The University of Queensland
Australia
CREATE CHANGE

ECON2300 - Introductory Econometrics

Tutorial 11: Experiments and Quasi-Experiments

Tutor: Francisco Tavares Garcia

Quiz 5 is available!

Posted on: Monday, 16 October 2023 06:00:00 o'clock AEST

Dear ECON2300 Students,

Quiz 5 is now available in the "Quizzes: Problem Solving Exercises" folder, which you can access via the Assessment tab.

The due date for Quiz 5 is **Friday, October 20, 2023, 4pm**

Please read all instructions carefully before commencing the quiz. For convenience, a copy of the quiz instructions has been presented below.

==

Instructions:

Please pay close attention to the number of decimal places required for each answer. The required number of decimal places may differ from question to question.

Avoid rounding during intermediate calculations where possible.

Some answers in this quiz will involve percentages or dollar figures. **Do not include a % or $ sign in any of your answers.** For example, if the answer is 12%, just enter 12

The quiz is not timed. This means that you can open the quiz and return to it as many times as you need to (provided that you do not click submit).

There is only one attempt for this quiz.

The quiz is marked out of 7, but will contribute 10% towards your final grade if it is among the highest 3 of your 5 Quiz scores across the semester.

The closing time for this quiz is **4pm on Friday, October 20, 2023**. Please make sure that you have submitted your answers by this time. Remember that you need to click submit before the deadline for your quiz to be marked.

Please Note: If you encounter any technical issues with the quiz, please email the CML coordinator at cml.2300@uq.edu.au. Do not email quiz issues to the Course Coordinator or Course Administrator. Otherwise there may be a delay in responding to your enquiry.

Report 2 is available!

ECON 2300: INTRODUCTORY ECONOMETRICS
Coordinator: Professor Rodney Strachan
Research Project 2
Due: 4 pm, 6 November

Submission of your report

Your report must be single-spaced and in 12 Font size. You should give your answer to each of the following questions following a similar format of the solutions to the tutorial problem sets. When you are required to use R, you must show your R command and R `outputs` (screenshots or figures generated from R). You will lose **2 points** whenever you fail to provide R commands and `outputs`. For each question, when you are asked to discuss or interpret, your answer has to be brief and compact. You will lose **2 points** if your answer is needlessly wordy. You must upload your assignment on the course webpage (Blackboard) in PDF format. (Do not submit a hard copy.)

This project has two research questions. You are required to investigate both of them.

Problem 1: money, Growth, and Inflation (30 marks)

Background

To examine the quantity theory of `money`, Brumm (2005) ["Money Growth, Output Growth, and Inflation: A Reexamination of the Modern Quantity Theory's Linchpin Prediction," *Southern Economic Journal*, 71(3), 661–667] specifies the inflation equation

$$\texttt{inflat} = \beta_1 + \beta_2\texttt{money} + \beta_3\texttt{output} + u$$

SETutor is available!!!

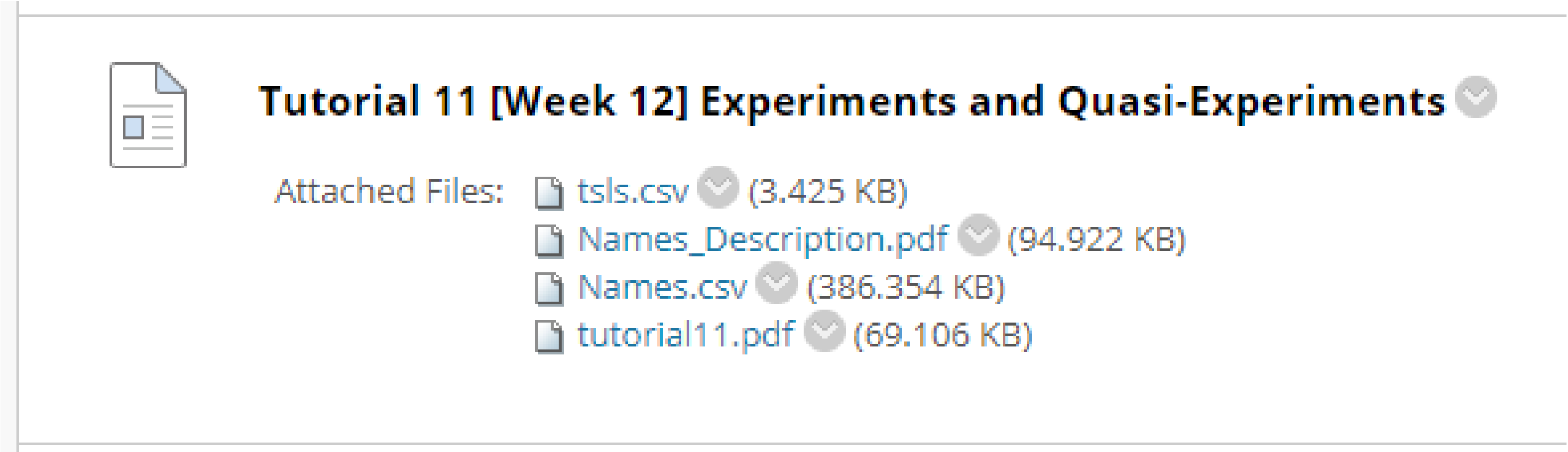
If you found these tutorials helpful, please answer the survey.

