Exam 4137 – Spring 2021 postponed

IMPORTANT: Always motivate your answers. Show knowledge and understanding of the concepts taught in the course. Your answers should be as short as possible and as long as necessary. Total marks of the examination are 130. Each subquestion carries 10 marks.

---

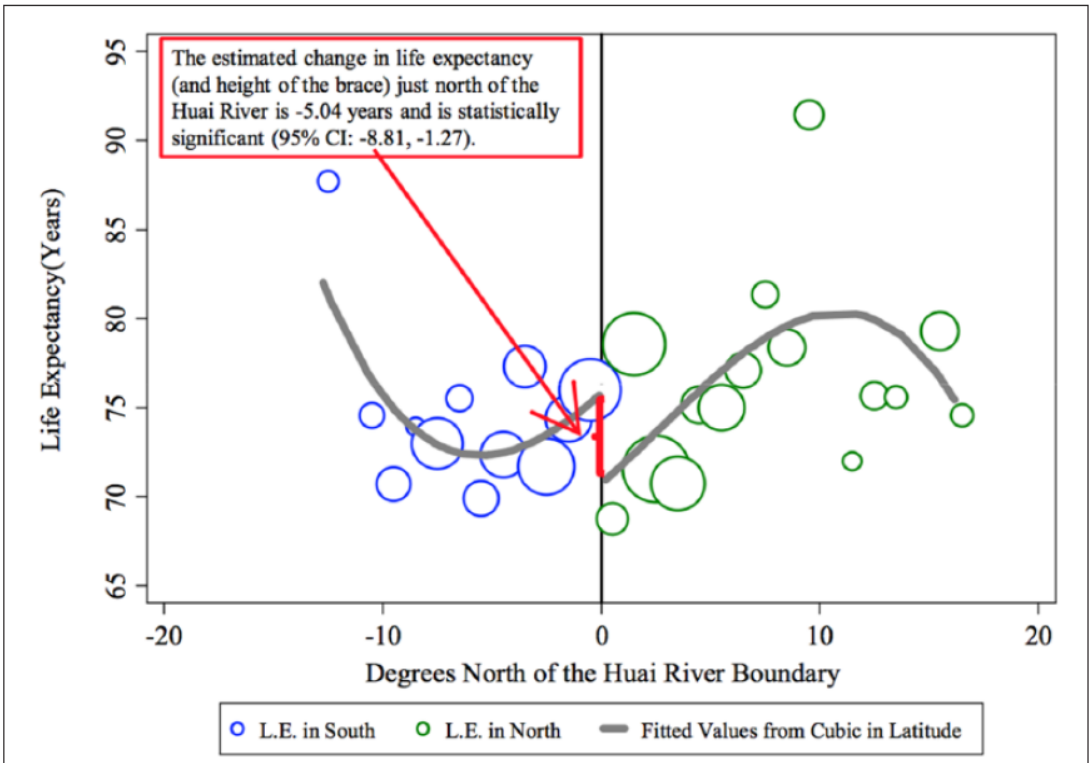
1. In a study of the marital status of a sample of Norwegian adults, we seek to study the relationship between marital status and individual characteristics such as age (x_1) and gender ($x_2 = 1$, if female and $= 0$ if male). Marital status is coded as (1) Single and never married, (2) Married, (3) Divorced or separated, and (4) Widowed, meaning the husband or wife is dead.

   (a) Write down the model you want to use to analyze this problem. How will you estimate the unknown parameters?

   (b) Suppose you want a single test of whether, controlling for gender, marital status is related to age. Specify the null hypothesis and use the attached Stata output to test this hypothesis.

   (c) Compute the partial effect of being female on the probability of being married for a person of 47 years old.

2. A paper published on PNAS (Proceedings of the National Academy of Sciences) estimates the effect of coal heating in China on life expectancy. There was a coal heating policy which dictated that areas to the north of the Huai River received free or highly subsidized coal for indoor heating. This led to the construction of a coal-powered centralized heating infrastructure only in cities north of the Huai River, and no equivalent system is in cities to the south; the legacy of that policy is evident even today, with very different rates of indoor heating north and south of the Huai River. Using a regression discontinuity design based on distance from the Huai River, the authors claim that coal heating is causing the 500 million residents of Northern China to lose more than 2.5 billion life years of life expectancy. The essence of the empirical analysis can be summarized using the following diagram:



The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

(a) The authors used a Regression Discontinuity Design (RDD) to identify the causal effect they want to study. Based on the diagram, did the author use a sharp or fuzzy design? What implicit assumption do the authors need to make for this choice (between sharp RD and fuzzy RD) to be correct?

