Econometric Modelling for Economic Insights: Predicting Gross National Income Using Socioeconomic indicators about Financial Inclusion.

Indian Institute of Foreign Trade.

Ahan Mazumdar IIFT roll- 6

Ayush Pal IIFT roll-20

1) Problem Statement –

The Gross National Income (GNI) is a broad measure of the economy of a nation, measuring the total income of all the resident citizens and businesses inside and outside the country's territory. The accurate forecasting of GNI is very important for policymakers, economists, and researchers etc., who are engaged in preparing good guidelines for economic development; but the relationship between GNI and influencing variables is rather complicated, thereby necessitating advanced statistical techniques to facilitate meaningful analysis.

This study aims to investigate whether financial inclusion indicators, such as Gender based inclusion, labor force participation in financial systems, account ownership by income groups, and account access among less-educated populations, can significantly predict the Gross National Income (GNI) of a country using econometric analysis.

Null Hypothesis (H0): $\beta1=\beta2=\beta3=\ldots=0$ (Financial inclusion indicators have no significant effect on GNI).

Alternative Hypothesis (H1): At least one **$\beta i\neq0$** (At least one financial inclusion indicator significantly affects GNI).

This project is focused on the application of regression techniques and on easily predicting GNI based on a small set of strong indicators, which are:

- **Financial Inclusion Indicators:** The percentage of people with accounts opened, disaggregated by their gender, income, and labor force participation.
- **Labor Force Dynamics:** Including male and female status of participation across various economic activities.
- **Income:** Representation of the richest 60% and poorest 40% of the population.
- **Level of Educational Attainment**: Primary or less compared to secondary or more education level.

2 – Dataset

2.1 - Data Source - World Bank Link

Data Processing Steps:

1. **Data Cleaning:**
    - Ensured consistency in column names and data formats.
    - Removed duplicates and handled erroneous entries to maintain data integrity.
2. **Feature Scaling:**
    - Applied **StandardScaler** from **sklearn.preprocessing** to standardize the data, ensuring all variables have a mean of 0 and a standard deviation of 1. This step was crucial for optimizing the performance of the regression model.
3. **Handling Missing Values:**
    - Used **K-Nearest Neighbours (KNN)** imputation to address missing values effectively. This method predicts and fills in missing data points based on the values of the nearest neighbours, preserving the data's overall structure and variance.

2.2 – Variable Description -

Variable Name	Type	Units of Measurement	Variable Description	Role in Analysis
Gross National Income (GNI)	Continuous	Current USD	Total income earned by a nation's residents and businesses, including any income earned abroad, over a specific period, usually a year.	Dependent
Account, male (% age 15+)	Continuous	Percentage (%)	Proportion of males aged 15 and above who have an account at a bank or other financial institution or use a mobile money service.	Independent
Account, in labor force (% age 15+)	Continuous	Percentage (%)	Percentage of individuals aged 15 and above, participating in the labor force, who have an account at a financial institution or use mobile money services.	Independent
Account, out of labor force (% age 15+)	Continuous	Percentage (%)	Percentage of individuals aged 15 and above, not participating in the labor force, who have an account at a	Independent

Variable Name	Type	Units of Measurement	Variable Description	Role in Analysis
			financial institution or use mobile money services.	
Account, income, poorest 40% (% ages 15+)	Continuous	Percentage (%)	Proportion of adults aged 15 and above in the poorest 40% of households who have an account at a bank or other financial institution or use a mobile money service.	Independent
Account, income, richest 60% (% ages 15+)	Continuous	Percentage (%)	Proportion of adults aged 15 and above in the richest 60% of households who have an account at a bank or other financial institution or use a mobile money service.	Independent
Account, primary education or less (% age 15+)	Continuous	Percentage (%)	Percentage of individuals aged 15 and above with primary education or less who have an account at a financial institution or use mobile money services.	Independent
Account, secondary education or more (% age 15+)	Continuous	Percentage (%)	Percentage of individuals aged 15 and above with secondary education or more who have an account at a financial institution or use mobile money services.	Independent
Account female (% age 15+)	Continuous	Percentage (%)	Proportion of females aged 15 and above who have an account at a bank or other financial institution or use a mobile money service.	Independent

