Econometrics Final Report Time Series Analysis of Price of Mango

HSUAN-YU CHEN

2017/6/1

研究動機:

在即將迎來的溽暑中，最具代表性的水果當屬「愛文芒果」，因具有速生、高產、營養價值高、經濟價值高等優點，目前也並列世界五大水果之一（葡萄、柑橘、香蕉、蘋果、芒果）。而台灣因為雨量豐富且氣溫合適，地雖小，但也在世界芒果產量中排名第十。過去受限於運輸成本問題，大部分芒果還是國內自產自食，但這拜所賜，除了直接食用外，也誕生了許多膾炙人口的副產品：芒果冰、芒果乾...，更有人將芒果運用入料理。但隨著運輸成本漸低、跨國市場漸大，對外輸出量也是逐年增加，而其中以日本更是對台灣的愛文芒果讚不絕口，來台觀光客中，永康街的芒果冰也是不可錯過的行程之一。
因此，此次研究目的在於，透過研究過往芒果的市場交易價以及交易量來觀察其價格變動模式，因芒果本身是屬於季節性極強的水果，但是否還有其他被我們平時所忽略的因素也在影響著芒果的價錢呢？我們將在接下來透過研究價格變動的歷史資料裡一探究竟。

```
## Loading required package: magrittr
```

```
## Loading required package: ggplot2
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: urca
```

```
## Loading required package: prais
```

```
## Loading required package: ecm
```

```
## Loading required package: FinTS
```

```
## Loading required package: car
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

研究主文:

我們的以農糧署的公開平台上芒果果價資料為標的, 再決定解釋變數前, 先檢驗果價自身的相關性, 用來決定是否以原始資料當作應變數或是需進行差分消除相關性, 在此我們已 Dicky - Fuller 的 unit root test 來檢驗, 結果如下

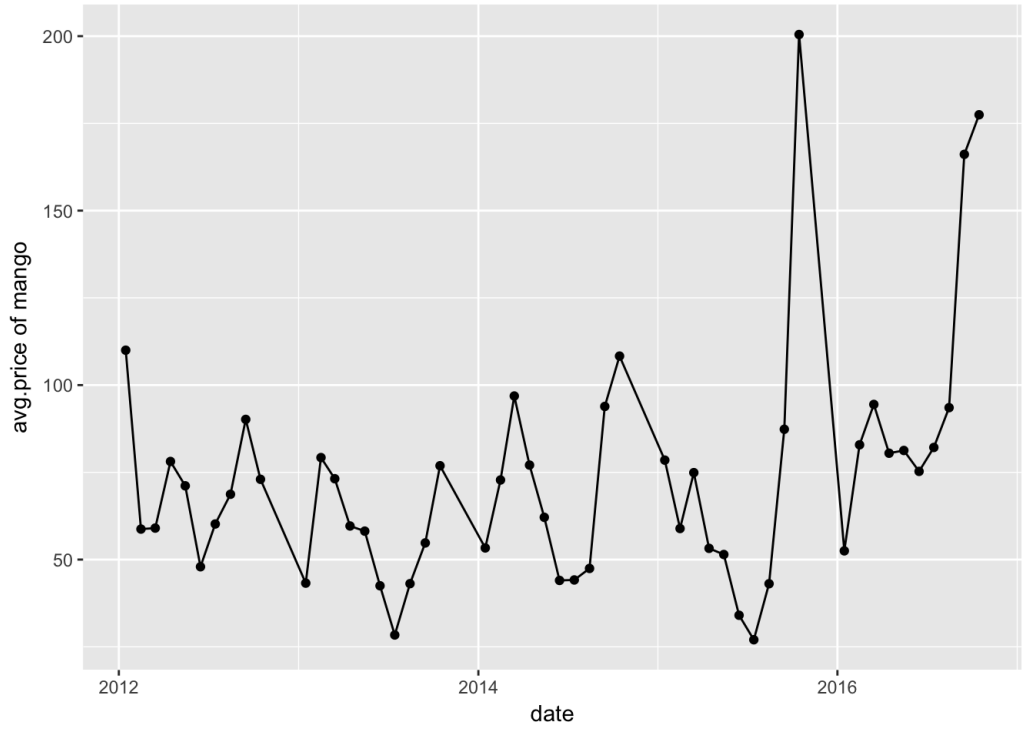
```
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression none
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.075  -13.780   -0.897   16.975  125.008
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## z.lag.1    -0.02037    0.06670  -0.305    0.761
