Econometrics II TA Session # GMM[*]

Kan Pang [†]

January 21, 2021

Contents

[*] All comments welcome!

[†] E-mail: member_1363710747@yahoo.co.jp

1 Before Estimating

1.1 Data

Brief Background. The data we use today originates from Harrison, D. and Rubinfeld, D.L.(1978)[1] which investigates the methodological problems associated with the use of housing market data to measure the willingness to pay for clean air. It drew a conclusion that marginal air pollution damages are found to increase with the level of air pollution and with household income. But here we just use some variables of the original data to construct a multiple linear regression model focusing on how the attributes of communities affect housing prices(value) of Boston city. And this time, we will concentrate on the endogenous problem.

Data. We use an open data which is called as "Boston Neighboorhood Housing Prices Dataset[1]". Although there exist many variables, this time, we take 6 of them: `value`, `crime`, `industrial`, `distance`, `black` and `ptratio`. And here come the descriptions.

- `value`: a continuous variable meaning value of owner-occupied homes in $1000's.
- `crime`: a continuous variable representing per capita crime rate by town.
- `industrial`: a continuous variable showing proportion of non-retail business acres per town.
- `distance`: a continuous variable revealing weighted distances to five Boston employment centres.
- `black`: a continuous variable which is the black proportion of population, used as an instrument for `crime`.
- `ptratio`: a continuous variable defined as pupil-teacher ratio by town school district, measuring public sector benefits and used as an instrument for `crime` as well.

As same as the last time, `value` is the dependent variable. Now, let's quickly build the dataset to use today.

```r
library(AER) # Including lmtest, sandwich and ivreg(2SLS)
library(gmm) # Including gmm function
library(MASS) # Using ginv: generalized inverse operation
library(stargazer)

# Preparations

dt = read.csv(
file = "~/boston.csv",
header = TRUE, sep = ",", row.names = NULL, stringsAsFactors = FALSE)
dt = dt[complete.cases(dt),]
# complete.cases returns TURE if one row doesn't contain NA

n = nrow(dt)
head(dt[, c("value", "crime", "industrial", "distance", "black", "ptratio")])
```

And the first few rows are shown as below.

```
  value   crime    industrial distance black     ptratio
1  24.0 0.00632       2.31    4.0900 0.0000000    15.3
2  21.6 0.02731       7.07    4.9671 0.0000000    17.8
3  34.7 0.02729       7.07    4.9671 0.3238495    17.8
4  33.4 0.03237       2.18    6.0622 0.1804159    18.7
5  36.2 0.06905       2.18    6.0622 0.0000000    18.7
6  28.7 0.02985       2.18    6.0622 0.2210220    18.7
```

1.2 The idea of finding an instrumental variable

We say that when exogeneity condition is no longer satisfied(relaxization of Gauss Markov basic assumption), the ols estimator will lose consistency. To deal with such problem, IV(instrumental variable) method

[1] data source: `http://biostat.mc.vanderbilt.edu/DataSets`. This linkage seems not to work very well so I have uploaded the csv file on CLE.

is a good solution. But how do we find an instrumental variable? Here are two main criteria for defining an IV:

(i) The instrument must be correlated with the endogenous explanatory variables, conditionally on the other covariates.[2]

(ii) The instrument cannot be correlated with the error term in the explanatory equation, conditionally on the other covariates. The explanatory equation is that, taking a linear example, an equation which is given by $\underline{\mathbf{Y}} = \underline{\mathbf{X}}\theta + \underline{\mathbf{U}}$[3].

To initialize today's empirical application, we think that `crime` may be an endogenous variable and want to deal with the problems resulting from endogeneity. Following the previous 2 criteria, `black` and `ptratio` are, in a way, correlated with crime rate(the original paper is published in 1978 and I mean no offense here). However, these two variables remain a probability of not functioning very well and that's called weak instruments problem which won't be talked today. If you are interested in such stuffs, please check Nichols, Austin(2006)[6] on your own.

Then, let's prepare the explanatory variables and instrumental variables in a matrix form with R for further estimations. Here, we denote the dimensions that $dim(g) = L$ and $dim(\theta) = d$. In this case, L = d + 1 = 5(containing constant). Remember that we detected heteroskedasticity and autocorrelation in TA session GLS with this dataset. So it provides us a good example to learn about 2SGMM(2 step GMM) method.

```
# Extract varibales and instruments
name_x = c(”constant”, ”crime”, ”industrial”, ”distance”)
name_z = c(”constant”, ”black”,  ”ptratio”, ”industrial”, ”distance”) # means L = d + 1 = 5
 constant = rep(1, n)

X = cbind(constant, dt[, c(”crime”, ”industrial”, ”distance”)])
 colnames(X) = name_x

Z = cbind(constant, dt[, c(”black”, ”ptratio”, ”industrial”, ”distance”)])
 colnames(Z) = name_z

Y = dt[, ”value”]
```

