Case study: Did itaewon class cause โคชูจัง?

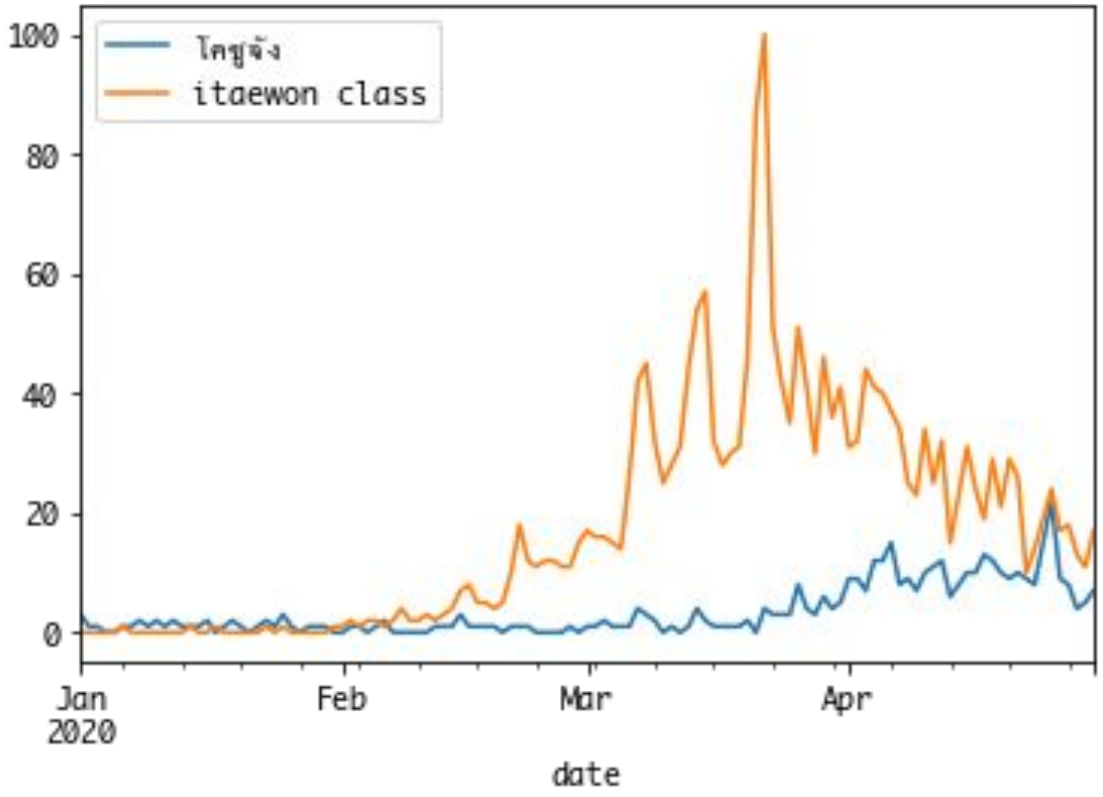
Granger Causality with Google Trend

By

cstorm125, ben-mj

Time Series Data

- 'itaewon class' and 'โคชูจัง' are a time series data that are collected from google trend since 2020-01-01 to 2020-04-30.



Non-stationary Vs. Stationary

- Time series data that can be used for forecasting must be stationary time series.