(b) What is the key assumption of this identification strategy? What causal effect is estimated using this strategy? Can you specify a simple regression similar to the one in the above diagram (you don't need to consider the weights and other control variables). Remember to define your variables.

(c) As we can see from the diagram, the estimated causal effect is around -5 years. The conclusion that air pollution caused by coal heating is causing a loss of 2.5 billion life years in China is obtained by simply multiply the estimate (-5) by the number of residents of Northern China. What do you think about the estimate of the causal effect and the total loss of lifeyears?

(d) The authors also checked several other specifications. The results are summarized in the following table. What can you learn from this table? Note that you are supposed to comment on the effect on life expectancy (second row) only.

Table S9

Robustness checks of choice of functional form for latitude

	Linear & Controls	Quadratic & Controls	Cubic & Controls	Quartic & Controls	Quintic & Controls
	(1)	(2)	(3)	(4)	(5)
Panel 1: Impact of "North" on the Listed Variable, Ordinary Least Squares					
TSP (100 μg/m^3)	2.89***	2.63***	1.84***	1.95***	1.52**
	(0.56)	(0.49)	(0.63)	(0.59)	(0.72)
Life Expectancy (years)	-1.62	-1.29	-5.52**	-5.67**	-5.43*
	(1.66)	(1.68)	(2.39)	(2.36)	(2.94)

3. Growing concerns over the inadequate achievement of pupils have led to proposals to reward good teachers and penalize (or fire) bad ones. The leading method for assessing teacher quality is "value added" modeling (VAM), which decomposes students' test scores into components attributed to student heterogeneity and to teacher quality. In a recent study, administrative data on elementary pupils is used to address this problem. The data contains end-of-grade math test scores from grades 3 through 5 for 2,415 pupils taught by 116 different teachers. Denote the test score for pupil i at grade t (let $t = 1$ corresponds to grade 3, and etc) as y_{it} ($i = 1, 2, ..., 2415$, $t = 1, 2, 3$) and similarly her/his teacher as J_{it} ($J_{it} \in \{1, 2, ..., 116\}$).

   (a) What type of data is this? What are the benefits of having such data?

   (b) One of your fellow students suggested to run the following regression using OLS:

$$y_{it} = \alpha_t + \sum_{j=1}^{116} \beta_j I(J_{it} = j) + \varepsilon_{it} \tag{1}$$

   where $I(.)$ is an indicator function which takes value 1 if pupil i 's teacher is j at grade t, and 0 otherwise. What do you think of his suggestion?

   (c) Another student believes that unobserved pupil effects are important to be controlled. She suggests to run the following regression

$$y_{it} = \alpha_t + \gamma_i + \sum_{j=1}^{116} \beta_j I(J_{it} = j) + e_{it} \tag{2}$$

   using a random effect model. What assumptions are needed for her estimates to be consistent? If these conditions hold, what would you say about estimates from (1)?

   (d) However, the majority of your fellow students prefer to use a fixed effect setup. What do you think their main arguments will be? Explain how would you estimate this model, if you think that there are some unobserved time-varying factors (not correlated with J_{it}) that have impacts on pupils' grades and correlated over time, for example, influences from family, non-classroom peers, and etc.

(e) Suppose that some teachers taught the same group of pupils during the whole observation periods. What can you expect to find when you estimate the regression (2) using the fixed effect model? Would your answer change if you estimate the regression using instead the random effect model?

(f) Some argue that pupil achievement depends on the full history of inputs received to date, even after pupil and teacher effects are controlled for. In other words, everything else (including all observed and unobserved factors) being the same, at current period, pupils who had high math scores previously tend to score higher than those who had low math scores previously. If this is true, what would you say about the estimation method discussed above in (c) and (d)?

