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.

---

1. (10 points) 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.

   (a) TRUE. Homoskedasticity cannot hold in a linear probability model. The variance of y given x from the LPM is $Var(y \mid x) = \Pr(y \mid x)(1 - \Pr(y \mid x)) = E(y \mid x)(1 - E(y \mid x)) = x'\beta(1 - x'\beta)$, which is a function of x unless all the coefficients happen to be zero. This violates homoskedasticity, which implies that the variance of the outcome does not depend on x.

   (b) FALSE. The incidental parameters problem imply that we cannot consistently estimate the fixed effects themselves, because the number of observations per individual is fixed and small. We can however estimate the variance of the fixed effects by decomposing the error term and backing out the variance of the individual fixed effects.

2. (20 points) 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.

The Moulton factor describes the ratio of the true variance to the estimated variance assuming homoskedasticity:

$$M = \sqrt{\frac{Var(\hat{\beta})_T}{Var(\hat{\beta})_h}} = \sqrt{1 + (\frac{Var(n_g)}{\bar{n_g}} + \bar{n_g} - 1)\rho_e \rho_x}$$

, where n_g is the group size of group g and ρ_e and ρ_x are the intraclass correlation coefficient of the error term and the regressor x, respectively. The Moulton factor increases in

$Var(n_g)$: If classes are very different sized, the problem is worse. In Norwegian high schools, the variance in the class size is typically relatively small, so this term should perhaps not be the major concern.

$\bar{n_g}$: If the average group size is larger, the problem is worse, because with a fixed sample you have fewer independent groups.

ρ_e: If the error term is mostly determined at the group level ($\rho_e \to 1$), there is less independent variation in the outcome. In this case it is very unlikely that $\rho_e = 1$, as there is variation in tests scores within class, but it is also unlikely that $\rho_e = 0$ as there is correlation within class coming from shared school, class and neighborhood that affect test scores.

ρ_x: If the regressor is mostly determined at the group level ($\rho_x \to 1$), there is less independent variation in the regressor, and the problem is worse. In this case, a lot of the variation in the share of teachers that a student experienced that has a master degree is determined at the class level, because a class share a lot of teachers. Still, students take different courses, and thus $\rho_x < 1$.

In the limit where $\rho_e = \rho_x = 1$ (all observations within a group are identical) and $Var(n_g) = 0$ (fixed group size), the true variance is $\bar{n}$ times as large as the estimated variance! We are essentially multiplying our group-level dataset with the average class size $\bar{n}$ to get more observations when in reality there is no extra information in repeated identical observations. Then the true standard errors will be $\sqrt{\bar{n}} = 5$ times larger than the estimated standard errors! In reality, the difference is likely to be smaller than this because $\rho_e < 1$ and $\rho_x < 1$, counteracted by $Var(n_g) > 0$ which increases the ratio compared to the limit case.
Explanations based on the simplified Moulton factor with x fixed at the group level and equal group size should be given partial credit. Students need to provide interpretation of why the different factors increase standard errors, not only list the formula and what the terms are.

(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?

The most straightforward thing to do is apply clustered standard errors, clustering by class. This allows for arbitrary correlation of errors within class, but not across classes.
We could also cluster bootstrap by resampling classes from our data and estimating the parameter of interest over and over again. Then our estimated variance would be the variance of the parameter across many bootstrap repetitions, and confidence intervals could be computed using the percentile method.
Finally we could collapse data to the class level and regress class level average test scores on class level average exposure to teachers with master degrees.
Unless data is perfectly identical within classes, we should prefer cluster-robust standard error or cluster bootstrap, because they doesn't discard the variation within class.

3. (80 points) 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.

   (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.

