定量政治分析方法 _hw002

吴温泉

目录

1 载入数据 2sls.dta

```
library(foreign)
library(stargazer)
library(parallel)
library(systemfit)
options(scipen = 200)
data <- read.dta('./2sls.dta')
attach(data)
```

2 使用 OLS 分析 y 和 x1, x2, x3 的关系。x3 的回归系数为何？

```
OLS <- lm(y ~ x1 + x2 + x3, data = data)
summary(OLS)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73820 -0.30346  0.00444  0.30536  1.57935
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -0.98447    0.21283  -4.626          0.00000423 ***
## x1           0.71478    0.05923  12.068 < 0.0000000000000002 ***
## x2          -2.14751    0.08569 -25.062 < 0.0000000000000002 ***
## x3           0.75718    0.03078  24.601 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

3 如果 X3 与 Y 存在内生性关系(互为因果),请使用工具变量 Z1, Z2,应用 2SLS 重新分析 Y 和 X1, X2, X

```
## Residual standard error: 0.482 on 996 degrees of freedom
## Multiple R-squared:  0.8815, Adjusted R-squared:  0.8812
## F-statistic:  2470 on 3 and 996 DF,  p-value: < 0.00000000000000022
```

x_3 的回归系数为 0.757。

3 如果 x3 与 y 存在内生性关系（互为因果），请使用工具变量 z1, z2，应用 2SLS 重新分析 y 和 x1, x2, x3 的关系。x3 的回归系数为何？

```
eq1 <- y ~ x1 + x2 + x3
eq2 <- x3 ~ x1 + x2 + y
system <- list(eq1, eq2)
inst <- ~ x1 + x2 + z1 + z2
TSLS <- systemfit(system, method="2SLS", inst=inst)
summary(TSLS)
```

```
##
## systemfit results
## method: 2SLS
##
##            N   DF     SSR  detRCov   OLS-R2 McElroy-R2
## system 2000 1992 980.135 0.000005 0.999813          1
##
##        N  DF     SSR      MSE     RMSE       R2   Adj R2
## eq1 1000 996 825.697 0.829013 0.910501 0.577207 0.575934
## eq2 1000 996 154.438 0.155058 0.393775 0.999971 0.999970
##
## The covariance matrix of the residuals
##           eq1       eq2
## eq1  0.829013 -0.358525
```

3 如果 *X3* 与 *Y* 存在内生性关系(互为因果),请使用工具变量 *Z1, Z2*,应用 *2SLS* 重新分析 *Y* 和 *X1, X2, X*

```
## eq2 -0.358525  0.155058
##
## The correlations of the residuals
##          eq1      eq2
## eq1  1.00000 -0.99998
## eq2 -0.99998  1.00000
##
##
## 2SLS estimates for 'eq1' (equation 1)
## Model Formula: y ~ x1 + x2 + x3
## Instruments: ~x1 + x2 + z1 + z2
##
##              Estimate Std. Error  t value              Pr(>|t|)
## (Intercept)  8.906199   0.756456  11.7736 < 0.000000000000000222 ***
## x1           3.616310   0.218756  16.5312 < 0.000000000000000222 ***
## x2          -6.481324   0.324087 -19.9987 < 0.000000000000000222 ***
## x3           2.313832   0.116407  19.8770 < 0.000000000000000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.910501 on 996 degrees of freedom
## Number of observations: 1000 Degrees of Freedom: 996
## SSR: 825.696825 MSE: 0.829013 Root MSE: 0.910501
## Multiple R-Squared: 0.577207 Adjusted R-Squared: 0.575934
##
##
## 2SLS estimates for 'eq2' (equation 2)
## Model Formula: x3 ~ x1 + x2 + y
## Instruments: ~x1 + x2 + z1 + z2
##
##                 Estimate   Std. Error   t value              Pr(>|t|)
## (Intercept) -3.873070926  0.143027013  -27.0793 < 0.000000000000000222 ***
## x1          -1.565789074  0.019298200  -81.1365 < 0.000000000000000222 ***
```

```
## x2           2.800958009  0.000979941 2858.2919 < 0.000000000000000222 ***
## y            0.428049439  0.021653538   19.7681 < 0.000000000000000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.393775 on 996 degrees of freedom
## Number of observations: 1000 Degrees of Freedom: 996
## SSR: 154.438255 MSE: 0.155058 Root MSE: 0.393775
## Multiple R-Squared: 0.999971 Adjusted R-Squared: 0.99997
```

x_3 的回归系数为 2.313832。

4 检验 z1, z2 之于 x3 是否为弱工具变量。

```
M1 <- lm(x3 ~ x1 + x2 + z2)
x3hat <- predict(M1)
TM1 <- lm(y ~ x1 + x2 + x3 + x3hat)
summary(TM1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x3hat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98772 -0.18483  0.01239  0.19969  0.97181
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 10.00976    0.31885   31.39 <0.0000000000000002 ***
## x1           3.94005    0.09269   42.51 <0.0000000000000002 ***
## x2          -6.96487    0.13773  -50.57 <0.0000000000000002 ***
```