Total Observations –

3- Data Exploration

	GNI(Y)	Account, female (% age 15+)	Account, in labor force (% age 15+)	Account, income, poorest 40% (% ages 15+)	Account, income, richest 60% (% ages 15+)	Account, male (% age 15+)	Account, out of labor force (% age 15+)	Account, primary education or less (% ages 15+)	Account, secondary education or more (% ages 15+)
0	16,106.34	2.6	15.0	1.1	14.2	15.4	2.1	5.3	30.0
1	76,665.53	22.7	36.1	17.5	35.4	33.7	14.9	15.1	36.2
2	20,564.53	20.4	51.1	23.1	40.5	46.1	17.7	30.1	40.3
3	52,302.72	38.9	45.8	..	..	39.5	30.9	34.8	44.4
4	33,168.56	13.9	33.2	16.4	26.7	30.7	11.5	14.2	31.0
5	37,214.13	31.8	36.8	19.7	42.1	34.6	27.6	28.3	36.9
6	79,579.38	18.2	23.9	15.0	19.1	16.7	11.1	7.0	20.3
7	32,337,027.96	98.6	99.7	98.0	99.8	99.6	97.8	100.0	99.2
8	26,860.48	96.6	97.8	94.7	98.6	97.7	96.0	88.9	98.3
9	4,252,085.48	14.3	16.8	8.2	19.4	15.6	12.3	8.7	16.9

3.1 - Descriptive Statistics –

3.2 – Key Insights -

1. GNI (Gross National Income, YYY)

- **Trend**: The GNI varies significantly across observations
    - Example: The lowest GNI is 16,106.34, and the highest is 32,337,027.96
- **Insight**: There is substantial variation in economic prosperity across observations, as measured by GNI.

---

2. Account Ownership: Female (% age 15+)

- **Trend**: Account ownership among females shows significant variation.
    - The lowest rate is 2.6%, while the highest is 98.6%.
- **Insight**: Financial inclusion for females is extremely low in some regions (e.g., 2.6%), indicating gender disparity, while in others (e.g., 98.6%), it is near universal.

---

3. Account Ownership: In Labor Force (% age 15+)

- **Trend**: The percentage of account ownership among individuals in the labor force also varies widely.
    - The lowest is 15.0%, and the highest is 99.7%.

- **Insight**: Higher labor force participation often correlates with higher account ownership, though exceptions exist (e.g., Row 6 has only 23.9%.

---

4. Account Ownership: Income Poorest 40% (% age 15+)

- **Trend**: The poorest 40% exhibit low account ownership in many cases.
    - The lowest rate is 1.1%, and the highest is 98.0%.
- **Insight**: Financial exclusion is severe among the poorest in some regions, but there are cases of financial inclusion reaching near universality.

---

5. Account Ownership: Income Richest 60% (% age 15+)

- **Trend**: Account ownership among the richest 60% is consistently higher than the poorest 40%, but variation exists.
    - The lowest rate is 14.2%, and the highest is 99.9%.
- **Insight**: Financial inclusion is skewed towards the wealthier populations, emphasizing income-based inequality in financial access.

---

6. Account Ownership: Male (% age 15+)

- **Trend**: Account ownership among males also varies widely across observations.
    - The lowest is 15.4%, and the highest is 99.6%.
- **Insight**: Similar to females, account ownership among males ranges from low to near universal, showing both disparities and progress.

---

7. Account Ownership: Out of Labor Force (% age 15+)

- **Trend**: Account ownership for individuals out of the labor force shows a stark contrast.
    - The lowest rate is 0.0%, and the highest is 97.8%.
- **Insight**: This variable captures financial inclusion among non-working populations and shows cases where financial inclusion is either completely absent or highly advanced.

---

8. Account Ownership: Primary Education or Less (% age 15+)

- **Trend**: Individuals with primary education or less have lower account ownership.

- The lowest rate is 5.3, and the highest is 100.0%

- **Insight**: Education level strongly correlates with account ownership, as those with less education are generally excluded.

