Homework 2 Questions

Note: Please respect intellectual property and do not make this document available in the public domain

1. You would like to know how the years of education attained by individuals is determined. In particular, you would like to know whether the average school fees charged by schools in a person's neighbourhood influences the years of education attained by the person. To investigate this, you collect data on peoples' education, average fees charged by schools in the person's neighbourhood, the distance of the individual's residence (home) to the nearest tertiary institution, the combined income earned by the individual's parents, the education of the individual's father, and the individual's gender and race among others. The variable *education* represents the total number of months of education received by the individual; *fees* represents the average school fees charged per year by schools in the individual's residential neighbourhood, measured in thousands of dollars (In other words, a value of $1,000 shows up as 1 in the dataset); *fees2* denotes the square of fees; *high_income* represents a binary variable equal to one if the individual has high income parents (his/her parents have a combined income greater than or equal to $80,000 per year) and equal to zero if the individual has low income parents (his/her parents have a combined income of less than $80,000 per year); *dad_grad* represents a binary variable equal to one if the individual's father is a university graduate and equal to zero if the individual's father is not a university graduate; *female* represents a binary variable equal to one if the individual is female and equal to zero if the individual is male. *minority* represents a binary variable indicating whether the student belongs to a minority race (equal to 1 if the individual belongs to a minority race and equal to zero if the individual belongs to the majority race). *distance* is a continuous variable representing the distance from an individual's home to the nearest tertiary institution, measured in kilometres.

   You log transform *distance (*using natural logs*)* and you call this variable *lndistance*. Further, you interact the variables *female* and *high_income* and you call this variable *int_fhighinc*.

   You then run the regression below:

```
. regress education fees fees2 high_income dad_grad female minority lndistance int_fhighinc, robust

Linear regression                               Number of obs     =      2,473
                                                F(8, 2464)        =      40.75
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1139
                                                Root MSE          =     20.237
```

education	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
fees	37.12135	9.735449	3.81	0.000	18.03084	56.21186
fees2	-19.67065	5.288026	-3.72	0.000	-30.04009	-9.301219
high_income	3.46022	1.380232	2.51	0.012	.7536847	6.166755
dad_grad	13.62843	1.142116	11.93	0.000	11.38883	15.86804
female	-.0233956	.9559493	-0.02	0.980	-1.897943	1.851151
minority	-3.665875	.9077404	-4.04	0.000	-5.445888	-1.885862
lndistance	-.7379352	.3640536	-2.03	0.043	-1.451818	-.0240526
int_fhighinc	2.358989	1.842266	1.28	0.200	-1.253561	5.971539
_cons	145.3011	4.240155	34.27	0.000	136.9865	153.6157

You then performed a test using the command "test *high_income int_fhighinc*" and obtained the following result:

```
. test high_income int_fhighinc

 ( 1)  high_income = 0
 ( 2)  int_fhighinc = 0

       F(  2,  2464) =   11.71
            Prob > F =    0.0000
```

Use this information to answer Questions 1 to 12. **Be careful of units.**

Person A is male, with high income parents, whose father is a university graduate, who belongs to the majority race, who lives 0.2km from the nearest tertiary institution, and whose average school fees charged per year by schools in the residential neighbourhood amounts to $889.15. Person B is female, with low income parents, whose father is not a university graduate, who belongs to the majority race, who lives 0.5km from the nearest tertiary institution, and whose average school fees charged per year by schools in the residential neighbourhood amounts to $900. Round your final answer to the following question to 1 decimal place and choose the closest option.

Complete the statement. Person A's education is predicted to be:

2. Trump is male, with low income parents, whose father is a university graduate, who belongs to the majority race, who lives 0.6km from the nearest tertiary institution, and whose average school fees charged per year by schools in the residential neighbourhood amounts to $931.87. Trump's has 192 months of education. What is the residual specific to Trump? Round your final answer to 2 decimal places and choose the closest option.

3. Which of the following statements is incorrect? The regression output implies that:

4. Refer to the regression in question 1. What would the estimated coefficient on the variables *fees* and *fees2* be if the variable *fees* were now representing the average school fees charged per year by schools in the individual's residential neighbourhood, and measured in hundreds of dollars?

