Microeconometrics
Sample exam questions

The questions provided in this document represents some examples of questions in preparation of the exam, not of a full exam.

1. A Government is launching a new nation-wide project providing public funds to municipalities in order to build new roads as locally-managed project. However, it is worried that municipality governments will capture large part of the grant for their private returns. To reduce the risk of corruption, they contact you to understand how public funds affect corruption at the municipality level. To answer this question to the government, you have access to a previous dataset containing the following information at municipality-level (municipality is indicated by i) for the 1534 municipalities of the country:

   - Number of corruption cases (per 1000 inhabitants) (Y_i)
   - Municipality expenditure on public infrastructure (in thousands of US$) ($E_i$)
   - Municipality characteristics - k variables covering the characteristics of the municipality
   - Municipality leader characteristics - m variables covering the characteristics of the political leader of each municipality

   (a) You run the following OLS specification:

   $$Y_i = \alpha + \beta * E_i + \epsilon_i \tag{1}$$

   Show how you can identify β (i.e. the causal effect of expenditure on corruption) and discuss what assumptions are needed and whether these are credible in this case.

   (b) The estimates of β are reported in the following table. Column (1) use the specification of equation (1), while column (2)–(3) adds control variables as reported in the bottom panel of the table. How do you interpret these results?

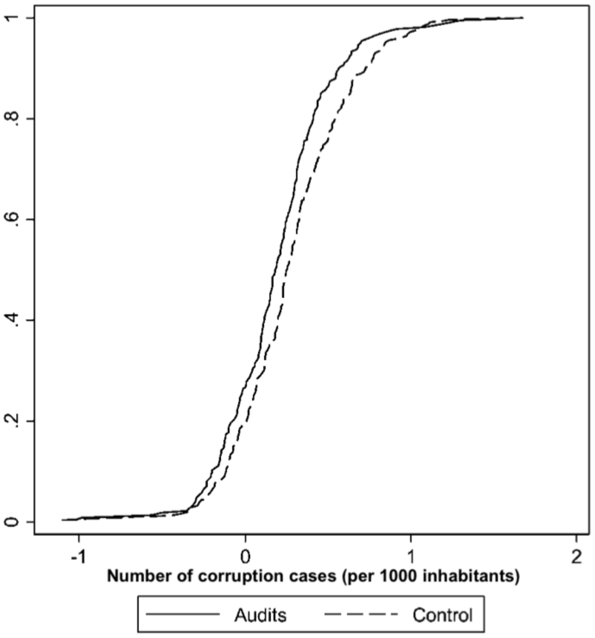
Table 1: Expenditure on public infrastructure and corruption cases

	Dep. variable: Corruption cases (per 1000 inhabitants)		
	(1)	(2)	(3)
Expenditure on public infrastructure (E_i)	0.085*	0.076**	0.048
	(0.044)	(0.036)	(0.031)
Observations	1534	1534	1534
Controls:			
Municipality characteristics	No	Yes	Yes
Municipality leader characteristics	No	No	Yes

Note. Estimates refer to OLS regressions. Standard errors in parenthesis. * significant at 10 percent ** significant at 5 percent *** significant at 1 percent.

(c) You are given the possibility to design a randomized controlled trial focusing on 500 municipalities in the country. You can randomly allocate municipalities in two groups: one group is kept as a control group (no intervention) and another group is characterized by monthly audits from external auditors sent by the central Government. These two groups are defined by a dummy variable T_i equal to 1 if the municipality is in the group receiving the audit and 0 if it is in the control group. Discuss how you would estimate the effect of audits on corruption cases (Y_i), including a discussion of the assumptions needed for identifying the effect.

(d) The following figure shows the (cumulative) distribution of the number of corruption cases in the treatment group (Audits) and in the control group (Control). Can you conclude that audits were effective? Explain.



