参考书《金融数据分析导论——基于 R 语音》，P96——1(a)、(b)和 2

所需数据可以在 http://faculty.chicagobooth.edu/ruey.tsay/teaching/introTS/这个网址找到

（注：单位根的检验可以使用 adfTest 或是 ur.df 函数实现）

除非特别声明，在以下习题中都用 5%的显著性水平来得出结论。

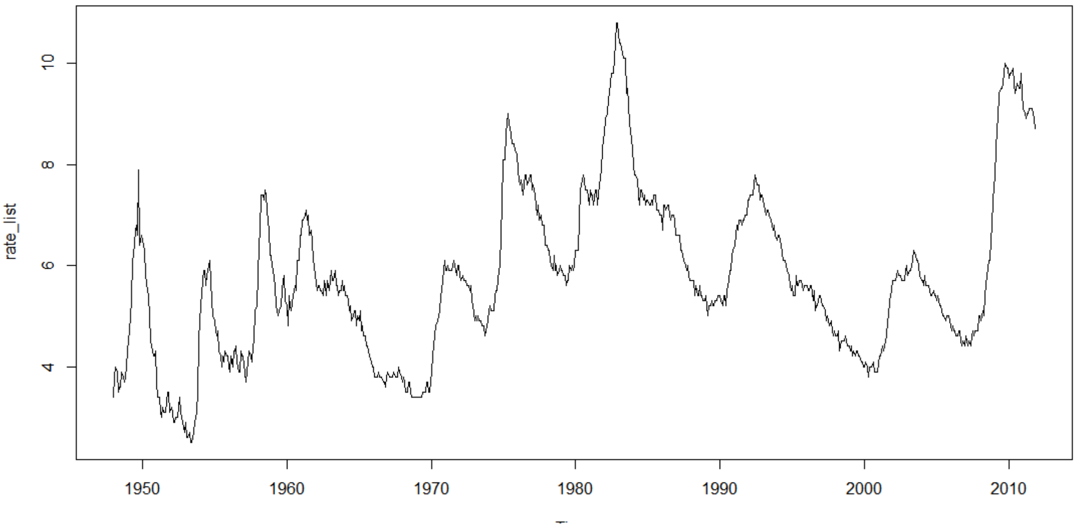
1.考虑从 1948 年 1 月到 2011 年 11 月美国失业率的月数据（见文件 m-unrate-4811.txt），数据来自美国圣路易斯的联邦储备银行。

（a）该失业率的月数据是否存在单位根？为什么？

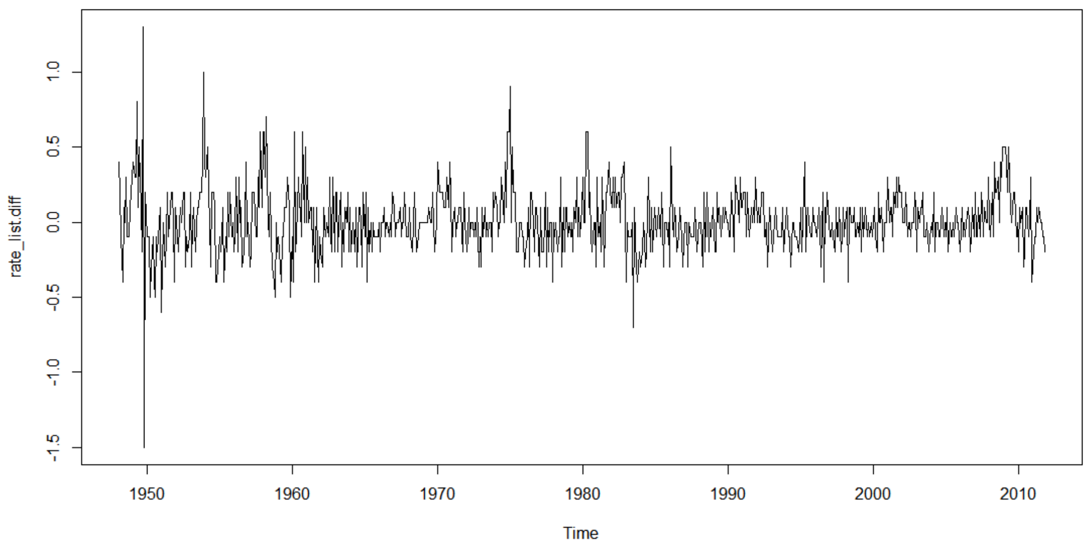
（b）根据该数据建立一个时间序列模型并检验模型是否已充分拟合数据。然后，根据所建立的模型对美国 2011 年 12 月和 2012 年前 3 个月的失业率进行预测。（注意：适合该数据的模型不止一个，只要模型充分即可。）

