# 北京航空航天大学

2021－2022 学年 第二学期期末

# 《R 语言编程及应用》

# 考 试 A 卷

任课教师：马 杰

班 级_____学 号 _____

姓 名_____成 绩 _____

考试日期:2022 年 6 月 15 日

A

班号 ____ 学号___ 姓名____ 成绩_____

# 《R 语言编程及应用》期末考试卷

注意事项：

1、因疫情原因，本次考试为在线开卷考试，允许看书、看课件、看参考资料，也允许调用或改编课堂上讲过的代码或函数，答题时请“标清题号”。

2、绝不允许相互抄袭，一旦发现有雷同卷、文字/用词/举例等雷同作弊情形，抄袭与提供抄袭答案的同学，都记为作弊 0 分。

试题：

一、数据分析与编程应用题……(5 小题，共 100 分)

**中国股票市场指数的实证数据分析**

股票市场是中国金融市场的核心组成部分，也是机构投资者与个人投资者展开投资行为的主战场。现搜集了上证综指（股票代码：000001）、深圳成指（股票代码：399001）、创业板指（股票代码：399006）等三只主要股票市场指数的交易数据，分别存于文件“000001.csv、399001.csv、399006.csv”中，请针对这些数据，在线编程展开数据分析、并计算解决以下问题（提示：1）所有分析与计算，均应写出相应的 R 代码、编程解决，其他操作无效；2）输出的表格或图表，请直接使用 R 计算结果的截图，必要时按要求辅之以“简洁的文字说明”）：

1．读取三只股票指数的交易数据：

```
#读取交易数据
SHANG <- read.csv("000001.csv")
SHEN <- read.csv("399001.csv")
CHUANG <- read.csv("399006.csv")
```

1）以创业交易有效数据的最大时段为准，截取上综指、深成指交易日期，使得三只指数的交易时间可匹配、均有交易数据，然后输出展示上综指的前 6 行交易数据信息；

```
#按照创业板指有效数据最大时段选取上证综指和深证成指数据
ndays <- nrow(CHUANG)
SHANG <- na.omit(SHANG[(nrow(SHANG)-ndays+1:nrow(SHANG)),])
SHEN <- na.omit(SHEN[(nrow(SHEN)-ndays+1:nrow(SHEN)),])
head(SHANG)
```

```
> head(SHANG)
          日期 股票代码     名称    收盘价    最高价    最低价    开盘价    前收盘   涨跌额
4769 2010/6/1  '000001 上证指数 2568.283 2598.897 2534.267 2577.763 2592.147 -23.864
4770 2010/6/2  '000001 上证指数 2571.423 2572.343 2521.057 2548.542 2568.283    3.14
4771 2010/6/3  '000001 上证指数 2552.656 2596.137 2551.529 2575.770 2571.423 -18.767
4772 2010/6/4  '000001 上证指数 2553.593 2556.969 2527.873 2536.231 2552.656   0.937
4773 2010/6/7  '000001 上证指数 2511.729 2527.104 2491.661 2508.329 2553.593 -41.864
4774 2010/6/8  '000001 上证指数 2513.947 2531.882 2491.647 2510.084 2511.729   2.218
      涨跌幅    成交量     成交金额
4769 -0.9206 74666113 79814661762
4770  0.1223 64866634 69692518811
4771 -0.7298 69502328 74462434052
4772  0.0367 56518495 60645956510
4773 -1.6394 65697988 70153783991
4774  0.0883 63032468 69342653328
```

