Exam 4137 – Spring 2026

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 in addition to using math. Your answers should be as short as possible and as long as necessary. Each lettered subtask carries 10 points, for a total of 150.

---

1. (20 points) Short questions.

   (a) TRUE or FALSE - No points without an explanation!

      i. In order for OLS regression coefficients to be unbiased, we must assume that the error term is homoskedastic

      FALSE: Homoskedasticity is required for estimation of standrard errors, not regression coefficients. For coefficients to be unbiased, we need the error term to be uncorrelated with the regressors.

      ii. The common trends assumption can be tested by comparing pre-trends

      FALSE: You can never test the common trends assumption, it is fundamentally about unobservables - about how trends would have evolved after treatment in the absence of treatment. We often evaluate trends before treatment to substantiate that it is plausible, but parallel pre-trends can hold when common trends assumption fails and vice versa.

   (b) Interpret the following assumptions. $Y(d,z)$ and Y_{it}^{d} are potential outcomes under treatment state d and instrument value$z$, D_i is actual treatment and X_{it} are covariates for observation i at time t.

      i. $Y_i(d,z) = Y_i(d)$

      This is the exclusion restriction for IV. Potential outcomes do not depend on the instrument other than through the realized treatment.

      ii. $\mathbb{E}(Y_{i1}^{0} - Y_{i0}^{0} \mid D_i, X_{it}) = \mathbb{E}(Y_{i1}^{0} - Y_{i0}^{0} \mid X_{it})$

      This is the (conditional) common trend assumption: It says that the change in potential outcomes from period 0 to period 1 does not depend on treatment status - i.e. units who are and aren't actually treated would have change on average the same over time if neither were treated and their covariates changed the same.

2. (80 points) You work with the application and admissions system for high schools in the Oslo Education Agency. You are tasked with analyzing the causal effect of attending different schools because your employer is worried about differences in school quality. You have a dataset on an entire cohort of Oslo students at academic track high schools that contains information on which high school $D \in (1,2,...,15)$

they attended, and annual earnings at age 30, Y_i. You are interested in well-defined average causal effects on earnings of attending one school j relative to another j'.

(a) Using potential outcomes notation, define and interpret the potential and realized outcomes.

Let Y_i^j be the potential outcome of student i if he or she attended school j. It is the (unobserved) earnings that this student would have if he or she had attended school j. Potential and realized outcomes are linked through $Y_i = \sum_j (D_i = j) Y_i^j$, so that if student i actually attend school 13 ($D_i = 13$), the potential outcome attending school 13 for this student is observed. Students are expected to be able to generalize from two to multiple treatments in this simple case.

(b) What do you need to assume about potential outcomes for a simple difference in means between students attending school j and j' to be causal? What causal effect is that?

Independence $(Y^1, Y^2, ...Y^{15}) \perp D_i$ - all potential outcomes are independent of actual school attendance. In this case: What you would earn if you attended school \$j\$ is independent of what school you actually attended. It cannot be the case that individuals who would make more at any school j would be more or less likely to attend that or any other school. Alternatives with mean instead of full independence are also accepted. For any two schools j and j', a simple mean difference identifies the average treatment effect of attending school j instead of j' - with independence there can be neither selection on gains nor levels, so all treatment effect averages (ATT/ATE/ATUT) are the same.

(c) In the reasonable case that independence does not hold, will you over- or underestimate the true effect? Use potential outcomes notation to express the bias and explain why.

If independence fails, the mean comparison or regression identifies $\underbrace{\mathbb{E}(Y^j - Y^{j'} \mid D_i = j)}_{\text{ATT}} + \underbrace{\mathbb{E}(Y^{j'} \mid D = j) - \mathbb{E}(Y^{j'} \mid D = j')}_{\text{selection bias}}$. In words, it identifies the average treatment effect of attentding school j instead of j' for those who attended j, plus the average difference in the potential outcomes in school j' between students who attend school j and j'.
The likely direction of selection bias depend on which school is better. If school j attracts better students from upper secondary school, these students would probably have done better than students at school j' even if they all attended j'. In that case, selection bias is positive and we would overestimate the true effect because ability or past grades are omitted variables.
The direction of the selection bias is not important - students can get full points if they convincingly argue for one or more omitted variables that could cause negative selection bias. The response should contain both the math and the verbal reasoning for full marks.

