

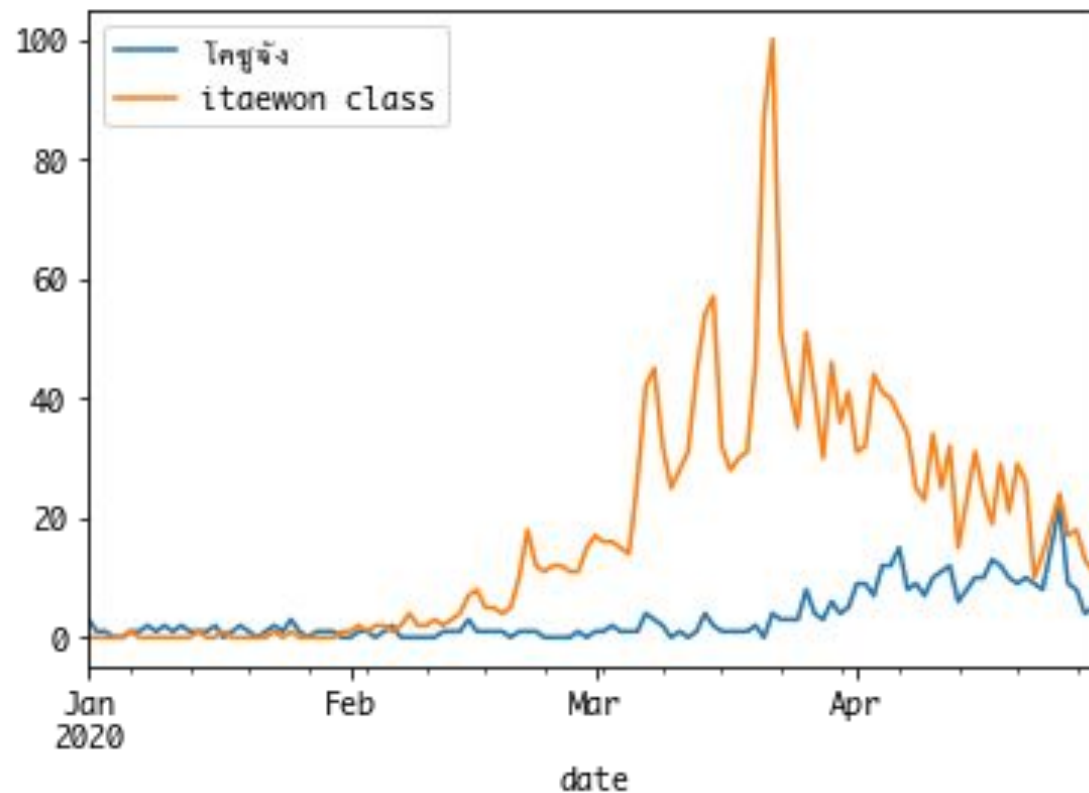


Case study: Did itaewon class cause โคชูจัง?

Granger Causality with Google
Trend

Time Series Data

- 'itaewon class' and 'โคชจิ้ง' 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.

