11/12/2024

Cointegration: An Applied Perspectives

An Overview on Pakistan Data

Satyaki Basak
INDIAN INSTITUTE OF FOREIGN TRADE

Applied Econometrics Assignment

Name- Satyaki Basak
Roll- 48A

What is Stationarity of a time series?
The stationarity of a time series refers to the property that its statistical characteristics, such as mean, variance, and autocovariance, are constant over time.

What do *I(0)* and *I(1)* mean?
If x_t is *stationary*, then it is integrated of order zero written as *I(0)*.If not, then x_t *is non-stationary*. If we take the first order difference i.e. $x_t - x_{t-1} = \Delta x_t$. If Δx_t is *stationary* then x_t is called integrated of order 1 or *I(1)*. After differencing Δx_t becomes I(0)

Problem Statement : Collect at least 30 years data on Aggregate consumption and Aggregate income for a country and perform unit root test. Subsequently, check for cointegration using the 2 step Granger ,Johansen test and Impulse response.

Data Source : World Development Indicators, World Bank Databank. Data for per capita GDP and final consumption expenditure were calculated for the years 1993-2023 for Pakistan.

Results
Consumption and Income variables were both non-stationary as was observed by the plot of the values of consumption and income, as shown in figure 1.

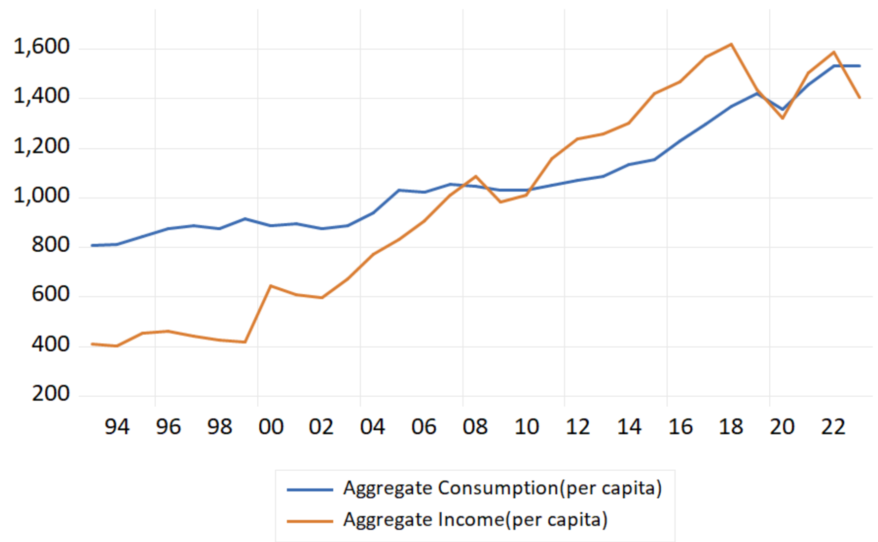

Fig 1. Line Plot of Consumption and Income series

Augmented Dickey Fuller test was performed on Consumption and Income variables to check stationarity of the series. The results are tabulated in Table 1.

Table 1. ADF results for Consumption and Income variables

Series	ADF statistics value	p-value	1%	5%	10%
Income	-0.82	0.80	-3.69	-2.97	-2.63
Consumption	1.17	0.99	-3.67	-2.96	-2.62

For ADF test null hypothesis$(\boldsymbol{H_0})$ **:** $(\boldsymbol{\rho - 1}) = \boldsymbol{0}$
The alternative hypothesis $(\boldsymbol{H_1})$ **:** $(\boldsymbol{\rho - 1}) \neq \boldsymbol{0}$
Clearly, the calculated values of ADF for both Consumption and Income are smaller than the critical values. Besides, we also have high p-values (> 0.05). Thus, the null hypothesis cannot be rejected and we can conclude that both the series are **non-stationary**.

Correlogram of Individual Series

A correlogram is visual representation that shows autocorrelation at different lags. ACF shows correlation with its own lags and PACF shows the correlation with lagged values while controlling for the effects of intermediate lags.