[2] But a weak correlation may provide misleading inferences about parameter estimates and standard errors.[6]

[3] ”explanatory” is used because we want to know the causal effects between $\underline{\mathbf{Y}}$ and $\underline{\mathbf{X}}$

2 Numerical 2SGMM Method

2.1 Formulations and Recalling HA-Robusted Ols

Firstly, let's declare the linear model. The multiple regression model is easily written as below.

$$value_i = \beta_0 + \beta_1 crime_i + \beta_2 industrial_i + \beta_3 distance_i + u_i \quad \forall i = 1, \ldots, n.$$

Remember that when there exist heteroskedasticity and autocorrelation in the error term, we need to apply a NeweyWest(HAC) variance covariance estimator. Here, I just show you with the commands and don't display the results. The results will be summarized in a table as well as the 2 step estimates for comparision.

```r
# Numerical 2SGMM Method to A Simple Linear Over-identified case

# Since we tested out the heteroskedasticity,
# this dataset provides an appropriate example for studying 2SGMM.


# Firstly, recall HA-robust ols estimates for later comparison
model = value ~ crime + industrial + distance

ols = lm(model, data = dt)
#summary(ols)
coeftest(ols, vcov. = NeweyWest)

cov_hac = NeweyWest(ols)
se_hac = sqrt(diag(cov_hac))

t_hac = coef(ols)/se_hac
p_hac = pt(abs(t_hac), df = nrow(dt) - ncol(X), lower.tail = FALSE)*2
```

2.2 Numerical 2SGMM

2.2.1 Step 1 of the 2SGMM Method

Firstly, let's review the procedures of the 2 step estimation[4].

[4] For a better understanding of the codes I provided, I use the notations from class and modified them a little bit. Of course, stacked form is used as well.

Step 1 of the 2SGMM Method

Choose the metric in the first step. The optimal metric is the inverse of the variance of the orthogonality conditions which are $g(K_i,\theta) = (\mathbf{Z}_i^{IV})^T(Y_i - \mathbf{X}_i\theta) = (\mathbf{Z}_i^{IV})^T u_i$ for all i. And the population orthogonality condition(or so-called moment condition) is given by $\mathbb{E}\left(g(K_i,\theta)\right) = \mathbf{0}$.

Thus it is better to choose a metric in the first step, which is close to the optimal metric. To do so, we assume that homoskedasticity assumption is satisfied. For example,

$$\mathbb{E}[u_i^2|\mathbf{Z}_i^{IV}] = \mathbb{E}[u_i^2] = \sigma^2.$$

Then, the first step metric would be

$$\begin{aligned}
\widehat{Var}(g(K_i,\widehat{\theta}))_1 &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Z}_i^{IV})^T(\mathbf{Z}_i^{IV}),\\
&= \frac{1}{n}(\underline{\mathbf{Z}}^{IV})^T(\underline{\mathbf{Z}}^{IV}),\\
\widehat{W}_1 &= \left(\widehat{Var}(g(K_i,\widehat{\theta}))_1\right)^{-1}.
\end{aligned}$$

When this metric is applied, the first step estimator is then given by

$$\begin{aligned}
\hat{\theta}_1 = &\left\{\left(\frac{1}{n}\sum_{i=1}^{n}((\mathbf{Z}_i^{IV})^T\mathbf{X}_i)\right)^T \widehat{W}_1 \left(\frac{1}{n}\sum_{i=1}^{n}((\mathbf{Z}_i^{IV})^T\mathbf{X}_i)\right)\right\}^{-1}\\
&\times \left(\frac{1}{n}\sum_{i=1}^{n}((\mathbf{Z}_i^{IV})^T\mathbf{X}_i)\right)^T \widehat{W}_1 \left(\frac{1}{n}\sum_{i=1}^{n}((\mathbf{Z}_i^{IV})^T\mathbf{Y}_i)\right),\\
= &\left\{\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)^T \widehat{W}_1(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right\}^{-1}\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)^T \widehat{W}_1(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{Y}},
\end{aligned}$$

where $\underline{\mathbf{Y}} = (Y_1,\ldots,Y_n)^T$, $\underline{\mathbf{X}} = (\mathbf{X}_1^T,\ldots,\mathbf{X}_n^T)^T$ and $\underline{\mathbf{Z}}^{IV} = ((\mathbf{Z}_1^{IV})^T,\ldots,(\mathbf{Z}_n^{IV})^T)^T$. Such an estimator with the first step metric $\widehat{W}_1$ is called as the 2 step instrumental variable(2S-IV) estimator or the 2 step least square(2SLS) estimator. After being calculated, this 2S-IV is then plugged in the second step of GMM.

