Analysis of various testing methodologies and estimation procedures in Econometrics

Debarghya Jana[*] Shailza Sharma[†] Dasari Charithambika[‡]
Roll: 221306 Roll: 221416 Roll: 210302

Supervisor: Dr. Sharmishtha Mitra[§]

A Project Report Submitted in the Requirements of the course MTH676 for the Degree of

MASTER OF SCIENCE

in

STATISTICS

to

DEPARTMENT OF MATHEMATICS AND STATISTICS

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

[*]final year Student, M.Sc. Statistics, IIT Kanpur
[†]final year Student, M.Sc. Statistics, IIT Kanpur
[‡]3rd year Student, B.S. Statistics and Data Science, IIT Kanpur
[§]Professor, Department of Mathematics and Statistics, IIT Kanpur

ACKNOWLEDGEMENT

We want to extend a sincere and heartfelt gratitude towards all of the group members Shailza Sharma and Dasari Charithambika, without whom the successful completion of the project would have been a distant dream. We express our profound thankfulness and deep regards to Dr. Sharmishtha Mitra, Professor, IIT Kanpur for her guidance, valuable feedback and constant encouragement throughout the project.

Contents

1 Tests for Heteroscedasticity :

1.1 Dataset:

- Dataset Used : We have used Real Estate price prediction dataset for testing heteroscedasticity. The source of our dataset is "Kaggle".
- Data Description : It contains 1 response variable and 6 predictors. The total number of data points are 414.
  - Y : House price per unit area
  - X1 : Transaction date of the property
  - X2 : Age of the house
  - X3 : Distance to the nearest MRT (Mass Rapid Transit) station
  - X4 : Number of convenience stores nearby
  - X5 : Latitude of the location
  - X6 : Longitude of the location
- Following is the summary of the dataset :

```
>
> summary(Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.60   27.70   38.45   37.98   46.60  117.50
>
> summary(X)
       V1             V2               V3                V4              V5
 Min.   :2013   Min.   : 0.000   Min.   :  23.38   Min.   : 0.000   Min.   :24.93
 1st Qu.:2013   1st Qu.: 9.025   1st Qu.: 289.32   1st Qu.: 1.000   1st Qu.:24.96
 Median :2013   Median :16.100   Median : 492.23   Median : 4.000   Median :24.97
 Mean   :2013   Mean   :17.713   Mean   :1083.89   Mean   : 4.094   Mean   :24.97
 3rd Qu.:2013   3rd Qu.:28.150   3rd Qu.:1454.28   3rd Qu.: 6.000   3rd Qu.:24.98
 Max.   :2014   Max.   :43.800   Max.   :6488.02   Max.   :10.000   Max.   :25.01
       V6
 Min.   :121.5
 1st Qu.:121.5
 Median :121.5
 Mean   :121.5
 3rd Qu.:121.5
 Max.   :121.6
>
```

Figure 1: Summary of the dataset

1.2 Tests:

1.2.1 Glejser Test:

Step 1: Estimate original regression with ordinary least squares and find the sample residuals e_i. Step 2: Regress the absolute value e_j l on the explanatory variable that is associated with the heteroscedasticity.

$$|e_i| = \gamma_0 + \gamma_1 X_i + v_i$$
$$|e_i| = \gamma_0 + \gamma_1 \sqrt{X_i} + v_i$$
$$|e_i| = \gamma_0 + \gamma_1 \frac{1}{X_i} + v_i$$

Step 3: Select the equation with the highest R^2 and lowest standard errors to represent heteroscedasticity.
Step 4: Perform a t-test on the equation selected from step 3 on γ_1. If γ_1 is statistically significant, reject the null hypothesis of homoscedasticity.

```
> # Glejser test
> glejser(model)
# A tibble: 1 × 4
  statistic   p.value parameter alternative
      <dbl>     <dbl>     <dbl> <chr>
1      28.1 0.0000892         6 greater
>
```

Figure 2: Performance of Glejser test

The test statistic is 28.1 with a p-value of 0.0000892. This indicates evidence against the null hypothesis of homoscedasticity (equal variance of residuals) in favor of the alternative hypothesis of heteroscedasticity (unequal variance of residuals).Since the p-value is less than 0.05, we reject the null hypothesis of homoscedasticity in favor of heteroscedasticity.

1.2.2 Breusch Pagan Godfrey test:

Under the classical assumptions, ordinary least squares is the best linear unbiased estimator (BLUE), i.e., it is unbiased and efficient. It remains unbiased under heteroskedasticity, but efficiency is lost. Before deciding upon an estimation method, one may conduct the Breusch-Pagan test to examine the presence of heteroskedasticity. The Breusch-Pagan test is based on models of the type $\sigma_i^2 = h\left(z_i'\gamma\right)$ for the variances of the observations where $z_i = (1, z_{2i}, \ldots, z_{pi})$ explain the difference in the variances. The null hypothesis is equivalent to the $(p-1)$ parameter restrictions:

$$\gamma_2 = \cdots = \gamma_p = 0.$$

The following Lagrange multiplier (LM) yields the test statistic for the Breusch-Pagan test:

$$\mathrm{LM} = \left(\frac{\partial \ell}{\partial \theta}\right)^{\top} \left(-E\left[\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right]\right)^{-1} \left(\frac{\partial \ell}{\partial \theta}\right).$$

This test can be implemented via the following three-step procedure: - Step 1: Apply OLS in the model

$$y_i = X_i \beta + \varepsilon_i, \quad i = 1, \ldots, n$$

- Step 2: Compute the regression residuals, $\hat{\varepsilon}_i$, square them, and divide by the Maximum Likelihood estimate of the error variance from the Step 1 regression, to obtain what Breusch and Pagan call g_i :

$$g_i = \hat{\varepsilon}_i^2 / \hat{\sigma}^2, \quad \hat{\sigma}^2 = \sum \hat{\varepsilon}_i^2 / n$$

- Step 2: Estimate the auxiliary regression

$$g_i = \gamma_1 + \gamma_2 z_{2i} + \cdots + \gamma_p z_{pi} + \eta_i .$$

where the z terms will typically but not necessarily be the same as the original covariates x. - Step 3: The LM test statistic is then half of the explained sum of squares from the auxiliary regression in Step 2:

$$\mathrm{LM} = \frac{1}{2}(\mathrm{TSS} - \mathrm{RSS}).$$

where TSS is the sum of squared deviations of the g_i from their mean of 1 , and RSS is the sum of squared residuals from the auxiliary regression. The test statistic is asymptotically distributed as χ^2_{p-1} under the null hypothesis of homoskedasticity and normally distributed ε_i, as proved by Breusch and Pagan in their 1979 paper.

Methodology:

```
> #Bruesh-Pagan Test
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 8.4591, df = 6, p-value = 0.2064
```

Figure 3: Performance of Breusch Pagan Godfrey test

The test statistic is 8.4591 with a p-value of 0.2064. This indicates weak evidence against the null hypothesis of homoscedasticity, suggesting that the variance of the residuals may not be constant across all levels of the independent variables. Since the p-value is greater than 0.05, we fail to reject the null hypothesis of homoscedasticity.

1.2.3 Harvey Test:

```
> #Harvey's test
> harvey(model)
# A tibble: 1 × 4
  statistic    p.value parameter alternative
      <dbl>      <dbl>     <dbl> <chr>
1      33.2 0.00000978         6 greater
>
```

Figure 4: Performance of Harvey Test

The test statistic is 33.2 with a p-value of 0.00000978. This provides strong evidence against the null hypothesis of homoscedasticity in favor of the alternative hypothesis of heteroscedasticity. Since the p-value is less than 0.05, we reject the null hypothesis of homoscedasticity in favor of heteroscedasticity.

1.2.4 White test:

```
> #White's test
> white_test(model)
White's test results

Null hypothesis: Homoskedasticity of the residuals
Alternative hypothesis: Heteroskedasticity of the residuals
Test Statistic: 0.55
P-value: 0.758388
>
```

Figure 5: Performance of White's Test

The test statistic is 0.55 with a p-value of 0.758388. This provides no evidence against the null hypothesis of homoscedasticity. Since the p-value is greater than 0.05, we fail to reject the null hypothesis of homoscedasticity.

Conclusion: Overall, the results suggest that there is some evidence of heteroscedasticity in the residuals of our model, particularly supported by the Glejser and Harvey tests, while the Breusch-Pagan test provides weaker evidence. White's test, however, indicates homoscedasticity.

