# Exam 4137 – Spring 2020 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. You are supposed to answer all questions. The total marks of the examination are 180. Each subquestion in problem 1 carries 5 marks. Sub-questions in problem 2, 3 and 4 carry 10 marks each.

---

1. Give brief answers to the following questions.

(a) In 2005, following a recommendation from an expert group, many high schools started offering a homework help (Hhelp) program to their last year students. In schools that offered this program, some students participated in this program while others did not. Suppose that you have data on 50 high schools and 100 students per school for 11 years (2000-2010). You have also information on whether these students have received college degrees. What casual question can you study based on the data available?

ANSWER HINT:
The impact of Hhelp program on college completion.

(b) What are the possible sources of variation in who participated in the Hhelp program? Name at least two.

ANSWER HINT:
Here we want students to realize that the variation can come from both schools and students.
1. some schools offered the program, some did not.
2. within the same school:
a) maybe only the motivated participated.
b) maybe only those students who had difficulties in homework participated.

(c) Suppose that you run the following OLS regression:

$$D_{ist} = \alpha + \beta H_{ist} + \varepsilon_{ist}$$

where $i$,$s$,$t$ index student, school and year respectively. $D_{ist}$ is a dummy variable which takes value 1 if the student received a college degree and 0 otherwise. $H_{ist}$ is the dummy variable that denotes whether the student participated in the Hhelp program or not. What is the OLS estimate of $\beta$?

ANSWER HINT:
OLS estimate $\hat{\beta}$ is simply the mean difference in the college completion rates of those who participated and those who did not participate.

(d) What is a potential problem of the OLS standard error estimate for $\hat{\beta}$? How would you solve this problem?

ANSWER HINT:
This is a linear probability model, so the OLS standard error estimate for $\hat{\beta}$ is wrong. One can use the Heteroskadasticity robust standard errors. Or one can simply replace LPM with probit or logit.

(e) Would you consider the OLS estimate of $\beta$ as a consistent estimator of the causal effect of the Hhelp program on the probability of college completion? Why?

ANSWER HINT:
No, potential selection problem.
For example, highly motivated students may have higher participation rates and also higher college completion rates.
Or the schools which offered the program are systematically different from the schools did not.

(f) Some of your fellow students suggested that you should include school-by-year fixed effects $\omega_{st}$,

$$D_{ist} = \alpha + \beta H_{ist} + \omega_{st} + \varepsilon_{ist}.$$

Would this help? Explain your answer.

ANSWER HINT:
Maybe, this will control for the systematic differences in the the school-year combinations.
This is a within estimator.
However, this will not solve the selection problem within the same school year combination.

(g) Others suggested to aggregate $D_{ist}$ and $H_{ist}$ to the school level and estimate the following model instead

$$\bar{D}_{st} = \alpha + \beta \bar{H}_{st} + u_{st}.$$

What is the identifying variation used here?

ANSWER HINT:
Between estimator. The fraction of students who participated in the program varies over school and year.

(h) Would you prefer model suggested in $(g)$ or model suggested in $(f)$? Discuss. Do you have better suggestions?

ANSWER HINT:
Both solve only partially the problem. So hard to decide.
IV would be one way to deal with the selection problem: for example, we can instrument $H_{ist}$ in (c) by the dummy variable whether the school is offering the program.

2. Suppose that you want to study the effect of a weight-loss drug. We know that the effect of the drug is different for different people. In a study, 200 people are randomly assigned into two groups: Some people are given the weight-loss drugs (the treatment group) while the rest are not (the control group). Suppose that only 60% of the people assigned to take the drug actually took the drug, and all people who are not assigned did not take the drug.

   (a) You observe that the average weight loss for the control group is 7 kg while the average weight loss for the treatment group is 13 kg. One of your fellow students claims that the average treatment effect of the drug is 6 kg, since the treatment assignment is random. Do you agree? Explain your answer.

   ANSWER HINT:
   No. Not every people in the treatment group is treated. What he gets is only ITT not ATE.

   (b) You suggest using the IV technique as the identification strategy. In IV analysis, it is often useful to classify people into different groups, such as always takers, etc. Define these groups and classify the people who participate in the study into these groups. How many people are in each group?

