Instrumental Variables analysis

Cerny David

2022-12-06

Analysis of course Fundamentals of statistics

We have cross-sectional data about Fundamental of Statistics (FoS course) 2018 collected among students in Uzbekistan. Data are stored as `.xlsx` file. The legend related to this data could be found in file `Codebook.pdf`. The goal of this analysis is to inspect, whether there is a relationship between number of attended seminars and final grades.

Data

Firstly, we wil import packages and the data. We can see that NA data are present in the dataset, but we will follow carefuly, since we have just 101 observation - if we remove all NA obs we would have just a few obs left. Notice 101 is approximately minimum required number of observation for assymptotic features to hold.

```
# import packages
library("readxl")
library("dplyr")
library("lmtest")
library("systemfit")
library("ivreg")
library("tseries")
library("stargazer")

FOS = read_excel("dataForStudents.xlsx")
str(FOS)
```

```
## tibble [101 x 19] (S3: tbl_df/tbl/data.frame)
##  $ id2         : num [1:101] 1 2 3 4 5 6 7 8 9 10 ...
##  $ nlesson     : num [1:101] 6 13 9 13 23 3 3 8 22 9 ...
##  $ nfile       : num [1:101] 34 28 14 31 122 41 21 73 117 44 ...
##  $ course      : chr [1:101] "BSc (Hons) in Economics" "BSc (Hons) in Business Information Systems" "
##  $ ent_math    : num [1:101] 40 53 63 NA 51 NA 89 73 67 54 ...
##  $ ent_ielts   : num [1:101] 6 7 44686 44686 6 ...
##  $ course_c    : chr [1:101] "BSc (Hons) in Economics" "BSc (Hons) in Business Information Systems" "
##  $ gender      : chr [1:101] NA "Male" "Male" "Male" ...
##  $ age         : num [1:101] NA 19 20 21 20 21 20 21 20 19 ...
##  $ ethnicity   : chr [1:101] NA "Uzbek" "Others" "Uzbek" ...
##  $ with_parents: chr [1:101] NA "Yes" "Yes" "Yes" ...
##  $ married     : chr [1:101] NA "Not Married" "Not Married" "Not Married" ...
##  $ working     : chr [1:101] NA "Part time" "Full time" "Part time" ...
##  $ work_travel : chr [1:101] NA "No" "No" "No" ...
```

```
##  $ chapters    : num [1:101] NA 0 10 5 NA NA 0 0 4 5 ...
##  $ mark_e1_FoS : num [1:101] 47 85 48 20 49 28 64 47 86 30 ...
##  $ mark_e2_FoS : num [1:101] 44 86 33 35 64 49 48 28 94 35 ...
##  $ mark_p_FoS  : num [1:101] 69 84 75 59 80 71 80 78 73 55 ...
##  $ mark_t_FoS  : num [1:101] 50 85 47 34 61 45 61 46 87 37 ...
```

```
summary(FOS)
```

```
##       id2         nlesson          nfile           course
##  Min.   :  1   Min.   : 1.00   Min.   :  1.00   Length:101
##  1st Qu.: 26   1st Qu.: 6.00   1st Qu.: 32.00   Class :character
##  Median : 51   Median :11.00   Median : 59.00   Mode  :character
##  Mean   : 51   Mean   :10.88   Mean   : 66.37
##  3rd Qu.: 76   3rd Qu.:15.00   3rd Qu.: 90.00
##  Max.   :101   Max.   :23.00   Max.   :164.00
##
##     ent_math       ent_ielts       course_c            gender
##  Min.   :40.00   Min.   :    6   Length:101         Length:101
##  1st Qu.:49.00   1st Qu.:    6   Class :character   Class :character
##  Median :56.00   Median :    7   Mode  :character   Mode  :character
##  Mean   :57.61   Mean   :21683
##  3rd Qu.:67.00   3rd Qu.:44686
##  Max.   :91.00   Max.   :44687
##  NA's   :5
##       age          ethnicity         with_parents         married
##  Min.   :19.00   Length:101         Length:101         Length:101
##  1st Qu.:20.00   Class :character   Class :character   Class :character
##  Median :20.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :20.43
##  3rd Qu.:21.00
##  Max.   :30.00
##  NA's   :11
##    working          work_travel           chapters       mark_e1_FoS
##  Length:101         Length:101         Min.   : 0.000   Min.   : 8.00
##  Class :character   Class :character   1st Qu.: 1.000   1st Qu.:30.00
##  Mode  :character   Mode  :character   Median : 4.000   Median :43.00
##                                        Mean   : 3.253   Mean   :45.09
##                                        3rd Qu.: 5.000   3rd Qu.:56.00
##                                        Max.   :12.000   Max.   :86.00
##                                        NA's   :22
##   mark_e2_FoS      mark_p_FoS     mark_t_FoS
##  Min.   : 7.00   Min.   :43.0   Min.   :28.00
##  1st Qu.:40.00   1st Qu.:63.0   1st Qu.:44.00
##  Median :52.00   Median :71.0   Median :50.00
##  Mean   :52.84   Mean   :70.2   Mean   :53.18
##  3rd Qu.:64.00   3rd Qu.:80.0   3rd Qu.:61.00
##  Max.   :94.00   Max.   :93.0   Max.   :87.00
##
```

