西南财经大学

Southwestern University of
Finance and Economics

R 语言课程作业

学生姓名：杨双杰

所在学院：会计学院

专　　业：财务管理

学　　号：**216120Z6024**

任课教师：李　伊

2018 年 1 月

一、数据说明

$$\mathrm{Y_{it}} = \beta_0 + \beta_1 \mathrm{downturn_{it}} + \beta_2 controls_{it} + \sum ind + \sum \mathrm{year} + u_i + \varepsilon_{it}$$

这个模型用于研究经济周期 X（downturn）对审计质量 Y（opinion_dum 和 lnfee）具有怎么的影响关系，此外模型中除了行业（ind）、年度（year）外，还包含其他 10 个控制变量（controls），各变量具体说明如下表。

为了得到各个变量的数据，从 CSMAR，DIB 数据库里一共下载了近 20 份原始数据，现在利用 R 对这些数据进行清洗最后合并为一份最终数据。具体流程主要包括：①清洗各零散的控制变量，然后合并为一份“control_data”。②清洗解释变量得到含有经济周期虚拟变量 downturn 的数据“GDP_grate”，清洗被解释变量得到数据“audit”。③最后将所有变量合并为一份“final_data”，并结合前面清洗的数据绘制 3 张图。

变量选取列表

被解释变量	变量说明
Opinion_dum	审计意见，虚拟变量，事务所出具非标准审计意见则取值 1，否则为 0
lnfee	审计费用，取对数
解释变量	
downturn	经济周期，虚拟变量，审计当年处于经济衰退期取值 1，否则为 0
控制变量	
lnasset	企业规模，年末总资产对数
roa	盈利能力，资产收益率
lev	财务杠杆，资产负债率
loss	盈亏状况，虚拟变量，公司盈利取值 1，亏损为 0
growth	公司成长性，营业收入增长率
ndts	非债务税盾，固定资产折旧、无形资产摊销等资产折耗占总资产比率
soe	控制人性质，虚拟变量，公司实际控制人为国有，取值为 1，否则为 0
sanction	监管力度，每年被证监会处罚的事务所数量
complex	审计复杂程度，公司存货与应收账款占总资产的比例
peratio	投资需求，市盈率
ind	行业，虚拟变量，若为所属行业，取值为 1，否则为 0

二、清洗流程

（一）清洗控制变量

首先设置项目工作路径，加载清洗所需要的 R 包。

```
#Tidy Control Variables ------------------------------------------------
setwd("./control")
library(tidyverse)
library(lubridate)
```

1. 第一份数据读取主要包含资产负债表变量的 balance.csv 文件，并自定义变量名，筛选出合并报表数据（A）以及年度数据（月份“12”），并删除不需要的四个变量（typrept,fasset,fadisposal,iasset），数据存为 balance。然后让年度变量 year 只显示年度，同时由于股票代码 stkcd 前面的 0 被省去，此处统一股票代码的数据格式为 6 位数，以便后续数据合并。

```
# the first data
col_names <- c("stkcd", "year", "typrept", "recei",
               "inv", "fasset", "fadisposal", "iasset", "assets")
balance <- read_csv("balance.csv",col_names = col_names) %>%
               filter(typrept=="A",str_sub(year,6,7)=="12") %>%
                 select(-typrept,-fasset,-fadisposal,-iasset)
balance$year <- year(ymd(balance$year))
balance$stkcd <- sprintf("%06d",as.numeric(balance$stkcd))
```

2.第二份数据读入 grevenue.csv 文件，并自定义变量名，筛选合并报表和年度数据，之后删除不需要的变量，对年度 year 和股票代码 stkcd 的处理同上。

```
#the second data
growth <- read_csv("grevenue.csv",col_names = c(
             "stkcd", "year", "typrept", "ind", "grev", "susgrate")) %>%
              filter(typrept=="A",str_sub(year,6,7)=="12") %>%
               select(-typrept,-susgrate)
growth$year <- year(ymd(growth$year))
growth$stkcd <- sprintf("%06d",as.numeric(growth$stkcd))
```

