農產價格分析

黃肉西瓜時間序列分析與價格預測

農經碩一 R09627023 吳柏叡

ARIMA vs ARIMA_with_volumn

ARIMA(0,0,2)

```
ARIMA regression

Sample:  1999m2 - 2016m12                       Number of obs     =        215
                                                Wald chi2(2)      =       0.79
Log likelihood =  -68.4266                      Prob > chi2       =     0.6751

------------------------------------------------------------------------------
             |                 OPG
  train_true |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
train_true   |
       _cons |   -.001333   .0007524    -1.77   0.076    -.0028076    .0001416
-------------+----------------------------------------------------------------
ARMA         |
          ma |
         L1. |  -.7383323   827.4268    -0.00   0.999    -1622.465    1620.988
         L2. |  -.2616683   216.4773    -0.00   0.999    -424.5493    424.026
-------------+----------------------------------------------------------------
       /sigma |   .3288135   136.033     0.00   0.499           0    266.9485
------------------------------------------------------------------------------
Note: The test of the variance against zero is one sided, and the two-sided
      confidence interval is truncated at zero.
```

```
test MSE: 0.0942788
test RMSE: 0.3070485
train MSE: 0.1108279
train RMSE: 0.3329083
```

除此頁外報告後也新增:
1.將差分還原為ln(price)並進行預測(p.15)
2.Whitw noise test (p.13)

註:加入量後MA(2)亦為最佳model

ARIMA(0,0,2) with volumn

```
ARIMA regression

Sample:  1999m2 - 2016m12                       Number of obs     =
>    215
                                                Wald chi2(3)      =    1
> 048.32
Log likelihood = -66.53629                      Prob > chi2       =
> 0.0000

------------------------------------------------------------------------------
> ------
             |                 OPG
v_train_true |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Int
> erval]
-------------+----------------------------------------------------------------
> ------
v_train_true |
      volumn |   1.75e-06   8.55e-07     2.05   0.041     7.32e-08    3.
> 42e-06
       _cons |  -.0160744   .0072731    -2.21   0.027    -.0303294   -.0
> 018193
-------------+----------------------------------------------------------------
> ------
ARMA         |
          ma |
         L1. |  -.7276717   .0616395   -11.81   0.000    -.8484829   -.6
> 068606
         L2. |  -.2360928   .0689736    -3.42   0.001    -.3712786   -.
> 100907
-------------+----------------------------------------------------------------
> ------
       /sigma |    .327832   .0141306    23.20   0.000     .3001365    .3
> 555274
------------------------------------------------------------------------------
```

亦存在白噪音

```
v_test MSE: 0.0917578
v_test RMSE: 0.3029155
v_train MSE: 0.1087181
v_train RMSE: 0.3297243
```

```
. //對殘差檢定是否有白噪音--WN檢定
. wntestq v_e

Portmanteau test for white noise
---------------------------------------
Portmanteau (Q) statistic =    45.3272
Prob > chi2(40)           =     0.2595
```

outline

- 動機
- 資料處理
- 敘述統計
- 單根檢定
- 模型選擇
- 預測結果
- 深度學習模型比較(LSTM)
- 結論

假議題?真問題?