(If you didn’t, please let me know how to improve them through the survey too ☺)

This is very valuable for us tutors!

https://eval.uq.edu.au/eus.onlinesurveyportal/Home/Survey?surveyid=768118861

- Download the files for tutorial 11 from Blackboard,
- save them into a folder for this tutorial.



- Copy the code from Codeshare,
- https://codeshare.io/tut11
- Paste the code in a new script in RStudio,
- Save the script in the same folder as the data.

E13.1 A prospective employer receives two resumes: a resume from a white job applicant and a similar resume from an African American applicant. Is the employer more likely to call back the white applicant to arrange an interview? Marianne Bertrand and Sendhil Mullainathan carried out a randomized controlled experiment to answer this question. Because race is not typically included on a resume, they differentiated resumes on the basis of "white-sounding names" (such as Emily Walsh or Gregory Baker) and "African American-sounding names" (such as Lakisha Washington or Jamal Jones). A large collection of fictitious resumes was created, and the presupposed "race" (based on the "sound" of the name) was randomly assigned to each resume. These resumes were sent to prospective employers to see which resumes generated a phone call (a "call back") from the prospective employer. Use the data file `Names.csv` to answer the following questions. See `Names_Description.pdf` for more details about the data.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Variable Descriptions

Variable Name	Description
Key Variables	
firstname	applicant's first name
female	1 = female
black	1 = black
high	1= high quality resume
call_back	1= applicant was called back
chicago	1 = data from Chicago
Detailed Information on Resume	
ofjobs	number of jobs listed on resume
yearsexp	number of years of work experience on the resume
honors	1=resume mentions some honors
volunteer	1=resume mentions some volunteering experience
military	1=applicant has some military experience
empholes	1=resume has some employment holes
workinschool	1=resume mentions some work experience while at school
email	1=email address on applicant's resume
computerskills	1=resume mentions some computer skills
specialskills	1=resume mentions some special skills
college	applicant has college degree or more

(a) Define the "call-back rate" as the fraction of resumes that generate a phone call from the prospective employer. What was the call-back rate for whites? For African Americans? Construct a 95% confidence interval for the difference in the call-back rates. Is the difference statistically significant? Is it large in a real-world sense?

```
> summary(reg1)

Call:
lm_robust(formula = call_back ~ black, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
(Intercept)  0.09651   0.005985  16.124 5.045e-57  0.08478  0.10824 4868
black       -0.03203   0.007785  -4.115 3.941e-05 -0.04729 -0.01677 4868

Multiple R-squared:  0.003466 , Adjusted R-squared:  0.003261
F-statistic: 16.93 on 1 and 4868 DF,  p-value: 3.941e-05
```