3.第三份数据处理过程同前。

```
#the third data
lev <- read_csv("LEV.csv",col_names = c(
             "stkcd", "year", "typrept", "ind", "lev")) %>%
               filter(typrept=="A",str_sub(year,6,7)=="12") %>%
                 select(-typrept)
lev$year <- year(ymd(lev$year))
lev$stkcd <- sprintf("%06d",as.numeric(lev$stkcd))
```

4.第四份数据处理过程同前。

```
#the fourth data
depamor <- read_csv("depamor.csv",col_names = c(
               "stkcd", "year", "typrept", "ind", "depamor")) %>%
                  filter(typrept=="A",str_sub(year,6,7)=="12") %>%
                    select(-typrept)
depamor$year <- year(ymd(depamor$year))
depamor$stkcd <- sprintf("%06d",as.numeric(depamor$stkcd))
```

5.第五份数据处理过程同前。

```
#the fifth data
netincome <- read_csv("NI.csv",col_names = c(
              "stkcd", "year", "typrept", "ni")) %>%
                filter(typrept=="A",str_sub(year,6,7)=="12") %>%
                 select(-typrept)
netincome$year <- year(ymd(netincome$year))
netincome$stkcd <- sprintf("%06d",as.numeric(netincome$stkcd))
```

6.第六份数据处理过程同前。

```
#the sixth data
roa <- read_csv("ROAE.csv",col_names = c(
      "stkcd", "year", "typrept", "ind", "roa","roe","roic")) %>%
        filter(typrept=="A",str_sub(year,6,7)=="12") %>%
          select(-typrept,-roe,-roic)
roa$year <- year(ymd(roa$year))
roa$stkcd <- sprintf("%06d",as.numeric(roa$stkcd))
```

7.第七份数据处理过程同前。

```
#the seventh data
roegrate <- read_csv("roegrate.csv",col_names = c(
               "stkcd", "year", "typrept", "ind", "roegrate")) %>%
                 filter(typrept=="A",str_sub(year,6,7)=="12") %>%
                   select(-typrept)
roegrate$year <- year(ymd(roegrate$year))
roegrate$stkcd <- sprintf("%06d",as.numeric(roegrate$stkcd))
```

8.第八份数据不含子公司数据，所以不需要删选出合并报表数据（A），其余过程同前。

```
#the eighth data
peratio <- read_csv("PEratio.csv",col_names = c(
           "stkcd", "year", "ind", "peratio")) %>%
               filter(str_sub(year,6,7)=="12")
peratio$year <- year(ymd(peratio$year))
peratio$stkcd <- sprintf("%06d",as.numeric(peratio$stkcd))
```

9.第九份数据为对审计师的行政处罚（saction）数据，这里只需读入并删除一个不需要的数据（discipline）。

```
#the nineth data
saction <- read_csv("saction.csv") %>%
           select(-discipline)
```

