1. Give brief answers to the following question.

   (a) Explain why Random Controlled Trial (RCT) is sometimes argued to be the "gold standard" for empirical causal research. Suppose that you want to study the impact of number of kids on household income for married couples in Norway. Can RCT be employed to study this particular question?

   (b) Instead, you collect some information from a sample of married couples. For couple i, you observe household income y_i, the number of kids x_i and average education years for the couple z_i. One of your friends suggests to run the following regression:

   $$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i. \tag{1}$$

   What is the main assumption needed for the Ordinary Least Square (OLS) estimator of β to have a causal interpretation?

   (c) Suppose that the assumption you made in (b) holds and $cov(x_i, z_i) < 0$,what can you say about the OLS estimate from a bi-variate regression of y_i on x_i in this case?

   (d) Suppose that the variation of income for couples with more than two kids is much larger than those with two or less kids, what potential problem will this cause when estimating (1)? How will you deal with this problem?

   You are worried that some other variables such as religion, region and country of origin may affect both the household's decisions to have kids and its income. Unfortunately, you do not have data on these variables.

   (e) Suppose that you have information on whether the couple had twin births. Would this information help you to identify the casual effect of interest? What assumptions will you need to make? Do you think these assumptions are likely satisfied? Explain.

   (f) Explain briefly in words how you would proceed to estimate the model in case (e).

   (g) Suppose instead that you have observations of household income both before and after couples having an extra kid. How will you estimate the causal effect in this case?

   (h) Discuss when your method in case (g) will lead to inconsistent estimate of the causal effect of interest.

2. We want to study the relationship between hospital visits and individual characteristics. Suppose that conditional on the individual characteristics X_i, the number of hospital visits Y_i follows a Poisson distribution with parameter λ_i:

$$Prob[Y_i = y_i] = exp(-\lambda_i)\lambda_i^{y_i}/y_i!$$

where

$$\lambda_i = exp(\beta X_i).$$

Note that if $Y \sim Poisson(\lambda)$, then $f(y;\lambda) = \lambda^y e^{-\lambda}/y!$, and $E[Y] = Var(Y) = \lambda$.

We have collected information of hospital visits from 27326 individuals. We also have information on the following individual characteristics.

Edu	Years of education
Female	Female dummy (1=Female, 0=Male)
Married	Marital status (1 if married, 0 otherwise)
Age	Age
Age^2	Age square

(a) Outline in words how you would estimate β using maximum likelihood, and explain the intuition behind the method.

(b) Assume that $\lambda_i = exp(\beta_0 + \beta_1 Edu_i)$. Write down the log-likelihood function excluding constant terms that do not depend on λ (i.e. the denominator $y_i!$) and derive the score function.

(c) Use the result you obtained in (b), show that the sample mean of the estimated λ_i (that is, the estimates of λ_i when you plug in the data and the maximum likelihood estimates of the parameters) equals the sample mean of hospital visits.

In the following, we assume that $\lambda_i = exp(\beta_0 + \beta_1 Edu_i + \beta_2 Female_i + \beta_3 Married_i + \beta_4 Age_i + \beta_5 Age_i^2)$.

(d) Based on the attached Stata output, test the hypothesis that the number of visits is unrelated to marriage status using the Wald test. You can use the attached critical value table.

(e) Compute the average marginal effect of an additional year in education on the expected number of visits.

(f) Carry out a likelihood ratio test of the hypothesis that the five coefficients on Edu, $Female$, $Married$, Age and Age^2 are all zero.

Stata Output

```
. su hospvis educ female married age age2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     hospvis |     27,326    .1382566     .884339          0         51
        educ |     27,326    11.32063    2.324885          7         18
      female |     27,326    .4787748    .4995584          0          1
     married |     27,326    .7586182    .4279291          0          1
         age |     27,326    43.52569    11.33025         25         64
        age2 |     27,326    2022.855    1004.078        625       4096
-------------+---------------------------------------------------------


. ml model lf loglikelihoodi (hospvis = educ female married age age2)
. ml maximize

                                                Number of obs     =     27,326

Log likelihood = -11168.252
------------------------------------------------------------------------------
     hospvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0734901   .0084125    -8.74   0.000    -.0899782   -.0570019
      female |   .0802982   .0333027     2.41   0.016      .015026    .1455703
     married |  -.0384616   .0396869    -0.97   0.332    -.1162465    .0393234
         age |   -.013886   .0125443    -1.11   0.268    -.0384724    .0107004
        age2 |   .0002544   .0001393     1.83   0.068    -.0000185    .0005274
       _cons |  -1.089854   .2807546    -3.88   0.000    -1.640123   -.5395851
------------------------------------------------------------------------------


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
```