(d) The dataset also contains GPA_i, the grade point average of each student from

compulsory school. Explain how you could use this to weaken the independence assumption from b)

Now we could rely on conditional independence instead: $(Y^1, Y^2, \ldots, Y^{15}) \perp D_i \mid GPA_i$. Now, we need actual school assignment to be independent of potential outcomes only within cells of GPA, i,e, we want to be comparing students with equal scores from upper secondary school. This is likely to pick up a lot of the selection bias we described above.

(e) What regression would you run to exploit the weaker assumption from d)? What average treatment effect would you estimate?

For a single comparison between school j and j', I would focus on the sample of students who attended one of them , and estimate

$$Y_i = \alpha + \beta 1(D_i = j) + \delta(GPA_i - \bar{GPA}) + \theta(GPA_i - \bar{GPA})1(D_i = j) + \varepsilon_i$$

Where 1() is the indicator function and δ controls (linearly) for GPA and θ for the interaction between GPA and treatment. The latter is important to ensure that the regression actually estimates the average treatment effect on the treated - omitting the θ term would lead to a weighted average of GPA-specific average treatment effects that does not necessarily equal the overall ATE. REcentering at other values to get the ATT or ATUT is also fine, but students should clearly state which one the regression would produce.
This specification assumes a linear functional form in GPA, which is not necessarily correct. We could make the functional form less restrictive by using more flexible functional forms.
Students should get points, but not full marks, for correctly specifying a regression without an interaction term, unless they also explicitly argue why the precision-weighted average of GPA-specific effects are the target parameter. More flexible functional forms are also fine. Students should also note the functional form for full marks - it is a restriction over the general conditional independence assumption.

(f) Still unconvinced about the assumption you need to make, you dig up data on the admissions thresholds c_j to each of the 15 schools. Let schools now be indexed by their cutoff so that school 1 is the highest-threshold school, $c_1 \geq c_2 \geq \ldots \geq c_{15}$.You can also construct $Z_{ij} = GPA_i \geq c_j$, a dummy for whether student i's grades are good enough to be accepted to school j. Focus on the causal affect of attending the highest-ranked school ($D_i = 1$), among students with GPA better than the cutoff of the second-highest ranked school ($GPA_i > c_2$), who applied to school 1 above school 2, using the instrument for threshold crossing to the highest-ranked school ($Z_{i1} = GPA_i > c_1$). Explain what assumptions you need to make to utilize this in a regression discontinuity framework.

Continuity of potential outcomes $lim_{x\downarrow c_1} E(Y^j \mid GPA_i = x) = lim_{x\uparrow c_1} E(Y^j \mid GPA_i = x)$. Mean potential outcomes do not change discontinuously at the threshold, i.e. mean earnings when attending school j or j' does not jump at the admission threshold. $E(Y^j \mid X = x)$ continuous in x are also sufficient. Students could note that this assumption implies exclusion.

Relevance $lim_{c\downarrow 0} Pr(D_i = 1 \mid GPA_i = c) \neq lim_{c\uparrow 0} Pr(D_i = 1 \mid GPA_i = c_1)$.. Treatment is discontinuous at the threshold - the probability to attend school 1 does jump discontinuously at the admission threshold.

Monotonicity $D_1 \geq D_0$. Crossing the threshold to the higher ranked school does not induce someone who would otherwise go to the higher ranked school not to go. Here D_z are potential treatments indicating whether a student would go to the higher ranked school if threshold crossing is exogenously set to z.

We did not explicitly discuss monotonicity in the RD lecture, so you should give students some leeway here, but since we discussed extensively the link to IV, students should be able to connect.

(g) You estimate the following system using 2SLS

$$\begin{aligned} Y_i &= \alpha + \beta 1(D_i = 1) + \gamma(GPA_i - c_1) + \theta(GPA_i - c_1)Z_{ij} + \varepsilon_i \\ 1(D_i = 1) &= \zeta + \pi Z_{i1} + \delta(GPA_i - c_1) + \eta(GPA_i - c_1)Z_{i1} + \mu_i \end{aligned}$$

Explain the roles of the $\gamma, \theta, \delta$and η terms. How do you interpret the coefficients π and β?

The parameters γ and θ controls for mean differences in earnings for people with different GPA's, making the estimates local to the cutoff.
The parameters δ and η does the same for the probability to attend the higher ranking school
π: Average effect of threshold crossing on the probability to attend the higher-ranking school
β: Local average treatment effect of attending the higher ranking school for people with GPA at the threshold, and who go to the higher-ranking school if above, but not if below.
Students need both of the ways in which β is local (local to threshold, compliers) to get full marks.

