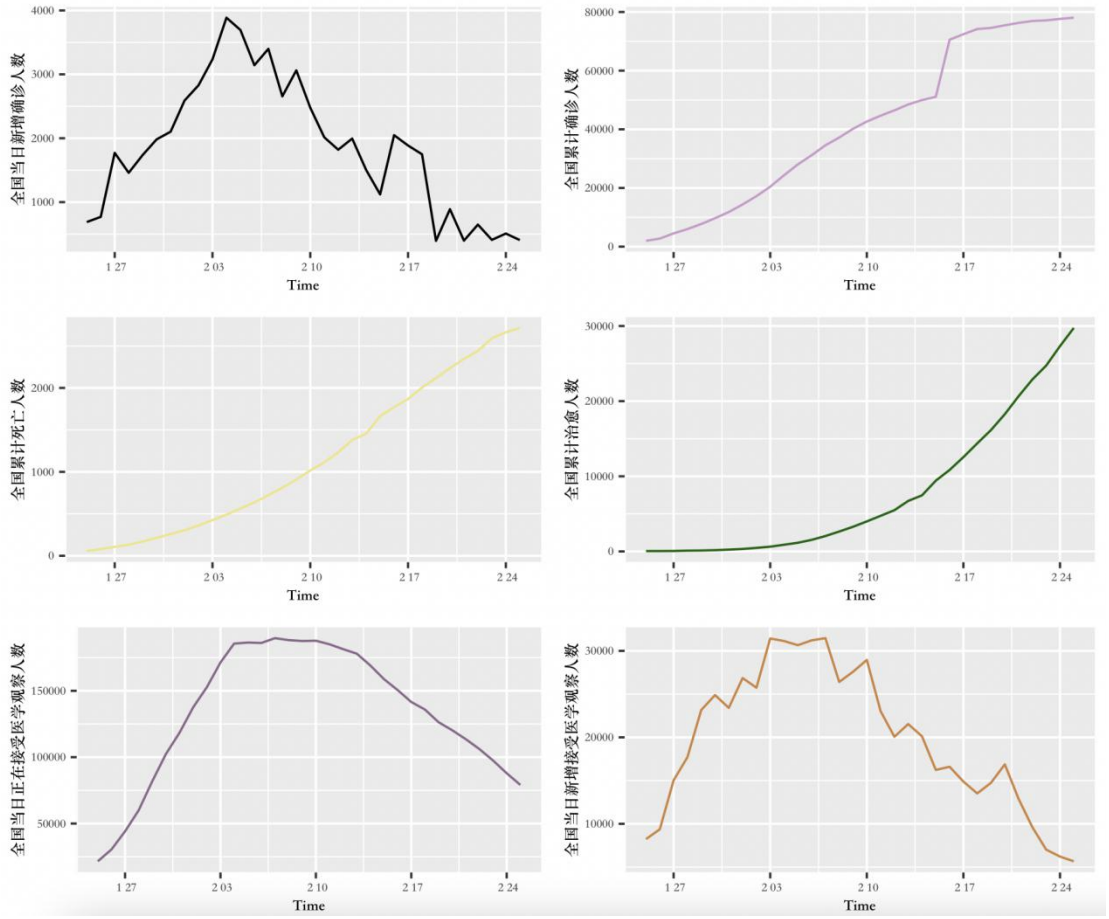

图 1 原始数据的时间序列变化图

由图 1 可以得到随着疫情的发展和防控措施的跟进，新增确诊人数出现了急速上升后又快速下降的趋势，累积确诊病例的走势在逐渐平稳，累积治愈人数在迅速上升，累积死亡人则上升的比较缓慢。正在医学观察的人数在 2 月 5 日左右出现了波峰，而后逐渐下降；当日新增接受医学观察的人数和新增确诊人数走势相同，出现了急速上升后有迅速下降，且对新增确诊人数具有一定的滞后性。这说明在我国政府强有力的防控措施下，在全国人民众志成城全力抗疫的努力下，我们国家的疫情发展趋势得到了有效遏制并且在逐步好转。