5. What is the adjusted R-squared for the regression in Question 1? Round your answers to 4 decimal places. Pick the closest option.

6. What would the root mean squared error of the regression from Question 1 be if we divided all values of *education* by 2.5, before running a regression of this modified *education* variable on all the regressors that appear in the regression in Question 1? Round your final answer to 3 decimal places.

7. Refer to the regression in Question 1 again. Had you removed the binary variable *minority* from that regression, your estimated regression would have been:

```
. regress education fees fees2 high_income dad_grad female lndistance int_fhighinc, robust

Linear regression                               Number of obs     =      2,473
                                                F(7, 2465)        =      43.58
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1084
                                                Root MSE          =     20.295

------------------------------------------------------------------------------
             |               Robust
   education |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        fees |   36.85774   9.766442     3.77   0.000     17.70646    56.00902
       fees2 |  -19.05966   5.302802    -3.59   0.000    -29.45807   -8.661255
 high_income |   3.769938   1.383674     2.72   0.006     1.056655     6.48322
    dad_grad |   14.11593   1.136272    12.42   0.000     11.88779    16.34408
      female |  -.1517545   .9596826    -0.16   0.874    -2.033622    1.730113
  lndistance |  -.5243943     .36021    -1.46   0.146     -1.23074    .1819511
int_fhighinc |    2.49199   1.849202     1.35   0.178     -1.13416    6.118139
       _cons |    143.794   4.233083    33.97   0.000     135.4933    152.0948
------------------------------------------------------------------------------
```

Which of the following statements is correct?

8. Refer to the regression in Question 7, where the variable *minority* is excluded as a regressor. Suppose that the true causal effect of being of a minority race is to decrease educational attainment (perhaps because of racial discrimination). Then what do the regression results in Question 1 and the regression results in Question 7 together imply?

9. Refer to the regression output in Question 1 again. Suppose you are now thinking of creating another binary variable which is equal to one if the individual has low income parents (his/her parents have a combined income of less than $80,000 per year) and equal to zero if the individual has high income parents (his/her parents have a combined income of greater than or equal to $80,000 per year). You call this binary variable *low_income*. If you ran the same regression in question 1, but replace the variable *high_income* with the variable *low_income*, and replace the variable *int_fhighinc* with the variable *int_flowinc*, where *int_flowinc* is obtained by interacting the variables *female* and *low_income*. what would the coefficients and standard errors on *low_income* and *int_flowinc* be?

10. Refer to the regression output in Question 1 again. Which of the following statements is true?

11. Refer to the regression output in Question 1 again. Rounded to 2 decimal places, what is the 79.6% confidence interval for the coefficient on the variable *dad_grad*?

12. Refer to the regression output in Question 1 again. Suppose you include a binary variable *mon_grad* which is equal to one if the individual's mother is a university graduate and equal to zero if the individual's mother is not a university graduate as a regressor. And you obtain the following regression output:

```
. regress education fees fees2 high_income dad_grad female minority lndistance int_flowinc mom_grad, robust

Linear regression                               Number of obs     =      2,473
                                                F(9, 2463)        =      41.55
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1241
                                                Root MSE          =     20.125

------------------------------------------------------------------------------
             |               Robust
   education |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        fees |   34.75205   9.678221     3.59   0.000     15.77376    53.73034
       fees2 |  -18.34573   5.266235    -3.48   0.001    -28.67244   -8.019029
 high_income |   2.709981   1.385492     1.96   0.051    -.0068692    5.426831
    dad_grad |   11.09158   1.254174     8.84   0.000     8.632234    13.55092
      female |   2.626507   1.558911     1.68   0.092    -.4304041    5.683418
    minority |  -3.583195    .903216    -3.97   0.000    -5.354337   -1.812054
  lndistance |  -.6673087   .3618306    -1.84   0.065    -1.376832    .0422148
 int_flowinc |  -2.737482   1.829484    -1.50   0.135    -6.324969    .8500042
    mom_grad |   7.102339   1.376777     5.16   0.000     4.402579    9.802099
       _cons |   145.9511    4.20086    34.74   0.000     137.7135    154.1886
------------------------------------------------------------------------------
```

Suppose that you now conduct the following test:

```
. test dad_grad = mom_grad

 ( 1)  dad_grad - mom_grad = 0

       F(  1,  2463) =    3.21
            Prob > F =    0.0733
```

What do your results imply?