（a）

```
> #读取数据，观察时序图
> data = read.table("E:/DATA/data mining/fts03/data/m-unrate-4811.txt",header=T)
> dim(data)
[1] 767   1
> rate_list = ts(data$rate,frequency=12,start=c(1948,1))
> plot.ts(rate_list)
>
```

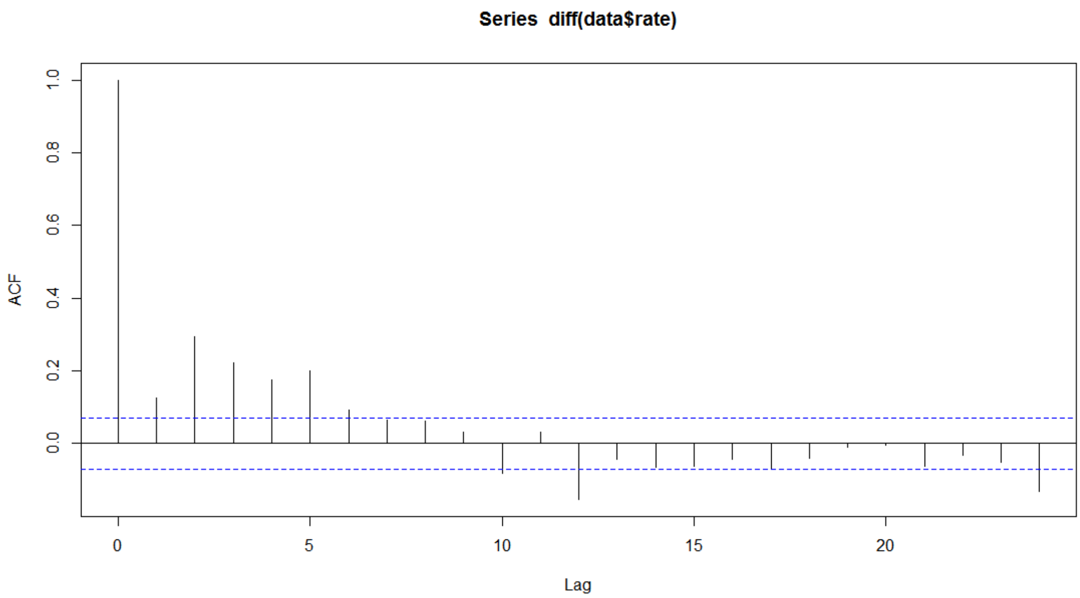



```
> #差分处理，一阶差分观察
> rate_list.diff = diff(rate_list)
> plot(rate_list.diff)
```



通过肉眼观察，图形基本是平滑的。在一个固定的均值水平附近波动。

```
> #通过ACF观察自相关性
> acf(diff(data$rate),lag.max=24)
```



可见 lag 值在 1，2，3，4，5，6，12，23 都有较明显的相关性

选择 lag =5 通过单位根检验其平滑性

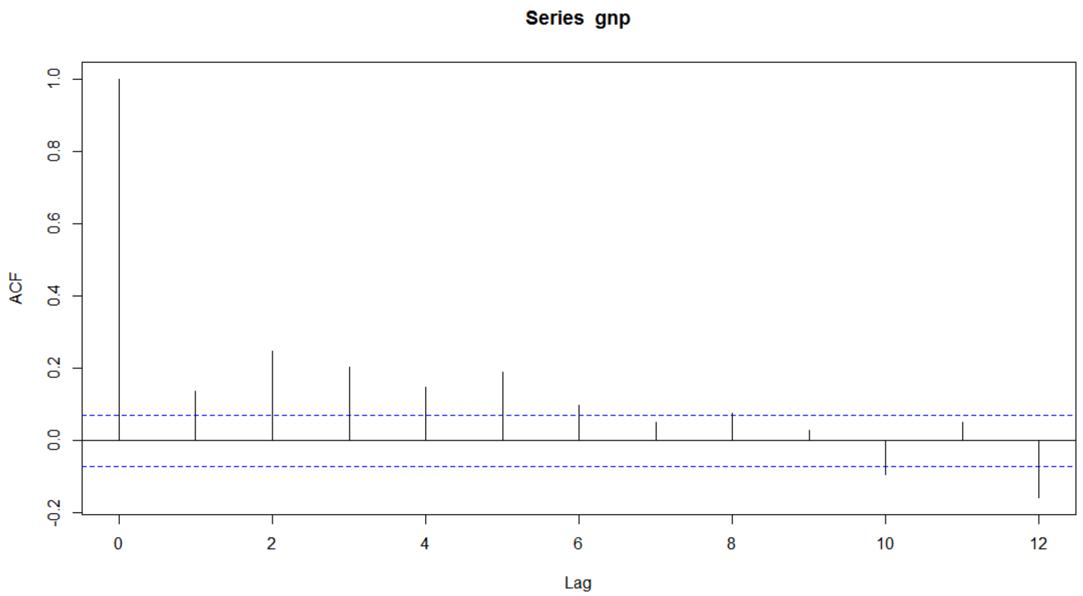
```
> #通过扩展的Dickey-Fuller做单位根检验
> library(fUnitRoots)
> #ar命令对AIC准则值进行调整，使得AIC的值最小
> m1 = ar(diff(data$rate),method='mle')
> m1$order
[1] 12
> adfTest(diff(data$rate),lags=5,type=c("c"))

Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 5
  STATISTIC:
    Dickey-Fuller: -7.5061
  P VALUE:
    0.01

Description:
 Tue Apr 21 12:42:50 2015 by user: Administrator

Warning message:
In adfTest(diff(data$rate), lags = 5, type = c("c")) :
  p-value smaller than printed p-value
```