```
## x3           0.43361    0.02139   20.27 <0.0000000000000002 ***
## x3hat        2.05391    0.05390   38.11 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3075 on 995 degrees of freedom
## Multiple R-squared:  0.9518, Adjusted R-squared:  0.9516
## F-statistic:  4915 on 4 and 995 DF,  p-value: < 0.00000000000000022
```

F>10, 因此可以拒绝 z_2 之于 x_3 为弱工具变量的原假设。

5 使用 2SLS 的余数检验 z1, z2 之于 y 的关系是否为外生。

```
n <- nrow(data)
res <- resid(TSLS)$eq1
TM2 <- lm(res ~ x1 + x2 + z1 + z2)
summary(TM2)
```

```
##
## Call:
## lm(formula = res ~ x1 + x2 + z1 + z2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4655 -0.6051  0.0060  0.5863  2.4379
##
## Coefficients:
##               Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  0.24418976  0.20549914   1.188    0.235
## x1           0.00289290  0.02783732   0.104    0.917
## x2           0.00005823  0.00110684   0.053    0.958
```

```
## z1          -0.04343556  0.02794052  -1.555    0.120
## z2          -0.00650060  0.00554155  -1.173    0.241
##
## Residual standard error: 0.9092 on 995 degrees of freedom
## Multiple R-squared:  0.003831,   Adjusted R-squared:  -0.0001736
## F-statistic: 0.9567 on 4 and 995 DF,  p-value: 0.4305
```

```
R2 <- summary(TM2)$r.squared
cQuant <- n * R2
pchisq(cQuant, 3, lower.tail=FALSE)
```

```
## [1] 0.2802862
```

因为 pchisq 为 0.2802862 不趋近于 0，因此无法拒绝原假设。z_1,z_2 之于 y 的关系为外生。

6 使用 Hausmen test 检验 x3 和 y 的关系是否为内生。

```
B1 <- coef(TSLS)[1:4]
B2 <- coef(OLS)[1:4]
VB1 <- vcov(TSLS)[1:4,1:4]
VB2 <- vcov(OLS)[1:4,1:4]
hQuant <- t(B1 - B2) %*% solve(VB1-VB2) %*% (B1 - B2)
pchisq(hQuant, 4, lower.tail=FALSE)
```

```
##                                                          [,1]
## [1,] 0.00000000000000000000000000000000000001729733
```

因为 pchisq 趋近于 0，因此拒绝原假设，IV 回归与原来的 OLS 回归显著不同。

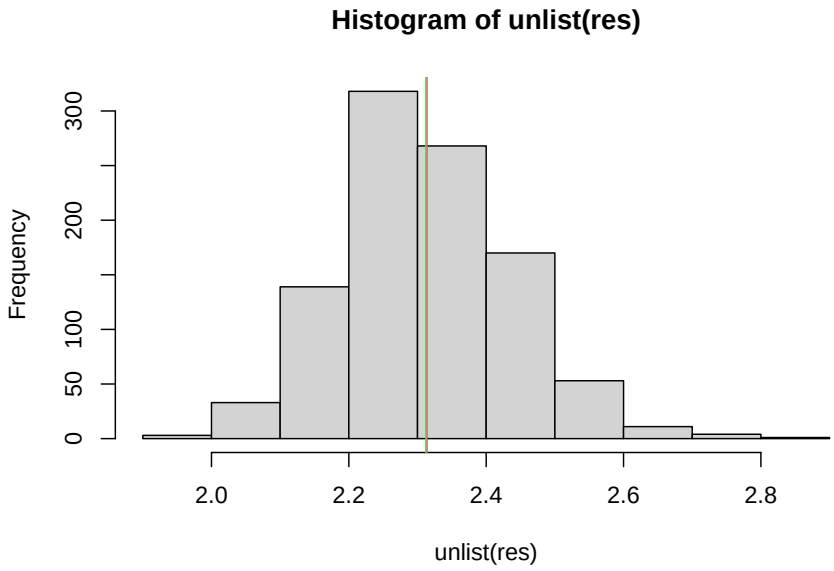
7 使用重复抽样（bootstrap 1000 次）的方法获取 x3 系数的标准误，并比较该标准误与 2SLS 的差异，说明为什么会有差异？重复抽样 5000 次和 10000 次是否会缩小差异？如果不能，是说明为什么？

```
bootFUN <- function(){
    n <- nrow(data)
    idx <- sample(1:n, n, replace=TRUE)
    tmp <- as.data.frame(data[idx, ])
    M1 <- lm(x3 ~ x1 + x2 + z1 + z2, data=tmp)
    x3hat <- predict(M1)
    M2 <- lm(y~x1+x2+x3hat, data=tmp)
    return(coef(M2)["x3hat"])
}

# 1000 times
foo <- function() replicate(125, bootFUN())

cl <- makeCluster(parallel::detectCores())
clusterExport(cl = cl, c("data", "bootFUN")) # export object to each thread
tryCatch(res <- clusterCall(cl=cl, fun = foo), finally = stopCluster(cl))
sd(unlist(res))
```