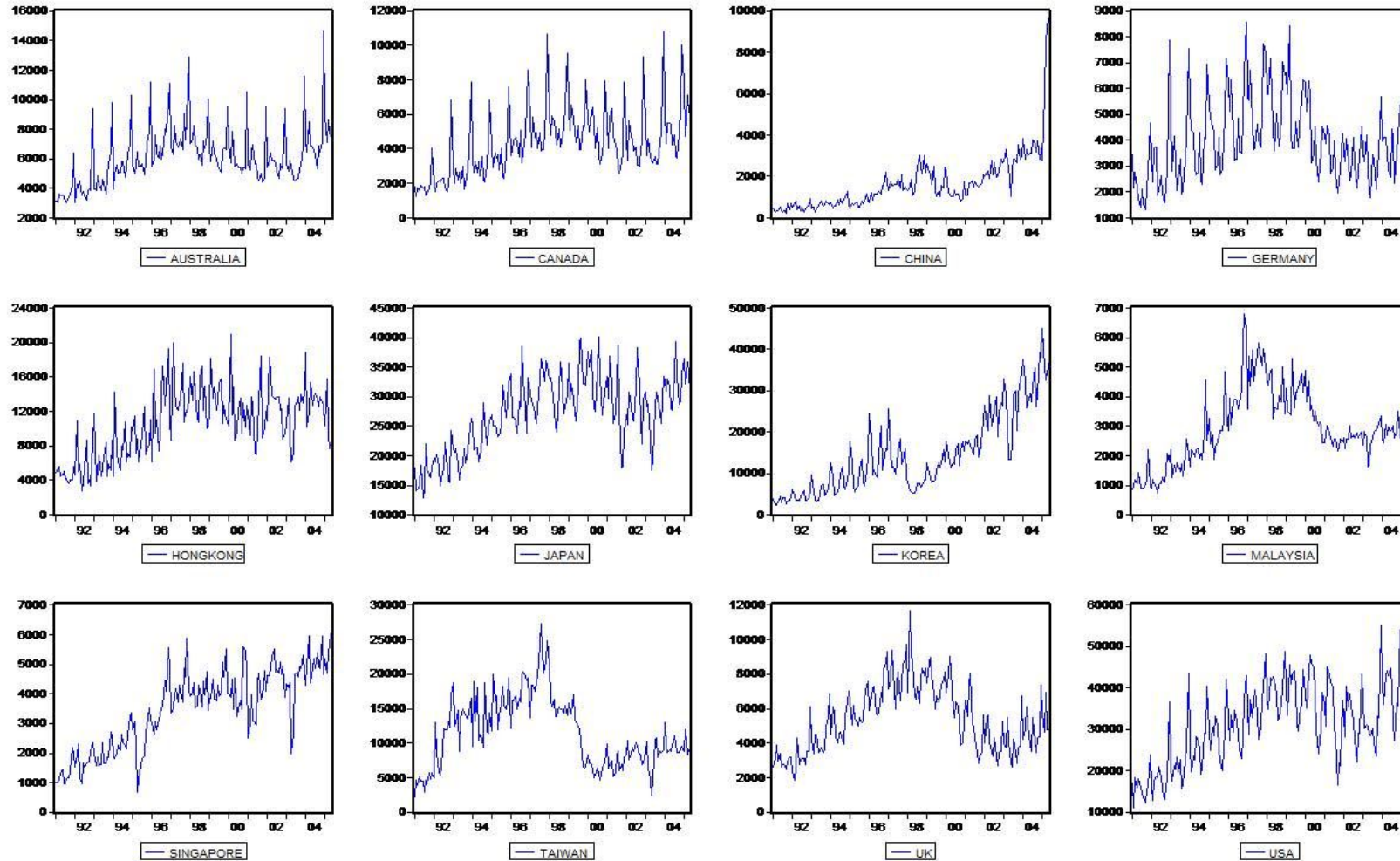
Mean : $E(Y_t) = \mu$

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

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

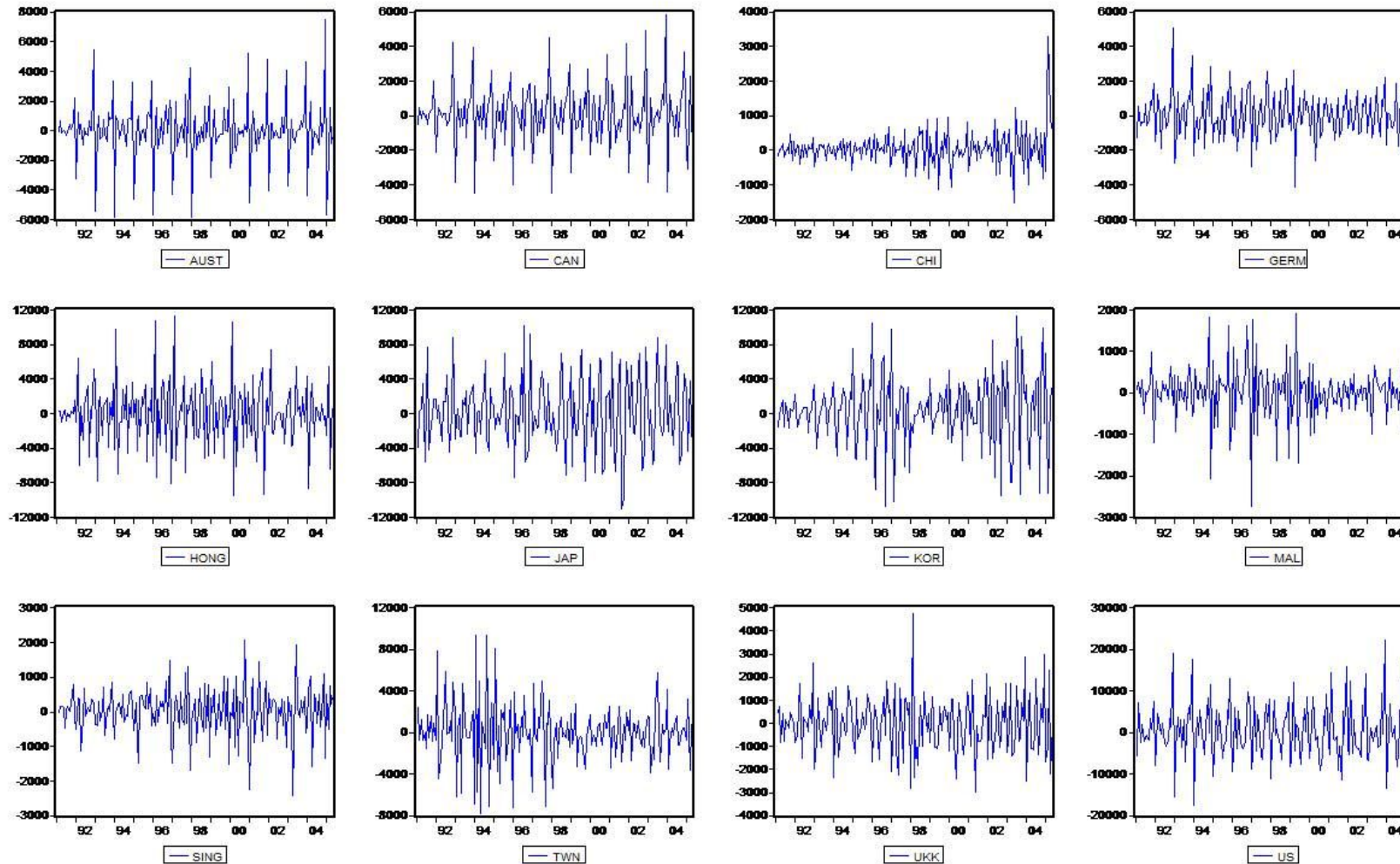
- Stationary time series has the properties as mean, variance and covariance are constant (same value) across time.
- Normally, time series data is a non-stationary.
- In statistics, using non-stationary time series for analysis can cause spurious regression and assumptions for analysis not being valid.

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 "โคชู้จ้ง"

Null Hypothesis: Data has unit root. Non-Stationary.

Significance Level = 0.05

Test Statistic = -0.3921

No. Lags Chosen = 5

Critical value 1% = -3.489

Critical value 5% = -2.887

Critical value 10% = -2.58

=> P-Value = 0.9114. Weak evidence to reject the Null Hypothesis.

=> "โคชู้จ้ง" is Non-Stationary.

Augmented Dickey-Fuller Test on "itaewon class"

Null Hypothesis: Data has unit root. Non-Stationary.