2 Tests for Autocorrelation:

2.1 Dataset:

For testing the Autocorrelation we used two tests i.e. Durbin Watson test and BreuschGodfrey test.For Durbin Watson test we used ”Vehicle dataset” from https://www.kaggle.

com/datasets/nehalbirla/vehicle-dataset-from-cardekho/data where it contains information about used cars listed on different websites. The columns in the given dataset are as follows:Car_Name, year, selling_price, Present_Price, Kms_Driven, Fuel_Type, Seller_Type, Transmission, Owner. We first made a linear regression model using selling_price as response variable and Present_Price and Kms_Driven as covariates. Then we perform Durbin Watson test and For BreuschâGodfrey test both on that model. Summary of the both tests are given below.

2.2 Tests:

2.2.1 Durbin Watson test:

If e_t is the residual given by $e_t = \rho e_{t-1} + \nu_t$, the Durbin-Watson test statistic is

$$d = \frac{\sum_{t=2}^{T} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{T} e_t^2},$$

where T is the number of observations. For large T, d is approximately equal to $2(1-\hat{\rho})$, where $\hat{\rho}$ is the sample autocorrelation of the residuals at lag 1. $d = 2$ therefore indicates no autocorrelation. The value of d always lies between 0 and 4 . If the Durbin-Watson statistic is substantially less than 2 , there is evidence of positive serial correlation. As a rough rule of thumb, if Durbin-Watson is less than 1.0, there may be cause for alarm. Small values of d indicate successive error terms are positively correlated. If $d > 2$, successive error terms are negatively correlated. In regressions, this can imply an underestimation of the level of statistical significance. To test for positive autocorrelation at significance α, the test statistic d is compared to lower and upper critical values ($d_{L,\alpha}$ and $d_{U,\alpha}$) : - If $d < d_{L,\alpha}$, there is statistical evidence that the error terms are positively autocorrelated. - If $d > d_{U,\alpha}$, there is no statistical evidence that the error terms are positively autocorrelated. - If $d_{L,\alpha} < d < d_{U,\alpha}$, the test is inconclusive.

Positive serial correlation is serial correlation in which a positive error for one observation increases the chances of a positive error for another observation.

To test for negative autocorrelation at significance α, the test statistic $(4-d)$ is compared to lower and upper critical values ($d_{L,\alpha}$ and $d_{U,\alpha}$) : - If $(4-d) < d_{L,\alpha}$, there is statistical evidence that the error terms are negatively autocorrelated. - If $(4-d) > d_{U,\alpha}$, there is no statistical evidence that the error terms are negatively autocorrelated. - If $d_{L,\alpha} < (4-d) < d_{U,\alpha}$, the test is inconclusive.

Negative serial correlation implies that a positive error for one observation increases the chance of a negative error for another observation and a negative error for one observation increases the chances of a positive error for another. If the design matrix $\mathbf{X}$ of the regression is known, exact critical values for the distribution of d under the null hypothesis of no serial correlation can be calculated. Under the null hypothesis, d

is distributed as

$$\frac{\sum_{i=1}^{n-k} \nu_i \xi_i^2}{\sum_{i=1}^{n-k} \xi_i^2},$$

where n is the number of observations and k is number of regression variables; the ξ_i are independent standard normal random variables; and the ν_i are the nonzero eigenvalues of $\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{A}$, where $\mathbf{A}$ is the matrix that transforms the residuals into the d statistic, i.e. $d = \mathbf{e}^T\mathbf{A}\mathbf{e}$.A number of computational algorithms for finding percentiles of this distribution are available.

Methodology:

```
> model

Call:
lm(formula = Selling_Price ~ Present_Price + Kms_Driven, data = car_data)

Coefficients:
  (Intercept)  Present_Price     Kms_Driven
    1.331e+00      5.356e-01     -2.043e-05

> durbinWatsonTest(model) ## Performing Durbin Watson test
 lag Autocorrelation D-W Statistic p-value
   1        0.215915       1.56752       0
 Alternative hypothesis: rho != 0
```

Figure 6: Performance of Durbin Watson test

From the output we can see that the Durbin Watson test statistic is 1.56752 and the corresponding p-value is 0.034. Since this p-value is less than 0.05 in Durbin Watson test, we can reject the null hypothesis and conclude that the residuals in this regression model are autocorrelated.

2.2.2 BreuschGodfrey test :

Consider a linear regression of any form, for example

$$Y_t = \beta_1 + \beta_2 X_{t,1} + \beta_3 X_{t,2} + u_t$$

where the errors might follow an AR(p) autoregressive scheme, as follows:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t.$$

The simple regression model is first fitted by ordinary least squares to obtain a set of sample residuals $\hat{u}_t$. Breusch and Godfrey[citation needed] proved that, if the following auxiliary regression model is fitted

$$\hat{u}_t = \alpha_0 + \alpha_1 X_{t,1} + \alpha_2 X_{t,2} + \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + \cdots + \rho_p \hat{u}_{t-p} + \varepsilon_t$$

and if the usual Coefficient of determination (R^2 statistic) is calculated for this model:

$$R^2 := \frac{\sum_{j=1}^{T-p} \left(u_{T-j} - \hat{u}_{T-j}\right)^2}{\sum_{j=1}^{T-p} \left(u_{T-j} - \bar{u}\right)^2},$$

where $\bar{u}$ stands for the arithmetic mean over the last $n = T - p$ samples, where T is the total number of observations and p is the number of error lags used in the auxiliary regression.

The following asymptotic approximation can be used for the distribution of the test statistic:

$$nR^2 \sim \chi_p^2,$$

when the null hypothesis $H_0 : \{\rho_i = 0 \text{ for all } i\}$ holds (that is, there is no serial correlation of any order up to p).

Methodology:

```
> ## Performing Breusch-Godfrey test for serial correlation of order up to 3
> bgtest(Selling_Price ~ Present_Price + Kms_Driven, order = 3, data = car_data)

	Breusch-Godfrey test for serial correlation of order up to 3

data:  Selling_Price ~ Present_Price + Kms_Driven
LM test = 15.344, df = 3, p-value = 0.001545
```

Figure 7: Performance of BreuschGodfrey test

From the output we can see that the test statistic is $\chi^2 = 15.344$ with 3 degrees of freedom. The corresponding p-value is 0.001545. Since this p-value is less than 0.05, we can reject the null hypothesis and conclude that autocorrelation exists among the residuals at some order less than or equal to 3.

3 Seemingly Unrelated Regression Equations:

3.1 Description:

Assume K series are observed for N time periods. Denote the stacked observations of series i as Y_i and its corresponding regressor matrix X_i. The complete model spanning all series can be specified as

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_K \end{bmatrix} = \begin{bmatrix} X_1 & 0 & 0 & 0 \\ 0 & X_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & X_K \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_K \end{bmatrix}.$$

The OLS estimator of β is then

$$\hat{\beta} = (X'X)^{-1} X'Y$$

A GLS estimator can be similarly defined as

$$\hat{\beta}_{GLS} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y$$

where $\Omega = \Sigma \otimes I_N$ is the joint covariance of the residuals. In practice Σ is not known as so a feasible GLS (FGLS) is implemented in two steps. The first uses OLS to estimate $\hat{\epsilon}$ and then estimates the residual covariance as

$$\hat{\Sigma} = N^{-1} \begin{bmatrix} \hat{\epsilon}_1 & \hat{\epsilon}_2 & \dots & \hat{\epsilon}_N \end{bmatrix}' \begin{bmatrix} \hat{\epsilon}_1 & \hat{\epsilon}_2 & \dots & \hat{\epsilon}_N \end{bmatrix}.$$

The feasible GLS estimator is then

$$\hat{\beta}_{FGLS} = \left(X'\hat{\Omega}^{-1}X \right)^{-1} X'\hat{\Omega}^{-1}Y$$

where $\hat{\Omega} = \hat{\Sigma} \otimes I_N$.

3.2 Dataset:

For SURE estimation we have downloaded data from https://pages.stern.nyu.edu/~wgreene/Text/Edition7/tablelist8new.htm. We worked on Munnell Productivity Data, 48 Continental U.S. States, 17 years,1970 to 1986,Source: Baltagi (2005), Munnell (1990). It is a panel data. So, we performed SURE estimation upon this dataset. In this data set there are 50 States of USA and there are 17 years from 1970 to 1986 and coloumns are STATE, ST_ABB, YR, P_CAP, HWY, WATER, UTIL, PC, GSP, EMP, UNEMP. Now, We have taken Gross state product(GSP) as the response and Private capital(PC), Water utility capital(WATER), Public capital(P_CAP), state unemployment rate(UNEMP), Highway capital(HWY), Utility capital(UTIL) are covariates.