And the asmptotic variance covariance matrix can, according to the properties of GMM estimator(appendix A), be written as

$$\begin{aligned}
\mathbb{V}(\hat{\theta}_1) = &\left\{\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]^T W_0\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]\right\}^{-1}\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]^T W_0 Var\left(g(K_i,\theta_0)\right)\\
&\times W_0\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]\left\{\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]^T W_0\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]\right\}^{-1},
\end{aligned}$$

where θ_0 is the true parameter with $Var\left(g(K_i,\theta_0)\right) = \mathbb{E}\left[g(K_i,\theta_0)g(K_i,\theta_0)^T\right]$ and $\widehat{W}_1 \xrightarrow[n\to\infty]{\mathbb{P}} W_0$.

And if we just take $\widehat{W}_1$ and $\widehat{Var}(g(K_i,\theta))_1$ as the empirical counterparts of W_0 and $Var\left(g(K_i,\theta_0)\right)$, respectively. We will have the covariance estimator as

$$\begin{aligned}
\widehat{\mathbb{V}}(\hat{\theta}_1) = &\left\{\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)^T \widehat{W}_1(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right\}^{-1}\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)^T \widehat{W}_1\widehat{W}_1^{-1}\\
&\times \widehat{W}_1\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)\left\{\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)^T \widehat{W}_1(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right\}^{-1},\\
= &\left\{\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)^T \widehat{W}_1(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right\}^{-1}.
\end{aligned}$$

Now, let's check the commands and runing results.

```
# Secondly, perform 2SGMM and
# the 1st step is like follows (looks like 2SLS).

Z = as.matrix(Z)
X = as.matrix(X)

ZX = t(Z) %*% X/n # or ZX = crossprod(Z, X)/n
ZY = t(Z) %*% Y/n

var_g_1 = t(Z) %*% Z/n

theta_1 = solve(t(ZX) %*% solve(var_g_1) %*% ZX) %*%
t(ZX) %*% solve(var_g_1) %*% ZY

cov_1   = solve(t(ZX) %*% solve(var_g_1) %*% ZX)


# Caculate the estimates and check
se_1   = sqrt(diag(cov_1))
t_1    = theta_1/se_1
p_1    = pt(abs(t_1), df = nrow(dt) - ncol(X), lower.tail = FALSE)*2

print("1st␣step␣estimates:"); theta_1

print("1st␣step␣se␣estimates:"); se_1

print("T␣statistics␣:"); t_1

print("P␣values:"); p_1
```

And the results are listed as

```
## [1] "1st step estimates:"
##                [,1]
## constant   37.7720297
## crime      -1.1413414
## industrial -0.4293433
## distance   -1.6688765
##
## [1] "1st step se estimates:"
## constant    crime     industrial distance
## 4.7137367  0.3971785  0.2482251  0.7364254
##
## [1] "T statistics :"
##                [,1]
## constant    8.013182
## crime      -2.873623
## industrial -1.729653
## distance   -2.266185
##
##[1] "P values:"
##                [,1]
## constant   7.881782e-15
## crime      4.229911e-03
## industrial 8.430688e-02
## distance   2.386493e-02
```

2.2.2 Step 2 of 2SGMM Method

Let's review the remaing part of the 2SGMM method,

Step 2 of the 2SGMM Method

From step 1, we have the estimator $\hat{\theta}_1$ in hand and if moments are assumed to be IID distributed, we have

$$\begin{aligned}
\widehat{Var}(g(K_i,\widehat{\theta}_1))_2 &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Z}_i^{IV})^T(Y_i-\mathbf{X}_i\hat{\theta}_1)^2(\mathbf{Z}_i^{IV}),\\
&= \frac{1}{n}(\underline{\mathbf{Z}}^{IV})^T(\underline{\mathbf{Y}}-\underline{\mathbf{X}}\hat{\theta}_1)(\underline{\mathbf{Y}}-\underline{\mathbf{X}}\hat{\theta}_1)^T(\underline{\mathbf{Z}}^{IV}),\\
\widehat{W}_2 &= \left(\widehat{Var}(g(K_i,\widehat{\theta}_1))_2\right)^{-1}.
\end{aligned}$$

