Econometrics Problem Set 4

Gigi Lin

October 10, 2018

Question 1

Derivation

Derive the 2SLS estimator in matrix algebra only in terms of endogeneous regressors **X** and instruments **Z**. Show the final equivalence between 2SLS and GIVE.

Starting from the multivariate linear regression model and the first stage auxiliary regression:

We have $y_i = x_1'\beta + \epsilon_i$ and $E(z\epsilon) = 0$ by analogy principle. Then $\frac{1}{N}\sum\limits_{i=1}^{N} z_i(y_i - x_i'\beta) = 0$

With $R > K$, we assume as close to zero as possible.

$$Q_N(\beta) = [\frac{1}{N}\sum_{i=1}^{N} z_i(y_i - x_i'\beta)]'\, W_N[\frac{1}{N}\sum_{i=1}^{N} z_i(y_i - x_i'\beta)]$$

In matrix notation we have

$$Q_N(\beta) = [\frac{1}{N}Z(Y - X\beta)]'\, W_N[\frac{1}{N}Z(Y - X'\beta)]$$

and $\frac{\partial Q_N(\beta)}{\partial\beta} = -2X'ZW_NZ'y + 2X'ZW_NZX\hat{\beta_{IV}} = 0$

$\hat{\beta_{IV}} = (X'ZW_NZ'X)^{-1}X'ZW_NZ'y$

We see W_N is proportional to the inverse of the **covariance matrix** of the sample moments.

Assume that $\epsilon_i \sim iid(0, \sigma^2)$, independent of Z.

Then, the sample moments: $\frac{1}{N}\sum\limits_{i=1}^{N}[\epsilon_i Z_i]$ has asymptotic covariance matrix

$\text{Var}(\epsilon Z) = \text{E}[\epsilon^2 ZZ'] = \sigma^2\, plim\, \frac{1}{N}\sum\limits_{i=1}^{N} Z_i Z_i'$

$W_N = (\frac{1}{N}\sum\limits_{i=1} NZ_iZ_i')^{-1} = [\frac{1}{N}Z'Z]^{-1}$

$\hat{\beta_I}V = (XZ(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$

We can also obtain this estimator via Two Stage Least Squares

1. For any endogeneous x_k, regress x_k on z in the model

   $$x_{k_i} = z_i\pi_k + v_{k_i}$$

   to get:

   $$\hat{\pi_k} = (Z'Z)^{-1}Z'x_k$$

   and

   $$\hat{x_k} = Z(Z'Z)^{-1}Z'x_k$$

2. Regress y on $\hat{x_k}$ instead of x_k to get:

   $$\hat{\beta_I}V = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

and

$$\hat{X} = Z(Z'Z)^{-1}Z'X$$

$$\hat{\beta_{IV}} = [Z(Z'Z)^{-1}Z'X)'(Z(Z'Z)^{-1}Z'X)]^{-1}(Z(Z'Z)^{-1}Z'X)'y$$
$$Z(X'Z(Z'Z)Z'X)^{-1}Z'(ZZ')^{-1}ZX'Y = \hat{\beta_{IV}}$$

And we get the equivalence between 2SLS and GIVE.

Question 2

Avoiding Invalid Instruments and Coping with Weak Instruments [Michael P. Murray]

a) As described in Murray (2006), Levitt (1996) in his study of incarceration rates attempts to anticipate and test possible arguments about why his lawsuit instruments might be invalid. Give one example of these arguments. How did Levitt counter it?

Levitt tries to justify that estimates of the effect of prison populations on crime rates are not affected by the presence of simultaneity. Such as increases in crime will lead to larger prison populations, and increased incarceration will likely reduce the amount of crime.

To control for this, Levitt uses the IV approach of prison overcrowding litigation status, that is correlated with changes in the size of the prison population, but otherwise unrelated to crimes. His results show a larger impact of prison population on crime than previous estimates imply.

b) As described in Murray (2006), Levitt in his both studies on police and crime made sure to control for key explanatory variables that, if omitted, would severely bias the 2SLS estimator. Which variables were these and why?