3.3 Methodology:

```
systemfit results
method: SUR

Coefficients:
eq1_(Intercept)      eq1_P_CAP        eq1_HWY      eq1_WATER       eq1_UTIL
    3.25357e+03   -5.24883e+04    5.24895e+04    5.24935e+04    5.24894e+04
         eq1_PC      eq1_UNEMP
    4.13365e-01   -1.21889e+03
```

Figure 8: Performance of SURE model

First, the estimation method is reported and a few summary statistics for the entire system and for each equation are given. Then, the covariance matrix used for estimation and the covariance matrix as well as the correlation matrix of the (final) residuals are printed. Finally, the estimation results of each equation are reported: the formula of the estimated equation, the estimated coefficients, their standard errors, t values, P values and codes indicating their statistical significance, as well as some other statistics like the standard error of the residuals and the R^2 value of the equation. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables. The adjusted R^2 of the model is 0.981487 which is less than it's Multiple R-squared = 0.981623 which is very close to 1 that implies predictors improved the model as expected. Other summaries are given below.

```
> summary(Sure_model)

systemfit results
method: SUR

         N  DF         SSR  detRCov   OLS-R2 McElroy-R2
system 816 809 73333632909 90647259 0.981623   0.981623

      N  DF         SSR      MSE    RMSE       R2   Adj R2
eq1 816 809 73333632909 90647259 9520.89 0.981623 0.981487

The covariance matrix of the residuals used for estimation
         eq1
eq1 90646526

The covariance matrix of the residuals
         eq1
eq1 90647259

The correlations of the residuals
    eq1
eq1   1


SUR estimates for 'eq1' (equation 1)
Model Formula: GSP ~ P_CAP + HWY + WATER + UTIL + PC + UNEMP

                Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)  3.25357e+03  1.11866e+03  2.90844  0.0037318 **
P_CAP       -5.24883e+04  5.65352e+04 -0.92842  0.3534669
HWY          5.24895e+04  5.65352e+04  0.92844  0.3534567
WATER        5.24935e+04  5.65352e+04  0.92851  0.3534198
UTIL         5.24894e+04  5.65352e+04  0.92844  0.3534573
PC           4.13365e-01  1.49275e-02 27.69152 < 2.22e-16 ***
UNEMP       -1.21889e+03  1.52694e+02 -7.98261  4.885e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9520.885435 on 809 degrees of freedom
Number of observations: 816 Degrees of Freedom: 809
SSR: 73333632908.5861 MSE: 90647259.466732 Root MSE: 9520.885435
Multiple R-Squared: 0.981623 Adjusted R-Squared: 0.981487
```

Figure 9: Performance of SURE model

4 Panel Data model:

4.1 Dataset:

For Panel data analysis, we have used the dataset **Cost Data of U.S Airlines** consisting of 90 Observations on 6 firms for 15 years, $1970 - 1984$

Predictors:

- I = Airline
- T = Year
- Q = Output, in revenue passenger miles, index number
- PF = fuel price
- LF = Load factor, the average capacity utilization of the fleet.

Response:

- C = Total cost, in \$1000

We are making a matrix over Airline and Time.

- Y = vector of Total cost, in \$1000
- X = matrix of Q, PF, and LF.

4.2 Methodology:

```
> summary(Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  68978  292046  637001 1122524 1345968 4748320 
> summary(X)
       V1                 V2                V3        
 Min.   :0.03768   Min.   : 103795   Min.   :0.4321  
 1st Qu.:0.14213   1st Qu.: 129848   1st Qu.:0.5288  
 Median :0.30503   Median : 357434   Median :0.5661  
 Mean   :0.54499   Mean   : 471683   Mean   :0.5605  
 3rd Qu.:0.94528   3rd Qu.: 849840   3rd Qu.:0.5947  
 Max.   :1.93646   Max.   :1015610   Max.   :0.6763  
```

Figure 10: Summary of Y and X

Ways to handle a pooled model

❊ Pooling model:

```
Pooling Model

Call:
plm(formula = Y ~ X, data = pdata, model = "pooling")

Balanced Panel: n = 6, T = 15, N = 90

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
-520654 -250270   37333  208690  849700

Coefficients:
               Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)  1.1586e+06  3.6059e+05  3.2129   0.00185 **
X1           2.0261e+06  6.1807e+04 32.7813 < 2.2e-16 ***
X2           1.2253e+00  1.0372e-01 11.8138 < 2.2e-16 ***
X3          -3.0658e+06  6.9633e+05 -4.4027 3.058e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1.2647e+14
Residual Sum of Squares: 6.8177e+12
R-Squared:      0.94609
Adj. R-Squared: 0.94421
F-statistic: 503.118 on 3 and 86 DF, p-value: < 2.22e-16
```

Figure 11: Pooling model

❊ Between model:

```
Oneway (individual) effect Between Model

Call:
plm(formula = Y ~ X, data = pdata, model = "between")

Balanced Panel: n = 6, T = 15, N = 90
Observations used in estimation: 6

Residuals:
      1       2       3       4       5       6
 -38528   58079  -44440   98838   32460 -106409

Coefficients:
               Estimate  Std. Error t-value Pr(>|t|)
(Intercept)  6.2014e+06  1.8381e+07  0.3374  0.76795
X1           1.8183e+06  5.4493e+05  3.3367  0.07928 .
X2          -9.4242e+00  4.2111e+01 -0.2238  0.84370
X3          -2.8987e+06  3.9612e+06 -0.7318  0.54045
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    5.0464e+12
Residual Sum of Squares: 2.8978e+10
R-Squared:      0.99426
Adj. R-Squared: 0.98564
F-statistic: 115.432 on 3 and 2 DF, p-value: 0.008601
```

Figure 12: Between model

❊ Within model:

```
Oneway (individual) effect Within Model

Call:
plm(formula = Y ~ X, data = pdata, model = "within")

Balanced Panel: n = 6, T = 15, N = 90

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-551783 -159259    1796       0  137226  499296

Coefficients:
      Estimate  Std. Error t-value  Pr(>|t|)
X1  3.3190e+06  1.7135e+05 19.3694 < 2.2e-16 ***
X2  7.7307e-01  9.7319e-02  7.9437 9.698e-12 ***
X3 -3.7974e+06  6.1377e+05 -6.1869 2.375e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    5.0776e+13
Residual Sum of Squares: 3.5865e+12
R-Squared:      0.92937
Adj. R-Squared: 0.92239
F-statistic: 355.254 on 3 and 81 DF, p-value: < 2.22e-16
```

Figure 13: Within model

❊ Random model:

```
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = Y ~ X, data = pdata, model = "random")

Balanced Panel: n = 6, T = 15, N = 90

Effects:
                   var   std.dev share
idiosyncratic 4.428e+10 2.104e+05 0.793
individual    1.154e+10 1.074e+05 0.207
theta: 0.5486

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
-535726 -238494   49890  207491  722934

Coefficients:
              Estimate  Std. Error z-value  Pr(>|z|)
(Intercept)  1.0743e+06  3.7747e+05  2.8461  0.004427 **
X1           2.2886e+06  1.0949e+05 20.9015 < 2.2e-16 ***
X2           1.1236e+00  1.0344e-01 10.8622 < 2.2e-16 ***
X3          -3.0850e+06  7.2568e+05 -4.2512 2.126e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    6.6198e+13
Residual Sum of Squares: 5.8721e+12
R-Squared:      0.91129
Adj. R-Squared: 0.9082
Chisq: 883.501 on 3 DF, p-value: < 2.22e-16
```

Figure 14: Random model

❊ First Difference Model:

```
Oneway (individual) effect First-Difference Model

Call:
plm(formula = Y ~ X, data = pdata, model = "fd")

Balanced Panel: n = 6, T = 15, N = 90
Observations used in estimation: 84

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
-232631  -58504  -25086   31884  493212

Coefficients:
               Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)  7.3268e+04  1.6544e+04  4.4286 2.975e-05 ***
X1           1.1493e+06  2.1346e+05  5.3842 7.099e-07 ***
X2           5.7761e-01  1.3371e-01  4.3197 4.449e-05 ***
X3          -1.7024e+06  4.7366e+05 -3.5941  0.000561 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1.5613e+12
Residual Sum of Squares: 9.9318e+11
R-Squared:      0.36388
Adj. R-Squared: 0.34002
F-statistic: 15.2539 on 3 and 80 DF, p-value: 6.1466e-08
```

Figure 15: First Difference Model

Testing:

❊ Lagrange Multiplier Test for Random effects v/s OLS:

```
        Lagrange Multiplier Test - (Honda)

data:  Y ~ X
normal = 0.783, p-value = 0.2168
alternative hypothesis: significant effects
```

Figure 16: LM test for random effects v/s OLS

The outcomes of the Lagrange Multiplier Test could suggest that there would be random effects.

❊ F test for individual effects:

```
        F test for individual effects

data:  Y ~ X
F = 14.595, df1 = 5, df2 = 81, p-value = 3.467e-10
alternative hypothesis: significant effects
```

Figure 17: LM test for fixed effects v/s OLS

The outcomes of the F-Test could suggest that there would be fixed effects.

