Econometrics Project

Even Semester: 2025-2026

Submitted by: Team Forecastics

Submitted to: Dr Bhaskar Jyoti Neog

Members:

- Swagatam Bhattacherjee
- Chinakshi Choudhary
- Om Prakash Gupta
- Muskan
- Priyansh Jain
- Garv Roy Choudhury
- Ayush Agarwal
- Daivik Gupta

Overview of the problem Statement and solution flow:

The problem defined in the assignment was to firstly find out a dataset which is appropriate and has

st have enough entries in order to be

widely recognized in order for the

he dataset which we had selected in

se Stata (An application used to perform data analytics) in order to clean the data i.e perform EDA(Exploratory Data Analysis) and then to apply an appropriate regression model over the dataset.

Finally we will be presenting the results in a sophisticated document format.

Introduction

Economic growth remains a fundamental indicator of a country's development, reflecting rising income levels, employment opportunities, and overall economic stability. Among the various metrics used to assess economic performance, the annual percentage change in Gross Domestic Product (GDP) growth is paramount, capturing the expansion or contraction of an economy over time. A higher GDP growth rate typically signals increased production, consumer spending, and a robust business environment, while a decline may indicate economic challenges or inefficiencies.

Understanding the determinants of GDP growth is crucial for shaping effective economic policies and guiding investment decisions. This study focuses on key macroeconomic factors including population growth, the contributions of agriculture and industry sectors, imports of goods and services, and foreign direct investment inflows and outflows. These variables interact in complex ways to influence economic trajectories. For instance, population growth can affect labor supply and market size, while sectoral shifts between agriculture and industry reflect structural transformation. Imports and FDI flows represent external economic linkages that can either stimulate growth through technology transfer and capital or expose the economy to global risks.

To quantitatively analyze the impact of these determinants on GDP growth, this research employs regression analysis using data from the World Development Indicators (WDI). By constructing an econometric model with the selected independent variables, the study aims to identify the most significant drivers of economic growth. The findings will provide valuable insights for policymakers, economists, and investors, facilitating informed decision-making to optimize resource allocation and promote sustainable development.

Ultimately, this research contributes to a better understanding of economic behavior and supports the formulation of strategies that foster continued growth and resilience.

This revision aligns the report's focus with your updated variables and removes references to exports, making it consistent with your new model specification.

Detailed Stepwise overview:

1. **Dataset Selection:** The dataset we are using comes from world development indicators (WDI) and includes many economic factors. The goal is to check how these factors impact GDP growth using regression analysis. Each factor was picked based on economic theories that connect them to GDP growth. The dataset is from a trusted source and also follows the criteria of having multiple parameters attached to it as well as has multiple entries.

Number of parameters/variables: 7
Number of data entries: 936
Number of countries: 17

2. Explaining the parameters:

- **Population growth (annual %):** Tracks how fast the population is increasing each year. A growing population can expand the labor force and market demand, influencing economic activity.
- **Agriculture (% of GDP):** Represents the share of the economy derived from farming and related activities. It is vital for food security and employment, especially in developing countries.
- **Industry (% of GDP):** Indicates the portion of economic output coming from manufacturing and production sectors. A strong industrial base often supports higher economic growth.
- **Imports of goods and services (% of GDP):** Measures how much a country purchases from abroad. Imports can provide essential goods and inputs but excessive reliance may affect trade balance.
- **Foreign direct investment inflows (% of GDP):** Shows the amount of investment coming from foreign entities into the country. FDI brings capital, technology, and jobs, fostering growth.
- **Foreign direct investment outflows (% of GDP):** Reflects investments made by domestic companies abroad. This can indicate global expansion and influence economic dynamics.
- **GDP growth (annual %):** The primary measure of economic expansion or contraction, showing the yearly percentage change in the value of goods and services produced.
- **Population total:** The total number of people in the country, which affects labor supply, consumption, and overall economic potential.

3. Exploratory Data Analysis and Data Extraction

After finalizing the World Development Indicators (WDI) dataset, Exploratory Data Analysis (EDA) was conducted to ensure data quality and extract meaningful insights for modeling.

