1. (10 points)

   (a) Consider the following cross-sectional stochastic frontier regression model

$$y_i = \alpha + X_i'\beta - \mu_i + v_i, \quad i = 1, 2, \ldots, N, \tag{1}$$

   where y_i is the output, X_i is a vector of inputs, v_i is the regression error term, and $\mu_i > 0$ is the technical inefficiency of the firm i. In order to identify μ_i, what assumptions are typically needed?

   (b) Consider the following panel data stochastic frontier regression model

$$y_{it} = \alpha + X_{it}'\beta - \mu_i + v_{it}, \quad i = 1, 2, \ldots, N; \quad t = 1, 2, \ldots, T, \tag{2}$$

   In order to identify μ_i, what assumptions are typically needed? In particular, do we have to assume the distribution of μ_i or not? If not, how to estimate the technical inefficiency μ_i? Please describe the procedure.

2. (10 points) Consider the following linear regression model

$$Y_i^* = \alpha + \beta X_i^* + u_i, \quad i = 1, 2, \ldots, n, \tag{3}$$

   where Y_i^* and X_i^* are the true values.

   (a) For the regression model in Equation (3), suppose we observe the true value X_i^*. However, we do not observe the true value Y_i^*. Instead, we observe a "contaminated" variable Y_i that contains measurement error. The relationship between Y_i and Y_i^* is:

$$Y_i = Y_i^* + \epsilon_i, \tag{4}$$

   where the measurement error ϵ_i has mean 0 and variance σ_ϵ^2. Using equations (3) and (4), derive Y_i as a function of X_i^*. What happens to the estimator of β from a OLS regression of Y_i on X_i^*? Is it unbiased? consistent? efficient?

   (b) For the regression model in Equation (3), suppose we observe the true value Y_i^*. However, we do not observe the true value X_i^*. Instead, we observe a "contaminated" variable X_i that contains measurement error. The relationship between X_i and X_i^* is:

$$X_i = X_i^* + e_i, \tag{5}$$

   where the measurement error e_i has mean 0 and variance σ_e^2. Using equations (3) and (5), derive Y_i^* as a function of X_i. What happens to the estimator of β from a OLS regression of Y_i^* on X_i? Is it unbiased? consistent? efficient?

3. (20 Points)

   (a) Consider the following panel data Logit/Probit regression model

$$Y_{it}^* = \alpha + X_{it}'\beta + \mu_i + v_{it}, \tag{6}$$

   where Y_{it}^* is a latent variable that is unobservable. Instead, we observe $Y_{it} = 1$ if $Y_{it}^* > 0$; and $Y_{it} = 0$ if $Y_i^* \leq 0$. For example, consider a regression of female workers' union membership `union` over worker's age `age`, years of education `edu` and a dummy variable of from south or not `south`. The *fixed effects* Logit regression result is as follows.

```
. xtlogit union age edu south, fe

Conditional fixed-effects logistic regression      Number of obs     =     12,035
Group variable: id                                 Number of groups  =      1,690

                                                   Obs per group:
                                                                 min =          2
                                                                 avg =        7.1
                                                                 max =         12

                                                   LR chi2(3)        =      68.09
Log likelihood = -4516.1412                        Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       union | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0170305   .004146     4.11   0.000     .0089045    .0251564
         edu |   .0853099   .0418742    2.04   0.042     .0032379    .1673819
       south |  -.7474497   .1249313   -5.98   0.000    -.9923105   -.5025889
------------------------------------------------------------------------------
```

Brief interprete the meaning of these (positive/negative) sign of the coefficients. In particular, can we interprete these coefficients as the marginal effects?