❊ Hausman Test:

```
        Hausman Test

data:  Y ~ X
chisq = 60.87, df = 3, p-value = 3.832e-13
alternative hypothesis: one model is inconsistent
```

Figure 18: Hausman test for fixed v/s random effects model

The outcome of the Hausman test is an alternative hypothesis, which is one model is inconsistent. It recommends using the Random Effect Model.

5 Augmented Dickey-Fuller test

The testing procedure for the ADF test is the same as for the Dickey-Fuller test but it is applied to the model

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

where α is a constant, β the coefficient on a time trend and p the lag order of the autoregressive process. Imposing the constraints $\alpha = 0$ and $\beta = 0$ corresponds to modelling a random walk and using the constraint $\beta = 0$ corresponds to modeling a random walk with a drift. Consequently, there are three main versions of the test, analogous to the ones discussed on Dickey-Fuller test (see that page for a discussion on dealing with uncertainty about including the intercept and deterministic time trend terms in the test equation.)

By including lags of the order p the ADF formulation allows for higher-order autoregressive processes. This means that the lag length p has to be determined when applying the test. One possible approach is to test down from high orders and examine the t-values on coefficients. An alternative approach is to examine information criteria such as the Akaike information criterion, Bayesian information criterion or the Hannan-Quinn information criterion.

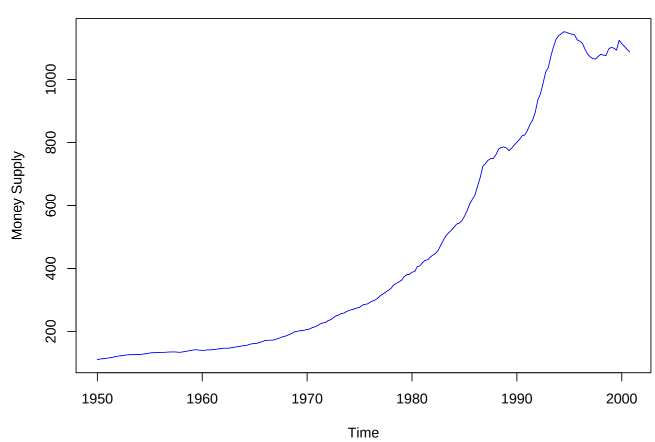
The unit root test is then carried out under the null hypothesis $\gamma = 0$ against the alternative hypothesis of $\gamma < 0$. Once a value for the test statistic

$$\mathrm{DF}_\tau = \frac{\hat{\gamma}}{\mathrm{SE}(\hat{\gamma})}$$

is computed it can be compared to the relevant critical value for the Dickey-Fuller test. As this test is asymmetrical, we are only concerned with negative values of our test statistic DF_τ. If the calculated test statistic is less (more negative) than the critical value, then the null hypothesis of $\gamma = 0$ is rejected and no unit root is present.

5.1 Methodology:

- **Objective:** To check the stationarity in the time-series data.
- Data description:
  - US Macroeconomic Data (1950â2000, Greene)
  - This is the Time series data on 12 US macroeconomic variables for 1950â2000.
  - A quarterly multiple time series from 1950 to 2000 with 12 variables.
  - Variables are GDP, consumption, invest, government, dpi, cpi, m1, tbill, unemp, population, inflation, interest.
  - we have worked on with the m1 (money supply) and unemp (Unemployment rate) variable.
- Null hypothesis H_0 : Money supply data has a unit root i.e. the data is not stationary.



- From the previous slide we can observe Money Supply has been continuously increased up to year 1990 and there is a fall in between 1990 to 2000.

```
> adf.test(money_supply)

        Augmented Dickey-Fuller Test

data:  money_supply
Dickey-Fuller = -2.0744, Lag order = 5, p-value = 0.5449
alternative hypothesis: stationary
```

- Also from the ADF test we can observe that the p-value is 0.5449 which is greater than 5% level of significance. So, we accept the null hypothesis that the Money supply data does not have a unit root, i.e., the data is non-stationary.
- Null hypothesis H_0 : Unemployed rate data has a unit root i.e. the data is not stationary.
- Alternative hypothesis H_A : Unemployed rate data does not have a unit root i.e. the data is stationary.

```
> adf.test(unemp_us)

        Augmented Dickey-Fuller Test

data:  unemp_us
Dickey-Fuller = -2.4984, Lag order = 5, p-value = 0.3673
alternative hypothesis: stationary
```

- From the ADF test we can observe that the p-value is 0.3673 which is greater than 5% level of significance. So, we accept the null hypothesis that the Unemployment rate data does not have a unit root i.e. here the data is non-stationary.

6 Phillips Perron test:

6.1 Description:

In statistics, the Phillips Perron test (named after Peter C. B. Phillips and Pierre Perron) is a unit root test. That is, it is used in time series analysis to test the null hypothesis that a time series is integrated of order 1. It builds on the Dickey–Fuller test of the null hypothesis $\rho = 1$ in

$$\Delta y_t = (\rho - 1)y_{t-1} + u_t,$$

where Δ is the first difference operator. Like the augmented Dickey–Fuller test, the Phillips–Perron test addresses the issue that the process generating data for y_t might have a higher order of autocorrelation than is admitted in the test equationâmaking y_{t-1} endogenous and thus invalidating the Dickey–Fuller test. Whilst the augmented Dickey–Fuller test addresses this issue by introducing lags of Δy_t as regressors in the test equation, the Phillips–Perron test makes a non-parametric correction to the t-test statistic. The test is robust with respect to unspecified autocorrelation and heteroscedasticity in the disturbance process of the test equation.

6.2 Methodology:

❊ **Objective:** To check the stationarity in the time-series data.

❊ Data description:

- ➢ US Macroeconomic Data (1950â2000, Greene)
- ➢ This is the Time series data on 12 US macroeconomic variables for 1950â2000.
- ➢ A quarterly multiple time series from 1950 to 2000 with 12 variables.
- ➢ Variables are GDP, consumption, invest, government, dpi, cpi, m1, tbill, unemp, population, inflation, interest.
- ➢ we have worked on with the unemp (Unemployment rate) and inflation (inflation rate) variable.

❊ Our null hypothesis H_0 : The employment rate data has a unit root or the data is not stationary.

❊ Our alternative hypothesis H_A : The employment rate data does not have a unit root or the data is stationary.

```
> pp.test(USMacroG[,9]) ### Unemployment

        Phillips-Perron Unit Root Test

data:  USMacroG[, 9]
Dickey-Fuller Z(alpha) = -15.289, Truncation lag parameter = 4, p-value = 0.2449
alternative hypothesis: stationary
```

❊ Here, we are getting a p-value is 0.2499 which is greater than 5% level of significance. So, we accept our null hypothesis that the data(Unemployment rate) has a unit root i.e. here the data is not stationary.

❊ Our null hypothesis H_0 : The inflation rate data has a unit root or the data is not stationary.

❊ Our alternative hypothesis H_A : The inflation rate data does not have a unit root or the data is stationary.

```
> pp.test(USMacroG[-1,11]) ### Inflation

        Phillips-Perron Unit Root Test

data:  USMacroG[-1, 11]
Dickey-Fuller Z(alpha) = -68.274, Truncation lag parameter = 4, p-value = 0.01
alternative hypothesis: stationary
```

❊ Here, we are getting a p-value is 0.01 which is less than 5% level of significance. So, we reject our null hypothesis that the data(inflation rate) does not have a unit root i.e. here the data is stationary.

