Lab 12

Nikita Grabher-Meyer

11/12/2020

# Lab 12

## OLS regression on foot

*Setup*

setwd("/Users/nikitagrabher-meyer/Desktop/PHD/Econometrics/Labs/Lab 12, Homework")  
  
library(data.table)  
library(stargazer  
 )

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

df.OLSdat <- read.table("OLSqData.RData")  
DT <- data.table(df.OLSdat)

*Take a look at the data, which variables do you have?*

summary(df.OLSdat)

## y1 const x1 x2   
## Min. :-11.7353 Min. :1 Min. :-1.0813748 Min. :-3.626672   
## 1st Qu.: -0.2118 1st Qu.:1 1st Qu.:-0.2214267 1st Qu.:-0.686488   
## Median : 2.2171 Median :1 Median : 0.0001322 Median : 0.000400   
## Mean : 2.1698 Mean :1 Mean :-0.0038033 Mean :-0.006917   
## 3rd Qu.: 4.5833 3rd Qu.:1 3rd Qu.: 0.2136286 3rd Qu.: 0.670995   
## Max. : 17.5787 Max. :1 Max. : 1.1481061 Max. : 4.024930   
## x3 eps1 eps2   
## Min. :-3.75367 Min. :-5.86698 Min. :-53.6324   
## 1st Qu.:-0.69583 1st Qu.:-0.95432 1st Qu.: -9.4102   
## Median :-0.03440 Median : 0.03125 Median : 0.1831   
## Mean :-0.02018 Mean : 0.01829 Mean : 0.1411   
## 3rd Qu.: 0.66101 3rd Qu.: 0.97468 3rd Qu.: 9.8168   
## Max. : 3.77398 Max. : 5.35372 Max. : 59.9260

summary(DT)

## y1 const x1 x2   
## Min. :-11.7353 Min. :1 Min. :-1.0813748 Min. :-3.626672   
## 1st Qu.: -0.2118 1st Qu.:1 1st Qu.:-0.2214267 1st Qu.:-0.686488   
## Median : 2.2171 Median :1 Median : 0.0001322 Median : 0.000400   
## Mean : 2.1698 Mean :1 Mean :-0.0038033 Mean :-0.006917   
## 3rd Qu.: 4.5833 3rd Qu.:1 3rd Qu.: 0.2136286 3rd Qu.: 0.670995   
## Max. : 17.5787 Max. :1 Max. : 1.1481061 Max. : 4.024930   
## x3 eps1 eps2   
## Min. :-3.75367 Min. :-5.86698 Min. :-53.6324   
## 1st Qu.:-0.69583 1st Qu.:-0.95432 1st Qu.: -9.4102   
## Median :-0.03440 Median : 0.03125 Median : 0.1831   
## Mean :-0.02018 Mean : 0.01829 Mean : 0.1411   
## 3rd Qu.: 0.66101 3rd Qu.: 0.97468 3rd Qu.: 9.8168   
## Max. : 3.77398 Max. : 5.35372 Max. : 59.9260

*Now use cbind to build 2 matrices: a) Using DTx1, DTx3 to build matrix X and b) Using DT$y5 to build matrix y*

X <- cbind(c(DT$const), c(DT$x1), c(DT$x2), c(DT$x3))  
  
y <- cbind(c(DT$y1))

*Now compute X’ (transpose of X)*

Xt <- t(X)

*Next compute X’X . Report your result*

XtX <- Xt%\*%X

*Next compute the inverse of X’X. Report it*

invXtX <- (XtX)^-1

*Next compute X’y. Report it*

Xty <- Xt%\*%y

*Lastly multiply the inverse of X’X with X’y. Report your result. What is this?*

OLS <- invXtX%\*%Xty  
OLS

## [,1]  
## [1,] -263.43677  
## [2,] 439.83816  
## [3,] -136.03906  
## [4,] 21.26173

*Finally run a few comparisons: Regress y1 on x1, x2, and x3 in a linear model*

OLS1 <- lm(y1 ~ x1 + x2 + x3, data=DT)  
summary(OLS1)