小玉西瓜盛產價慘跌 1台斤5元

讚 0 分享 Tweet 分享

2020/04/28 12:44

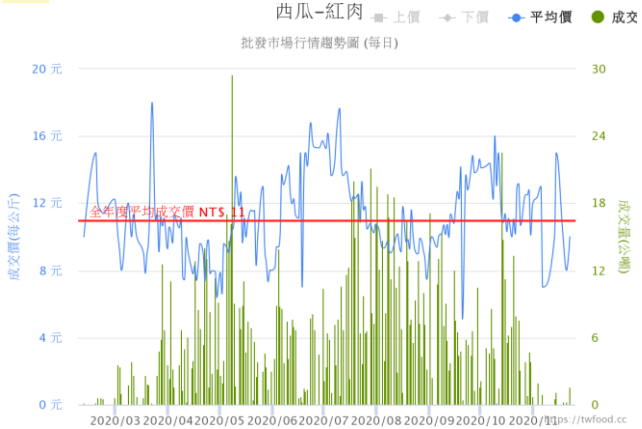
50年來最便宜 西瓜1公斤跌到剩5元

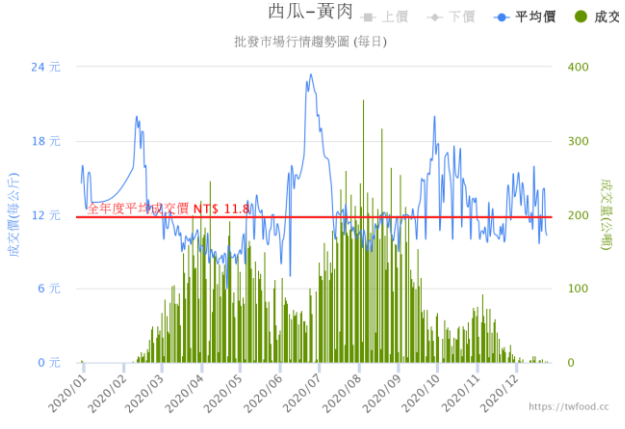
西瓜跌破50年來最低價　陳吉仲：考慮全國茶飲協助收購榨成汁

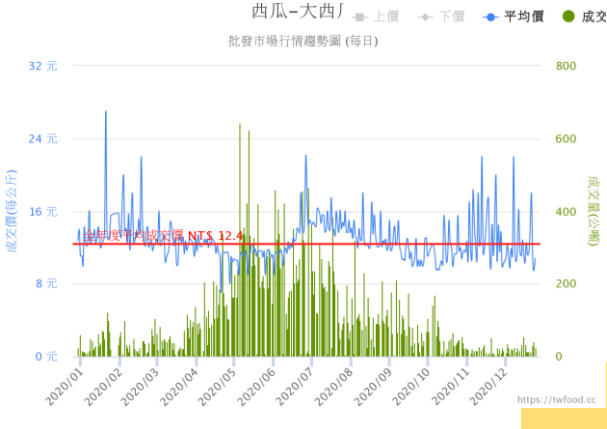
西瓜價跌至50年來最低 農民嘆血本無歸

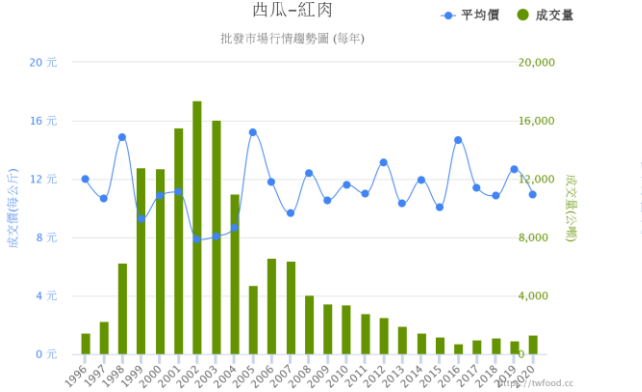
綜合報導　2020-04-27 19:53　最後更新：2020-04-27 21:57

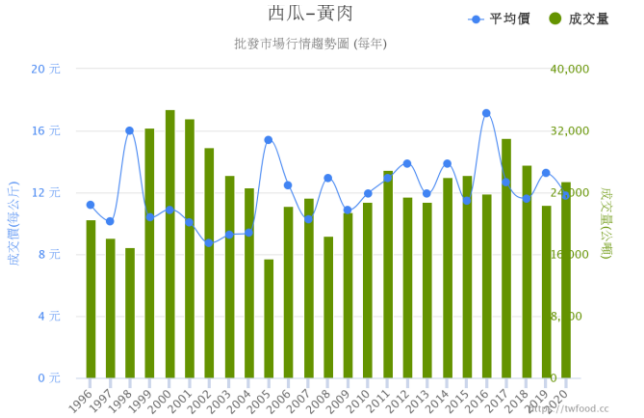
假議題?真問題?

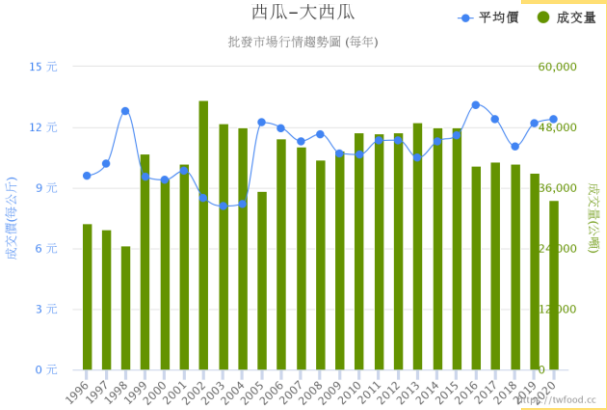

西瓜-紅肉
上價
下價
平均價
成交
批發市場行情趨勢圖 (每日)
全年度平均成交價 NT$ 11.
成交價(每公斤)
成交量(公噸)
https://twfood.cc


西瓜-黃肉
上價
下價
平均價
成交
批發市場行情趨勢圖 (每日)
全年度平均成交價 NT$ 11.8
成交價(每公斤)
成交量(公噸)
https://twfood.cc


西瓜-大西瓜
上價
下價
平均價
成交
批發市場行情趨勢圖 (每日)
全年度平均成交價 NT$ 12.4
成交價(每公斤)
成交量(公噸)
https://twfood.cc


西瓜-紅肉
平均價
成交量
批發市場行情趨勢圖 (每年)
成交價(每公斤)
成交量(公噸)
https://twfood.cc


西瓜-黃肉
平均價
成交量
批發市場行情趨勢圖 (每年)
成交價(每公斤)
成交量(公噸)
https://twfood.cc


西瓜-大西瓜
平均價
成交量
批發市場行情趨勢圖 (每年)
成交價(每公斤)
成交量(公噸)
https://twfood.cc

資料處理

缺失值處理

1.以月平均數取代缺失值
2.若當月無資料則以台北一之同月平均取代

日期處理

1.將年/月/日格式改為年/月
2.將民國年改為西元年

平均價格

以日平均價格計算月平均

```
new_df = []
avg_price = []
avg_vol = []
for year in year_list:
    for month in month_list:
        price = []
        vol=[]
        for i in df_list:
            date = i[0].split('/')
            if date[1] == month and date[0] ==year:
                price.append(i[3])
                vol.append(i[4])
        if len(price)== 0:
            price.append(avg_price[-1])
        if len(vol) == 0:
            vol.append(avg_vol[-1])
        avg_price.append(avg(price))
        avg_vol.append(avg(vol))
        #ew_df.append([year+'/'+month,year,month,round(avg(price),1),round(avg(vol),1)])
        new_df.append([ch_to_west(year)+'-'+month,ch_to_west(year),month,round(avg(price),1),round(avg(vol),1)])
new_df
```