7 Vector Auto regressive Model[VAR(p)]

7.1 Description:

The vector autoregression (VAR) model extends the idea of univariate autoregression to k time series regressions, where the lagged values of all k series appear as regressors. Put differently, in a VAR model we regress a vector of time series variables on lagged vectors of these variables. As for AR(p) models, the lag order is denoted by p so the VAR(p) model of two variables X_t and $Y_t (k = 2)$ is given by the equations

$$Y_t = \beta_{10} + \beta_{11} Y_{t-1} + ... + \beta_{1p} Y_{t-p} + \gamma_{11} X_{t-1} + ... + \gamma_{1p} X_{t-p} + u_{1t}$$
$$X_t = \beta_{20} + \beta_{21} Y_{t-1} + ... + \beta_{2p} Y_{t-p} + \gamma_{21} X_{t-1} + ... + \gamma_{2p} X_{t-p} + u_{2t}$$

7.2 Dataset:

- We have consider a dataset called `denmark` from the package `urca` in `R`.
- Data collection: A data frame with 55 observations on the following 6 variables.
  - `period`: Time index from 1974:Q1 until 1987:Q3.
  - `LRM`: Logarithm of real money, M2.
  - `LRY`: Logarithm of real income.
  - `LPY`: Logarithm of price deflator.
  - `IBO`: Bond rate.
  - `IDE`: Bank deposit rate.
- we have considered how to estimate a VAR model of `LRM`, real money and `LPY`, price deflator.

$$LRM_t = \beta_{10} + \beta_{11} LRM_{t-1} + ... + \beta_{1p} LRM_{t-p} + \gamma_{11} LPY_{t-1} + ... + \gamma_{1p} LPY_{t-p} + u_{1t}$$
$$LPY_t = \beta_{20} + \beta_{21} LRM_{t-1} + ... + \beta_{2p} LRM_{t-p} + \gamma_{21} LPY_{t-1} + ... + \gamma_{2p} LPY_{t-p} + u_{2t}$$
$$p = 2$$

- We estimate both equations separately by OLS and use `coeftest()` to obtain robust standard errors.

```
> VAR_eq1 <- dynlm(LRM ~ L(LRM,1:2) + L(LPY,1:2),
+                  start =  c(1947,1), end = c(1987,3))
> names(VAR_eq1$coefficients) <- c("Intercept","LRM_t-1","LRM_t-2","LPY_t-1","LPY_t-2")
> coeftest(VAR_eq1)

t test of coefficients:

           Estimate Std. Error t value  Pr(>|t|)
Intercept  1.122358   0.421623  2.6620  0.008582 **
LRM_t-1    1.202527   0.126023  9.5421 < 2.2e-16 ***
LRM_t-2   -0.297854   0.128488 -2.3182  0.021740 *
LPY_t-1   -0.065759   0.056534 -1.1632  0.246529
LPY_t-2    0.077145   0.055227  1.3969  0.164430
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 19: VAR model equation-1 in when LRM is response variable

```
> VAR_eq2 <- dynlm(LPY ~ L(LRM,1:2) + L(LPY,1:2),
+                  start =  c(1947,1), end = c(1987,3))
> names(VAR_eq2$coefficients) <- names(VAR_eq1$coefficients)
> coeftest(VAR_eq2)

t test of coefficients:

          Estimate Std. Error t value  Pr(>|t|)
Intercept  2.21440    0.94667  2.3392   0.02060 *
LRM_t-1    0.38418    0.28296  1.3577   0.17651
LRM_t-2   -0.57211    0.28849 -1.9831   0.04911 *
LPY_t-1    0.77574    0.12694  6.1113 7.587e-09 ***
LPY_t-2    0.18442    0.12400  1.4873   0.13896
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 20: VAR model equation-2 in when LPY is response variable

- We end up with the following results:
  - $LRM_t = (1.12\pm0.42)+(1.20\pm0.12)LRM_{t-1}+(-0.29\pm0.12)LRM_{t-2}+(-0.065\pm 0.05)LPY_{t-1} + (0.077 \pm 0.05)LPY_{t-2} + u_{1t}$
  - $LPY_t = (2.21\pm0.94)+(0.38\pm0.28)LRM_{t-1}+(-0.57\pm0.288)LRM_{t-2}+(0.77\pm 0.126)LPY_{t-1} + (0.18 \pm 0.124)LPY_{t-2} + u_{2t}$
- The function VAR() can be used to obtain the same coefficient estimates as presented above since it applies OLS per equation.

```
> VAR_est <- VAR(y = VAR_data, p = 2)
> VAR_est

VAR Estimation Results:
=======================

Estimated coefficients for equation LRM:
========================================
Call:
LRM = LRM.l1 + LPY.l1 + LRM.l2 + LPY.l2 + const

     LRM.l1      LPY.l1      LRM.l2      LPY.l2       const 
 1.20252722 -0.06575943 -0.29785383  0.07714547  1.12235843 


Estimated coefficients for equation LPY:
========================================
Call:
LPY = LRM.l1 + LPY.l1 + LRM.l2 + LPY.l2 + const

    LRM.l1     LPY.l1     LRM.l2     LPY.l2      const 
 0.3841836  0.7757426 -0.5721125  0.1844239  2.2144021 
```

Figure 21: VAR model of both equations using VAR() function

- For instance, if an increase in LRM leads to an increase in LPY in subsequent periods, it may indicate inflationary pressure resulting from changes in the money supply.
- Conversely, if changes in LPY affect LRM, it may imply feedback effects between inflation and monetary policy. The significance and direction of these relationships can provide insights into the monetary dynamics of the economy and help policymakers understand the drivers of inflation.

8 Vector Moving Averages[VMA(q)]

8.1 Description:

Given the n-dimensional vector White Noise ϵ_t a vector moving average of order q is defined as

$$Y_t = \mu + \epsilon_t + C_1\epsilon_{(t-1)} + ... + C_q\epsilon_{(t-q)}$$

where C_j are n x n matrices of coefficients and μ is the mean of Y_t

- For estimating the VMA(q) model, we have generated the data

$$q = 3$$
$$Y_t = \mu + \epsilon_t + C_1\epsilon_{(t-1)} + C_2\epsilon_{(t-2)} + C_3\epsilon_{(t-3)}$$
$$\mu = sin(x\pi) \; ; \; x \in \mathbf{R}$$
$$\epsilon_t \sim \text{white noise}(0, \sigma^2 = 0.5^2) \text{ for all } t$$

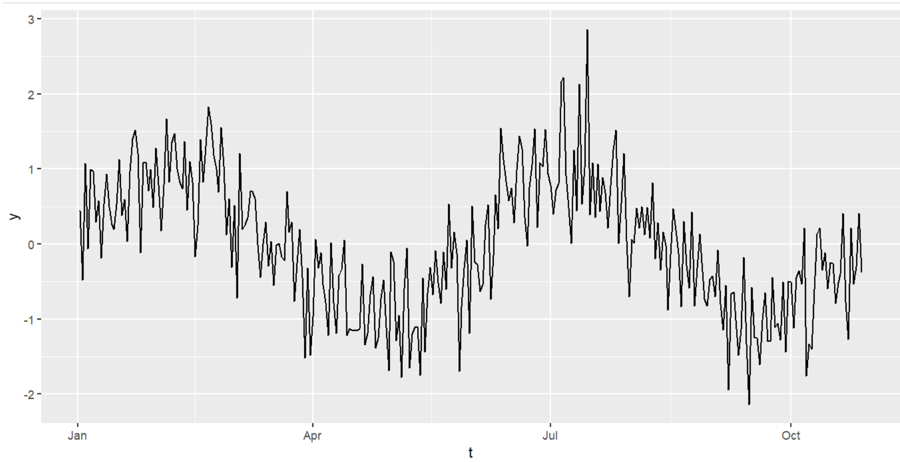


Figure 22: Graph of Y variable

❊ Calculating for VMA($q = 3$) model, we have used `filter` function from `stats` package in `R`.

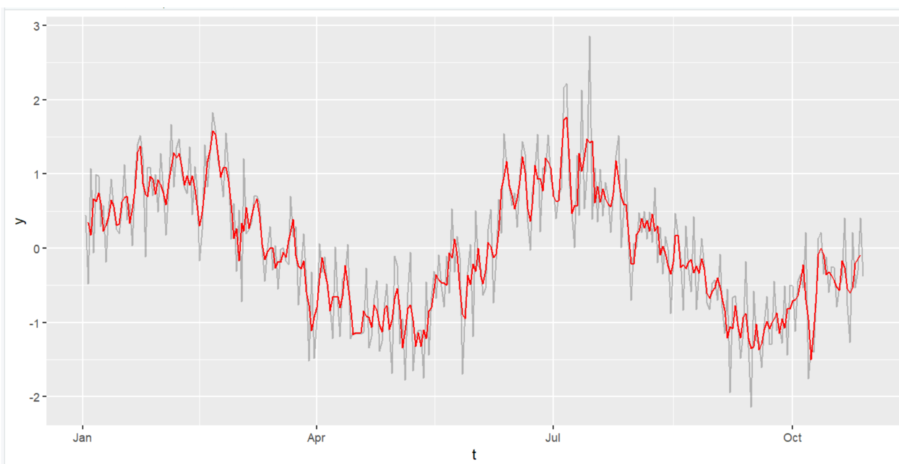


Figure 23: Estimate of VMA(3) on Y variable

❊ We can also use other functions for calculating moving averages

- ➢ `runmean`() from the caTools package
- ➢ `frollmean`() from the data.table package

❊ One reason to calculate a moving average is to smooth out day-to-day variation.

9 Impulse Response function

9.1 Description:

Impulse response analysis is an important step in econometric analysis, which employs vector autoregressive models. Their main purpose is to describe the evolution of a model's variables in reaction to a shock in one or more variables. This feature allows us to trace the transmission of a single shock within an otherwise noisy system of equations.

