**实证知识点总结**

目录

[一、数据预处理 2](#_Toc104400411)

[二、 Stata中数据的基本处理 4](#_Toc104400412)

[(1)取年份 4](#_Toc104400413)

[(2)剔除数据中的样本 4](#_Toc104400414)

[(3)生成新的变量 4](#_Toc104400415)

[(4)给变量添加标签 4](#_Toc104400416)

[(5)生成行业、年度虚拟变量 4](#_Toc104400417)

[(6)基本数据匹配 4](#_Toc104400418)

[（7）Winsor处理，剔除异常值 5](#_Toc104400419)

[（8）剔除缺失值 5](#_Toc104400420)

[（9）回归结果输出 5](#_Toc104400421)

[三、 描述性统计 6](#_Toc104400422)

[输入论文的代码写法（很实用） 6](#_Toc104400423)

[四、相关性分析\* 6](#_Toc104400424)

[五、实证模型分析 6](#_Toc104400425)

[1.单变量分析 6](#_Toc104400426)

[2.最小二乘法 7](#_Toc104400427)

[面板数据 7](#_Toc104400428)

[3.分位数回归 7](#_Toc104400429)

[4.logit模型 7](#_Toc104400430)

[5.Tobit模型 7](#_Toc104400431)

[6.probit模型 8](#_Toc104400432)

[7.系统GMM 8](#_Toc104400433)

[8.双重差分模型（DID） 8](#_Toc104400434)

[9.豪斯曼检验 8](#_Toc104400435)

[10.时间固定效应检验 8](#_Toc104400436)

[11.个体固定效应检验 8](#_Toc104400437)

[12.行业固定效应检验 9](#_Toc104400438)

[六、内生性问题解决方法（附具体模型） 9](#_Toc104400439)

[1.PSM模型 9](#_Toc104400440)

[PSM1————近邻匹配 使用最近邻匹配1:1原则 9](#_Toc104400441)

[PSM2- 核匹配 10](#_Toc104400442)

[PSM匹配3-半径匹配 10](#_Toc104400443)

[2.滞后期模型 10](#_Toc104400444)

[3.工具变量：两阶段最小二乘 11](#_Toc104400445)

[4.Heckman两阶段 11](#_Toc104400446)

[七、调节效应检验：机制分析 11](#_Toc104400447)

[八.中介效应检验：路径分析（三种方法） 11](#_Toc104400448)

[1.回归模型三阶段法 11](#_Toc104400449)

[2.Sobel-Goodman检验 12](#_Toc104400450)

[3.Bootstrap检验 12](#_Toc104400451)

[九．其他命令 13](#_Toc104400452)

[\*中值分组 13](#_Toc104400453)

[\* gen & egen 13](#_Toc104400454)

[字符转化为数值的stata命令 13](#_Toc104400455)

[Split分割函数 13](#_Toc104400456)

[字符串处理函数 14](#_Toc104400457)

[合并数据 14](#_Toc104400458)

# 一、数据预处理

\* 1.如何导入数据到Stata里

\* （国泰安TXT原始文件、Excel原始文件+万德原始文件）

\*=======================================================

\*导入国泰安的TXT原始文件方式

insheet using E:\stata16\FS\_Combas.txt,clear // 其中 “E:\stata16\FS\_Combas.txt”可自行根据电脑中的stata安装位置调整,“FS\_Combas”为你需要导入的文件名称

\*导入国泰安的Excel原始文件

import excel 股东股权质押统计表.xlsx, firstrow clear // 其中“股东股权质押统计表”为你需要导入的文件名称

\*=========================================================

\* 导入Wind原始文件方式

\* （由于Wind数据处理比较复杂，所以以下分步骤展示）

=======================================================

\* 其一，将wind数据导入STATA之前的准备工作

\*=========================================================

\*第一，在STATA中安装github的下载接口

net install github, from("https://haghish.github.io/github/")

