All students have to hand in this assignment by 11:59pm Sunday 18 October 2020. **Please submit both your answers and program**.

1. Apartment developers sometimes add extra greenery, like more trees, to the premises of an apartment project, hoping that it leads to higher apartment prices. Suppose you estimate a simple linear regression with dependent variable $\ln price$, log price of an apartment, and the lone explanatory variable is *greenery*, which equals 1 if the project to which the apartment belongs has extra greenery and equals 0 otherwise:

$$\ln price_i = \gamma_0 + \gamma_1 greenery_i + u_i, \quad i = 1, \ldots, n$$

   (a) Suggest an omitted variable that would cause the OLS estimator $\hat{\gamma}_1$ to be inconsistent for β_1, the true effect of *greenery* on $\ln price$.

   (b) Use the omitted variable bias formula to sign the direction of the bias:

$$plim\, \hat{\gamma}_1 = \beta_1 + \beta_2 \frac{\operatorname{cov}(x_1, x_2)}{\operatorname{var}(x_1)}$$

   (c) Suppose that you didn't realize that omitted variable bias was a problem. How would the bias that you find in part (b) affect your conclusion regarding the effect of greenery on apartment prices?

2. Assume that the relationship between unemployment rate u and inflation rate i is determined by the following equation

$$u_t = 0.3i_t + 0.1i_{t-1} - 0.02i_{t-2} \tag{1}$$

   where u_t is the unemployment rate in year t, i_t is the inflation rate in year t, and $i \sim iidN(1, 4)$.

   (a) Compute the mean and variance of u_t.

(b) Compute the first three autocovariances of u_t: $\mathrm{cov}(u_t, u_{t-1}), \mathrm{cov}(u_t, u_{t-2}), \mathrm{cov}(u_t, u_{t-3})$.

(c) Compute the first three autocorrelations of u_t.

(d) Is u_t stationary? Is it weakly dependent?

3. The file `beer` contains monthly beer sales for a manufacturer over a number of years.

   (a) The data includes a year and a month variable. Let's combine this into a single variable that we'll call `period`. The command is `ym`. After creating `period`, format the variable using the command `format period %tm`. Use the `tsset` to tell Stata that the relevant time variable is `period`.

   (b) Create two new variables, log of beer sales and quarter of the year, which we will use throughout the analysis.

   (c) Graph the log of beer sales over time. Comment on the pattern.

   (d) Estimate a regression that accounts for seasonality (at the quarterly level) and statistically test if there is seasonality in the data. Comment.

   (e) Compute the residuals from the regression and plot them against time. Comment on the pattern.

   (f) Estimate an AR(1) of the residuals. Do the results indicate that the residuals are serially correlated?

   (g) Re-estimate the model now adding the lag of log beer sales as an additional right-hand side variable. Comment on the results.

   (h) Re-do parts (e) and (f) using the modified model. What do you conclude? Which model do you prefer?

   (i) Estimate an ARCH(1) model. Comment on your findings.

1.

(a) There is only one dummy variable for "greenary" to explain the lnprice. Also lots of possible variable could be omitted, such as size, location, the distance from city CBD or downtown. May be overall quality with greenary is greater than non-green, it will lead $\hat{\gamma}_1$ drift and inconsistent for β_1. The equation's "u_i" will include something omitted variables, which is correlated to greenary varrble. And this model will exist endogenous issues.

(b) $$\beta_2 \frac{cov(\text{greenary}, \text{greater location})}{Var(\text{greenary})} > 0$$

$\because$ $\beta_2 > 0$ (greater location leads price $\uparrow$) and $cov(X_1 X_2) > 0$.

$\downarrow$

(greater location, the real estate company will pay more attention to greenary)

$\therefore$ The bias is positive

(c) I will see the "greenary" as more powerful and higher $\hat{\gamma}_1$, that is, overestimates greenary effect than it should be.

2.