Which variables might affect students performance? **nfile** - number of files given student downloaded. With more downloaded files student might perform better at the final exam.

nlesson - number of seminars given student attended. The seminar attendence might positively affect the performance.

ent_math - entrance score in math. The statistics could be considered as applied math, therefore the math skill could be correlated with stats skill.

Are there any factors, which could affect seminar attendance? **working** - whether given student worked during the semestr. If student works, he could be more busy and therefore could pass the seminar from time to time.

Inspecting the data at hand

We will inspect the data at hand to find some usefull relationships or outliers.

Could working affect the score and attendence? There seems to be an impact of working on the seminar attendence and final grade. The more student works, the less seminar he on average attend and the worse grade he on average obtain.

```
work_fultime = FOS %>%
  filter(working == "Full time") %>%
  summarise(mean(mark_t_FoS), mean(nlesson))

work_parttime = FOS %>%
  filter(working == "Part time") %>%
  summarise(mean(mark_t_FoS), mean(nlesson))

workn = FOS %>%
  filter(working == "No") %>%
  summarise(mean(mark_t_FoS), mean(nlesson))

working = FOS %>%
  filter(working != "No") %>%
  summarise(mean(mark_t_FoS), mean(nlesson))

work_effect = as.matrix(rbind(work_fultime, work_parttime, working, workn))
rownames(work_effect) = c("Full time", "Part time", "Job", "No job")
work_effect
```

```
##           mean(mark_t_FoS) mean(nlesson)
## Full time         46.71429      7.428571
## Part time         50.20930      9.813953
## Job               49.72000      9.480000
## No job            56.97500     12.100000
```

To catch the work effect we will create three dummy variables: **full_time**, **part_time**, **job**

$$full\_time = \begin{cases} 1 & \text{if } working = \ Full\ time \\ 0 & \text{otherwise} \end{cases}$$

$$part\_time = \begin{cases} 1 & \text{if } working = \ Part\ time \\ 0 & \text{otherwise} \end{cases}$$

$$job = \begin{cases} 1 & \text{if } working \neq No \\ 0 & \text{otherwise} \end{cases}$$

```
FOS$full_time =0
FOS$full_time[FOS$working == "Full time"] = 1
FOS$part_time = 0
FOS$part_time[FOS$working == "Part time"] = 1
FOS$job = 0
FOS$job[FOS$working != "No"] = 1
```

Simple Linear Regression model

Firstly, we will estimate the simple linear regression model. The dependent variable is final score, the explanatory variables are number of lessons attended and number of files downloaded. Both explanatory variables are statisticaly significant, especially nlesson. $R^2 = 0.29$, which is definitely not bad

```
simple_model = lm(mark_t_FoS ~ nlesson + nfile, data = FOS)
summary(simple_model)
```

```
##
## Call:
## lm(formula = mark_t_FoS ~ nlesson + nfile, data = FOS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.616  -8.224  -0.728   8.388  32.482
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.17229    2.65394  14.383  < 2e-16 ***
## nlesson      0.95304    0.21697   4.392 2.84e-05 ***
## nfile        0.06985    0.03089   2.262   0.0259 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.88 on 98 degrees of freedom
## Multiple R-squared:  0.2909, Adjusted R-squared:  0.2764
## F-statistic:  20.1 on 2 and 98 DF,  p-value: 4.832e-08
```