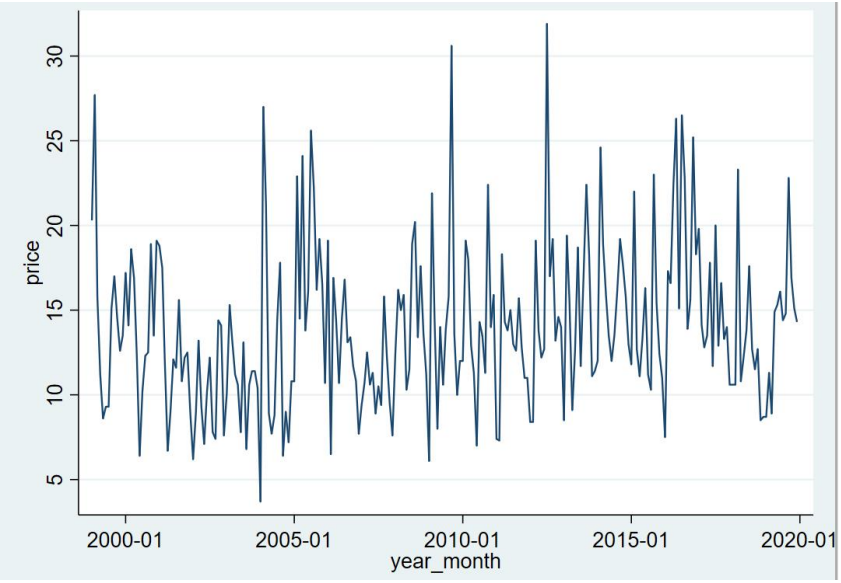
```
['2000-09', '2000', '09', 12.5, 9826.8],
['2000-10', '2000', '10', 18.9, 2823.4],
['2000-11', '2000', '11', 13.5, 2265.5],
['2000-12', '2000', '12', 19.1, 264.7],
['2001-01', '2001', '01', 18.8, 607.4],
['2001-02', '2001', '02', 17.5, 4559.5],
['2001-03', '2001', '03', 11.9, 11706.7],
['2001-04', '2001', '04', 6.7, 32992.0],
```

```
gen y_m = date(date,"YM")
format y_m %tdCCYY-NN
```

date	year	month	price	volumn
Jan-99	1999	1	20.3	230.3
Feb-99	1999	2	27.7	2248.7
Mar-99	1999	3	15.8	18235.1
Apr-99	1999	4	11.2	26201.8
May-99	1999	5	8.6	27782.8
Jun-99	1999	6	9.3	31282.5
Jul-99	1999	7	9.3	21000.6
Aug-99	1999	8	15.1	10110.7
Sep-99	1999	9	17	10218.2
Oct-99	1999	10	14.6	3904.2
Nov-99	1999	11	12.6	3044

Price

看起來較接近定態



敘述統計

```
. sum price ,detail

                            price
-------------------------------------------------------------
      Percentiles      Smallest
 1%           6.2           3.7
 5%           7.4           6.1
10%           8.5           6.2       Obs                 252
25%          10.8           6.4       Sum of Wgt.         252

50%         13.25                     Mean           13.92817
                        Largest       Std. Dev.       4.72832
75%         16.25            27
90%          19.8          27.7       Variance       22.35701
95%          22.9          30.6       Skewness       .9019533
99%          27.7          31.9       Kurtosis       4.102355
```

```
price
-------------------------------------------------------------

        type:  numeric (float)

       range:  [3.7,31.9]                   units:  .1
unique values:  129                  missing .:  0/252

        mean:   13.9282
     std. dev:  4.72832

  percentiles:        10%       25%       50%       75%       90%
                      8.5      10.8     13.25     16.25      19.8
```

Dickey–Fuller

```
. dfuller lnp if train ==0,trend reg

Dickey-Fuller test for unit root                    Number of obs   =       215

                                 ---------- Interpolated Dickey-Fuller ----------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic         Value             Value            Value
------------------------------------------------------------------------------
 Z(t)           -11.538          -4.002            -3.435           -3.135
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000

------------------------------------------------------------------------------
       D.lnp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnp |
         L1. |   -.765185   .066318   -11.54   0.000    -.8959171   -.6344579
       _trend |   .0011096   .0003727     2.98   0.003     .0003749    .0018443
       _cons |   1.847642   .1673168    11.04   0.000     1.517824    2.17746
------------------------------------------------------------------------------
```

無trend

```
. dfuller lnp if train ==0 ,reg

Dickey-Fuller test for unit root                    Number of obs   =       215

                                 ---------- Interpolated Dickey-Fuller ----------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic         Value             Value            Value
------------------------------------------------------------------------------
 Z(t)           -10.951          -3.472            -2.882           -2.572
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000

------------------------------------------------------------------------------
       D.lnp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnp |
         L1. |  -.7188516   .0656454   -10.95   0.000    -.8482495   -.5894537
             |
       _cons |   1.848315   .1703773    10.85   0.000     1.512473    2.184156
------------------------------------------------------------------------------
```