Significance Level = 0.05

Test Statistic = -1.2427

No. Lags Chosen = 7

Critical value 1% = -3.49

Critical value 5% = -2.887

Critical value 10% = -2.581

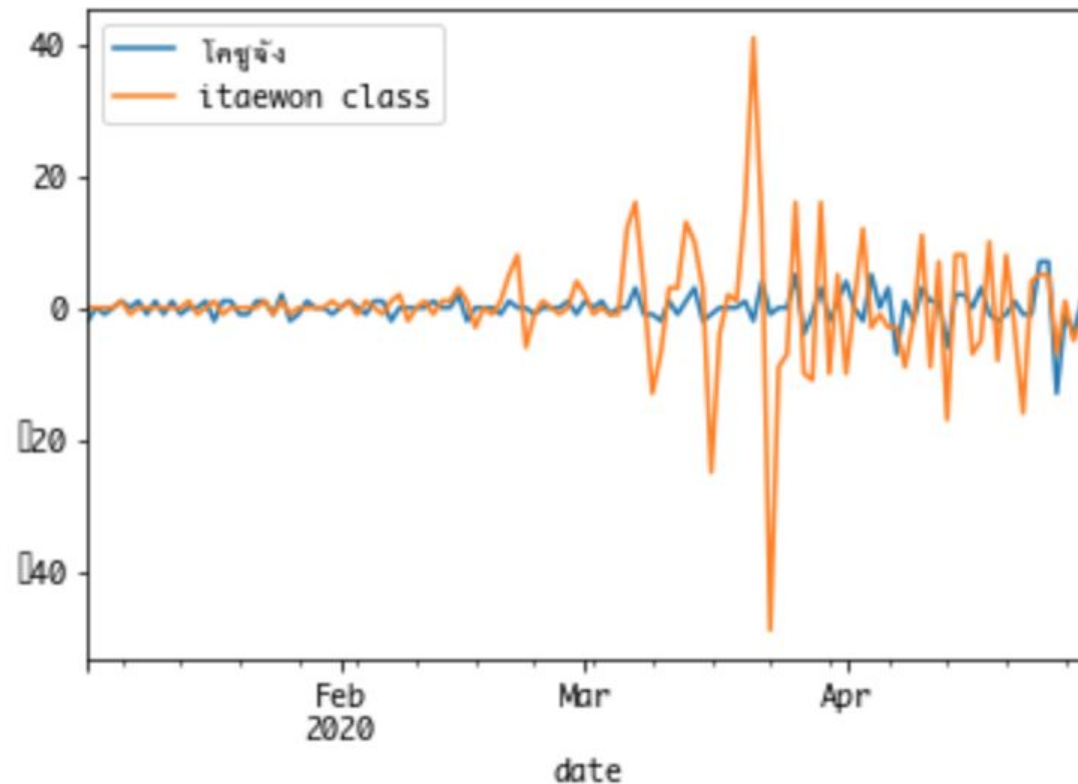
=> P-Value = 0.6549. Weak evidence to reject the Null Hypothesis.

=> "itaewon class" is Non-Stationary.

- p-value > 0.05: Fail to reject the null hypothesis (H_0).
- So, "โคชู้จ้ง" and "itaewon class" have a unit root. They are non-stationary.

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 "โคชู้จ้ง"

Null Hypothesis: Data has unit root. Non-Stationary.

Significance Level = 0.05

Test Statistic = -8.9465

No. Lags Chosen = 4

Critical value 1% = -3.489

Critical value 5% = -2.887

Critical value 10% = -2.58

=> P-Value = 0.0. Rejecting Null Hypothesis.

=> Series is Stationary.

Augmented Dickey-Fuller Test on "itaewon class"

Null Hypothesis: Data has unit root. Non-Stationary.

Significance Level = 0.05

Test Statistic = -5.4473

No. Lags Chosen = 6

Critical value 1% = -3.49

Critical value 5% = -2.887

Critical value 10% = -2.581

=> P-Value = 0.0. Rejecting Null Hypothesis.

=> Series is Stationary.

- p-value ≤ 0.05 : Reject the Null hypothesis (H_0).
- So, "โคชู้จ้ง" and "itaewon class" have no unit root. They are Stationary.



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	5.692	5.742	296.4	5.712
1	5.505	5.655	245.9	5.566
2	5.360	5.610	212.8	5.461
3	5.350	5.700	210.8	5.492
4	5.271	5.720	194.7	5.453
5	5.031	5.580*	153.3	5.254*
6	5.042	5.692	155.2	5.305
7	5.029	5.778	153.3	5.333
8	4.993	5.842	148.1	5.337
9	5.029	5.979	154.0	5.414
10	5.032	6.081	154.9	5.458
11	4.979*	6.128	147.3*	5.445
12	5.012	6.261	152.9	5.518
13	4.994	6.343	150.8	5.541