```
## [1] 0.125486
```

```
hist(unlist(res))
abline(v = 2.313832, col = 2)
abline(v = mean(unlist(res)), col = 3)
```

7 使用重复抽样(BOOTSTRAP 1000 次)的方法获取 X3 系数的标准误,并比较该标准误与 2SLS 的差异,

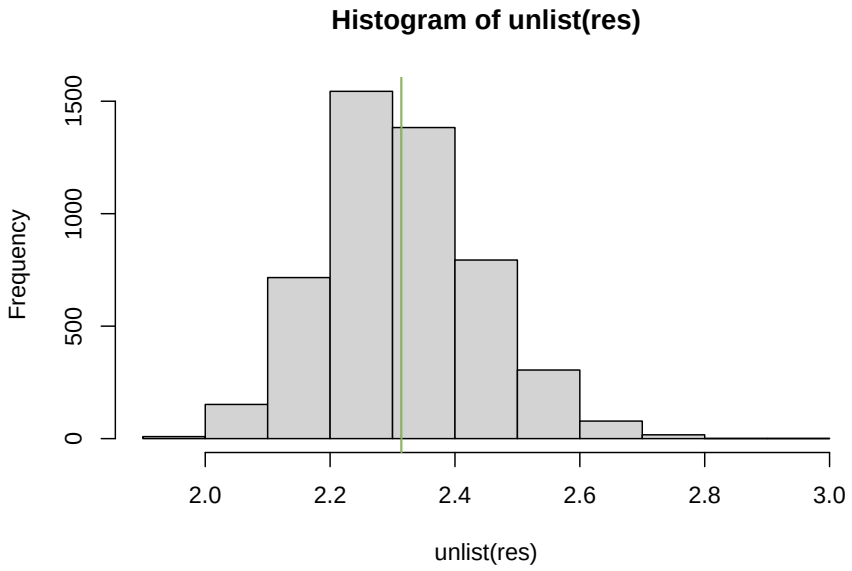



```
# 5000 times
foo <- function() replicate(625, bootFUN())

cl <- makeCluster(parallel::detectCores())
clusterExport(cl = cl, c("data", "bootFUN")) # export object to each thread
tryCatch(res <- clusterCall(cl=cl, fun = foo), finally = stopCluster(cl))
sd(unlist(res))
```

```
## [1] 0.125956
```

```
hist(unlist(res))
abline(v = 2.313832, col = 2)
abline(v = mean(unlist(res)), col = 3)
```

7 使用重复抽样(BOOTSTRAP 1000 次)的方法获取 X3 系数的标准误,并比较该标准误与 2SLS 的差异,

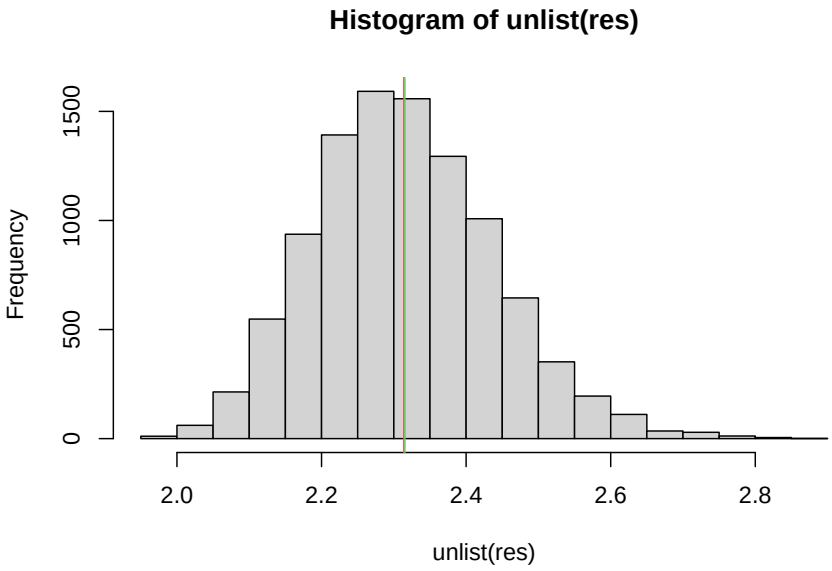



```
# 10000 times
foo <- function() replicate(1250, bootFUN())

cl <- makeCluster(parallel::detectCores())
clusterExport(cl = cl, c("data", "bootFUN")) # export object to each thread
tryCatch(res <- clusterCall(cl=cl, fun = foo), finally = stopCluster(cl))
sd(unlist(res))
```

```
## [1] 0.1247878
```

```
hist(unlist(res))
abline(v = 2.313832, col = 2)
abline(v = mean(unlist(res)), col = 3)
```

7 使用重复抽样(*BOOTSTRAP 1000* 次)的方法获取 *X3* 系数的标准误,并比较该标准误与 *2SLS* 的差异,



2SLS 中 x_3 的标准误为 0.116407, bootstrap 1000 次为 0.1212649, 1000 次为 0.1211651，1000 次为 0.1246845。增加重抽样的次数，能使得结果更接近正态分布，结果更加接近真实结果。