Table 1: Race and Resume Call-Back Rate

	(1)	(2)	(3)	(4)
(Intercept)	0.0965***	0.0965***	0.0734***	0.0850***
	(0.0060)	(0.0060)	(0.0053)	(0.0080)
black	−0.0320***	−0.0382**		−0.0231*
	(0.0078)	(0.0117)		(0.0106)
female.black		0.0080		
		(0.0115)		
high			0.0141	0.0229
			(0.0078)	(0.0120)
high.black				−0.0178
				(0.0156)
R^2	0.0035	0.0035	0.0007	0.0044
Adj. R^2	0.0033	0.0031	0.0005	0.0038
Num. obs.	4870	4870	4870	4870
RMSE	0.2716	0.2717	0.2720	0.2716

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

From (1) in the table, the call-back rate for whites is 0.0965 and the call-back rate for blacks is $0.0965 - 0.032 = 0.0645$. The difference is -0.032 is statistically significant at the 1% level (t-statistic = -4.11). This result implies that 9.65% of resumes with white-sounding names generated a call back. Only 6.45% of resumes with black-sounding names generated a call back. The difference is large in both statistical and economic sense.

(b) Is the African American/white call-back rate differential different for men than for women?

```
> summary(reg2)

Call:
lm_robust(formula = call_back ~ black + female.black, se_type = "stata")

Standard error type:  HC1

Coefficients:
             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   0.09651   0.005986 16.1227 5.178e-57  0.08477  0.10824 4867
black        -0.03822   0.011657 -3.2790 1.049e-03 -0.06107 -0.01537 4867
female.black  0.00799   0.011527  0.6931 4.883e-01 -0.01461  0.03059 4867

Multiple R-squared:  0.003541 , Adjusted R-squared:  0.003132
F-statistic: 8.805 on 2 and 4867 DF,  p-value: 0.0001524
```

Table 1: Race and Resume Call-Back Rate

	(1)	(2)	(3)	(4)
(Intercept)	0.0965***	0.0965***	0.0734***	0.0850***
	(0.0060)	(0.0060)	(0.0053)	(0.0080)
black	−0.0320***	−0.0382**		−0.0231*
	(0.0078)	(0.0117)		(0.0106)
female.black		0.0080		
		(0.0115)		
high			0.0141	0.0229
			(0.0078)	(0.0120)
high.black				−0.0178
				(0.0156)
R^2	0.0035	0.0035	0.0007	0.0044
Adj. R^2	0.0033	0.0031	0.0005	0.0038
Num. obs.	4870	4870	4870	4870
RMSE	0.2716	0.2717	0.2720	0.2716

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

From (2) in the table, the call-back rate for male blacks $0.0965 - 0.0382 = 0.0583$, and for female blacks is $0.0965 - 0.0382 + 0.008 = 0.0663$. The difference is 0.008, which is not significant at the 5% level (t-statistic $= 0.69$).

(c) What is the difference in call-back rates for high-quality versus low-quality resumes? What is the high-quality/low-quality difference for white applicants? For African American applicants? Is there a significant difference in this high-quality/low-quality difference for whites versus African Americans?

```
Call:
lm_robust(formula = call_back ~ high, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)  CI Lower CI Upper   DF
(Intercept)  0.07343   0.005299  13.857 7.404e-43  0.063044  0.08382 4868
high         0.01406   0.007793   1.804 7.132e-02 -0.001221  0.02934 4868

Multiple R-squared:  0.0006675 ,        Adjusted R-squared:  0.0004622
F-statistic: 3.254 on 1 and 4868 DF,  p-value: 0.07132
> summary(reg4)

Call:
lm_robust(formula = call_back ~ black + high + high.black, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)  CI Lower  CI Upper   DF
(Intercept)  0.08498   0.008013  10.605 5.397e-26  0.069274  0.100693 4866
black       -0.02310   0.010590  -2.182 2.919e-02 -0.043864 -0.002341 4866
high         0.02295   0.011958   1.919 5.505e-02 -0.000496  0.046392 4866
high.black  -0.01778   0.015561  -1.143 2.532e-01 -0.048286  0.012725 4866

Multiple R-squared:  0.0044 ,   Adjusted R-squared:  0.003787
F-statistic: 6.613 on 3 and 4866 DF,  p-value: 0.0001868
```