(e) Now imagine that you don't have access to Y_i, but only to a dummy variable equal to one if there is at least one corruption case in the municipality and zero otherwise. You decide to estimate a probit model using the variables detailed in the randomized controlled trial and assuming error terms are homoskedastic. Detail the latent variable model behind your estimation, derive the response probability and the likelihood function.

2. You have access to a dataset providing data about health status for a group of individuals, their smoking habit and their characteristics. The data come in panel format, meaning you observe these variables for the same individual at multiple points in time over 10 years. Let $H_{i,t}$ be a variable measuring the health status of person i at time t, and let $S_{i,t}$ be the number of cigarettes smoked per day by person i at time t. $X_{i,t}$ is a vector of individual (time-varying) characteristics (example: education, labour supply, etc.). Your objective is to estimate the causal effect of smoking on the health status.

   (a) What assumptions are needed to estimate this causal effect using pooled OLS?

   (b) Propose a panel data model to measure this causal effect and discuss the difference with the pooled OLS approach presented in (a), specifying the assumptions.

3. You are interested in measuring the effect of a scholarship to attend a training course on the probability to find a job. The scholarship is randomly assigned to students, and is measured by the

variable S_i, which reports the amount received by the person. In the dataset, you also observe whether the person found a job in the year following the course. This variable, JOB_i, is equal to 1 if the person found a job, and zero otherwise. In addition, the dataset includes a vector X_i containing observable characteristics including gender, age, grades, etc.

(a) What are the pros and cons of using a OLS model like the following one to estimate the effect of the scholarship on the probability to find a job?

$$JOB_i = \alpha + \beta_S S_i + \gamma X_i + u_i \tag{2}$$

(b) Assume you want to estimate the effect using a probit model. You do that assuming a latent variable model, in which JOB_i is observed only if JOB_i^* is positive, and the error term u_i is distributed normally with mean zero and standard deviation. In other words, these are the main assumptions you are taking:

$$\begin{aligned} JOB_i^* &= \alpha + \beta_S S_i + \gamma X_i + u_i \\ & \quad u_i|\mathbf{x}_i \sim Normal(0, \sigma^2) \\ JOB_i &= 1 \text{ if } y_i^* > 0 \\ &= 0 \text{ if } y_i^* \leq 0 \end{aligned}$$

Derive the probability that the individual i finds a job on the scholarship received and on its observable characteristics.

(c) Assume you estimate the model using the model derived in the previous point. However, the original data generating process include an error term distributed as

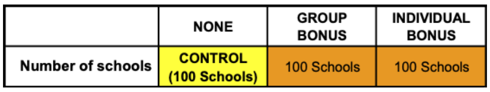
$$u_i|\mathbf{x}_i \sim Normal(0, \gamma\sigma^2 + \gamma^2)$$

What is the consequence for your estimates? Show the steps to reach your conclusion.

4. You are hired to work with the main economic advisor of the country of Krakozhia to study whether providing financial incentives to school teachers help improving the performance of students. The advisor proposes to improve the quality of teaching provided by teachers by introducing two types of financial bonuses for teachers: a bonus based on the average grade of the class where the teacher is teaching (*individual bonus*) versus a bonus based on the average grade of the school (*group bonus*). To study whether this is effective, the advisor proposes to work with 300 schools selected to be representative of the country's schools and then allocate schools to incentives according to their address number, i.e., schools whose address number starts with 1, 2 or 3 will receive no bonus, schools whose address number starts with 4, 5, 6 will receive the group bonus, and schools whose address number starts with 7, 8, 9 will receive the individual bonus. The advisor proposes to look at the grades after one year by estimating the following model with OLS:

$$y_i = \alpha + \beta_{IND} IND_i + \beta_{SCH} SCH_i + u_i \tag{3}$$

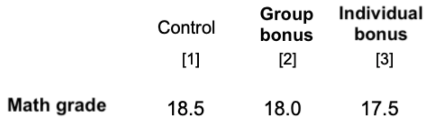
where y_i is the average grade in the school i, IND_i is a dummy variable equal to 1 if the school receives individual incentive and zero otherwise, SCH_i is a dummy variable equal to 1 if the school receives group incentive, and u_i is a residual error term. The division in groups prepared by the Minister is presented in the following table.