\*第二，通过github接口安装github上的命令

github install zhbsis/TwoSort

\*第三，安装tidy命令

ssc install tidy

\*=========================================================

\* 其二，在Wind库下载需要的数据（比如下载产权性质）

\*=========================================================

wind数据库下载原始数据.xlsx

\*=========================================================

\* 其三，调整EXCEL格式，使其满足readWind代码需求

\*=========================================================

\*需要按照以下样式调整EXCEL格式： 股票代码 股票名称 时间1 时间2 ... 时间n

E:\stata16\如何将WIND数据库下载的数据导入stata\wind数据库原始数据和调整后的数据\调整合适的格式满足导入stata.xlsx

\*=========================================================

\* 其四，使用readWind命令导入

\*=========================================================

\*使用readWind命令导入

cd E:\stata16\如何将WIND数据库下载的数据导入stata\wind数据库原始数据和调整后的数据

clear

readWind, var(调整合适的格式满足导入stata) timeType(y) t0(2004)

/\*readWind, var(excel名称) timeType(时间类型) t0(样本起始时间) [split splitN(样本分割数) erase]

\*1.var括号里的excel名称就是从wind下载下来的是excel名称，比如我这个excel就命令为“调整合适的格式满足导入stata”

\*2.timeType括号里的时间类型是指下载下来的数据是年度、季度还是月度，所以取值分别为y/q/m（注意：不支持日）

\*3.t0括号里的是样本起始时间，比如你选择2004开始，那么t0()里就是2004；如果你用了月度数据，比如从2005，如果是年度数据以年为变化单位那就写1995

\*4.\*最后面的split splitN(n) erase这3个option是可选择的命令，主要是当我们的样本中的数据量太大时，会影响运算速度，所以这个命令可以使我们将样本分割成n个小样本，分别进行处理，最后合并在一起，也就是说，n取值越大分割的样本越小。【注意】n不要取值太小也不要太大。

\*/

\*基本处理

gen id=real(substr(stkcd,1,6))

gen year=real(time)

order id year

tsset id year

rename 调整合适的格式满足导入stata 公司属性

rename id 股票代码

rename year 年份

rename comp 公司名称

order 股票代码 年份 公司属性

drop time

\*关于如何导入Wind数据到Stata中，详细介绍请参考 https://bbs.pinggu.org/thread-10721029-1-1.html

\*==========================================================

\*

# 二、 Stata中数据的基本处理

# (1)取年份

gen year=real(substr(accper,1,4))

gen a=real(substr(accper,6,2))

drop if a!=12

drop accper a

// 将accper表示的年月日转换为年份，且格式转换为数值型

(2)剔除数据中的样本

drop if typrep=="B"

// 将typrep表示的B样本剔除掉

(3)生成新的变量

gen Size=ln(a001000000)

// 生成新的变量，且对老变量a001000000取ln

(4)给变量添加标签

label var year 年度

// 给year添加标签年度，以便于在stata中读懂每个英文的含义

(5)生成行业、年度虚拟变量

tab Industry,gen(indu) //生成行业虚拟变量

drop indu1 //为了避免共线性，删掉indu1

tab year,gen(time) //生成年度虚拟变量

drop time1 //为了避免共线性，删掉time1

// 作用——控制行业和年度效应

(6)基本数据匹配

\*一种方法：

merge 1:1 id year using 行业代码与公司年度对应至2020.dta

keep if \_merge==1 | \_merge==3

drop \_merge

//使用 merge 1:1 匹配数据 1:1指的是1对1匹配;id year 对应的是匹配变量,id是股票代码，year是年份;“行业代码与公司年度对应至2020.dta”是匹配数据的名称

\*另一种方法：

merge 1:1 id year using 行业代码与公司年度对应至2020.dta, nogen keep(1 3)

/\*nogen 表示不生成\_merge变量

keep(1 3) 相当于 keep if \_merge==1 | \_merge==3

keep里面具体对应

numeric equivalent

code word (results) description

-------------------------------------------------------------------

1 master observation appeared in master only

2 using observation appeared in using only

3 match observation appeared in both

\*/

（7）Winsor处理，剔除异常值

winsor2 变量, replace cuts(1 99) by(year)

// by(year)为可选择项，可加或者不加，（1 99）也可以调整

（8）剔除缺失值

\*foreach i in 变量1 变量2 变量3 {

drop if `i'==.