有截距

取一階差分

```
. dfuller newlnp if train ==0,trend reg

Dickey-Fuller test for unit root                    Number of obs   =       214

                                 ---------- Interpolated Dickey-Fuller ----------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic         Value             Value            Value
------------------------------------------------------------------------------
 Z(t)           -20.928          -4.002            -3.435           -3.135
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000

------------------------------------------------------------------------------
    D.newlnp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      newlnp |
         L1. |  -1.349839   .064498   -20.93   0.000    -1.476982   -1.222696
       _trend |  -.0001933   .0004366    -0.44   0.658    -.0010539    .0006674
       _cons |   .0223587   .0541274     0.41   0.680    -.084341    .1290584
------------------------------------------------------------------------------
```

無trend

```
. dfuller newlnp if train ==0,reg

Dickey-Fuller test for unit root                    Number of obs   =       214

                                 ---------- Interpolated Dickey-Fuller ----------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic         Value             Value            Value
------------------------------------------------------------------------------
 Z(t)           -20.964          -3.472            -2.882           -2.572
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000

------------------------------------------------------------------------------
    D.newlnp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      newlnp |
         L1. |  -1.349232   .064361   -20.96   0.000    -1.476102   -1.222363
             |
       _cons |   .0015842   .0269138     0.06   0.953    -.0514687    .0546371
------------------------------------------------------------------------------
```

無截距

Dickey–Fuller

時間序列模型:
Random Walk with Drift (RWWD)

```
. dfuller lnp if train ==0,noconstant reg

Dickey-Fuller test for unit root                   Number of obs   =       215

                         ------------ Interpolated Dickey-Fuller ------------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic         Value             Value            Value
------------------------------------------------------------------------------
 Z(t)            -1.201            -2.585            -1.950           -1.618
------------------------------------------------------------------------------

------------------------------------------------------------------------------
       D.lnp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnp |
         L1. |  -.0131664   .0109663    -1.20   0.231    -.0347822    .0084494
------------------------------------------------------------------------------
```