	NONE	GROUP BONUS	INDIVIDUAL BONUS
Number of schools	CONTROL (100 Schools)	100 Schools	100 Schools

(a) The advisor claims that the coefficients β_{IND} and β_{SCH} identifies the causal effect of providing incentives to teachers on average grades. Explain what these coefficient identify.

(b) The following table reports the means of the student performance in math one year after the introduction of the incentives for each of group (control, group incentive and individual incentive). How can you interpret the difference between the value in column [2] and the values in column [1]?



	Control	Group bonus	Individual bonus
	[1]	[2]	[3]
Math grade	18.5	18.0	17.5

5. You are the CEO of a large bank thinking about opening a new micro-finance institution in the city of Lucknow, India. You have access to a study implemented in a similar city (Hyderabab), where another micro-finance institution (MFI) opened a large number of new branches selecting their location in order to attract more customers. Your dataset comprises a random sample of citizens in the city of Hyderabad and the following variables:

- D_i: dummy variable equal to 1 if the person i has opened a micro-finance account, and 0 otherwise;
- B_i: a variable indicating the number of branches of MFI in the neighborhood where person i lives.

You are interested in estimating the marginal effect of an extra MFI branch in the neighborhood on the conditional mean of D, i.e., $\frac{\partial E[D|B]}{\partial B}$. You estimate the following regression using OLS:

$$D_i = \alpha + \beta B_i + u_i \tag{4}$$

(a) Can you identify the causal effect of B on D?

(b) How can you estimate $\frac{\partial E[D|B]}{\partial B}$ using a probit model? Detail the model you plan to apply, derive the response probability and show how this relates to $\frac{\partial E[D|B]}{\partial B}$. Assume that the error term follows a Normal distribution with mean 0 and that the variance is heteroskedastic with variance equal $\{\sigma^2 \cdot (\gamma B_i)^2\}$.

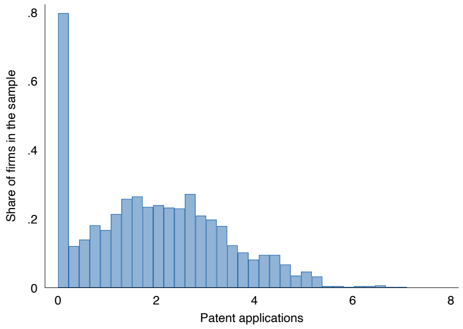
(c) The following table shows estimates of equation (4) estimated with OLS (column 1) and as a linear index model using probit (column 2). What do you learn by looking at the results?

	Dep. variable: D_i	
Estimation procedure:	OLS	Probit
	(1)	(2)
B_i	0.010***	0.030***
	(0.001)	(0.004)
Observations	11,459	11,459

Note. The constant term is omitted from the table. Standard errors in parenthesis assumes heteroskedasticity. * significant at 10 percent ** significant at 5 percent *** significant at 1 percent.

6. You are interested in studying how investments in R&D (k_i) translates into patent applications (y_i). You have access to a random sample of N firms ($i = 1, ..N$) extracted from the population of all

firms in a country. The dataset also provides you with 10 variables covering characteristics of each firms ($\mathbf{M_i} = [\ 1 \quad M_{1,i} \quad ... \quad M_{10,i}\ ]$). The following figure shows the distribution of the number of patent applications for the sampled firms, indicating that the distribution is (left) censored at zero.



(a) You are interested in computing the derivative of $E[y_i|k, X]$ with respect to k. Assume you estimate the following equation using OLS:

$$y_i = \alpha + \beta k_i + \mathbf{X_i}\gamma + v_i \tag{5}$$

where v_i is assumed to be a firm-specific unobservable. In addition, you estimate a Tobit model assuming censoring at zero and maintaining the standard assumptions of the model. Explain how do you plan to compute $\frac{\partial E[y_i|k,X]}{\partial k}$ in both models and how you would compare them.