}

// 将以上变量转换为自己论文的变量值

（9）回归结果输出

reg y x x1 x2 x3

est store a1

xtreg y x x1 x2 x3,re

est store a2

xtreg y x x1 x2 x3,fe

est store a3

esttab a1 a2 a3 using 回归结果.rtf, replace b(%6.4f) t(%6.4f) nogap ar2 star(\* 0.1 \*\* 0.05 \*\*\* 0.01)

//以上将回归结果记为a1、a2及a3，输出。

# 三、 描述性统计

\*简单列表（不实用）

tabstat y x1 x2 x3, stat(max min mean p50 sd n)

## 输入论文的代码写法（很实用）

sum2docx 变量1 变量2 变量3 using C:\Users\我的论文.docx,replace stats(N mean(%9.4f) sd min(%9.2f) median(%9.2f) max(%9.2f)) title("Table 2: Summary Statistics")

//该代码直接输出论文格式，简单调整即可

# 四、相关性分析\*

logout,save(相关系数分析) word replace:pwcorr\_a 变量1 变量2 变量3, star1(.01) star5(.05) star10(.1)

# 五、实证模型分析

\* （由于实证模型种类比较多，所以以下分步骤展示）

## 1.单变量分析

logout,save (单变量分析) word replace:ttable3 被解释变量,by(解释变量虚拟形式) f(%8.4f) notitle

//解释变量虚拟形式为01虚拟变量形式

## 2.最小二乘法

\*基本形式

reg y x x1 x2 x3

\*论文跑数据形式

xi:reg 被解释变量 解释变量 控制变量1 控制变量2 i.year i.industry

est store a1

esttab a1 using 主变量回归结果.rtf,replace nogap ar2 b(%6.4f) t(%6.4f) star(\* 0.1 \*\* 0.05 \*\*\* 0.01)

### 面板数据（时间序列考虑多重共线性）

\*基本形式

xtset id year

xtreg y x x1 x2 x3,fe //固定效应（fe）

xtreg y x x1 x2 x3,re //随机效应（re）

\*论文跑数据形式（以固定效应模型为例，下面模型较多，就不一一举例，也可参考论文跑数据形式）

xi:xtreg 被解释变量 解释变量 控制变量1 控制变量2 i.year i.industry,fe

est store fe\_1

esttab fe\_1 using 固定效应方法.rtf,replace b(%6.4f) t(%6.4f) nogap ar2 star(\* 0.1 \*\* 0.05 \*\*\* 0.01)

## 3.分位数回归

sqreg y x1 x2 x3 ,q( .1 .25 .5 .75 .9)

//分别为0.1分位 0.25分位 0.5 0.75 0.9，可根据研究问题自行调整

## 4.logit模型

logit y x x1 x2 x3

## 5.Tobit模型

xttobit y x x1 x2 x3 , ll(0) nolog tobit

## 6.probit模型

probit y x x1 x2 x3

## 7.系统GMM

xtabond2 y L.y x x1 x2 x3, iv(x1 x2 x3) gmm(L.y L.(x), lag(1 2) c) robust twostep

## 8.双重差分模型（DID）

xi:xtreg 被解释变量 c.Ifexpgro#c.Iftime 控制变量,fe cl(id)

est store m1

esttab m1 using 双重差分模型主检验.rtf,replace ar2 b(%6.4f) t(%6.4f) star(\* 0.1 \*\* 0.05 \*\*\* 0.01)

// Ifexpgro为是否实验组，若是则取值为1，否则为0；Iftime为是否政策实施前后变量，若政策前则取值为0，若政策后，取值为1。c.Ifexpgro#c.Iftime 为Iftime与Ifexpgro的交乘项