(h) Explain the logic behind at least two different robustness checks you could perform to increase your trust in the assumptions behind this estimate.

McCrary test: Check for local discontinuities in the density of GPA at the threshold - evidence of manipulation
Covariate balance checks: See if predetermined covariates change discontinuously at the threshold - then we would worry that potential outcomes could, too
Placebo checks: Test at other points in the GPA distribution where there was no threshold and see if you get effects of the same magnitude - then we would worry about violation of continuity
Robustness to polynomial order, bandwidth or weights - check whether your model is flexible and/or local enough to capture differences across GPA

3. (50 points) In order to fully understand how the application system to high school shapes outcomes, you also need to understand students' (and parents) preferences for schools. Next, you are tasked with surveying the literature on school preferences, and you find the influential paper *"What Parents Want: School Preferences and School Choice"* by Simon Burgess, Ellen Greaves, Anna Vignoles, Deborah Wilson, published in the Economic Journal in 2015. In this paper, the authors use high school application data from England to estimate a conditional logit model of preferences for school characteristics. Their model is the classic conditional/multinomial logit model, where

$$P_{is} = \frac{e^{X_{is}\beta + W_i X_{iz}\gamma}}{\sum_{l=1}^{n} e^{X_{il}\beta + W_i X_{il}\gamma}}$$

governs the probability that student i chooses school s. The specification includes school characteristics such as school quality and distance (in X_{is}), and indicators for family socioeconomic status (in W), which allows different SES-group families to have different preferences.

   (a) Derive an expression for the log likelihood from this probability, and explain how you would use maximum likelihood to estimate the parameters based on this expression.

Assuming observations are i.i.d, the probability to observe the exact choices we see in the data is simply equal to the product of the probabilities to observe each of the individual-school choices. Thus

$$
\begin{aligned}
\mathscr{L} &= \prod_i \prod_s P_{is}^{1(y_i=s)} \\
\ell &= \sum_i \sum_s 1(y_i = s) \log(P_{is}) \\
&= \sum_i \sum_s 1(y_i = s) \log\left(\frac{e^{X_{is}\beta + W_i X_{iz}\gamma}}{\sum_{l=1}^{n} e^{X_{il}\beta + W_i X_{il}\gamma}}\right) \\
&= \sum_i \sum_s 1(y_i = s) \left[X_{is}\beta + W_i X_{is}\gamma - \log\left(\sum_{l=1}^{n} \exp(X_{il}\beta + W_i X_{il}\gamma) \right) \right].
\end{aligned}
$$

Which is a function of unknown parameters and data (X, W, y). Maximum likelihood finds the solution by maximizing the log likelihood. While we could take derivatives of ℓ with respect to all parameters and set them to zero to find the first order condition, these generally cannot be solved analytically. Therefore, we need to use numerical optimization algorithms, that essentially tries lots of different values for the parameters and chooses the combination that produces the largest log likelihood.

(b) A simplified excerpt of their results is found in Table 1. Focusing on the first column of the table and one or two coefficients of choice, what can you learn about the preferences of parents in the highest SES group from signs and relative magnitudes? How are preferences of the lowest SES group different?

Coefficients must be interpreted on a utility scale. Magnitudes are relative to the magnitudes of unobserved factors, and thus a bit hard to interpret, but signs can be interpreted: Parents like schools with higher academic quality as measured by mean grades in past years. Parents dislike schools with more disadvantaged students, as measured by special education needs, share of students with free meals or share non-white. Parents do not like to travel, as evident by the negative coefficient on distance rank
We can also interpret relative magnitudes. Parents value one unit of academic quality about half as much as they dislike one of more unit of free school meal students, as evident by the coefficient being about half the size.
Preferences of low-SES parents are different - they dislike travel less and care less about academic quality, as evident by the negative coefficient on academic quality and the positive coefficient on distance rank in the second column.
Students need to discuss sign, relative magnitudes and the contrast between SES for full marks.

(c) How much longer is a high-SES parent willing to send their kids to get a 1 SD increase in academic quality?

