UNIVERSITY OF EDINBURGH
Business School

2021-22

Industrial Organisation (CMSE11450)

Individual Assignment

B193613

Word count: 2600

<u>Ques 1.1</u> Consider this study of crime rate in the US represented by the model below:

$$crime_i = \beta_0 + \beta_1 \log educ\ av_i + \beta_2 Black\ male_i + \beta_3 age\ range_i + \beta_4 FBI\ offence_i + e_i$$

a) Estimate the model 1, using OLS (with standard errors clustered at the State level - variable State). Could we argue that educm is an endogenous regressor? Why?

In our model, we strive to identify the explanatory factors that contribute to the arrest rate. The log of the arrest rate is our dependent variable in this case. The log value of the average education, the proportion of the community that is black, the age group at the time of the arrest, and the offense committed are our independent variables.

A simple OLS of the findings yields the results as shown in *Fig 1*:

```
. reg crime lneducm blackm rage off, cluster (state)

Linear regression                               Number of obs     =     10,458
                                                F(4, 50)          =     518.88
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2456
                                                Root MSE          =     1.4232

                                   (Std. err. adjusted for 51 clusters in state)
```

crime	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
lneducm	-1.575748	.1839358	-8.57	0.000	-1.945194	-1.206302
blackm	1.531634	.2434341	6.29	0.000	1.042682	2.020586
rage	-.3657432	.0106885	-34.22	0.000	-.3872118	-.3442746
off	.0073975	.0007388	10.01	0.000	.0059137	.0088814
_cons	13.7119	.5335477	25.70	0.000	12.64023	14.78356

Fig 1 – OLS regression implementation

The equation comes out to be as below:

$crime_i$ = 13.71 - 1.58 log $educ\ av_i$ + 1.53 $Black\ male_i$ - 0.37 $age\ range_i$ + 0.0074 $FBI\ offence_i$ + e_i

The stats show that the log of education has a significant negative correlation with the log of the arrest rate. With each marginal increase in log education, the log arrest rate is reduced by a factor of 1.58. The t-value of -8.57 is very significant. The black people community has a very high coefficient in the positive direction. The log arrest rate increases by 1.53 for every additional one percent and at 6.29, this becomes much more significant. On considering the FBI offense code, it shows that with every offense change the log of arrest is impacted/increased by a small factor of 0.0074 but the high t-value of 10.01 means it has a significant impact on the log arrest rate. Finally, with a t-value of -25.70, the age range exhibits a highly statistically significant negative outcome. The log arrest rate decreases with each marginal rise in the age group. It is suggested that as people age, their proclivity to commit crime lessens. If the additional detailed inference is required for each type of offense code and a better-fit regression model, then we can investigate the results of [1].

A variable can be said to be an endogenous variable when it is affected by other variables. There must be a correlation between the variable and the error term for a variable to be endogenous. The existence of a relationship between the variable and other observable and non-observable variables results in this correlation. In this model, the education in years for any offender depends on several factors like educational facilities and services in the state, the offender's family financial status, personal motivation of the offender to study, and age. Therefore, we can say that the ***educm*** is an endogenous variable.

b) Discuss the possible endogeneity of the other control variables.

As discussed earlier, the exogenous variables are the ones that are not affected by other variables and the endogenous variable are the ones that are affected by the other variables. In the above-mentioned model, the percentage of black people (blackm) can be affected by the geographical location (US State), and thus, the ***blackm*** variable is endogenous. The age of the offender(rage) and the offense code are the variables that are not impacted by any factors and thus these variables are exogenous variables.