For constructing the IRF, we realize that the VAR (p) model can be written in the equivalent Vector Moving Average VMA(∞) representation as

$$Y_t = \eta + \epsilon_t + \Psi_1\epsilon_{t-1} + \Psi_2\epsilon_{t-2} + ...$$
$$\frac{\partial Y_{t+s}}{\partial \epsilon_t} = \Psi_s$$

The plot of $\frac{\partial Y_{i,t+s}}{\partial \epsilon_{j,t}}$ as a function of s is called the impulse response plot of variable Y_i for shocks in Y_j.

9.2 Data description:

http://www.jmulti.de/download/datasets/e1.dat. It is the data containing quarterly, seasonally adjusted, West German fixed investment(`invest`), disposable income(`income`), consumption expenditures (`cons`) in billions of DM, 1960Q1-1982Q4

9.3 Methodology:

❊ Plotting all the variables time series plot

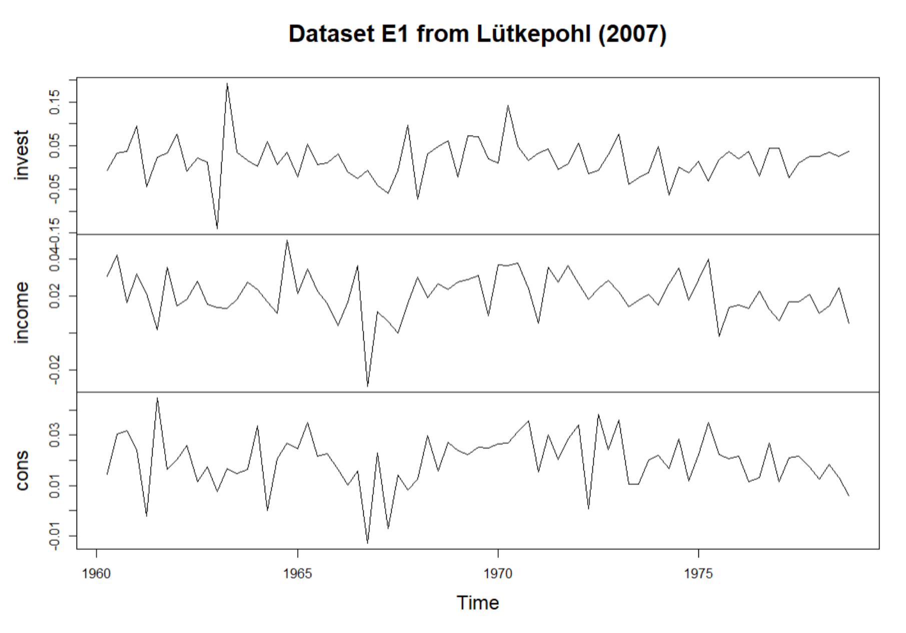


Figure 24: Three variables time series plots

❊ This data is used to estimate a VAR(2) model with a constant term.

```
> # Estimate model
> model <- VAR(data, p = 2, type = "const")
> # Look at summary statistics
> summary(model)

VAR Estimation Results:
=========================
Endogenous variables: invest, income, cons
Deterministic variables: const
Sample size: 73
Log Likelihood: 606.307
Roots of the characteristic polynomial:
0.5705 0.5513 0.5513 0.4917 0.4917 0.3712
Call:
VAR(y = data, p = 2, type = "const")


Estimation results for equation invest:
=======================================
invest = invest.l1 + income.l1 + cons.l1 + invest.l2 + income.l2 + cons.l2 + const

          Estimate Std. Error t value Pr(>|t|)
invest.l1 -0.31963    0.12546  -2.548   0.0132 *
income.l1  0.14599    0.54567   0.268   0.7899
cons.l1    0.96122    0.66431   1.447   0.1526
invest.l2 -0.16055    0.12491  -1.285   0.2032
income.l2  0.11460    0.53457   0.214   0.8309
cons.l2    0.93439    0.66510   1.405   0.1647
const     -0.01672    0.01723  -0.971   0.3352
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.04615 on 66 degrees of freedom
Multiple R-Squared: 0.1286,     Adjusted R-squared: 0.04934
F-statistic: 1.623 on 6 and 66 DF,  p-value: 0.1547
```

Figure 25: VAR(2) model when invest as response

```
Estimation results for equation income:
=======================================
income = invest.l1 + income.l1 + cons.l1 + invest.l2 + income.l2 + cons.l2 + const

           Estimate Std. Error t value Pr(>|t|)
invest.l1  0.043931   0.031859   1.379 0.172578
income.l1 -0.152732   0.138570  -1.102 0.274378
cons.l1    0.288502   0.168700   1.710 0.091936 .
invest.l2  0.050031   0.031720   1.577 0.119512
income.l2  0.019166   0.135752   0.141 0.888156
cons.l2   -0.010205   0.168899  -0.060 0.952004
const      0.015767   0.004375   3.604 0.000602 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.01172 on 66 degrees of freedom
Multiple R-Squared: 0.1142,     Adjusted R-squared: 0.03367
F-statistic: 1.418 on 6 and 66 DF,  p-value: 0.221
```

Figure 26: VAR(2) model when income as response

```
Estimation results for equation cons:
=====================================
cons = invest.l1 + income.l1 + cons.l1 + invest.l2 + income.l2 + cons.l2 + const

           Estimate Std. Error t value Pr(>|t|)
invest.l1 -0.002423   0.025676  -0.094 0.925114
income.l1  0.224813   0.111678   2.013 0.048191 *
cons.l1   -0.263968   0.135960  -1.942 0.056467 .
invest.l2  0.033880   0.025564   1.325 0.189631
income.l2  0.354912   0.109407   3.244 0.001851 **
cons.l2   -0.022230   0.136120  -0.163 0.870772
const      0.012926   0.003526   3.666 0.000493 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.009445 on 66 degrees of freedom
Multiple R-Squared: 0.2513,     Adjusted R-squared: 0.1832
F-statistic: 3.692 on 6 and 66 DF,  p-value: 0.003184



Covariance matrix of residuals:
          invest    income      cons
invest 2.130e-03 7.162e-05 1.232e-04
income 7.162e-05 1.373e-04 6.146e-05
cons   1.232e-04 6.146e-05 8.920e-05

Correlation matrix of residuals:
       invest income   cons
invest 1.0000 0.1324 0.2828
income 0.1324 1.0000 0.5553
cons   0.2828 0.5553 1.0000
```

Figure 27: VAR(2) model when cons as response and covariance & correlation

- Forecast error impulse response

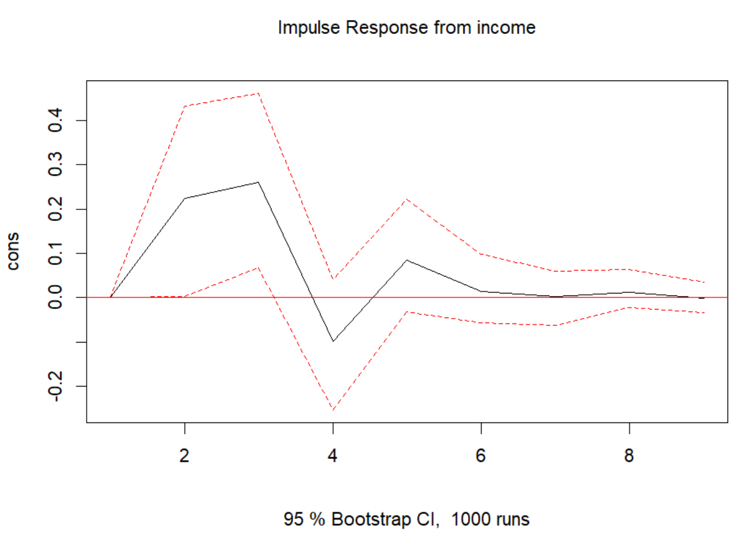


Figure 28: Forecast impulse response plot

- A caveat of FEIRs is that they cannot be used to assess contemporaneous reactions of variables. This can be seen in the previous plot, where the FEIR is zero in the first period.

- A common approach to identify the shocks of a VAR model is to use orthogonal impulse responses (OIR). The basic idea is to decompose the variance-covariance matrix so that $\Sigma = PP^T$, where P is a lower triangular matrix with positive diagonal elements, which is often obtained by a Choleski decomposition.