Stata Output

```
. su id status age female

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          id |      1,000       500.5    288.8194          1       1000
      status |      1,000       2.582    .7898733          1          4
         age |      1,000      39.709    5.702529         30         50
      female |      1,000        .482    .4999259          0          1

.  . mlogit status female, baseoutcome(1)

Iteration 0:   log likelihood = -1153.7144
Iteration 1:   log likelihood = -1142.1744
Iteration 2:   log likelihood = -1142.0874
Iteration 3:   log likelihood = -1142.0873

Multinomial logistic regression                 Number of obs     =      1,000
                                                LR chi2(3)        =      23.25
                                                Prob > chi2       =     0.0000
Log likelihood = -1142.0873                     Pseudo R2         =     0.0101

------------------------------------------------------------------------------
              status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+--------------------------------------------------------
Single_and_never_married |  (base outcome)
---------------------+--------------------------------------------------------
Married              |
              female |   .5077406   .3151699     1.61   0.107     -.109981    1.125462
               _cons |   1.977163   .1778312    11.12   0.000      1.62862    2.325705
---------------------+--------------------------------------------------------
Divorced_or_separated |
              female |   .9676716   .3186213     3.04   0.002     .3431854    1.592158
               _cons |   1.522426   .1839521     8.28   0.000     1.161887    1.882966
---------------------+--------------------------------------------------------
Widowed              |
              female |   1.162325   .3466513     3.35   0.001     .4829008    1.841749
               _cons |   .4595322   .2128888     2.16   0.031     .0422778    .8767865 ----------------------------------------

.  . mlogit status age female, baseoutcome(1)
Iteration 0:   log likelihood = -1153.7144
Iteration 1:   log likelihood = -1118.8063
Iteration 2:   log likelihood = -1116.8325
Iteration 3:   log likelihood = -1116.8259
Iteration 4:   log likelihood = -1116.8259

Multinomial logistic regression                 Number of obs     =      1,000
                                                LR chi2(6)        =      73.78
                                                Prob > chi2       =     0.0000
Log likelihood = -1116.8259                     Pseudo R2         =     0.0320
------------------------------------------------------------------------------
              status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+--------------------------------------------------------
Single_and_never_married |  (base outcome)
---------------------+--------------------------------------------------------
Married              |
                 age |   .1644902   .0308803     5.33   0.000     .1039659    .2250144
              female |   .4329018   .3203811     1.35   0.177    -.1950336    1.060837
               _cons |  -4.265608   1.130526    -3.77   0.000    -6.481398   -2.049818
---------------------+--------------------------------------------------------
Divorced_or_separated |
                 age |   .1341297   .0311311     4.31   0.000     .0731138    .1951455
              female |   .9025774   .3221191     2.80   0.005     .2712355    1.533919
               _cons |  -3.504963   1.140289    -3.07   0.002    -5.739887   -1.270038
---------------------+--------------------------------------------------------
Widowed              |
                 age |   .0783897   .0335091     2.34   0.019     .0127131    .1440664
              female |   1.120498   .3479925     3.22   0.001     .4384457    1.802551
               _cons |  -2.414512   1.231514    -1.96   0.050    -4.828236   -.0007882 ----------------------------------------

.  . predict double p_*
(option pr assumed; predicted probabilities)

.  . list id status age female p_* if _n<=10

     +-------------------------------------------------------------------------------------------+
     | id                  status   age   female         p_1         p_2         p_3         p_4 |
     |-------------------------------------------------------------------------------------------|
  1. |  1                 Married    37        0   .07635091   .47147619   .32805298   .12411993 |
  2. |  2   Divorced or separated    49        0   .01406425   .62518453    .3021813   .05856991 |
  3. |  3                 Married    50        0   .01212509   .63535116   .29791186   .05461189 |
  4. |  4                 Married    38        1   .03363601   .37748015    .4075402   .18134364 |
  5. |  5                 Married    49        0   .01406425   .62518453    .3021813   .05856991 |
     |-------------------------------------------------------------------------------------------|
  6. |  6                 Married    40        0   .05102921   .51614998   .32787169   .10494912 |
  7. |  7                 Married    47        1   .00982768   .48470233   .39818308   .10728692 |
  8. |  8   Divorced or separated    46        0   .02183508   .59256355     .313726   .07187538 |
  9. |  9   Divorced or separated    47        0   .01887477   .60380835   .31011965   .06719723 |
 10. | 10                 Married    39        0   .05847646   .50176576   .32855981   .11119797 |
     +-------------------------------------------------------------------------------------------+
```

Critical Values of Chi-square

df	.50	.25	.10	.05	.025	.01	.001
1	0.45	1.32	2.71	3.84	5.02	6.63	10.83
2	1.39	2.77	4.61	5.99	7.38	9.21	13.82
3	2.37	4.11	6.25	7.81	9.35	11.34	16.27
4	3.36	5.39	7.78	9.49	11.14	13.28	18.47
5	4.35	6.63	9.24	11.07	12.83	15.09	20.52
6	5.35	7.84	10.64	12.59	14.45	16.81	22.46
7	6.35	9.04	12.02	14.07	16.01	18.48	24.32
8	7.34	10.22	13.36	15.51	17.53	20.09	26.12
9	8.34	11.39	14.68	16.92	19.02	21.67	27.88
10	9.34	12.55	15.99	18.31	20.48	23.21	29.59