$\mu \ \sigma^2$

(a) $E(u_t) = E(0.3 i_t + 0.1 i_{t-1} - 0.02 i_{t-2})$ $\because N(1,4)$

$= 0.3 E i_t + 0.1 E i_{t-1} - 0.02 E(i_{t-2}) = 0.4 - 0.02 = 0.38$

$var(u_t) = var(0.3 i_t + 0.1 i_{t-1} - 0.02 i_{t-2})$

$= 0.3^2 var\, i_t + 0.1^2 var\, i_{t-1} + 0.02^2 var\, i_{t-2}$

$= 0.1004 \times 4$

$= 0.4016$

(b) $cov(u_t, u_{t-1}) = cov(0.3 i_t + 0.1 i_{t-1} - 0.02 i_{t-2},\ 0.3 i_{t-1} + 0.1 i_{t-2} - 0.02 i_{t-3})$

$\because$ i_t i_{t-1} i_{t-2} ..... both independent

& $\because$ $cov(A,B) = 0$ when A,B 独立

$\therefore$ 原式 $= cov(0.1 i_{t-1}, 0.3 i_{t-1}) + cov(-0.02 i_{t-2}, 0.1 i_{t-2})$

$= 0.1 \times 0.3 \times var(i_{t-1}) + (-0.02) \times (0.1) \times var(i_{t-2})$

$= 0.028 \times 4 = 0.112$

$cov(u_t, u_{t-2}) = cov(0.3 i_t + 0.1 i_{t-1} - 0.02 i_{t-2},\ 0.3 i_{t-2} + 0.1 i_{t-3} - 0.02 i_{t-4})$

$= cov(-0.02 i_{t-2}, 0.3 i_{t-2})$

$= -0.02 \times 0.3 \times Var(i_{t-2})$

$= -0.006 \times 4 = -0.024$

$cov(u_t, u_{t-3}) = cov(0.3 i_t + 0.1 i_{t-1} - 0.02 i_{t-2},\ 0.3 i_{t-3} + 0.1 i_{t-4} - 0.02 i_{t-5})$

$\because cov(u_t, u_{t-3}) = 0$

$\therefore$ 原式 $= 0$

(c) $\text{Correlation}(u_t, u_{t-1}) = \dfrac{\text{cov}(u_t, u_{t-1})}{\sqrt{\sigma^2 u_t \, \sigma^2 u_{t-1}}} = \dfrac{0.112}{0.4016} = \dfrac{70}{251} = 0.27888$

$\text{Correlation}(u_t, u_{t-2}) = \dfrac{\text{Cov}(u_t, u_{t-2})}{\sigma u_t \, \sigma u_{t-2}} = \dfrac{-0.024}{0.4016} \approx -\dfrac{15}{251} = -0.05976$

$\text{correlation}(u_t, u_{t-3}) = \dfrac{\text{cov}(u_t, u_{t-3})}{\sigma u_t \, \sigma u_{t-3}} = 0$

(d). Yes, u_t is stationary and weakly dependent.

$\because$ $E(u_t)$ $\text{Var}(u_t)$ $\text{cov}(u_t, u_{t-1})$ don't depend on time.

$\therefore$ stationary

& $\because$ $\text{Correlation}(u_t, u_{t-3}) = 0$

so with any lagging time larger than 2 times equals to 0.

$\therefore$ weakly dependent.

Appendix

3.