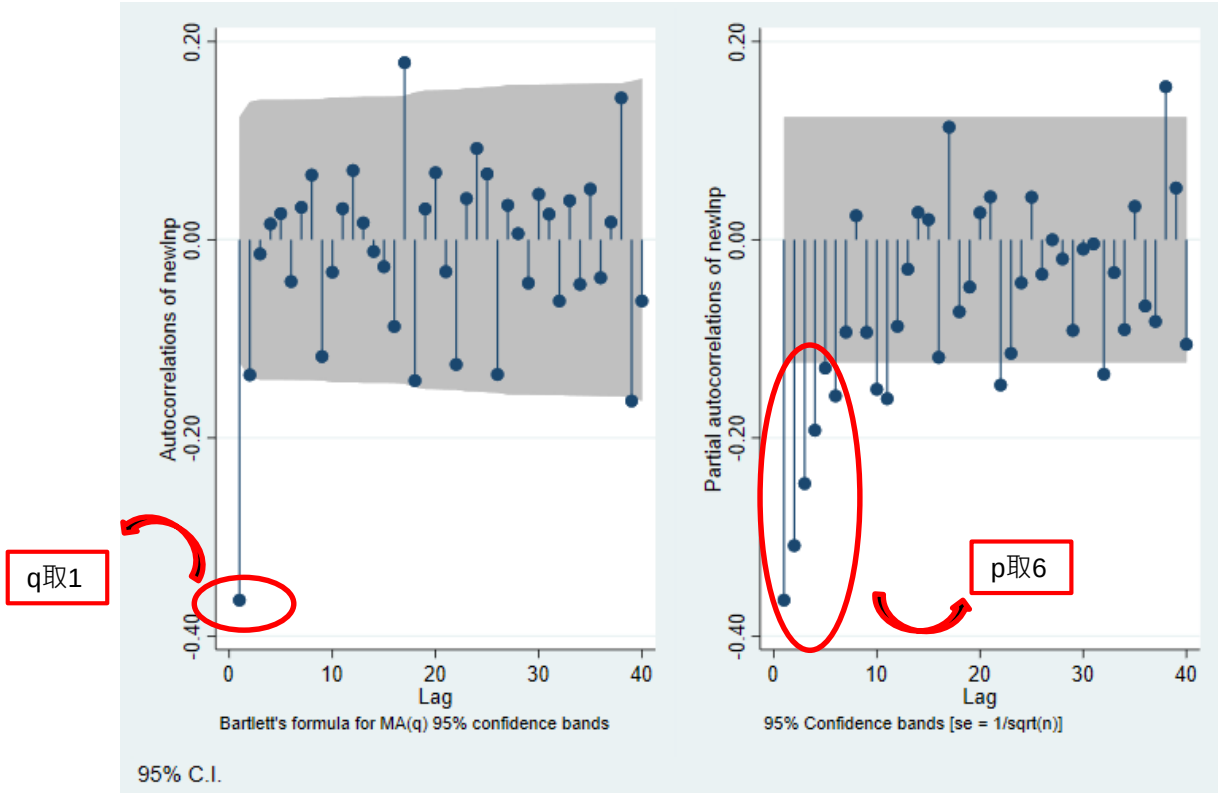
有單根

取一階差分

```
. dfuller newlnp if train ==0,noconstant reg

Dickey-Fuller test for unit root                   Number of obs   =       214

                         ------------ Interpolated Dickey-Fuller ------------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic         Value             Value            Value
------------------------------------------------------------------------------
 Z(t)           -21.013            -2.585            -1.950           -1.618
------------------------------------------------------------------------------

------------------------------------------------------------------------------
    D.newlnp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      newlnp |
         L1. |  -1.349241   .0642101   -21.01   0.000    -1.47581   -1.222673
------------------------------------------------------------------------------
```

無單根

Acf&Pacf


Autocorrelations of newlnp
0.20
0.00
-0.20
-0.40
0
10
20
30
40
Lag
Bartlett's formula for MA(q) 95% confidence bands
Partial autocorrelations of newlnp
95% Confidence bands [se = 1/sqrt(n)]
95% C.I.
q取1
p取6

AIC,BIC to find best model

ARIMA regression

Sample: 1999m2 - 2016m12 | Number of obs = 215
Wald chi2(6) = 1016.94
Log likelihood = -66.27992 | Prob > chi2 = 0.0000

train_p	Coef.	OPG Std. Err.	z	P>|z|	[95% Conf.	Interval]
train_p						
_cons	-.0013361	.000827	-1.62	0.106	-.0029569	.0002847
ARMA						
ar						
L1.	-.712499	.0806946	-8.83	0.000	-.8706575	-.5543404
L2.	.1749633	.0924536	1.89	0.058	-.0062424	.3561689
L3.	-.025643	.1077966	-0.24	0.812	-.2369204	.1856345
L4.	.0126169	.0750153	0.17	0.866	-.1344103	.1596441
ma						
L1.	1.05e-06	890.1886	0.00	1.000	-1744.738	1744.738
L2.	-1	987.459	-0.00	0.999	-1936.384	1934.384
/sigma	.3240561	159.9952	0.00	0.499	0	313.9089

不顯著

Note: The test of the variance against zero is one sided, and the two-sided confidence interval is truncated at zero.

4.0.1

df	AIC	BIC
7	150.546	174.1405

BEST

4.0.2

df	AIC	BIC
8	148.5598	175.525

5.0.1

df	AIC	BIC
8	151.4846	178.4497

5.0.2

df	AIC	BIC
9	149.8891	180.2248

6.0.1

df	AIC	BIC
9	153.4748	183.8105

6.0.2

df	AIC	BIC
9	152.9415	183.2773

Auto-ARIMA

```
model = pm.auto_arima(df.newlnp, start_p=4,start_q=2,
                      information_criterion='aic',
                      test='adf',
                      max_p=6, max_q=2,
                      m=1,
                      d=None,
                      seasonal=False,

                      trace=True,
                      error_action='ignore',
                      suppress_warnings=True,
                      stepwise=True)

print(model.summary())
```

Best model

```
Performing stepwise search to minimize aic
 ARIMA(4,0,2)(0,0,0)[0]             : AIC=159.363, Time=0.40 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=258.730, Time=0.02 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=225.169, Time=0.02 sec
 ARIMA(0,0,1)(0,0,0)[0]             : AIC=162.353, Time=0.03 sec
 ARIMA(3,0,2)(0,0,0)[0]             : AIC=inf, Time=0.37 sec
 ARIMA(4,0,1)(0,0,0)[0]             : AIC=158.052, Time=0.17 sec
 ARIMA(3,0,1)(0,0,0)[0]             : AIC=156.627, Time=0.12 sec
 ARIMA(2,0,1)(0,0,0)[0]             : AIC=154.832, Time=0.07 sec
 ARIMA(1,0,1)(0,0,0)[0]             : AIC=153.110, Time=0.06 sec
 ARIMA(1,0,2)(0,0,0)[0]             : AIC=154.778, Time=0.11 sec
 ARIMA(0,0,2)(0,0,0)[0]             : AIC=152.852, Time=0.06 sec
 ARIMA(0,0,2)(0,0,0)[0] intercept   : AIC=inf, Time=0.30 sec

Best model:  ARIMA(0,0,2)(0,0,0)[0]
Total fit time: 1.744 seconds
```

