实证 Stata 代码命令汇总

作者：马克数据网
官网：www.macrodatas.cn
更新时间：2023.12

目录

一、数据导入和管理

```
* 清除内存中的所有现有数据
clear

* 设置工作路径（根据你的文件位置进行调整）
cd "C:\马克数据网\实证代码命令大全 202311 版'"
```

（1）数据导入

```
* 从 Excel 文件导入数据
import excel "example.xlsx", firstrow

* 从 CSV 文件导入数据
import delimited "example.csv", delimiter(",")

* 从 Stata 文件（.dta 格式）导入数据
use "example.dta", clear

* 检查导入的数据
describe
list in 1/5
```

（2）数据导出

```
* 导出数据到 Excel 文件
export excel using "exported_data.xlsx", firstrow(variables)

* 导出数据到 CSV 文件
export delimited using "exported_data.csv", delimiter(",")

* 保存为 Stata 格式的数据文件
save "exported_data.dta", replace
```

二、数据的处理

（1）生成新变量

```
gen new_var = var1 * var2
gen new_var = ln(var)
```

（2）格式转换

```
* 将字符串日期转换为 Stata 日期
gen date_var = date(date_string, "DMY")
* 年份生成
gen year=real(substr("统计日期",1,4))
* 字符转为数字格式
destring year,replace
```

（3）缺失数据

```
* 如果变量 var1 和 var2 的任何行存在缺失值，则删除该行
drop if missing(var1) | missing(var2)
* 或者通过循环删除变量缺失的数据
foreach i in  变量 1 变量 2 变量 3 {
    drop if `i'==.
}
```

（4）异常数据

```
* 将 var2 中不合理的负值设为 0
replace var2 = 0 if var2 < 0
* 缩尾处理
winsor2 last_income, replace cuts(0 99) //缩尾代替
winsor2 last_income, replace cuts(0 99) trim //缩尾删除
```

（5）重命名变量

```
rename var3 new_var3
```

（6）编码分类变量

```
* 将字符串变量 gender 转换为数字
encode gender, gen(gender_code)
* 生成行业虚拟变量，为了避免共线性，删掉 indu1
tab Industry,gen(indu)
drop indu1
tab year,gen(time)
drop time1
```

（7）设定面板数据

```
* 假设 id 和 year 是面板数据的两个维度
xtset id year
```

（8）数据合并

```
* 根据 id、year 合并另一个数据集“raw_data.data”
merge 1:1 id year using raw_data
```

（9）数据追加

```
* 追加另一个数据集“extra_data.dta”
append using extra_data
```

三、描述性统计

（1）基本统计

```
summarize //或者 sum
```

（2）变量的详细统计

```
summarize income, detail
```

（3）变量的频率表

```
tabulate gender
```

（4）变量间的相关性

```
correlate income education
```

（5）回归分析及其描述性统计

```
regress income education age
estat summarize
```

（6）简单统计

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

四、相关性分析

（1）绘制直方图

```
histogram income
```

（2）绘制散点图

```
scatter income education
```

（3）矩阵散点图

```
graph matrix var1 var2 var3
```

（4）相关图

```
pwcorr var1 var2 var3, sig star(0.05) matrix(corr_matrix)
matrix plot corr_matrix
```

（5）回归拟合图

```
twoway (scatter var1 var2) (lfit var1 var2)
```

（6）相关系数

法 1：pwcorr varname
pwcorr varname,sig //看相关性是否显著

法 2：pwcorr_a varname

其中 pwcorr 是命令，varname 是分析变量，pwcorr 命令需要下载。
使用 pwcorr_a 可以输出带*的相关性系数，*表示显著性水平，需要安装

（7）相关系数矩阵

```
graph matrix price wei len
```

五、实证模型

（1）单变量分析

用途：分析单个变量的基本统计特性，如均值、中位数、方差等，以获取对数据的基本理解。
例子：在研究家庭收入时，单变量分析可以用来计算整个样本的平均家庭收入。

```
* 安装 ttable3 命令：解释变量虚拟形式为 01 虚拟变量形式
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
```

（2）OLS 回归

用途：分析一个或多个自变量如何线性影响因变量。
例子：研究教育水平（自变量）如何影响个人收入（因变量）

```
reg y x x1 x2 x3

*导出结果
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 reg1
esttab  reg1  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)
```

（3）分位数回归

用途：分析自变量对因变量不同分位数的影响，提供比 OLS 更全面的视角。
例子：研究培训项目（自变量）对员工工资（因变量）在不同工资水平（如中位数、上四分位数）的影响。

```
*分位数为 0.1 0.25 0.5 0.75  0.9，可根据研究问题自行调整
sqreg y x1 x2 x3 ,q( .1 .25 .5 .75 .9)
```