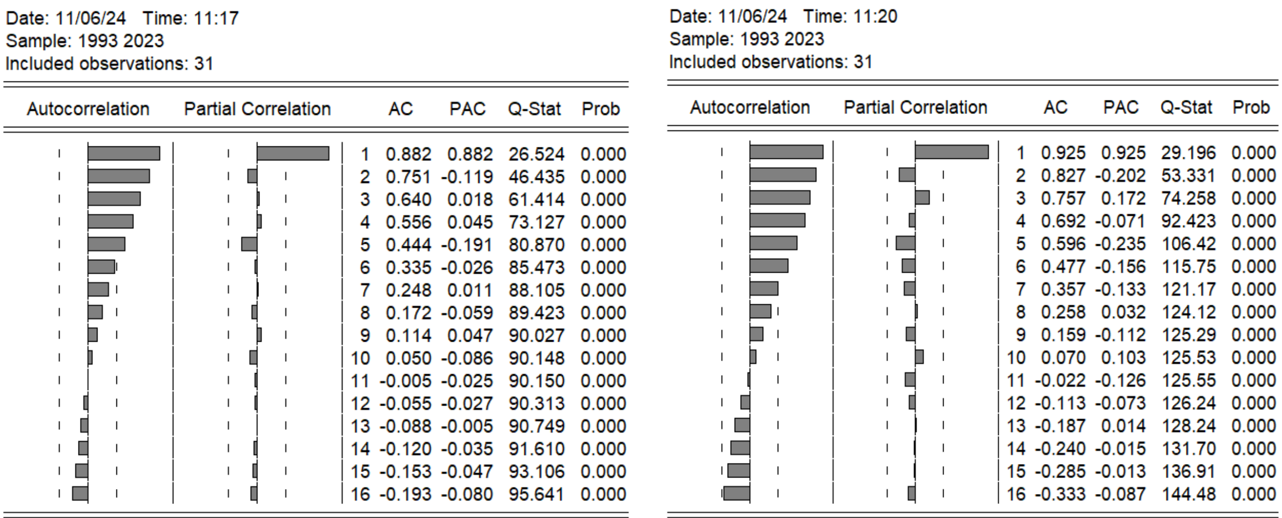

Fig 2. Correlogram of Individual series

What is cointegration?

When x_t and y_t are both $I(1)$ and if there is a θ, such that $y_t - \theta x_t$ is $I(0)$, then x_t and y_t are cointegrated. *Cointegration* is the long run equilibrium relationship between two variables.

Cointegration test was performed on Consumption and income series to check the long run relationship. The result is tabulated in table 2.

Table 2: Cointegration test result for Income and Consumption

Series	t-statistics value	p-value
Aggregate Consumption	-0.09	0.98
Aggregate Income	-0.83	0.90

Clearly, the calculated values of t-statistics for both Consumption and Income are smaller than the critical values. Besides, we also have high p-values (> 0.05). Thus, t we can conclude that both the series are **not cointegrated**.

The Granger Two Step Procedure

Step 1. Check for the level of Integration of Consumption, C_t and Income, y_t. If any one of them is not *I(1)*, then do not proceed to cointegration, else proceed to *Step 2*.

Step 2. Run OLS on the regression equation : $C_t = \beta_0 + \beta_1 y_t. + \epsilon_t$.

Step 3. Save the series of residuals and check if it is *I(0)*, using the Augmented Dickey Fuller (ADF) test.

Step 4. If ϵ_t is *I(0)*, then C_t and y_t are cointegrated or they have a long run equilibrium relationship.

Results: ***Step 1***

Now, to check if they are both I(1), we take the first order difference of both the series and perform ADF test (Table 2.) as well as visualize the series (Figure 2.).