<u>Ques 1.2</u> Consider the variable dropm. It is the percent of high school drop-out in the State.

a) Argue about the validity of $dropmi$ as a possible instrument for educmi. Discuss its exogeneity

An instrumental variable (sometimes known as an "instrument" variable) is a third variable, Z, that is employed in regression analysis when endogenous variables are present. If in the above-mentioned model, we use dropm as an instrument for educm then we get the details as shown in *Fig2.*

The lower p-value in the Hansen test is the evidence against the null hypothesis (H0: instrument is valid) and thus indicating that the ***dropm*** is not a valid instrument for ***educm***.

```
. ivreg2  crime (lneducm=dropm) blackm rage off, cluster(state)

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity and clustering on state

Number of clusters (state) =         51                Number of obs =     10458
                                                       F(  4,      50) =   475.52
                                                       Prob > F      =    0.0000
Total (centered) SS     =  28062.97945                 Centered R2   =    0.2452
Total (uncentered) SS   =  193987.4667                 Uncentered R2 =    0.8908
Residual SS             =  21181.27018                 Root MSE      =     1.423

------------------------------------------------------------------------------
             |               Robust
       crime | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     lneducm |   -1.88964   .7314375    -2.58   0.010    -3.323231   -.4560489
      blackm |    1.47992   .3063595     4.83   0.000     .8794663    2.080374
        rage |   -.371092   .0162711   -22.81   0.000    -.4029828   -.3392012
         off |   .0074654   .0007403    10.08   0.000     .0060145    .0089163
       _cons |   14.58567   2.054575     7.10   0.000     10.55878    18.61257
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):             34.026
                                                   Chi-sq(1) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):              297.739
                         (Kleibergen-Paap rk Wald F statistic):        106.929
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.000
                                                 (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         lneducm
Included instruments: blackm rage off
Excluded instruments: dropm
------------------------------------------------------------------------------
```

Fig 2 Instrument variable weak test

To check the exogeneity of the educm variable, we did the endogeneity test (Durbin-Wu Hausman test) as shown in *Fig 3*.

```
.  estat endog

  Tests of endogeneity
  H0: Variables are exogenous

  Durbin (score) chi2(1)          =  .134568  (p = 0.7137)
  Wu-Hausman F(1,10452)           =  .134492  (p = 0.7138)
```

Fig 3 Durbin-Wu-Hausman test (Endogeniety test)

The results of the test done above show high p values, which supports the null hypothesis indicating that the variable ***educm*** is exogenous. It also signifies that the OLS model is appropriate for this and ***dropm*** as an instrument is not appropriate.

b) Produce your own two stage least squares estimator of the coefficient $\beta 1$ using $dropmi$ as an instrument for log-prices. (You can use the command *ivregress 2sls* or *ivreg2*). Interpret the results.

For implementing the 2 stage least square, we need to follow 2 steps. Firstly, apply regression making average education year **(educm)** as a dependent variable and using percentage of high school dropouts **(dropm)** as an independent variable. Doing so will remove any correlation between ***educm*** and the error term. As a second step, we need to regress the crime rate over educm and other variables.

```
. ivreg2  crime (lneducm=dropm) blackm rage off, robust

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                   Number of obs =     10458
                                                   F(  4, 10453) =    751.97
                                                   Prob > F      =    0.0000
Total (centered) SS     =  28062.97945             Centered R2   =    0.2452
Total (uncentered) SS   =  193987.4667             Uncentered R2 =    0.8908
Residual SS             =  21181.27018             Root MSE      =     1.423
```

crime	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
lneducm	-1.88964	.8639416	-2.19	0.029	-3.582934	-.1963457
blackm	1.47992	.1913123	7.74	0.000	1.104955	1.854885
rage	-.371092	.0160148	-23.17	0.000	-.4024805	-.3397035
off	.0074654	.00074	10.09	0.000	.006015	.0089157
_cons	14.58567	2.406412	6.06	0.000	9.869191	19.30215

