Forecasting project for EC3304 Sem 1, 2016/2017

By A0102680R

1. Model can be delta vacancy rate with no lags
   Alternative models can be: vacancy with lag(1) OR current vacancy rate
2. Sample period is 1988 Q1 to 2016 Q2
3. Additional data: Refer to appendix
4. Abstract:
   Using STATA: After realising time series may be a unit root, reduced the time period to year >=2000; however did not obvious much difference from plot. So using full sample period, did a AIC&BIC check. Result suggests to use lag(1). However, later noticed that lag(1) and lag(2) may be highly correlated and jointly significant. Hence, using AR(2) model may be probable.
   But since Dickey Fuller test for both AR(1) and AR(2) is unit root, decided to look into delta vacancy rate. Result from AIC and BIC for delta vacancy rate is no lag. Dickey Fuller test suggests a stationary time series for zero lag for delta vacancy which is great. Performing regression on the model,forecast result obtained is 8.84. Using AR(1) would give very similar results but since DF test suggests stationary time series for delta vacancy with zero lag, we would use that instead.

   Using R, did a decomposition to observe for trend/seasonality. Using the forecast package on the decomposed, forecast result obtained is 8.92.

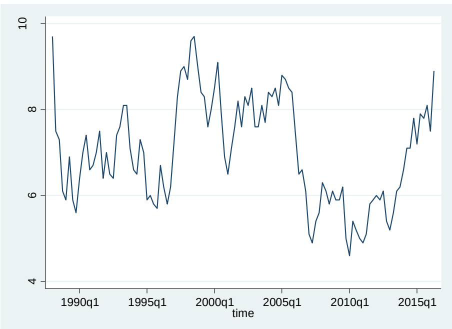
   Both results are similar, and in fact, close to 2016 Q2 vacancy rate.
   Best forecast for 2016 Q3 can be current period 2016 Q2.

5. 2016 Q3 forecast: 8.84
   95% confidence interval: (7.62,10.1)

In the following my analysis will be done in STATA unless specified otherwise.
Cleaning of data and generating time series

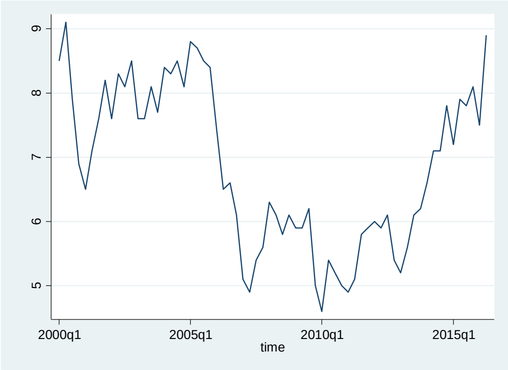
```
gen year = substr(time,1,4)
gen quarter = substr(time,6,1)
drop in 115/120
drop v3
destring year, replace
destring quarter, replace
drop time
gen time = yq(year,quarter)
format time %tq
tsset time
gen vacancy=vacancyrateofprivateresidentialu
destring vacancy, replace
tsline vacancy
```

Started off plotting the time series for vacancy rate for private residential Units(Excluding EC). I will be using the term “vacancy” to replace the variable in my following explanations.
Time period of data used is 1988 Q1 to 2015



From my preliminary graph, it shows evidence of stochastic trend as the time series changes direction abruptly and unexpectedly. I don’t suspect a drift since I don’t see any obvious upward or downwards trend.
Similarly for just year >=2000,

tsline vacancy if year >=2000



So returning to full sample period, 1988 Q1 to 2016 Q2,

varsoc vacancy

```
. varsoc vacancy

   Selection-order criteria
   Sample:  1989q1 - 2016q2                      Number of obs      =       110
```

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-176.184				1.46763	3.22153	3.23148	3.24608
1	-93.5246	165.32*	1	0.000	.332518*	1.73681*	1.75673*	1.78591*
2	-93.3984	.25254	1	0.615	.337846	1.7527	1.78257	1.82635
3	-93.2764	.24384	1	0.621	.343289	1.76866	1.80849	1.86686
4	-93.265	.02292	1	0.880	.349526	1.78664	1.83642	1.90939

```
   Endogenous:  vacancy
    Exogenous:  _cons
```