## z.diff.lag -0.22816    0.14591  -1.564    0.125
##
## Residual standard error: 33.96 on 46 degrees of freedom
## Multiple R-squared:  0.06256,    Adjusted R-squared:  0.0218
## F-statistic: 1.535 on 2 and 46 DF,  p-value: 0.2263
##
##
## Value of test-statistic is: -0.3054
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau1 -2.62 -1.95 -1.61
```

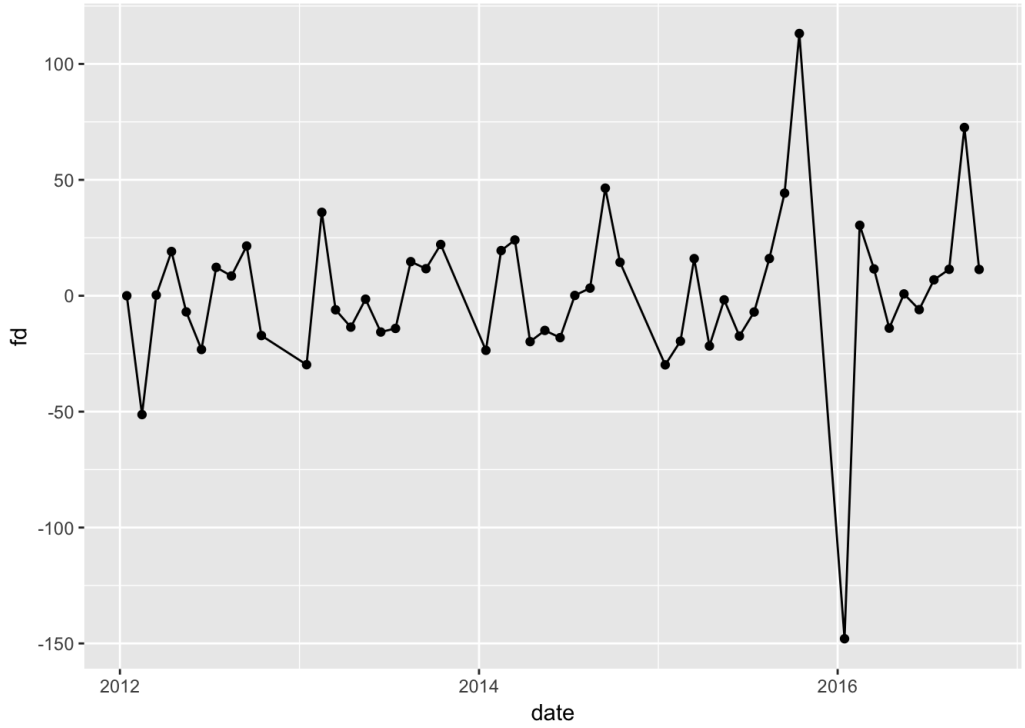
可看到 unit root test 不拒絕虛無假設, 代表果價非穩定, 形式為:

y(t) = phi*y(t-1) + error(t), 其中 |phi| < 1 ,

因此月平均售價資料需做差分。

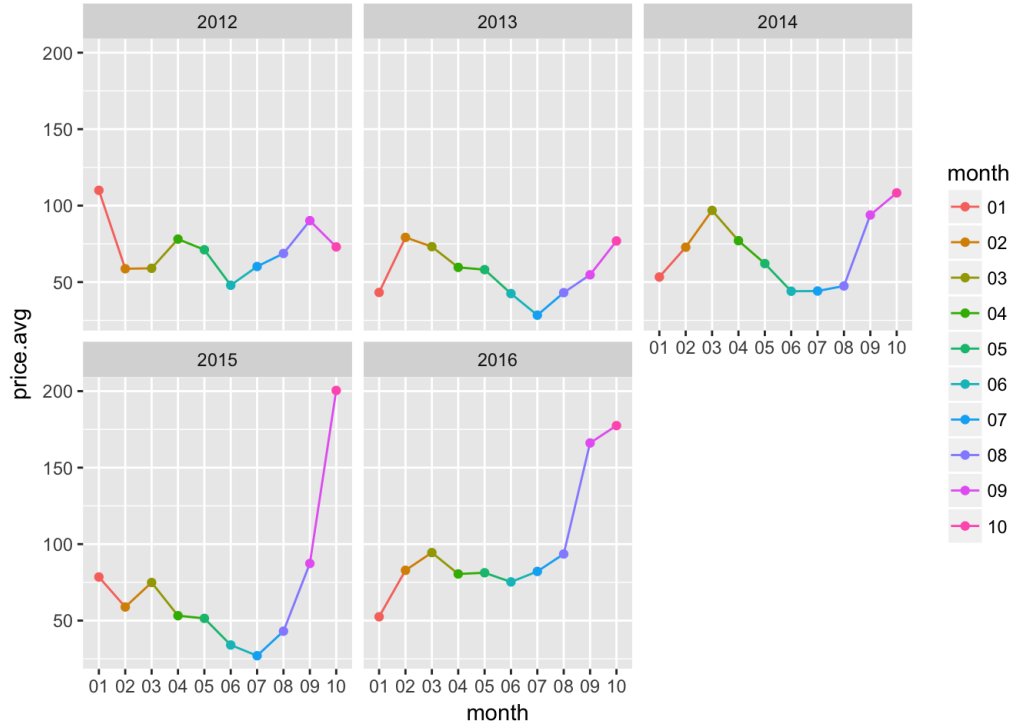
接下來我們分別對原始資料與差分後資料作時間序列, 發現原始資料有三點值得注意的地方:時間趨勢 ,非同質變異數 ,季節性, 而這些因素在差分後的資料中以大部分被消除,





由兩者時間序列資料可看到，原始資料相較差分後的資料不穩定的許多，因此我們確定以差分後資料當作被解釋變數。觀察上面差分的時間序列圖可發現，雖然在前半部有穩定的結果，但 2015 後期及 2016 前期有較大幅度的變動，因此我們回到原始資料來檢視發生的原因。

可看到下圖從 2013 開始，9 & 10 月的價格變化的幅度持續增加，到 2015 達到高峰，2016 時 幅度則稍減，我們不確定是否為特定因素所造成的影響，以及放入模型內是否有非常大的差異，因此我們先配模後再進行檢測。

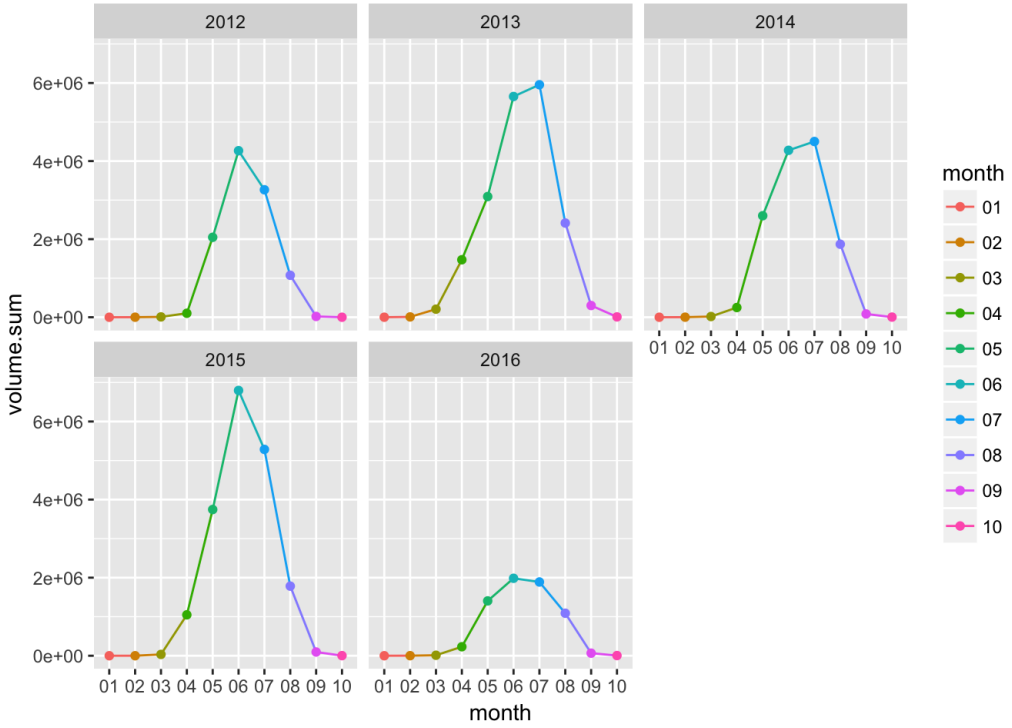


在決定價格模型時，我們通常考慮兩個主要因素：「需求量」、「供給量」，因此我們假設模型的形式為：

果價變化(t) = Bo + B1^需求量變化(t) + B2^供給量變化(t) + error(t),

而對於這兩個無法實際捕捉的變數，分別以其他代理變數來檢驗。

在高需求量部分，我們以交易量的變化做為代表，但交易量同時含括供給及需求，因此其實不能算是非常好的代理，需謹慎解釋，先觀察累積交易量未差分前的各年每個月的時間序列圖，我們將焦點放在交易量較高的 5.6.7.8 月



可看到每年都是以夏季為交易量高峰季節，若由原先交易價的時間圖來推側的話，旺季時 2013.2014.2015 都有較低的價格，2012次之，2016最高，因此推側交易量的變動模式 2013.2014.2015應該會有相近的型態，觀察實際資料後也確實如此，而且旺季平均價格最高的 2016 也確實交易量最少，雖然乍看交易量可能是「需求量」好的代理變數，但其實如水果及蔬菜如此價格變動大的目標，通常供給量才是主要造成市場價格變化的來源，因此應該是先由供給量決定售價後，才達到最終的交易量。

接下來我們回來觀察交易量變化所造成的影響，以回歸模型來檢驗是否該期交易量的變化或落後一期交易量的變化有解釋力