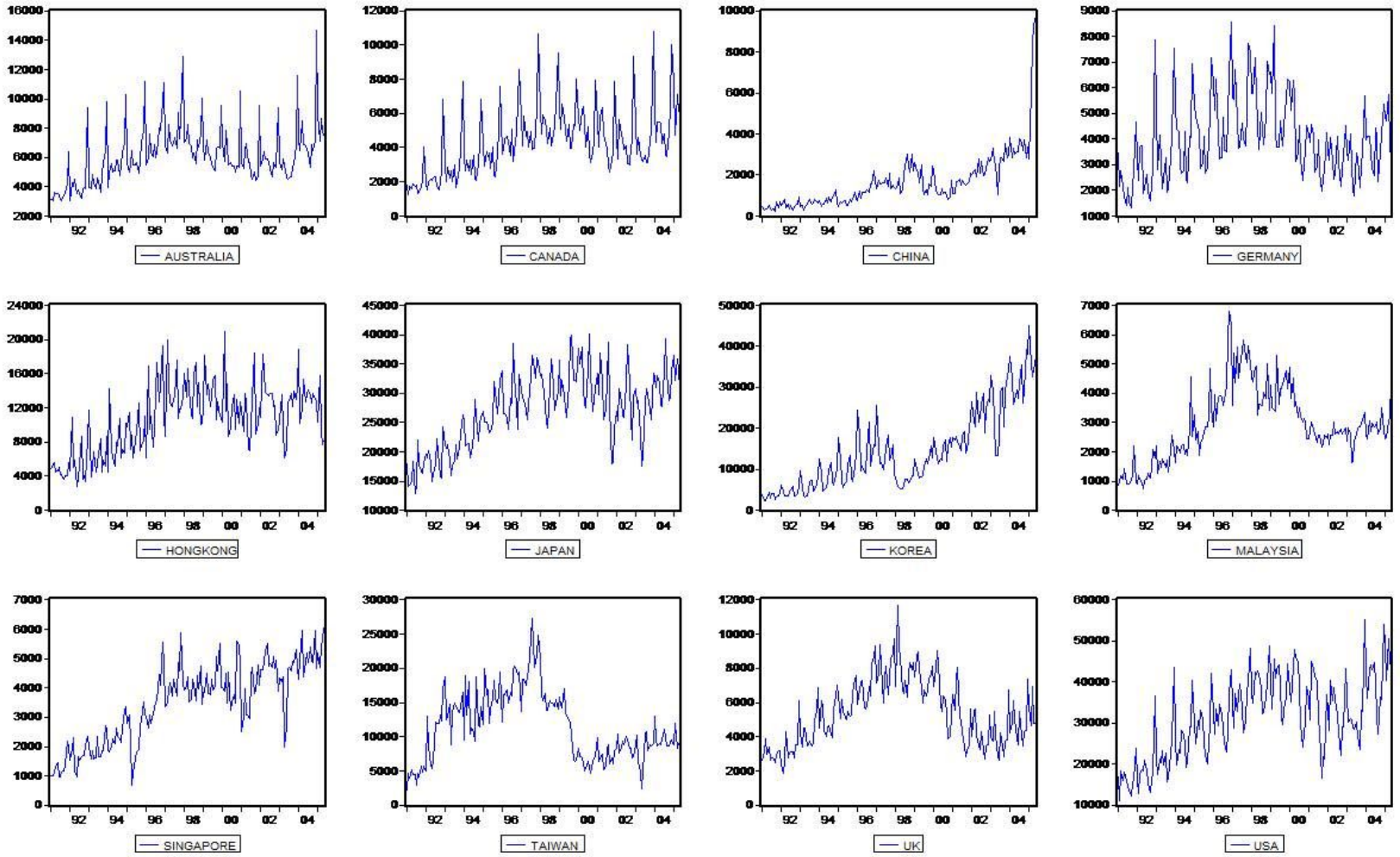
$$\text{Mean :} \quad E(Y_t) = \mu$$

$$\text{Variance :} \quad Var(Y_t) = E(Y_t - \mu)^2 = \sigma^2$$

$$\text{Covariance :} \quad E\left[(Y_t - \mu)(Y_{t+k} - \mu)\right] = \gamma_k$$