四、新指标的构建与说明

我们的目标是基于 t 日的数据预测 t+1 日的疫情发展，因此，我们需要选择一个合适的指标作为因变量来反应疫情发展的情况。

每日新增确诊人数是人们最关注的一个数字，其直接反映了疫情的发展态势，而其增长率指标可以更为直观的反映新增确诊的走势，对日后疫情发展预测有重要意义，因此，我们考虑用 t+1 日新增确诊人数的增长率视作因变量，据此预测疫情的发展。而 t 日的新增确诊人数、累计确诊人数、累计治愈人数、正在医学观察人数，新增接受医学观察人数、新增密切接触人数 6 个指标的增长率作为新指标充当自变量的角色。其中，具体变量说明如表 2，增长率的计算公式如下：

t+1 日新增确诊增长率=(t+1 日新增确诊人数-t 日新增确诊人数)/ t 日新增确诊人数

表 2 建模数据变量说明

	变量名称	详细说明	取值范围
因变量	全国 t+1 日新增确诊人数	连续变量	-0.77473~1.30299
自变量	全国 t 日新增确诊人数	连续变量	-0.77473~1.30299
	全国 t 日累计确诊人数	连续变量	0.002782~0.645408
	全国 t 日累计死亡人数	连续变量	0.01068~0.42857
	全国 t 日累计治愈人数	连续变量	0.04082~0.71667
	全国 t 日正在接受医学观察人数	连续变量	-0.10720~0.54335
	全国 t 日新增接受医学观察人数	连续变量	-0.347165~0.604825

其中 t+1 日新增确诊人数增长率反映了 t+1 日较 t 日新增确诊人数的增长情况，而剩下的 6 个指标均是 t 日相较于 t-1 日时的增长率。

五、新指标的描述性分析

（一）新指标的变化趋势

为了进一步建模，我们对新指标形成的数据进行描述性分析。继续利用时间序列变化图进行变化趋势的描述，如图 2。

由图 2 可看出各个增长率指标的时序图都呈现下降趋势，其中新增确诊增长率波动较大，在 1 月 23 日至 1 月 27 日期间变化剧烈，至 2 月 16 日走势平稳，而后至 2 月 28 日变化加剧。累计确诊增长率在 2 月 16 日数据激增，此后缓步下降。累计死亡增长率、累计治愈增长率、正在接受医学观察人数增长率的走势呈