10.第十和第十一份数据共同得出需要的公司性质变量（soe），由两份不同来源的数据经过清洗合并得出。首先读入 DIB 来源的数据，由于含有中文，读入时指定编码格式为 GB18030，筛选所需的 1，4，5 列的变量，并重命名，再生成一个缺失值变量 soe 用于根据公司性质 type 变量是否含有关键字段而分别赋予 0 和 1，即二值变量。第十一份数据处理过程与此类似。最后将两份 soe 数据进行 full_join,再利用 unique 去重。

```
#the tenth and eleventh data
soe1 <- read_csv("境内公司基本信息-公司性质-DIB.csv",
                 locale = locale(encoding = "gb18030")) %>%
                 select(c(1,4,5)) %>%
                 rename(
                    stkcd = 证券代码,
                    type = 公司性质,
                    controller = 实际控制人
                  ) %>%
                    mutate(soe=NA)
soe1$stkcd <- str_sub(soe1$stkcd,1,6)

soe1$soe[soe1$type=="国有企业"] <- 1
soe1$soe[soe1$type=="其他" & str_detect(
                        soe1$controller,"国有资产监督管理")] <- 1
soe1$soe[is.na(soe1$soe)] <- 0
soe1 <- soe1 %>%
       select(stkcd,soe)

soe2 <- read_csv("上市公司实际控制人.csv",
                 locale = locale(encoding = "gb18030")) %>%
                  select(c(1,4)) %>%
                   rename(
                    stkcd = 代码,
                     contr_type = 控制人类型
                     ) %>%
                     mutate(soe=NA)

soe2$stkcd <- str_sub(soe2$stkcd,1,6)
soe2$soe[str_detect(
               soe2$contr_type,"国资委|国有企业|中央|政府")] <- 1
soe2$soe[is.na(soe2$soe)] <- 0
```

```
soe <- soe2 %>%
       select(-contr_type) %>%
         full_join(soe1) %>%
           unique()
```

11. 第十二份和第十三份数据分别为公司 ST 年份和 IPO 年份数据，这里处理过程依然与前述相同。只是进行了保存输出到 final 文件夹下。

```
#the twelveth and thirteenth data
ST <- read_csv("ST.csv",col_names = c("stkcd", "year", "typrept")) %>%
       filter(typrept=="A",str_sub(year,6,7)=="12") %>%
         select(-typrept) %>%
           mutate(st=1)
ST$year <- year(ymd(ST$year))
ST$stkcd <- sprintf("%06d",as.numeric(ST$stkcd))
write_csv(ST,"../final/ST.csv")

IPO <- read_csv("IPO.csv",skip = 1,col_names = c(
         "stkcd", "initial", "ipoyear", "listyear")) %>%
           filter(initial=="A") %>%
             select(stkcd,ipoyear)
IPO$ipoyear <- year(ymd(IPO$ipoyear))
IPO$stkcd <- sprintf("%06d",as.numeric(IPO$stkcd))
write_csv(IPO,"../final/IPO.csv")
```

12.这里将前述所有清洗的控制变量利用 reduce 函数进行 full 合并，并利用其他变量生成新的所需变量 lnassets、ndts、complex 和缺失值变量 loss，然后根据净利润 ni 大于还是小于 0 分别对 loss 赋值 0 或 1。最后保存输出为 control_data.csv。

```
#integrated controls
control_data  <- list(balance,depamor,growth,lev,
              netincome,roa,roegrate,peratio,soe,saction) %>%
               reduce(full_join) %>%
                mutate(
                  lnassets=log(as.numeric(.$assets)),
                  loss = NA,
                  ndts = as.numeric(.$depamor)/as.numeric(.$assets),
                  complex=(
         (as.numeric(.$inv)+as.numeric(.$recei))/as.numeric(.$assets)
                     ) %>%
                    select(-recei,-inv,-assets,-depamor)
control_data$loss[control_data$ni>=0] <- 1
control_data$loss[control_data$ni<0] <- 0
```

```
control_data <- control_data %>%
                  select(-ni)
write_csv(control_data,"../final/control_data.csv")
```

（二）清洗解释变量和被解释变量

先将工作路径切换到 XY 文件夹。

```
#Tidy XY ------------------------------------------------------------
setwd("../XY")
```

1. 处理 X。读入 GDP 数据，根据各年的 GDP 指数计算 GDP 增长率（gdpgrate），然后根据本年增长率相较于去年是否下降生成虚拟变量 downtown。然后选取所需变量存为 GDP_cycle 数据，并输出为 X.csv。

```
GDP <- read_csv("GDP.csv",col_names = c("year","gdpindex"),
                locale=locale(encoding = "gb18030"),skip = 7) %>%
                  select(-3) %>%
                     mutate(gdpgrate=(as.numeric(.$gdpindex)-100)/100 ,
                            downturn = NA)
GDP$downturn[GDP$gdpgrate<lag(GDP$gdpgrate)] <- 1
GDP$downturn[is.na(GDP$downturn)] <- 0

GDP_cycle <- select(GDP,'year',"downturn")
write_csv(GDP_cycle,"../final/X.csv")
```