Table 1: School choice preference estimates from Burgess et al., 2015

	Reference group (highest SES)	Interaction: lowest SES
Academic quality: KS2 mean points score over 3 years (SD units)	0.270 (0.087)	-0.363 (0.108)
Share free school meals (SD units)	-0.539 (0.161)	0.939 (0.185)
Share English as additional language (SD units)	0.171 (0.109)	-0.044 (0.152)
Share white (SD units)	0.233 (0.115)	-0.272 (0.154)
Any special educational need (SD units)	-0.230 (0.075)	0.165 (0.087)
Faith school (binary indicator)	-0.817 (0.111)	-0.071 (0.144)
Distance rank	-0.644 (0.091)	0.261 (0.108)
N	42,047	
No. households	7,905	

Willingness to pay can be calculated from negative the ratio of the coefficients, so willingness to travel in this case is $WTP = -\frac{\beta_Q}{\beta_d} = -\frac{0.267}{-0.644} = 0.42$, so they are willing to travel around 0.42 of a distance rank more to get another unit of academic quality.

(d) An elasticity of demand measures how much relative demand for a school increase when a characteristic x_k increase by 1%, e.g. $\frac{\partial P_{is}}{\partial x_{isk}} \frac{x_{isk}}{P_{is}} = \beta_k x_{isk}(1-P_{is})$. The authors report elasticity estimates based on Partial Effects at the Average (PEA) , i.e. evaluating the elasticity for the average school and individual. An excerpt of the authors' elasticity estimates is found in Table 2, which shows that the elasticity of demand with respect to school quality is 4 for high-SES students in row 1. Use the numbers in the Tables to compute this elasticity and interpret it.

Table 2: Demand elasticity estimates from Burgess et al. (2015)

	Mean school characteristic X	Elasticity (Lowest SES)	Elasticity (Highest SES)
Academic quality (KS2, SD units)	17.7	-1.38	4.00
		(0.631)	(0.80)
Free school meals (SD units)	1.2	0.40	-0.54
		(0.057)	(0.096)
English as additional language (SD units)	0.9	0.10	0.13
		(0.052)	(0.070)
White (SD units)	3.9	-0.13	0.76
		(0.259)	(0.308)
Any special educational need (SD units)	1.9	-0.10	-0.37
		(0.062)	(0.084)
Distance rank	4.0	-0.26	-0.55
		(0.063)	(0.061)

The mean number of schools in the choice set is 6.1. We therefore fix $\bar{P} = 1/6.1 = 0.164$.

From table 1, we have that β_k for academic quality is 0.270
From table 2 we have that the average academic quality is $\bar{x_{isk}} = 17.7$
From the footnote, we have that the average choice set has 6.1 schools, so $\bar{P_{is}}$ for one of 6.1 identical schools is $P_{is} = \frac{e^{X_{is}\beta + W_i X_{iz}\gamma}}{\sum_{l=1}^{n} e^{X_{il}\beta + W_i X_{is}\gamma}} = \frac{e^{\bar{X_s}\beta}}{6.1 e^{\bar{X_s}\beta}} = \frac{1}{6.1} = 0.164$. Unsuprisingly, when all schools are average on all characteristics the choice probabilities for all schools are the same and equal to 1/N.
Thus, the elasticity is $\frac{\partial P_{is}}{\partial x_{isk}} \frac{x_{isk}}{P_{is}} = \beta_k \bar{x_{isk}}(1 - \bar{P_{is}}) = 0.270 * 17.7 * (1 - 0.164) = 3.995$, very close to the point estimate of 4 reported in the table. This means that demand for any school increases by 4 percent when school quality of that school increase by 1 percent and all schools are average for an individual who faces the average choice set..

(e) Do you think that this partial elasticity at the average is a reasonable way to summarize the heterogeneous effects from this model? What is the alternative, and what is the difference between the two?

In reality, all schools are not the same, and the size of the choice set differ across individuals. Instead of evaluating the elasticity at the average school and for the average choice set, we could evaluate it for each actual school in each individual's choice set, and then take the average across all schools and individuals. This is an elasticity based on the Average Partial Effect (APE) rather than the Partial Effect at the Average (PEA). Because of the nonlinearity of the conditional logit model, these generally differ. For example, people will typically not face 6.1 identical average schools, but will have some schools with much larger predicted choice probabilities and school quality, and others with much lower. This means that the elasticities for each of the schools could be very different, and averaging over them could produce a very different average than the PEA-based elasticity. Furthermore, individuals' choice sets could differ in size, meaning that the P_{is} are different across individuals. PEA-based elasticities could also evaluate marginal effects at weird combinations of covariates, such as at a school that is 10% faith-based, which does not exist. In general, we often prefer the average partial effects to the partial effects at the average to predict effects of changes in the covariates.