However, we already know that there exist heteroskedasticity and autocorrelation in this case, it's suggested by Chaussé, P.(2010)[2] that we should use a HAC variance covariance estimator. In this situation, denote $\underline{\hat{\mathbf{u}}} = \underline{\mathbf{Y}} - \underline{\mathbf{X}}\hat{\theta}_1$, we have

$$\begin{aligned}
Var(g(\underline{\mathbf{K}},\widehat{\theta}))_2 &= Var\left((\underline{\mathbf{Z}}^{IV})^T\underline{\hat{\mathbf{u}}}\right),\\
&= (\underline{\mathbf{Z}}^{IV})^T Var(\underline{\hat{\mathbf{u}}})\underline{\mathbf{Z}}^{IV}.
\end{aligned}$$

Denote that $\Omega = Var(\underline{\hat{\mathbf{u}}})$, if we have the NeweyWest covariance estimate of $\hat{\theta}_1$ as Q(check appendix B for definition), then

$$Q = \underline{\mathbf{X}}^T\widehat{\Omega}\underline{\mathbf{X}}.$$

Therefore with the generalized inverse operation, $\widehat{\Omega}$ can be calculated numerically by

$$\widehat{\Omega} = (\underline{\mathbf{X}}^T)^{-1}Q(\underline{\mathbf{X}})^{-1}.$$

Or to make it more comfortable to see[a],

$$\widehat{\Omega} = \left(\underline{\mathbf{X}}\underline{\mathbf{X}}^T\right)^{-1}\underline{\mathbf{X}}Q\underline{\mathbf{X}}^T\left(\underline{\mathbf{X}}\underline{\mathbf{X}}^T\right)^{-1}.$$

With these conditions above, it's able to get

$$\begin{aligned}
\widehat{Var}(g(K_i,\widehat{\theta}_1))_2^{HAC} &= \frac{1}{n}(\underline{\mathbf{Z}}^{IV})^T\widehat{\Omega}\underline{\mathbf{Z}}^{IV},\\
\widehat{W}_2^{HAC} &= \left(\widehat{Var}(g(K_i,\widehat{\theta}_1))_2^{HAC}\right)^{-1}.
\end{aligned}$$

Here, $\widehat{W}_2$ is just the metric in sildes P201. But we finally use $\widehat{W}_2^{HAC}$. The results of these two metrics are provided at the end of this section.
The remaining parts are similar to those in step 1.

$$\begin{aligned}
\hat{\theta}_2 &= \left\{\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)^T\widehat{W}_2(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right\}^{-1}\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)^T\widehat{W}_2(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{Y}},\\
\mathbb{V}(\hat{\theta}_2) &= \left\{\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]^T W_0\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]\right\}^{-1}\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]^T W_0 Var\left(g(K_i,\theta_0)\right)\\
&\quad\times W_0\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]\left\{\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]^T W_0\mathbb{E}\left[(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right]\right\}^{-1},\\
\widehat{\mathbb{V}}(\hat{\theta}_2) &= \left\{\left((\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right)^T\widehat{W}_2(\underline{\mathbf{Z}}^{IV})^T\underline{\mathbf{X}}\right\}^{-1},
\end{aligned}$$

where θ_0 is the true parameter with $Var\left(g(K_i,\theta_0)\right) = \mathbb{E}\left[g(K_i,\theta_0)g(K_i,\theta_0)^T\right]$ and $\widehat{W}_2 \xrightarrow[n\to\infty]{\mathbb{P}} W_0$.

[a] Almostly same values were returned.

As usual, I show you the codes and results as below.

```r
# Do 2nd step 2SGMM

# Theoretically, we should calculate the estimate of the variance with theta_1 by

Zu_1    = t(Z) %*% (Y - X %*% theta_1)
var_g_2 = Zu_1 %*% t(Zu_1)/n
# or var_g_1 = tcrossprod(Zu_1, Zu_1)/n

# But the previous var_g_2 is derived under the iid assumption.
# It is equal to set vcov = "iid" in gmm function, which would slow down the calculation.

# In practice, let's build a HAC covariance estimate from 2sls residuals and X.
# For simplicity, set L = 10(<= n).

u_1 = Y - X %*% theta_1
i = 1
l = 1
L = 10
weight = rep(0, L)

Q = matrix(0, nrow = ncol(X), ncol = ncol(X))
Q_1 = matrix(0, nrow = ncol(X), ncol = ncol(X))
Q_2 = matrix(0, nrow = ncol(X), ncol = ncol(X))

for (l in 1:L) {
    weight[l] = 1 - l/(L + 1)
}

for (i in 1:n) {
    Q_1 = Q_1 + u_1[i]^2 * (X[i,] %*% t(X[i,])) # X[i,] here returns column vector
}
Q_1 = Q_1/n

for (l in 1:L) {
    for (i in (l+1):n) {
        Q_2 = Q_2 + weight[l]*u_1[i]*u_1[i-l]*
        (X[i,] %*% t(X[i-l,]) + X[i-l,] %*% t(X[i,]))
    }
}
Q_2 = Q_2/n
Q = Q_1 + Q_2

XXT = X %*% t(X)
omega = ginv(XXT) %*% (X %*% Q %*% t(X)) %*% ginv(XXT)
# Use ginv here to suppress error warnings from traditional inverse operation.
# omega = ginv(t(X)) %*% Q %*% ginv(X) returns almostly same values.
var_g_2 = t(Z) %*% omega %*% Z /n

theta_2 = ginv(t(ZX) %*% ginv(var_g_2) %*% ZX) %*%
t(ZX) %*% ginv(var_g_2) %*% ZY

cov_2   = ginv(t(ZX) %*% ginv(var_g_2) %*% ZX)
```

```


# Calculate the estimates and check
se_2   = sqrt(diag(cov_2))
se_2   = as.vector(se_2)
t_2    = theta_2/se_2
p_2    = pt(abs(t_2), df = nrow(dt) − ncol(X), lower.tail = FALSE)∗2

print(”2nd␣step␣estimates:”); theta_2

print(”2nd␣step␣se␣estimates:”); se_2

print(”T␣statistics␣:”); t_2

print(”P␣values:”); p_2

# Finally, theta_1 is not optimal
se_1 >= se_2
```