2. 处理 Y。读入数据后，选取所需列并重命名，再过滤掉非年度数据，生成审计费用的对数（lnfees）。之后根据审计意见 opinon 变量是否为“标准无保留意见”对 opinion_dum 分别赋值 0 和 1。最后删除第 3 和 4 列不需要的变量输出为 Y.csv。

```
audit <- read_csv("audit.csv",locale=locale(encoding = "gb18030")) %>%
                select(1,3,5,13) %>%
                rename(
                  stkcd = 证券代码,
                  year = 会计截止日期,
                  opinion = 审计意见类型,
                  fees = 审计费用合计
                ) %>%
                  filter(str_sub(year,6,7)=="12") %>%
                  mutate(
                    lnfees = log(as.numeric(fees)),
                    opinion_dum = NA
                  )
audit$year <- year(ymd(audit$year))
audit$stkcd <- sprintf("%06d",as.numeric(audit$stkcd))
```

```
audit$opinion_dum[audit$opinion=="标准无保留意见"] <- 0
audit$opinion_dum[is.na(audit$opinion_dum)] <- 1

audit <- select(audit,-3,-4)
write_csv(audit,"../final/Y.csv")
```

（三）将所有变量合并清洗为最终数据

```
#Tidy Final Data ----------------------------------------------------------
setwd("../final")
```

这里先读取行业（ind）数据，这份数据较全，有利于减少最终数据中行业的缺失值。过程同样如前大多数处理，读入-命名-选取合并报表数据（A）-年度数据（12）-删除不需要的报告类型变量（-typre）。

```
ind <- read_csv("ind.csv",col_names = c(
                "stkcd","year","typre","ind"),skip=1) %>%
                  filter(typre=="A",str_sub(year,6,7)=="12") %>%
                    select(-typre)
ind$year <- year(ymd(ind$year))
ind$stkcd <- sprintf("%06d",as.numeric(ind$stkcd))
```

将前面各步骤处理的数据作为列表传入 reduce 进行 full_join 操作，然后分别剔除 ST 年和 IPO 年度数据，再剔除金融业（行业代码 J 开头，股票代码 2 或 9 开头），最后剔除资产负债率（lev）大于 1 的数据，最终再剔除缺失值和重复值。