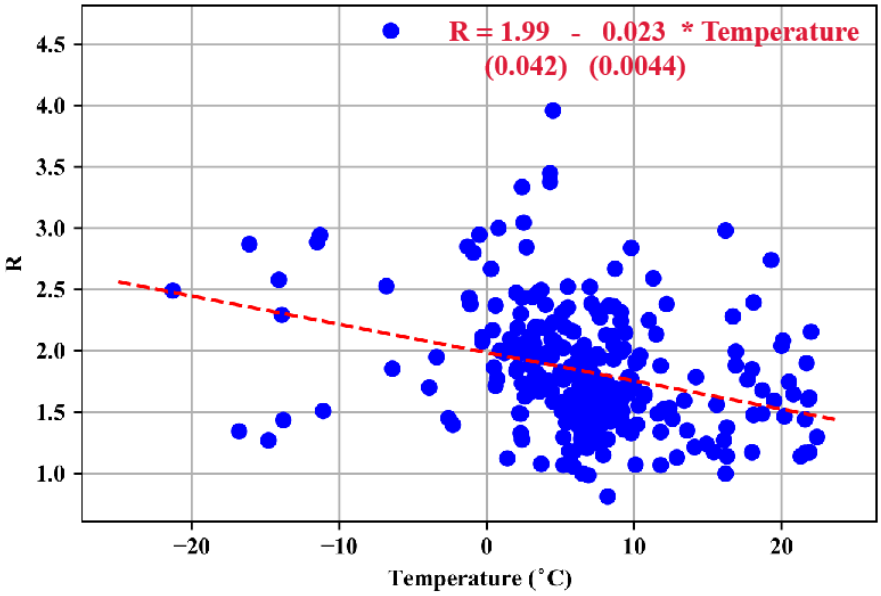
3. A recent paper investigated how air temperature influences the transmission of COVID-19. The authors calculated the daily effective reproductive number, R_{it}, for 100 Chinese cities from January 21 to January 23, 2020. They also collected the average daily temperature x_{it} and humidity z_{it} for these cities during the same periods. They then run the following regression:

$$R_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \varepsilon_{it} \tag{2}$$

   (a) What is the parameter of interest? What variation in the data did the authors use to estimate it in (2)?

   (b) The authors believe that medical condition and crowdedness inside a city are important factors that influence the reproductive number R. Thus, they also included population density and the GDP per capita (both measured in year 2018) in the regression. Do you think this is necessary given their purpose?

   (c) The following diagram gives the scatter plot of R_{it} and x_{it} based on the data for the analysis. What can you learn from this diagram? Do you think the authors' conclusion that higher temperature reduce R reasonable, why?



Utilizing the panel structure of data, the authors also estimated a fixed effect and a random effect model, and the estimation results are given in the following table:

	Fixed Effects		*Random Effects*	
	Estimate	*T-statistics*	Estimate	*T-statistics*
Temperature	-0.0383	-3.27	-0.024	-3,97
Humidity	-0.0224	-10.18	-0.020	-3.06
constant	3.968	19.04	3.877	19.60
Additional controls			✓	
R^2	17%		19%	

(d) What type of variation is used to identify the parameter of interest when estimating a fixed effect model? What assumptions are needed for the estimate from the fixed effect model to be consistent?

(e) Interpret the estimates for the fixed effect model, assuming that the assumptions required are satisfied.

(f) Explain the difference between a random effect model and a fixed effect model. The authors reported that p value of a Hausman test is 0.06. Explain briefly what a Hausman test is. Based on the test result, which model would you choose?

(g) The authors claimed that the results from the fixed effect model provide a way to predict the R values of a city given its temperature x and humidity z:

$$R = 3.97 - 0.038x - 0.022z$$

They then proceeded to use this equation to predict the R values for cities worldwide, using the 2019 values of March and July temperature and humidity for these cities. For example, they estimated the R value for Tokyo in July will be below 1. Suppose for now that the assumptions for the fixed effect model are satisfied. If you work for the International Olympic Committee (IOC), what suggestion would you give to IOC w.r.t the Tokyo 2020 Olympic Games? Explain your answer.