Table 1: Race and Resume Call-Back Rate

	(1)	(2)	(3)	(4)
(Intercept)	0.0965***	0.0965***	0.0734***	0.0850***
	(0.0060)	(0.0060)	(0.0053)	(0.0080)
black	−0.0320***	−0.0382**		−0.0231*
	(0.0078)	(0.0117)		(0.0106)
female.black		0.0080		
		(0.0115)		
high			0.0141	0.0229
			(0.0078)	(0.0120)
high.black				−0.0178
				(0.0156)
R^2	0.0035	0.0035	0.0007	0.0044
Adj. R^2	0.0033	0.0031	0.0005	0.0038
Num. obs.	4870	4870	4870	4870
RMSE	0.2716	0.2717	0.2720	0.2716

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

From (3) in the table, the call-back rate for low-quality resumes is 0.0734 and the call-back rate for high-quality resumes is $0.0734 + 0.0141 = 0.0875$. The difference is 0.0141, which is not significant at the 5% level, but is at the 10% level (p-value $= 0.071$). From (4) the (high-quality)-(low-quality) difference for whites is 0.0229 and for blacks is $0.0229 - 0.0178 = 0.0051$; the black-white difference is -0.0178 which is not statistically significant at the 5% level (t-statistic = -1.14).

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

(d) The authors of the study claim that race was assigned randomly to the resumes. Is there any evidence of nonrandom assignment?

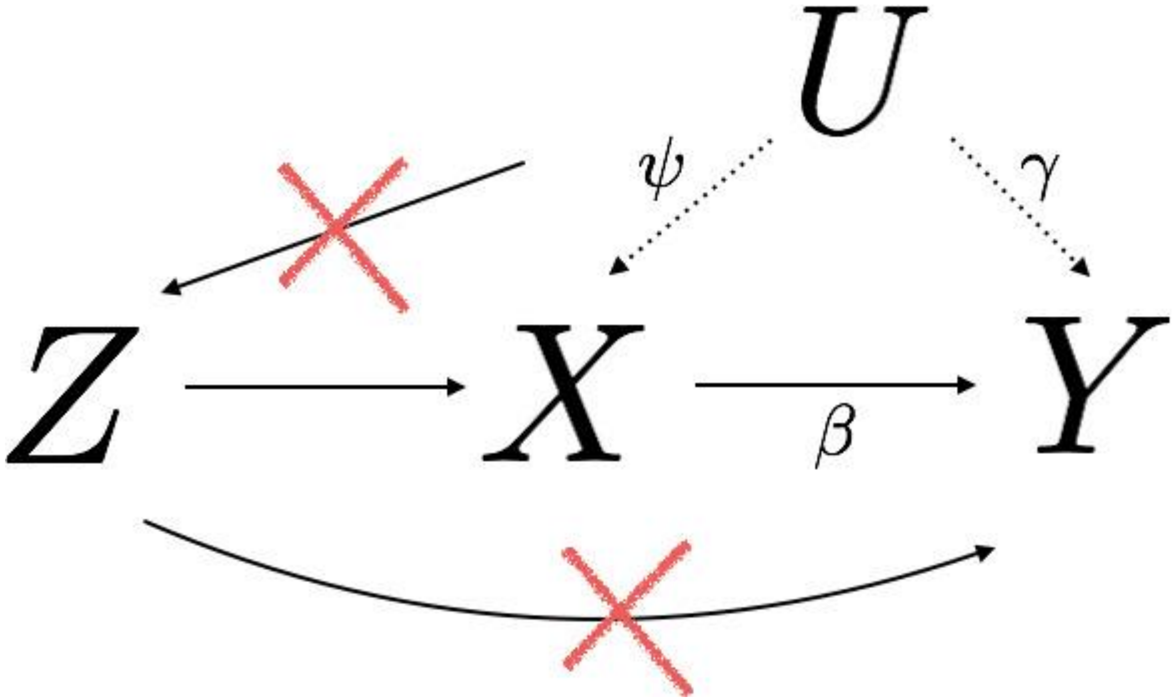
```
> Tests = lm_robust(cbind(ofjobs, yearsexp, honors, volunteer, military, empholes,
+                  workinschool, email, computerskills, specialskills, eoe, manager,
+                  supervisor, secretary, offsupport, salesrep,
+                  retailsales, req, expreq, comreq, educreq, compreq, orgreq,
+                  manuf, transcom, bankreal, trade, busservice, othservice,
+                  missind, chicago, high, female, college, call_back) ~ black,
+               se_type = "stata")
> tidy(Tests)
            term      estimate   std.error     statistic       p.value     conf.low    conf.high   df        outcome
13 (Intercept)  5.581109e-01 0.010065993  5.544519e+01  0.000000e+00  0.538376993  0.577844773 4868   workinschool
14       black  2.874743e-03 0.014230522  2.020125e-01  8.399154e-01 -0.025023504  0.030772990 4868   workinschool
15 (Intercept)  4.788501e-01 0.010125602  4.729103e+01  0.000000e+00  0.458999353  0.498700853 4868          email
16       black  8.213552e-04 0.014320252  5.735620e-02  9.542638e-01 -0.027252803  0.028895513 4868          email
17 (Intercept)  8.086242e-01 0.007973639  1.014122e+02  0.000000e+00  0.792992298  0.824256162 4868 computerskills
18       black  2.381930e-02 0.010994740  2.166427e+00  3.032693e-02  0.002264649  0.045373955 4868 computerskills
19 (Intercept)  3.301848e-01 0.009532257  3.463868e+01 3.283966e-235  0.311497278  0.348872332 4868  specialskills
20       black -2.874743e-03 0.013465635 -2.134874e-01  8.309558e-01 -0.029273467  0.023523980 4868  specialskills
21 (Intercept)  2.911704e-01 0.009208402  3.162008e+01 9.531045e-200  0.273117807  0.309223056 4868            eoe
22       black -2.163638e-16 0.013022647 -1.661443e-14  1.000000e+00 -0.025530266  0.025530266 4868            eoe
```