*Fig 4. - 2SLS model (**dropm** as instrument)*

```
. correlate state year off rage obsm lneducm blackm crime dropm work_age
(obs=9,829)
```

	state	year	off	rage	obsm	lneducm	blackm	crime	dropm	work_age
state	1.0000									
year	0.0158	1.0000								
off	0.0089	0.0673	1.0000							
rage	-0.0129	0.1421	0.0155	1.0000						
obsm	-0.0311	0.3550	0.0212	-0.0525	1.0000					
lneducm	-0.0025	0.6683	0.0464	-0.3459	0.2789	1.0000				
blackm	-0.2268	0.0090	-0.0134	-0.0294	0.0255	-0.1589	1.0000			
crime	-0.0517	-0.1270	0.0926	-0.4559	0.0053	0.0546	0.1324	1.0000		
dropm	0.0039	0.1269	-0.0058	-0.0546	0.0632	0.1820	-0.0395	-0.0014	1.0000	
work_age	0.0540	0.1703	0.0105	-0.2020	0.1967	0.1522	0.0890	0.0568	0.0264	1.0000

Fig 5. - Correlation

It shows correlation between lneducm and dropm to be 0.18. It is not huge but still shows that there is link.

The equation comes out to be as below:

$crime_i$ = 14.58 - 1.89 log $educ\ av_i$ + 1.48 $Black\ male_i$ - 0.37 $age\ range_i$ + 0.0075 $FBI\ offence_i$ + e_i

It is clearly visible that the average years of education ***educm*** and the age range ***rage*** are the variables that have a negative relationship with the dependent variable log of arrest ***crime***. This implies that the arrest rate decrease with the older age group and the group of highly educated individuals. On the other hand, the percentage population of the black people and offense codes have positive relationship with the log of arrest. This means that the states/ areas with higher black population corresponds to high crime and also for certain offenses the crime is higher.
According to the p-values, all of the independent variables had a significant influence on the arrest rate ***(crime)*** at a significance threshold of 1%, with the exception of the variable ***lneducm***, which is significant at 5%. According to this, the percentage of black population, crime codes, age range, and education years all have an impact on the arrest rate in the United States.

The result shows that the coefficient β1 has a value of -1.889. This indicates that every unit increase in the education years leads to 1.889 unit decrease in the arrest rate. Therefore, it can be implied that the arrest rate is less among the highly educated individuals group.

c) Compare IV and OLS results

The outcomes of IV and OLS are very comparable. The fit of both is nearly identical. All the non-instrumented variables are still significant, with roughly comparable coefficients. However, there is a difference in the degree of significance of average years of education (***lneducm***) on the arrest rate (***crime***) between the OLS and IV model results.

The average education years (***lneducm***) had a substantial influence on the arrest rate in the OLS model, with = 1 percent. However, when we used the 2SLS regression model, the average education years (***lneducm***) had a significance level of = 5%. This means that because we used percentage high school dropouts (***dropm***) as an instrument, we may have eliminated some bias between ***lneducm*** and the error.

There is also a change in the coefficient of lneducm and the blackm variables between OLS and IV. This also confirms that using dropm as an instrument has surely eliminated the bias with lneducm hence, causing the decrease in the coefficients. All other coefficients remain the same.

On concluding it, we can say that OLS is a better regression model than the IV 2SLS regression model.

d) Now consider the variable *work_age*: the minimum allowed working age in State i. *Discuss the validity as instrument and report and discuss the results. Compare with previous instrument. Which IV would you choose?*

For validating the variable work age as an instrument, we will implement the 2 stage least square, which requires 2 steps to be followed. Firstly, apply regression making average education year **(educm)** as a dependent variable and using legal work age **(work_age)** as an independent variable. Doing so will remove any correlation between ***educm*** and the error term. As a second step, we need to regress the crime rate over educm and other variables. The results can be seen in Fig 6. From Fig 5 we can see that the correlation between work_age and lneducm is 0.15. Once again it is not significant but shows some link between them. From Fig 6, we can see that the p-value is pretty low. This once again indicates that it is against the null hypothesis (H0: the instrument is valid). Thus we can say that the work_age as an instrument is weak.

```
. ivregress 2sls crime (lneducm=work_age) blackm rage off

Instrumental variables 2SLS regression          Number of obs   =      9,829
                                                Wald chi2(4)    =    2606.54
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.1156
                                                Root MSE        =     1.5302

------------------------------------------------------------------------------
       crime | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     lneducm |  -7.457785   1.478358    -5.04   0.000    -10.35531   -4.560257
      blackm |   .5637426   .2793033     2.02   0.044     .0163183    1.111167
        rage |  -.4702525   .0265586   -17.71   0.000    -.5223064   -.4181987
         off |    .008836   .0007488    11.80   0.000     .0073684    .0103036
       _cons |   30.15014   4.122084     7.31   0.000       22.071    38.22928
------------------------------------------------------------------------------
Instrumented: lneducm
 Instruments: blackm rage off work_age
```