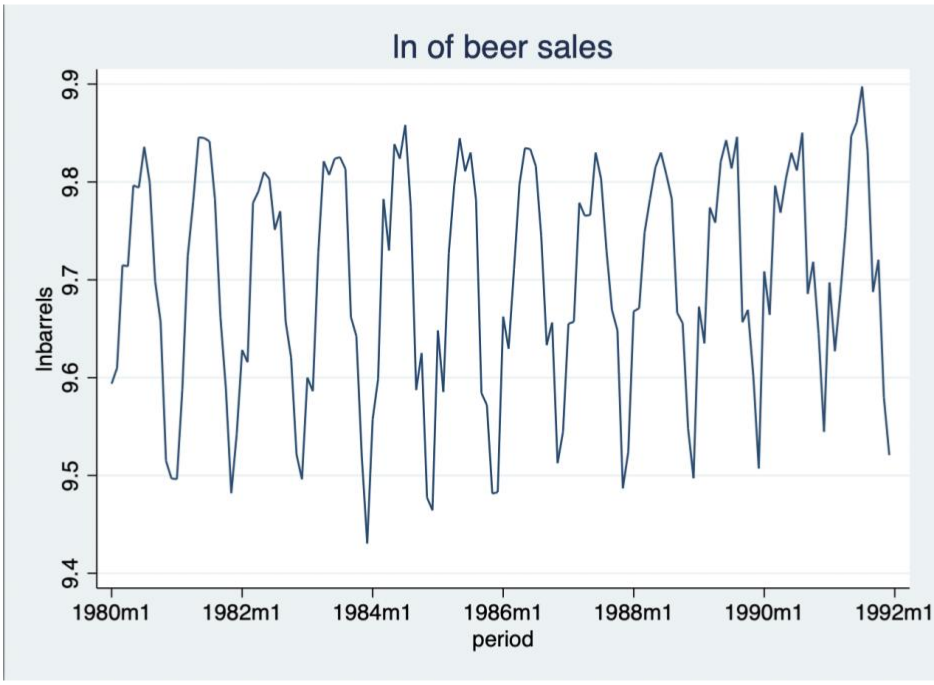
(a)

```
use beer.dta, clear //导入文件,默认调用CD路径

gen period = ym(year,month) //生成年月时期 ym表明数据内含年月

format period %tm

tsset period //声明period是时间变量
       time variable:  period, 1980m1 to 1991m12
               delta:  1 month
```

(b)

```
. //(b)
. gen lnbarrels=log(barrels) //生成对数形式

. gen qoy=1 if month<=3
(108 missing values generated)

. replace qoy=2 if month<=6 & month>3
(36 real changes made)

. replace qoy=3 if month<=9 & month>6
(36 real changes made)

. replace qoy=4 if month<=12 & month>9 //生成季度, 备用3虚拟变量
(36 real changes made)
```

(c)



This time series fluctuate with potential seasonal trend. But average mean seems like constant.

(d)

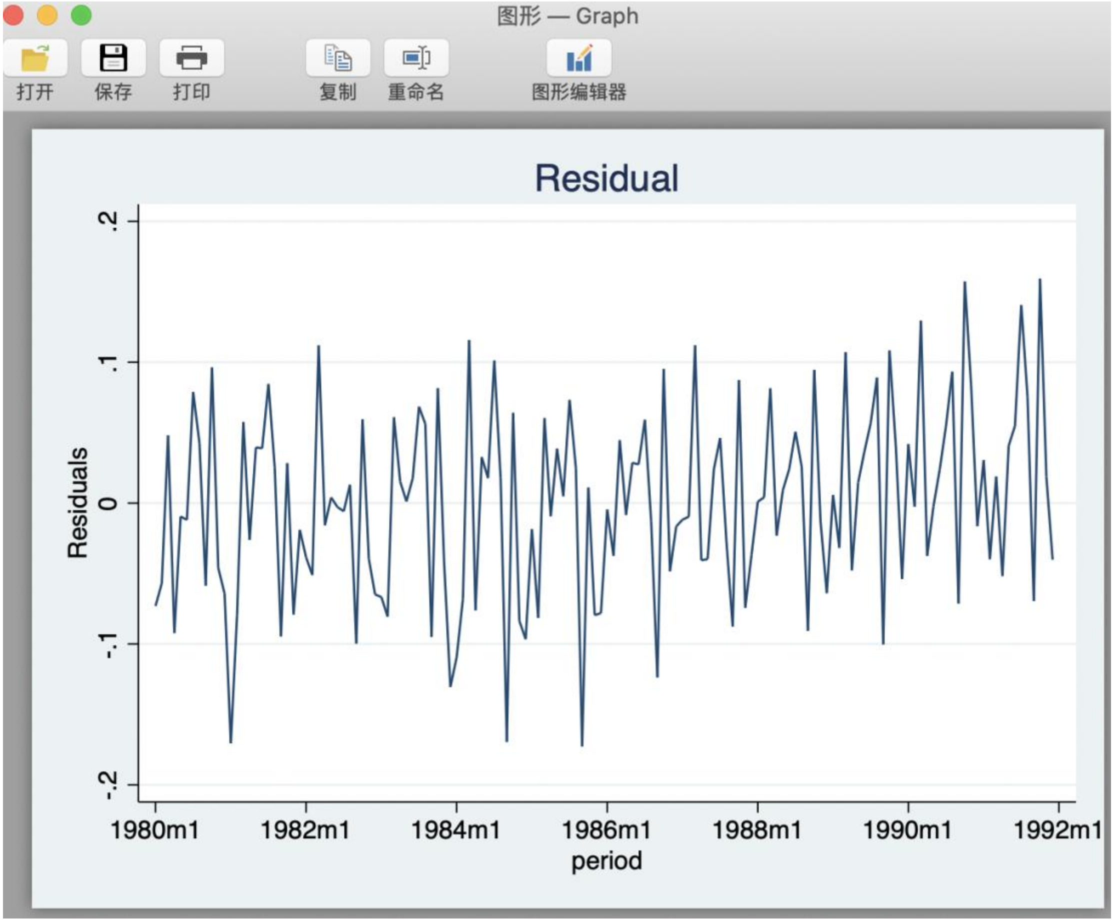
VARIABLES	(1) Seasonaltest lnbarrels
q_dummy2	0.139***
	(0.0163)
q_dummy3	0.0901***
	(0.0163)
q_dummy4	-0.106***
	(0.0163)
Constant	9.667***
	(0.0115)
Observations	144
R-squared	0.654