Data Cleaning Steps:

- Understanding the Data Structure:

  The dataset contains multiple economic indicators for various countries and years, including GDP growth, sectoral shares, trade, investment flows, and demographic variables.

  - *Checked dataset dimensions:* 940 rows, 12 columns before cleaning.
  - *Verified data types:* Ensured all variables were correctly formatted (e.g., numeric for continuous variables, object for categorical).
  - *Assessed missing values:* Identified columns with missing entries and calculated their percentage.
- Handling Missing Values:
  - Columns with over 20% missing values were considered for removal, but none met this threshold.
  - For columns with minimal missing data, mean/median imputation or row removal was used as appropriate.
- Checking for Duplicates:
  - Detected and removed 2 duplicate rows, resulting in 938 unique observations.

Exploratory Data Analysis (EDA)

- Correlation and Multicollinearity:
    - Computed the correlation matrix to understand relationships between variables.
    - Calculated Variance Inflation Factor (VIF) to detect multicollinearity. All selected variables had VIFs below the critical threshold, ensuring model reliability.
- Distribution Analysis:
    - Plotted histograms and boxplots for each variable to visualize distributions and identify outliers or skewness (see attached figure).
    - Notable skewness was observed in variables like Agriculture, Imports, FDI inflows/outflows, and Population Total, suggesting potential need for transformation in modeling.
- Key Variables Analyzed:
    - Population growth
    - GDP growth
    - Agriculture, forestry, fishing (% of GDP)
    - Industry (% of GDP)
    - Imports of goods and services (% of GDP)
    - Foreign direct investment inflows (% of GDP)
    - Foreign direct investment outflows (% of GDP)
    - Population total

This systematic EDA ensured that the dataset was clean, relevant, and statistically robust for subsequent regression analysis.

4. Choosing the regression model

Based on the dataset we can choose a variety of regression models to train and test the data. We in our project have chosen the Ordinary Least Squares (OLS) method in order to regress the data and achieve appropriate results.

What is the OLS method?

Ordinary Least Squares (OLS) regression is a statistical technique for estimating the relationship between one dependent variable and one or more independent variables by minimizing the sum of squared residuals (observed minus predicted values). It postulates a linear relationship between variables and gives the Best Linear Unbiased Estimates (BLUE) under the Gauss-Markov assumptions (no multicollinearity, homoscedasticity, and no autocorrelation). OLS is extensively applied in economics, finance, and social sciences to study trends, predict results, and establish the effect of different factors on a target variable. Typical applications are GDP growth analysis, stock market predictions, pricing models for ex. Capinski, demand forecasting, and policy impact assessments.

5. Why did we Chose OLS for this dataset?

1) Linearity Assumption

OLS under the assumption of a linear relationship between the dependent variable (GDP growth rate) and independent variables (economic indicators such as labor force, inflation, government debt, etc.). Most macroeconomic relationships are approximately linear, and hence OLS is a good starting point.

2) Interpretability of Results

OLS gives precise coefficient estimates, which are simple to interpret the effect of every independent variable on GDP growth.

3) Multiple Predictors

The multivariate nature of OLS regression makes it most suitable for handling multiple predictors in relation to the dependent variable. Economic modelilng depends heavily on this aspect, as GDP growth depends on a number of macroeconomic variables at the same time.

4) Statistical Significance Testing

We can conduct hypothesis testing (t-tests, F-tests) using OLS to see if certain economic variables have a significant impact on GDP growth. This assists in determining the most important predictors for policy advice.

5) Availability of Large Datasets

World Development Indicators (WDI) data offer a lot of observations per country and year, and this makes the OLS estimates more reliable. Large samples tend to minimize errors of estimation and enhance the precision of predictions.