*Fig 6. - 2SLS model (**work_age** as instrument)*

I argue that work age has an impact on crime rates because those who are incentivized to study are less likely to commit crimes. A high minimum work age, on the other hand, may contribute to greater crime rates since people who are idle at a young age are unable to earn an income and may seek money through illicit methods. In terms of validity, work age is more likely to be valid/ uncorrelated with 0 than dropout rate because, unlike dropout rate, which can be decided by background influences such as income, work age is an element of policy that is actively selected at a high level.

Despite being a completely distinct variable as ***dropm***, work age has many of the same problems.

If I had to choose, I would remain with OLS or attempt to discover another instrumental variable. When we compare the two instruments, we see that they are both incorrect and have some association with the error term. If I had to choose, I would go with the dropout rate because, while both are terrible instruments, ***dropm***, as demonstrated by the Stock-Yogo test, is at least a stronger instrument with more relation to the crime rate. This makes logical sense as well, because, unlike employment age, which has inconsistent consequences on crime, dropouts have a clearer understanding of increasing crime via inactivity.

Ques 2.1

a) What're the main concerns that might cause a biased estimation if we run a simple regression? And write down your model.

As we are interested in modelling the financial consequences of Brexit on aggregate trade value, i.e. trade value is the dependent variable, we will include two independent variables that indicate their financial influence (consider trade value as TV):
1. Total Production – represents Gross Domestic Product (GDP) - TP
2. Sterling Pound value - currency conversion rate – SPER

As the production also depends on the changes in demand on seasonal basis therefore we can take the demand (D) as an instrument for the production. (Note: ε is error)

Equation can be written as below:

$$TV = \beta_0 + \beta_1 TP * D + \beta_2 SPER + \varepsilon \qquad (1)$$

In this model, the economic shock of BREXIT is represented by the currency exchange rate changes post-Brexit and the change in the production due to changes and disruptions in the trade agreements from EU post-Brexit. We have also considered the production as an endogenous variable dependent on and instrument named **demand**.

However, some bias will be included in the error terms in this model. Many visible and non-observable factors will contribute to this biassed reaction. However, the impact of the COVID 19 epidemic on many businesses will have a substantial impact on the model's reaction. Because the pandemic outbreak occurred at the same time as Brexit, it disrupted global trade by causing a labour shortage and shipping delays. As a result, a thorough examination of its implications for UK commerce is required.

b) If the government wants to implement some policy against Brexit shock to support the economy how we could conduct a causal inference about how the shock affected the aggregate trade value?; which study method (DiD, IV, RDD, and etc.)

are you going to use, explain the reasons and limitation. And write down your model.

The difference-in-difference method is the greatest way to understand the impact of Brexit (D-i-D). This strategy is used when certain groups are exposed to a causative variable of interest (such as Brexit) while others are not. The exposed group is referred to as a treatment group, while the rest are referred to as control groups. The principle behind the D-i-D is to compare the two groups across two different time periods. Neither group is exposed during the first period, whereas one group is exposed while the other is not during the second period. Following that, the two groups' performance is compared, and the difference is assessed as the treatment effect. D-i-D is a transparent tool that may be used to investigate rapid changes in government or economic conditions.