   ANSWER HINT:

   i. **Compliers**: $d_i(1) = 1$ and $d_i(0) = 0$;
   people who take the drug if assigned to the treatment group, but not if not assigned.

   ii. **Always-takers**: $d_i(1) = 1$ and $d_i(0) = 1$;
   people who always take the drug,regardless of their assignment status

   iii. **Never-takers**: $d_i(1) = 0$ and $d_i(0) = 0$;
   people who never take the drug, regardless of their assignment status

   iv. **Defiers**: $d_i(1) = 0$ and $d_i(0) = 1$;
   people who don't take the drug if assigned to treatment group, but will take if not assigned..

   Since all people in the control group don't take the drugs – so there are no always taker and defiers in this study.
   $P(\text{never taker}) = P(D = 0|Z = 1) = 0.4$, number of never takers=200*0.4=80
   and
   $P(\text{complier}) = 1 - P(\text{never taker}) = 0.6$, number of compliers=200*0.6=120

   (c) What treatment (causal) effects you can identify in this study? Use the information given above to derive your estimate.

   ANSWER HINT:
   We can identify LATE, however, since in this study, we have one-sided compliance, ATT=LATE.
   So we can identify ATT.
   $ATT = LATE = \frac{E(y|z=1)-E(y|z=0)}{E(d|z=1)-E(d|z=0)} = \frac{6}{0.6} = 10$

(d) Suppose that now only 2% of the people assigned to take the drug actually take the drug, will it be problematic for your identification strategy? Explain.

ANSWER HINT:
There are too few compliers for our instrument ⇒ the relationship between the treatment and the instrument is weak!
Weak instrument problem:
Here we have only one endogenous variable and one IV, the weak instrument does not create bias per se. If the treatment assignment is truly random, and the sample size is large enough, the IV estimate will still give you consistent estimate of ATT. However, small violations can lead to large bias. So find a better IV!

3. There are four suppliers (Company $A$,$B$, $C$, and $D$) for a semi-conduct product. A market analysis firm collected data from $n = 88$ customers of this product. They asked each customer to provide assessments of these four companies on several attributes using a nine point scale (the higher the score, the better the performance of the company on that particular attribute.)

   The data include:

   *CustomerID, Choice, Price, Support, Quality,* and *CompanyID*

   (a) Specify a simple logit model using only dummy variables for companies. What parameters can be identified in this model? Write down the log-likelihood functions, and outline how one can estimate the model using MLE.

   ANSWER HINT:
   Define $y_{ij} = 1$,if individual $i$ chooses company $j$
   define the utility of choosing company $j$ for customer $i$ as

   $$U_{ij} = \alpha_j + \varepsilon_{ij}$$

   where $\varepsilon_{ij}$ i.i.d extreme value distributed. Then we have

   $$sPr(chosen = j) = \frac{\exp(\alpha_j)}{\sum_{k=1}^{4} \exp(\alpha_k)}$$

   One cannot identify all $\alpha_k$, need to normalize one of the parameters.
   log-likelihood function: $\log L = \sum_{i=1}^{N} \sum_{j=1}^{4} [y_{ij} \log P(y_{ij} = 1)]$
   write down log-likelihood, and maximize. ie numerically iterate until first order conditions evaluated at the parameter values are zero, and second derivative is negative. the main idea is that there are no other parameters under which our data are more likely to occur.

   (b) Explain briefly what is meant by the term IIA and state its main implication.

   ANSWER HINT:
   IIA. Independence from Irrelevant Alternatives: the relative probability of someone choosing between two options (j and h) is independent of any additional alternatives in the choice set. In other words, the attributes of other alternatives than j and h do not matter for choosing j over h.

   (c) Suppose that IIA holds, use the model you specified in (a) to predict the expected market shares for the other companies, if company $A$ had bankrupted and exited the market (Hint: all necessary numbers are given in the attached Stata output).