❊ Orthogonal impulse responses

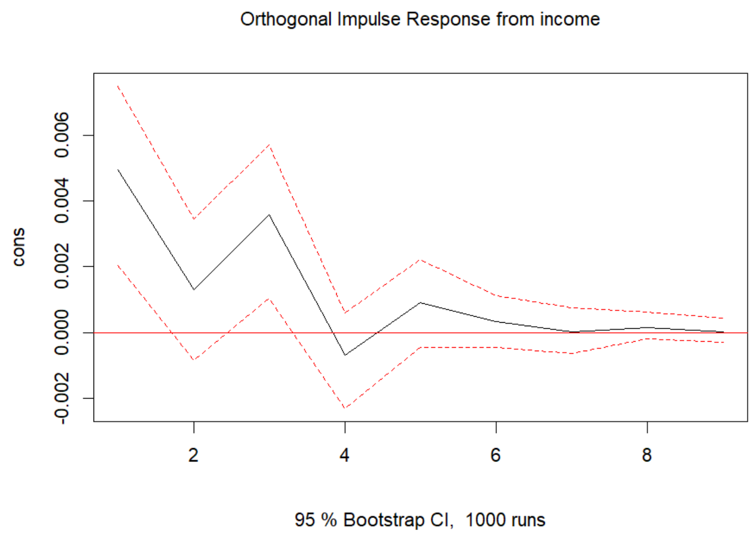


Figure 29: Orthogonal impulse response plot

10 Error Correction Model

Error Correction model examines whether a cointegration exists between GC and NQ for the April Data series. If it exists, we can estimate the speed of adjustment of GC to NQ when there is a shock.

10.1 Plot of the data

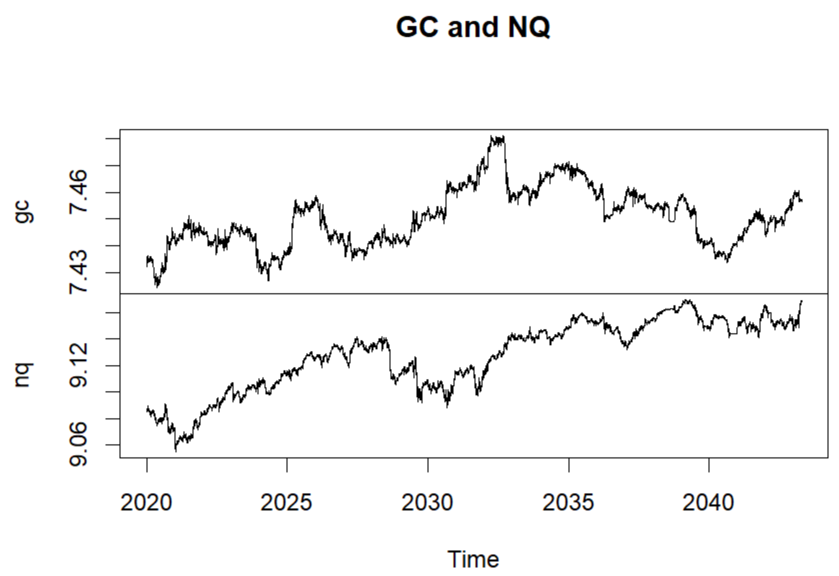


Figure 30: Plot of data between GC, NQ with time

10.2 Stationary test for both GC and NQ

➢ GC

```
> adf.test(ln_gc)
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
      lag   ADF p.value
 [1,]   0 0.399   0.759
 [2,]   1 0.405   0.761
 [3,]   2 0.417   0.764
 [4,]   3 0.427   0.767
 [5,]   4 0.440   0.771
 [6,]   5 0.444   0.772
 [7,]   6 0.464   0.778
 [8,]   7 0.464   0.778
 [9,]   8 0.471   0.780
[10,]   9 0.467   0.778
[11,]  10 0.479   0.782
[12,]  11 0.482   0.783
[13,]  12 0.475   0.781
[14,]  13 0.476   0.781
[15,]  14 0.469   0.779
```

```
Type 2: with drift no trend
       lag   ADF p.value
 [1,]    0 -2.48   0.135
 [2,]    1 -2.47   0.141
 [3,]    2 -2.40   0.167
 [4,]    3 -2.37   0.179
 [5,]    4 -2.35   0.187
 [6,]    5 -2.35   0.189
 [7,]    6 -2.34   0.190
 [8,]    7 -2.33   0.194
 [9,]    8 -2.32   0.202
[10,]    9 -2.31   0.204
[11,]   10 -2.31   0.203
[12,]   11 -2.33   0.195
[13,]   12 -2.35   0.190
[14,]   13 -2.34   0.194
[15,]   14 -2.36   0.182
Type 3: with drift and trend
       lag   ADF p.value
 [1,]    0 -2.57   0.336
 [2,]    1 -2.55   0.344
 [3,]    2 -2.48   0.375
 [4,]    3 -2.44   0.391
 [5,]    4 -2.41   0.403
 [6,]    5 -2.40   0.406
 [7,]    6 -2.39   0.413
 [8,]    7 -2.38   0.417
 [9,]    8 -2.35   0.427
[10,]    9 -2.35   0.428
[11,]   10 -2.35   0.430
[12,]   11 -2.37   0.422
[13,]   12 -2.38   0.415
[14,]   13 -2.37   0.420
[15,]   14 -2.40   0.406
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

Figure 31: ADF test for ln(GC)

- NQ

```
> adf.test(ln_nq)
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
      lag  ADF p.value
 [1,]   0 1.15   0.934
 [2,]   1 1.13   0.931
 [3,]   2 1.15   0.933
 [4,]   3 1.16   0.934
 [5,]   4 1.15   0.933
 [6,]   5 1.16   0.934
 [7,]   6 1.17   0.935
 [8,]   7 1.17   0.936
 [9,]   8 1.18   0.937
[10,]   9 1.20   0.939
[11,]  10 1.20   0.940
[12,]  11 1.20   0.940
[13,]  12 1.19   0.938
[14,]  13 1.21   0.941
[15,]  14 1.21   0.940
```

```
Type 2: with drift no trend
      lag   ADF p.value
 [1,]   0 -1.29   0.599
 [2,]   1 -1.34   0.583
 [3,]   2 -1.33   0.583
 [4,]   3 -1.32   0.587
 [5,]   4 -1.31   0.590
 [6,]   5 -1.31   0.593
 [7,]   6 -1.31   0.593
 [8,]   7 -1.30   0.596
 [9,]   8 -1.30   0.595
[10,]   9 -1.30   0.596
[11,]  10 -1.30   0.596
[12,]  11 -1.32   0.589
[13,]  12 -1.30   0.596
[14,]  13 -1.29   0.600
[15,]  14 -1.30   0.596
Type 3: with drift and trend
      lag   ADF p.value
 [1,]   0 -2.66   0.297
 [2,]   1 -2.73   0.269
 [3,]   2 -2.71   0.276
 [4,]   3 -2.69   0.286
 [5,]   4 -2.69   0.286
 [6,]   5 -2.68   0.291
 [7,]   6 -2.66   0.297
 [8,]   7 -2.65   0.303
 [9,]   8 -2.64   0.305
[10,]   9 -2.61   0.317
[11,]  10 -2.61   0.318
[12,]  11 -2.63   0.312
[13,]  12 -2.63   0.312
[14,]  13 -2.59   0.327
[15,]  14 -2.61   0.320
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

Figure 32: ADF test for ln(NQ)

- Stationary test tells us that GC and NQ are I(d); that is they contain a unit root. Now, we regress GC on NQ.

```
> regression <- lm(ln_gc ~ ln_nq)
> summary(regression)

Call:
lm(formula = ln_gc ~ ln_nq)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0205196 -0.0077152 -0.0004225  0.0059443  0.0309139

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.129400   0.021000  291.88   <2e-16 ***
ln_nq       0.144762   0.002301   62.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0104 on 29343 degrees of freedom
Multiple R-squared:  0.1189,	Adjusted R-squared:  0.1188
F-statistic:  3959 on 1 and 29343 DF,  p-value: < 2.2e-16
```