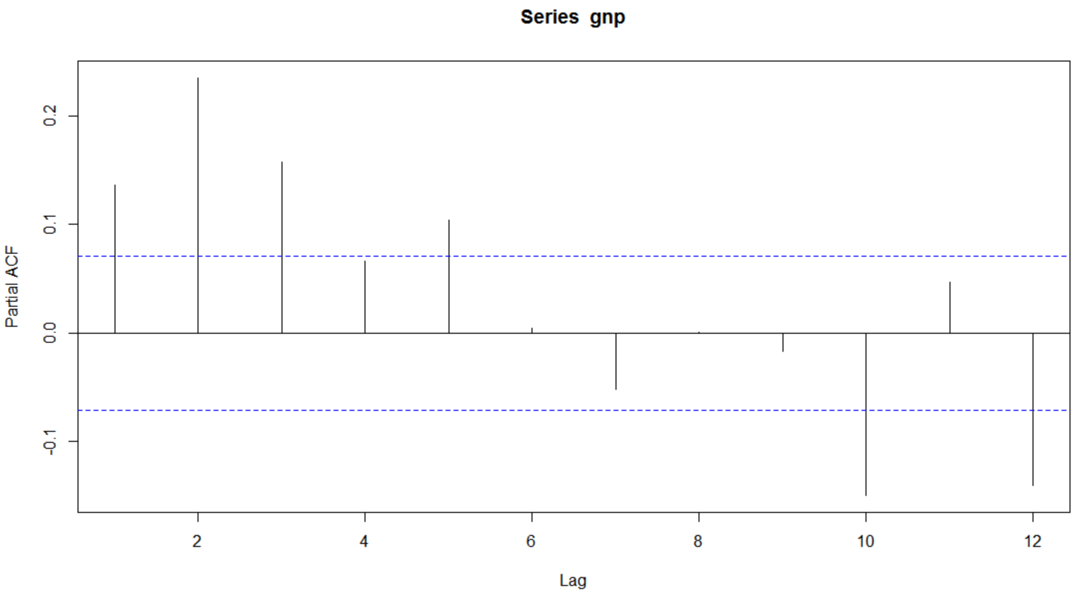
当 lag=5 时，ADF 检验统计量是-7.5061，p 值是 0.01。

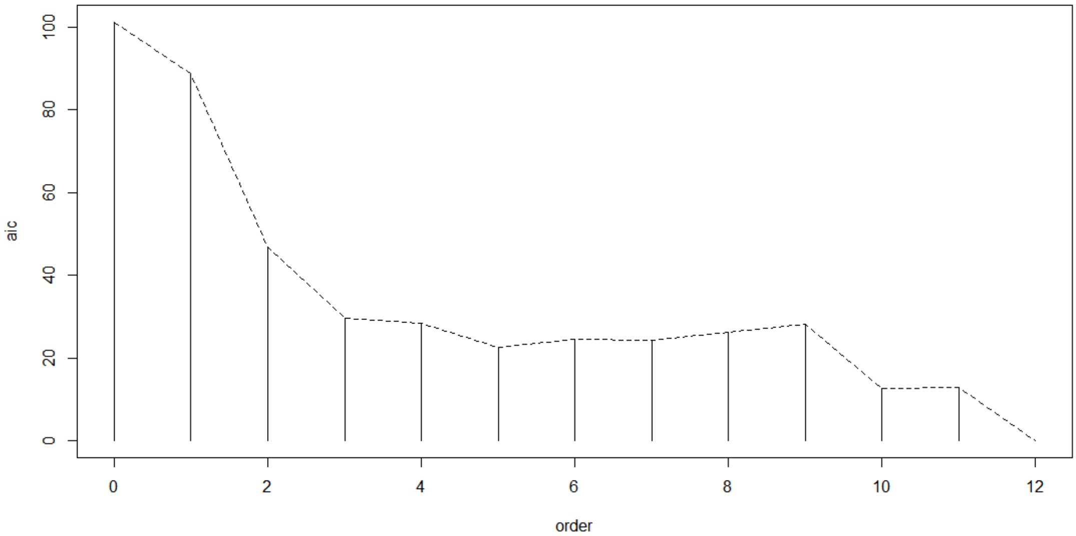
这表明单位根假设不能被拒绝，即接受。存在单方根。

（b）

```
> ########构建AR models#########
> G=data$rate
> LG=log(G)  #求对数
> gnp=diff(LG) #LG数列的后一项减去前一项得到的差数列，即gnp[i]=LG[i+1]-LG[i]
> #定阶
> acf(gnp,lag=12)  # compute ACF
> pacf(gnp,lag=12) # compute PACF
> mm1=ar(gnp,method='mle')
> mm1$order # Find the identified order
[1] 12
> names(mm1)
 [1] "order"        "ar"           "var.pred"     "x.mean"       "aic"          "n.used"
 [7] "order.max"    "partialacf"   "resid"        "method"       "series"       "frequency"
[13] "call"         "asy.var.coef"
> print(mm1$aic,digits=3)
    0     1     2     3     4     5     6     7     8     9    10    11    12