Results of a series of t-tests (via linear regressions, see the log-file) shows estimated means of other characteristics for black and white sounding names. There are only two significant differences in the mean values: the call-back rate (the variable of interest) and computer skills (for which black-named resumes had a slightly higher fraction than white-named resumes). Thus, there is no evidence of non-random assignment.

TSLS In this question, we fit the following regression model to the data `tsls.csv`

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \tag{1}$$

We are interested in studying the causal effect of X_2 on Y, i.e., β_2.



DOI: https://doi.org/10.1145/3178876.3186151

TSLS In this question, we fit the following regression model to the data `tsls.csv`

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \tag{1}$$

We are interested in studying the causal effect of X_2 on Y, i.e., β_2.

```r
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Data")
tsls <- read_csv("tsls.csv")
attach(tsls)

reg1 = lm_robust(y ~ x1 + x2, se_type = "stata")
reg2 = ivreg(y ~ x1 + x2 | x1 + z1 )
reg3 = ivreg(y ~ x1 + x2 | x1 + z1 + z2)

texreg(list(reg1, reg2, reg3), include.ci = F, caption.above = T, digits = 4,
       caption = "TSLS",
       custom.model.names = c("(1)", "(2)", "(3)"))
```

(a) Estimate (1) using OLS. Write out the estimated regression equation along with standard errors and one measure of fit in a standard form.

```
> summary(reg1)

Call:
lm_robust(formula = y ~ x1 + x2, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
(Intercept)   1.0445     0.1154   9.051 1.499e-14  0.81550   1.2736 97
x1            0.3034     0.1759   1.725 8.770e-02 -0.04567   0.6525 97
x2           -0.5430     0.0561  -9.679 6.616e-16 -0.65434  -0.4317 97

Multiple R-squared:  0.505 ,   Adjusted R-squared:  0.4948
F-statistic: 49.34 on 2 and 97 DF,  p-value: 1.651e-15
```

Table 2: TSLS

	(1)	(2)	(3)
(Intercept)	1.0445***	1.0244***	1.0174***
	(0.1154)	(0.1383)	(0.1551)
x1	0.3034	0.8307**	1.0123***
	(0.1759)	(0.3126)	(0.2888)
x2	−0.5430***	−0.9289***	−1.0618***
	(0.0561)	(0.1775)	(0.1357)
R^2	0.5050	0.2688	0.0781
Adj. R^2	0.4948	0.2537	0.0591
Num. obs.	100	100	100
RMSE	0.8109		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The estimated model is