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
```

We need randomization: $(work_i^0, work_i^1) \perp morekids_i$, which says that joint potential outcomes are independent from treatment. In this application, it implies that having three or more rather than two children is unrelated to potential labor supply you would have had with either 2 or 3 or more children - fertility is as good as random. Under full randomization, the coefficient on morekids estimates the average treatment effect (ATE).

An alternative (and, strictly speaking, weaker) way to express this is $\mathbb{E}(work_i^0 \mid morekids_i = 0) = \mathbb{E}(work_i^0 \mid morekids_i = 1)$, which is a mean independence condition on untreated potential outcomes only. Under the weaker mean independence condition β estimates the average treatment effect on the treated (ATT).

The estimate suggest that having three or more kids reduces the labor force participation of the average mother (average mother with 3 or more kids under the weaker mean independence condition) with 3.7 percentage points. Students should be given full marks for either of these two formulations, but need to interpret the assumptions in light of the fertility and labor supply example.

An alternative that should be given some credit is the assumption that $\mathbb{E}(\varepsilon_i morekids_i) = 0$ or $\varepsilon_i \perp morekids_i$. In a structural world with homogenous effects, this allows us to estimate the (one and only) treatment effect β. In this case students should interpret ε_i and d_i as unobservables that affect labor supply and realized fertility, respectively. An assumption like $\mathbb{E}(\varepsilon_i morekids_i) = 0$ or $\varepsilon_i \perp morekids_i$ with an explicit interpretation of β as an ATE should give full credit.

(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.

This is an open ended question and it does not matter what direction they argue the bias goes, what matters is the argument. Below is one example that should be given full marks.
When the randomization assumption fails, the regression above estimates $\underbrace{\mathbb{E}(w_i^1 - w_i^0 \mid m_i = 1)}_{\text{ATT}} + \underbrace{\mathbb{E}(w_i^0 \mid m_i = 1) - \mathbb{E}(w_i^0 \mid m_i = 0)}_{\text{selection bias}}$, where w and m is short for work and morekids. This is the sum of the ATT and a selection bias term. The latter captures how much more mothers with three or more children would work than mothers with 2, if both of them had 2 children.
In this example, it may be argued that selection bias is negative: Women who have three or more children would have had lower labor supply than women who would stop at two, even when both had stopped at two. If that is the case, our estimate from a) will be too negative.

(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).

Now we rely on the conditional independence assumption - $(work_i^0, work_i^1) \perp morekids_i \mid age_i$. This means that *conditional on age,* potential labor supply with 2 and 3 or more kids is unrelated to actual fertility.
The regression in column 2 estimates the average treatment effect, and tells us that on average, a mother would reduce labor supply by around 2 percentage points when going from 2 to 3 or more children. This suggests that the selection bias was negative - our original estimate overestimated the negative effects of fertility.
Students need to interpret this as an average treatment effect for full credits.

(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?

These regressions estimate the Average treatment effect (ATE, col 2), the average treatment effect for the treated (ATT, col 3), and the average treatment effect for the untreated (ATUT, col 4), respectively. Because of the conditional independence assumption, the only reason these are different is because 2- and 3+ child mothers are of different ages, and age may affect the treatment effects. Appropriately recentering at the mean of the group of interest (all, treated, untreated) allows us to interpret the coefficient on *morekids* as either the ATE, ATT or ATU.
In this case it is most reasonable for policymakers to be interested in the ATU. If their policy works by inducing some families to get the third child, it must be families who currently stop at 2!
In practice, there is very little evidence of heterogeneity in treatment effects across age, so that all three treatment effect parameters are very similar.

(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.

This is an open ended question, where students are expected to briefly explain a few of the tools we covered on specification.
First, it is unreasonable for age to have linear effects on labor supply. Age profiles in earnings and labor supply are typically nonlinear, as few mothers work in their early 20s when they e.g. study, then gradually enter employment. Nonlinear age profiles may also be present later in life, for instance due to early retirement.
To investigate the specification for age, we can plot mean labor supply by age to get a visual representation of the relationship. Because there are two regressors in the model (*morekids* and *age*), we need to residualize age and labor supply from *morekids* and plot the (mean of the) residuals.
We can also try alternative specifications, such as higher order polynomials. When comparing specifications, it may be useful to compare adjusted R^2, which is a goodness of fit measure, and we would prefer models with higher adjusted R^2. We can also formally test the statistical significance of a new coefficient, for instance by testing whether the coefficient on the higher order term is significant.
Students may also mention J-tests, RESET tests and non-nested F-tests.
In the end, the most flexible specification we can use in this setting is a separate dummy for each 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.

In the potential outcomes framework we have the following potential outcomes, indexed by values of treatment ($d = 0$: 2 children, $d = 1$: 3 or more children) and the instrument ($z = 0$: First two children of different gender, $z = 1$: first two children of the same gender)
$y_i(d,z)$ is person i's potential labor supply with morekids set to d and with samesex set to z.
$d_i(z)$ is person i's potential fertility with samesex set to z

Exclusion: $y_i(d,1) = y_i(d,0)$. The instrument only matters for the outcome through treatment, in this case getting kids of the same sex (vs. opposite) only affects labor supply through whether or not a mother gets additional children. This is violated if for instance having two boys is harder to combine with a career than one boy and one girl, in which case having the first two of the same sex affects labor supply even for mothers who would stop at 2 no matter what or mothers who would have the third no matter what. Another possible violation is if the instrument affects fertility at later margins - increasing the probability to have the fourth and fifth child. This does not seem too unreasonable - people with very strong preferences for both genders could go on to have the fourth and the fifth if they fail at the third or fourth, which is caused by failing at the second attempt, our instrument.