顯著

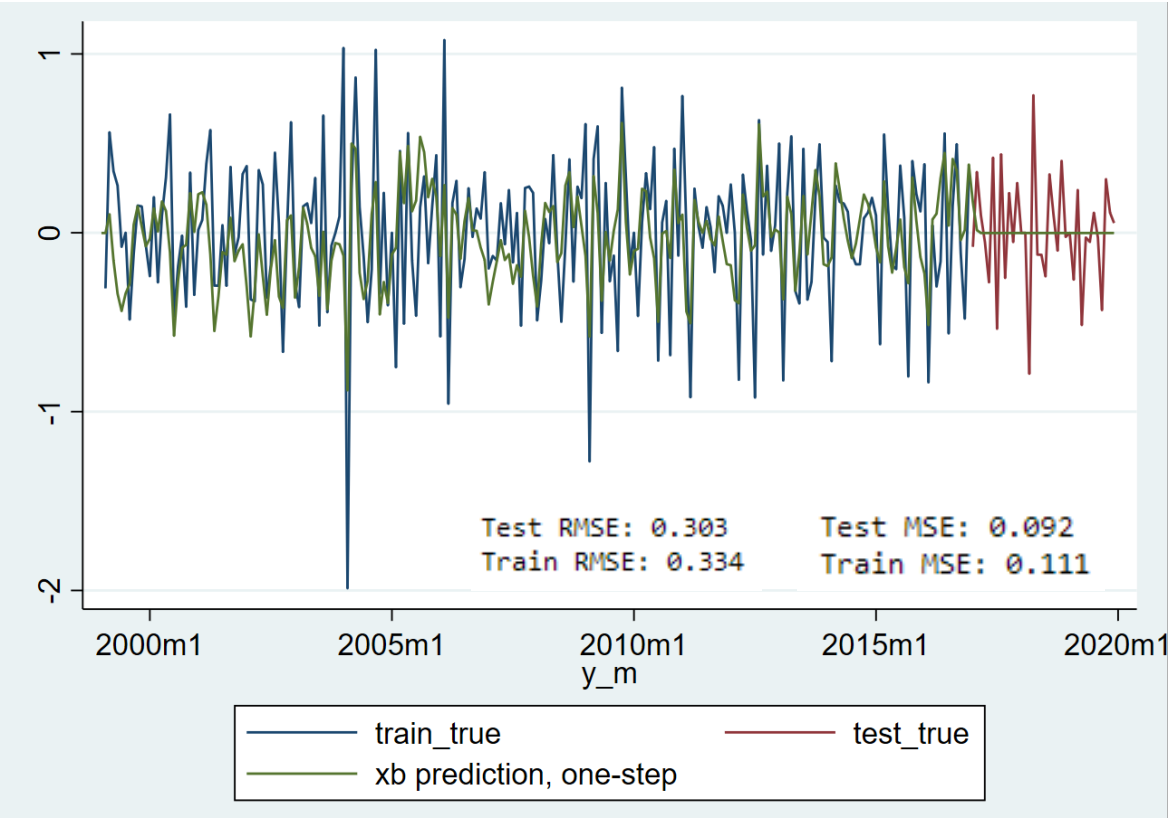
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  251
Model:               SARIMAX(0, 0, 2)   Log Likelihood                -73.426
Date:                Wed, 06 Jan 2021   AIC                            152.852
Time:                        23:52:07   BIC                            163.428
Sample:                             0   HQIC                           157.108
                                  - 251
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7228      0.055    -13.058      0.000      -0.831      -0.614
ma.L2         -0.2300      0.059     -3.866      0.000      -0.347      -0.113
sigma2         0.1042      0.008     12.526      0.000       0.088       0.120
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 8.51
Prob(Q):                              0.95   Prob(JB):                         0.01
Heteroskedasticity (H):               0.60   Skew:                            -0.19
Prob(H) (two-sided):                  0.02   Kurtosis:                         3.82
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

White Noise Q test

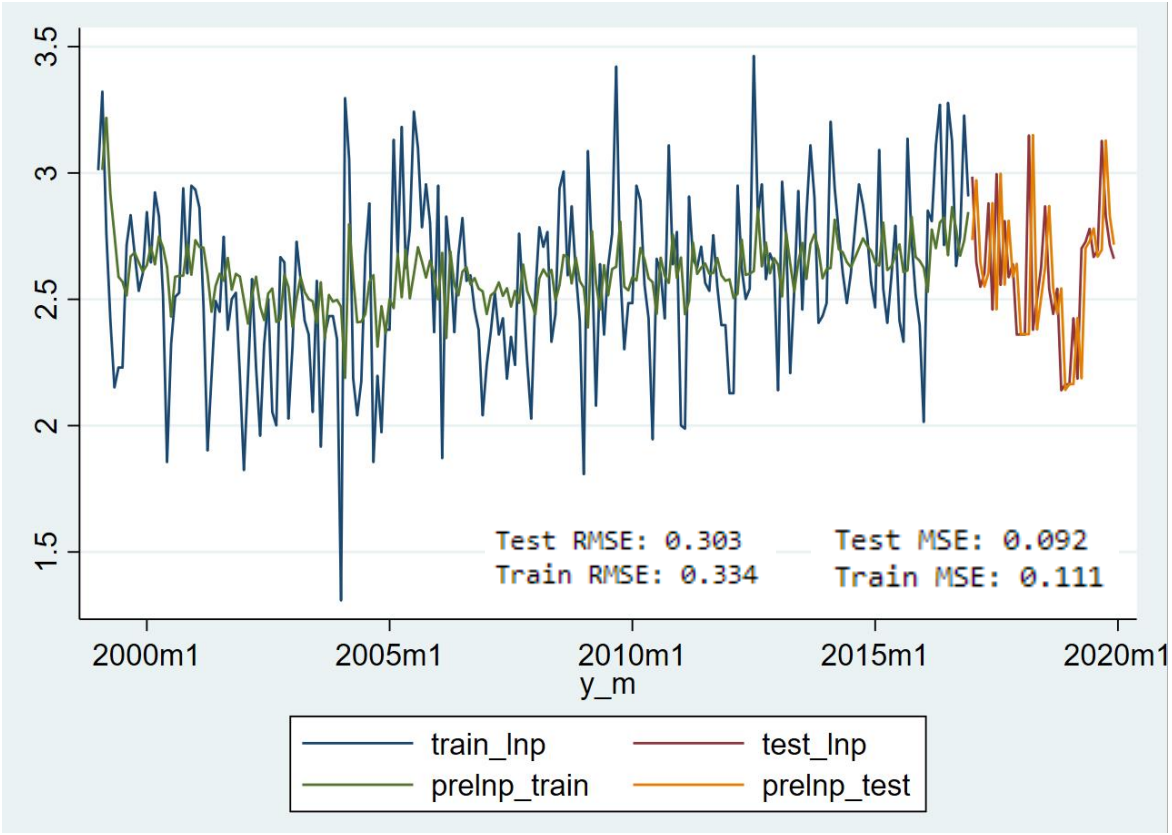
```
. //對殘差檢定是否有白噪音--WN檢定
. wntestq e

Portmanteau test for white noise
---------------------------------------------------------------------------
 Portmanteau (Q) statistic =     43.6446
 Prob > chi2(40)           =     0.3193
```