(b) Suppose now that you have access to panel data, in which the N firms are observed for T periods, $t = 1, .., T$. Suppose you wish to estimate the following model

$$y_{it} = \alpha + \beta k_{it} + \mathbf{M_{it}}\gamma + v_{it} \tag{6}$$

using a fixed effects (FE) estimator. How does these estimates compare to the ones obtained in OLS?

7. Consider an observational (cross-sectional) study where the objective is to estimate, using a random sample of 22,501 children, whether the consumption of micro-nutrient has an effect on their nutritional status, measured by the weight-for-height (y_i). In addition to the nutritional status of the children, you have access to information about the quantity of micro-nutrient consumed in the month previous to the interview (m_i) and to mother and child characteristics at the moment of the interview ($\mathbf{X_i}$). You are interested in the following specification:

$$y_i = \alpha + \beta m_i + \mathbf{X_i}\gamma + u_i \tag{7}$$

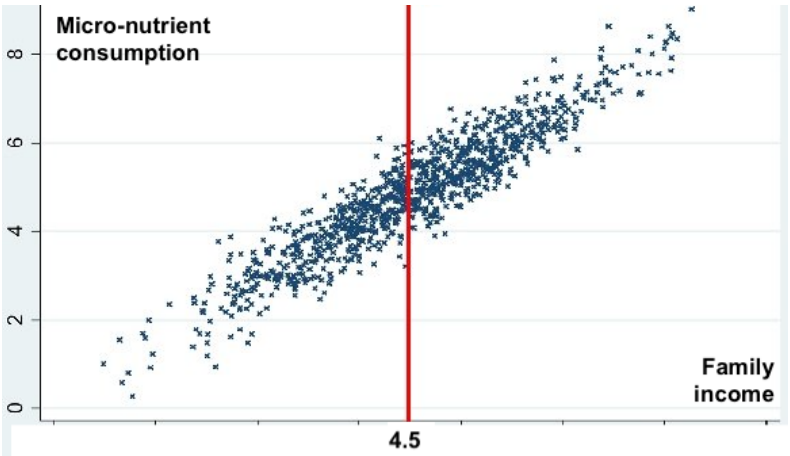
(a) With or without adding control variables, the orthogonality assumption in equation (7) is violated, i.e. $cov(m, u) \neq 0$. How can you identify the parameters in the equation using an instrumetal variable approach, (including proposing one or more variables that can be used to achieve your goal)?

(b) The following table shows estimates of equation (7) using OLS and using the IV strategy you proposed in a). What can you learn from these results?

	Dep. variable: y_i	
Estimation procedure:	OLS	IV
	(1)	(2)
Micro-nutrients consumed (m)	-0.10***	0.25***
	(0.001)	(0.002)
Observations	22,501	22,501

Note. Standard errors in parenthesis assumes clustering at the municipality level. * significant at 10 percent ** significant at 5 percent *** significant at 1 percent.

(c) According to a local administrator, there is a law that provided micro-nutrients for free for households with income less than 4.5. The consumption of micro-nutrients as function of family income is presented in the following figure. Explain whether the egression discontinuity approach would identify a different parameter and whether you use a regression discontinuity approach to estimate the impact of micro-nutrient consumption on y_i for households around the discontinuity.



8. Consider an individual i facing a choice of going to graduate school ($Y_i = 1$) or participating in the labor market ($Y_i = 0$). Let X_i be a vector of her/his observable characteristics including gender, age, grades, etc. Let

$$U_{1i} = X_i'\beta_1 + \epsilon_{1i}$$

be the utility level that i can get by going to graduate school and

$$U_{0i} = X_i'\beta_0 + \epsilon_{0i}$$

be the utility level that i can get by participating in the labor market, where ϵ_{1i} and ϵ_{0i} are random (unobserved) components of utility obtained for each choice. Assume that the individual chooses $Y_i = 1$ and $Y_i = 0$ based on which gives a higher utility.