101.3  88.9  46.9  29.6  28.3  22.6  24.6  24.3  26.3  28.1  12.7  12.9   0.0
> aic=mm1$aic  # For plotting below.
> length(aic)
[1] 13
> plot(c(0:12),aic,type='h',xlab='order',ylab='aic')
> lines(0:12,aic,lty=2)
```







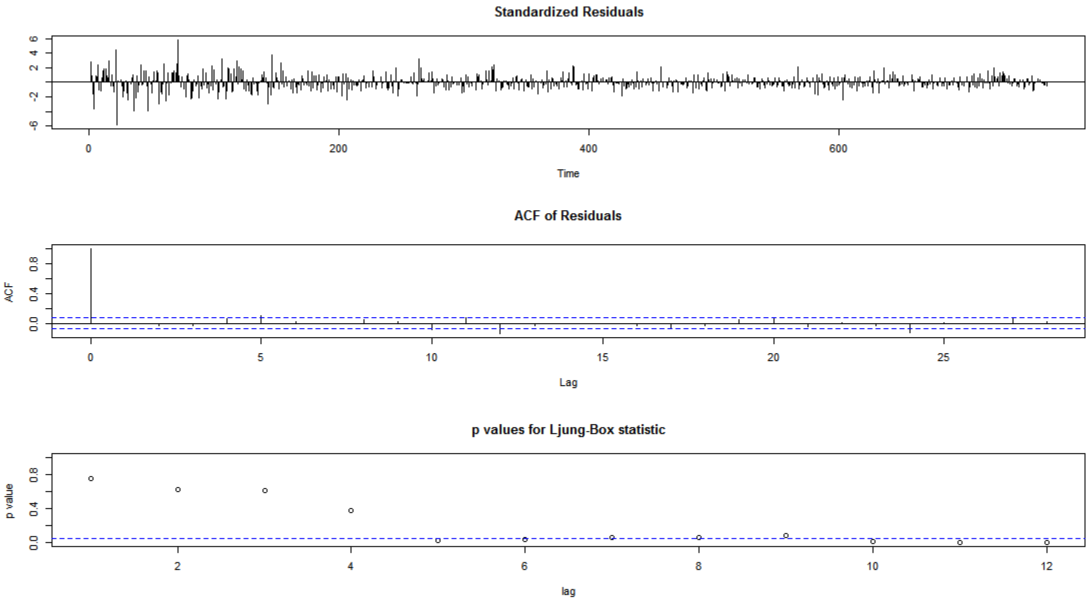
通过图形可以看出，lag=3 也是不错的选择

```
> #构建方程
> m1=arima(gnp,order=c(3,0,0))  #构建AR(3)模型
> m1

Call:
arima(x = gnp, order = c(3, 0, 0))

Coefficients:
         ar1     ar2     ar3  intercept
      0.0674  0.2213  0.1585     0.0013
s.e.  0.0358  0.0350  0.0359     0.0024

sigma^2 estimated as 0.001379:  log likelihood = 1435.49,  aic = -2860.99
>
> #模型检验
> tsdiag(m1,gof=12)  # 检验残差是否白噪声
> Box.test(m1$residuals,lag=12,type='Ljung')

 Box-Ljung test

data:  m1$residuals
X-squared = 41.8975, df = 12, p-value = 3.465e-05
```