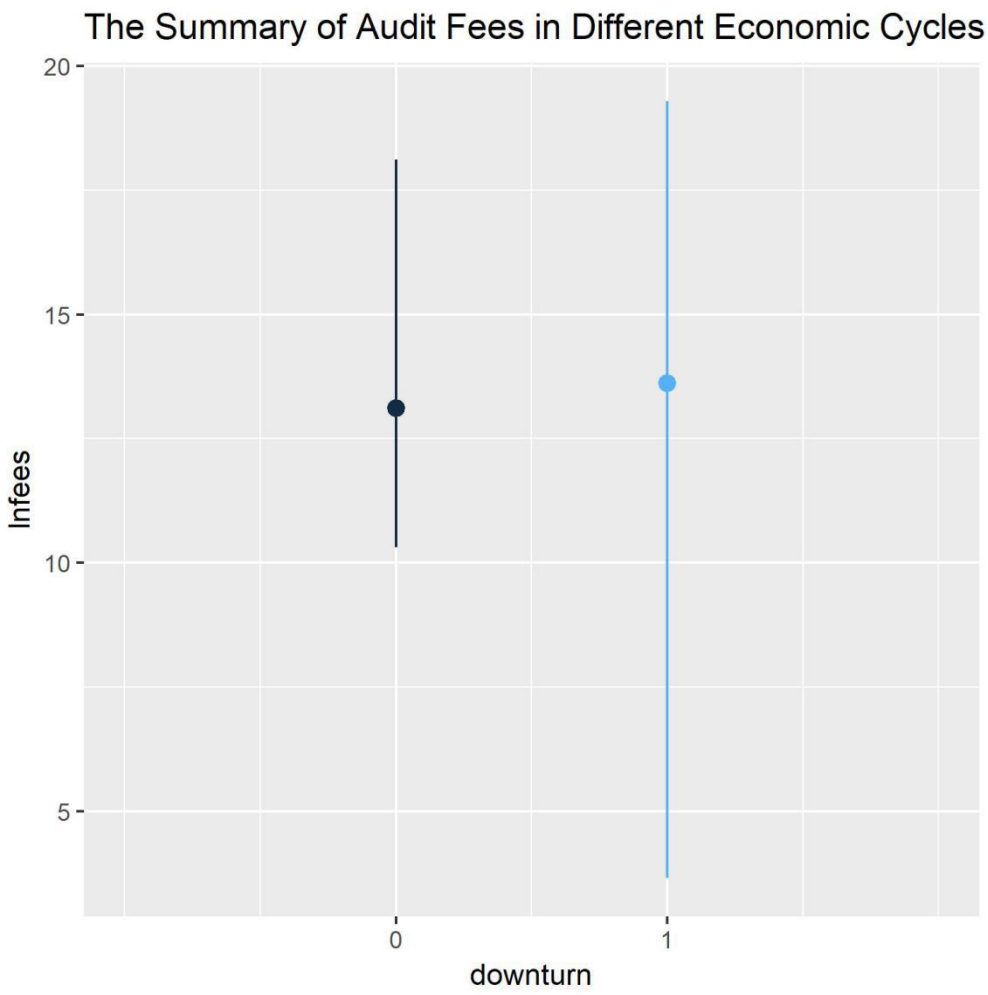
```
final_data <- list(audit,GDP_cycle,control_data,ST,IPO,ind) %>%
                  reduce(full_join) %>%
                   filter(st!=1|ipoyear!=year) %>%
                     filter(!str_detect(ind,"^J")) %>%
                       filter(!str_detect(stkcd,"^(2|9)")) %>%
                        filter(lev<1) %>%
                          select(-st,-ipoyear) %>%
                           drop_na() %>%
                             unique()

write_csv(final_data,"finaldata.csv")
```

三、利用清洗的数据作图

（一）第一张图为不同经济周期对数审计费用的统计差异

```
#picture 1
final_data %>%
     ggplot() +
        stat_summary(aes(x=downturn,y=lnfees,color=downturn),
```

