Exam 4137 – Spring 2024

IMPORTANT: Always motivate your answers. Show knowledge and understanding of the concepts taught in the course. Make sure not only to repeat material from lectures and curriculum, but engage with the empirical examples and show that you understand what it means in the particular setting. Your answers should be as short as possible and as long as necessary. Each lettered subtask carries 10 points, for a total of 160 points.

---

1. Short questions

   True or False? No points without explanation

   a) You should always use robust standard errors with a linear probability model

   b) The incidental parameters problem imply that we cannot consistently estimate the variance of the fixed effects.

2. You estimate a regression of student test scores on the share of high school teachers that have master degrees. The average class size is 25.

   (a) Would you be worried about correlation between errors within class? Use the Moulton factor to explain what factors determine how much you underestimate standard errors, and argue whether or not these are important in this example.

   (b) Explain how cluster-robust standard errors, clustered bootstrap or collapsing your data to class averages would correct standard errors in this case. Which method would you prefer, and why?

3. The Norwegian government is worried about dropping fertility rates, and proposes a pro-natalist policy to induce Norwegian families to get more children. They are particulary worried about fewer families having three or more children. You work in the Ministry of Children and Families, and are asked to analyse the degree to which female labor supply would be reduced if more families had three rather than two children.

   You have a dataset of a representative sample of Norwegian women who all have two or more children with a dummy variable for whether they work or not ($work_i$), and whether or not they have three or more children ($morekids_i = 1(numkids_i \geq 3)$). In addition, you observe the age of the woman (age_i) and whether or not she has a university degree ($educ_i$). You run the simple regression

   $$work_i = \alpha + \beta morekids_i + \varepsilon_i$$

   The results are reported in column 1 of Table 1.