```
##
## Call:
## lm(formula = diff(price.avg)[2:49] ~ diff(volume.sum)[2:49] +
##     diff(volume.sum)[1:48], data = month)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -151.111   -9.362    2.037   10.971   98.509
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.483e+00  4.681e+00   0.531   0.5984
## diff(volume.sum)[2:49] -3.762e-06  3.718e-06  -1.012   0.3171
## diff(volume.sum)[1:48] -6.986e-06  3.718e-06  -1.879   0.0667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.43 on 45 degrees of freedom
## Multiple R-squared:  0.1591, Adjusted R-squared:  0.1217
## F-statistic: 4.256 on 2 and 45 DF,  p-value: 0.02028
```

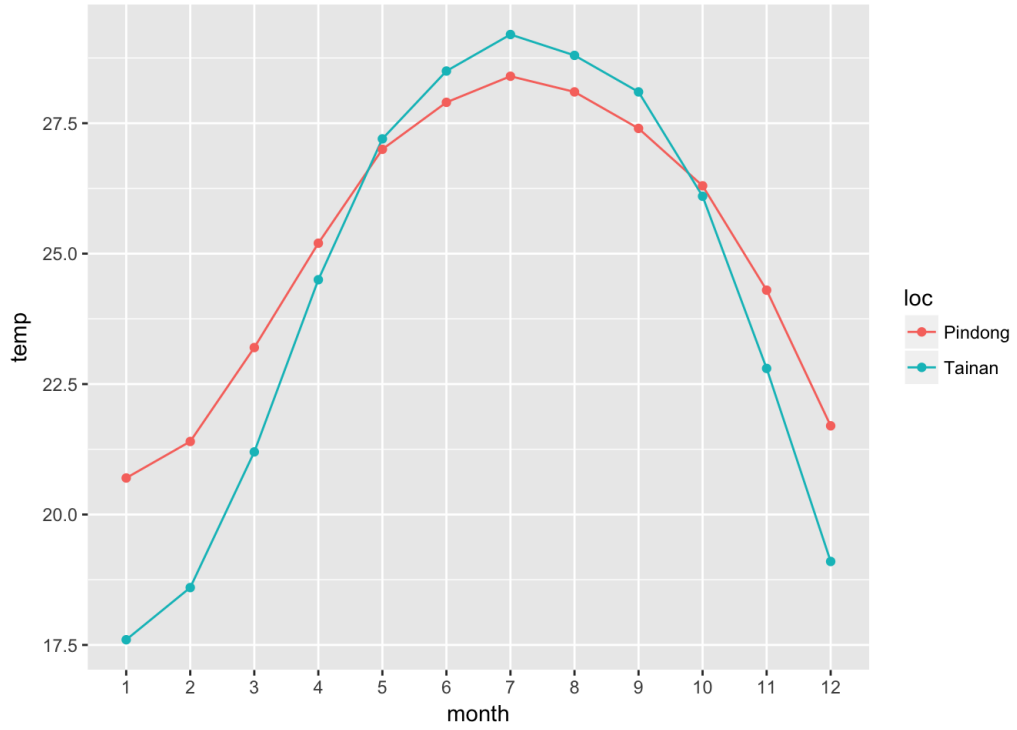
```
##
## Call:
## lm(formula = diff(price.avg) ~ diff(volume.sum), data = month)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -149.366  -13.305    0.621   12.478  111.082
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.378e+00  4.836e+00   0.285   0.7770
## diff(volume.sum) -7.215e-06  3.374e-06  -2.138   0.0377 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.86 on 47 degrees of freedom
## Multiple R-squared:  0.08866,    Adjusted R-squared:  0.06927
## F-statistic: 4.573 on 1 and 47 DF,  p-value: 0.03771
```