## 9.豪斯曼检验

xtreg y x x1 x2 x3,re

est store re

xtreg y x x1 x2 x3,fe

est store fe

hausman fe re // y为被解释变量，x为解释变量，x1、x2及x3为控制变量；

//最终，选择标准为显著性在0.1及其以下选择固定效应，否则选择随机效应。

## 10.时间固定效应检验

xi:xtreg y x x1 x2 x3 i.year,fe

\*或者

xtreg y x x1 x2 x3 i.year,fe

## 11.个体固定效应检验

xtreg y x x1 x2 x3 i.id,fe

## 12.行业固定效应检验

xi:xtreg y x x1 x2 x3 i.industry,fe

\*或者

xtreg y x x1 x2 x3 c.industry,fe

# 六、内生性问题解决方法（附具体模型）

\* （由于该种类比较多，所以以下分步骤展示）

## 1.PSM模型

\* 由于该方法形式很多，所以以下举例

\*首先，调用数据

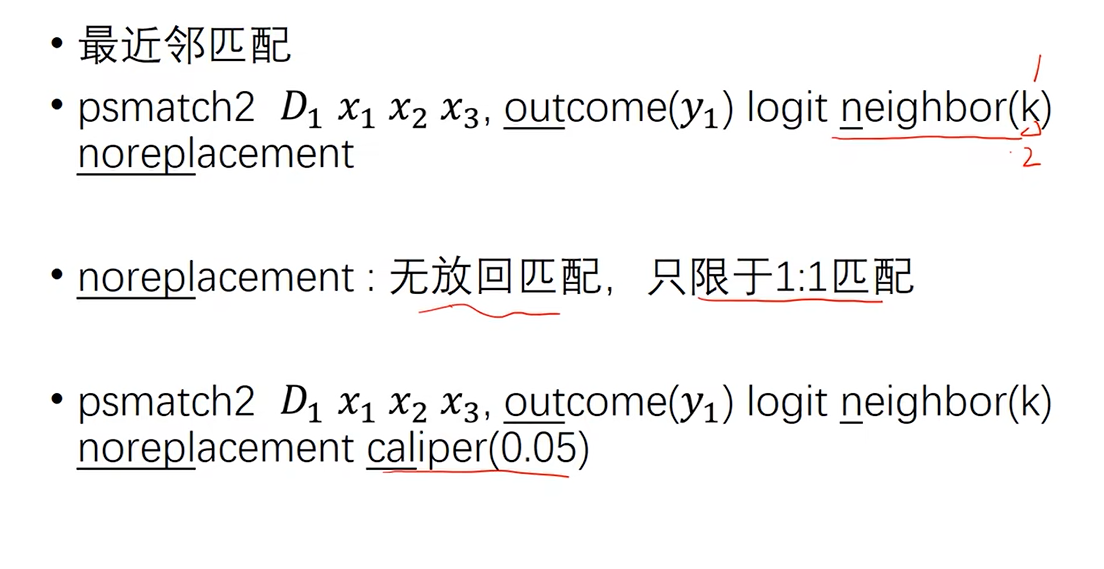
use 主回归数据.dta,clear

\*生成解释变量的虚拟变量形式，如下

bys year Industry:egen PLD\_RATE1\_meidan=median(PLD\_RATE1)

gen ifPLD\_RATE1 = (PLD\_RATE1>= PLD\_RATE1\_meidan) if !missing(PLD\_RATE1)

### PSM1————近邻匹配 使用最近邻匹配1:1原则



psmatch2 ifPLD\_RATE1（解释变量的虚拟变量形式） 匹配变量 ,outcome(被解释变量) logit neighbor(1) common ate ties