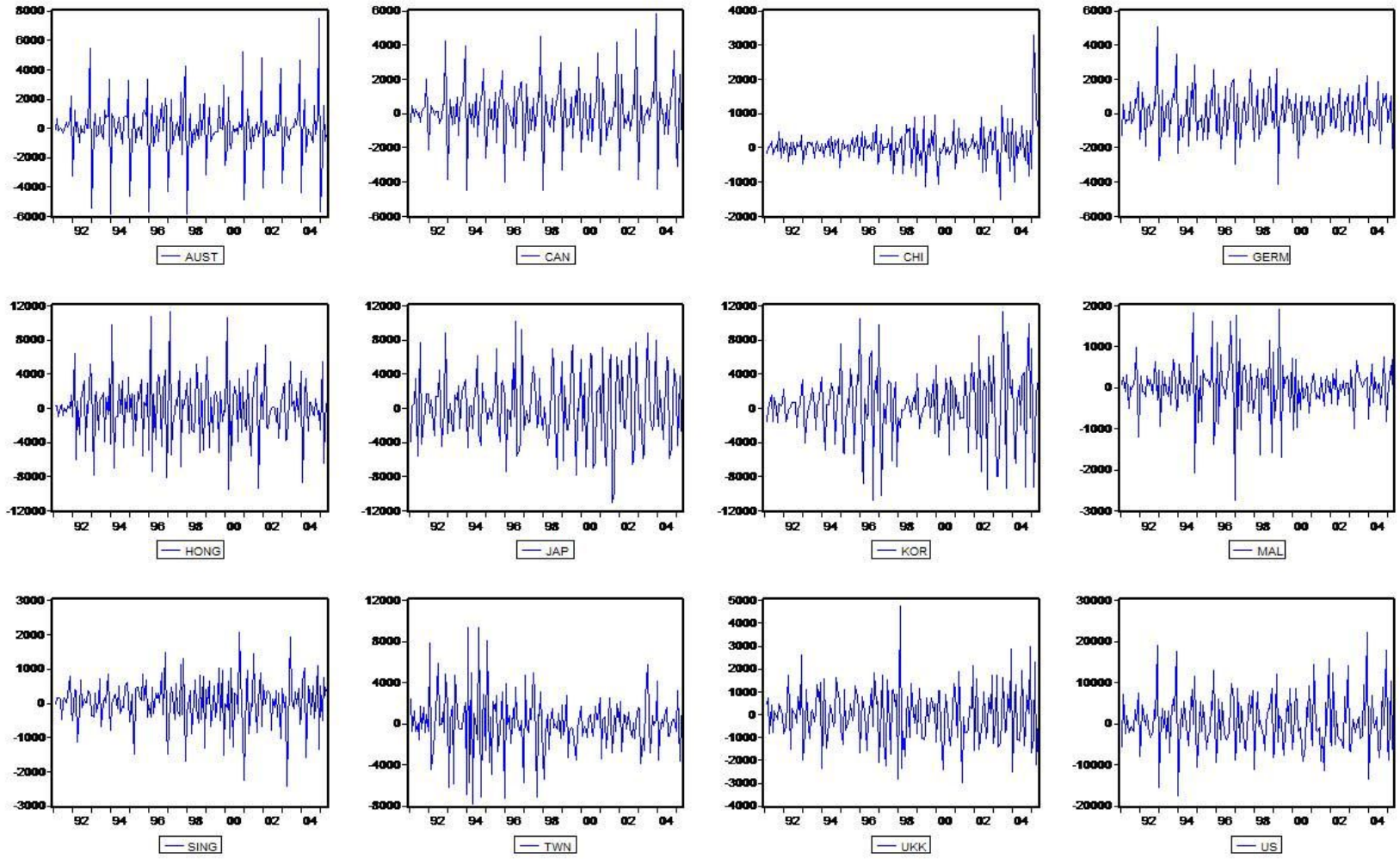
- Stationary time series has the properties as mean, variance and covariance are constant (same value) across time.
- Normally, time series data is a non-stationary.
- In statistics, using non-stationary time series for analysis can cause spurious regression and assumptions for analysis not being valid.

Examples of Non-Stationary Time Series



Source: https://drsifu.wordpress.com/2012/11/27/time-series-econometrics/

Examples of Stationary Time Series



Source: https://drsifu.wordpress.com/2012/11/27/time-series-econometrics/

What is Spurious regression?