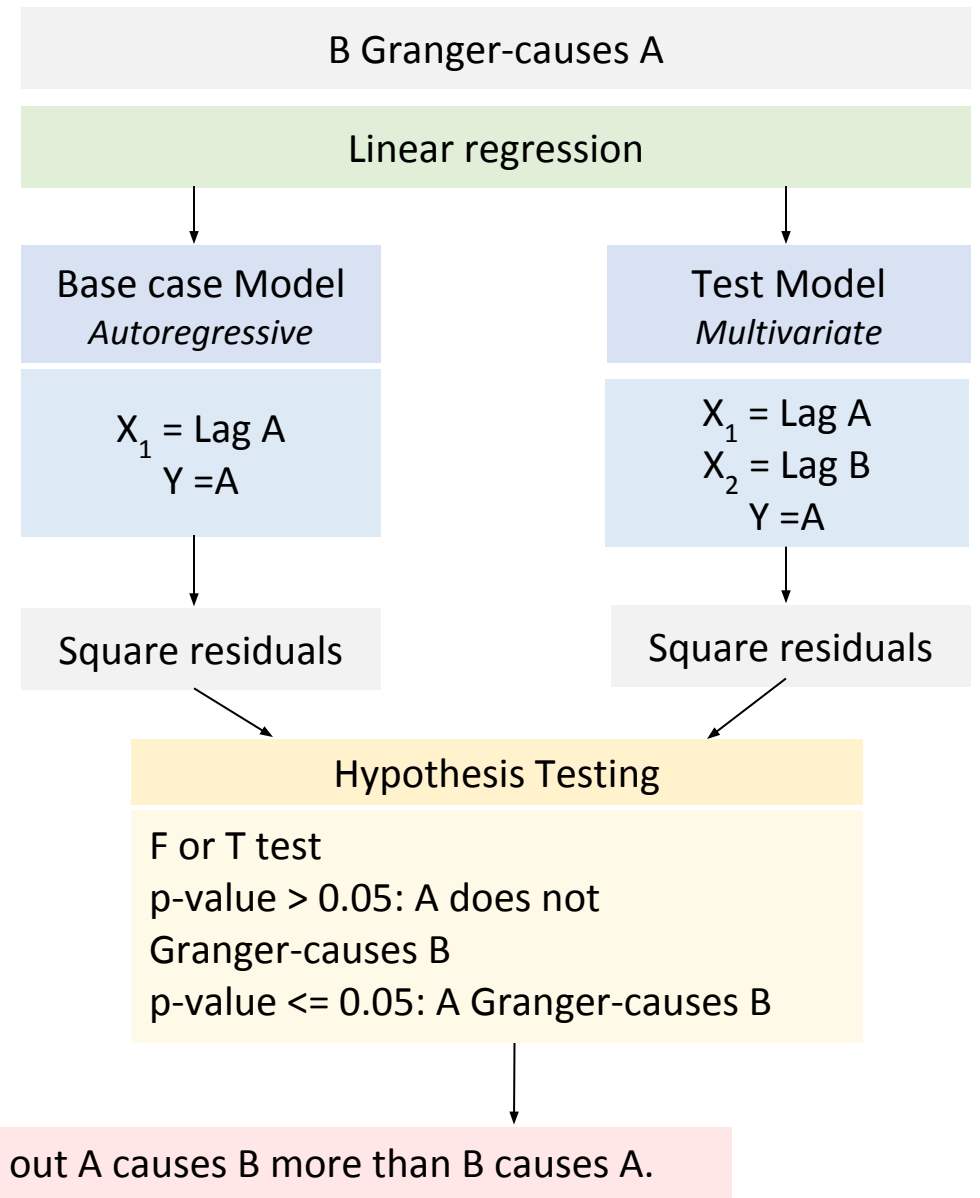
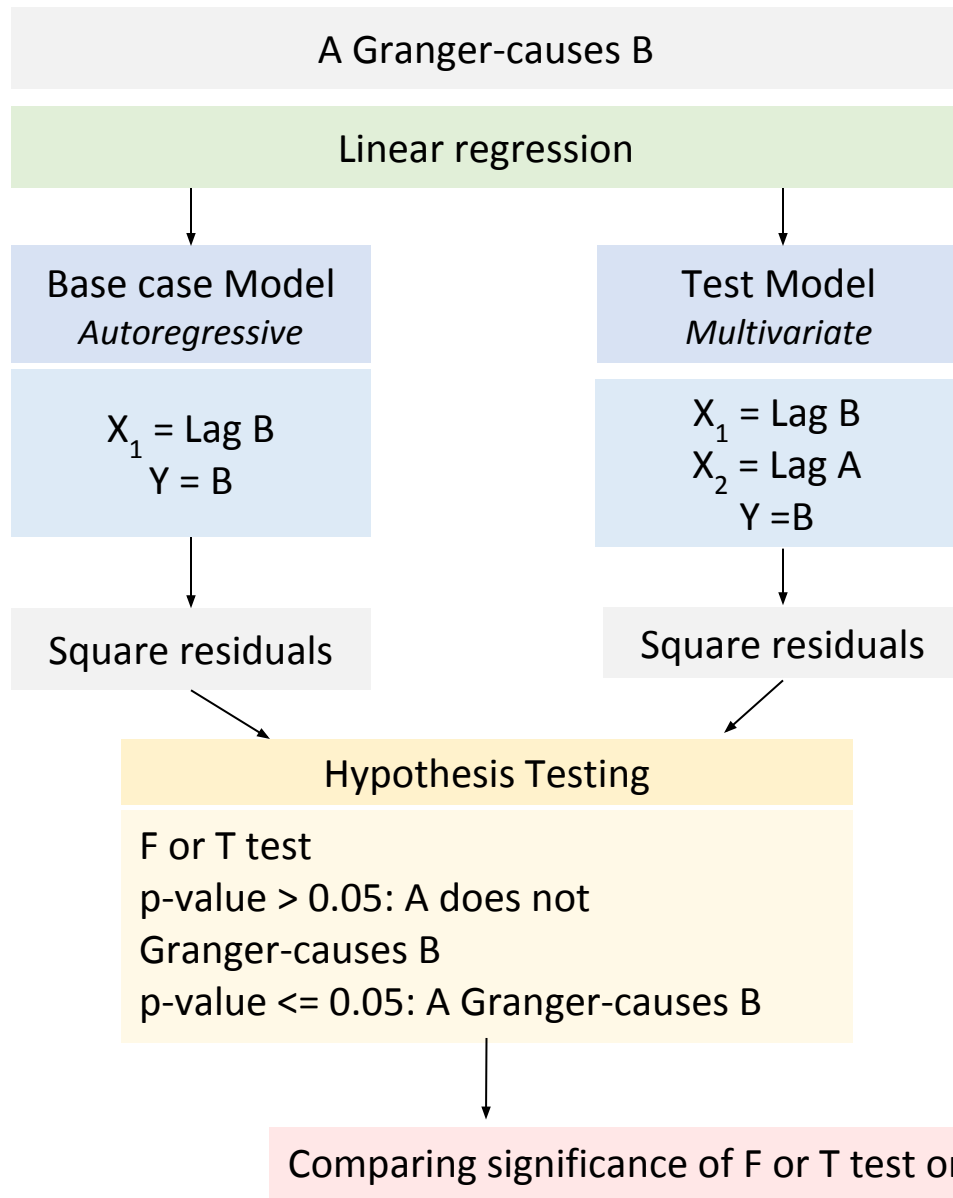
Granger causality


- 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.





Granger causality - Interpret result

- Hypothesis:
 - $H_0: \beta_1 = \beta_2 = \dots = \beta_m = 0$: 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 \leq 0.05: Reject the null hypothesis (H_0), A Granger-causes B.

Granger causality results

	โคชจ้ง_x	itaewon class_x	test
โคชจ้ง_y	1.0000	0.0100	ssr_ftest
itaewon class_y	0.4093	1.0000	ssr_ftest
โคชจ้ง_y	1.0000	0.0002	ssr_chi2test
itaewon class_y	0.3220	1.0000	ssr_chi2test
โคชจ้ง_y	1.0000	0.0013	lrtest
itaewon class_y	0.3415	1.0000	lrtest
โคชจ้ง_y	0.0000	0.0100	params_ftest
itaewon class_y	0.4093	0.0000	params_ftest

- p-value ≤ 0.05 : Reject the Null hypothesis (H_0)
- "itaewon class" causes "โคชจ้ง" but "โคชจ้ง" doesn't causes "itaewon class"
- So, "itaewon class" and "โคชจ้ง" 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>