不拒絕虛無(存在白噪音)

Predict

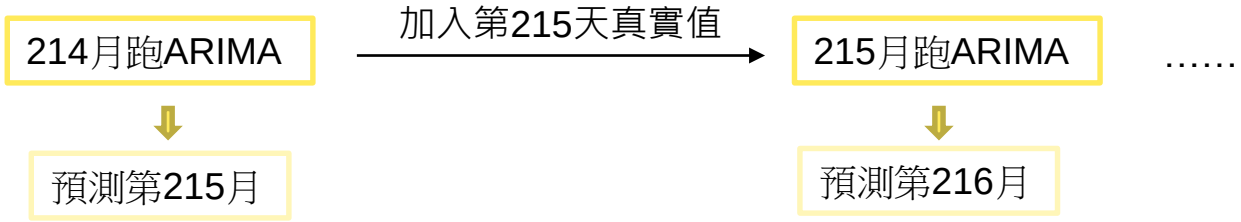

1
0
-1
-2
Test RMSE: 0.303
Train RMSE: 0.334
Test MSE: 0.092
Train MSE: 0.111
2000m1
2005m1
2010m1
2015m1
2020m1
y_m
train_true
test_true
xb prediction, one-step

Predict(還原lnPrice)

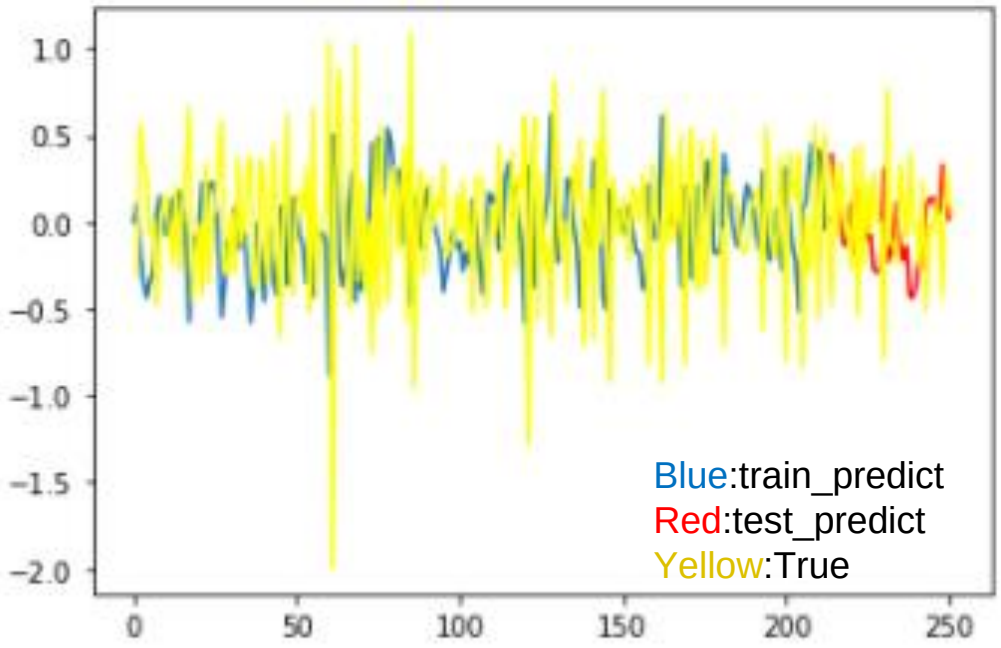


ARIMA by roll

```
predictions = []
for t in range(len(test)):
    model = ARIMA(history, order=(0,0,2))
    model_fit = model.fit(disp=0)
    output = model_fit.forecast()
    yhat = output[0]
    predictions.append(yhat)
    obs = test[t]
    history.append(obs)
    print('predicted=%f, expected=%f' % (yhat, obs))
```