```
##
## Call:
## lm(formula = diff(price.avg)[2:49] ~ diff(volume.sum)[1:48],
##     data = month)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -151.28  -10.05    0.59   12.73   95.72
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.486e+00  4.682e+00   0.531  0.59806
## diff(volume.sum)[1:48] -8.845e-06  3.233e-06  -2.736  0.00881 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.44 on 46 degrees of freedom
## Multiple R-squared:  0.1399, Adjusted R-squared:  0.1212
## F-statistic: 7.485 on 1 and 46 DF,  p-value: 0.00881
```

在考慮交易量部分，先一次放入當期與延遲變數，可看到整體模型有顯著但兩個係數都不顯著，可能是因為共線性因素，因此分別做回歸來檢定，發現在落後一期的差分變數解釋程度較高，因此在變數選擇上，我們已落後一期的交易量差分值當對象，這其實也較符合一般市場的想法，供應商再決定下一期的價格前，會考慮的應該是前一期的交易量，隨著其上升或下降來決定下一期價格的高低。

接下來從「供給面」來探討，許多相關芒果種植的文獻都指出，有三點對種植產生影響：「溫度」、「水分」、「濕度」。其中的「水分」與「濕度」，其實可以看作雨量的「長期」、「短期」影響。因芒果有分「早熟」、「晚熟」種，以防高度密集生產下降低經濟價值，但主要都在夏季、秋季，在長期面，此時因西南季風，主要種植地區（台南、高雄、屏東）等都在迎風面的雨季當中，因此水分不是太大的影響，而在短期的「濕度」面，主要為結果期的後期，不能有短期內有太大的雨量，但就算產地相同，各個果園在結果期可能也有數天的差距，因此，此影響較不易觀測，我們暫時不考慮。於是我們將重點放在「溫度」對於產量的影響，但在台灣芒果產量中，雖然名為「愛文芒果」，其實產地除了台南玉井之外，還有屏東枋山，而兩地的產季也因冬季溫度不同，分別主要供應不同台灣前後期的芒果需求。

我們先觀察從氣象局從 1981 統計到2010 台南與高雄的月均溫資料，



可看到台南的月均溫較屏東有集中的趨勢，而這對於芒果產量的影響主要在於「開花季時的需求溫度」。芒果生長最合適的年均溫為 21 ~ 27度，台南與屏東都大致符合，但芒果樹的生長溫度則為 18 ~ 35度，其中枝梢的生長最適溫度為 24 ~ 29 度（在採果完成後就要開始），而最重要的決定性因素為開花期與幼果發育初期的氣溫，此時溫度的要求為 20 度以上，因此在經歷完收成與發芽後，屏東的芒果大約可在12月便開始進行開花與幼果發育，而台南的芒果則較晚能才開始，因此在旺季供應上，也造成台南主要供應 6 月下旬至 7 月下旬而屏東供應 5 月下旬至 6 月下旬。於是我們先分別檢測兩地氣溫的解釋力。

一開始，先同時將兩地氣溫的延遲差分變數都放入模型，檢驗後發現雖然模型整體有幫助，但係數卻分別都不顯著，分別做檢定後，發現台南的氣溫變化對於價差有稍高的解釋力，且參考相關文獻後，目前雖然兩地都有生產，但主力還是以台南供應的為主，佔整體超過一半的比例，除此之外，屏東的溫差也較小，對於變動量大的交易價自然解釋程度也很低，因此我們在此以台南的月均溫當作解釋變數。