Firstly, let me show you what will happen if $\widehat{W}_2$ in the IID case is used.

```
## [1] "2nd step estimates:"
##           [,1]
## [1,]  0.1411261
## [2,] -2.0908670
## [3,] -0.1512835
## [4,]  0.9792031
##
## [1] "2nd step se estimates:"
## [1] 0.4575280 6.7785496 0.4904583 3.1745572
##
## [1] "T statistics :"
##          [,1]
## [1,]  0.3084534
## [2,] -0.3084534
## [3,] -0.3084534
## [4,]  0.3084534
##
## [1] "P values:"
##          [,1]
## [1,] 0.7578653
## [2,] 0.7578653
## [3,] 0.7578653
## [4,] 0.7578653
```

As we can see, after the second step estimation, efficiency get worse than before and due to the sigularity of variance covariance matrix, we obtain the same but bad p values after generalized inverse operation is performed. Then, let's check the results using the NeweyWest covariance estimator.

```
## [1] "2nd step estimates:"
##          [,1]
## [1,] 31.2124282
## [2,] -0.7333242
## [3,] -0.3499875
## [4,] -0.6303447
##
## [1] "2nd step se estimates:"
## [1] 3.0612074 0.3907887 0.1751742 0.4126052
##
## [1] "T statistics :"
##           [,1]
## [1,] 10.196117
## [2,] -1.876524
## [3,] -1.997940
## [4,] -1.527719
##
## [1] "P values:"
##           [,1]
## [1,] 2.589106e-22
## [2,] 6.116328e-02
## [3,] 4.626223e-02
## [4,] 1.272123e-01
```

And we can check the optimality by logical operation command as

```
## se_1 >= se_2
## constant    crime industrial   distance
## TRUE        TRUE        TRUE        TRUE
```

2.3 Displaying the results

All the estimation results are given in table(1) and I omit the interpretations because they are not different from those in OLS estimation.

```
# Display the results
stargazer(
ols, ols, ols,
coef = list(coef(ols), theta_1, theta_2), se = list(se_hac, se_1, se_2),
t = list(t_hac, t_1, t_2), p = list(p_hac, p_1, p_2),
t.auto = FALSE, p.auto = FALSE,
report = "vcstp", keep.stat = c("n"),
add.lines = list(
c("Type", "HA-Roubusted␣OLS", "1st␣step␣GMM", "2nd␣step␣GMM")),
title = "Results␣of␣rols␣and␣numerical␣2␣step␣GMM",
label = "Numeric",
type = "latex", header = FALSE, font.size = "small",
table.placement = "htb", omit.table.layout = "n"
)
```

Table 1: Results of rols and numerical 2 step GMM

	Dependent variable:		
	value		
	(1)	(2)	(3)
crime	−0.273	−1.141	−0.733
	(0.055)	(0.397)	(0.391)
	t = −4.926	t = −2.874	t = −1.877
	p = 0.00001	p = 0.005	p = 0.062
industrial	−0.730	−0.429	−0.350
	(0.142)	(0.248)	(0.175)
	t = −5.153	t = −1.730	t = −1.998
	p = 0.00000	p = 0.085	p = 0.047
distance	−1.016	−1.669	−0.630
	(0.375)	(0.736)	(0.413)
	t = −2.710	t = −2.266	t = −1.528
	p = 0.007	p = 0.024	p = 0.128
Constant	35.505	37.772	31.212
	(2.984)	(4.714)	(3.061)
	t = 11.899	t = 8.013	t = 10.196
	p = 0.000	p = 0.000	p = 0.000
Type	HA-Roubusted OLS	1st step GMM	2nd step GMM
Observations	506	506	506