现震荡下行趋势，整体数据都可以看出疫情在这期间发生了明显的好转。

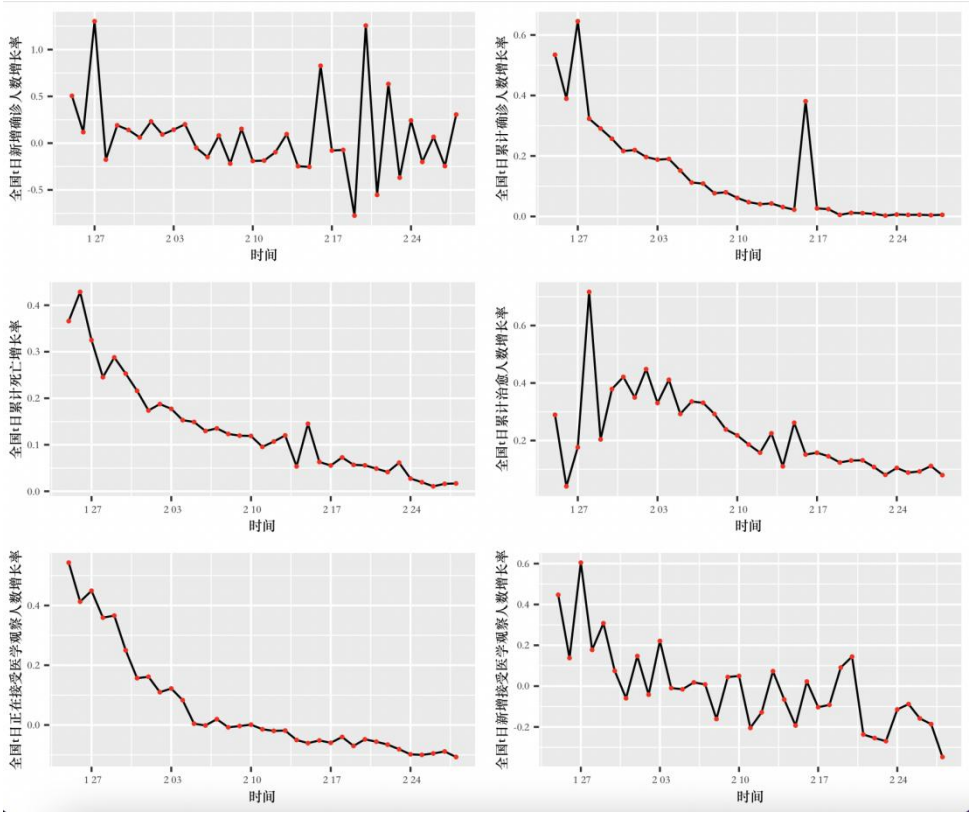

图 2 新指标的时间序列变化图

（二）新指标的与因变量之间的关系

接下来我们对新指标和因变量之间的关系进行研究。以因变量新增确诊增长率的中位数为阈值，将因变量分成高、低两组，绘制分组箱线图，如图 3。

由图 3 看出 t+1 日新增确诊增长率较高时，其他增长率指标也较高，其中只有 t 日新增观察医学观察人数增长率相差不大。累计确诊增长率在因变量的高低两组之间分布差异最明显，这说明累积确诊增长率和新增确诊增长率保持了较为一致的变化趋势。而累计死亡人数增长率、累计治愈人数增长率、正在接受医学观察人数增长率也都是高增长率组取值较高。正在医学观察的人群是未来确诊病例的主要来源，因此可能会导致更多的新增确诊，从而导致疫情更迅速地发展。而累计死亡增长率反映了疫情的严峻性，取值较高代表了新冠疫情处于高发期，因此新增确诊人数也可能走高。