- Spurious regression is a problem that arises when regression analysis indicates a strong relationship between two or more variables but in fact they are totally unrelated.
- Regression characteristics expected to be Spurious Regression.
  - R^2 is typically very high.
  - t-statistic value most often is significant.
  - Durbin-Watson statistic (DW) is low.
  - R^2 of the regression is greater than the Durbin-Watson Statistic.

Unit Root Test

- Therefore, it needs to check whether Time series is stationary.
- Hypothesis:
  - Null Hypothesis (H_0): time series has a unit root, meaning it is non-stationary.
  - Alternate Hypothesis (H_1): time series does not have a unit root, meaning it is stationary.
- Unit Root Test is a test for checking stationary of data that are various methods:
  - Dickey Fuller (DF)
  - Augmented Dickey and Fuller (ADF)
  - Etc.

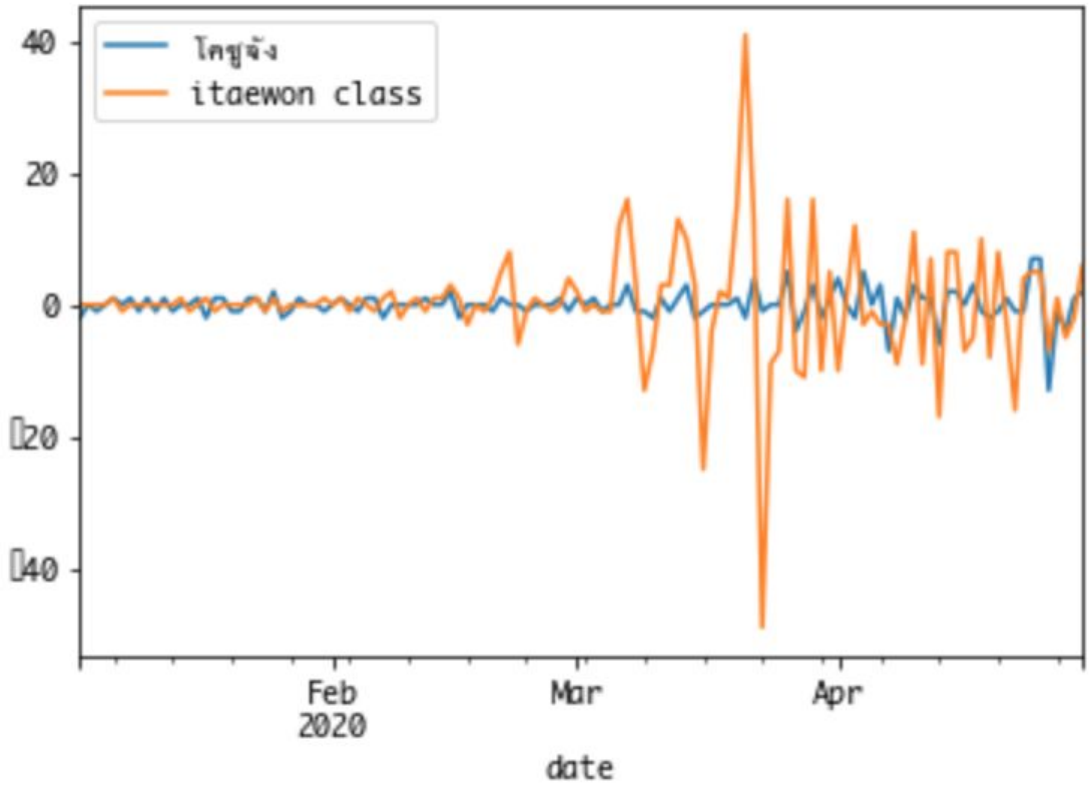
```
Augmented Dickey-Fuller Test on "โคชูจัง"
-----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Observation             = 107
Significance Level      = 0.05
Test Statistic          = -0.4375
No. Lags Chosen         = 13
Critical value 1%       = -3.493
Critical value 5%       = -2.889
Critical value 10%      = -2.581
=> P-Value = 0.9036. Weak evidence to reject the Null Hypothesis.
=> "โคชูจัง" is Non-Stationary.
```

```
Augmented Dickey-Fuller Test on "itaewon class"
-----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Observation             = 113
Significance Level      = 0.05
Test Statistic          = -1.3504
No. Lags Chosen         = 7
Critical value 1%       = -3.49
Critical value 5%       = -2.887
Critical value 10%      = -2.581
=> P-Value = 0.6058. Weak evidence to reject the Null Hypothesis.
=> "itaewon class" is Non-Stationary.
```