```
##
## Call:
## lm(formula = diff(price.avg) ~ diff(PDtemp.lag4) + diff(TNtemp.lag8),
##     data = month)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -134.658  -14.875   -2.651   12.021  106.660
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.7354     4.7131   0.156    0.877
## diff(PDtemp.lag4)   3.7082     3.0903   1.200    0.236
## diff(TNtemp.lag8)  -2.5422     1.8048  -1.409    0.166
##
## Residual standard error: 32.93 on 46 degrees of freedom
## Multiple R-squared:  0.156,  Adjusted R-squared:  0.1193
## F-statistic: 4.252 on 2 and 46 DF,  p-value: 0.0202
```

```
##
## Call:
## lm(formula = diff(price.avg) ~ diff(PDtemp.lag4), data = month)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -137.706  -16.867   -0.123    9.785   97.493
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.106      4.755   0.233   0.8171
## diff(PDtemp.lag4)    6.317      2.500   2.527   0.0149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.28 on 47 degrees of freedom
## Multiple R-squared:  0.1196, Adjusted R-squared:  0.1009
## F-statistic: 6.387 on 1 and 47 DF,  p-value: 0.01492
```

```
##
## Call:
## lm(formula = diff(price.avg) ~ diff(TNtemp.lag8), data = month)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -137.482  -15.431   -1.898   14.671  116.704
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.648      4.735   0.137   0.8917
## diff(TNtemp.lag8)   -3.840      1.452  -2.646   0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.09 on 47 degrees of freedom
## Multiple R-squared:  0.1296, Adjusted R-squared:  0.1111
## F-statistic: 6.999 on 1 and 47 DF,  p-value: 0.01106
```

決定變數後, 目前我們的模型為:

果價變化(t) = Bo + B1 ^ 交易量變化(t-1) + B2 ^ 台南氣溫變化(t-8)

因我們的應變數與解釋變數都是差分後的變數, 若 X 也為 I(1) process, 且模型殘差為定態, 則可能是有共整合的問題, 因此我們先以 unit-root test 分別檢定兩個解釋變數是否為 I(1) , 下方報表結果顯示兩變數都拒絕虛無假設, 代表兩變數皆為 I(0), 因此不會有共整合的問題。