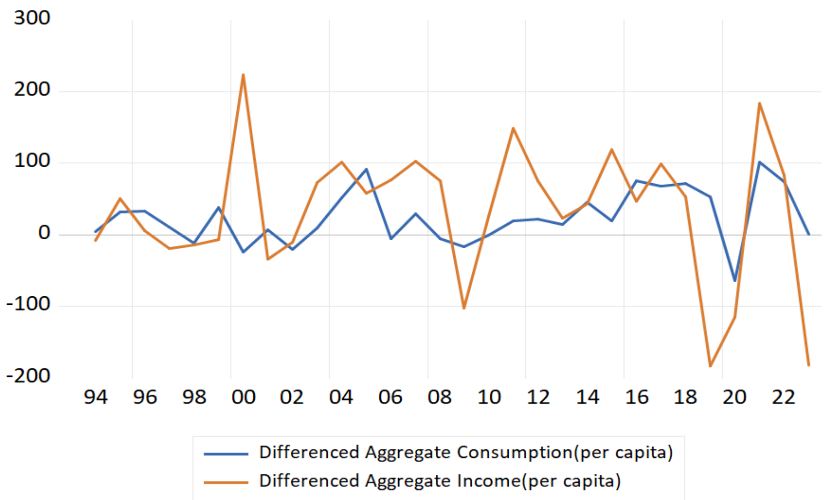

Fig 3: Line plot of differenced series

Though our data doesn't look horizontal, some degree of variance is observed for the Income & Consumption series. But data varies and fluctuate over a horizontal range. However, table 3 suggests that both the series are stationary, i.e. ***C_t* and *y_t* are both *I(1)***.

Table 3: ADF results for first differenced Consumption and Income variables

Series	ADF statistics value	p-value	1%	5%	10%
Differenced Income	-3.18	0.02	-3.69	-2.98	-2.63
Differenced Consumption	-5.02	0.00	-3.67	-2.97	-2.62

Now we can see that p-values of differenced series <0.05 hence we can say that they are stationary.

Step 2:
Next, we regress Ct on yt using Ordinary Least Square (OLS). The results are shown in figure 3.

Dependent Variable: AGGREGATE_CONSUMPTION_PER_CAPITA_
Method: Least Squares
Date: 11/06/24 Time: 13:45
Sample: 1993 2023
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	615.3656	43.95610	13.99955	0.0000
AGGREGATE_INCOME_PER_CAPITA_	0.471526	0.041224	11.43824	0.0000

R-squared	0.818561	Mean dependent var	1078.782
Adjusted R-squared	0.812304	S.D. dependent var	219.1176
S.E. of regression	94.93019	Akaike info criterion	12.00650
Sum squared resid	261340.5	Schwarz criterion	12.09902
Log likelihood	-184.1008	Hannan-Quinn criter.	12.03666
F-statistic	130.8332	Durbin-Watson stat	0.265336
Prob(F-statistic)	0.000000		

Table 4. OLS results obtained after regressing C_t on Y_t

Step 3:
We have regressed Consumption on Income variable and from that regression equation saved the residual and plot it to check stationarity visually. Then perform ADF test on residual.

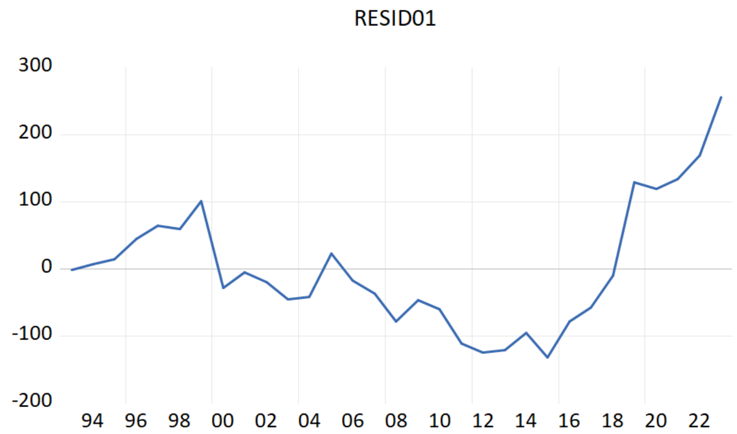

Figure 4. Residual plot

We perform ADF test on the residuals. Results are summarised in table 4. and visualized in figure 4. As can be observed from the table, the p-value is low (<0.05)

Table 5: ADF result for residuals

Series	ADF Statistics value	p-value	1%	5%	10%
Resid01	0.012	0.95	-3.67	-2.96	-2.62

Step 4:
From the above result we can conclude that resid01 is non-stationary series, i.e. resid01 is not I(0). Therefore, we can say that both consumption and income are not sharing any stable long run equilibrium relationship between them.