\* 匹配效果检验

pstest, both

\* 核密度图

\* 匹配前

tw (kdensity \_pscore if \_treat==0)(kdensity \_pscore if \_treat==1)

graph export 匹配前.png, as(png) replace

\* 匹配后

tw (kdensity \_pscore if \_treat==0 & \_wei!=.) (kdensity \_pscore if \_treat==1 & \_wei!=.)

graph export 匹配后.png, as(png) replace

\* 匹配后基本回归结果

xi:reg 被解释变量 解释变量 控制变量 i.indcd i.year if \_weight!=.

est store a2

esttab a2 using PSM结果1即近邻匹配.rtf,replace b(%6.4f) t(%6.4f) ar2 nogaps star(\* 0.1 \*\* 0.05 \*\*\* 0.01)

### PSM2- 核匹配

\*首先，调用数据

use 主回归数据.dta,clear

\*生成解释变量的虚拟变量形式，如下

bys year Industry:egen PLD\_RATE1\_meidan=median(PLD\_RATE1)

gen ifPLD\_RATE1 = (PLD\_RATE1>= PLD\_RATE1\_meidan) if !missing(PLD\_RATE1)

psmatch2 ifPLD\_RATE1（解释变量的虚拟变量形式） 匹配变量 ,outcome(被解释变量) logit kernel common ate ties

\* 匹配后基本回归结果

xi:reg 被解释变量 解释变量 控制变量 i.indcd i.year if \_weight!=.

est store a2

esttab a2 using PSM结果2即核匹配.rtf,replace b(%6.4f) t(%6.4f) ar2 nogaps star(\* 0.1 \*\* 0.05 \*\*\* 0.01)

### PSM匹配3-半径匹配

\*首先，调用数据

use 主回归数据.dta,clear

\*生成解释变量的虚拟变量形式，如下

bys year Industry:egen PLD\_RATE1\_meidan=median(PLD\_RATE1)

gen ifPLD\_RATE1 = (PLD\_RATE1>= PLD\_RATE1\_meidan) if !missing(PLD\_RATE1)

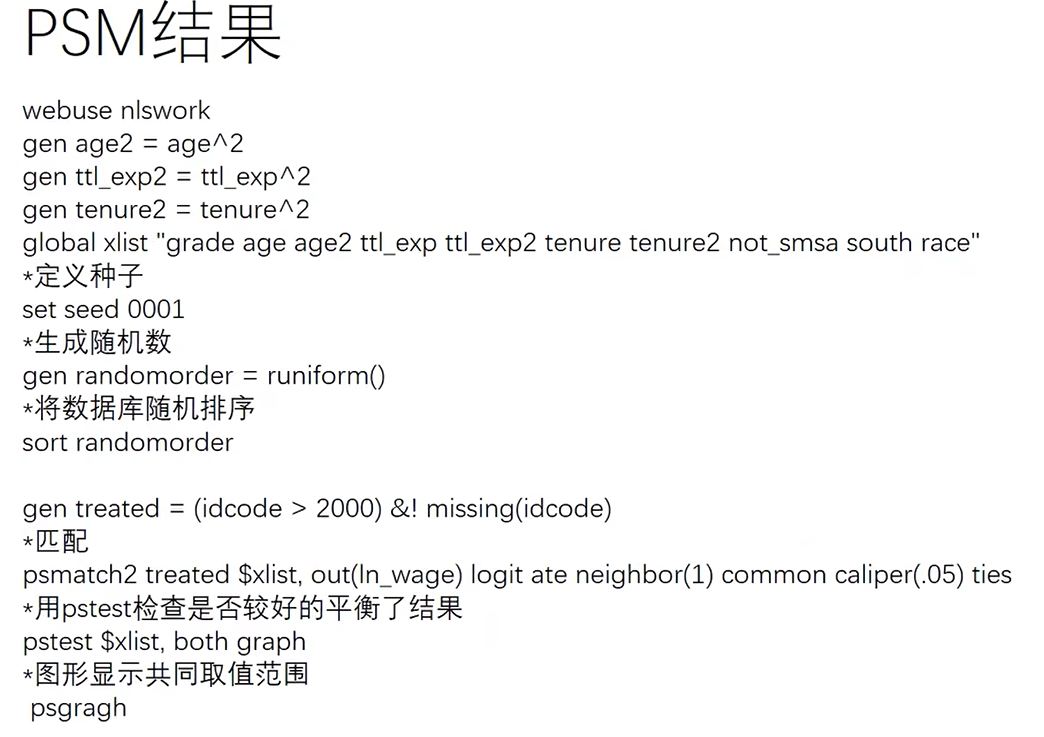
psmatch2 ifPLD\_RATE1（解释变量的虚拟变量形式） 匹配变量 ,outcome(被解释变量) logit common radius caliper(.01) ate ties