(b) We can run a *random effects* Logit regression as well. Its regression result is shown below.

```
. xtlogit union age edu south, re

Random-effects logistic regression                 Number of obs     =     26,200
Group variable: id                                 Number of groups  =      4,434

Random effects u_i ~ Gaussian                      Obs per group:
                                                                 min =          1
                                                                 avg =        5.9
                                                                 max =         12

Integration method: mvaghermite                    Integration pts.  =         12

                                                   Wald chi2(3)      =     211.72
Log likelihood = -10549.561                        Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       union | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0266316   .0036669     4.54   0.000     .0094447    .0238185
         edu |   .1923225   .017606      5.24   0.000     .0578153    .1268297
       south |  -.9657657   .0801091   -12.06   0.000    -1.122777   -.8087547
       _cons |  -3.708043   .2444072   -15.17   0.000    -4.187073   -3.229014
-------------+----------------------------------------------------------------
```

```
    /lnsig2u |   1.754189   .0469723                      1.662125    1.846254
-------------+----------------------------------------------------------------
     sigma_u |   2.403906   .0564585                      2.295757    2.517149
         rho |   .6372255   .0108586                      .6156862    .6582277
------------------------------------------------------------------------------
LR test of rho=0: chibar2(01) = 6033.56                  Prob >= chibar2 = 0.000
```

Note that the fixed effects Logit and random effects Logit are somewhat different. In this case, which one do we prefer? Why?

(c) In Stata, the fixed effects Probit regression is not available. In other words, the following codes will cause an error message.

```
. xtprobit union age edu south, fe
fixed-effects model not available
```

Please explain why fixed effects Probit is not available.

(d) Since the fixed effects Probit is not available in Stata, a researcher suggests to "manually" obtain it by using the following code.

```
. probit union age edu south i.id
```

Please discuss if this code produces the correct fixed effects Probit regression results. In particular, please explain what the code `i.id` does.

4. (20 Points)

(a) Consider an AR(1) process

$$y_t = \rho y_{t-1} + e_t, \tag{7}$$

where e_t is a independent white noise process with variance σ_e^2. When will y_t be a stationary process and when will y_t be a nonstationary process? Please briefly explain what a nonstationary process is.

(b) Variables x_t and y_t are both nonstationary $I(1)$ processes, i.e.,

$$x_t = x_{t-1} + \varepsilon_t, \tag{8}$$

and

$$y_t = y_{t-1} + e_t, \tag{9}$$

where ε_t and e_t are independent white noise process with variance σ_ε^2 and σ_e^2, respectively. A researcher uses the following regression model

$$y_t = \alpha + x_t\beta + u_t, \quad t = 1, \ldots, T \tag{10}$$

where α and β are scalars. The OLS regression results are as follows.

```
. reg y X

      Source |       SS       df       MS      Number of obs   =       100
-------------+------------------------------   F(1, 98)        =     91.43
       Model |  1205.11227     1  1205.11227   Prob > F        =    0.0000
    Residual |  1291.65165    98  13.1801188   R-squared       =    0.4827
```

```
-------------+------------------------------           Adj R-squared   =    0.4774
       Total |  2496.76392        99  25.2198376      Root MSE        =    3.6304

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           X |   .6605148   .0586183     9.56   0.000     .4441888    .8768409
       _cons |   9.908345   .3664315    27.04   0.000     9.181173    10.63552
------------------------------------------------------------------------------
```

Please discuss if the regression results are reliable. Do you think if there is a significant causal effect of x on y?

(c) Similarly, x_{it} and error term y_{it} are both nonstationary $I(1)$ processes, i.e.,

$$x_{it} = x_{i,t-1} + \varepsilon_{it}, \tag{11}$$

and

$$y_{it} = y_{i,t-1} + e_{it}, \tag{12}$$

where ε_{it} and e_{it} are independent white noise process with variance σ_ε^2 and σ_e^2, respectively. Consider the following panel data regression model

$$y_{it} = \alpha + x_{it}\beta + u_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T \tag{13}$$

where $u_{it} = \mu_i + \nu_{it}$, and α and β are scalars. The panel data regression results are as follows.

```
. xtreg y X, fe

Fixed-effects (within) regression               Number of obs      =     5,000
Group variable: id                              Number of groups   =       100

R-squared:                                      Obs per group:
     Within  = 0.0006                                         min =        50
     Between = 0.0160                                         avg =      50.0
     Overall = 0.0082                                         max =        50

                                                F(1,4899)          =      2.78
corr(u_i, Xb) = 0.0832                          Prob > F           =    0.0955

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           X |  -.0233038   .0139777    -1.67   0.096    -.0507064    .0040988
       _cons |   .4509879   .0433459    10.40   0.000     .3660105    .5359652
-------------+----------------------------------------------------------------
     sigma_u |  4.0718206
     sigma_e |  2.9365069
         rho |   .65785252   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(99, 4899) = 95.47                     Prob > F = 0.0000
```