Error Correction Model
The Error Correction Model (ECM) is a type of time series model used to understand the short-term and long-term dynamics of two or more cointegrated variables. It captures how deviations from this long-term equilibrium are corrected over time through short-term adjustments. The general ECM equation is written as

$$\Delta y_t = a + b\Delta x_t + c(y_{t-1} - \alpha - \beta x_{t-1}) + \varepsilon_t$$

y_t, x_t are two non-stationary series i.e. I(1)

where ,
Δy_t = change in y_t or first difference of $y_t \sim I(0)$
Δx_t= change in x_t or first difference of $x_t \sim I(0)$
$y_{t-1} - \alpha - \beta x_{t-1}$= error correction term
c = error correction coefficient that measures how quickly the short-run deviations are corrected (speed of adjustment). -1<c<0

Here series are not cointegrated in Long run so it's not required to perform ECM instead of having non cointegrated series . But we will do ECM to check the ECM coefficient.

Then we run OLS regression of differenced income and differenced for ECM where the model would be

$$\Delta cons_t = a + b\Delta inc_t + c(cons_{t-1} - \alpha - \beta inc_{t-1}) + \varepsilon_t$$

Dependent Variable: D(AGGREGATE_CONSUMPTION_PER_CAPITA_)
Method: Least Squares
Date: 11/12/24 Time: 20:06
Sample (adjusted): 1994 2023
Included observations: 30 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(AGGREGATE_INCOME_PER_CAPITA_)	0.117937	0.073921	1.595439	0.1223
RESID01(-1)	-0.070187	0.084685	-0.828797	0.4145
C	19.70672	7.124736	2.765958	0.0101

R-squared	0.123890	Mean dependent var	24.21431
Adjusted R-squared	0.058993	S.D. dependent var	37.81836
S.E. of regression	36.68590	Akaike info criterion	10.13730
Sum squared resid	36338.10	Schwarz criterion	10.27742
Log likelihood	-149.0595	Hannan-Quinn criter.	10.18213
F-statistic	1.909022	Durbin-Watson stat	1.857833
Prob(F-statistic)	0.167704		

Table 6: OLS for error correction model

From the above OLS result we got
a= 19.71
b=0.12
c=-0.07 (error correction coefficient)

Implication of c

Coefficient of error correction term implies that only **7%** of the deviation from the long-run equilibrium is corrected in each time period . The system is adjusting very slowly towards equilibrium after a short-term deviation. It suggests that if there is a shock that moves the system away from equilibrium, it will take a significant amount of time for the system to return to equilibrium.

Granger Causality Test

Granger causality tests whether the inclusion of past values of one variable x_t improves the prediction of another variable y_t.In mathematical terms, let's consider two time series, x_t and y_t . We want to test if x_t "Granger-causes" y_t , meaning that x_t provides useful information for forecasting y_t .

$$y_t = \sum_{i=1}^{k} \alpha_i y_{t-i} + \sum_{i=1}^{k} \beta_i x_{t-i} + u_t$$

Null hypothesis $(\boldsymbol{H_0})$: $\beta_i = 0 \quad \forall$ i=(1.....k)
Alternative hypothesis $(\boldsymbol{H_1})$: $\beta_1, \beta_2, \ldots . \beta_k \neq 0$ (at least one coefficient $\neq 0$)

F-statistic would be : $\frac{(RSS_1-RSS_2)/(v_2-v_1)}{RSS_2/(n-v_2)} \sim F_{(v_2-v_1, n-v_2)}$

We perform **Granger Causality test** to see that income Granger causes consumption or not. Here we take number of lags=4 and the test results tabulated as follows

Table7: p-values of tests are recorded as follows.

