IV

工具变量回归(2SLS法)

Instrumental Variables in R

Copyright 2013 by Ani Katchova

公式里的外生变量需要全部包含在工具变量的子公式里，例如： 如果公式里包含一个外生变量ex和内生变量en（工具变量iv的对象）， 正确地公式写法是：y ~ ex + en | ex + iv 另外一种等价的写法是：y ~ ex + en | . - en + iv . - en表示除了公式左边内生变量之外的所有变量，这个写法适合外生变量较多的时候。

作者Ani喜欢先定义好不同的变量，我不太喜欢这个方式，直接定义公式系统开销低

Defining variables (Y1 dependent variable, Y2 endogenous variable)

(X1 exogenous variables, X2 instruments, X2 instruments, overidentified case)

Y1 <- cbind(logmedexpense)

Y2 <- cbind(healthinsu)

X1 <- cbind(illnesses, age, logincome)

X2 <- cbind(ssiratio)

X2alt <- cbind(ssiratio, firmlocation)

mydata <- read.csv("d:/download/iv\_health.csv") %>%  
 tbl\_df %>%  
 print

## Source: local data frame [10,089 x 29]  
##   
## medexpense healthinsu age female blackhisp income illnesses ssiratio  
## 1 595 1 74 1 0 94.546 0 0.1499  
## 2 1783 1 73 0 0 35.796 3 0.3959  
## 3 176 0 80 1 0 9.600 1 1.0000  
## 4 2437 1 70 0 0 38.715 5 0.2066  
## 5 330 0 91 0 0 8.725 3 0.5372  
## 6 922 1 82 1 0 43.601 2 0.3263  
## 7 1294 1 81 0 0 44.432 4 0.1350  
## 8 616 0 90 1 0 12.000 2 1.0000  
## 9 3708 1 73 0 0 24.706 1 0.0000  
## 10 701 0 79 1 0 13.511 3 1.0000  
## .. ... ... ... ... ... ... ... ...  
## Variables not shown: lowincome (int), firmsize (dbl), firmlocation (int),  
## educyr (int), private (int), hisp (int), marry (int), vegood (int), good  
## (int), fair (int), poor (int), poverty (int), midincome (int), msa  
## (int), priolist (int), black (int), logmedexpense (dbl), age2 (int),  
## logincome (dbl), vgh (int), fph (int)

#定义公式  
  
#OLS  
f\_ols<-logmedexpense~healthinsu+illnesses+age+logincome  
  
#IV,其中healthinsu内生，ssiratio为工具变量,注意.-的中的点  
f\_iv1<-logmedexpense~healthinsu+illnesses+age+logincome|.-logmedexpense+ssiratio  
  
#IV的等价写法  
f\_iv2<-logmedexpense~healthinsu+illnesses+age+logincome|illnesses+age+logincome+ssiratio  
  
#额外的IV变量  
f\_iv3<-logmedexpense~healthinsu+illnesses+age+logincome|.-healthinsu+ssiratio+firmlocation  
  
# 描述性统计  
summary(select(mydata,logmedexpense,healthinsu,  
 illnesses, age, logincome,ssiratio, firmlocation))

## logmedexpense healthinsu illnesses age   
## Min. : 0.00 Min. :0.000 Min. :0.00 Min. :65   
## 1st Qu.: 5.74 1st Qu.:0.000 1st Qu.:1.00 1st Qu.:70   
## Median : 6.68 Median :0.000 Median :2.00 Median :74   
## Mean : 6.48 Mean :0.382 Mean :1.86 Mean :75   
## 3rd Qu.: 7.43 3rd Qu.:1.000 3rd Qu.:3.00 3rd Qu.:80   
## Max. :10.18 Max. :1.000 Max. :9.00 Max. :91   
## logincome ssiratio firmlocation   
## Min. :-6.91 Min. :0.000 Min. :0.000   
## 1st Qu.: 2.23 1st Qu.:0.238 1st Qu.:0.000   
## Median : 2.74 Median :0.505 Median :0.000   
## Mean : 2.74 Mean :0.537 Mean :0.062   
## 3rd Qu.: 3.31 3rd Qu.:0.909 3rd Qu.:0.000   
## Max. : 5.74 Max. :9.251 Max. :1.000

# OLS 回归  
olsreg <- lm(f\_ols,data=mydata)  
summary(olsreg)

##   
## Call:  
## lm(formula = f\_ols, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.279 -0.677 0.147 0.852 3.780   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.78013 0.15089 38.31 <2e-16  
## healthinsu 0.07496 0.02601 2.88 0.004  
## illnesses 0.44065 0.00957 46.04 <2e-16  
## age -0.00259 0.00188 -1.38 0.167  
## logincome 0.01724 0.01379 1.25 0.211  
##   
## Residual standard error: 1.24 on 10084 degrees of freedom  
## Multiple R-squared: 0.175, Adjusted R-squared: 0.175   
## F-statistic: 534 on 4 and 10084 DF, p-value: <2e-16