(a) Derive the condition for enrollment in the graduate school as a function of X_i, e.g. $U_{1i} - U_{0i} > 0$

(b) Define $v_i = \epsilon_{1i} - \epsilon_{0i}$, what is the assumption on the distribution of v_i such that we can estimate a Probit model?

(c) Assume v_i is distributed normally with mean zero and stardard deviation equal to 1, e.g. $v_i \sim \phi(0, 1)$. Derive the probability that the individual i is enrolled conditional on its characteristics [HINT: prove that $Pr(Y_i = 1|X_i) = \Phi(X_i'\gamma)$ where $\Phi()$ is the cumulative distribution function of the standard normal distribution].

9. Consider the population of beer companies in Europe. Let Y_i be the logarithm of company i's (potential) export measured by sales of export, and let S_i be the size of the firm and X_i be a vector of other company i's observable characteristics (years in business, etc.). Your objective is to estimate the effect of firm size on exports using the following model:

$$Y_i = \alpha + \beta S_i + X_i' \gamma + \epsilon_i$$

   (a) What assumptions are needed to estimate this model using OLS?

   (b) Show how to identify β using the stated assumptions in point a).

   (c) Now assume that, in data, we observe X_i for all the sampled companies, while we observe Y_i only for those which actually export. Let D, indicate whether company i exports ($D_i = 1$) or not ($D_i = 0$). Explain what would be the problem if you were to estimate the model using OLS and using only the observations for which Y_i is observed?

   (d) In order to take into account the selection problem, we consider the Heckman's two-step estimation. Briefly explain how does the procedure work and how it corrects for selection.

   (e) Assume Z_i is an instrument that is assumed to affect D_i while it is assumed not to affect the sales from export. Propose an example of the instrument Z that you think it is plausible in this context, and justify its validity by providing your reasoning.

10. Suppose that you have panel data with which to study the productivity on N firms labeled $i = 1, ..N$ at each time period from $t = 1, .., T$. For each firm i in each period t you observe the (log) output y_{it}, (log) capital k_{it} and (log) labour l_{it}. Your goal is to measure the productivity in terms on output of firms in this industry as a function of their capital and labour inputs.

   (a) Suppose you wish to estimate the model

$$y_{it} = \alpha + \beta k_{it} + \gamma l_{it} + v_{it}$$

   where v_{it} is assumed to be a firm and time specific unobservable such that $E[v_{it}|k_{it}, l_{it}] = 0$. How can the model parameters (α, β, γ) be estimated?

   (b) Suppose that the unobservable v_{it} can be decomposed into two components, c_i and u_{it}, where c_i is a time-persistent component to the productivity of firm i (i.e. it depends only on the firm and not on time). The model is now

$$y_{it} = \alpha + \beta k_{it} + \gamma l_{it} + c_i + u_{it}$$

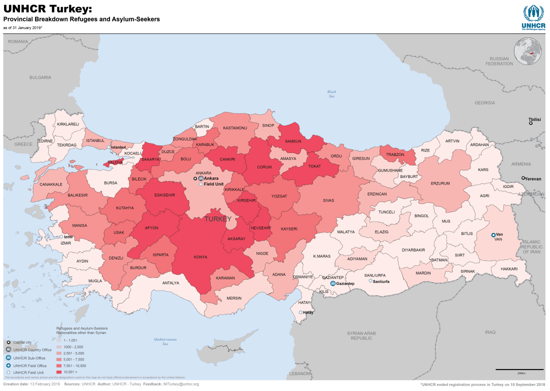
   What does the assumption of strict exogeneity of u_{it} mean? What does it rule out?