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

```
. reg lnbarrels q_dummy2-q_dummy4 //回归季度性 变量若有序号规律可缩写联结

      Source |       SS           df       MS      Number of obs   =       144
-------------+----------------------------------   F(3, 140)       =     88.04
       Model |  1.25557373         3  .418524578   Prob > F        =    0.0000
    Residual |  .665535587       140  .004753826   R-squared       =    0.6536
-------------+----------------------------------   Adj R-squared   =    0.6461
       Total |  1.92110932       143  .013434331   Root MSE        =    .06895

------------------------------------------------------------------------------
   lnbarrels |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    q_dummy2 |   .1392218   .0162512     8.57   0.000     .1070923    .1713513
    q_dummy3 |   .0901391   .0162512     5.55   0.000     .0580096    .1222686
    q_dummy4 |  -.1057724   .0162512    -6.51   0.000    -.1379019   -.0736429
       _cons |   9.666808   .0114913   841.23   0.000     9.644089    9.689527
------------------------------------------------------------------------------

. est store Seasonaltest //结果存于dta中，方便调用
```

The quarter seasonal dummy variables is significant. And P=0.0000 that shows the seasonality in the beer data.

(e)



Residual showing in gragh has a slightly up trend with times especially after 1988m1. But is not apparently. Series correlation seems doesn't exist but need to further discussions.

(f)

```
. reg res L.res //AR(1) 一阶自回归

      Source |       SS           df       MS      Number of obs   =       143
-------------+----------------------------------   F(1, 141)       =      1.72
       Model |  .007936528         1  .007936528   Prob > F        =    0.1924
    Residual |  .652226345       141  .004625719   R-squared       =    0.0120
-------------+----------------------------------   Adj R-squared   =    0.0050
       Total |  .660162873       142  .004649034   Root MSE        =    .06801

------------------------------------------------------------------------------
         res |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         res |
         L1. |  -.1093355    .083471    -1.31   0.192     -.274352    .0556809
             |
       _cons |   .0005415   .0056876     0.10   0.924    -.0107024    .0117854
------------------------------------------------------------------------------
```

(1) (2)

VARIABLES	AR1_1 res	AR1_2 res_modified
L.res	-0.109	
	(0.0835)	
L.res_modified		-0.437***
		(0.0757)
Constant	0.000542	0.000506
	(0.00569)	(0.00451)
Observations	143	142
R-squared	0.012	0.192