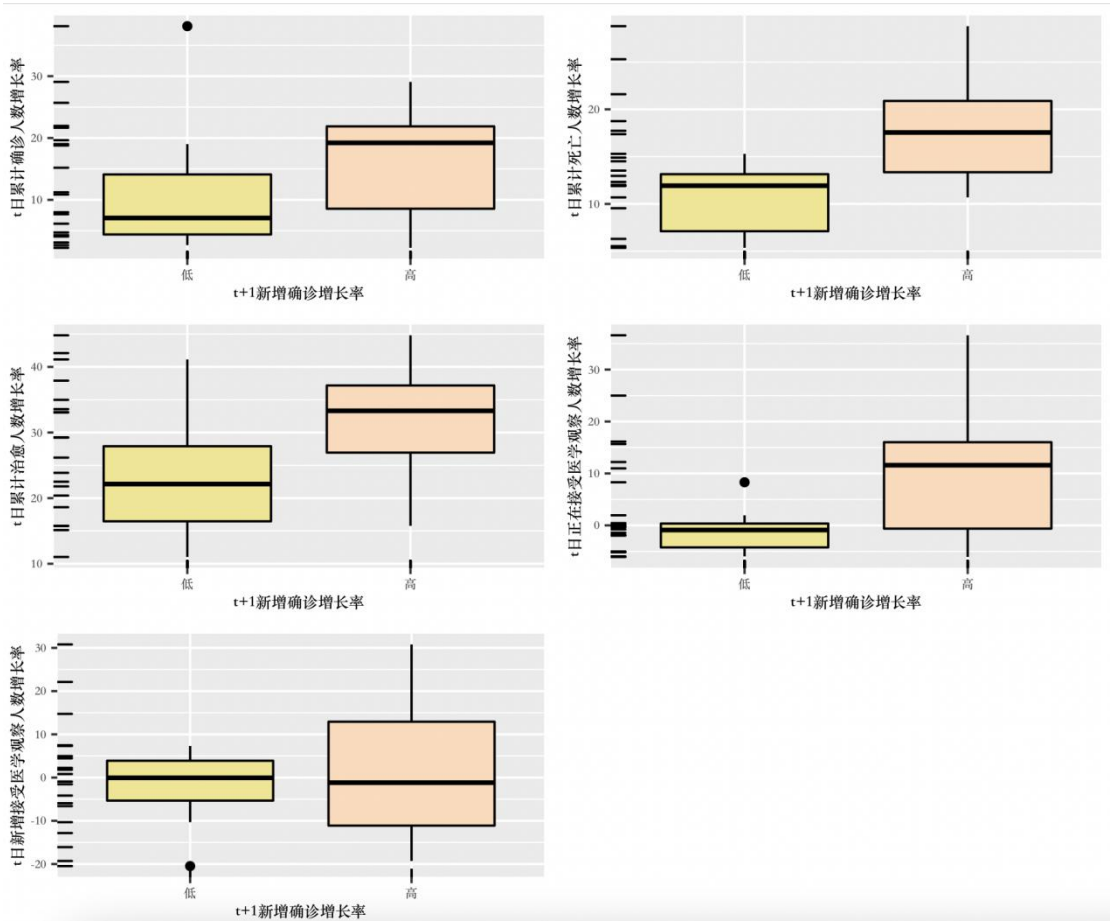

图 3 新指标在因变量高低两组上的分布情况

六、模型的建立

（一）ADF 单位根检验

ADF 检验主要特征是假设数据生成过程服从有单位根的 p 阶自回归过程。一般用于检验的模型为:

$$Y_t = \rho Y_{t-1} + \xi_1 \Delta Y_{t-1} + \xi_2 \Delta Y_{t-2} + \ldots + \xi_{p-1} \Delta Y_{t-p+1} + \varepsilon_t$$

或

$$Y_t = \alpha + \rho Y_{t-1} + \xi_1 \Delta Y_{t-1} + \xi_2 \Delta Y_{t-2} + \ldots + \xi_{p-1} \Delta Y_{t-p+1} + \varepsilon_t$$

在以上两式中都假设 $\varepsilon_t \sim \text{iid}(0, \sigma^2)$。

进行检验的原假设为序列中含有一个单位根,即 $H_0 : \rho = 1$；备择假设是序列是平稳的,即 $H_1 : |\rho| < 1$。用OLS对其进行估计,利用ρ的最小二乘估计 $\hat{\rho}$ 构造如下统计量，并根据Phillips(1987)的结论得到其极限

$$\frac{T(\hat{\rho}-1)}{1-\hat{\xi}_1-\hat{\xi}_2-\ldots-\hat{\xi}_{p-1}} \Rightarrow \frac{1/2\{W^2(1)-1\}-W(1)\int_0^1 W(r)dr}{\int_0^1 W^2(r)dr-(\int_0^1 W(r)dr)^2}$$

$$t \Rightarrow \frac{1/2\{W^2(1)-1\}-W(1)\int_0^1 W(r)dr}{\{\int_0^1 W^2(r)dr-(\int_0^1 W(r)dr)^2\}^{1/2}}$$

对原假设 $H_0:\rho=1,a=0$ 构造F统计量

$$F=\frac{(\tilde{R}^2-\hat{R}^2)/2}{\hat{R}^2/(T-P-1)}$$

其中，$\tilde{R}^2$ 和 $\hat{R}^2$ 分别为有约束的和无约束的残差平方和

$$\tilde{R}^2=\sum_{t=1}^{T}(Y_t-\hat{\xi}_1\Delta Y_{t-1}-\ldots-\hat{\xi}_{p-1}\Delta Y_{t-p+1})^2$$

$$\hat{R}^2=\sum_{t=1}^{T}(Y_t-\hat{\xi}_1\Delta Y_{t-1}-\ldots-\hat{\xi}_{p-1}\Delta Y_{t-p+1}-\hat{a}-\hat{\rho}Y_{t-1})^2$$