##   
## Call:  
## lm(formula = y1 ~ x1 + x2 + x3, data = DT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.1144 -2.1371 0.0641 2.1496 10.7707   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.19198 0.03656 59.954 <2e-16 \*\*\*  
## x1 0.12240 0.16701 0.733 0.464   
## x2 1.30648 0.03599 36.301 <2e-16 \*\*\*  
## x3 0.62897 0.05367 11.720 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.183 on 7580 degrees of freedom  
## Multiple R-squared: 0.1794, Adjusted R-squared: 0.179   
## F-statistic: 552.2 on 3 and 7580 DF, p-value: < 2.2e-16

*Divide XtX by n (number of observations) and compare it to cov(DT), comment*

XtX/7584

## [,1] [,2] [,3] [,4]  
## [1,] 1.000000000 -0.003803269 -0.006916885 -0.02018028  
## [2,] -0.003803269 0.103658354 0.001331209 0.23662823  
## [3,] -0.006916885 0.001331209 1.031653365 0.01050547  
## [4,] -0.020180277 0.236628234 0.010505469 1.00428137

covDT <- cov(DT)  
covDT

## y1 const x1 x2 x3 eps1  
## y1 12.34255205 0 0.1631956969 1.354630521 0.67399025 -0.0345977698  
## const 0.00000000 0 0.0000000000 0.000000000 0.00000000 0.0000000000  
## x1 0.16319570 0 0.1036575573 0.001305074 0.23658268 -0.0009250523  
## x2 1.35463052 0 0.0013050743 1.031741563 0.01036725 0.0044524846  
## x3 0.67399025 0 0.2365826780 0.010367251 1.00400651 -0.0149200056  
## eps1 -0.03459777 0 -0.0009250523 0.004452485 -0.01492001 2.0521287710  
## eps2 0.35828556 0 0.0845584145 -0.105670109 0.22875506 -0.1687865058  
## eps2  
## y1 0.35828556  
## const 0.00000000  
## x1 0.08455841  
## x2 -0.10567011  
## x3 0.22875506  
## eps1 -0.16878651  
## eps2 199.70843926

## IV – Exercise

dt.IV <- read.table("IV\_Data.RData")  
dt.IV <- data.table(dt.IV)   
IV <- lm(y5 ~ x1 + x2, data=dt.IV)  
stargazer(IV, type = "text")

##   
## ==================================================  
## Dependent variable:   
## ------------------------------  
## y5   
## --------------------------------------------------  
## x1 2.930\*\*\*   
## (0.007)   
##   
## x2 1.680\*\*\*   
## (0.007)   
##   
## Constant 0.490\*\*\*   
## (0.007)   
##   
## --------------------------------------------------  
## Observations 17,584   
## R2 0.932   
## Adjusted R2 0.932   
## Residual Std. Error 0.965 (df = 17581)   
## F Statistic 120,701.300\*\*\* (df = 2; 17581)  
## ==================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

*Looking at the Var-Cov Matrix, what bias would you expect for x1 and what bias for x2?, calculate! (Hint: the X – matrix for calculating the bias consists of x1 and x2., i.e. you can disregard the z’s and the constant for this calculation. E(X,u) is given by the “eps” - column)* The coefficients of x1 and x2 are both positive. Also the two variable have positive correlation. The bias should be positive as well.

*Which of the X Variables is endogenous with y (e.g. a simultaneity problem)? If you can mend only one of the variables, which one would you tackle and why? You have 4 candidates that you can use as an instrumental variable for X1, but one is itself endogenous, one is a weak instrument, one is irrelevant and only one is valid. Which is which?* 1) x1 seems to be correlated with the error term, therefore it’s endogenous. 2) 2) x2 also seems to be correlated with the error term, but the correlation is quite weak (0.02) 3) z1 is also endogenous (for the same reasons as above) 4) z2 is not correlated with x1, therefore it’s irrelevant 5) z3 is correlated with x1, and the covariance seems to be the highest within the different alternatives. It might be considered the best instrument 6) z4 is correlated with x1, but it is only 0.01, therefore it might be quite weak

In summary, z1, z2 and z4 should not be used as instruments, as they violate the 2 assumptions of relevance and exclusion restriction.