Figure 33: Summary of regression model

- now, we extract the residuals and subject them to unit root test. The result confirms that it is stationary. In other words, cointegration exists.

```
> adf.test(ect)
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
      lag   ADF p.value
 [1,]   0 -2.54  0.0115
 [2,]   1 -2.53  0.0123
 [3,]   2 -2.46  0.0150
 [4,]   3 -2.43  0.0166
 [5,]   4 -2.40  0.0177
 [6,]   5 -2.40  0.0179
 [7,]   6 -2.39  0.0184
 [8,]   7 -2.37  0.0190
 [9,]   8 -2.35  0.0200
[10,]   9 -2.35  0.0200
[11,]  10 -2.34  0.0204
[12,]  11 -2.36  0.0194
[13,]  12 -2.39  0.0182
[14,]  13 -2.38  0.0188
[15,]  14 -2.39  0.0180
```

```
Type 2: with drift no trend
      lag   ADF p.value
 [1,]   0 -2.54   0.110
 [2,]   1 -2.53   0.117
 [3,]   2 -2.46   0.143
 [4,]   3 -2.43   0.157
 [5,]   4 -2.40   0.168
 [6,]   5 -2.40   0.170
 [7,]   6 -2.39   0.174
 [8,]   7 -2.37   0.180
 [9,]   8 -2.35   0.189
[10,]   9 -2.35   0.189
[11,]  10 -2.34   0.193
[12,]  11 -2.36   0.183
[13,]  12 -2.39   0.173
[14,]  13 -2.38   0.178
[15,]  14 -2.39   0.170
Type 3: with drift and trend
      lag   ADF p.value
 [1,]   0 -2.53   0.354
 [2,]   1 -2.51   0.361
 [3,]   2 -2.44   0.389
 [4,]   3 -2.41   0.405
 [5,]   4 -2.38   0.417
 [6,]   5 -2.37   0.419
 [7,]   6 -2.36   0.424
 [8,]   7 -2.35   0.430
 [9,]   8 -2.32   0.440
[10,]   9 -2.32   0.440
[11,]  10 -2.31   0.445
[12,]  11 -2.33   0.435
[13,]  12 -2.36   0.423
[14,]  13 -2.35   0.429
[15,]  14 -2.37   0.420
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

Figure 34: ADF test for residuals

- Error Correction model
  - The first model is based on Engle-Granger (1987).

```
> ecm1 <- lm(diff(ln_gc) ~ diff(ln_nq) + l1_ect)
> summary(ecm1)

Call:
lm(formula = diff(ln_gc) ~ diff(ln_nq) + l1_ect)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0053071 -0.0001226  0.0000021  0.0001266  0.0063217 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.818e-07  1.743e-06   0.276  0.78223    
diff(ln_nq) 8.133e-02  4.275e-03  19.024  < 2e-16 ***
l1_ect      4.333e-04  1.676e-04   2.585  0.00973 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0002986 on 29341 degrees of freedom
Multiple R-squared:  0.01245,	Adjusted R-squared:  0.01238 
F-statistic: 184.9 on 2 and 29341 DF,  p-value: < 2.2e-16
```

Figure 35: Summary of ECM-1

➢ The second is an alternative and gives us short-run effects of GC reflected in NQ.

```
> ecm2 <- lm(diff(ln_gc) ~ diff(ln_nq) + l1_ln_gc + l1_ln_nq)
> summary(ecm2)

Call:
lm(formula = diff(ln_gc) ~ diff(ln_nq) + l1_ln_gc + l1_ln_nq)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0053036 -0.0001231  0.0000030  0.0001266  0.0063210 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.737e-03  1.191e-03  -1.459  0.14470    
diff(ln_nq)  8.138e-02  4.275e-03  19.036  < 2e-16 ***
l1_ln_gc     4.333e-04  1.676e-04   2.585  0.00974 ** 
l1_ln_nq    -1.633e-04  7.036e-05  -2.320  0.02032 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0002986 on 29340 degrees of freedom
Multiple R-squared:  0.01253,	Adjusted R-squared:  0.01243 
F-statistic: 124.1 on 3 and 29340 DF,  p-value: < 2.2e-16
```

Figure 36: Summary of ECM-2

❊ Auto correlation and Heteroscedasticity
ECM-2 model is free from Auto Correlation.

```
> ncvTest(ecm2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1372.879, Df = 1, p = < 2.22e-16
> bgtest(ecm2)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  ecm2
LM test = 2.0941, df = 1, p-value = 0.1479
```

Figure 37: Auto correlation and Heteroscedasticity tests

❄ Short Run and Long Run equilibrium

```
> ## short term run
> cons_sr_eq <- c1
> cons_sr_eq
diff(ln_nq)
 0.08138454
> ## long run term
> cons_lr_eq <- c2/-c3
> cons_lr_eq
 l1_ln_nq
0.3768115
```

Figure 38: Short Run and Long Run equilibrium

11 Granger Causality Test:

11.1 Description:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \ldots + \gamma_q X_{t-q} + \epsilon_t$$

The Granger causality test assesses whether the past values of one time series can help predict another time series. It tests whether the past values of variable X (X_t) have a statistically significant effect on the current values of variable Y (Y_t), beyond what can be explained by the past values of Y itself. The test is typically formulated using lagged values of the variables, where the null hypothesis is that lagged values of X do not have any additional explanatory power for Y, beyond what is already captured by the lagged values of Y. The alternative hypothesis is that lagged values of X do have additional explanatory power for Y. **11.1.1 Introduction**

:Objective: To investigate the causal relationship between FDI and GDP in Germany. Importance: Understanding this relationship can provide insights into the impact of FDI on economic growth.

11.2

Data Collection:

- Source: World Development Indicators (WDI) database.
- Variables: FDI (BX.KLT.DINV.CD.WD) and GDP (NY.GDP.MKTP.CD).
- Period: 1990-2020.
- Country: Germany (DEU).

	country	iso2c	iso3c	year	BX.KLT.DINV.CD.WD Foreign direct investment, net inflows (BoP, current US$)	NY.GDP.MKTP.CD GDP (current US$)
1	Germany	DE	DEU	1990	2556702846	1.771671e+12
2	Germany	DE	DEU	1991	4741534934	1.868945e+12
3	Germany	DE	DEU	1992	-2137728434	2.131572e+12
4	Germany	DE	DEU	1993	479814189	2.071324e+12
5	Germany	DE	DEU	1994	7517248751	2.205074e+12
6	Germany	DE	DEU	1995	12041505213	2.585792e+12
7	Germany	DE	DEU	1996	15591797829	2.497245e+12
8	Germany	DE	DEU	1997	18638443894	2.211990e+12
9	Germany	DE	DEU	1998	29526509277	2.238991e+12
10	Germany	DE	DEU	1999	86035665071	2.194945e+12

Figure 39: FDI and GDP Dataset for Germany (1990-2020)

11.3 Data Preparation:

- Loaded `WDI` and `lmtest` packages in R.

- ➢ Downloaded and extracted FDI and GDP data for Germany.
- ➢ Created a dataframe (`df`) to store the FDI and GDP data.

11.4 Methodology:

❊ **Stationarity Check**:

- ➢ Conducted Augmented Dickey-Fuller (ADF) tests.

```
> adf_test_fdi

###############################################################
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test #
###############################################################

The value of the test statistic is: -2.4847 3.282

> adf_test_gdp

###############################################################
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test #
###############################################################

The value of the test statistic is: -1.0331 1.5605
```

Figure 40: Stationarity Check by ADF test

- ➢ ADF test for FDI: p-value = 3.282
- ➢ ADF test for GDP: p-value = 1.5605
- ➢ Both variables are stationary at a 5% significance level because both p-values are much greater than 0.05.

11.5 Granger Causality Test:

➢ FDI Granger Causes GDP:

- ➡ Null Hypothesis (Model 2): FDI does not Granger cause GDP
- ➡ Alternative Hypothesis (Model 1): FDI Granger causes GDP.

```
> # Test if FDI Granger causes GDP
> grangertest(GDP ~ FDI, data = df, order = 1)
Granger causality test

Model 1: GDP ~ Lags(GDP, 1:1) + Lags(FDI, 1:1)
Model 2: GDP ~ Lags(GDP, 1:1)
  Res.Df Df      F Pr(>F)
1     27
2     28 -1 0.3175 0.5777
>
```

Figure 41: Testing if FDI Granger causes GDP

- ➡ Result: The F-statistic is 0.3175 with a p-value of 0.5777. Since the p-value is **greater** than the significance level (e.g., 0.05), we fail to reject the null hypothesis. Therefore, there is **no evidence to suggest that lagged values of FDI Granger cause GDP**.

➢ **GDP Granger Causes FDI:**

- ➡ Null Hypothesis (Model 2): GDP does not Granger cause FDI
- ➡ Alternative Hypothesis (Model 1): GDP Granger causes FDI.

```
> # Test if GDP Granger causes FDI
> grangertest(FDI ~ GDP, data = df, order = 1)
Granger causality test

Model 1: FDI ~ Lags(FDI, 1:1) + Lags(GDP, 1:1)
Model 2: FDI ~ Lags(FDI, 1:1)
  Res.Df Df      F Pr(>F)
1     27
2     28 -1 2.0418 0.1645
```

Figure 42: Testing if GDP Granger causes FDI

- ➡ Result: The F-statistic is 2.0418 with a **p-value of 0.1645**. Again, since the p-value is greater than the significance level, we fail to reject the null hypothesis. Therefore, there is **no evidence to suggest that lagged values of GDP Granger cause FDI**.

➢ **Conclusion:** Neither GDP granger cause FDI, nor FDI granger cause GDP incase of Germany.