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?

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

   (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 β?

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

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

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

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

   Would this help? Explain your answer.

   (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?

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

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.

   (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?

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

   (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.

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.

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

   (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).

   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?

   (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.

   (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.

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
```

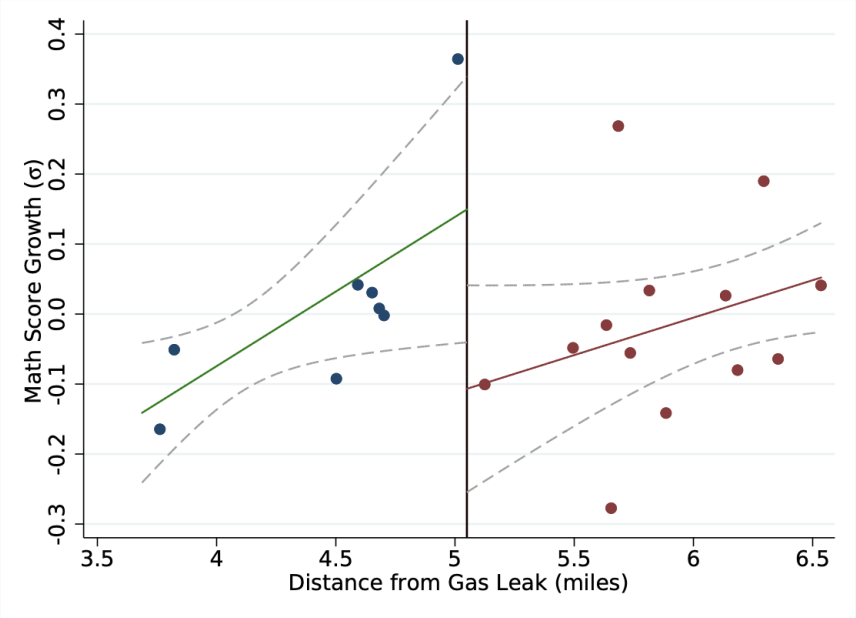
4. A recent paper investigated the achievement impact of installing air filters in classrooms. The author utilized a unique setting arising from a gas leak in the United States, whereby the offending gas company installed air filters in every classroom, office and common area for all schools within five 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?

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

   (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



   (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?