*Consider using Z4 as an instrument for X2: Which of the two assumptions can you test? Is it satisfied?* We can test the relevance assumption using Pearson’s product-moment correlation test

cor.test(dt.IV$z4, dt.IV$x1)

##   
## Pearson's product-moment correlation  
##   
## data: dt.IV$z4 and dt.IV$x1  
## t = 2.2537, df = 17582, p-value = 0.02423  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.002214089 0.031766931  
## sample estimates:  
## cor   
## 0.01699422

*Pick your most preferred instrument and run the 2SLS*

first\_step <- lm(x1 ~ x2 + z3, data=dt.IV)  
summary(first\_step)

##   
## Call:  
## lm(formula = x1 ~ x2 + z3, data = dt.IV)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8181 -0.6625 0.0070 0.6684 3.6077   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.002372 0.007369 -0.322 0.747   
## x2 0.128994 0.007294 17.685 <2e-16 \*\*\*  
## z3 0.186035 0.007397 25.150 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9771 on 17581 degrees of freedom  
## Multiple R-squared: 0.05069, Adjusted R-squared: 0.05059   
## F-statistic: 469.4 on 2 and 17581 DF, p-value: < 2.2e-16

dt.IV <- dt.IV[, x1hat:=predict(first\_step, newdata=dt.IV)]  
  
dt.IV  
  
second\_step <- lm(y5 ~ x1hat + x2, data=dt.IV)  
stargazer(second\_step, IV, type="text")

##   
## =============================================================  
## Dependent variable:   
## ----------------------------  
## y5   
## (1) (2)   
## -------------------------------------------------------------  
## x1hat 2.115\*\*\*   
## (0.124)   
##   
## x1 2.930\*\*\*   
## (0.007)   
##   
## x2 1.784\*\*\* 1.680\*\*\*   
## (0.028) (0.007)   
##   
## Constant 0.488\*\*\* 0.490\*\*\*   
## (0.023) (0.007)   
##   
## -------------------------------------------------------------  
## Observations 17,584 17,584   
## R2 0.325 0.932   
## Adjusted R2 0.325 0.932   
## Residual Std. Error (df = 17581) 3.044 0.965   
## F Statistic (df = 2; 17581) 4,231.712\*\*\* 120,701.300\*\*\*  
## =============================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

*Consider the MM-IV Estimator that we saw in class. What variables will you include in the Z matrix? Write a code that directly implements the estimator in matrix notation, using either the data.table command or the matrix command.Build a Z-matrix consisting of the instruments you want to use (including all exogeneous x); build an X-matrix consisting of the X variables; build a y-matrix that just consists of the vector y*

z <- cbind(c(dt.IV$x2), c(dt.IV$z3))  
x <- cbind(c(dt.IV$x1), c(dt.IV$x2))  
y5 <- cbind(c(dt.IV$y5))  
  
solve(t(z)%\*%x)%\*%(t(z)%\*%y5)

## [,1]  
## [1,] 2.115358  
## [2,] 1.780586

library(ivreg)  
IVz1 <- ivreg(y5~x1+x2|x2 + z1, data=dt.IV)  
IVz2 <- ivreg(y5~x1+x2|x2 + z2, data=dt.IV)  
IVz3 <- ivreg(y5~x1+x2|x2 + z3, data=dt.IV)  
IVz4 <- ivreg(y5~x1+x2|x2 + z4, data=dt.IV)  
  
stargazer (dt.IV, type="text")

##   
## =================================================================  
## Statistic N Mean St. Dev. Min Pctl(25) Pctl(75) Max   
## -----------------------------------------------------------------  
## y4 17,584 0.471 3.156 -12.021 -1.654 2.585 12.833  
## y5 17,584 0.467 3.705 -14.442 -2.023 2.950 13.584  
## const 17,584 1.000 0.000 1 1 1 1   
## x1 17,584 -0.003 1.003 -3.940 -0.681 0.678 3.737   
## x2 17,584 -0.008 1.010 -3.936 -0.688 0.685 3.921   
## z1 17,584 0.003 1.007 -4.046 -0.677 0.682 3.814   
## z2 17,584 0.005 1.002 -5.136 -0.675 0.676 4.217   
## z3 17,584 0.0001 0.996 -4.464 -0.666 0.666 4.006   
## z4 17,584 -0.003 1.001 -3.911 -0.674 0.681 4.009   
## x1hat 17,584 -0.003 0.226 -0.964 -0.156 0.146 0.842   
## -----------------------------------------------------------------