这里的 $T(\hat{\rho}-1)$、t和F统计量不能再按照传统的临界值来进行检验。必须用Dickey一Fuller分布表。

我们通常在建模前利用 ADF 单位根检验时间序列的平稳性。这是因为对于一个时间序列，线性回归模型要求建立在稳定数据变量的基础上，否则会出现伪回归的情况。如果把非平稳序列直接应用，会破坏古典线性模型的基本假设，得到的统计量是没有意义的，因此，在建模之前我们首先要对每个指标进行平稳性检验，将非平稳序列转化为平稳序列。在这里，我们利用 ADF 单位根检验来进行平稳性分析，根据 AIC 值、SC 值和 HQ 值最小为准则选择截距项，得到结果如表 3：

表 3　ADF单位根检验

变量	ADF检验值	P值	结果
全国 t+1 日新增确诊人数增长率	-2.8419	0.2462	不平稳
全国 t 日新增确诊人数增长率	-3.321	0.08516	平稳*
全国 t 日累计确诊人数增长率	-3.1832	0.1137	不平稳
全国 t 日累计死亡人数增长率	-1.3938	0.8084	不平稳
全国 t 日累计治愈人数增长率	-2.456174	0.1359	不平稳
全国 t 日正在接受医学观察人数增长率	-3.1403	0.1303	不平稳
全国 t 日新增接受医学观察人数增长率	-3.2752	0.09178	平稳*
D(全国 t+1 日新增确诊人数增长率)	-5.4512	<0.01	平稳***

D(全国 t 日累计确诊人数增长率)	-3.9871	0.02184	平稳**
D(全国 t 日累计死亡人数增长率)	-4.845	<0.01	平稳***
D(全国 t 日正在接受医学观察人数增长率)	-3.0025	0.1846	不平稳
D(全国 t 日正在接受医学观察人数增长率，2)	-7.2936	<0.01	平稳***

注：***表示在1%的显著性水平下拒绝原假设，**表示在5%的显著性水平下拒绝原假设，*表示在10%的显著性水平下拒绝原假设，D(X)表示变量的一阶差分,D(X,2)表示变量的二阶差分

由表 3可以得到结论：在数据中，全国t日新增确诊人数、全国t日新增接受医学观察人数都是10%的显著性水平下平稳，全国t+1日新增确诊人数、全国t日累计确诊人数、全国t日累计死亡人数在一阶差分下平稳，全国t日正在接受医学观察人数在二阶差分下平稳。

（二）线性回归模型

将时间序列变成平稳序列后，我们可以进一步建立线性回归模型。通过AIC准则进行变量选择，得到相关参数估计以及检验结果如表 4。模型调整后R^2 为0.8045，说明拟合程度较好。进一步对AIC选择的模型进行了多重共线性检验，发现保留下来的各个变量的方差膨胀因子均小于10，说明当前模型中不存在多重共线性问题。

表 4 线性回归结果（因变量：全国 t+1 日新增确诊人数）

变量名称	标准化系数估计	标准差	统计量	P 值
截距项	-0.033508817	0.09737	-0.344	0.733
全国 t 日新增确诊人数增长率	-1.616691149	0.1526	-10.589	1.78e-11 ***
全国 t 日累计确诊人数增长率	0.781875668	0.55664	1.405	0.171
全国 t 日累计死亡人数增长率	1.02041	0.70874	1.440	0.161

通过 AIC 模型，最终保留了三个变量，分别是： 全国 t 日新增人数增长率、全国 t 日累计确诊人数增长率和 t 日累计死亡人数增长率。其中 t 日新增人数增长率对因变量有显著的负向影响，这意味着 t 日新增人数增长率越高，下一日的新增人数增长率可能反而会降低，这说明新增病例在此时的疫情防控期间每天的增长速度越来越慢，符合实情，也印证了疫情防控的有效性。