Table 1: Code and output from your fertility and labor supply analysis

```
> col1=feols(work~morekids,data=data)
> col2=feols(work~morekids+age+I(morekids*(age-mean(age))),data=data)
> col3=feols(work~morekids+age+I(morekids*(age-mean(age[morekids==1]))),data=data)
> col4=feols(work~morekids+age+I(morekids*(age-mean(age[morekids==0]))),data=data)
> col5=feols(morekids~samesex,data=data)
> col6=feols(work~1|morekids~samesex,data=data)
> col7=feols(morekids~samesex,data=data[data$educ==1,])
> etable(col1,col2,col3,col4,col5,col6,col7,se.below=T,order="!Constant")
                                          col1       col2       col3       col4       col5       col6       col7
Dependent Var.:                           work       work       work       work   morekids       work   morekids

morekids                           -0.0372***  -0.0202***  -0.0199***  -0.0204***                0.0330
                                   (0.0028)    (0.0027)    (0.0028)    (0.0027)                (0.0629)
age                                             -0.0080***  -0.0080***  -0.0080***
                                                (0.0002)    (0.0002)    (0.0002)
I(morekids x (age-mean(age)))                    0.0002
                                                (0.0003)
I(morekids x (age-mean(age[morekids==1])))                   0.0002
                                                            (0.0003)
I(morekids x (age-mean(age[morekids==0])))                                0.0002
                                                                         (0.0003)
samesex                                                                             0.0425***                0.0393***
                                                                                    (0.0034)                 (0.0051)
Constant                            0.8360***   1.223***    1.223***    1.223***    0.3835***  0.8076***   0.3674***
                                   (0.0017)    (0.0079)    (0.0079)    (0.0079)    (0.0024)   (0.0255)    (0.0036)
___________________________________ __________ __________ __________ __________ __________ __________ __________
S.E. type                          Hete.-rob. Hete.-rob. Hete.-rob. Hete.-rob. Hete.-rob. Hete.-rob. Hete.-rob.
Observations                           82,876     82,876     82,876     82,876     82,876     82,876     35,877
R2                                    0.00227    0.04758    0.04758    0.04758    0.00187   -0.00582    0.00163
Adj. R2                               0.00226    0.04755    0.04755    0.04755    0.00186   -0.00583    0.00160
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Which assumption do you need to make for the coefficient β on the explanatory variable morekids to be interpreted as an average causal effect? Interpret the estimated coefficient.

(b) You are worried about the validity of the assumption you made, and in particular the bias coming from omitting the age of the mother. Explain how this omitted variable biases your results above.

(c) Since you have access to age, you want to control for this in the above regression. In the heterogeneous effects framework, what assumption do you need to make for the parameter on morekids to have a causal interpretation when controlling for age? Interpret the estimate in column 2 of Table 1, and compare to a).

(d) You run two additional regressions and get the output in Table 1, columns 3-4, where the parameter of interest in all cases is the coefficient on *morekids*. Which of the average treatment effects in columns 2-4 do you think policymakers should be interested in, and why?

(e) In the above regressions, you controlled for age linearly. Is this reasonable? Explain how you would approach the problem of choosing the appropriate specification of age.

(f) You are still worried about omitted variable bias. Fortunately, you remember from your time studying applied microeconometrics that having two children of the same sex is often used as an instrument for fertility, in particular having the third child. Discuss exclusion and monotonicity in this setting, providing at least one example of a violation for each.

Table 2: Coefficients from a logit model, Nieboer et al. (2010)

Attribute	**Coeff**	***p*-value**
Hours of care per week	0.039	0.000
Hours of organized social activities per week	0.071	0.000
Transportation service	0.572	0.000
Non-regular care provider	-0.173	0.000
Individualized care	0.145	0.003
Coordinated by care recipient	-0.169	0.000
Maximum care delay, hours	-0.119	0.000
Waiting list, per month	-0.064	0.000
Co-payment per week (euros)	-0.005	0.000
Living arrangement (baseline: Living alone)		
Apartment building	0.171	0.040
Sheltered accomodation	0.169	0.000
Elderly/nursing home	-0.013	0.628
Log likelihood	-1277	
Pseudo R2	0.160	

(g) Confident in the assumptions needed for IV, you estimate the following model using 2SLS, and get the output in Table 1, columns 5-6. Interpret these estimates. Is the instrument relevant? Compare your estimated treatment effect to the one obtained in a).

(h) You proudly present your estimates of causal effects of fertility on female labor supply to the Minister of Children and Families. She wants to know how representive the treatment effect you have estimated for compliers is for all mothers that would be affected by the policy. Column 7 of Table 1 shows the first stage estimated separately in the group of college educated women. Explain how you would use this to estimate the share of college educated women among compliers. What more could you do to understand who your compliers are?

4. In "*Preferences for long-term care services: Willingness to pay estimates derived from a discrete choice experiment*" (Social Science & Medicine 70, 2010), Anna P. Nieboer, Xander Koolman and Elly A. Stolk investigate a sample of 50-64 year old Dutch taxpayer's preferences over elderly care services. The authors asked the respondents to imagine a hypothetical elderly relative in need of care, and randomly presented them with two care packages with different attributes and prices. They analyse the choice experiment data using a logit model.

(a) The results from estimating the logit model are reported in Table 2. Focus on the coefficients for number of hours of care per week, the hours of organized social activities per week an and co-payment. How can you interpret the signs and magnitudes of the the three coefficients in the latent utility framework?

(b) Calculate the willingness to pay for another weekly hour of care. Interpret the result.

(c) The probability to choose care package B ($y_i = 1$) and the marginal effect of an increase in an attribute x_k in the logit model can be expressed as

$$\begin{aligned} P(y_i = 1 \mid x) = &\quad p(x) = \tfrac{1}{1+\exp(-x'\beta)} \\ \frac{\partial P(y_i = 1 \mid x)}{\partial x_k} = &\quad p(x)(1-p(x))\beta_k \end{aligned}$$

Where x is the vector of differences in characteristics between option B and the baseline option A and β refer to the estimated coefficients in Table 2.

Consider two individuals. Both face two scenarios with exactly the same characteristics except for price, which is 50 euros higher for option B for individual 1 and 100 euros higher for option B for individual 2. Calculate the marginal effect on the probability of choosing option B of a 1 euro increase in price of option B for these two individuals. Why is one larger than the other? Assume that the constant is zero.

(d) Would you prefer to summarize the marginal effects using Average Partial Effects (APE) or Partial Effects at the Average (PEA)? Why?