Monotonicity: $d_i(1) \geq d_i(0)$ or vice verca. The instrument shifts people only into treatment, not both ways. This could be violated in this setting if some parents have active preference for two boys or two girls, which could very well be the case. The monotonicity assumption thus puts restrictions on what kind of preferences parents may have for the gender of their children.

(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).

The estimate of *samesex* in column 5 suggest that 4.2% of the population are compliers: They get the third child if the first two are same sex, but otherwise would not. This is strongly significant, with an F-statistic well above 100 ($(0.0425/0.0034)^2$), supporting the existence of a first stage.
The estimate in column 6 for morekids suggest that there is a *positive*, but insignificant average treatment effect of having three or more kids on labor supply. Taken at face value, the point estimate suggest that on average, compliers increase their labor supply by 3 percentage points when getting the third child. This is a local average treatment effect: The average causal effect of the third child for mothers who get it because the first two children are same sex, and wouldn't otherwise.
This estimate is of opposite sign than the estimated treatment effects that rely on CIA or randomization, but very imprecise.
Students should get limited points for just repeating what is in the table - they need to correctly interpret these as complier shares and a LATE.

(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?

To understand who the compliers are, we can characterize them in terms of their background variables. There are multiple ways of doing this, but the most straightforward is using first stage ratios. Since the first stage measures the share of compliers, a higher first stage in a subgroup than in the whole sample suggest that people of that subgroup are overrepresented among compliers. The first stage in the group of college educated women is 0.0393. This implies that the college share among compliers is 3.9%, smaller than in the sample as a whole. Consequently, the share of college educated women among compliers is $\frac{0.0393}{0.0425} = 0.925$ times the share in the sample as a whole.
We can also estimate mean potential outcomes for compliers, $\mathbb{E}(work_i^0 \mid morekids_i(1) - morekids_i(0) > 0)$ and $\mathbb{E}(work_i^1 \mid morekids_i(1) - morekids_i(0) > 0)$. These can help us understand whether the treatment effects for compliers are large because they work very little when they have 3 or more children, or whether they are large because they work a lot when they have two. We can also compare these to mean potential outcomes for women who always choose to have 3 or more kids $\mathbb{E}(work_i^1 \mid morekids_i(1) = morekids_i(0) = 1)$ or women who always choose to have 2 kids $\mathbb{E}(work_i^0 \mid morekids_i(1) = morekids_i(0) = 0)$, allowing us to think about how the LATE could extrapolate to other groups. Mean potential outcomes for compliers can be estimated by replacing the outcome in the IV model with the outcome multiplied by treatment for the mean treated potential outcome, and analogously for the mean untreated potential outcome.

4. (40 points) 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?

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	

We imagine an underlying linear utility $U_j = x_j'\beta + \varepsilon_j$, where x_j is the attributes of choice j, β are the preference parameters and ε_j is unobserved utility. Because utility does not have a location, we need to normalize the utility of one of the options to 0 (here: Care package a), with the consequence that the utility and the attributes of the other option is measured relative to this baseline option. Neither does utility have a scale, and we normalize this by choosing a distribution for ε, here the logistic distribution.
The implication of this is that the coefficients β can be interpreted in utility terms - relative to the variance of the random part of utility. The sign of the coefficients are indicative of preferences - and thus the negative coefficient on co-payment indicates that people do not like to pay. The positive coefficient on hours of care and organized social activities indicate that they like more care and more activities. Moreover, the relative magnitudes of two coefficients can be interepreted as relative utility. Thus, respondents like an hour of social activities for the elderly twice as much as an hour of care because the coefficient on hours of social activities are about twice the size of the coefficient on hours of care.
Students do not need to fully lay out the latent utility framework, but should be able to interpret signs and express that β must be interpreted in utility terms.

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

Marginal rates of substitution that on average keep respondents indifferent may be expressed as the ratio of two coefficients. Willingness to pay is just a special case of a marginal rate of substitution where the normalization is with respect to money, and thus we measure the utility value of a particular amenity in monetary terms. To calculate WTP for another weekly hour of care, we take the negative of the ratio of the coefficient on weekly hour of care to the coefficient on co-payment:

$$WTP = -\frac{\beta_h}{\beta_p} = -\frac{0.039}{-0.005} = 7.8$$

This WTP suggest that the respondents are willing to pay 8.11 euros for another hour of care for their elderly relative.

(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

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

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.

The probabilities $p(x) = \frac{1}{1+\exp(-x'\beta)} = \frac{1}{1+\exp(-p\beta_p)}$, where p is the price difference and β_p is the coefficient on price. Thus the two probabilities are $p_1 = \frac{1}{1+\exp(0.05*50)} = 0.438$ and $p_2 == \frac{1}{1+\exp(0.05*100)} = 0.378$, and the marginal effects are $p_1(1-p_1)\beta_p = -0.00123 < -0.00118 = p_2(1-p_2)\beta_p$.
The marginal effect for person 1 is more negative because he has a predicted probability closer to 0.5 - people who face similar scenarios as him tend to choose both scenarios with more equal probability than people who face scenarios like person 2. It makes sense that a 1 euro increase in price has a larger effect on the probability of choice for someone who is facing a scenario where a lot of people choose both options. Indeed, the marginal effects are larges for $p(x) = 0.5$!

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

The marginal effects differ across people who face different scenarios and thus different x. We can either evaluate the average of these marginal effects for all observations in the sample (the APE), or we can evaluate the marginal effect for an average scenario (the PEA). These may be very different in practice.
In this case it is most reasonable to summarize using the APE. Particulary in cases with dummy variables like this one, the PEA will evaluate a scenario with for instance a living arrangement that is 0.25 apartment building, 0.25 living on his/her own, 0.25 living in sheltered accomodation and 0.25 living in an elderly home, if these are the average differences across scenarios. That is not representative for anyone in the sample, and makes little sense. In contrast, the APE takes a well defined average over scenario-specific marginal effects.