# 2SLS 工具变量护轨  
ivreg1 <- ivreg(f\_iv1,data=mydata)  
summary(ivreg1)

##   
## Call:  
## ivreg(formula = f\_iv1, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.279 -0.677 0.147 0.852 3.780   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 5.78013 0.15089 38.31 <2e-16  
## healthinsu 0.07496 0.02601 2.88 0.004  
## illnesses 0.44065 0.00957 46.04 <2e-16  
## age -0.00259 0.00188 -1.38 0.167  
## logincome 0.01724 0.01379 1.25 0.211  
##   
## Residual standard error: 1.24 on 10084 degrees of freedom  
## Multiple R-Squared: 0.175, Adjusted R-squared: 0.175   
## Wald test: 534 on 4 and 10084 DF, p-value: <2e-16

ivreg2 <- ivreg(f\_iv2,data=mydata)  
summary(ivreg2)

##   
## Call:  
## ivreg(formula = f\_iv2, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.714 -0.747 0.129 0.891 4.090   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 6.58984 0.23468 28.08 < 2e-16  
## healthinsu -0.85220 0.19839 -4.30 1.8e-05  
## illnesses 0.44851 0.01029 43.58 < 2e-16  
## age -0.01180 0.00279 -4.23 2.4e-05  
## logincome 0.09769 0.02246 4.35 1.4e-05  
##   
## Residual standard error: 1.31 on 10084 degrees of freedom  
## Multiple R-Squared: 0.0709, Adjusted R-squared: 0.0706   
## Wald test: 477 on 4 and 10084 DF, p-value: <2e-16

# 2SLS 分步估计 (details)  
#第一第一步的回归公式  
f\_stage1<-healthinsu~illnesses+age+logincome+ssiratio  
#第一阶段OLS回归，拟合值代入Y2hat  
olsreg1 <- lm (f\_stage1,data=mydata)  
summary(olsreg1)

##   
## Call:  
## lm(formula = f\_stage1, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.682 -0.388 -0.241 0.517 2.592   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.959158 0.056878 16.86 <2e-16  
## illnesses 0.011351 0.003634 3.12 0.0018  
## age -0.008530 0.000712 -11.97 <2e-16  
## logincome 0.054425 0.005643 9.64 <2e-16  
## ssiratio -0.199754 0.014158 -14.11 <2e-16  
##   
## Residual standard error: 0.469 on 10084 degrees of freedom  
## Multiple R-squared: 0.0684, Adjusted R-squared: 0.068   
## F-statistic: 185 on 4 and 10084 DF, p-value: <2e-16

Y2hat <- fitted(olsreg1)  
  
#第二阶段回归的公式  
f\_stage2<-logmedexpense~Y2hat+illnesses+age+logincome  
  
olsreg2 <- lm(f\_stage2,data=mydata)  
summary(olsreg2)

##   
## Call:  
## lm(formula = f\_stage2, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.292 -0.668 0.152 0.851 3.688   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 6.58984 0.22102 29.82 < 2e-16  
## Y2hat -0.85220 0.18684 -4.56 5.1e-06  
## illnesses 0.44851 0.00969 46.27 < 2e-16  
## age -0.01180 0.00263 -4.49 7.1e-06  
## logincome 0.09769 0.02116 4.62 3.9e-06  
##   
## Residual standard error: 1.24 on 10084 degrees of freedom  
## Multiple R-squared: 0.176, Adjusted R-squared: 0.176   
## F-statistic: 538 on 4 and 10084 DF, p-value: <2e-16

# 2SLS 估计的过度识别  
ivreg\_o <- ivreg(f\_iv3,data=mydata)  
  
summary(ivreg\_o,diagnostics = TRUE)

##   
## Call:  
## ivreg(formula = f\_iv3, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.769 -0.766 0.118 0.907 4.178   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 6.69239 0.22871 29.26 < 2e-16  
## healthinsu -0.96962 0.18639 -5.20 2.0e-07  
## illnesses 0.44951 0.01043 43.11 < 2e-16  
## age -0.01296 0.00273 -4.75 2.0e-06  
## logincome 0.10788 0.02182 4.94 7.8e-07  
##   
## Diagnostic tests:  
## df1 df2 statistic p-value  
## Weak instruments 2 10083 116.53 < 2e-16  
## Wu-Hausman 1 10083 37.41 9.9e-10  
## Sargan 1 NA 2.38 0.12  
##   
## Residual standard error: 1.33 on 10084 degrees of freedom  
## Multiple R-Squared: 0.0429, Adjusted R-squared: 0.0426   
## Wald test: 466 on 4 and 10084 DF, p-value: <2e-16

# Hausman test for endogeneity of regressors   
#AER包新版的summary的diganostic选项包含了下面的检验，所以没什么意义了  
cf\_diff <- coef(ivreg1) - coef(olsreg)  
vc\_diff <- vcov(ivreg1) - vcov(olsreg)  
x2\_diff <- as.vector(t(cf\_diff) %\*% solve(vc\_diff) %\*% cf\_diff)  
pchisq(x2\_diff, df = 2, lower.tail = FALSE)

## [1] 1