ANSWER HINT:
We could have used the estimates from the model in (a) directly to predict the market share when A goes bankrupt. However, the results are not included.
IIA holds implies that odds ratio between any two alternatives does not depend on other alternatives in the choice set. So the conditional logit model above predicts that the market share for company A gets divided by the rest 3, proportional to their original market share.
Let the market share of company $j$ as $P_j$ then we have

$$\begin{cases} P_2/P_3 = 0.26/0.30 \\ P_4/P_3 = 0.24/0.30 \\ P_2 + P_3 + P_4 = 1 \end{cases}$$

So

$$\begin{cases} P_2 \simeq 0.325 \\ P_3 \simeq 0.375 \\ P_4 \simeq 0.3 \end{cases}$$

Now add the attribute information into the logit model specified in (a). Based on the stata output and the critical value table for chi-square, answer the following questions:

(d) Are these attributes jointly significant?

ANSWER HINT:
Joint hypothesis that all attributes parameters=0.
Wald test is included in the stata output.
degree of freedom 3, wald statistic 47.69.
Reject the hypothesis.

(e) You are hired by company $D$ to find the most effective way to increase its market share, suppose that the costs of improvement are the same for different attributes, what attribute would you recommend company $D$ to improve.

ANSWER HINT:
the marginal effects of attribute $k$ which is proportional to $\beta_k$ , since

$$\frac{\partial \Pr(y_{ij} = 1|x_{ij})}{\partial x_{ijk}} = p_j(x_i)(1 - p_j(x_i))\beta_k$$

and the cost is the same. So we just pick the attribute with largest $\hat{\beta}$, which is “support”.

(f) What is the probability for customer 1 to choose company $D$? Calculate the marginal effect of attribute “support” of company $D$ on this probability.

ANSWER HINT:
the probability is 0.032

$$ME_i(x_{ij}) = \frac{\partial \Pr(y_{ij} = 1|x_{ij})}{\partial x_{ij}} = p_j(x_i)(1 - p_j(x_i))\beta$$

ME=$0.032 * (1 - 0.032) * 1.53 \simeq 0.05$

## Stata Output

```
. su customerid choice price support quality companyid
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  customerid |        352        44.5    25.43793          1         88
      choice |        352         .25    .4336291          0          1
       price |        352    4.511364    1.244881          1          7
     support |        352    5.857955    1.140922          3          9
     quality |        352    4.735795    1.112515          2          7
   companyid |        352         2.5    1.119625          1          4

 . su price support quality if companyid==4
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         88    4.193182    1.337848          1          7
     support |         88    5.534091    1.212472          3          8
     quality |         88    4.431818    1.266686          2          7


. tabulate companyid, su(choice)
            |          Summary of choice
  companyid |        Mean   Std. Dev.       Freq.
------------+------------------------------------
  Company A |   .20454545   .40568067          88
  Company B |   .26136364   .44189556          88
  Company C |   .29545455   .45886143          88
  Company D |   .23863636    .4286927          88
------------+------------------------------------
      Total |         .25   .43362909         352

. asclogit choice price support quality, case(customerid) alternative(companyid) base(4) noheader

 ------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
companyid    |
       price |    1.15406   .2594101     4.45   0.000     .6456259    1.662495
     support |   1.527176   .2690022     5.68   0.000     .9999415    2.054411
     quality |   .7111528   .2642696     2.69   0.007     .1931938    1.229112
-------------+----------------------------------------------------------------
Company_A    |
       _cons |  -.7128381   .4507301    -1.58   0.114    -1.596253    .1705766
-------------+----------------------------------------------------------------
Company_B    |
       _cons |  -.6052823   .4392791    -1.38   0.168    -1.466254    .2556889
-------------+----------------------------------------------------------------
Company_C    |
       _cons |  -.6579015    .447554    -1.47   0.142    -1.535091    .2192882
-------------+----------------------------------------------------------------
Company_D    |  (base alternative)
------------------------------------------------------------------------------

. test price=support=quality=0

 ( 1)  [companyid]price - [companyid]support = 0
 ( 2)  [companyid]price - [companyid]quality = 0
 ( 3)  [companyid]price = 0

           chi2( ?) =   47.69

. predict p_pred, pr
. list customerid companyid p_pred in 1/4
     +--------------------------------+
     | custom~d   companyid     p_pred |
     |--------------------------------|
  1. |        1   Company A   .2270369 |
  2. |        1   Company B   .2528176 |
  3. |        1   Company C   .4884329 |
  4. |        1   Company D   .0317126 |
     +--------------------------------+


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
```

4. A recent paper investigated the achievement impact of installing air filters in classrooms. The author utilized 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 five miles of the leak (but none beyond).