而t日累计确诊人数增长率和t日累计死亡人数增长率对因变量有正向影响，这是因为累计确诊增长率代表着过去一段时间内确诊人数累计值的增长情况，累计确诊增长率越高，形势越严峻，因此下一日新增确诊增长率也可能走高。

（三）t+1 日的预测与结果解读

最后，我们通过建立的 AIC 模型基于最后一天 2 月 28 日的数据对 t+1 日的新增病例增长率进行预测，也就是 2 月 29 日的新增病例增长率，得到预测的新增病例增长率为 0.3775457，预测的新增病例人数为，置信区间为[-0.2957659 1.050857]，预测的新增病例人数为 588 人。而真实的新增病例人数为 573，位于置信区间以内，这说明 AIC 模型可以对未来一天的新增确诊人数进行较为精准的预测。

七、结论与展望

本文基于国家卫健委发布的疫情数据，试图对未来一日的新增确诊人数增长率进行预测。基于增长率指标，我们建立线性回归模型，最终将新增确诊人数增长率、累计确诊人数增长率以及累计死亡人数增长率作为主要因素建立模型，得到了比较好的结果。模型对人们了解疫情走势有着重要的作用，可以帮助人们进行每日生活计划的制定，帮助国家进行疫情防控措施的制定，为新冠疫情预测局面提供良好的结果与解决方案。