```
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression none
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1917771     7787   219770  1046678  2459769
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## z.lag.1    -0.30556    0.06848  -4.462 5.21e-05 ***
## z.diff.lag  0.64704    0.11242   5.755 6.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1074000 on 46 degrees of freedom
## Multiple R-squared:  0.4726, Adjusted R-squared:  0.4496
## F-statistic: 20.61 on 2 and 46 DF,  p-value: 4.072e-07
##
##
## Value of test-statistic is: -4.4621
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau1 -2.62 -1.95 -1.61
```

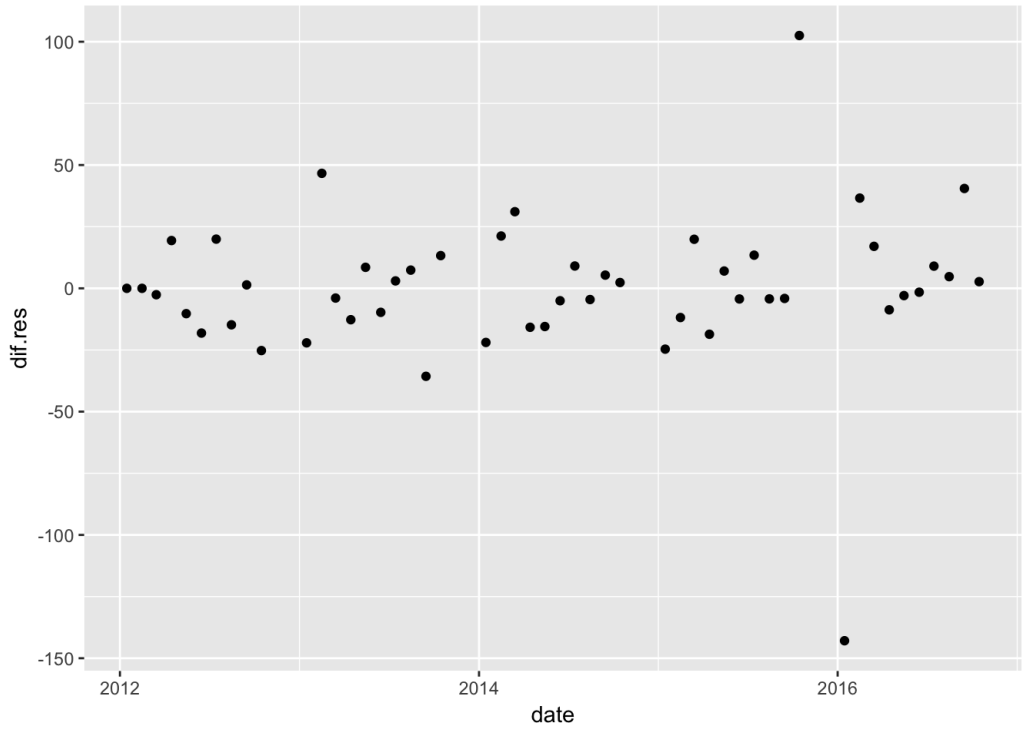
```
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression none
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.516 -1.126  1.012  2.238  5.830
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## z.lag.1    -0.01735    0.01765  -0.983   0.3307
## z.diff.lag  0.32429    0.13728   2.362   0.0225 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.134 on 46 degrees of freedom
## Multiple R-squared:  0.1236, Adjusted R-squared:  0.08547
## F-statistic: 3.243 on 2 and 46 DF,  p-value: 0.04813
##
##
## Value of test-statistic is: -0.9832
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau1 -2.62 -1.95 -1.61
```

我們開始配飾模型, 並檢驗其解釋力, 可看到 R 平方接近 20 %, 雖然看起來不高, 但因我們的目標為差分後值, 還在可接受的範圍內。

```
dif.model <- lm(data = month, diff(price.avg)[2:49] ~ diff(volume.sum)[1:48] + diff(TNtemp.lag8)[2:49])
dif.model %>% summary
```