3 2SLS and 2SGMM by R

3.1 2SLS and Hausman Tests for Endogeneity

3.1.1 2SLS in R

The works we did in section(2.2.1) can be simply performed by `AER::ivreg` function in R. I just show you the codes here and the result remains to be talked about a little later.

```
# Built-in R Functions

# 2SLS and Hausman Wu Test for Endogeneity (By ivreg).
model_iv = value ~ crime + industrial + distance | black + ptratio + industrial + distance

twoSLS = ivreg(model_iv, data = dt)
```

3.1.2 Durbin-Wu-Hausman Test for Endogeneity

The DWH test looks similar to the Hausman test in chapter 5, talking about how to choose between FE model and RE model. Here, let me share a simple definition with you.

Durbin-Wu-Hausman Test

Assume that

$$y_i = \mathbf{X}_i\beta + u_i \quad \forall i = 1, \ldots, N,$$

where u_i is a 0-meaned error term and $\mathbf{X}_i^T \in \mathbb{R}^K$.
Declare the null hypothesis $H_0 : \mathbb{E}[\mathbf{X}_i^T u_i] = 0$[a] and suppose that we have the ols estimator $\hat{\beta}_{ols}$ and gmm estimator $\hat{\beta}_{GMM}$, respectively. The test statistic is given by

$$\hat{H} = \left(\hat{\beta}_{ols} - \hat{\beta}_{GMM}\right)^T \left\{\widehat{Var}(\hat{\beta}_{GMM}) - \widehat{Var}(\hat{\beta}_{ols})\right\}^{-1} \left(\hat{\beta}_{ols} - \hat{\beta}_{GMM}\right) \xrightarrow[n\to\infty]{d} \chi^2(K)$$

It is also known that

- Under the null hypothesis, the estimator $\hat{\beta}_{ols}$ is efficient and the estimator $\hat{\beta}_{GMM}$ is consistent but typically not efficient.
- Under the alternative hypothesis, the estimator $\hat{\beta}_{ols}$ is inconsistent, the estimator $\hat{\beta}_{GMM}$ is consistent.

[a] Exogeneity condition meaning that all explanatory variables are exogenous.

Apply this method in R[5], we have the follows,

```
# Durbin-Wu-Hausman Test
dwh.test = function(model.iv, model.ols){
    cf_diff = coef(model.iv) - coef(model.ols)
    vc_diff = vcovHC(model.iv, ”HC0”) - vcovHC(model.ols, ”HC0”)
    # NeweyWest() doesn't fit well to model.iv so we use the White estimator.
    x2_diff = as.vector(t(cf_diff) %*% solve(vc_diff) %*% cf_diff)
    pvalue = pchisq(q = x2_diff, df = dim(vc_diff)[1], lower.tail = F)

    result = list(x2_diff, dim(vc_diff)[1], pvalue)
    names(result) = c(”DWH␣Statistic”, ”df”, ”P␣value”)
    return(result)
}
dwh.test(twoSLS, ols)
```

[5] The original code is posted at ”http://klein.uk/R/myfunctions.R” from professor Thilo Klein. And I modified it a little bit.

And then, the results of applying this test to our framework are shown as

```
## $`DWH Statistic`
## [1] 10.77423
##
## $df
## [1] 4
##
## $P value
## [1] 0.02922208
```

From this result, we know that the null hypothesis is rejected and there exists at least one endogenous variable.

3.1.3 Wu-Hausman Test in R

Indeed, R has already encompassed the Wu-Hausman 2 step test for endogeneity. I don't clearly specify this test and anyone interested is suggested to refer to Wooldridge(2010)[9] for details.[6]

```
dwh.test(twoSLS, ols) # ”http://klein.uk/R/myfunctions.R” from professor Thilo Klein.

# Use ”diagnostics = TRUE” calls a Wu-Hausman(2 step) F test(available in Wooldridge(2003)).
summary(twoSLS, vcov. = NeweyWest, df = Inf, diagnostics = TRUE)
```