Ljung-Box 统计量 Q(m)渐近服从自由度为 df-lag 的卡方分布

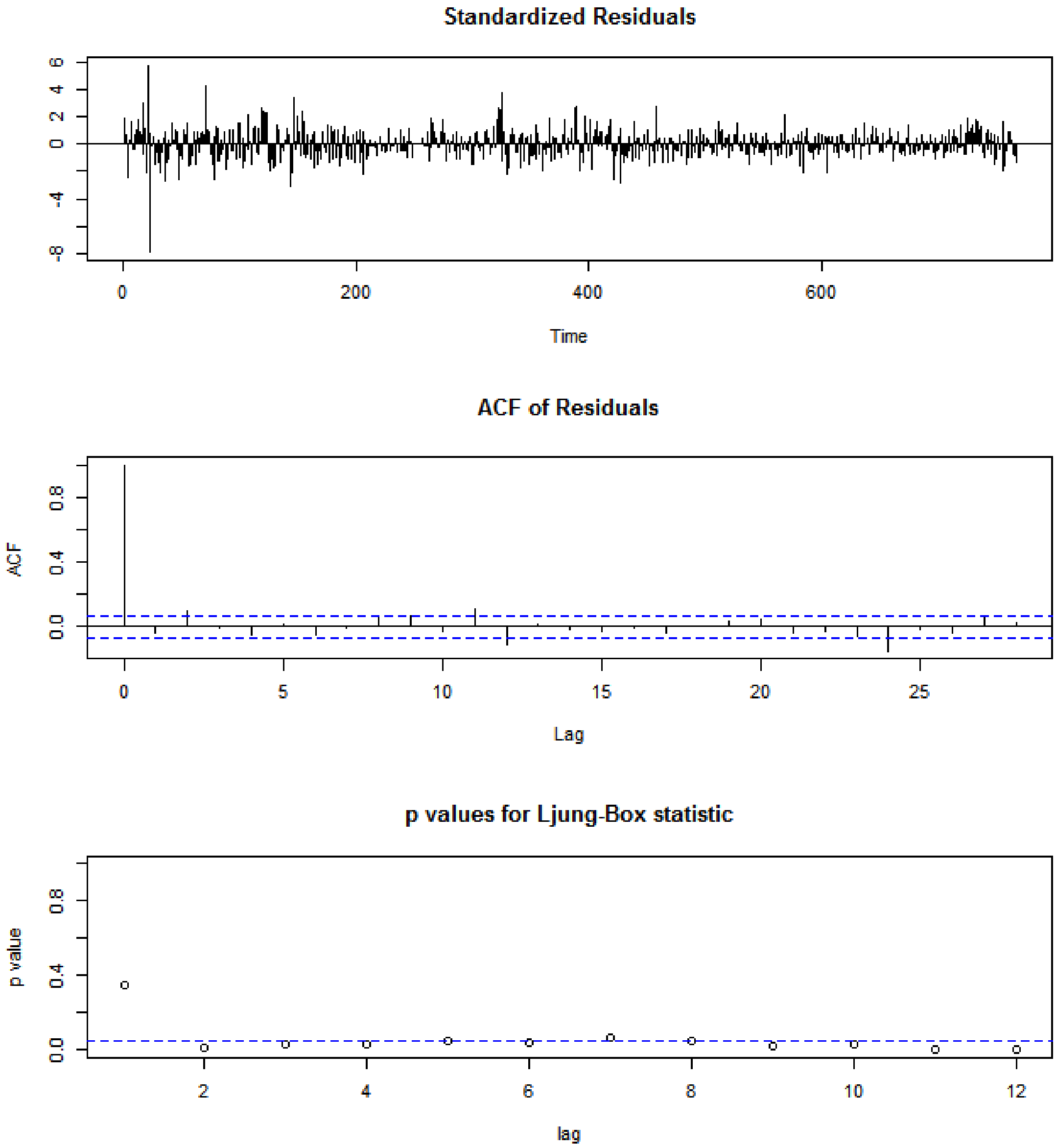
针对 AR(3)模型进行残差序列分析，可算的 Q(12)=41.8975,p 值=3.465e-05