   (a) The author used a Regression Discontinuity Design (RDD) to identify the causal effect he wants to study. Is this a sharp or fuzzy RDD? What variation does the author use?

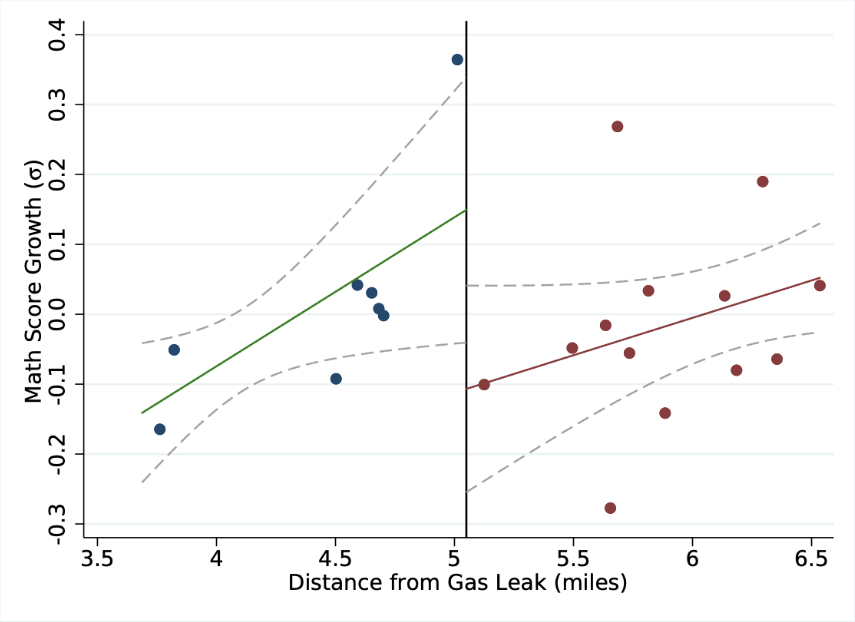
   > ANSWER HINT:
   > Sharp RDD, since the air filters are installed for all schools within the cutoff, but none beyond.
   > The author uses variation in the installation of air filters to compare student achievement in schools receiving air filters relative to those that did not.

   (b) What is the key assumption of this identification strategy? What causal effect is estimated using this strategy?

   > ANSWER HINT:
   > The key assumption is that the schools/students on either side of the cutoff are similar. More formally, this is an assumption that continuity of potential outcomes at the cutoff.
   > Sharp RDD actually gives a very specific parameter, the $ATE$ at the cut off point.

   (c) Although the actual empirical setup is more sophisticated, the essence of the author's empirical analysis can be summarized using the following diagram, which gives the scatter plot of Math score growth rates before and after the gas leak (aggregated on school level) against the distances of school from the gas leak. Dashed lines represent 95% confidence intervals with standard errors clustered at the school level. How would you estimate the effect of the air filter? Based on the diagram, what can you say about the impact of the air filter on math achievement?

(a) Math Score Growth



ANSWER HINT:
1. cut off is 5 miles from the gas leak.
2. Linear regression is used on both sides of the cutoff point.
3. The causal effect is represented by the distance between two intersection points generated by: the green line and the cut off line vs the red line and the cut off line.
4. Although the overlapping of 95% CI does not necessarily imply that the difference is insignificant, the diagram does not provide very strong evidence that air filter has a positive effect.
5. The linear regression seem to suggest that the score growth increases with the distance to gas leak (both within and outside the cutoff). Why? Discussions on this should be given additional credit.

(d) You find some documents that show that there have been multiple waves of temporary reassignments of students within the 5 miles of the gas leak to different schools after the accidents. In the light of this finding, what will be your evaluation of the above analysis?

ANSWER HINT:
The internal validity of RDD may fail since the cutoff is used for other policies than the air filter. The observed effect, if any, could be the effect of both factors. It is wrong to interpret this as solely from the air filter.