Let's check the result.

```
## Call:
## ivreg(formula = model_iv, data = dt)
##
## Residuals:
##    Min      1Q    Median    3Q     Max
## -16.515  -6.069  -1.965   2.639  84.315
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  37.7720     3.3464  11.287  < 2e-16 ***
## crime        -1.1413     0.4339  -2.630 0.008527 **
## industrial   -0.4293     0.2126  -2.019 0.043452 *
## distance     -1.6689     0.4852  -3.440 0.000582 ***
##
## Diagnostic tests:
##                   df1 df2 statistic  p-value
## Weak instruments   2  501   5.921   0.00287 **
## Wu-Hausman         1  501   15.498  9.43e-05 ***
## Sargan             1   NA   17.923  2.30e-05 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 10.25 on Inf degrees of freedom
## Multiple R-Squared: -0.2352,Adjusted R-squared: -0.2425
## Wald test: 35.98 on 3 DF,  p-value: 7.561e-08
```

Look at the Wu-Hausman row, the null hypothesis is rejected at 0.1% level. And the endogenous variable is indeed `crime` because we specify the structure in the previous model_iv.

[6] And be careful that the second step is replaced by the Fisher test in R but not the t test.

3.2 2SGMM and The Sargan-Hansen (Sargan's J) Test in R

3.2.1 2SGMM in R

The works we did in section(2) can also be conducted by `gmm::gmm` function in R. As usual, I just display the codes here.

```
# 2SGMM and Sargan's J Test (By gmm)
value = dt[, "value"]
crime = dt[, "crime"]
industrial = dt[, "industrial"]
distance = dt[, "distance"]

instruments = dt[, c("black", "ptratio", "industrial", "distance")]

twoSGMM_iid = gmm(g = model, x = instruments, vcov = "iid")
twoSGMM_HAC = gmm(g = model, x = instruments, vcov = "HAC")

summary(twoSGMM_iid)
summary(twoSGMM_HAC)
```

3.2.2 The Sargan-Hansen Test

The Sargan-Hansen Test

The Sargan-Hansen test is a statistical test used for testing over-identifying restrictions in a statistical model. It was proposed by John Denis Sargan in 1958[7], and several variants were derived by him in 1975[8]. Lars Peter Hansen re-worked through the derivations and showed that it can be extended to general non-linear GMM in a time series context[3].

As same as the notations in the slides, we denote that $dim(g) = L$ and $dim(\theta) = d$. When we have more instruments than we need to identify an equation, we can test whether the additional instruments are valid in the sense that they are uncorrelated with the error term.

In our case, we have the maybe endogenous explanatory variables and the exogenous instrumental variables as

$$\mathbf{X}_i = (constant_i, crime_i, industrial_i, distance_i),$$
$$\mathbf{Z}_i = (constant_i, black_i, ptratio_i, industrial_i, distance_i),$$

respectively. So conducting the Sargan-Hansen test means that we want to check whether `black` and `ptratio` are good instruments. By the way, you can take `industrial` and `distance` as instruments of their own. For more details, please refer to Wooldrige(2010)[9].

And then, with the statistical criterion $\hat{\theta}_{GMM} = \underset{\theta\in\Theta}{argmax}\left\{-(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{K}_i,\theta))^T\widehat{W}(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{K}_i,\theta))\right\}$, we are able to build the J statistic.

The null hypothesis is that $H_0 : \exists\theta, \mathbb{E}[g(\mathbf{K}_i,\theta)] = \mathbf{0}$, and the statistic is

$$\hat{J} = (\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(\mathbf{K}_i,\hat{\theta}))^T\widehat{W}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(\mathbf{K}_i,\hat{\theta})) \xrightarrow[n\to\infty]{d} \chi^2(L-d)$$

The statistic is then based on the optimal GMM estimator.

This test is already included in `gmm::gmm` function, to look for the result, please use `base::summary` function.

```
## Call:
## gmm(g = model, x = instruments, vcov = "HAC")
##
## Method:  twoStep
##
## Kernel:  Quadratic Spectral(with bw =  1.54322 )
```

```
##
## Coefficients:
##                Estimate     Std. Error   t value      Pr(>|t|)
## (Intercept)   3.8101e+01   3.2027e+00   1.1897e+01   1.2331e-32
## crime        -1.1011e+00   3.4308e-01  -3.2095e+00   1.3297e-03
## industrial   -4.6190e-01   1.8771e-01  -2.4607e+00   1.3867e-02
## distance     -1.7307e+00   4.4494e-01  -3.8898e+00   1.0032e-04
##
## J-Test: degrees of freedom is 1
##                 J-test    P-value
## Test E(g)=0:    5.698567  0.016979
##
##
## Initial values of the coefficients
## (Intercept)    crime     industrial   distance
## 37.7720297  -1.1413414  -0.4293433  -1.6688765
```

Focus on the ”J-Test” part, since the p value is smaller than 0.02, we can say that the null hypothesis is rejected at 5% level and this model specification is indeed not correct. Better instruments are wanted for making a proper estimation.