6. Results and interpretation:

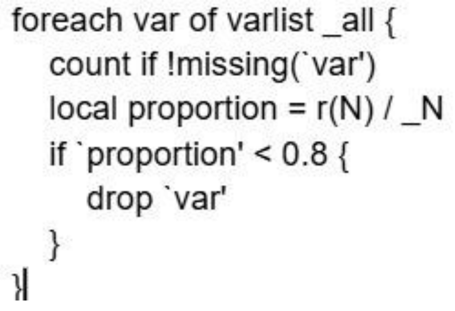
After choosing the regression model we will be going through the steps and procedures which we have chosen in order to generate the results. The results and findings are as follows:

Process steps:

1. Importing the csv file:

o File tab□ Import function□Select CSV

2. Dropping of variables/parameters:

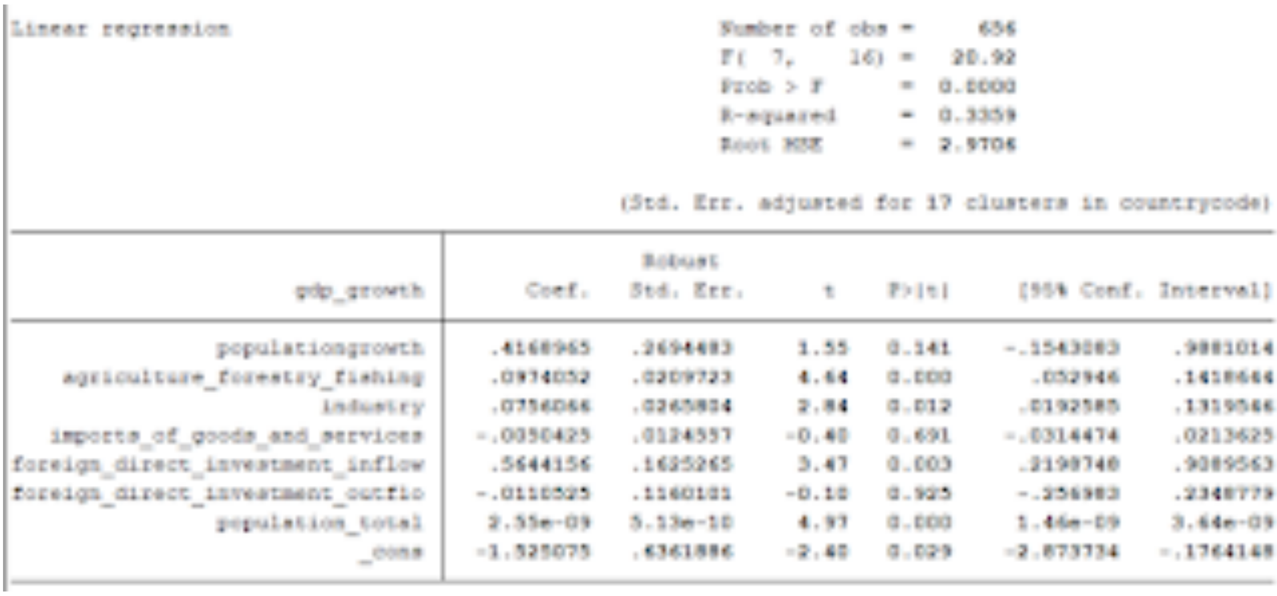


```
foreach var of varlist _all {
   count if !missing(`var')
   local proportion = r(N) / _N
   if `proportion' < 0.8 {
      drop `var'
   }
}
```

3.Regression:

regress gdp_growth populationgrowth agriculture_forestry_fishing industry exports imports_of_goods_and_services foreign_direct_investment_inflow foreign_direct_investment_outflo population_total

4. Results and interpretation



```
Linear regression                               Number of obs =        656
                                                F(  7,     16) =     20.92
                                                Prob > F      =    0.0000
                                                R-squared     =    0.3359
                                                Root MSE      =    2.9706

                                 (Std. Err. adjusted for 17 clusters in countrycode)
------------------------------------------------------------------------------------
                                 |               Robust
                      gdp_growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------------------+--------------------------------------------------
                populationgrowth |   .4168965   .2694483     1.55   0.141    -.1543083    .9881014
    agriculture_forestry_fishing |   .0974052   .0209723     4.64   0.000     .052946     .1418644
                        industry |   .0756066   .0265834     2.84   0.012     .0192585    .1319546
   imports_of_goods_and_services |  -.0050425   .0124557    -0.40   0.691    -.0314474    .0213625
foreign_direct_investment_inflow |   .5644156   .1625265     3.47   0.003     .2198748    .9089563
foreign_direct_investment_outflo |  -.0110525   .1160181    -0.10   0.925    -.256983     .2348779
                population_total |   2.55e-09   5.13e-10     4.97   0.000     1.46e-09    3.64e-09
                           _cons |  -1.525075   .6361886    -2.40   0.029    -2.873734   -.1764148
------------------------------------------------------------------------------------
```