Since entry level of math could also affext final grade, we will add it as an explanatory variable. As we can see the ent_math is significant at 0.01 significance level and the R^2 adjusted, also increased (Note that the increase in R^2 does not imply improvement of the model, because the R^2 always increase, when we add another explanatory variable, therefore we use adjusted R^2, which punish us for the unwanted complexity, as an indicator.)

```
lmmath = lm(mark_t_FoS ~ nlesson + nfile + ent_math, data = FOS)
summary(lmmath)
```

```
##
## Call:
```

```
## lm(formula = mark_t_FoS ~ nlesson + nfile + ent_math, data = FOS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.026  -8.220  -0.184   8.863  33.210
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.18820    6.29738   3.365  0.00112 **
## nlesson      1.07672    0.20762   5.186 1.27e-06 ***
## nfile        0.07479    0.02953   2.532  0.01302 *
## ent_math     0.27378    0.09684   2.827  0.00576 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.17 on 92 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.387,  Adjusted R-squared:  0.367
## F-statistic: 19.36 on 3 and 92 DF,  p-value: 8.152e-10
```

```
stargazer(simple_model, lmmath, header=FALSE,
          title='Simple Linear Regression model',
          single.row=TRUE, type='text')
```

```
##
## Simple Linear Regression model
## ================================================================
##                                 Dependent variable:
##                     ---------------------------------------------
##                                       mark_t_FoS
##                               (1)                    (2)
## ----------------------------------------------------------------
## nlesson                 0.953*** (0.217)       1.077*** (0.208)
## nfile                   0.070** (0.031)        0.075** (0.030)
## ent_math                                       0.274*** (0.097)
## Constant               38.172*** (2.654)      21.188*** (6.297)
## ----------------------------------------------------------------
## Observations                101                     96
## R2                         0.291                  0.387
## Adjusted R2                0.276                  0.367
## Residual Std. Error    11.882 (df = 98)       11.174 (df = 92)
## F Statistic         20.103*** (df = 2; 98) 19.361*** (df = 3; 92)
## ================================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

Are all explanatory variables exogenous? One of the crucial assumtion of OLS estimator is Strict Exogeneity. $\mathbb{E}[u_t|X] \neq 0$. When this assumption is violated, the OLS estimater cannot be unbiased. The less stricter assumption is Contemporaneous Exogeneity. $\mathbb{E}[u_t|X_t] \neq 0$. This is assumption of asymptotic process and if is violated the OLS estimator can no longer be either unbiased or consistent.

The problem arises when $Cov(x, u) \neq 0 \implies \mathbb{E}[u_t|X_t] \neq 0$. The explanatory variable is in this case endogeneos. There are several solution. We can use Proxy Model, Two-staged Least squares(TSLS) or Instrumental variable(special case of TSLS, when just one IV is used).

In our model we could suspect the nlesson variable to be endogenoues(see above). If it is so we should use different model.

Inspect nlesson Whether explanatory variable is exogenous or not cannot be straightforwardly tested. Firstly we will try to check, whether correlation between nlesson and working is present in the dataset. We will use correlation matrix to test it. When we will focus on the job dummy, we can see that there is negative correlation between having job and seminar attendance.

```
cormat = cor(cbind(FOS$nlesson, FOS$full_time, FOS$part_time, FOS$job))
labels = c("nlesson", "full_time", "part_time","job")
colnames(cormat) = labels
rownames(cormat) = labels
cormat
```