3.2.3 Displaying the results

```
# Display
stargazer(
twoSLS, twoSGMM_iid, twoSGMM_HAC,
report = "vcstp", keep.stat = c("n"),
add.lines = list(
c("Type", "2SLS", "2SGMM(iid)", "2SGMM(HAC)")),
title = "Results␣of␣2SLS␣and␣2SGMM␣by␣R",
label = "byR",
type = "latex", header = FALSE, font.size = "small",
table.placement = "htb", omit.table.layout = "n"
)
setwd("~")
```

Table 2: Results of 2SLS and 2SGMM by R

	Dependent variable:		
	value		
variable			
	(1)	(2)	(3)
crime	−1.141	−1.141	−1.101
	(0.181)	(0.180)	(0.343)
	t = −6.305	t = −6.330	t = −3.209
	p = 0.000	p = 0.000	p = 0.002
industrial	−0.429	−0.429	−0.462
	(0.113)	(0.113)	(0.188)
	t = −3.795	t = −3.810	t = −2.461
	p = 0.0002	p = 0.0002	p = 0.014
distance	−1.669	−1.669	−1.731
	(0.336)	(0.334)	(0.445)
	t = −4.972	t = −4.992	t = −3.890
	p = 0.00000	p = 0.00000	p = 0.0002
Constant	37.772	37.772	38.101
	(2.148)	(2.140)	(3.203)
	t = 17.582	t = 17.652	t = 11.897
	p = 0.000	p = 0.000	p = 0.000
Type	2SLS	2SGMM(iid)	2SGMM(HAC)
Observations	506	506	506

Look at table(2), you may find that 2SLS and 2SGMM(iid) returned similar estimates for that IID assumption is not satisfied in this case. And the reason why 2SGMM(HAC) didn't give the optimal estimates is that, by the Sargan-Hansen test, we know that our model specification is indeed not correct, therefore such results were returned.

Appendices

A Properties of the GMM estimator

Under suitable conditions, the GMM estimator satisfies

$$\hat{\theta} \xrightarrow[n\to\infty]{\mathbb{P}} \theta_0,$$
$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow[n\to\infty]{d} \mathcal{N}_{\mathbb{R}^{dim(\theta)}}(\mathbf{0}, \mathbb{V}(\hat{\theta})),$$

with

$$\mathbb{V}(\hat{\theta}) = \left\{G^T W_0 G\right\}^{-1} G^T W_0 Var\left(g(K_i, \theta_0)\right) W_0 G \left\{G^T W_0 G\right\}^{-1},$$

and

$$Var\left(g(K_i, \theta_0)\right) = \mathbb{E}\left[g(K_i, \theta_0) g(K_i, \theta_0)^T\right], \widehat{W} \xrightarrow[n\to\infty]{\mathbb{P}} W_0.$$

Be careful that G is the population level Jacobian

$$G = G(\theta_0) = \mathbb{E}\left[\nabla_\theta g(K_i, \theta_0)\right] \in \mathcal{M}_{dim(g)\times dim(\theta)}(\mathbb{R})$$

B NeweyWest Variance Covariance Estimator

According to Whitney K. Newey and Kenneth D. West(1987), if $\mathbf{X}_i$ is the 1×d row vector and e_i is the residual, for all i = 1, . . . , n. Then NeweyWest estimator should be

$$Q = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \mathbf{X}_i^T \mathbf{X}_i + \frac{1}{n}\sum_{l=1}^{L}\sum_{i=l+1}^{n} w_l e_i e_{i-l}\left(\mathbf{X}_i^T \mathbf{X}_{i-l} + \mathbf{X}_{i-l}^T \mathbf{X}_i\right),$$
$$w_l = 1 - \frac{l}{L+1},$$

where w_l can be thought of as a ”weight”. Disturbances that are farther apart from each other are given lower weight, while those with equal subscripts are given a weight of 1.

References

[1] Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air, *J. Environ. Economics & Management,* 5: 81-102.

[2] Chaussé, P. (2010). Computing generalized method of moments and generalized empirical likelihood with R. *Journal of Statistical Software*, 34(11), 1-35.

[3] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029-1054.

[4] Hlavac, Marek (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables.*

[5] Newey, Whitney K; West, Kenneth D(1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica.* Vol. 55, No. 3, pp. 703-708.

[6] Nichols, Austin(2006). *Weak Instruments: An Overview and New Techniques.*

[7] Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, 393-415.

[8] Sargan, J. D. (1988)[1975]. Testing for misspecification after estimating using instrumental variables. *Contributions to Econometrics.* New York: Cambridge University Press. ISBN 0-521-32570-6.

[9] Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data,* MIT Press.