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

AR1_1 not shows serial correlations. And P=0.1924>0.05 is not significant

(g)

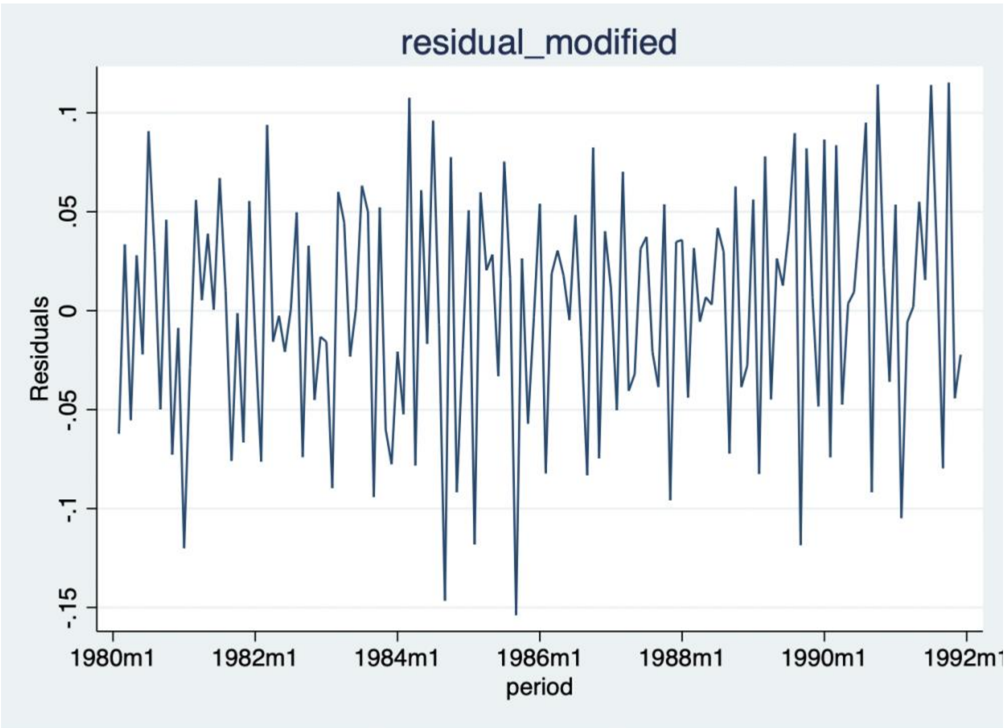
VARIABLES	(1) Regnew lnbarrels
L.lnbarrels	0.577***
	(0.0882)
q_dummy2	0.0272
	(0.0221)
q_dummy3	-0.0425*
	(0.0246)
q_dummy4	-0.121***
	(0.0145)
Constant	4.141***
	(0.845)
Observations	143
R-squared	0.736

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

```
. //(g)
. reg lnbarrels L.lnbarrels q_dummy2-q_dummy4

      Source |       SS           df       MS      Number of obs   =       143
-------------+----------------------------------   F(4, 138)       =     96.28
       Model |  1.40631583         4  .351578958   Prob > F        =    0.0000
    Residual |  .503914217       138  .003651552   R-squared       =    0.7362
-------------+----------------------------------   Adj R-squared   =    0.7286
       Total |  1.91023005       142  .013452324   Root MSE        =    .06043

------------------------------------------------------------------------------
   lnbarrels |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lnbarrels |
         L1. |   .5765389   .0881697     6.54   0.000     .4022006    .7508771
             |
    q_dummy2 |   .0271872   .0221016     1.23   0.221    -.0165143    .0708888
    q_dummy3 |  -.0425256   .0245872    -1.73   0.086    -.0911419    .0060907
    q_dummy4 |  -.1210239    .014485    -8.36   0.000    -.1496651   -.0923827
       _cons |   4.140944   .8454473     4.90   0.000     2.469238     5.81265
------------------------------------------------------------------------------
```

The second quarter seasonal effect is not significant. Added one time lag shows it has own up trend on selling beers with times. Whole R-squared increases, which may represent this model is better.