Every OLS analysis is essential that consideration of explanatory variables is taken into consideration, as the estimatino would be biased if the omitted variables were correlated with the included explanators. Even in the case of the IV approach, the estimation would be biased if an omitted explanatory that belongs in the model is correlated with either the included nontroublesome explanators (X vars) or the instrumental variables (Z vars).

In Levitt's first police study (1996), the instrument was mayoral and gubernatorial election cycles. He was careful to include local welfare expenditures as part of his explanators, as otherwise they are plausibly correlated with election cycles and may lower crime rates. Even if the correlation between welfare expenditures and numbers of police officers were zero (so that omitting welfare expenditures as an explanator would not bias OLS), welfare expenditures' correlation with mayoral and gubernatorial election cycles could be huge, in which case omitting such a variable could seriously bias the IV estimate of the effect of police officers on crime rates.

In Levitt's second police study (2002), Levitt chose to add a string of city-specific explanatory variables to his crime rate model to reduce the change that the number of firefighters is correlated with omitted relevant variables.

In both studies, Levitt chose to use panel data to estimate fixed effects models, or first-differences models to further reduce the severity that omitted relevant variables may bias the IV approach's results.

Question 3

Model of medical expenditure

$$ldrugexp = \beta_0 + \beta_1 hi_e mpunion + \beta_2 totchr + \beta_a ge + \beta_4 female + \beta_5 blhisp + \beta_6 linc + u$$

Define:

ldrugexp: log of total out-of-pocket expenditures on prescribed med- ications; $hi_e mpunion$: indicator for whether the individual holds either employer or union sponsored health insurance; **considered as endogenous**; *totchr*: number of chronic conditions; *age*: age in years; $female$: indicator for female; *blhisp*: indicate for black or hispanic; *linc*: log of annual household income (\$'000s) *ssiratio*: ratio of an individualís social security income to the individualís income from all sources, with high values indicating a signiÖcant income constraint; *multlc*: indicator for whether the Örm is a large operator with multiple locations

Part A)

Percentage of people in the sample that have either employer or union sponsored health insurance is 37.966%.

```
## [1] 37.96555
```

Part B)

OLS regression of the model (robust standard errors):

```
##
## Call:
## lm(formula = ldrugexp ~ hi_empunion + totchr + age + female +
##     blhisp + linc, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3295 -0.6754  0.1516  0.8559  3.7343
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.861131   0.153184  38.262  < 2e-16 ***
## hi_empunion  0.073879   0.026109   2.830  0.00467 **
## totchr       0.440381   0.009573  46.002  < 2e-16 ***
## age         -0.003529   0.001886  -1.871  0.06132 .
## female       0.057806   0.025163   2.297  0.02163 *
## blhisp      -0.151307   0.033808  -4.475 7.71e-06 ***
## linc         0.010482   0.013952   0.751  0.45251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 10082 degrees of freedom
##   (302 observations deleted due to missingness)
## Multiple R-squared:  0.177,  Adjusted R-squared:  0.1765
## F-statistic: 361.3 on 6 and 10082 DF,  p-value: < 2.2e-16
```

As researched more about the robust standard errors method used by the default behavior in Stata in a call to regress, I use the HC1 robust variance-covariance matrix. Displayed is the output:

```
coeftest(ols_wage2, vcov = vcovHC(ols_wage2, "HC1"))  # robust; HC1 (Stata default)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  5.8611305  0.1571037 37.3074 < 2.2e-16 ***
## hi_empunion  0.0738788  0.0259848  2.8432  0.004476 **
## totchr       0.4403807  0.0093633 47.0327 < 2.2e-16 ***
## age         -0.0035295  0.0019370 -1.8221  0.068466 .
## female       0.0578055  0.0253651  2.2789  0.022692 *
## blhisp      -0.1513068  0.0341264 -4.4337  9.36e-06 ***
## linc         0.0104815  0.0137126  0.7644  0.444664
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part C)

The variable *ssiratio* as an IV for $hi_empunion$.