To use this method in our case study, we used Brexit as the incidental variable and Value Trade as the dependent variable. The treatment group is the United Kingdom, while the control group might be any nation in the European Union. To choose our group control nation, we must look for a country with similar trade tendencies to the UK. A T-test must be performed between the EU nation and the UK GDP trend. The nation that backs up the hypothesis is referred to as the control country. There are two time periods to consider: before Brexit and after Brexit. For both eras, the trade value of both groups will be computed. The difference found after comparison is the Brexit effect.

$TV_{it} = \beta_0 + \beta_1 Treat_{it} + \beta_2 Post_{it} + \beta_3 Treat_{it} \times Post_{it} + \varepsilon_{it}$ (2)

Difference between the trade value of UK and the control group country post BREXIT will be calculated:

$E(TV_{it} \mid Post_{it} = 1, Treat_{it} = 1) - E(TV_{it} \mid Post_{it} = 1, Treat_{it} = 0) = \beta_1 + \beta_3$ (3)

Difference between the trade value of UK and the control group country pre BREXIT will be calculated:

$E(TV_{it} \mid Post_{it} = 1, Treat_{it} = 1) - E(TV_{it} \mid Post_{it} = 1, Treat_{it} = 0) = \beta_1$ (4)

After comparing the equations (3) and (4), it can be seen that the Brexit effect on value trade is β_3

D-i-D approach is based on two assumptions:

i) Control group and treatment group have parallel trend prior to the Brexit shock. And this would have continued in the similar fashion if there has been no intervention

ii) The Brexit shcok would have impacted both the groups equally.

c) If the government want to investigate the effect on different sectors which study method (DiD, IV, RDD, and etc.) are you going to use, explain the reasons and

describe how you are going to proceed with the analysis. And write down your model.

To analyse the impact of Brexit on various industries, we must apply the D-i-D approach to several treatment and control groups. In our scenario, the treatment groups will be different sectors in the UK to be studied, whereas the control groups will be the same sectors in various EU nations.

Matching approaches employing covariates may be used to match the treatment subject, i.e. a specific sector to be investigated, with the corresponding control subjects based on pre-treatment characteristics of a given sector. This ensures that the pre-treatment groups have similar patterns.

Such model can be defined as below:

$Y_{igt} = \alpha_g + \mu_t + \gamma I_{gt} + \theta Z_{igt} + \S_{igt}$ (5)

Y_{igt} --- Response value of a sector
α_g --- Group fixed effects
μ_t --- Time fixed effects
I_{gt} --- Treatment dummy
θZ_{igt} --- Control dummy
$\S_{igt}$ - Error

Appendix

[1]

```
. reg crime lneducm blackm rage i.off, cluster (state)

Linear regression                               Number of obs     =     10,458
                                                F(10, 50)         =    1611.80
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8089
                                                Root MSE          =     .71647

                                  (Std. err. adjusted for 51 clusters in state)
-------------------------------------------------------------------------------
             |               Robust
       crime | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+-----------------------------------------------------------------
     lneducm |  -.6593943   .2105061    -3.13   0.003    -1.082208   -.2365805
      blackm |   2.064106   .2835678     7.28   0.000     1.494543    2.633669
        rage |  -.3716213   .0086517   -42.95   0.000    -.3889987   -.3542438
             |
         off |
         20  |   .3367215   .0612253     5.50   0.000     .2137468    .4596962
         30  |   1.085955   .0607502    17.88   0.000     .9639342    1.207975
         40  |    2.39588   .0704617    34.00   0.000     2.254353    2.537406
         50  |    2.06303   .0543846    37.93   0.000     1.953795    2.172264
         60  |   3.416471   .0719637    47.47   0.000     3.271928    3.561014
         70  |   .8801451   .0645996    13.62   0.000     .7503929    1.009897
         90  |  -.6408688   .0694171    -9.23   0.000    -.7802971   -.5014404
             |
       _cons |   10.47811   .6145618    17.05   0.000     9.243724    11.71249
-------------------------------------------------------------------------------
```