- p-value > 0.05: Fail to reject the null hypothesis (H_0).
- They are not significant.
- So, "โคชูจัง" and "itaewon class" have a unit root.

Unit Root problem

- Unit Root problem can solve by taking different.
- After taking different at level(1), can plot graph as following graph.



```
Augmented Dickey-Fuller Test on "โคชูจัง"
-----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Observation             = 107
Significance Level      = 0.05
Test Statistic          = -2.8133
No. Lags Chosen         = 12
Critical value 1%       = -3.493
Critical value 5%       = -2.889
Critical value 10%      = -2.581
=> P-Value = 0.0564. Weak evidence to reject the Null Hypothesis.
=> "โคชูจัง" is Non-Stationary.

Augmented Dickey-Fuller Test on "itaewon class"
-----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Observation             = 113
Significance Level      = 0.05
Test Statistic          = -4.5131
No. Lags Chosen         = 6
Critical value 1%       = -3.49
Critical value 5%       = -2.887
Critical value 10%      = -2.581
=> P-Value = 0.0002. Rejecting Null Hypothesis.
=> Series is Stationary.
```

Different at level(1)

- p-value > 0.05: Fail to reject the null hypothesis (H_0).
- "โคชูจัง" is not significant.
- But "itaewon class" is significant.
- So, "โคชูจัง" still has a unit root.

```
Augmented Dickey-Fuller Test on "โคชูจัง"
-----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Observation             = 107
Significance Level      = 0.05
Test Statistic          = -2.9325
No. Lags Chosen         = 11
Critical value 1%       = -3.493
Critical value 5%       = -2.889
Critical value 10%      = -2.581
=> P-Value = 0.0417. Rejecting Null Hypothesis.
=> Series is Stationary.
```

Different at level(2)

- p-value > 0.05: Fail to reject the null hypothesis (H_0).
- "โคชูจัง" is significant.
- So, "โคชูจัง" has no unit root now.

Lag length

- A time lag is a delay between an economic action and a consequence.
- Very often, the dependent variable responds to an independent variable with a lapse of time.
- For Granger causality test, it needs to define an optimal lag for testing.
- The optimal lag is selected from considering p-value of following criterias:
  - AIC: Akaike information criterion
  - BIC: Bayesian information criterion
  - FPE: Final prediction error criterion
  - HQIC: Hannan-Quinn information criterion

Lag length: criterion

$$AIC = n\left[\log\left(\frac{SS_{error(k)}}{n}\right) + \frac{2p_k}{n}\right]$$

$$BIC = n\left[\log\left(\frac{SS_{error(k)}}{n}\right) + \frac{p_k \log(n)}{n}\right]$$

$$HQ = n\left[\log\left(\frac{SS_{error(k)}}{n}\right) + \frac{2p_k \log(\log n)}{n}\right]$$

$$FPE = \left(\frac{SS_{error(k)}}{n - p_k}\right) \times \left(1 + \frac{p_k}{n}\right)$$

Note:

$SS_{error(k)} = sum\ of\ squared\ errors\ for\ k^{th} model\ in\ a\ set\ of\ models$

$p_k = number\ of\ coefficients\ in\ the\ k^{th} model\ plus\ 1$

VAR Order Selection (* highlights the minimums)