13. Please see Canvas (under Files -> Homework -> Homework 2). The file pisa_2015_apr2023.dta contains 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, *escs* is a continuous composite variable measuring a student's economic, social and cultural (i.e. socioeconomic) status. The higher the level of *escs*, the higher a person's socioeconomic status. Finally, *pared* is a continuous variable measuring the years of education received by the student's father. Use this information to answer Questions 13 to 16.

    Note that you may have to create new variables in stata before you can run some of the regressions in the questions that follow.

    Run a regression of the natural logarithm of Science test score, *ln(science)*, on *immigrant*, *outhours*, and *escs*. Holding immigration status and socioeconomic status constant, if number of hours that a student spends per week doing self-study outside of school increases from 20 to 23, which of the following is true?

14. Now, run a regression of the natural logarithm of Science test score, *ln(science)*, on *outhours*, *immigrant*, *escs*, as well as the natural logarithm of *pared* (i.e. *ln(pared)*). Holding immigration status, socioeconomic status constant, and out of school self-study constant, if parental education increases from 15 years to 16 years, which of the following is true?

15. Now, run a regression of the natural logarithm of Science test score, *ln(science)*, on *outhours*, *immigrant*, *escs*, as well as *pared* and $pared^2$. Conduct additional tests to examine the following statements, if necessary. Which of the following statements is correct?

16. Interact the variables *immigrant* and *outhours* and include this interaction term in the regression specified in Question 15. What do the results tell you?

17. In this question (and question 18), we build on what we learnt about fixed effects in Tutorial 5. Specifically, we are once again interested in 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 five years from 1982 to 1986. These data are contained in the dataset fatality_apr2023.dta, which can be located in Canvas (under Files -> Homework -> Homework 2), where our main variables of interest, *fatalityrate* is the number of annual traffic deaths per 10,000 people in the population in the state and *beertax* is the real tax on a case of beer in the state, in dollars. *year* represents the year in which the fatality rate and the beer tax is observed. Note that the dataset fatality_apr2023.dta is slightly different from the one we saw in the tutorial.

    Focus only on the data from the years 1983 and 1984. Create the first differenced fatalityrate and beertax variables. Run a regression of the first differenced fatality rate on the first differenced beer tax. Cluster your standard errors by state. What do the results tell you? Choose the best option.

18. Return to the dataset mentioned in question 17. Instead of using data only from the years 1983 and 1984, use data on all the years available (i.e. data on all years, from 1982 to 1986). Declare the dataset to be a panel. Use the xtreg command to run a state fixed effects regression with the main dependent variable being *fatalityrate* and the main regressor being *beertax*. Include the variables *perinc* (measuring per capita income in a state in a year) and *unrate* (measuring unemployment rate in a state in a year) as regressors. What do the results tell you?

19. In questions 19 and 20, we build on what we learnt about IV regression in Tutorial 5. Specifically, you want to know how much a woman's labour supply changes when she has an additional child. To learn about this, you will examine the data contained in the dataset fertility_apr2023.dta, which can be found in Canvas (under Files -> Homework -> Homework 2). The dataset contains information on married women aged 21-35 with at least 2 children. Note that the dataset fertility_apr2023.dta is slightly different from the one we saw in the tutorial.

Your variables of interest are *weeksworked,* which represents the number of weeks worked by a woman in a year, *morekids* which is a binary variable equal to 1 if a woman had more than 2 children and equal to zero if a woman had only 2 children, and *samesex*, which is a binary variable equal to 1 if the first two children are of the same sex (boy-boy or girl-girl) and equal to 0 otherwise. Your dataset consists of people belonging to 4 mutually exclusive and exhaustive race categories – white, black, Hispanic, and "all other races". The variable *black* is a binary variable equal to 1 if the person is black and equal to zero otherwise. The other race variables are defined similarly. *agem1* represents the age of the woman.

Run a regression of *weeksworked* on *morekids*, using *samesex* as an instrument for *morekids*. Include controls for race (*black*, *hispan*, *othrace*) and age of the woman. What do your results tell you?

20. Refer to the regression in question 19. What is the value of the F-statistic from the first stage regression?