From AIC and BIC, both suggests that lag(1) should be used

reg vacancy L1.vacancy,r

```
Linear regression                                   Number of obs =      113
                                                    F(  1,    111) =  308.56
                                                    Prob > F      =  0.0000
                                                    R-squared     =  0.7487
                                                    Root MSE      =   .6012
```

vacancy	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
vacancy L1.	.855442	.0486988	17.57	0.000	.758942	.9519419
_cons	1.003419	.3289862	3.05	0.003	.3515112	1.655327

AR(1) with coefficient 0.855 for lag(1) suggests lag(1) is seemingly close to current time. Hence it may behave a little like unit root. But overall it is good. R-sqaure is high 0.7487 and lag(1) is significant.

Checking for drift: if it has a drift: for delta Y(t)=drift+u(t), the intercept drift will not be close to zero.

Let's just try with AR(2)

reg vacancy L1.vacancy L2.vacancy,r
corr l1.vacancy l2.vacancy

```
Linear regression                                      Number of obs =     112
                                                       F(  2,   109) =  246.59
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.7707
                                                       Root MSE      =  .57913

------------------------------------------------------------------------------
             |               Robust
     vacancy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     vacancy |
         L1. |   .9386458    .084445    11.12   0.000     .7712784    1.106013
         L2. |   -.058029   .0869083    -0.67   0.506    -.2302784    .1142204
             |
       _cons |   .8452719   .2889446     2.93   0.004     .2725931    1.417951
------------------------------------------------------------------------------
```

```
             |       L.       L2.
             |  vacancy  vacancy
-------------+------------------
     vacancy |
         L1. |   1.0000
         L2. |   0.8700   1.0000
```

This suggests that lag(1) was underestimated in the AR(1) model since coefficient of lag(1) is larger in AR(2) model: lag(1) and lag(2) have large correlation
The regression further reaffirms the suspicion of stochastic trend. Amount of underestimation increases when data is more strongly correlated
Also lag(2) is not significant in AR(2) (p-value=0.506) even though the overall F test is significant (p value=0.000) suggesting they are jointly significant
To investigate,

test L1.vacancy L2.vacancy

```
 ( 1)  L.vacancy = 0
 ( 2)  L2.vacancy = 0

       F(  2,   109) =  183.14
            Prob > F =    0.0000
```

This confirms lag(1) and lag(2) are jointly significant suggesting that correlation between the lags are slightly large and in presence of collinearity.
Hence using AR(2) may be a better model than AR(1). An interesting note to keep in mind is that R-square for AR(2) is slightly higher and Root MSE for AR(2) is lower slightly.

Proceeding to check via Dickey-Fuller Test

dfuller vacancy,lag(1)

```
Augmented Dickey-Fuller test for unit root         Number of obs   =       112

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.515            -3.506            -2.889            -2.579
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.1119
```

dfuller vacancy,lag(2)

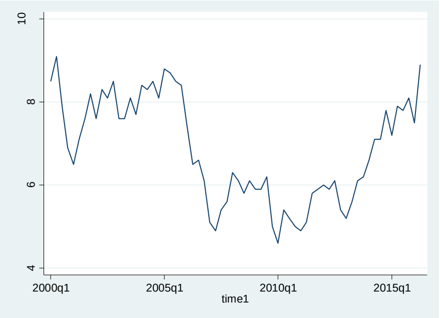
```
Augmented Dickey-Fuller test for unit root         Number of obs   =       111

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.464            -3.506            -2.889            -2.579
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.1245
```

Both Dickey Fuller test for lag 1 and 2 suggests vacancy rate demonstrates stochastic trend for its time series.