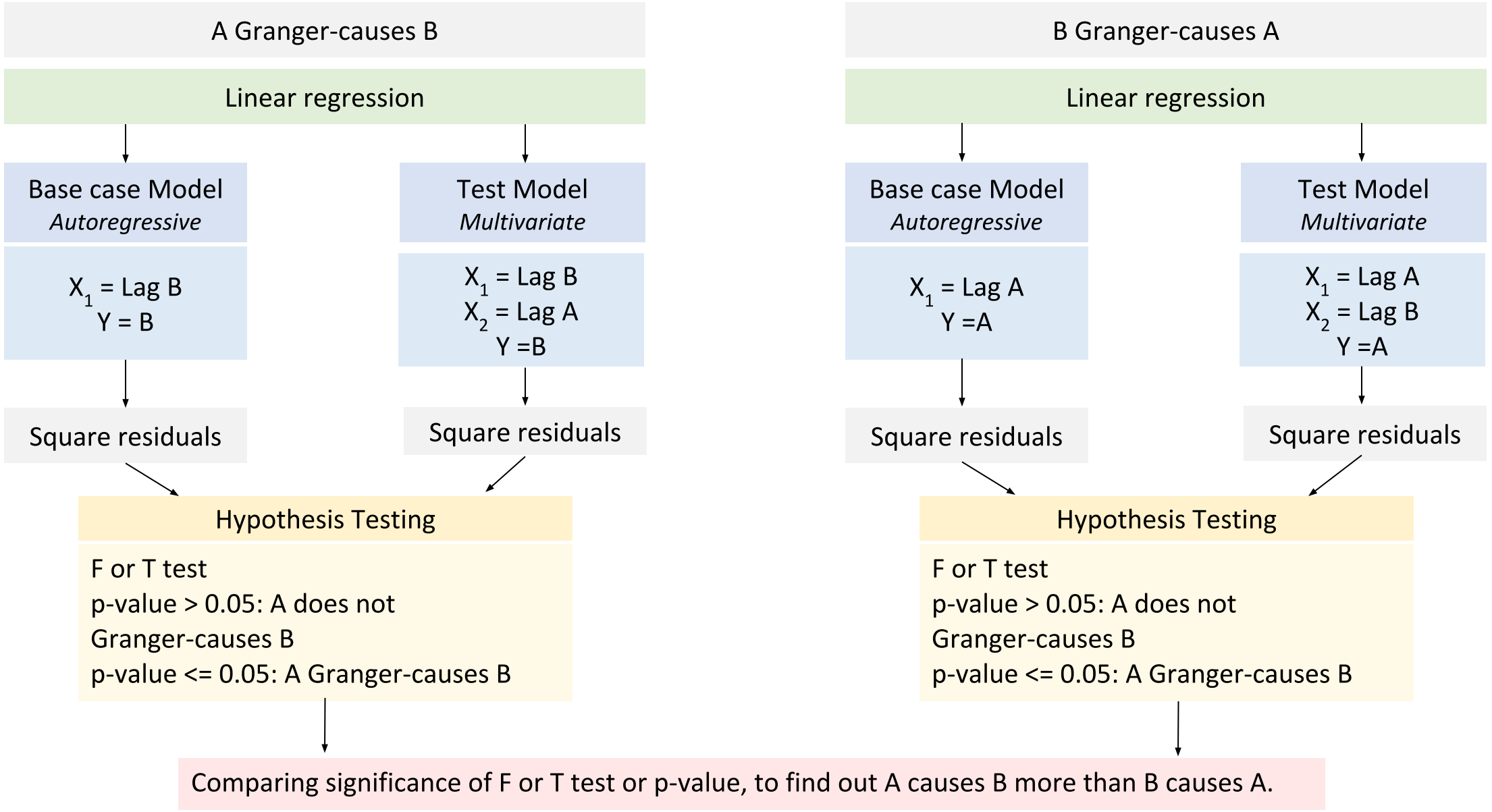
	AIC	BIC	FPE	HQIC
0	5.783	5.833	324.6	5.803
1	5.822	5.973	337.6	5.883
2	5.668	5.920	289.6	5.770
3	5.680	6.032	293.2	5.823
4	5.362	5.814	213.3	5.545
5	5.218	5.771*	184.9	5.442
6	5.274	5.927	195.6	5.539
7	5.129	5.883	169.5	5.434*
8	5.114	5.969	167.3	5.461
9	5.168	6.123	177.0	5.555
10	5.213	6.268	185.5	5.640
11	5.108	6.264	167.6	5.576
12	5.065*	6.322	161.3*	5.574
13	5.084	6.441	165.1	5.634