Please discuss if the regression results are reliable. Do you think if there is a significant causal effect of x on y?

(d) The following results show the panel unit root test. Based on the results, clearly write down the H_0, H_1 and your conclusion.

```
. xtunitroot breitung y, trend

Breitung unit-root test for y
------------------------------
H0: Panels contain unit roots                Number of panels  =    100
Ha: Panels are stationary                    Number of periods =     50

AR parameter: Common                         Asymptotics:  T,N -> Infinity
Panel means:  Included                                     sequentially
Time trend:   Included                       Prewhitening: Not performed
------------------------------------------------------------------------------
                     Statistic      p-value
------------------------------------------------------------------------------
 lambda               -0.1444       0.4426
------------------------------------------------------------------------------

. xtunitroot breitung X, trend

Breitung unit-root test for X
------------------------------
H0: Panels contain unit roots                Number of panels  =    100
Ha: Panels are stationary                    Number of periods =     50

AR parameter: Common                         Asymptotics:  T,N -> Infinity
Panel means:  Included                                     sequentially
Time trend:   Included                       Prewhitening: Not performed
------------------------------------------------------------------------------
                     Statistic      p-value
------------------------------------------------------------------------------
 lambda               0.5668       0.7146
------------------------------------------------------------------------------
```

5. (20 Points)

(a) To allow spatial correlation, consider the following fixed effects spatial error regression model is

$$y_t = X_t\beta + \mu + v_t, \quad t = 1, 2, \ldots, T, \tag{14}$$

and

$$v_t = \lambda W_N v_t + \epsilon_t, \tag{15}$$

where $y_t = (y_{t1}, \cdots, y_{tN})'$ is a $N \times 1$ vector. Similarly, X_t is a $N \times k$ matrix. v_t and ϵ_t are $N \times 1$ vectors. For example, the following Fixed Effects regression estimate a Cobb-Douglas production function that relates the gross social product (gsp) of a given state to the input of public capital (pcap), private capital (pc), labor (emp) and state unemployment rate (unemp). The fixed effects spatial error regression result is as follows.

```
. * Spatial Error Model (SEM)
. xsmle lngsp lnpcap lnpc lnemp unemp, ematrix(usaww) model(sem) fe hausman

SEM with spatial fixed-effects                    Number of obs =       816

Group variable: state                             Number of groups =     48
Time variable: year                               Panel length =         17

R-sq:     within  = 0.9401
          between = 0.9907
          overall = 0.9896

------------------------------------------------------------------------------
       lngsp | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
Main         |
      lnpcap |   .0051441   .0250122     0.21   0.837    -.0438789    .0541671
       lnpc |   .2053021   .0237393     8.65   0.000     .158774     .2518303
      lnemp |   .7822543   .0278151    28.12   0.000     .7277378    .8367708
      unemp |  -.0022317   .0010861    -2.05   0.040    -.0043603    -.000103
-------------+----------------------------------------------------------------
Spatial      |
     lambda |   .5574017   .0329539    16.91   0.000     .4928132    .6219902
-------------+----------------------------------------------------------------
Variance     |
    sigma2_e |   .0009765   .0000498    19.60   0.000     .0008789    .0010741
------------------------------------------------------------------------------
Ho: difference in coeffs not systematic chi2(5) = 10.99    Prob>=chi2 = 0.0516
------------------------------------------------------------------------------
```

Based on the Stata outputs, between the fixed effects estimator and fixed effects spatial error estimator, which one do we prefer? Why? If we ignore the spatial correlation and use the fixed effects estimator, will it be unbiased? consistent? efficient?