Predict by roll



Test MSE: 0.069
Train MSE: 0.111

Test RMSE: 0.263
Train RMSE: 0.334

LSTM

```
def buildTrain(train, pastDay, futureDay):
    X_train, Y_train = [], []
    for i in range(train.shape[0]-futureDay-pastDay):
        X_train.append(np.array(train.iloc[i:i+pastDay]))
        Y_train.append(np.array(train.iloc[i+pastDay:i+pastDay+futureDay]['newlnp']))
    return np.array(X_train), np.array(Y_train)
```

```
test = np.array([test])
print(len(test[0]))
```

```
37
```

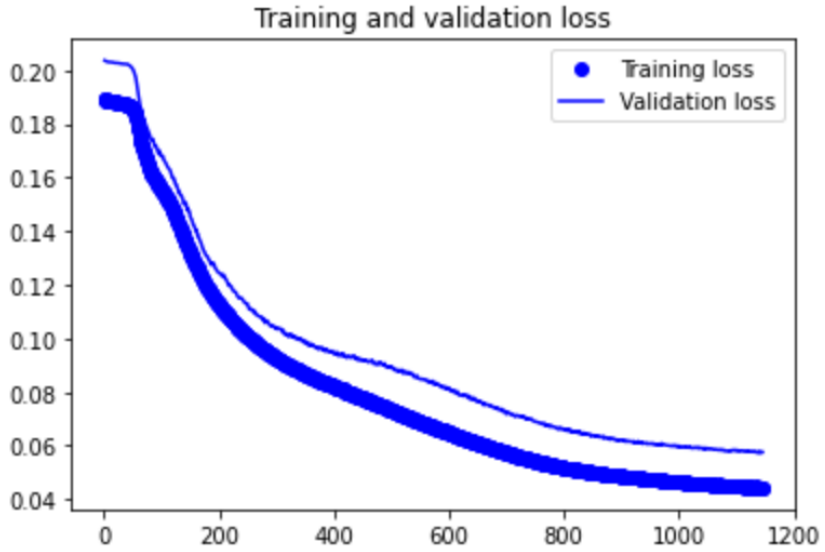
```
def buildManyToManyModel(shape):
    model = Sequential()
    model.add(LSTM(10, input_length=shape[1],
    model.add(TimeDistributed(Dense(1)))
    model.compile(loss="mse", optimizer="adam")
    model.summary()
    return model
```

1.214月以37月為單位分割

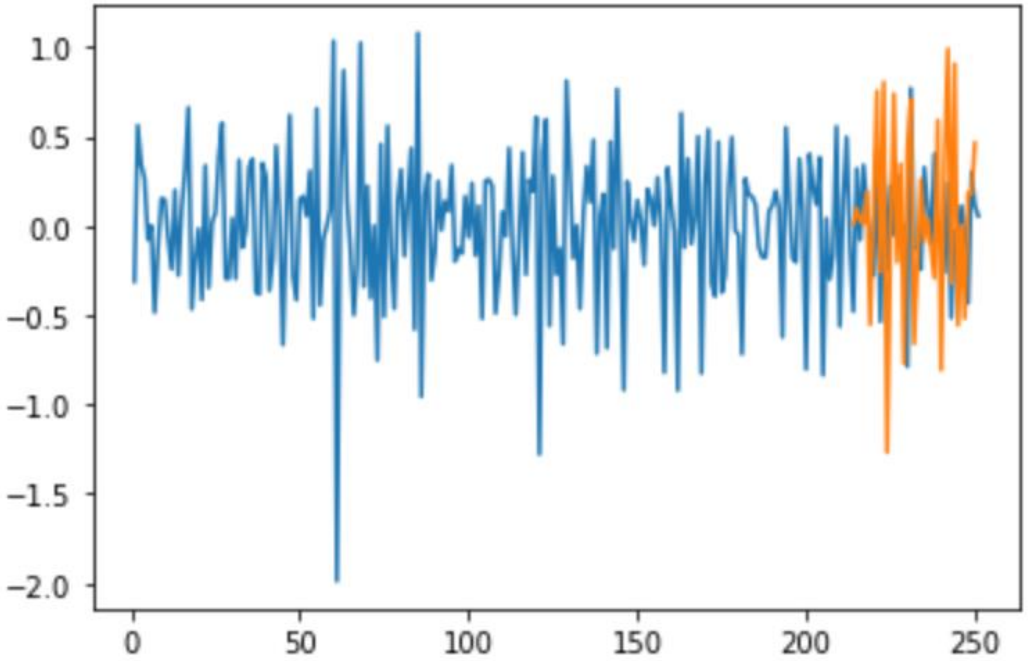
2.拆分為訓練與驗證資料

3.跑模型

4.輸入要預測的前37月做預測



LSTM



Test MSE: 0.334
Test RMSE: 0.578

Conclusion

1.ARIMA在預測短期資料有較好的表現

2.較長期資料的預測會出現無效預測

3.長期可用滾動預測的方式處理，但需逐步加入真實資料

4.若無真實資料又要進行長期預測，則LSTM會有較佳表現，但需要較大樣本

5.ACF圖若不好判斷pq，則可用auto方式找尋AIC最佳之模型

未解決問題:

1.若不對非定態進行處理，會有什麼問題?不同非定態因素是否有不同?

2.若模型AIC為最佳，但係數Pvalue較大(不顯著)是否有影響