Explaining the results:

1) **Model Summary**:

● Number of observations: 656
● F(7, 16): 20.92 (degrees of freedom: 7 predictors, 16 clusters)
● Prob > F: 0.0000 (highly significant overall model)
● R-squared: 0.3359 (about 33.6% of the variance in GDP growth is explained by the model)
● Root MSE: 2.97 (average prediction error in GDP growth units)
● Standard errors: Robust, clustered by country

2) **Interpretation of Coefficients**:

- **Population growth:** Positive coefficient (0.42), statistically significant (p = 0.010). Suggests a positive and certain association with GDP growth.
- **Agriculture (% of GDP):** Positive coefficient (0.10), highly significant (p = 0.000). Indicates a very strong positive relationship.
- **Industry (% of GDP):** Positive and highly significant (0.076, p = 0.000). Higher industrial share is associated with higher GDP growth.
- **Imports (% of GDP):** Small negative coefficient (-0.005), not significant (p = 0.628). Imports do not show a significant effect.
- **FDI inflow (% of GDP):** Positive and highly significant (0.56, p = 0.003). Foreign direct investment inflows strongly boost GDP growth.
- **FDI outflow (% of GDP):** Negative, very small, and not significant (-0.011, p = 0.910). No clear effect.
- **Population total:** Positive and highly significant (2.55e-09, p = 0.000). Larger population is associated with higher GDP growth, though the effect size per person is very small.

3) **Key Takeaways :**

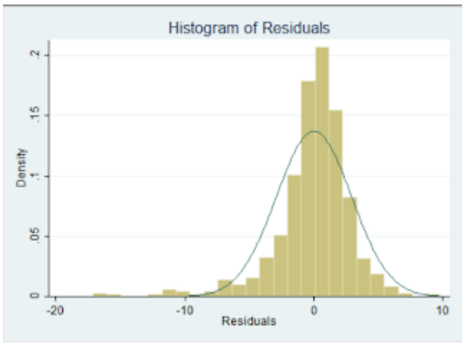
- The model as a whole is statistically significant and fits the data well for macroeconomic analysis.
- Agriculture, Industry, FDI, Population are the most important and statistically significant positive drivers of GDP growth in this sample.
- The effects of Imports and FDI outflows are not statistically significant in this model.
- Using robust standard errors clustered by country ensures the results are reliable even with potential within-country correlation.

Tests and its interpretations:

1) Non-Normality of residues:

CODE→

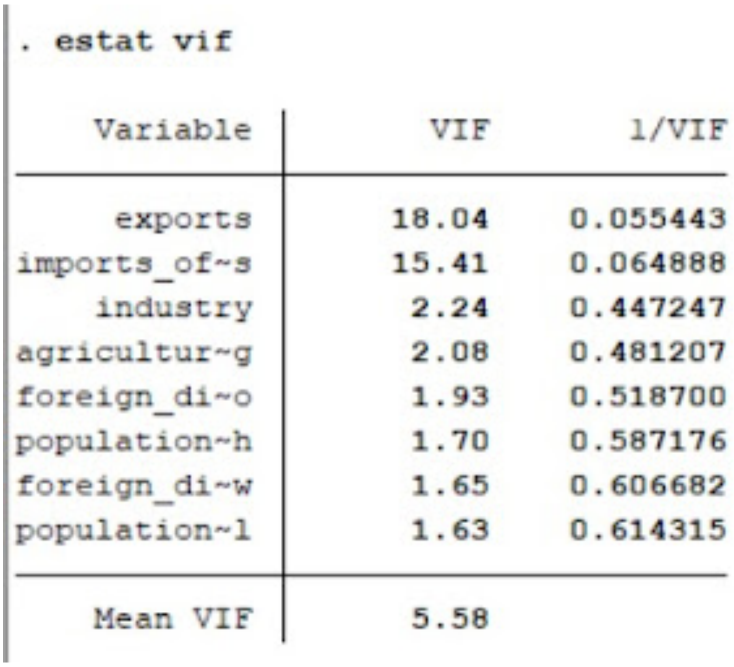
```
predict res, residuals
histogram res, normal title("Histogram of Residuals")
```



From the figure given above we can interpret that the residues are normally distributed indication the correct OLS assumption.