(h)



It fluctuates around 0 fiercely than old model.

```
. reg res_modified L.res_modified //继续重复一阶自回归

      Source |       SS           df       MS      Number of obs   =       142
-------------+----------------------------------   F(1, 140)       =     33.24
       Model |  .095945009         1  .095945009   Prob > F        =    0.0000
    Residual |  .404073062       140  .002886236   R-squared       =    0.1919
-------------+----------------------------------   Adj R-squared   =    0.1861
       Total |  .500018071       141  .003546227   Root MSE        =    .05372

------------------------------------------------------------------------------
res_modified |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
res_modified |
         L1. |  -.4365635   .0757185    -5.77   0.000    -.5862632   -.2868639
             |
       _cons |   .0005064   .0045084     0.11   0.911     -.008407    .0094197
------------------------------------------------------------------------------
```

VARIABLES	(1) AR1_1 res	(2) AR1_2 res_modified
L.res	-0.109	
	(0.0835)	
L.res_modified		-0.437***
		(0.0757)
Constant	0.000542	0.000506
	(0.00569)	(0.00451)
Observations	143	142
R-squared	0.012	0.192

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

When lag of log beers added in the model **(AR1_2)**, it has shown the self series correlation and P=0.0000<0.05 the lag variable is significant. As this new model R-sqaured is higher in both AR and regression, so it is more explainable and robust. Series correlation issues indeed exist. So, I prefer the latter one.

```
reg res_modified L.barrels L.res_modified //初步观察新模型序列相关是否存在

      Source |       SS           df       MS      Number of obs   =       142
-------------+----------------------------------   F(2, 139)       =     24.47
       Model |  .130207858         2  .065103929   Prob > F        =    0.0000
    Residual |  .369810213       139  .002660505   R-squared       =    0.2604
-------------+----------------------------------   Adj R-squared   =    0.2498
       Total |  .500018071       141  .003546227   Root MSE        =    .05158
```

(i)

VARIABLES	(1) Arch1 res_modifiedsqr
L.res_modifiedsqr	-0.0160
	(0.0847)
Constant	0.00358***
	(0.000465)
Observations	142
R-squared	0.000

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

```
. reg res_modifiedsqr L.res_modifiedsqr

      Source |       SS           df       MS      Number of obs   =       142
-------------+----------------------------------   F(1, 140)       =      0.04
       Model |  6.4299e-07         1  6.4299e-07   Prob > F        =    0.8501
    Residual |  .002510002       140  .000017929   R-squared       =    0.0003
-------------+----------------------------------   Adj R-squared   =   -0.0069
       Total |  .002510645       141  .000017806   Root MSE        =    .00423

------------------------------------------------------------------------------
es_modifie~r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
es_modifie~r |
         L1. |  -.0160324   .0846586    -0.19   0.850    -.1834069    .1513421
             |
       _cons |   .0035783   .0004651     7.69   0.000     .0026587    .0044979
------------------------------------------------------------------------------
```

It shows that in Arch1 model, it seems like no Heteroscedasticity. And P=0.04<0.05 on the edge of the boundary and the lag variable p>0.05 is not significant, which indicates there is no relationship between res^2 and lag of res^2. So we could not reject the H0, thus there is no heteroscedasticity.

VARIABLES	(1) heteroscedasticitytest res_modifiedsqr
L.lnbarrels	0.000262
	(0.00307)
Constant	0.000985
	(0.0298)
Observations	143
R-squared	0.000

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The heteroscedasticity test also prove that there is no significant and no heteroscedasticity.

	(1)
	Serial_Corr_Test
VARIABLES	res_modified
L.lnbarrels	0.153***
	(0.0440)
L.res_modified	-0.589***
	(0.0850)
Constant	-1.485***
	(0.427)
Observations	142
R-squared	0.257

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

But, in series correlation test, it is showing this model have a series correlation, so it has some issue to be modified.