Hence, I think the best forecast for 2016 q3 will be 2016 q2 even though R-square for the above two model are pretty high (>0.7)

However, I will like to look into it further.
Considering only period greater than or equal to 2000 as maybe considering a nearer past will act as better predictors for current vacancy rate.
Period: 2000Q1 to 2016Q2

gen time1=time if year >=2000
format time1 %tq
tsset time1
tsline vacancy



Still looks like a random walk

varsoc vacancy if year>=2000

Selection-order criteria
Sample: 2001q1 - 2016q2 Number of obs = 109

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-175.24				1.48571	3.23377	3.24378	3.25846
1	-151.078	48.325	1	0.000	.971313	2.80877	2.82879	2.85815
2	-144.93	12.296*	1	0.000	.883776*	2.71431*	2.74435*	2.78839*
3	-144.894	.07148	1	0.789	.899569	2.732	2.77206	2.83077
4	-144.308	1.1721	1	0.279	.906457	2.7396	2.78967	2.86306

Endogenous: vacancy
Exogenous: _cons

Using AIC and BIC to check, interestingly it suggests 2 lags for the AR model

Proceeding to check via Dickey-Fuller Test

dfuller vacancy,lag(2)

Augmented Dickey-Fuller test for unit root Number of obs = 63

	Test Statistic	Interpolated Dickey-Fuller 1% Critical Value	5% Critical Value	10% Critical Value
Z(t)	-1.501	-3.562	-2.920	-2.595

MacKinnon approximate p-value for Z(t) = 0.5329

Sadly, the test statistics on the lag of Y (-1.501) is greater than DF 10% critical value. Hence, we cannot reject null hypothesis: the time series has unit root.

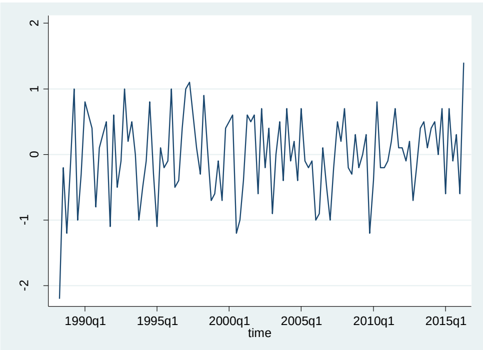
Personally, I suspect there is a structural break observing from the change in AC sign midway.

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial Autocor]
1	0.8372	0.8554	82.023	0.0000		
2	0.7238	-0.0580	143.87	0.0000		
3	0.6138	-0.0257	188.76	0.0000		
4	0.5420	0.0137	224.08	0.0000		
5	0.4762	-0.0251	251.59	0.0000		
6	0.3859	-0.1063	269.82	0.0000		
7	0.2868	-0.1267	279.98	0.0000		
8	0.2387	0.1141	287.09	0.0000		
9	0.2132	0.0830	292.82	0.0000		
10	0.1919	0.0045	297.5	0.0000		
11	0.1665	-0.0083	301.06	0.0000		
12	0.1385	-0.0733	303.55	0.0000		
13	0.0937	-0.1166	304.7	0.0000		
14	0.0696	0.0171	305.34	0.0000		
15	0.0675	0.0962	305.95	0.0000		
16	0.0844	0.0371	306.91	0.0000		
17	0.0559	-0.1518	307.33	0.0000		
18	0.0129	-0.0911	307.36	0.0000		
19	0.0002	0.1200	307.36	0.0000		
20	-0.0002	0.0195	307.36	0.0000		
21	0.0198	-0.0012	307.41	0.0000		
22	0.0126	-0.0563	307.43	0.0000		
23	-0.0223	-0.0545	307.51	0.0000		
24	-0.0541	-0.0945	307.94	0.0000		
25	-0.0770	-0.1123	308.82	0.0000		
26	-0.0770	-0.0248	309.71	0.0000		
27	-0.0827	0.0359	310.75	0.0000		
28	-0.1615	-0.3439	314.76	0.0000		
29	-0.2301	-0.0863	322.99	0.0000		
30	-0.2945	-0.0579	336.65	0.0000		
31	-0.3475	-0.2324	355.9	0.0000		
32	-0.3685	-0.0659	377.79	0.0000		
33	-0.4175	-0.1367	406.25	0.0000		
34	-0.4705	-0.1531	442.83	0.0000		
35	-0.4593	0.0884	478.15	0.0000		
36	-0.4144	-0.0813	507.26	0.0000		
37	-0.3711	-0.1972	530.91	0.0000		
38	-0.3317	-0.0929	550.06	0.0000		
39	-0.2812	0.1650	563.99	0.0000		
40	-0.2503	-0.0630	575.19	0.0000		

As we are not be taught about running Chow test. I shall not consider that but I think considering Chow test may give a better forecast.