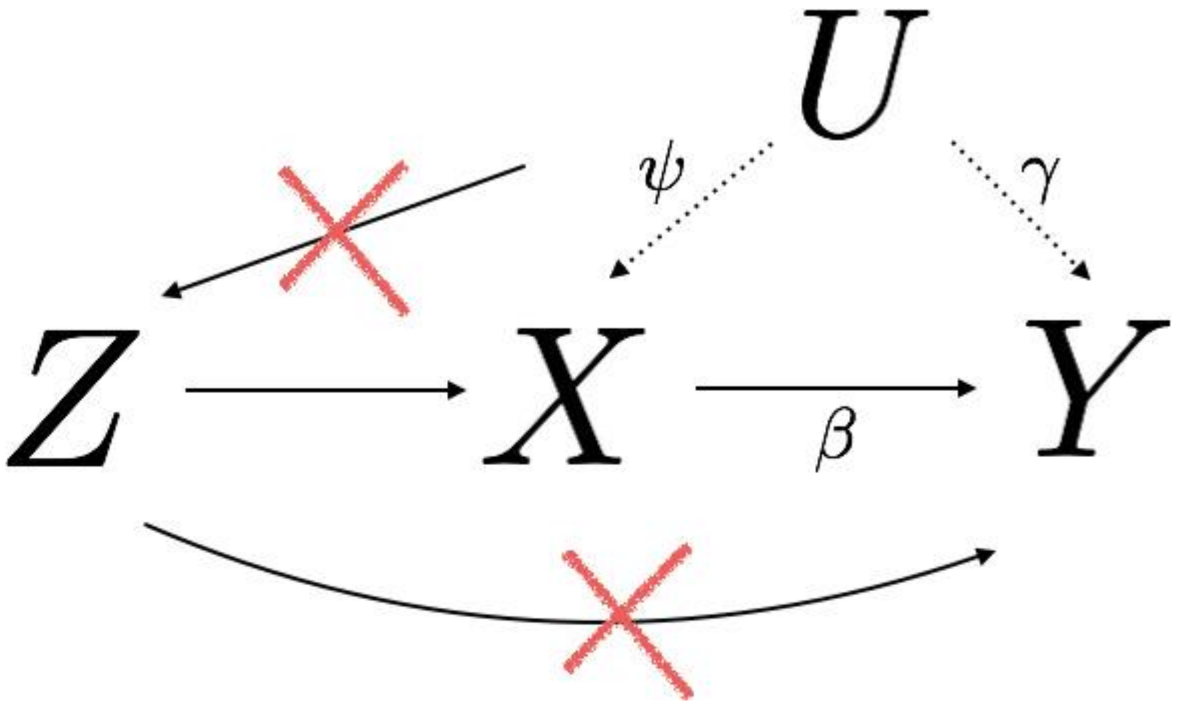
$$\hat{Y} = \underset{(0.115)}{1.045} + \underset{(0.176)}{0.303} X_1 - \underset{(0.056)}{0.543} X_2, \quad \bar{R}^2 = 0.495$$

(b) If X_2 were endogenous, which least squares assumption would be violated? What could be wrong with OLS if this assumption is indeed invalid?

The exogeneity assumption ($E[u|X_1, X_2] = 0$) would be violated. If this were the case, OLS would be biased and inconsistent.

(c) Estimate β_2 using two-stage least squares (TSLS), instead of OLS. Z_1 is one of our candidate instrumental variables (IV). What conditions must hold for Z_1 to be a valid IV for X_2?

Two conditions must hold: (1) $C(u, Z_1) = 0$ (exogeneity), and (2) $C(X_2, Z_1) \neq 0$ (relevance).