\* 匹配后基本回归结果

xi:reg 被解释变量 解释变量 控制变量 i.indcd i.year if \_weight!=.

est store a2

esttab a2 using PSM结果3即半径匹配.rtf,replace b(%6.4f) t(%6.4f) ar2 nogaps star(\* 0.1 \*\* 0.05 \*\*\* 0.01)



Pstest 值中的“bias”最好大多小于10 t-test最好不显著

## 2.滞后期模型

\*被解释变量滞后期

tsset stkcd year

xi:reg F.被解释变量 解释变量 控制变量1 控制变量2 i.year i.industry

est store a1

esttab a1 using 滞后期回归结果1.rtf,replace nogap ar2 b(%6.4f) t(%6.4f) star(\* 0.1 \*\* 0.05 \*\*\* 0.01)

\*解释变量滞后期

tsset stkcd year

xi:reg 被解释变量 L.解释变量 控制变量1 控制变量2 i.year i.industry

est store a1

esttab a1 using 滞后期回归结果2.rtf,replace nogap ar2 b(%6.4f) t(%6.4f) star(\* 0.1 \*\* 0.05 \*\*\* 0.01)

## 3.工具变量：两阶段最小二乘

ivregress 2sls y w (x1 x2 c.x1#c.x2 = z1 z2 c.z1#c.z2 )

// 其中，y是被解释变量，w是控制变量，x1 x2是内生变量，z1 z2是工具变量

## 4.Heckman两阶段

Heckman 被解释变量 控制变量, select (D(解释变量虚拟变量) =Z(工具变量其他影响因素) X(控制变量)) twostep

# 七、调节效应检验：机制分析

reg y x x1 x2 x3 //(1): y为被解释变量，x为解释变量，x1 x2 x3为控制变量

reg y x m x1 x2 x3 //(2): m为调节变量

reg y x m x\*m x1 x2 x3 //(3): (1)中的x要显著，(3)中的x和m的乘积项要显著，则存在调节效应

# 八.中介效应检验：路径分析（三种方法）

## 1.回归模型三阶段法

reg y x x1 x2 x3 //(1):y为被解释变量，x为解释变量，x1 x2 x3为控制变量

reg Z x x1 x2 x3 //(2):Z为中介变量

reg y x Z x1 x2 x3 //(3):(1)中的x要显著 (2)中的x要显著 (3)中的x和Z要显著，则存在中介效应

## 2.Sobel-Goodman检验

sgmediation y, mv(x) iv(Z) cv(x1 x2 x3) // 其中，y为被解释变量，x为解释变量，x1 x2 x3为控制变量，Z为中介变量，Indirect effect:中介效应，Direct effect:直接效应