Dropping the idea of looking into near future (year>=2000), now I want to work on the previously deduced conclusion that data is a unit root. I take the difference for vacancy rate to forecast instead so it would hopefully be stationary.

tsline d.vacancy



Looks stationary.

varsoc d.vacancy

```
Selection-order criteria
Sample:  1989q2 - 2016q2                          Number of obs      =       109
```

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-95.9242				.346646*	1.77843*	1.78844*	1.80312*
1	-95.9077	.03294	1	0.856	.35296	1.79647	1.8165	1.84585
2	-95.2974	1.2207	1	0.269	.355496	1.80362	1.83366	1.8777
3	-94.9242	.74626	1	0.388	.359615	1.81512	1.85518	1.91389
4	-94.9052	.03804	1	0.845	.366158	1.83312	1.88319	1.95658

```
Endogenous:  D.vacancy
 Exogenous:  _cons
```

AIC & BIC suggest no lag.

dfuller d.vacancy,lag(0)

```
Dickey-Fuller test for unit root                   Number of obs   =       112

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -10.775            -3.506            -2.889            -2.579
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

Using delta Y ensures a stationary time series since dickey fuller test for delta Y is rejected.
Using model delta Y with no lag gives better forecast

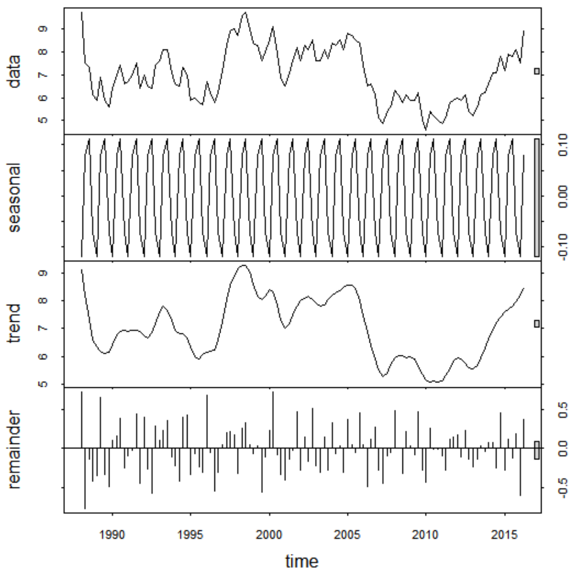
reg d.vacancy,r

```
Linear regression                               Number of obs =        113
                                                F(  0,    112) =      0.00
                                                Prob > F      =          .
                                                R-squared     =     0.0000
                                                Root MSE      =     .62346