```
(inst_var1 <- ivreg(ldrugexp ~ totchr + age + female + blhisp + linc + hi_empunion|totchr + age + fem
```

```
##
## Call:
## ivreg(formula = ldrugexp ~ totchr + age + female + blhisp + linc +     hi_empunion | totchr + age
##
## Coefficients:
## (Intercept)       totchr          age       female       blhisp
##     6.78717      0.45027     -0.01322     -0.02041     -0.21742
##        linc  hi_empunion
##     0.08700     -0.89759
```

To justify using this instrument, we need to assume: *ssiratio* is uncorrelated with the error term u ("endogeneity condition"), $Cov(ssiratio, hi_empunion) \neq 0$ ("relevance condition"), $ssiratio \neq regressors$ ("exclusion restriction").

Part D)

Using the variables *ssiratio* and *multlc* together as instruments for $hi_empunion$, use the GIVE (2SLS) estimator and report the output.

First stage: regress $hi_empunion$ on *ssiratio* and *multlc*

Second stage: regress *ldrugexp* on fitted values (Y-hat)

My GIVE (2SLS) estimation approach is:

$$ldrugexp = \beta_0 + \beta_1 hi_empunion + ssiratio + multlc$$

```
##
## Call:
## ivreg(formula = ldrugexp ~ totchr + age + female + blhisp + linc +     hi_empunion | totchr + age
##
## Coefficients:
## (Intercept)       totchr          age       female       blhisp
##     6.87519      0.45121     -0.01414     -0.02784     -0.22371
##        linc  hi_empunion
##     0.09427     -0.98993
```

I approached this IV regression in two ways. My first output statistic is for the 2SLS approach, but the standard errors are not robust; I was not able to apply the *robust* approach on the 2SLS coefficients. My second output statistic is under the IV approach, with robust standard errors.

To justify using this instrument, we need to assume: $Cov(ssiratio, u) \neq 0$ and $Cov(multlc, u) \neq 0$ ("endogeneity condition"), $Cov(ssiratio, hi_empunion) \neq 0$ and $Cov(multlc, hi_empunion) \neq 0$ ("relevance condition"), $ssiratio \neq regressors$ and $multlc \neq regressors$("exclusion restriction"). There is no additional assumptions to consider in the relationship between the two instruments as it is true that the above 3 conditions hold in addition (i.e. $Cov(ssiratio, hi_empunion, u_t)$) due to the linearity in properties.

Part E)

For parts (c) and (d), we can test for weak instruments by the F-test, and regressor endogeneity via the Wu-Hausman test. Interpretation of the outcomes will come after the output.

For part c)

```
summary(inst_var1, vcov = sandwich, df = Inf, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = ldrugexp ~ totchr + age + female + blhisp + linc +
##     hi_empunion | totchr + age + female + blhisp + linc + ssiratio,
##     data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7616 -0.7529  0.1275  0.8959  4.0723
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.787170   0.268845  25.246  < 2e-16 ***
## totchr       0.450266   0.010197  44.157  < 2e-16 ***
## age         -0.013218   0.002998  -4.409 1.04e-05 ***
## female      -0.020406   0.032611  -0.626 0.531491
## blhisp      -0.217424   0.039494  -5.505 3.69e-08 ***
## linc         0.087002   0.022636   3.844 0.000121 ***
## hi_empunion -0.897591   0.221127  -4.059 4.92e-05 ***
##
## Diagnostic tests:
##                   df1   df2 statistic  p-value
## Weak instruments    1 10082     65.81 5.56e-16 ***
## Wu-Hausman          1 10081     26.45 2.75e-07 ***
## Sargan              0    NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.318 on Inf degrees of freedom
## Multiple R-Squared: 0.06395, Adjusted R-squared: 0.0634
## Wald test:  2001 on 6 DF,  p-value: < 2.2e-16
```

Interpretation is that with the null of the F-test being the instrument is weak, since the p-value is quite low (< 2e-16 < 0.01) at the 1% confidence level, there is statistical significance that the instrument is sufficiently strong.

The null of the Wu-Hausman test is that the IV is equally as consistent as OLS. In this output, it shows that the p-value (1.17e-09 < 0.01) at the 1% confidence level, is statistically significant. Therefore, null is rejected and IV is consistent, whereas OLS is not.

For part d)