```
##              nlesson  full_time  part_time        job
## nlesson    1.0000000 -0.1563738 -0.1525143 -0.2302644
## full_time -0.1563738  1.0000000 -0.2349662  0.2756038
## part_time -0.1525143 -0.2349662  1.0000000  0.8696016
## job       -0.2302644  0.2756038  0.8696016  1.0000000
```

Instrumental variable

We observed correlation between job and nlesson in our data, therefore we decided to use IV to estimate robust model. We will estimate IV regression model with job as the instrumental variable. But before it we should test, whether the job dummy is suitable for IV model.

The instrumental variable have to meet two assumptions. Instrumental variable z should be uncorrelated with disturbances u: $Cov(z, u) = 0$ and z should be correlated with the endogeneous variable x: $Cov(z, x) \neq 0$. When both assumption are met we can estimate new coefficient $\beta_1 = \frac{Cov(z,y)}{Cov(z,x)}$. Since the exogeneity cannot be easily tested, we will at least test correlation with the endogeneos variable.

We will estimate model:

$x_i = \pi_0 + \pi_1 z_i + \gamma$.

$Cov(z, x) = 0$ only if $\pi_0 = 0$, which means when z is statistically significant.

As we can see z is significant and 95% significance level, there for we can reject the null hypothesis $H_0 : \pi_0 = 0$.

```
cov_model = lm(nlesson ~ job, data = FOS)
summary(cov_model)
```

```
##
## Call:
## lm(formula = nlesson ~ job, data = FOS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.255  -5.255   0.520   4.520  12.520
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.2549     0.8293  14.778   <2e-16 ***
## job          -2.7749     1.1786  -2.354   0.0205 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.922 on 99 degrees of freedom
## Multiple R-squared:  0.05302,    Adjusted R-squared:  0.04346
## F-statistic: 5.543 on 1 and 99 DF,  p-value: 0.02053
```

Now we will assume exogeneity of job and use it as IV. The procedure of IV model is following:

1. We have the simple regression model with endogeneous variable *nlesson*:

$$mark\_t\_FoS_i = \beta_0 + \beta_1 nlesson_i + \beta_2 nfiles_i + \beta_3 ent\_math + u_i$$

2. We will use IV *job*. From the first equation and IV assumptions above we can derive:

$$Cov(job, mark\_t\_FoS) = \beta_1 Cov(job, mark\_t\_FoS) + Cov(job, u) \implies \beta_1 = \frac{Cov(job, mark\_t\_Fos)}{Cov(job, nlesson)}$$

3. From the LLN:

$$\hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{n}(job_i - \overline{job})(mark\_t\_FoS - \overline{mark\_t\_FoS})}{\sum_{i=1}^{n}(job_i - \overline{job})(nlesson - \overline{nlesson})}$$

In R there is no need to derive these formulas, we can just use `tsls` or `ivreg` functions.

As we can see the newly estimated model results are worse than the results of the simple linear regression model, but the model should be unbiased and consistent. The R^2 is much worse than in the initial model and last but not least the only significant variable is *ent_math*.

```
IVreg = ivreg(mark_t_FoS ~ nlesson + nfile + ent_math | job + nfile + ent_math, data = FOS)
summary(IVreg)
```

```
##
## Call:
## ivreg(formula = mark_t_FoS ~ nlesson + nfile + ent_math | job +
##     nfile + ent_math, data = FOS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -32.3182  -7.9801   0.9772   9.3454  28.5323
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.289257  15.783940   0.589   0.5576
## nlesson      2.356231   1.514528   1.556   0.1232
## nfile        0.000311   0.093796   0.003   0.9974
## ent_math     0.323976   0.129175   2.508   0.0139 *
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1  92     2.509   0.117
## Wu-Hausman         1  91     1.036   0.311
## Sargan             0  NA        NA      NA
##
## Residual standard error: 13.28 on 92 degrees of freedom
## Multiple R-Squared: 0.134,   Adjusted R-squared: 0.1057
## Wald test: 8.166 on 3 and 92 DF,  p-value: 7.061e-05
```

Homoscedasticity testing

In order to obtain valid statistical inference and valit t and F tests, we need to assume homoscedasticity:

$Var(u) = \mathbb{E}[u^2|z] = \sigma^2$, in other words the variance of disturbances must be constant across the whole sample.

There are several approaches how to test homoscedasticity, but we decided to use **Goldfeld-Quandt Test** and **Breusch-Pagan test**. We will use `gqtest` and `bptest` functions from the package `lmtest`. H_0 : there is homoscedasticity present, H_A : variance increased between segments.

According the test results we cannot reject the H_0 for both models at any reasonoble level.