## 3.Bootstrap检验

bootstrap r(ind\_eff) r(dir\_eff), reps(1000): sgmediation y, mv(x) iv(Z) cv(x1 x2 x3) //y为被解释变量，x为解释变量，x1 x2 x3为控制变量，Z为中介变量，计算间接效应ind\_eff（\_bs\_1）、直接效应dir\_eff（\_bs\_2）

estat bootstrap, percentile bc

//计算间接效应（\_bs\_1）的置信区间（检验中介效应：若该置信区间不包括0，则拒绝H0）；若直接效应（\_bs\_2）的置信区间不包括0就表明是【部分中介效应】；若直接效应（\_bs\_2）的置信区间包括0就表明是【完全中介效应】。

di "中介（间接）效应 ab = "\_b[\_bs\_1] //此为中介效应

di "直接效应 c' = "\_b[\_bs\_2] //此为直接效应

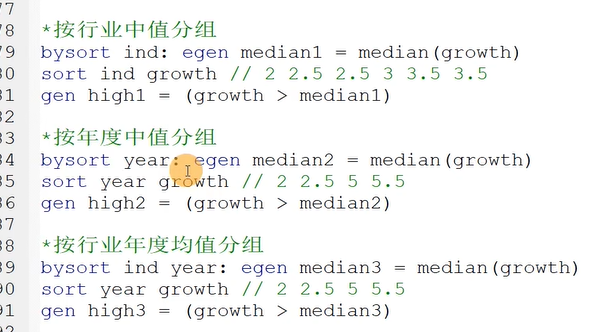
di "中介效应占比ab/c = "\_b[\_bs\_1]/(\_b[\_bs\_1]+\_b[\_bs\_2]) //此为中介效应的占比：\_bs\_1/(\_bs\_1+\_bs\_2)

# 九DID

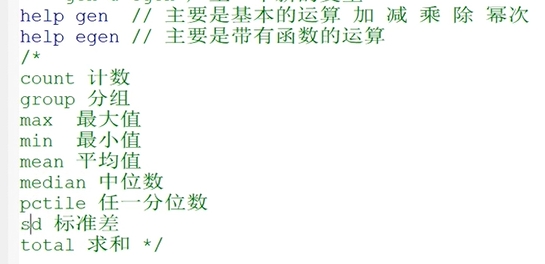
连玉君：https://mp.weixin.qq.com/s/v0gRrBktuqb4F5uIX1z8sA