(d) Suppose Z_1 is a valid IV for X_2. Run a TSLS regression using Z_1. Write out the estimated regression equations for the second-stage estimation. Are $(\beta_0, \beta_1, \beta_2)$ exactly identified, over-identified, or under-identified? What could be wrong if we run TSLS "manually" (i.e., use the `regress` command twice to replicate the TSLS procedure)?

```
> summary(reg2)

Call:
ivreg(formula = y ~ x1 + x2 | x1 + z1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3753 -0.7670  0.0596  0.5884  2.0363 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.0244     0.1383   7.407 4.79e-11 ***
x1            0.8307     0.3126   2.658   0.0092 ** 
x2           -0.9289     0.1775  -5.233 9.64e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2: TSLS

	(1)	(2)	(3)
(Intercept)	1.0445***	1.0244***	1.0174***
	(0.1154)	(0.1383)	(0.1551)
x1	0.3034	0.8307**	1.0123***
	(0.1759)	(0.3126)	(0.2888)
x2	−0.5430***	−0.9289***	−1.0618***
	(0.0561)	(0.1775)	(0.1357)
R^2	0.5050	0.2688	0.0781
Adj. R^2	0.4948	0.2537	0.0591
Num. obs.	100	100	100
RMSE	0.8109		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The estimated model is

$$\hat{Y} = \underset{(0.137)}{1.024} + \underset{(0.337)}{0.831}\,X_1 - \underset{(0.178)}{0.929}\,\hat{X}_2, \quad \bar{R}^2 = 0.254$$

As we have one IV for one endogenous regressor, β's are exactly identified. Running two OLS can replicates the TSLS estimates. However, this procedure tends to underestimate the SE of the IV estimator, which would make statistical inference (t-statistics, p-values, and confidence intervals, etc.) invalid.

(f) Is Z_1 is a weak IV? Test the relevance of Z_1.

```
> ols.1stage = lm_robust(x2 ~ x1 + z1)
> linearHypothesis(ols.1stage, c("z1 = 0"), test = c("F"))
Linear hypothesis test

Hypothesis:
z1 = 0

Model 1: restricted model
Model 2: x2 ~ x1 + z1

  Res.Df Df      F    Pr(>F)
1     98
2     97  1 17.825 5.463e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Run a regression of X_2 against $(1, X_1, Z_1)$. Compute the F-statistic for the coefficient on Z_1 being 0. The F-statistic $=$ $17.83 > 10$ and has essentially 0 p-value. Thus, we can conclude that Z_1 is relevant and sufficiently strong.

(g) Suppose we have another candidate IV, Z_2. Test the exogeneity of Z_2.

```
> summary(reg3, diagnostics = TRUE)

Call:
ivreg(formula = y ~ x1 + x2 | x1 + z1 + z2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.61717 -0.82661  0.07949  0.70318  2.28762

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0174     0.1551   6.559 2.64e-09 ***
x1            1.0123     0.2888   3.505 0.000692 ***
x2           -1.0618     0.1357  -7.823 6.38e-12 ***

Diagnostic tests:
                 df1 df2 statistic  p-value
Weak instruments   2  96    23.136 6.30e-09 ***
Wu-Hausman         1  96    68.293 7.83e-13 ***
Sargan             1  NA     0.855    0.355
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.107 on 97 degrees of freedom
Multiple R-Squared: 0.07808,    Adjusted R-squared: 0.05907
Wald test: 32.56 on 2 and 97 DF,  p-value: 1.519e-11
```

We conduct the overidentifying restrictions test. The resulting p-value $= 0.355$ is large. Hence, we do not reject the null hypothesis that Z_2 is exogenous.

(h) Now suppose both Z_1 and Z_2 are valid IV. Estimate (1) using both Z_1 and Z_2. How many IV do you want to use to estimate β_2? Explain your answer.

Table 2: TSLS

	(1)	(2)	(3)
(Intercept)	1.0445***	1.0244***	1.0174***
	(0.1154)	(0.1383)	(0.1551)
x1	0.3034	0.8307**	1.0123***
	(0.1759)	(0.3126)	(0.2888)
x2	−0.5430***	−0.9289***	−1.0618***
	(0.0561)	(0.1775)	(0.1357)
R^2	0.5050	0.2688	0.0781
Adj. R^2	0.4948	0.2537	0.0591
Num. obs.	100	100	100
RMSE	0.8109		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

It is better to use both Z_1 and Z_2. The two TSLS estimations give similar estimates of the two slope coefficients, while the one using both Z_1 and Z_2 has smaller SE.

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Thank you

Francisco Tavares Garcia

Academic Tutor | School of Economics

tavaresgarcia.github.io

Reference

Stock, J. H., & Watson, M. W. (2019). Introduction to Econometrics, Global Edition, 4th edition. Pearson Education Limited.