   (c) What assumptions are needed to justify the random effects estimator?

   (d) How is the random effects estimator implemented?

   (e) What assumptions does fixed effects estimation require? What would be its relative advantages and disadvantages compared to the random effects estimator?

11. You are interested in estimating the effect of a large refugee influx that happened in 2018 in Turkey on the share of votes for party P in the last elections in Turkey. You are given the following map, which indicates the distribution of these refugees across provinces. You gain access to only a partial information from the map: a dummy variable R_i, which is equal to 1 if in January 2019 the province i received more than 5000 refugees, and 0 otherwise. In addition, you obtain electoral outcomes at provincial level for three rounds of elections (2011, 2016, and 2021), and some additional characteristics of the provinces in the same years (2011, 2016 and 2021), such as population size, public expenditure, and crime rate.



(a) How can you use a two-way fixed effects (TWFE) approach to estimate the effect of refugee influx on the vote share for party P? Specify the estimating equation, the assumptions needed and their credibility in this setting.

(b) Column (1) in the following table shows estimates of the following model:

$$P_{it} = \alpha + \beta R_i + \mathbf{X_{it}}\gamma + \mu_t + \epsilon_{it} \tag{8}$$

where μ_t are time fixed effects (two dummies for year 2016 and 2021). How can you interpret the coefficient in column (1).

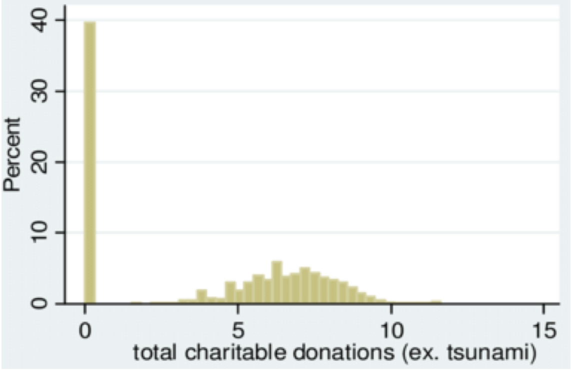
(c) The estimate in column column (2) is a difference-in-differences estimate comparing over time provinces where $R_i = 1$ with provinces where $R_i = 0$. Why do you think it is larger than in (1)?

Table 2: Effect of refugee influx

	Dependent variable: vote share for party (P_it)	
	OLS	DiD
	(1)	(2)
R_i	0.075***	0.152***
	(0.022)	(0.046)
Controls	Yes	Yes

Note. Standard errors in parenthesis. * significant at 10 % ** significant at 5 % *** significant at 1 %. Observations are at province level and cover all periods (2011, 2016 and 2021).

12. Following the 2004 Indian Ocean Tsunami, an international NGO collected donations among its members by sending individual letters asking for a donation. Since the NGO was interested in measuring the effect on donations of giving a free poster of the Indian Ocean to the members together with the letter requesting the donation, it designed a randomized experiment. Among the 2500 members, 1250 randomly selected members received the letter and the poster and 1250 received only the letter. The two groups are indicated by the variable P_i, which is equal to 1 if the member i received the letter and the poster, and equal to 0 if the member i received only the letter. In addition, you have data on donations, y, and one individual characteristics, x, which is an important determinant of donations. The following figure shows the distribution of donations.



(a) How can you estimate the average treatment effect (ATE) of providing the poster on donations?

(b) You decide to estimate the following Type I Tobit model:

$$\begin{aligned} y_i^* &= \alpha P_i + \beta x_i + u_i \\ y_i &= 1[y_i^* > 0] \end{aligned}$$

where u is distributed as a normal distribution $N(0, \sigma^2)$, and u is independent of x and P. Note that neither y_i^* nor u_i are observed. Why would you choose this model in the current setting?

(c) Derive the probability of zero donations as function of the parameters, $P[y = 0|P, x]$.