2）在此基础上，新创建一个数据框，其中仅包括交易日期信息、以及三只指数的收盘价与成交量信息，然后输出任意给定的时段内（如本题设定为：2011.4.1~2022.3.31）三只指数的后6行交易信息。[提示：编程中的变量名，可自行命名，以“简洁易识别”为原则]

```
#生成新数据框并筛选数据
new.data <- data.frame(Date = SHANG[,1], Close_Shanghai = SHANG[,4],
Close_Shenzhen = SHEN[,4],
                       Close_Chuang = CHUANG[,4], Amount_Shanghai =
SHANG[,11], Amount_Shenzhen =
                         SHEN[,11], Amount_Chuang = CHUANG[,11])
date <- as.Date(new.data$Date)
data.sample <- new.data[(which(date == "2011-04-01"):which(date == "2022-
03-31")),]
tail(data.sample)
```

```
> tail(data.sample)
          Date Close_Shanghai Close_Shenzhen Close_Chuang Amount_Shanghai Amount_Shenzhen
2871 2022/3/24       3250.264       12305.50     2706.215       329006654     12531702112
2872 2022/3/25       3212.240       12072.73     2637.944       340020006     12145753583
2873 2022/3/28       3214.503       11949.94     2594.128       344994873     12190273644
2874 2022/3/29       3203.939       11895.08     2592.666       316273649     11993282284
2875 2022/3/30       3266.596       12263.80     2696.826       366753677     14946579655
2876 2022/3/31       3252.203       12118.25     2659.492       398439934     14694076937
     Amount_Chuang
2871    1368002438
2872    1266519881
2873    1225967642
2874    1236946999
2875    1621278726
2876    1355159768
```

(25 分)

2．在股票市场上，存在着广泛的假说或猜想，认为存在“红周一、黑周五”现象，即周一易上涨、周五易下跌。请以给定时段 2011.4.1~2022.3.31 内的上综指的数据信息为例，计算并输出相应的对数收益率的描述性统计结果（至少应包括均值、方差、偏度、峰度等数值特征），简单验证下“红周一、黑周五”现象是否存在。

```
#加载需要的 R 包
library(moments)
library(fBasics)
library(tseries)
```

```
library(FinTS)
library(zoo)
library(xts)
library(quantmod)

#首先定义描述性统计函数
descrip.plot <- function(dat, ifqq, ifdensity) {
  res <- c(length(dat), mean(dat), median(dat), sd(dat), max(dat),
min(dat),
           skewness(dat), kurtosis(dat))
  jbtst <- jarque.test(dat)
  adftst <- adf.test(dat)
  archtst <- ArchTest(dat)
  res <- c(res, round(jbtst$statistic, 3), round(jbtst$p.value, 3),
           round(adftst$statistic, 3), round(adftst$p.value, 3),
           round(archtst$statistic, 3), round(archtst$p.value, 3))
  res <- as.list(res)
  names(res) <- c("Obs", "Mean", "Median", "sd", "max", "min",
                  "skewness", "kurtosis", "JBtest", "JB-p.value",
                  "adftest", "adf-p.value", "archtest", "arch-p.value")
  if (ifqq == 1) {
    qqnorm(dat, main = "QQPlot")
    qqline(dat)
  }
  if (ifdensity == 1) {
    plot(density(dat), col = "green", xlim = c(min(dat), max(dat)))
    s <- c(min(dat):max(dat))
    lines(s, dnorm(s, mean = mean(dat), sd = sd(dat)), col = "red", lty =
2)
  }
  res
}

#计算收益率并提取周一、周五数据，进行描述性统计
days <- weekdays(date.sample)
clsprc <- data.sample[,2]
cls <- zoo(clsprc, date.sample)
re <- returns(cls, method = "continuous", percentage = TRUE)
re.Mon <- re[which(days == "星期一")]
re.Fri <- re[which(days == "星期五")]
par(mfrow=c(1,2))
descrip.plot(as.vector(na.omit(re.Mon)), 1, 1)
descrip.plot(as.vector(na.omit(re.Fri)), 1, 1)
```

```
#t 检验：检验二者均值
t.test(re.Mon, re.Fri, alternative = "greater")
```

```
> descrip.plot(as.vector(na.omit(re.Mon)), 1, 1)
$`Obs`
[1] 520

$Mean
[1] 0.02125884

$Median
[1] 0.1307153

$sd
[1] 1.641353

$max
[1] 5.554206

$min
[1] -8.873175

$skewness
[1] -1.217092

$kurtosis
[1] 9.118312

$JBtest
[1] 939.445

$`JB-p.value`
[1] 0
```

```
$adftest
[1] -7.857

$`adf-p.value`
[1] 0.01

$archtest
[1] 46.09

$`arch-p.value`
[1] 0
```