Test	Lag=1	Lag=2	Lag=3	Lag=4
Prob (F-stat)	0.1349	0.1115	0.0258	0.0540

Pairwise Granger Causality Tests
Date: 11/06/24 Time: 23:24
Sample: 1993 2023
Lags: 3

Null Hypothesis:	Obs	F-Statistic	Prob.
INCOME does not Granger Cause CONSUMPTION	28	3.78575	0.0258
CONSUMPTION does not Granger Cause INCOME		2.34396	0.1021

Interpretation: We got p-value of F-stat at lag=3 , p=0.0258<0.05 therefore we can reject the null hypothesis for that lag and it implies that the income at lag=3 **Granger-causes** the consumption. As they are not cointegrated in long run we got uni-directional causality .

Optimal Lag Length Determination :

We are required to find the optimal lag length for the system for this we construct a VAR and then view the lag structure .

VAR Lag Order Selection Criteria
Endogenous variables: CONSUMPTION INCOME
Exogenous variables: C
Date: 11/13/24 Time: 21:39
Sample: 1993 2023
Included observations: 23

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-303.1276	NA	1.14e+09	26.53283	26.63157	26.55766
1	-247.9388	95.98049	13374608	22.08163	22.37785*	22.15613
2	-245.5789	3.693772	15597106	22.22425	22.71794	22.34841
3	-237.2093	11.64457*	10944662*	21.84429*	22.53546	22.01812*
4	-236.3467	1.050182	15089192	22.11710	23.00575	22.34060
5	-233.7149	2.746220	18442867	22.23608	23.32220	22.50924
6	-231.1137	2.261951	23746675	22.35771	23.64131	22.68053
7	-229.8797	0.858435	37134989	22.59823	24.07931	22.97072
8	-226.5009	1.762811	54527680	22.65225	24.33081	23.07441

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Table 9: Lag Order selection criteria for both series

From the above table we can see that according to AIC optimal lag length is 3 and according to SIC lag length is 1. With 30 years of data which provide moderate sample size ,a balance between complexity and parsimony is often ideal . Given this we use lag length 1 according to SIC.

Vector Autoregression (VAR) Model

In a Vector Autoregression model with two time series consumption and income , the aim is to capture the interdependencies and dynamics between them over time. Here we regressed each variable on its own lagged (1 periods) values and on the lagged(1 periods) values of the other variable.

If the lagged coefficients of income significantly affect consumption ,this suggests income has predictive power over future consumption values. The sign of each coefficient will tell whether an increase in tends to raise or lower future consumption.

Vector Autoregression Estimates
Date: 11/13/24 Time: 21:53
Sample (adjusted): 1994 2023
Included observations: 30 after adjustments
Standard errors in ( ) & t-statistics in [ ]

	CONSUMPT	INCOME
CONSUMPTION(-1)	0.921051	-0.227960
	(0.08389)	(0.21392)
	[10.9798]	[-1.06561]
INCOME(-1)	0.063280	1.063003
	(0.04105)	(0.10470)
	[1.54136]	[10.1532]
C	46.88625	214.5713
	(55.3846)	(141.240)
	[0.84656]	[1.51919]

Table 8: Vector Autoregression model estimates

From the above table we can see that consumption (-1) has a negative impact on income.

Johansen Test for cointegration

Date: 11/06/24 Time: 23:07
Sample (adjusted): 1995 2023
Included observations: 29 after adjustments
Trend assumption: Quadratic deterministic trend
Series: CONSUMPTION INCOME
Lags interval (in first differences): 1 to 1

Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None	0.429668	16.35092	18.39771	0.0945
At most 1	0.002285	0.066330	3.841465	0.7967

Trace test indicates no cointegration at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.**
None	0.429668	16.28459	17.14769	0.0664
At most 1	0.002285	0.066330	3.841465	0.7967

Max-eigenvalue test indicates no cointegration at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Table 10: Johansen Cointegration test results

In general Johansen test is used to check how many cointegration relationship present in the system when there are more than 2 variables. Here we have checked for cointegration using Johansen system Test , we got p-value> 0.05 at **None** that means there is no cointegration relationship.