（4）Probit 模型

用途：分析自变量如何影响二元因变量的概率（发生与不发生），基于正态分布。
例子：分析个人的某些特征（如年龄、教育水平）如何影响其是否选择退休（二元因变量：退休/不退休）

```
* y 为虚拟变量 01
probit y x x1 x2 x3
```

（5）Logit 模型

用途：与 Probit 模型类似，但基于逻辑分布，用于分析自变量对二元因变量的影响。
例子：研究信用评分（自变量）如何影响个人获得贷款的概率（二元因变量：批准/未批准）。

```
* y 为虚拟变量 01
logit y x x1 x2 x3
```

（6）Tobit 模型

用途：处理因变量受限的情况（如有下限或上限），常见于存在截断或下限数据。
例子：研究广告支出（自变量）如何影响产品销售（因变量），当部分产品销售为零（即数据被截断于零）时。

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

六、内生性解决

（1）工具变量法

用途：用于解决内生性问题，即当解释变量和误差项相关时。工具变量是与因变量无关但与内生解释变量相关的变量。
例子：研究教育对收入的影响时，教育年数可能与个人能力相关（内生性）。使用某些与个人能力无关但影响教育年数的变量（如地区教育政策）作为工具变量。

```
*两阶段最小二乘法：y 是被解释变量，x1  x2 是内生变量，z1  z2 是工具变量，w 是控制变量
ivregress 2sls y w (x1 x2 c.x1#c.x2 = z1 z2 c.z1#c.z2 )
```

（2）固定效应模型

用途：用于控制不随时间变化但可能影响因变量的未观测变量。
例子：分析公司政策对员工生产力的影响，固定效应模型可以控制每个公司的特定特征（如公司文化）。

```
* 设为面板数据
xtset id year
* 固定效应模型
xtreg y x x1 x2 x3,fe
```

```
est store reg1
```

（3）随机效应模型

用途：当个体效应（如个体、公司）被认为是随机且与其他解释变量无关时使用。
例子：在分析多个国家的经济增长数据时，每个国家的特定效应可能被视为随机。

```
* 随机效应模型
xtreg y x x1 x2 x3,re
est store reg2

* 导出回归结果
xtreg 被解释变量 var1 var2 var3  i.year i.industry,fe
est store reg2
esttab reg1 reg2 using 回归结果.rtf,replace b(%6.4f) t(%6.4f) nogap
ar2 star(* 0.1 ** 0.05 *** 0.01)
```

（4）系统 GMM 模型

用途：用于处理动态面板数据模型中的同时方程偏差和未观测变量偏差。
例子：研究企业投资行为对其未来收益的影响时，系统 GMM 可以有效控制内生性问题。

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

（5）DID 模型

用途：评估某项政策或事件对处理组和对照组之间影响的差异。
例子：评估“宽带中国”政策对受影响城市（处理组）和未受影响城市（对照组）创新水平的影响。

```
*post 为实验组，若是则取值为 1，否则为 0；after 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.post#c.after 为交乘项，根据 ID
进行聚类
xtreg 被解释变量 c.post#c.after 控制变量,fe cluster(id)
est store m1
esttab  m1 using DID 模型.rtf,replace  ar2 b(%6.4f) t(%6.4f) star(*
0.1 ** 0.05 *** 0.01)
```

（6）PSM 模型

用途：用于观测数据中的因果推断，通过匹配处理组和对照组来减少选择偏差。

例子：研究培训计划对员工晋升的影响时，PSM 可以用于匹配参加培训和未参加培训的员工。

```
*导入基础数据
use macrodata_basic.dta,clear

*生成解释变量的虚拟变量形式
bys year Industry:egen PLD_RATE1_meidan=median(PLD_RATE1)
gen ifPLD_RATE1 = (PLD_RATE1>= PLD_RATE1_meidan) if !missing(PLD_RATE1)


********** PSM——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

* 画核密度图
* 匹配之前
tw (kdensity _pscore if  _treat==0)(kdensity _pscore if _treat==1)
graph export 匹配前.png, as(png) replace

* 匹配之后
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
reg 被解释变量 解释变量 控制变量  i.indcd i.year if _weight!=.
est store m1
esttab  m1  using  PSM-近邻匹配结果.rtf,replace  b(%6.4f)  t(%6.4f)  ar2 nogaps star(* 0.1 ** 0.05 *** 0.01)


********** PSM——核匹配 **********
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
reg 被解释变量 解释变量 控制变量 i.indcd i.year if _weight!=.
est store m2
esttab  m2  using  PSM-核匹配结果.rtf,replace  b(%6.4f)  t(%6.4f)  ar2 nogaps star(* 0.1 ** 0.05 *** 0.01)
```

********** PSM——半径匹配 **********
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
reg 被解释变量 解释变量 控制变量 i.indcd i.year if _weight!=.
est store m3
esttab m3 using PSM-半径匹配结果.rtf,replace b(%6.4f) t(%6.4f) ar2 nogaps star(* 0.1 ** 0.05 *** 0.01)