(d) Can you compare α and β estimated with the Tobit model with the following estimates computed with OLS?

$$y_i = \alpha P_i + \beta x_i + u_i$$

13. You are asked to estimate the causal effect of marriage on earnings using individual-level data for a representative sample of the population. We are interested in the following equation:

$$Y_i = \alpha + \delta M_i + \mathbf{X_i}\beta + A_i + \gamma_i + \epsilon_i \tag{9}$$

where Y_i is the wage, M_i is the marital status at the time of the interview (dummy variable equal to 1 if the individual i is married, and 0 otherwise), X_i is a set of control variables, which include ethnicity and gender of individual i, job tenure, and duration of marriage in years at the time of the interview. The variables A_i and γ_i are the (fixed) cognitive ability and non-cognitive ability of individual i, both are unobserved. ϵ_i captures remaining unobserved determinants of wages.

(a) Explain why you cannot estimate equation (9) using OLS if you are interested in the causal effect of marriage.

(b) For the same sample of individuals, you obtain panel data, such that individual i is observed for four periods $t = 1, .., 4$. In Table 3 you estimate the panel version of equation (9) using in column (1) a random effect model, and in columns (2)–(4) a fixed effects model using alternative sets of control variables. Explains the main differences in terms of assumptions of these two methodologies.

(c) Focus on columns (2) to (4). Why the coefficient decreases when you control for job tenure and for the years of marriage and its squared term?

(d) From FE estimates, can you learn something about A_i and γ_i separately?

Table 3: Estimated wage regressions

	Dependent variable: wage (in logs)			
	RE	FE	FE	FE
	(1)	(2)	(3)	(4)
Married	0.083***	0.056**	0.051**	0.033
	(0.022)	(0.026)	(0.026)	(0.028)
Job tenure	No	No	Yes	Yes
Quadratics in years married	No	No	No	Yes

Note. Standard errors in parenthesis. * significant at 10 % ** significant at 5 % *** significant at 1 %. *Quadratics in years married* introduces as controls the years of marriage and its squared term.

14. A recent paper investigated the impact on schooling achievement of installing air filters in classrooms. The author used a unique setting arising from a gas leak in the United States, whereby the offending gas company installed air filters in every classroom, office and common area for all schools within 5 miles of the leak (but none beyond).

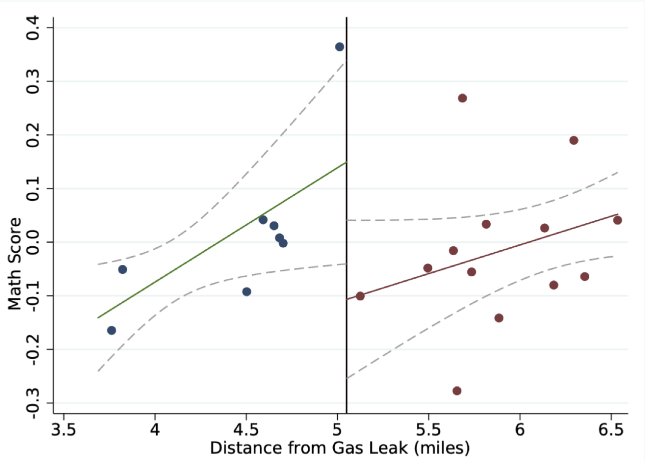
    (a) The author used a Regression Discontinuity (RD) design to identify the causal effect he wants to study. Is this a sharp or fuzzy RDD? Explain what variation does the author use and what are the keys assumptions of this identification strategy.

    (b) What causal effect is estimated using this strategy?

    (c) The author's empirical analysis can be summarized using the following figure, which gives the scatter plot of Math score two years after the gas leak (aggregated on school level) against the distances of school from the gas leak. Based on the diagram, how do you think the author estimated the impact of the air filter on math achievement?

    (d) How would you estimate the effect of the air filter?

(a) Math Score



Note. Dashed lines represent 95% confidence intervals with standard errors clustered at the school level.