```
##
## Call:
## lm(formula = diff(price.avg)[2:49] ~ diff(volume.sum)[1:48] +
##     diff(TNtemp.lag8)[2:49], data = month)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -142.906  -12.062   -2.082   10.105  102.520
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.848e+00  4.595e+00   0.402   0.6894
## diff(volume.sum)[1:48]  -6.889e-06  3.354e-06  -2.054   0.0458 *
## diff(TNtemp.lag8)[2:49] -2.604e+00  1.486e+00  -1.752   0.0866 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.73 on 45 degrees of freedom
## Multiple R-squared:  0.1949, Adjusted R-squared:  0.1591
## F-statistic: 5.446 on 2 and 45 DF,  p-value: 0.00762
```

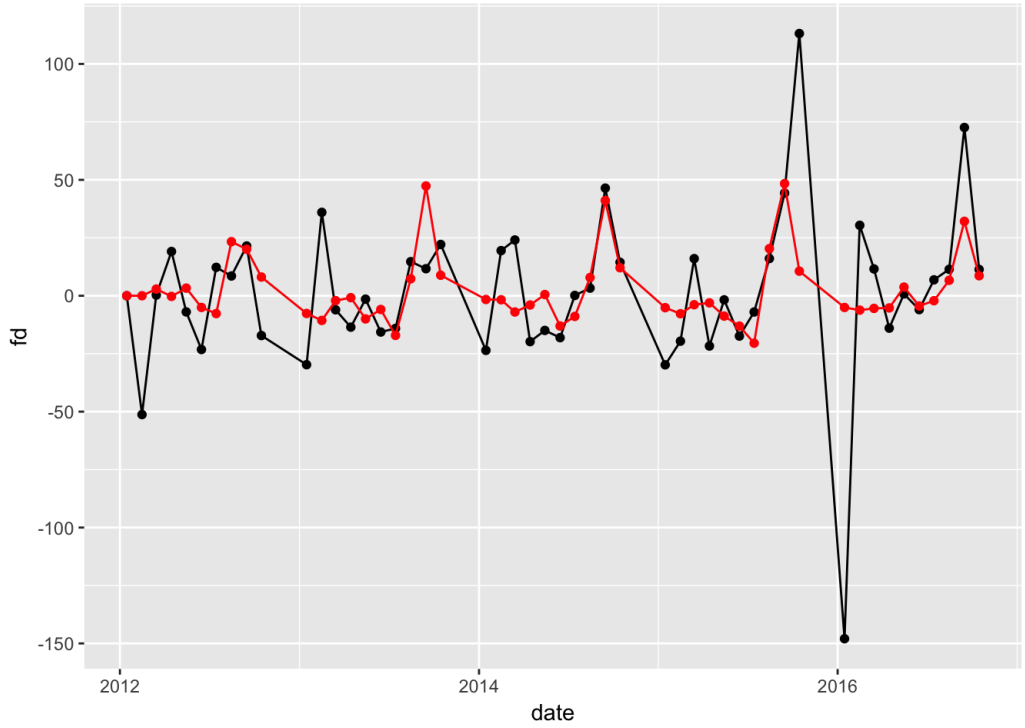
接著我們檢驗模型的殘差, 可看到有兩點, 明顯高於及低於其他殘差的散佈範圍中, 代表模型高估及低估, 而兩點分別為模型中第 38 & 39 的值,




```
##
## Call:
## lm(formula = dif.res[3:50] ~ dif.res[2:49], data = month)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.334  -9.643  -0.630   9.518 100.486
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.02836    3.91618  -0.007 0.994253
## dif.res[2:49] -0.50276    0.12747  -3.944 0.000271 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.13 on 46 degrees of freedom
## Multiple R-squared:  0.2527, Adjusted R-squared:  0.2365
## F-statistic: 15.56 on 1 and 46 DF,  p-value: 0.0002712
```

```
## [1] 38 39
```

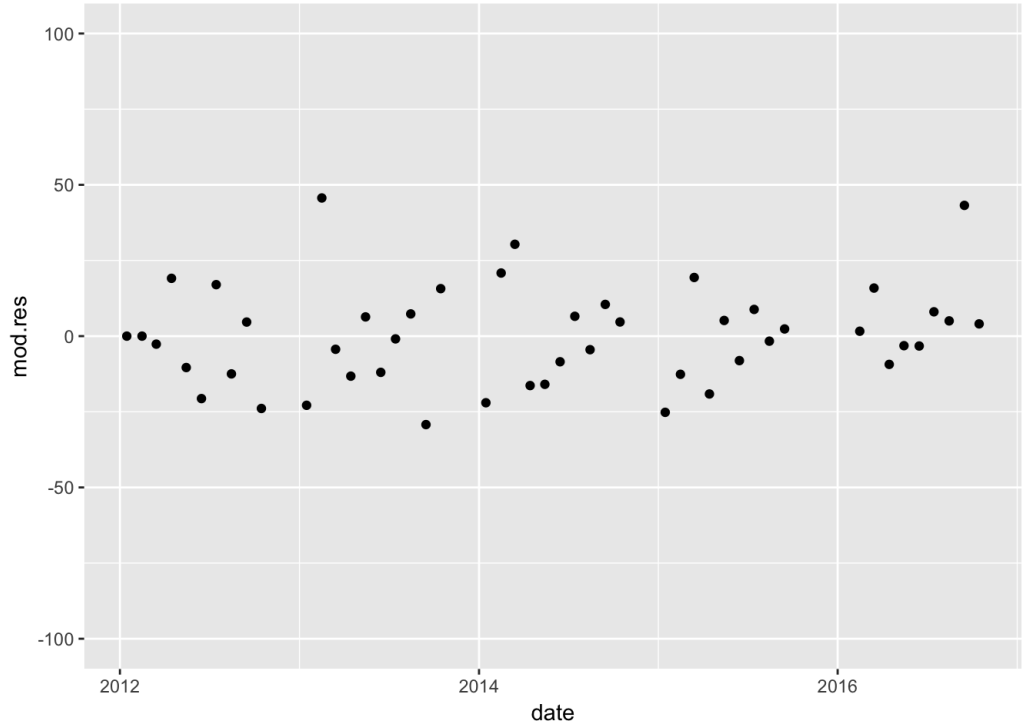
於是我們用原始資料及配飾值同時畫圖, 發現突起來源正是原始資料中有較大變動幅度的部分,



而據此推測, 這可能是造成 R 平方稍低的原因, 將極端值移除後, 或許能讓模型有更高的解釋力, 重新配飾後結果如下, R 平方接近 40 % , 且兩個變數都變得顯著, 代表移除極端值對於模型配飾非常有幫助,