# 十．其他命令

## \*中值分组



## \* gen & egen



## 字符转化为数值的stata命令

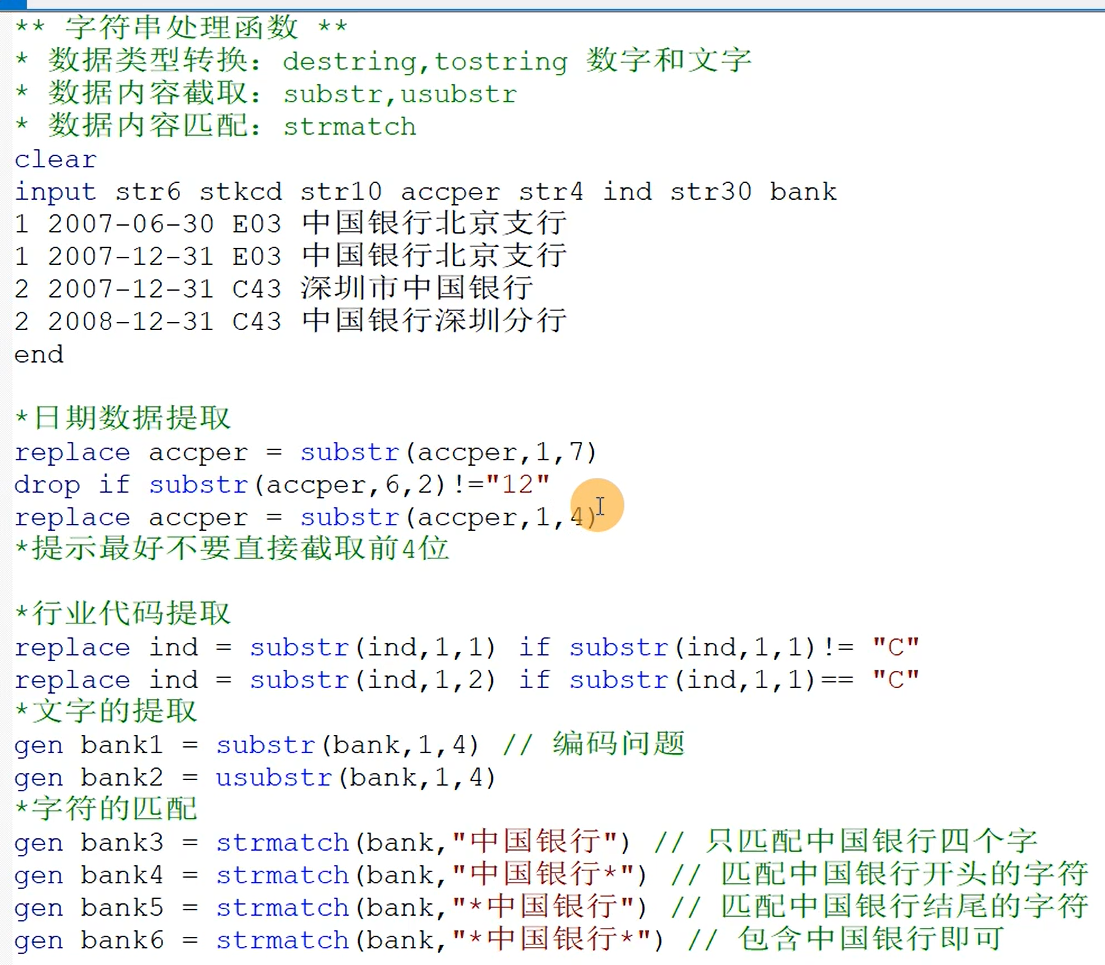
\*destring

\*例destring stkcd , replace ignore("%,.!")

## Split分割函数

\*split var,p(“”) destring

## 字符串处理函数

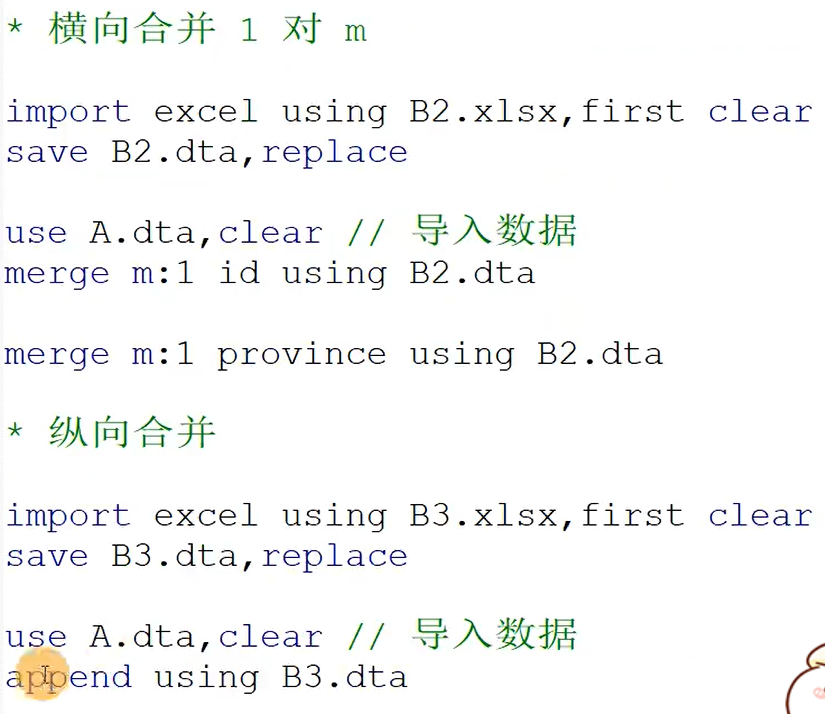


## 合并数据

导入表格保存为 ,dta格式

确定主表





文字型日期转化为“1sep2001”格式日期

gen double eventtime= clock(会计期间,"YMD")

gen eventdate=dofc(eventtime)

format eventdate %td