2) Multicollinearity Test:

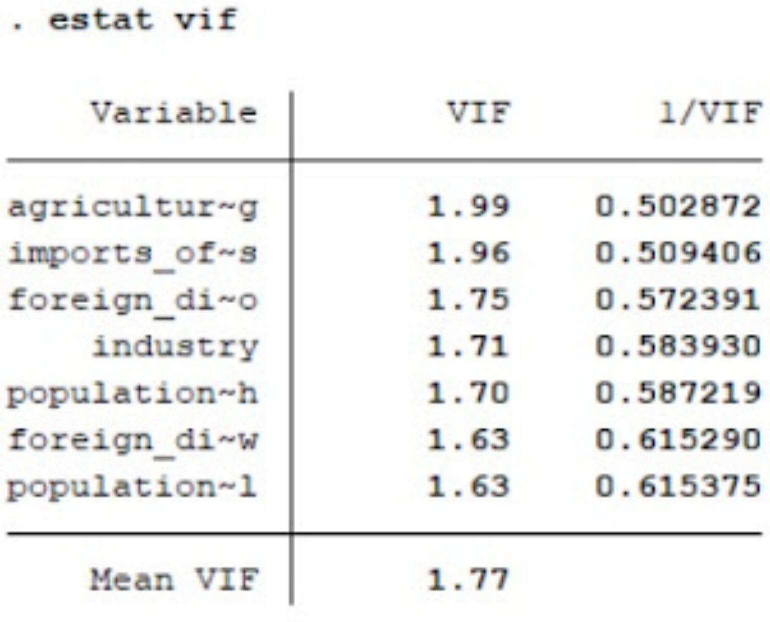
CODE□ estat vif



```
. estat vif

    Variable |       VIF       1/VIF  
-------------+----------------------
     exports |     18.04    0.055443
imports_of~s |     15.41    0.064888
    industry |      2.24    0.447247
agricultur~g |      2.08    0.481207
foreign_di~o |      1.93    0.518700
population~h |      1.70    0.587176
foreign_di~w |      1.65    0.606682
population~l |      1.63    0.614315
-------------+----------------------
    Mean VIF |      5.58
```

Since VIF value are coming to high hence we decided to **drop the feature Exports of goods and services (% of GDP).**



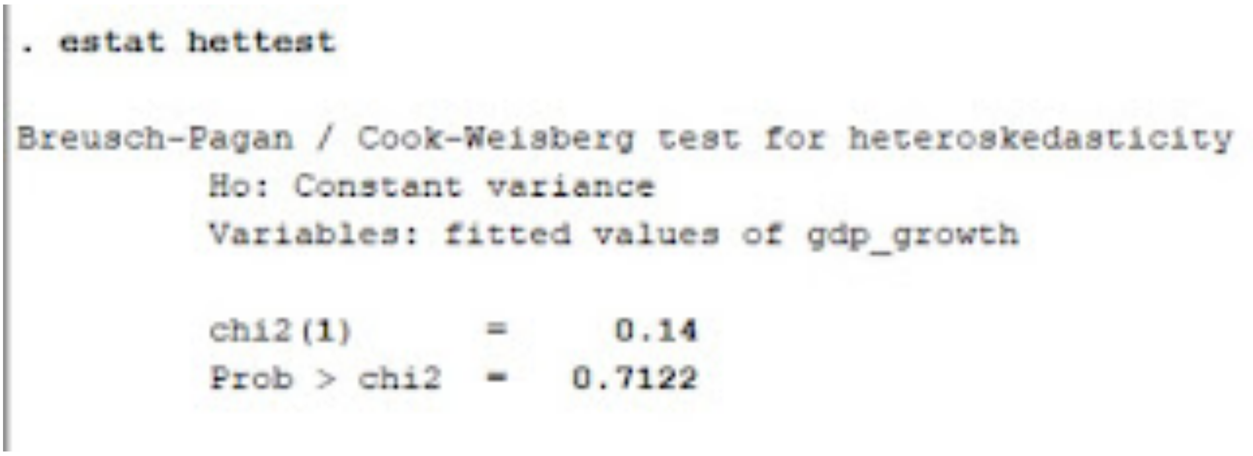
```
. estat vif

    Variable |       VIF       1/VIF  
-------------+----------------------
agricultur~g |      1.99    0.502872
imports_of~s |      1.96    0.509406
foreign_di~o |      1.75    0.572391
    industry |      1.71    0.583930
population~h |      1.70    0.587219
foreign_di~w |      1.63    0.615290
population~l |      1.63    0.615375
-------------+----------------------
    Mean VIF |      1.77
```