```
##
## Call:
## lm(formula = diff(price.avg)[2:48] ~ diff(volume.sum)[1:47] +
##     diff(TNtemp.lag8)[2:48], data = month[-c(40, 41), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.237 -12.353  -1.290   7.864  45.655
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.972e+00  2.547e+00   0.774  0.44318
## diff(volume.sum)[1:47]   -5.442e-06  1.788e-06  -3.043  0.00398 **
## diff(TNtemp.lag8)[2:48] -2.426e+00  7.695e-01  -3.153  0.00294 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.23 on 43 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3828, Adjusted R-squared:  0.3541
## F-statistic: 13.33 on 2 and 43 DF,  p-value: 3.123e-05
```

我們接著分別檢驗模型殘差的散佈、序列相關性、變異數、散佈圖中, 殘差都在 正負 50 內跳動, 沒有明顯的異質性問題。



而以當期殘差對前一期殘差作回歸也沒有顯著，不存在序列相關性。

```
##
## Call:
## lm(formula = mod.model$residuals[2:46] ~ mod.model$residuals[1:45])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.347 -10.560   0.773   7.254  44.319
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.03836    2.49961   0.015    0.988
## mod.model$residuals[1:45] -0.22817    0.14853  -1.536    0.132
##
## Residual standard error: 16.77 on 43 degrees of freedom
## Multiple R-squared:  0.05202,    Adjusted R-squared:  0.02998
## F-statistic:  2.36 on 1 and 43 DF,  p-value: 0.1318
```

最後，用 Breusch - Pagan test 檢驗也不存在異質變異數，因此我們使用以下模型當作最終的模型。

```
##
##  studentized Breusch-Pagan test
##
## data:  mod.model
## BP = 0.80379, df = 2, p-value = 0.669
```

果價變化(t) = 1.972 - 5.442e-06 ^ 交易量變化(t-1) - 2.426 ^ 台南溫度變化(t-8)

結論：

在模型中，因氣溫為較遠的歷史資料，我們將重點放在討論交易量變化下，在氣溫不變的情況下，當該期的交易量上升時，我們就可以預測下一期的果價可能會繼續在下跌，而達到7 .8 月交易量開始下降時，反之我們可以預測果價將會上升。雖然交易量變化的係數非常小，但這可以歸咎於交易量在變動部分的程度是遠大於價格變動的，因此有大幅度的交易量變化時，我們不只可以推測價格的變動，還能大略推算出可能變動的幅度。

而我認為最能利用此模型的，是一般直接面對消費者的店家，不論是水果商或是以芒果製品為主的冰店等，在台南氣溫變化已經是已知的情況下（延後8期），店家可以先大略推算出果價起始的價格，再依據模型給定的係數，觀察每期的交易量變化以用來決定下一期應該購入的價格，而不只是隨著消費者的購買量才決定要批入的數量，讓商家達到能主動掌握成本的效益，先控制住成本後，就可以間接在規劃其他種水果可能需購入的數量，可以再依照上述配模的方式計算出店內其他水果的價錢模型並用多維的線性規劃來找出最大利益點。但有些水果不像芒果有如此強的季節性，因此模型中可能應該考慮其他變數而非產地的氣溫，而擁有了各式水果的價錢模型後，或許可以搭配店內的庫存系統，讓系統自動抓取線上資料平台的交易量狀況，並計算出下一期的合適購入價格，讓店主隨時掌握應進貨的數量，已達到利益的最大化。但其實果價還有另一項很重要的控制因素並沒有被考慮在模型內，那就是「購入市場」，供應商會與店家在公開市場上交易，而每種水果在各個市場上其實也有不同的交易價格與佔整體的交易比例，因此可以將模型再拆成專屬於每個市場的模型，更精準地計算出合適的價格。讓成本的控制權從供應商回到水果商自己手上。

資料來源：

http://data.coa.gov.tw/Query/OpenData.aspx（行政院農業委員會開放平台）

http://www.cwb.gov.tw/V7/climate/monthlyData/mD.htm (氣象局)

文獻參考：

http://pugker.blogspot.tw/2007/03/blog-post_5059.html（芒果相關資料）

http://www.coa.gov.tw/ws.php?id=2448072（農委會出版品）