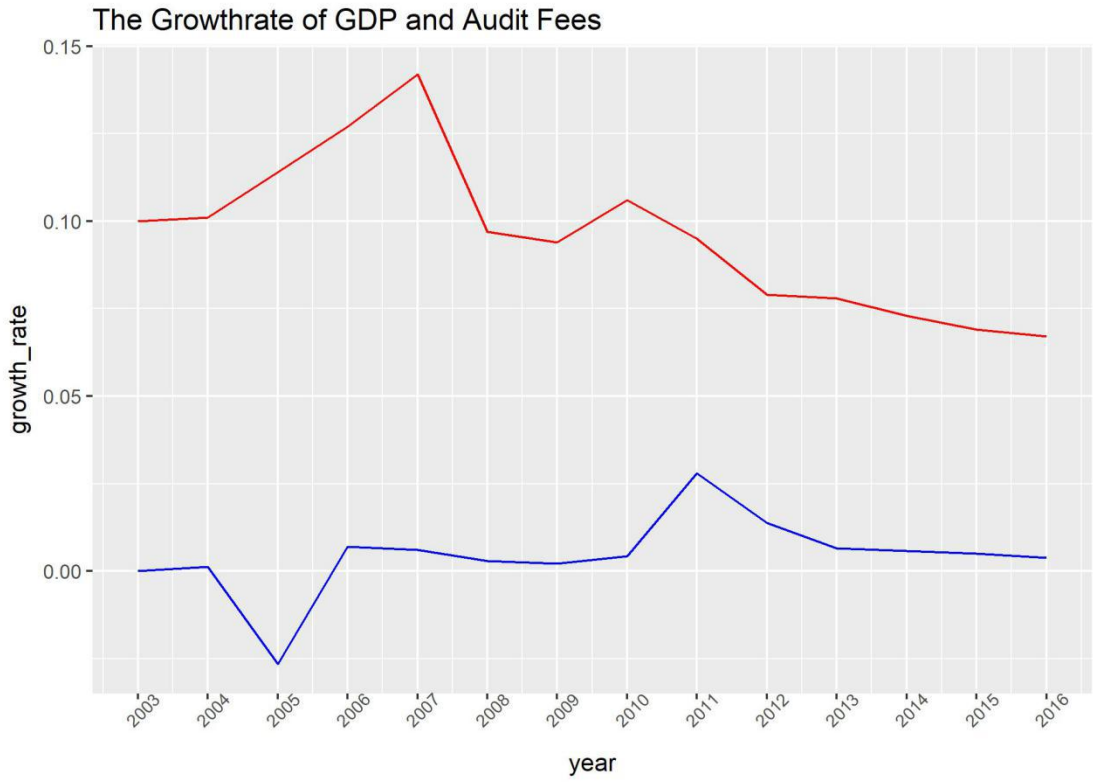
```
            show.legend = FALSE,
             fun.ymin = min,
             fun.ymax = max,
             fun.y = mean) +
        labs(title = paste(
        "The Summary of Audit Fees in Different Economice Cycles"
            )) +
        scale_x_continuous(limits=c(-1,2),breaks = seq(0,1))
    ggsave("diff_lnfees.jpeg")
```



（二）第二张图为 GDP 增长率和审计费用增长率的变化图，这里先根据年度分组计算对数审计费用的平均值，剔除平均后各年的重复值，进而计算对数审计费用的增长率，最后与处于不同表格的 GDP 增长率进行 left_join 操作。

```
#picture 2
GDP_lnfees <- audit %>%
          select(stkcd,year,lnfees) %>%
             group_by(year) %>%
                 mutate(mean_lnfees = mean(lnfees,na.rm=TRUE)) %>%
                    select(year,mean_lnfees) %>%
                        unique() %>%
```

```
                        ungroup() %>%
                          mutate(
          lnfeesgrate = (mean_lnfees-lag(mean_lnfees))/lag(mean_lnfees)
                                ) %>%
                            left_join(GDP) %>%
                              select(year,lnfeesgrate,gdpgrate) %>%
                                drop_na()
ggplot(data=GDP_lnfees) +
      geom_line(mapping = aes(x=year,y=gdpgrate),color="red") +
          scale_x_continuous(breaks = GDP_lnfees$year) +
             theme(axis.text.x = element_text(angle=45,size=8)) +
             geom_line(mapping=aes(x=year,y=lnfeesgrate),color="blue") +
                      labs(y="growth_rate",
                           title="The Growthrate of GDP and Audit Fees ")
ggsave("growth_rate.jpeg")
```



（三）第三张图为不同经济周期审计意见的分布图。

```
#picture 3
final_data %>%
  ggplot() +
    geom_bar(aes(x=opinion_dum,fill=downturn),show.legend = FALSE) +
      facet_wrap(~downturn) +
        labs(
            title="The Distribution of Auditing
                   Opinions in Different Cycles"
```

```
        ) +
      scale_x_continuous(breaks=seq(0,1,1))
ggsave("opinion_distribution.jpeg")
```