Vector Error Correction Model (VECM)

In a vector error correction model with two series income and consumption the aim is to capture both the short run dynamics and long run equilibrium relationship between the two variables assuming they are cointegrated.

As our model has no cointegrating relationship we don’t need to perform VECM test. But we still run VECM to interpret the result.

Vector Error Correction Estimates
Date: 11/12/24 Time: 21:30
Sample (adjusted): 1995 2023
Included observations: 29 after adjustments
Standard errors in ( ) & t-statistics in [ ]

Cointegrating Eq:	CointEq1
CONSUMPTION(-1)	1.000000
INCOME(-1)	0.362832 (0.35517) [1.02158]
C	-1430.808

Table 11: VECM Estimates table

We will be developing Johansen long run equation from VECM model

1*CONSUMPTION(-1) + 0.36*INCOME(-1) – 1430.81=0

- ⇨ CONSUMPTION(-1)= 1430.81 – 0.36*INCOME(-1)

Now we will be checking the long run adjustments and short run coefficient of the model . We consider consumption as the dependent variable.

Error Correction:	D(CONSUM	D(INCOME)
CointEq1	0.036079 (0.02236) [1.61322]	-0.051083 (0.05947) [-0.85894]
D(CONSUMPTION(-1))	-0.228175 (0.21182) [-1.07722]	-0.352611 (0.56327) [-0.62601]
D(INCOME(-1))	0.168521 (0.08416) [2.00244]	0.068167 (0.22379) [0.30460]
C	23.78292 (8.61455) [2.76079]	40.62748 (22.9078) [1.77352]

Table 12: Long run adjustment and short run dynamics parameters

So the equation which can be formed from the above result is:

D(CONSUMPTION)=23.78 + 0.036 ECT $_{t-1}$ – 0.228 D(CONSUMPTION $_{t-1}$) + D(INCOME $_{t-1}$)

The coefficient value of ECT is **0.036** This means that deviation from the long run relationship is corrected at the rate of 3.6% in the present period.

The Toda- Yamamoto Procedure

It is a technique used to test for Granger causality in VAR framework even if timeseries are non-stationary or have different order of integration. This method is useful because it bypasses the need for pretesting for cointegration , which can sometime complicate Granger causality test.

$$X_t = \alpha_1 + \sum_{i=1}^{k+d_{max}} \alpha_{2i} X_{t-i} + \sum_{i=1}^{k+d_{max}} \alpha_{3i} Y_{t-i} + \eta_1$$

$$Y_t = \beta_1 + \sum_{i=1}^{k+d_{max}} \beta_{2i} Y_{t-i} + \sum_{i=1}^{k+d_{max}} \beta_{3i} X_{t-i} + \eta_2$$

Step 1: We have identified the highest order of integration among the series. This is done by running unit root test such as ADF or KPSS test on each series, consumption and income.

Step 2: To determine the optimal lag length(p) for the VAR we use SIC criteria (as mentioned before)

Step 3: Estimated an augmented VAR model with p+d_max lags. In this case p is 1 and d_max is also 1 therefore p+d_max=2 , we use 2 period lag for both variables.

Step 4: Use wald test to check the statistical significance of the coefficients for the first p lags only.

VAR Granger Causality/Block Exogeneity Wald Tests
Date: 11/13/24 Time: 22:12
Sample: 1993 2023
Included observations: 29

Dependent variable: CONSUMPTION

Excluded	Chi-sq	df	Prob.
INCOME	4.501695	1	0.0339
All	4.501695	1	0.0339

Dependent variable: INCOME

Excluded	Chi-sq	df	Prob.
CONSUMPTION	0.703040	1	0.4018
All	0.703040	1	0.4018

Table 13: Toda Yamamoto wald test result

From the above result we can see that when consumption is dependent variable then income causes consumption as chi-sq value is significant. And when income is dependent variable ,consumption doesn't causes income as chi-sq value is insignificant. Therefore we got uni-directional causality. This result supports Granger Causality test result.