```
gqtest(IVreg)
```

```
##
##  Goldfeld-Quandt test
##
## data:  IVreg
## GQ = 0.98668, df1 = 44, df2 = 44, p-value = 0.5176
## alternative hypothesis: variance increases from segment 1 to 2
```

```
bptest(lmmath)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lmmath
## BP = 6.2571, df = 3, p-value = 0.09975
```

Test for endogeneity

In order to decide which model we will use, we will finally test endogeneity with **Hausman test**. H_0 : OLS and IV are both consistent, H_A : OLS is inconsistent and IV is consistent. We will use `hausman.systemfit`function from the `systemfit` package.

From the results of hausman test we cannot reject the H_0, therefore it is better stick with the simple linear regression model, which would be asiptotically better.

```
hausman.systemfit(IVreg, lmmath)
```

```
##
##  Hausman specification test for consistency of the 3SLS estimation
##
## data:  FOS
## Hausman = 0.7274, df = 4, p-value = 0.9479
```

```
stargazer(IVreg, lmmath, header=FALSE,
          title='IVreg vs lmmath',
          single.row=TRUE, type='text')
```

```
##
## IVreg vs lmmath
## ===============================================================
##                                     Dependent variable:
##                                 ------------------------------
##                                           mark_t_FoS
##                                  instrumental          OLS
##                                    variable
##                                       (1)               (2)
## ---------------------------------------------------------------
## nlesson                            2.356 (1.515)     1.077*** (0.208)
## nfile                             0.0003 (0.094)     0.075** (0.030)
## ent_math                          0.324** (0.129)    0.274*** (0.097)
## Constant                          9.289 (15.784)    21.188*** (6.297)
## ---------------------------------------------------------------
## Observations                          96                 96
## R2                                   0.134              0.387
## Adjusted R2                          0.106              0.367
## Residual Std. Error (df = 92)        13.281             11.174
## F Statistic                                    19.361*** (df = 3; 92)
## ===============================================================
## Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

OLS assumptions After the model testing we decided to use the simple OLS model:

$mark\_t\_FoS_i = \beta_0 + \beta_1 nlesson_i + \beta_2 nfiles_i + \beta_3 ent\_math + u_i$

After the testing above we can conclude that OLS1-5 holds:

OLS1 - Linear in Parameter (holds)

OLS2 - Random sampling (according to characteristics of data collection, shoul hold)

OLS3 - No perfect Colinearity (holds)

OLS4 - Zero Conditional Mean (we cannot reject it)

OLS5 - Homoscedasticity (we cannot reject it)

We need to test **OLS6** - Normality.

We will use `jarque.bera.test` from `tseries` package, with H_0 : Residuals are normaly distributed.

With $p\_value = 0.5962$ we cannot reject normality, thus OLS 1-6 are met, which implies, that t and F statistics are valid and our OLS estimator is **BLUE**(Best linear unbiased estimator).

```
res = lmmath$residuals
jarque.bera.test(res)
```

```
##
##  Jarque Bera Test
##
## data:  res
## X-squared = 1.4432, df = 2, p-value = 0.486
```

```
summary(lmmath)
```

```
##
## Call:
## lm(formula = mark_t_FoS ~ nlesson + nfile + ent_math, data = FOS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.026  -8.220  -0.184   8.863  33.210
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.18820    6.29738   3.365  0.00112 **
## nlesson      1.07672    0.20762   5.186 1.27e-06 ***
## nfile        0.07479    0.02953   2.532  0.01302 *
## ent_math     0.27378    0.09684   2.827  0.00576 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.17 on 92 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.387,  Adjusted R-squared:  0.367
## F-statistic: 19.36 on 3 and 92 DF,  p-value: 8.152e-10
```

Conclusion

After the several testing, we have chosed the Simple Linear Regression Model. The model met all OLS assumption, therefore is **BLUE**. When we look at results we can see, that all explanatory variables are statistically significant and have a positive affect on the final grade. Our goal was to inspect relationship between the number of lessons attended and the final grades. The increase in seminar attendance by one should cateris paribus increase final grade by one point (almost perfect linear relationship).