(b) The fixed effects spatial lag regression model is

$$y_t = \rho W_N y_t + X_t \beta + \mu + v_t, \quad t = 1, 2, \ldots, T, \tag{16}$$

Its regression result is shown as below.

```
. * Spatial Lag Model, also called Spatial Autoregressive Model (SAR)
. xsmle lngsp lnpcap lnpc lnemp unemp, wmatrix(usaww) model(sar) fe hausman
SAR with spatial fixed-effects                    Number of obs =       816

Group variable: state                             Number of groups =     48
Time variable: year                               Panel length =         17

R-sq:     within  = 0.9433
          between = 0.9557
          overall = 0.9547
```

```
------------------------------------------------------------------------------
       lngsp | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
Main         |
      lnpcap |  -.0465815   .0264001    -1.83   0.067    -.0963647    .0032017
        lnpc |   .1874323   .0233795     8.02   0.000     .1416094    .2332553
       lnemp |   .6250903   .028505     21.93   0.000     .5692215     .680959
       unemp |  -.0044816   .0008666    -5.17   0.000      -.00618   -.0027832
-------------+----------------------------------------------------------------
Spatial      |
         rho |   .2746886   .0210851    13.03   0.000     .2333625    .3160147
-------------+----------------------------------------------------------------
Variance     |
    sigma2_e |   .0011114   .0000551    20.16   0.000     .0010033    .0012194
------------------------------------------------------------------------------
Ho: difference in coeffs not systematic chi2(5) =  3.22    Prob>=chi2 = 0.6659
------------------------------------------------------------------------------
```

Based on the Stata outputs, between the fixed effects estimator and fixed effects spatial lag estimator, which one do we prefer? Why? If we ignore the spatial correlation and use the fixed effects estimator, will it be unbiased? consistent? efficient?

(c) Read the following Stata outputs of Pesaran CD test. Please clearly write down H_0, H_1. Based on the statitics and p-value, make your conclusion.

```
. xtreg lngsp lnpcap lnpc lnemp unemp, fe
<output omitted>
.
. * Pesaran CD test
. xtcsd, pesaran

Pesaran's test of cross sectional independence =    30.368, Pr = 0.0000
```

6. (20 points) A Difference-in-Difference regression model is

$$Y_{it} = \alpha + \beta_1 D_i + \beta_2 D_t + \beta_3 D_i \times D_t + u_{it}, \tag{17}$$

where D_i is a dummy variable of experimental group, D_t is a dummy variable of after the policy change, $D_i \times D_t$ is their interaction term.

(a) Among coefficients β_1, β_2 and β_3, which one is the 'Difference-in-Difference' estimator, i.e., which one do we care about?

(b) Alternatively, one can obtain the 'Difference-in-Difference' estimator from a 2×2 table. Comparing to the table, what are the advantages of a regression?

(c) What are the potential assumptions needed for a 'Difference-in-Difference' estimator?

(d) Di Tella and Schargrodsky (2004) published a paper "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack" in *The American Economic Review*. After a terrorist attack on the main Jewish center in Buenos Aires, Argentina, in July 1994, all Jewish institutions received police protection. Thus, does the police protection reduce crime, in

particular, car thefts? The regression model in Di Tella and Schargrodsky (2004) can be simplified as:

$$CarThefts_{it} = \alpha + \beta_1 Protected_i + \beta_2 After_t + \beta_3 Protected_i \times After_t + X'_{it}\beta + u_{it}, \quad (18)$$

where $CarThefts_{it}$ is the number of car thefts in street block i, at time t; $Protected_i = 1$ if there is a protected Jewish institution in the street block, =0 otherwise; $After_t = 1$ if after the terrorist attack, =0 otherwise. They find a large negative effect of police on crime. What are the assumptions needed in Di Tella and Schargrodsky (2004)? Please discuss if these assumptions hold or not.