对周一收益率序列的描述性统计如上所示。

```
> descrip.plot(as.vector(na.omit(re.Fri)), 1, 1)
$`Obs`
[1] 532

$Mean
[1] 0.03328465

$Median
[1] 0.07435283

$sd
[1] 1.259057

$max
[1] 4.71146

$min
[1] -7.684535

$skewness
[1] -1.004358

$kurtosis
[1] 9.559583

$JBtest
[1] 1043.231

$`JB-p.value`
[1] 0
```

```
$adftest
[1] -8.835

$`adf-p.value`
[1] 0.01

$archtest
[1] 148.732

$`arch-p.value`
[1] 0
```

对周五收益率的描述性统计如上图所示。

对二者进行均值 t 检验，原假设为均值相等，备择假设为周一收益率高于周五收益率。

```
> t.test(re.Mon, re.Fri, alternative = "greater")

        Welch Two Sample t-test

data:  re.Mon and re.Fri
t = -0.13312, df = 973.07, p-value = 0.5529
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.1607569        Inf
sample estimates:
 mean of x  mean of y
0.02125884 0.03328465
```

由 t 检验结果可知，p 值很大（超过 0.5），因此接受原假设。即认为这一时段上综指收益率的“红周一黑周五”现象不显著。

(15 分)

3．仍以给定时段 2011.4.1~2022.3.31 内三只股票指数的交易数据为对象，先测算输出三种股指对数收益率的相关性矩阵；然后考虑交互影响与价量结合，以上综指与深成指的收益率、及创业板成交量的对数变化率作为自变量，分析它们对创业板指数收益率的量化影响。

```
#计算三种股指的对数收益率并输出相关性矩阵
clsprc.1 <- as.numeric(data.sample[,3])
clsprc.2 <- as.numeric(data.sample[,4])
cls.1 <- zoo(clsprc.1, date.sample)
cls.2 <- zoo(clsprc.2, date.sample)
re.1 <- returns(cls.1, method = "continuous", percentage = TRUE)
re.2 <- returns(cls.2, method = "continuous", percentage = TRUE)
v1 <- na.omit(as.vector(re))
v2 <- na.omit(as.vector(re.1))
v3 <- na.omit(as.vector(re.2))
M <- cbind(v1, v2, v3)
(res <- cor(M))
```

```
> (res <- cor(M))
          v1        v2        v3
v1 1.0000000 0.9234944 0.7296356
v2 0.9234944 1.0000000 0.8400369
v3 0.7296356 0.8400369 1.0000000
```

相关性矩阵如上图所示。（其中 v1,v2,v3 分别表示上综指、深成指、创业板指）

可见，上综指和深成指的收益率相关性很强，深成指和创业板指的收益率相关性较强，但上综指和创业板指的收益率相关性一般。

```
#以上综指和深成指的收益率、创业板指成交量的对数变化率为自变量
#以创业板指收益率为因变量
#建立多元回归模型
amount <- data.sample[,7]
amount.log <- 100 * diff(log(amount))
```

```
M.lm <- data.frame(Chuang=v3, Shang=v1, Shen=v2, Amount=amount.log)
model.lm <- lm(Chuang~., data = M.lm)
summary(model.lm)
```

```
> summary(model.lm)

Call:
lm(formula = Chuang ~ ., data = M.lm)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4607 -0.4553  0.0097  0.4773  4.8188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.039238   0.019852   1.977   0.0482 *
Shang       -0.454485   0.039299 -11.565  < 2e-16 ***
Shen         1.364405   0.032544  41.926  < 2e-16 ***
Amount       0.006795   0.001059   6.415 1.66e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.026 on 2670 degrees of freedom
Multiple R-squared:  0.7244,    Adjusted R-squared:  0.7241
F-statistic:  2339 on 3 and 2670 DF,  p-value: < 2.2e-16
```