```
summary(inst_var2, vcov = sandwich, df = Inf, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = ldrugexp ~ totchr + age + female + blhisp + linc +
##     hi_empunion | totchr + age + female + blhisp + linc + ssiratio +
##     multlc, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8027 -0.7657  0.1166  0.9099  4.1359
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.87519    0.25789  26.660  < 2e-16 ***
## totchr       0.45121    0.01031  43.769  < 2e-16 ***
## age         -0.01414    0.00290  -4.875 1.09e-06 ***
## female      -0.02784    0.03217  -0.865    0.387
## blhisp      -0.22371    0.03959  -5.651 1.59e-08 ***
## linc         0.09427    0.02188   4.308 1.65e-05 ***
## hi_empunion -0.98993    0.20459  -4.839 1.31e-06 ***
##
## Diagnostic tests:
##                   df1   df2 statistic  p-value
## Weak instruments    2 10081    58.658  < 2e-16 ***
## Wu-Hausman          1 10081    37.081 1.17e-09 ***
## Sargan              1    NA     1.164    0.281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.334 on Inf degrees of freedom
## Multiple R-Squared: 0.04145, Adjusted R-squared: 0.04088
## Wald test:  1955 on 6 DF,  p-value: < 2.2e-16
```

We see similar results here. The null of the F-test is that the instrument is weak, since the p-value is quite low (< 2e-16 < 0.01) at the 1% confidence level, there is statistical significance that the instrument is sufficiently strong.

The null of the Wu-Hausman test is that the IV is equally as consistent as OLS. In this output, it shows that the p-value (1.17e-09 < 0.01) at the 1% confidence level, is statistically significant. Therefore, null is rejected and IV is consistent, whereas OLS is not.

Part F)

Looking at the Sargan test for the regression in part (d), the statistic is 1.164 with a p-value of 0.281. Since the Sargan tests all exogeneous instruments as truly exogeneous and uncorrelated with the model residuals, the statistic being insignificant (even at the 5% confidence level), implies that the instruments are valid. Given here, there should be no concern of overidentification.

g) Compare and comment on the estimates of β_1 of parts (b), (c), and (d).

The various β_1 are:

Part B) 0.073879

Part C) -0.897591

Part D) -0.98993

The OLS estimate is much further off from the IV approaches in c) and d). This supports the conclusion found in E) as the OLS approach was found to be inconsistent, whereas IV is the better approach to use.

Questions - do the IV regressions need robust standard errors too? (part d and onwards) - can't use robust standard errors on regression in d) due to atomic vector error

References: 1. Robust se's in R

2. Robust in R vs. Stata

3. Check validity of IV

4. Interpretation of various tests

5.

6. Hats in Math

Old code:

```
# Q3 part b)
# #Check robust se's results:
#
# # check that "sandwich" returns HC0
# coeftest(ols_wage2, vcov = sandwich)                # robust; sandwich
# coeftest(ols_wage2, vcov = vcovHC(ols_wage2, "HC0"))    # robust; HC0
#
# # check that the default robust var-cov matrix is HC3
# coeftest(ols_wage2, vcov = vcovHC(ols_wage2))           # robust; HC3
# coeftest(ols_wage2, vcov = vcovHC(ols_wage2, "HC3"))    # robust; HC3 (default)
```

```
# Q3 part d)
# partd_df <- wage2 %>%
#   select(ssiratio, multlc, hi_empunion, ldrugexp)
#
# # test for weak instruments
# cor(partd_df)
```

```
# Q3 part e)
partc_df <- wage2 %>%
  dplyr::select(ssiratio, hi_empunion, ldrugexp)

# # test for weak instruments
# cor(partc_df)

fe <- plm(ldrugexp ~ hi_empunion + ssiratio, data = partc_df, model = "within")
re <- pggls(ldrugexp ~ hi_empunion + $ssiratio, data = partc_df, "random")
phtest(fe,re)


(data=mydata, y=dependent variable,X1:X4: explanatory variables)
#step 1 : Estimate the FE model
 fe <- plm(y~X1+X2+X3+X4,data=mydata,model="within")
summary(model.fe)
#step 2 : Estimate the RE model
 re <- pggls(y~X1+X2+X3+X4,data=mydata,model="random")
summary(model.re)
#step 3 : Run Hausman test
phtest(fe, re)



#Bold
\begin{enumerate}\bfseries
\item \beta_0
\item The second item
\end{enumerate}
```