

EC3303 Econometrics I Tutorial Problem Set 5

1. You would like to know how various factors affect students' academic performance. To this end, you examine data collected from students across the World through the OECD's Programme for International Student Assessment study. This is a large-scale study conducted in 2015 comprising 15-year-old students from all across the world.

The variables in the dataset are defined as follows:

science is the student's Science test score, *immigrant* is a binary variable equal to 1 if the student was not born in the country of test and is equal to 0 if the student was born in the country of test, *outhours* is a continuous variable measuring the number of hours that a student spends per week doing self-study outside of school, *pared* is a continuous variable measuring the years of education received by the student's father.

You interact the variables *outhours* and *pared*. You call this variable *outhourspared_interact*. You run a regression of *science* on *immigrant*, *outhours*, *pared*, and *outhourspared_interact*. The output you obtained was:

```
. regress science immigrant outhours pared outhourspared_interact, robust
```

Linear regression	Number of obs	=	340,456
	F(4, 340451)	=	6985.01
	Prob > F	=	0.0000
	R-squared	=	0.0720
	Root MSE	=	94.425

science	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
immigrant	-2.919928	.4936796	-5.91	0.000	-3.887526 -1.95233
outhours	.4569629	.0605892	7.54	0.000	.3382098 .575716
pared	9.337509	.0895873	104.23	0.000	9.16192 9.513097
outhourspared~t	-.0937924	.0044286	-21.18	0.000	-.1024724 -.0851124
_cons	379.788	1.230575	308.63	0.000	377.3761 382.1999

- a. Interpret the estimated coefficient on *outhourspared_interact*. What does it tell you?
- b. Suppose that instead of using *pared*, another researcher instead created a binary variable named *highpared* which is equal to 1 if the years of education received by the student's father is 12 or greater and equal to zero if the years of education received by the student's father is less than 12. He then interacts the variables *outhours* and *highpared*. He calls this variable *outhourshighpared_interact*. He runs a regression of *science* on *immigrant*, *outhours*, *highpared*, and *outhourshighpared_interact*. The output he obtained was:

```
. regress science immigrant outhours highpared outhourshighpared_interact, robust
```

Linear regression

```
Number of obs   =   340,456
F(4, 340451)    =   4801.40
Prob > F         =   0.0000
R-squared        =   0.0514
Root MSE        =   95.466
```

science	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
immigrant	-2.412456	.4974035	-4.85	0.000	-3.387352	-1.43756
outhours	-.3858863	.0316267	-12.20	0.000	-.4478738	-.3238988
highpared	59.15278	.7192965	82.24	0.000	57.74298	60.56259
outhourshigh~t	-.510803	.035753	-14.29	0.000	-.5808779	-.4407281
_cons	458.2412	.6418338	713.96	0.000	456.9832	459.4992

Interpret the estimated coefficient on *outhourshighpared_interact*. What does it tell you?

- c. Suppose that instead of using *outhours*, the researcher in part b instead created a binary variable named *highouthours* which is equal to 1 if a student spends 20 or more hours per week doing self-study outside of school. Also, instead of using *pared*, he instead used *highpared* generated as described in part b. He then interacts the variables *highouthours* and *highpared*. He calls this variable *highouthourshighpared_interact*. He runs a regression of *science* on *immigrant*, *highouthours*, *highpared*, and *highouthourshighpared_interact*. The output he obtained was:

```
. regress science immigrant highouthours highpared highouthourshighpared_interac
> t, robust
```

Linear regression

```
Number of obs   =   340,456
F(4, 340451)    =   4662.69
Prob > F         =   0.0000
R-squared        =   0.0499
Root MSE        =   95.543
```

science	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
immigrant	-2.685839	.4974778	-5.40	0.000	-3.660881	-1.710797
highouthours	-7.290002	.7431192	-9.81	0.000	-8.746494	-5.83351
highpared	54.8654	.4983792	110.09	0.000	53.88859	55.84221
highouthours~t	-11.98708	.8380004	-14.30	0.000	-13.62954	-10.34462
_cons	454.3371	.4469388	1016.55	0.000	453.4611	455.2131

Interpret the estimated coefficient on *highouthourshighpared_interact*. What does it tell you?

Stata Exercise (to be done in tutorial with the tutor)

You will soon be learning about fixed effects regressions and instrumental variables regressions in the lectures (specifically, fixed effects regression will be covered in the week 11 lecture while instrumental variables regression will be covered in the week 12 lecture) and how these methods can be used to mitigate omitted variables bias. Today, we provide a preview on how to implement fixed effects regressions and instrumental variables regressions using Stata.

2. What are the effects of alcohol taxes on traffic fatalities? We address this question using data on traffic fatalities, alcohol taxes, and other related variables for the 48 contiguous U.S. states for each of the seven years from 1982 to 1988. These data are contained in the dataset *fatality.dta*, where our main variables of interest, *fatalityrate* is the number of annual traffic deaths per 1,000 people in the population in the state and *beertax* is the real tax on a case of beer in the state. *year* represents the year in which the fatality rate and the beer tax is observed. Below, we will learn how to perform fixed effects regressions.
 - a. Regress fatality rate in 1982 on real beer tax in 1982, using OLS. What is the coefficient on real beer tax?
 - b. Repeat (a) using fatality rate in 1988 and real beer tax in 1988. What is the coefficient on real beer tax?
 - c. Do the coefficients on real beer tax from (a) and (b) indicate that higher real beer tax is associated with more or fewer traffic fatalities?
 - d. Are estimates of the effect of real beer tax from (a) and (b) reliable? Why?
 - e. Focusing on the data from only 1982 and 1988 for now, create the first differenced variables ($fatalityrate_{i,1988} - fatalityrate_{i,1982}$) and ($beertax_{i,1988} - beertax_{i,1982}$). Run a regression of the first differenced fatality rate on the first differenced beer tax. What do the results tell you?
 - f. Declare the dataset to be a panel. Use the *xtreg* command to run a state fixed effects regression. What do the results tell you? What other variables would you want to include in this regression to obtain a reliable estimate of the effect of beer tax on fatality rate?

3. How does fertility affect labour supply? That is, how much does a woman's labour supply fall when she has an additional child? In this Stata exercise, you will estimate this effect using data for married women from the 1980 United States census. The data are contained in the dataset *fertility.dta*. The dataset contains information on married women aged 21-35 with at least 2 children. Use the dataset to answer these questions.
- Regress *weeksworked* on the indicator variable *morekids*, using OLS (where *weeksworked* represents the number of weeks worked by a woman in a year while *morekids* is a binary variable which is equal to 1 if a woman had more than 2 children and which is equal to zero if a woman had only 2 children). On average, do women with more than 2 children work more or less than women with only 2 children? How much more or how much less?
 - Explain why the OLS regression estimated in (a) is inappropriate for estimating the causal effect of fertility (*morekids*) on labour supply (*weeksworked*).
 - The dataset contains the variable *samesex*, which is equal to 1 if the first two children are of the same sex (boy-boy or girl-girl) and equal to 0 otherwise. Are couples whose first two children are of the same sex more likely to have a third child?
 - Explain why *samesex* is a valid instrumental variable for *morekids* in the regression of *weeksworked* on *morekids*.
 - Estimate the regression of *weeksworked* on *morekids*, using *samesex* as an instrument. What is the estimated effect of fertility on labour supply?