多元回归结果如上所示。

各变量与创业板指收益率的线性关系如参数所示，t 检验显示各变量与创业板指收益率的线性关系均显著，F 检验结果也显著，且调整后的 R^2 值变动不大，说明整体回归模型建构较好。但 R^2 值小于 0.8，整体拟合优度一般。

从具体估计值可得，深成指收益率对创业板指收益率的影响最大，上综指收益率与创业板指收益率之间呈负相关关系。

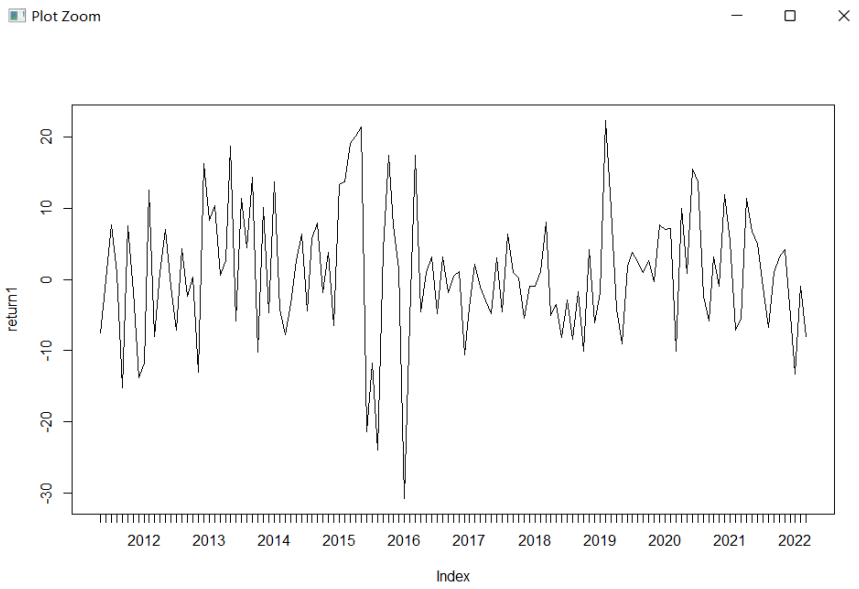
(15 分)

4．以给定时段 2011.4.1~ 2022.3.31 内创业板指数的收盘价数据为基础：

1）将之转化为月度信息，再计算月百分比对数收益、并输出其时序图；

2）对月百分比对数收益序列进行 Arch 效应检验，采用最基础的 Garch(1,1)模型对其建模，并分析建模的效果。

```
#将创业板指收盘价转为月度信息并绘制时序图
cls.month <- to.monthly(cls.2)
monthly.close <- cls.month[,4]
return1 <- 100 * diff(log(monthly.close))
par(mfrow=c(1,1))
plot(return1)
```




收益率的时序图如上所示。

```
#进行 ARCH 效应检验并建模
library(fGarch)
(archtst <- ArchTest(return1))
Garchmodel <- garchFit(~garch(1,1), data = return1, trace = FALSE)
summary(Garchmodel)
```

```
> (archtst <- ArchTest(return1))

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  return1
Chi-squared = 27.882, df = 12, p-value = 0.005756
```