在 5%的显著性条件下，前 12 个间隔无前后相关性的原假设不能拒绝。




```
> pv=1-pchisq(41.8975,9) # Compute p value using 1 degrees of freedom
> pv
[1] 3.431407e-06
```

```
> ###########构建ARMA模型################
> #根据信息准则定阶
> library("forecast")
> auto.arima(G,ic="bic")
Series: G
ARIMA(2,1,2)

Coefficients:
         ar1      ar2      ma1     ma2
      1.6625  -0.7869  -1.6367  0.8554
s.e.  0.0404   0.0459   0.0423  0.0514

sigma^2 estimated as 0.0395:  log likelihood=150.31
AIC=-290.63   AICc=-290.55   BIC=-267.42
> #模型拟合
> m2=arima(G,order=c(2,1,2))
> m2

Call:
arima(x = G, order = c(2, 1, 2))

Coefficients:
         ar1      ar2      ma1     ma2
      1.6625  -0.7869  -1.6367  0.8554
s.e.  0.0404   0.0459   0.0423  0.0514

sigma^2 estimated as 0.0395:  log likelihood = 150.31,  aic = -290.63
> #模型检验
> tsdiag(m2,gof=12)    #检查残差是否白噪声
> Box.test(m2$residuals,lag=12,type='Ljung')

 Box-Ljung test

data:  m2$residuals
X-squared = 38.2232, df = 12, p-value = 0.0001412
```




```
> #预测
> predict(m2,1)
$pred
Time Series:
Start = 768
End = 768
Frequency = 1
[1] 8.728842

$se
Time Series:
Start = 768
End = 768
Frequency = 1
[1] 0.1987515

> M.pre2 = forecast.Arima(m2,h=12)
> M.pre2
    Point Forecast    Lo 80    Hi 80    Lo 95     Hi 95
768       8.728842 8.474132 8.983552 8.339296  9.118387
769       8.716000 8.351114 9.080886 8.157954  9.274046
770       8.671956 8.206088 9.137825 7.959472  9.384441
771       8.608839 8.036989 9.180689 7.734270  9.483408
772       8.538564 7.852457 9.224672 7.489254  9.587875
773       8.471398 7.664310 9.278487 7.237063  9.705733
774       8.415032 7.484080 9.345984 6.991264  9.838801
775       8.374175 7.320919 9.427431 6.763359  9.984991
776       8.350603 7.180551 9.520655 6.561163 10.140042
777       8.343563 7.065090 9.622037 6.388308 10.298819
778       8.350408 6.973482 9.727335 6.244581 10.456236
779       8.367327 6.902323 9.832331 6.126797 10.607857
>
```

2.以 NYSE/AMEX/NASDAQ 的市场资本为基础考虑 CRSP Decile 1、2、5、9、10 投资组合的月简单收益率，该数据的时间区间是从 1961 年 1 月到 2011 年 9 月。