Impulse Response Function

The Impulse Response Function (IRF) describes how a dynamic system reacts over time to a sudden, one-time shock (or "impulse") in one of its variables. In economics and finance, it's commonly used to analyze how an unexpected change in one variable impacts other variables across future periods.

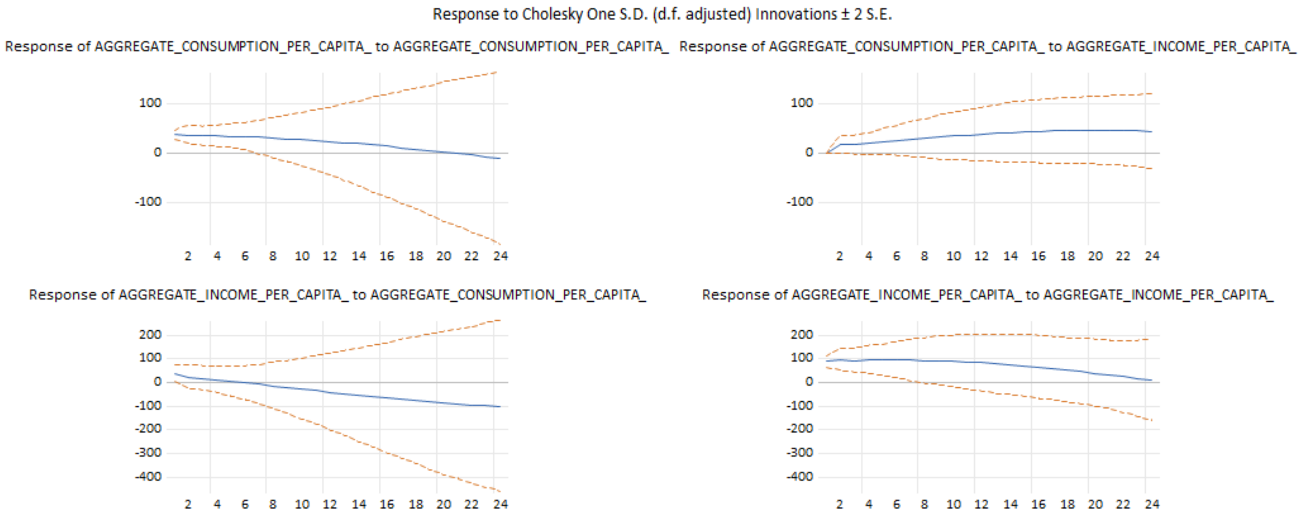

Fig 5: Impulse response to all variables

Impulse Response Function Analysis

- Magnitude of the shock: **One standard** deviation
- **Red Dots** are the standard error confidence bands . The confidence intervals are computed as +/-2 SE confidence bands.
- **X axis** represents the periods.
- **Y axis** shows the percentage variation.

From the above graphs , principle diagonal plots are "***own shocks***" and off-diagonal plots are "***shocks in other variable in the system***" . We can also interpret the plots as one standard deviation shock to income per capita increases consumption per capita over time (top right). Two "own shocks" gradually decreases its variable over time.

Interpretation:

Response of Income to Income

The response of Ct to Ct showing a positive range over the horizon. The magnitude reduces after 20th quarter showing a dampening effect.

Response of consumption to income

The response of Ct to Yt showing an effect on consumption to the one std shock to income. The magnitude is gradually increases over time but its showing some saturation over horizon.

Response of income to consumption

In response of income to consumption showing a decline in magnitude after 5th quarter. Therefore it's showing dampening effect over the horizon.

Response of income to income

The response of Yt to Yt is decreasing over time remaining above the horizontal indicates positive over time but showing slow gradual decreasing effect.