（7）滞后期模型

用途：用于探究先前时期的变量值（滞后期变量）对当前时期因变量的影响。
例子：分析上一年度的研发投入对本年度专利产出的影响。

*被解释变量滞后 1 期：使用 F.
xtset stkcd year
xtreg 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 m1
esttab m1 using 滞后期回归 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：使用 L.
xtsset stkcd year
xtreg 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 m2
esttab m2 using 滞后期回归 2.rtf,replace nogap ar2 b(%6.4f) t(%6.4f) star(* 0.1 ** 0.05 *** 0.01)

七、检验分析

（1）豪斯曼检验

* 豪斯曼检验：面板数据回归中，用于选择固定效应模型还是随机效应模型。一般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

```
xtreg y x x1 x2 x3,re
est store  re
xtreg y x x1 x2 x3,fe
est store  fe
hausman fe re
```

（2）Heckman 两阶段检验

*用途：Heckman 两阶段方法主要用于解决样本选择偏差问题。

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

（3）调节效应检验

*用途：节效应检验用于评估一个或多个变量（调节变量）如何影响其他变量之间关系的强度和方向

* y 为被解释变量，x 为解释变量，x1 x2 x3 为控制变量； m 为调节变量

```
reg y x x1 x2 x3     //回归 1
reg y x m x1 x2 x3   //回归 2
```

*若回归 1 中的 x 显著，回归 3 中的 x 和 m 的乘积项显著，则存在调节效应

```
reg y x m x*m x1 x2 x3  //回归 3
```

（4）中介效应检验

*用途：介效应检验用于评估某个变量（中介变量）在因变量和自变量之间的作用机制。

*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
```

*法 2：Sobel-Goodman 检验

```
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
```

*检验结果中：Indirect effect:中介效应，Direct effect:直接效应

*法 3：Bootstrap 检验
bootstrap r(ind_eff) r(dir_eff), reps(1000): sgmediation y, mv(x) iv(Z) cv(x1 x2 x3)
*检验结果：间接效应 ind_eff（_bs_1）、直接效应 dir_eff（_bs_2），中介效应占比=_bs_1/(_bs_1+_bs_2)

estat bootstrap, percentile bc
*结果分析：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”。

八、结果导出

（1）导出描述性统计

*输入论文的代码写法：该代码直接输出论文格式，简单调整即可
sum2docx 变量1 变量2 变量3 using 描述性统计.docx,replace stats(N mean(%9.4f) sd min(%9.2f) median(%9.2f) max(%9.2f)) title("Table 2: Summary Statistics")

*outreg2 导出
outreg2 using xxx.doc, replace sum(log) title(Decriptive statistics)
// xxx.doc 为输出文件名;sum(log) 即输出一般统计指标命令，一般统计指标包括样本数、中值、标准误、最大值和最小值

outreg2 using xxx.doc, replace sum(detail) title(Decriptive statistics)

（2）导出相关系数

* 导出相关系数，需要先下载 pwcorr_a 命令包
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)

（3）导出回归结果

展示回归结果

```
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)
```

线性回归结果

```
sysuse auto, clear
reg price mpg
outreg2 using xxx.doc, replace tstat bdec(3) tdec(2) ctitle(y)
```

/* ctitle 为自定义表格内标题命令，如果不进行设定则直接输出为被解释变量名；按照 outreg2 命令输出的表格内相关系数下括号内数字为标准误，因此我们一般利用命令 tstat 将其更改为 t 值；相关系数 bdec(3)保留 3 位有效数字；t 值 tdec(2)，保留 2 位有效数字 */

面板数据的回归结果

```
webuse grunfeld,clear
xtset company year
xtreg invest mvalue kstock,fe robust
outreg2 using xxx.doc, replace tstat bdec(3) tdec(2) ctitle(y) keep(invest mvalue kstock) addtext(Company FE, YES ) // addtext 为在表中增加信息命令，由于 Stata 进行固定效应回归后单纯利用 outreg2 命令输出不会展示是否控制固定效应，因此我们需要利用 addtext 命令追加。
```

工具变量法的回归结果

```
sysuse auto
ivregress2 2sls mpg weight (length=displacement),first
est restore first
outreg2 using xxx.doc, cttop(first)  tstat bdec(3) tdec(2) replace

ivregress2 2sls mpg weight (length=displacement), first

outreg2 using xxx.doc, cttop(two) tstat bdec(3) tdec(2)
```

按照 outreg2 命令输出的表格内相关系数下括号内数字为标准误，利用命令 tstat 将其更改为 t 值。
outreg2 命令输出时默认相关系数和 t 值都保留 3 位有效数字，而一般期刊要求相关系数保留 3 位有效数字，t 值保留 2 位有效数字，因此利用 bdec(3)和 tdec(2)命令限定。