（a）对于 Decile 2 和 Decile 10 的收益序列，在 5%的显著性水平下检验：原假设是滞后阶数为 1~12 的自相关系数均为 0，给出你的结论

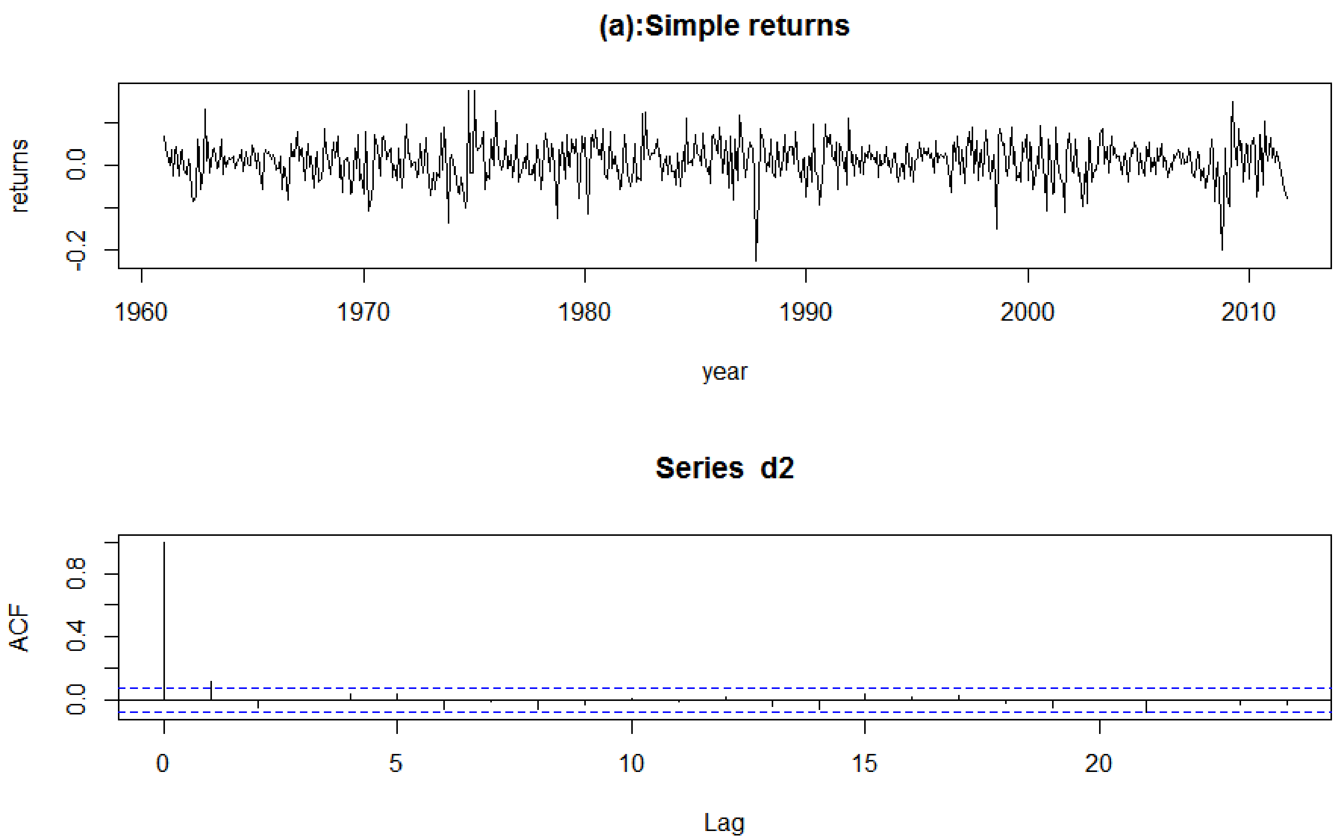
（b）对于 Decile 2 的收益率序列建立一个 ARMA 模型，对模型进行检验并写出拟合的模型

（c）利用拟合的 ARMA 模型对序列进行超前 1~12 步预测，并给出预测的相关标准误差。

（a）已经在上一次的作业中处理

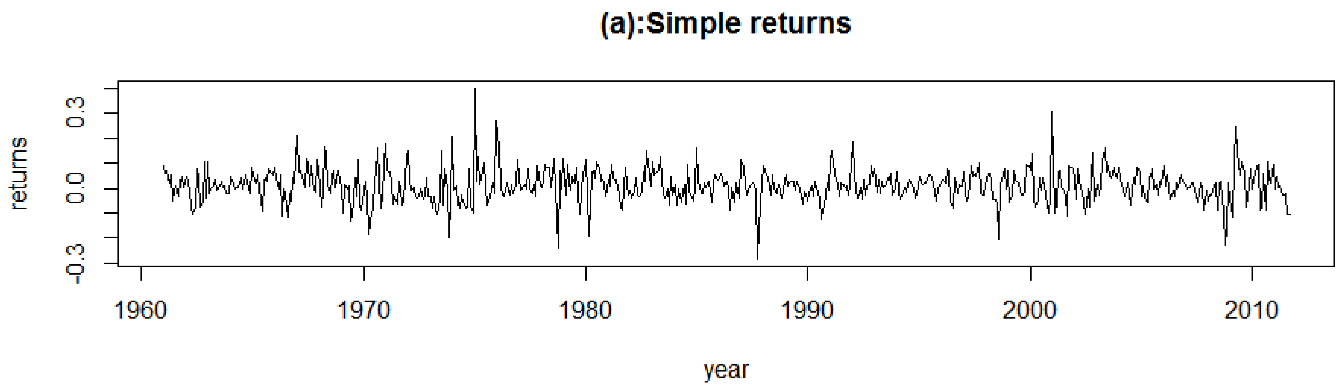
```
> #sample ACF
> da=read.table("E:/DATA/data mining/fts02/m-dec125910-6111.txt",header=T)
> head(da)
      date      dec1      dec2      dec5      dec9
1 19610131  0.058011  0.067392  0.081767  0.096754
2 19610228  0.029241  0.042784  0.055524  0.056564
3 19610330  0.025896  0.025474  0.041304  0.060563
4 19610428  0.005667  0.001365  0.000780  0.011911
5 19610531  0.019208  0.036852  0.049590  0.046248
6 19610630 -0.024670 -0.025225 -0.040046 -0.050651
     dec10
1  0.087207
2  0.060245
3  0.071875
4  0.023328
5  0.050362
6 -0.051434
```