ARCH 效应检验，p 值小于 0.01，因此拒绝原假设，认为序列存在 ARCH 自回归效应。

```
> summary(Garchmodel)

Title:
 GARCH Modelling

Call:
 garchFit(formula = ~garch(1, 1), data = return1, trace = FALSE)

Mean and Variance Equation:
 data ~ garch(1, 1)
<environment: 0x000002437463e448>
 [data = return1]

Conditional Distribution:
 norm

Coefficient(s):
      mu     omega    alpha1     beta1
 0.46415  10.43768   0.18871   0.68137

Std. Errors:
 based on Hessian

Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu       0.46415     0.69175    0.671   0.5022
omega   10.43768     5.91004    1.766   0.0774 .
alpha1   0.18871     0.08018    2.354   0.0186 *
beta1    0.68137     0.10163    6.705 2.02e-11 ***
---

Log Likelihood:
 -466.1936    normalized:  -3.55873

Description:
 Wed Jun 15 15:13:47 2022 by user: lenovo


Standardised Residuals Tests:
                                Statistic p-Value
 Jarque-Bera Test   R    Chi^2  0.2540125 0.8807282
 Shapiro-Wilk Test  R    W      0.9966261 0.990983
 Ljung-Box Test     R    Q(10)  8.300641  0.5994964
 Ljung-Box Test     R    Q(15)  12.92796  0.6078626
 Ljung-Box Test     R    Q(20)  21.39722  0.3740929
 Ljung-Box Test     R^2  Q(10)  2.751044  0.9866911
 Ljung-Box Test     R^2  Q(15)  6.635537  0.9669859
 Ljung-Box Test     R^2  Q(20)  9.013971  0.9827449
 LM Arch Test       R    TR^2   3.966699  0.9840298

Information Criterion Statistics:
     AIC      BIC      SIC     HQIC
7.178529 7.266321 7.176737 7.214203
```

GARCH(1,1)建模结果如上图所示。标准残差的 Ljung-Box 检验 p 值很大，认为 GARCH 模型拟合效果不佳。

(20 分)

5. 股市技术分析中，BOLL 线是约翰.布林基于标准差原理设计的常用指标，包含有中轨、上轨、下轨三条线，其计算原理可简单描述为：

$$Middle_t = \frac{P_{t-1} + P_{t-2} \cdots + P_{t-N}}{N}$$

$$Up_t = Middle_t + \alpha * Sd(P_t, N)$$

$$Down_t = Middle_t - \alpha * Sd(P_t, N)$$

其中$Sd(P_t, N)$为$t$时刻之前的N个交易日收盘价的标准差。

但通常交易软件中，经验参数$\alpha$只能设为整数（通常为 2），这并不完全符合正态分布假定下的分位数值；假定未来股价变化服从正态分布，保险起见，要求 BOLL 线上下轨以 99%的概率动态包括未来股价波动区间，此时参数$\alpha$的理论值应取 2.57583 较为精准。为此，请自行编写函数 Boll(x, date, N, alpha)，以绘制给定时段内的“股价及其布林线走势图”——其中，x 为收盘价$P_t$，date 为交易日期，N 为移动平均的窗宽，alpha 为给定置信水平下的分位数。

新设定时段“2022.1.4~2022.3.31”，利用自编的函数与新给定时段的样本数据，输出创业板指日收盘价的股价及其布林线走势图（要求的置信水平为 99%，N=20）。输出的结果中，因交易日期信息字段较长，要求日期刻度标签与横坐标垂直、并合理布局画布结构，以便能清楚标示每个交易日期。

```r
#Boll 线
Boll <- function(x, date, N, alpha) {
  miu <- c()
  up <- c()
  low <- c()
  for (i in (N+1):length(x)){
    miu <- c(miu, mean(x[i-N:i-1]))
    up <- c(up, miu[length(miu)] + alpha * sd(x[i-N:i-1]))
    low <- c(low, miu[length(miu)] - alpha * sd(x[i-N:i-1]))
  }
  plot(xts(miu, up, low, order.by = date))
}

Boll(clsprc.2, date.sample, 20, 2.57583)
```

(25 分)