此外，模型还可以在本文的基础上增加更多因素。例如，加入密切接触者、无症状感染者以及病毒的类型因素等，也可以增加预测的准确性，更好的进行建模。

附录 1　R 语言程序代码

```
##导入包
library(zoo)
library(tseries)
library(ggplot2)
library(scales) #修改刻度
library(gridExtra)
library(car)

data <- read.csv('/Users/yuqinhan1229/Desktop/time.csv')
##构造时间序列
data0 <- data[,-1]
ts <- ts(data0)
time <- as.Date('2020-01-20')+c(0:233)
tail(time)
dt.ts <- zoo(ts,time)
dt <- dt.ts[6:40,1:6]
###画时间序列图
# 对 zoo 类型的数据，用 fortify()转换成 data.frame 类型
##全国当日新增确诊人数
options(repr.plot.width=4, repr.plot.height=2)
g1<-ggplot(aes(x = Index, y = Value), data = fortify(dt$new_confirmed, melt =
TRUE))+geom_line() +xlab("Time") + ylab("全国当日新增确诊人数") +
    theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国累计确诊人数
options(repr.plot.width=4, repr.plot.height=2)
g2<-ggplot(aes(x = Index, y = Value), data = fortify(dt$confirmed_count, melt =
TRUE))+geom_line(color = '#CC99CC') +xlab("Time") + ylab("全国累计确诊人数")
+
  theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国累计死亡人数
options(repr.plot.width=4, repr.plot.height=2)
g3<-ggplot(aes(x = Index, y = Value), data = fortify(dt$dead_count, melt =
TRUE))+geom_line(color = 'khaki2') +xlab("Time") + ylab("全国累计死亡人数") +
  theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国累计治愈人数
```

```
options(repr.plot.width=4, repr.plot.height=2)
g4<-ggplot(aes(x = Index, y = Value), data = fortify(dt$cure_count, melt =
TRUE))+geom_line(color = 'darkgreen') +xlab("Time") + ylab("全国累计治愈人数")
+
  theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国当日正在接受医学观察人数
options(repr.plot.width=4, repr.plot.height=2)
g5<-ggplot(aes(x = Index, y = Value), data = fortify(dt$observe_count, melt =
TRUE))+geom_line(color = 'plum4') +xlab("Time") + ylab("全国当日正在接受医学
观察人数") +
  theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国当日新增接受医学观察人数
options(repr.plot.width=4, repr.plot.height=2)
g6<-ggplot(aes(x = Index, y = Value), data = fortify(dt$new_contact, melt =
TRUE))+geom_line(color = 'tan3') +xlab("Time") + ylab("全国当日新增接受医学观
察人数") +
  theme(text=element_text(family="Songti SC",size=7,face = "bold"))
grid.arrange(g1,g2,g3,g4,g5,g6,nrow=3,ncol=2)


##构造新的时间序列和指标
data <- data[5:41,]
new <- data[,2:7]
for (i in 1:37){
  for (j in 2:7){
  if (i==1){
    new[i,j-1]= 0
  }
  else{
    new[i,j-1] = (data[i,j]-data[i-1,j])/data[i-1,j]
  }
}
}
new$rate = c(new$new_confirmed[2:37],0)
new <- new[-c(1,37),]
ts <- ts(new)
time <- as.Date('2020-01-25')+c(0:34)
```

```
new$time=time
tail(time)
dt <- zoo(ts,time)
summary(dt)
# 对 zoo 类型的数据，用 fortify()转换成 data.frame 类型
##全国当日新增确诊人数
options(repr.plot.width=4, repr.plot.height=2)
g1 <- ggplot(new,aes(x = time, y =new_confirmed ))+geom_line() +xlab("时间") +
ylab("全国 t 日新增确诊人数增长率") +
   geom_point(colour = "red",size = 0.5,show_guide = FALSE)+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国累计确诊人数
options(repr.plot.width=4, repr.plot.height=2)
g2 <- ggplot(new,aes(x = time, y =confirmed_count ))+geom_line() +xlab("时间") +
ylab("全国 t 日累计确诊人数增长率") +
   geom_point(colour = "red",size = 0.5,show_guide = FALSE)+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国累计死亡人数
options(repr.plot.width=4, repr.plot.height=2)
g3 <- ggplot(new,aes(x = time, y =dead_count))+geom_line() +xlab("时间") + ylab("
全国 t 日累计死亡增长率") +
   geom_point(colour = "red",size = 0.5,show_guide = FALSE)+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国累计治愈人数
options(repr.plot.width=4, repr.plot.height=2)
g4 <- ggplot(new,aes(x = time, y =cure_count))+geom_line() +xlab("时间") + ylab("
全国 t 日累计治愈人数增长率") +
   geom_point(colour = "red",size = 0.5,show_guide = FALSE)+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国当日正在接受医学观察人数
options(repr.plot.width=4, repr.plot.height=2)