------------------------------------------------------------------------------
             |               Robust
   D.vacancy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -.0070796     .05865    -0.12   0.904     -.123287    .1091277
```

So 2016 Q3 =(-0.0070796*2016Q2)+2016Q2
=(-0.0070796*8.9)+8.9
=8.83699156
~=8.84

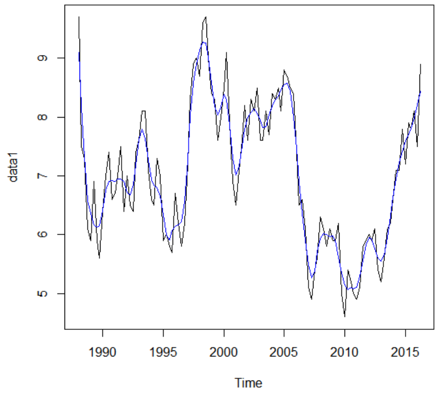
95% confidence interval= 8.83699156 +- 1.96* Root MSE
=8.83699156 +- 1.96 * (0.62346)
=(7.61500,10.05897)
~= (7.62,10.1)

Jumping to R for a while for a cross check,
by decomposing the data to observe perhaps for trend/seasonality
data=read.csv("C:/Users/woonling/Downloads/vacancy.csv",header=TRUE)
data1=ts(data[,1],start=c(1988,1),frequency=4)
data2=stl(data1,s.window="periodic")
plot(data2)

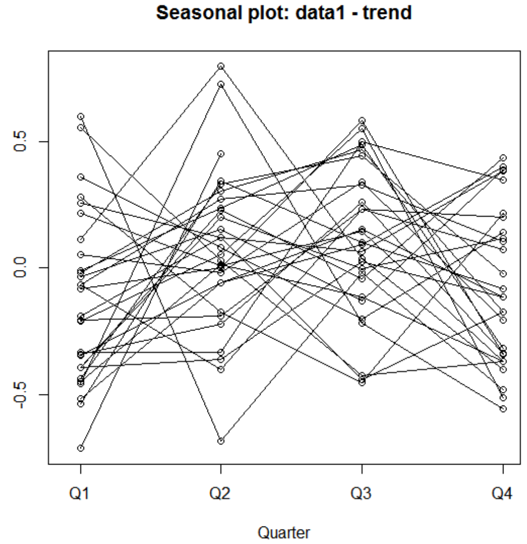


Interestingly, there is a clear seasonal component.
trend = data2$time.series[,"trend"]

seasonal = data2$time.series[,"seasonal"]
plot(data1)
lines(trend,col="blue")



The trend may closely mimic the time series, however, it doesn't mean forecast will be good.
Taking a look briefly at how the seasonal component looks
library(fpp)
seasonplot(data1-trend)



fcast=forecast(data2,method="naive",h=1)
fcast

```
        Point Forecast   Lo 80   Hi 80     Lo 95    Hi 95
2016 Q3       8.931915 7.97576 9.88807 7.469602 10.39423
```

Both methods generates similar results. Forecast for 2016 Q3 is almost same as result for 2016 Q2.

<u>Appendix</u>

I added in number of available private residential units excluding EC hoping it may be a possible predictor

gen year = substr(time,1,4)
gen quarter = substr(time,6,1)
drop in 115/120
drop v3
destring year, replace
destring quarter, replace
drop time
gen time= yq(year,quarter)
format time %tq

```
tsset time
merge time using "D:\AA- NUS\ec3304\project.dta"
drop _merge
gen available= availableprivateresidentialunits
destring available, replace
sort time
reg vacancy L1.available L1.vacancy,r
```

```
Linear regression                                  Number of obs =     113
                                                   F(  2,   110) =  158.35
                                                   Prob > F      =  0.0000
                                                   R-squared     =  0.7503
                                                   Root MSE      =  .60207
```

vacancy	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
available						
L1.	5.67e-07	7.41e-07	0.77	0.446	-9.01e-07	2.04e-06
vacancy						
L1.	.8606618	.0483704	17.79	0.000	.7648031	.9565205
_cons	.862989	.3635519	2.37	0.019	.1425144	1.583464

Did a few combinations and checking AIC etc. Realise that “number of available private residential excluding EC” doesn’t really help forecast vacancy rate.

```
newey d.vacancy L(1/4).available,lag(4)
```

```
Regression with Newey-West standard errors          Number of obs  =       110
maximum lag: 4                                      F(  4,   105)  =      0.42
                                                    Prob > F       =    0.7927
```

D.vacancy	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
available						
L1.	-4.63e-06	7.24e-06	-0.64	0.523	-.000019	9.72e-06
L2.	3.22e-06	.0000102	0.32	0.752	-.000017	.0000234
L3.	2.86e-06	6.07e-06	0.47	0.638	-9.17e-06	.0000149
L4.	-1.20e-06	6.80e-06	-0.18	0.860	-.0000147	.0000123
_cons	-.0061774	.1085329	-0.06	0.955	-.2213782	.2090234

So shall drop this analysis