Granger causality

- Granger causality is a statistical concept of causality that is based on a prediction.
- Granger's Causality test on time-series data gives evidence that variable A Granger-causes B.
- If A Granger-causes B, then past values of A should contain information that helps predict B above and beyond the information contained in past values of B alone.
- Type of causality
  - Unidirectional Causality
    - $A \rightarrow B$
  - Bi-directional Causality
    - $A \leftrightarrow B$
  - No directional Causality

Granger causality - Step

1. Develop a base case, autoregressive model, using a dependent variable and its lagged values as an independent variable.
2. Develop a test case, multivariate model, by adding a second lagged independent variable that you want to test.
3. Calculate the R-Square (the square of resident error) for two models and run F-test and t-test to check if the residuals are significantly lower when you added tested the second variable.
4. Redo step 1 to 3, but reverse the direction. By comparing the tests significance or p-value, you can see if A Granger-causes B more than B Granger-causes A.


A Granger-causes B
Linear regression
Base case Model
Autoregressive
Test Model
Multivariate
X1 = Lag B
Y = B
X1 = Lag B
X2 = Lag A
Y =B
Square residuals
Square residuals
Hypothesis Testing
F or T test
p-value > 0.05: A does not Granger-causes B
p-value <= 0.05: A Granger-causes B
B Granger-causes A
Linear regression
Base case Model
Autoregressive
Test Model
Multivariate
X1 = Lag A
Y =A
X1 = Lag A
X2 = Lag B
Y =A
Square residuals
Square residuals
Hypothesis Testing
F or T test
p-value > 0.05: A does not Granger-causes B
p-value <= 0.05: A Granger-causes B
Comparing significance of F or T test or p-value, to find out A causes B more than B causes A.

Granger causality - Interpret result

- Hypothesis:
  - H_0: $\beta_1 = \beta_2 = \ldots = \beta_m = 0$: no relation
  - H_1: at least one non-zero : have relation
- p-value > 0.05: Accept the null hypothesis (H_0), A does not Granger-causes B.
- p-value <= 0.05: Reject the null hypothesis (H_0), A Granger-causes B.

Granger causality results

	itaewon class_x	โคชูจัง_x	test
itaewon class_y	1.0000	0.3401	ssr_ftest
โคชูจัง_y	0.0052	1.0000	ssr_ftest
itaewon class_y	1.0000	0.2695	ssr_chi2test
โคชูจัง_y	0.0001	1.0000	ssr_chi2test
itaewon class_y	1.0000	0.2890	lrtest
โคชูจัง_y	0.0007	1.0000	lrtest
itaewon class_y	0.0000	0.3401	params_ftest
โคชูจัง_y	0.0052	0.0000	params_ftest

- p-value <= 0.05: Reject the Null hypothesis (H_0)
- "itaewon class" caused "โคชูจัง"
- But "โคชูจัง" didn’t caused "itaewon class"
- So,"itaewon class" and "โคชูจัง" are Unidirectional Causality

Reference

- Thurman, W. N., & Fisher, M. E. (1988). Chickens, eggs, and causality, or which came first. American journal of agricultural economics, 70(2), 237-238.
- Maitra, S., (2019). Time Series Forecasting using Granger's Causality and Vector Auto-regressive Model. Retrieved from https://towardsdatascience.com/granger-causality-and-vector-auto-regressive-model-for-time-series-forecasting-3226a64889a6