g5<-ggplot(new,aes(x = time, y =observe_count))+geom_line() +xlab("时间") +
ylab("全国 t 日正在接受医学观察人数增长率") +
   geom_point(colour = "red",size = 0.5,show_guide = FALSE)+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国当日新增接受医学观察人数
options(repr.plot.width=4, repr.plot.height=2)
```

```
g6<-ggplot(new,aes(x = time, y =new_contact))+geom_line() +xlab("时间") + ylab("
全国 t 日新增接受医学观察人数增长率") +
   geom_point(colour = "red",size = 0.5,show_guide = FALSE)+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
grid.arrange(g1,g2,g3,g4,g5,g6,nrow=3,ncol=2)

##选取中位数作为阈值进行划分
new[,1:7] <- new[,1:7]*100
new0 <- new[1:20,]
m <- median(new0$rate,na.rm = FALSE)
attach(new0)
new0$rate[new0$rate>=m]="高"
new0$rate[new0$rate<m]="低"
detach(new0)


##绘制箱线图
##全国累计确诊人数
g1 <- ggplot(data = new0,aes(x = rate, y = confirmed_count))+xlab("t+1 新增确诊增
长率") + ylab("t 日累计确诊人数增长率") +
   geom_boxplot(fill=c('khaki','peachpuff'),color="black")+
   geom_rug(color="black")+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国累计死亡人数
g2 <- ggplot(data = new0,aes(x = rate, y = dead_count))+xlab("t+1 新增确诊增长率")
+ ylab("t 日累计死亡人数增长率") +
   geom_boxplot(fill=c('khaki','peachpuff'),color="black")+
   geom_rug(color="black")+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国累计治愈人数
g3 <- ggplot(data = new0,aes(x = rate, y = cure_count))+xlab("t+1 新增确诊增长率")
+ ylab("t 日累计治愈人数增长率") +
   geom_boxplot(fill=c('khaki','peachpuff'),color="black")+
   geom_rug(color="black")+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国当日正在接受医学观察人数
g4 <- ggplot(data = new0,aes(x = rate, y = observe_count))+xlab("t+1 新增确诊增长
```

```
率") + ylab("t 日正在接受医学观察人数增长率") +
   geom_boxplot(fill=c('khaki','peachpuff'),color="black")+
   geom_rug(color="black")+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))
##全国当日新增接受医学观察人数
g5 <- ggplot(data = new0,aes(x = rate, y = new_contact))+xlab("t+1 新增确诊增长率
") + ylab("t 日新增接受医学观察人数增长率") +
   geom_boxplot(fill=c('khaki','peachpuff'),color="black")+
   geom_rug(color="black")+
   theme(text=element_text(family="Songti SC",size=7,face = "bold"))

grid.arrange(g1,g2,g3,g4,g5,ncol=2)




##ggplot2 绘制 ACF PACF
bacf <- acf(dt[, 13], plot = FALSE)
bacfdf <- with(bacf, data.frame(lag, acf))
ggplot(data = bacfdf, mapping = aes(x = lag, y = acf)) +
   geom_segment(mapping = aes(xend = lag, yend = 0)) +
   geom_hline(aes(yintercept = 0.05), linetype = 2, color = 'darkblue')
# pacf
bpacf <- pacf(dt[, 13], plot = FALSE)
bpacf <- with(bpacf, data.frame(lag, acf))
ggplot(data = bpacf, mapping = aes(x = lag, y = acf)) +
   geom_segment(mapping = aes(xend = lag, yend = 0)) +
   geom_hline(aes(yintercept = 0.05), linetype = 2, color = 'darkblue')


###ADF 检验
tseries::adf.test(dt[,7])
tseries::adf.test(diff(dt[,7]))
tseries::adf.test(dt[,1])##全国当日新增确诊人数增长率
tseries::adf.test(dt[,2])##全国累计确诊人数增长率
tseries::adf.test(diff(dt[,2]))
tseries::adf.test(dt[,3])##全国累计死亡人数增长率
```

```
tseries::adf.test(dt[,4])##全国累计治愈人数增长率
tseries::adf.test(diff(dt[,4]))
tseries::adf.test(dt[,5])##全国当日正在接受医学观察人数增长率
tseries::adf.test(diff(dt[,5]))
tseries::adf.test(diff(diff((dt[,5]))))
tseries::adf.test(dt[,6])##全国当日新增接受医学观察人数增长率


###线性回归
t1 <-diff(dt$rate)[2:34]
t2<- dt$new_confirmed[3:35]
t3<-diff(dt$confirmed_count)[2:34]
t4 <- dt$dead_count[3:35]
t5<-diff(dt$cure_count)[2:34]
t6 <- diff(diff(dt$observe_count))
t7 <- dt$new_contact[3:35]
fm<-lm(t1~t2+t3+t4+t5+t6+t7)
summary(fm)
AIC(fm)  #AIC
fm.AIC=step(fm,trace=0)
summary(fm.AIC)  #经过 AIC 选择后的变量系数输出表
##多重共线性检验
vif(fm.AIC)
##线性模型的预测
t2 <-t2[32]
t3 <-t3[32]
t4 <-t4[32]
new_p <- data.frame(t2,t3,t4)
lm.pred<-predict(fm.AIC, new_p, interval="prediction",level=0.95)
lm.pred
confint(fm.AIC)
```