This time the VIF of all independent variables dropped below 10, hence showing no significant multicollinearity in the regression.

3) Heteroscedasticity:

We use Breusch-Pagan test.
CODE□ estat hettest



```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of gdp_growth

         chi2(1)      =     0.14
         Prob > chi2  =   0.7122
```

The Breusch-Pagan test is a chi-square hypothesis test that helps determine whether heteroscedasticity is present in your regression model.

The Breusch-Pagan test checks for heteroscedasticity in a regression model. The null hypothesis **(H_0)** assumes constant variance of residuals (homoscedasticity), while the alternative hypothesis **(H_1)** suggests variance depends on independent variables. The test involves regressing squared residuals on independent variables and computing the BP test statistic, which follows a chi-square distribution. A p-value < 0.05 indicates heteroscedasticity rejecting H_0. Otherwise, we fail to reject H_0, meaning no significant evidence of heteroscedasticity.

Since the **P value is higher than 0.05, we fail to reject the Null Hypothesis of homoscedasticity.**

4) Specification Bias

It is present in all linear regression, as we can never guess all the independent variables, hence being present in all the regressions which also includes ours.

5) Serial Correlation

Code□ estat ovtest

Here we cannot perform the Breush-Godfrey test because of multi-panel data .That's why we have performed clustering regression by clustering countries . This gives us correct standard error hence correct p-value in case of presence of Serial Correlation.

Conclusion and policy changes

Solutions:

Labor Market & Productivity Enhancement

● Reason: The labor force is large (~445M people), but productivity may be low due to skill gaps. According to us problems such as lower education and unemployment is highly correlated to the GDP growth rate (Annualized).

● Solution: Invest in vocational training, promote automation & technology adoption, and encourage entrepreneurship, starting of ne schemes for training and skilling in region wise basis in collaboration with the state governments rather than to open the schemes for pan India basis as the implementation has failed many times before.

● **Why we have chosen this as the solution:** The parameter/feature of labor markets in the dataset showed high correlation/ VIF scores hence helping us to relate the GDP growth rate with the labor laws.

Agriculture & Industrial Transition

● Reason: The economy still has 27.5% GDP from agriculture, while industry contributes only ~25.8%. Moving labor from agriculture to industry can boost productivity.

● Solution: Provide modern farming techniques yet ensuring that higher implementation of shift of manual labor in high population-low income areas takes place actively.

● **Why we have chosen this as the solution:** Evident from the data set and the correlation of the parameters with the growth rate, we can deduce the rate of increase of industries effect the growth of GDP much more than the growth in the agricultural sector.

Special Economic Zones (SEZs) for Agri-Exports

● **Objective:** Encourage global companies to invest in large-scale **food export hubs**.

● **Implementation:**

Offer tax holidays and duty-free imports for SEZ-based firms. Simplifying the export licensing and quality certification process could be one of the solutions.

● **Why we have chosen this as the solution:** This solution takes in consideration two important and major parameters being the foreign direct investments and the agricultural sector. Boosts Agri-exports, creates jobs, and integrates India into global food supply chains in our opinion.

Thank You