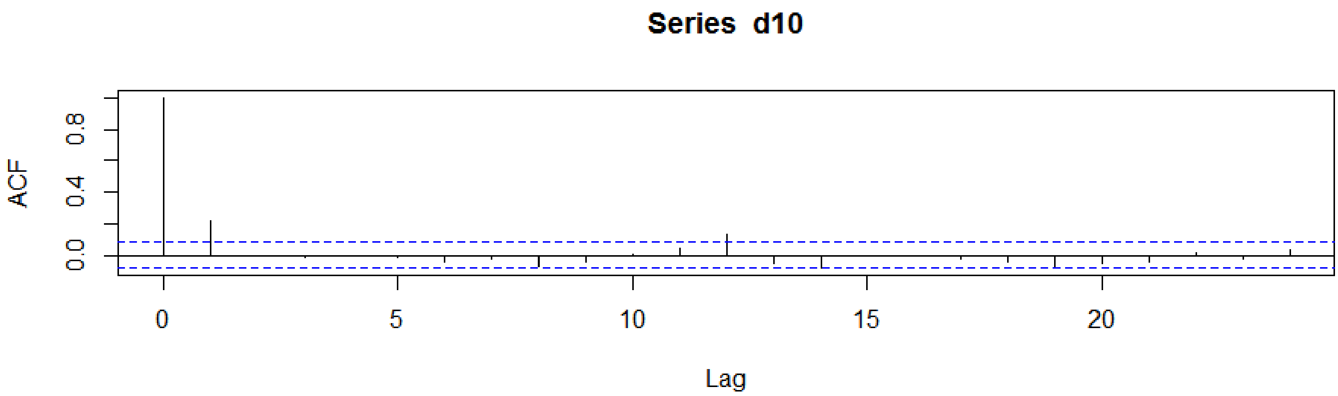
```
>
> d2=da$dec2#select the Decile 2 returns
> dec2=ts(d2,frequency=12,start=c(1961,1))
> par(mfcol=c(2,1))
> plot(dec2,xlab='year',ylab='returns')
> title(main='(a):Simple returns')
> acf(d2,lag=24)
```



由图表明。5%水平上间隔为 1 的 ACF 显著不为 0

```
> d10=da$dec10#selectthe Decile 10 returns
> dec10=ts(d10,frequency=12,start=c(1961,1))
> par(mfcol=c(2,1))
> plot(dec10,xlab='year',ylab='returns')
> title(main='(a):Simple returns')
> acf(d10,lag=24)
```





由图表明，5%水平上间隔为 1 和 12 的 ACF 显著不为 0

（b）（c）

```
> library(forecast)
> auto.arima(da$dec2,ic="bic")
Series: da$dec2
ARIMA(0,0,1) with non-zero mean

Coefficients:
         ma1  intercept
      0.1307     0.0093
s.e.  0.0425     0.0022

sigma^2 estimated as 0.002223:  log likelihood=996.04
AIC=-1986.08   AICc=-1986.04   BIC=-1972.84
```

forecast 包可以通过 auto.arima()函数来发现恰当的 ARIMA 模型

此处我们选择 BIC 准则进行选择

根据 BIC 准则，我们使用 ARIMA(0,0,1)的一阶移动平均模型

```
> #生成模型
> M.arima = arima(da$dec2,order=c(0,0,1))
> M.arima

Call:
arima(x = da$dec2, order = c(0, 0, 1))

Coefficients:
         ma1  intercept
      0.1307     0.0093
s.e.  0.0425     0.0022

sigma^2 estimated as 0.002223:  log likelihood = 996.04,  aic = -1986.08
> #进行预测
> M.pre2 = forecast.Arima(M.arima,h=12)
> M.pre2
    Point Forecast       Lo 80      Hi 80       Lo 95      Hi 95
610   -0.001330078 -0.06175141 0.05909126 -0.09373654 0.09107638
611    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
612    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
613    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
614    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
615    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
616    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
617    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
618    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
619    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
620    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
621    0.009281983 -0.05165357 0.07021754 -0.08391091 0.10247487
>
```

```
> #对上述模型进行检验，诊断残差序列是否为白噪声
> names(M.pre2)
 [1] "method"    "model"     "level"     "mean"      "lower"     "upper"     "x"         "xname"
 [9] "fitted"    "residuals"
> acf(M.pre2$residuals,lag.max=5)
> Box.test(M.pre2$residuals,lag=5,type='Ljung')

 Box-Ljung test

data:  M.pre2$residuals
X-squared = 3.228, df = 5, p-value = 0.6649
```