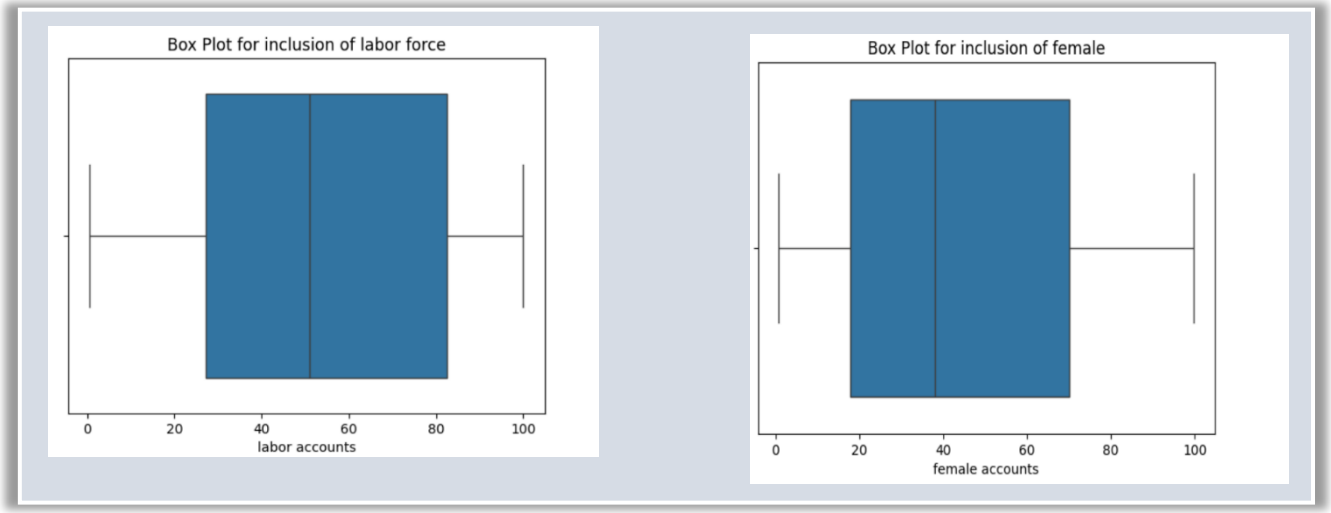
---

9. Account Ownership: Secondary Education or More (% age 15+)

- **Trend**: Individuals with secondary education or more show consistently higher account ownership.
    - The lowest rate is 16.9%, and the highest is 99.2%.
- **Insight**: Higher education levels are a strong predictor of financial inclusion, as this group consistently achieves high rates of account ownership.

Graphical Analysis -

Box Plot Analysis –



Box plots for two critical variables, "Inclusion of Labor Force" and "Inclusion of Females," provide insights into the variability and central tendency of financial inclusion across regions.

1. **Inclusion of Labor Force**:
    - The plot demonstrates a wide range (0% to 100%), indicating variability in access to financial accounts among the labor force across different countries.
    - The central tendency is balanced, with the median centrally located in the distribution, and no outliers suggest consistent data.
    - The broad interquartile range (IQR) reflects disparities in access within this demographic.
2. **Inclusion of Females**:

- The distribution reveals significant variability, with the median slightly below the center of the box, indicating relatively lower financial inclusion for half the dataset.
- Like the labor force data, this variable covers a broad range (0% to 100%), suggesting wide disparities across countries.
- The wide IQR further emphasizes these disparities in financial access among females.

4) Linearity checks: CCPR Plot

Detailed Analysis by Category -

1. **Educational Indicators -**

a) Secondary Education (15+ years)

- Moderate negative linear relationship
- Notable scatter around trend line
- Declining trend with increasing education percentage

b) Primary Education (15+ years)

- Strong positive linear relationship
- Consistent point distribution
- Clear upward trend with minimal deviation

2. **Labor Force Indicators**

a) Labor Force Participation (15+ years)

- Strong positive linear relationship
- Well-defined upward trend
- Tight clustering around trend line

b) Out of Labor Force (15+ years)

- Weak negative linear relationship
- Significant scatter
- Less predictable pattern

3. **Gender-Based Analysis -**

a) Male Account Holders (15+ years)

- Negative linear relationship

- Considerable variation from trend line
- Inconsistent pattern distribution

b) Female Account Holders (15+ years)

- Negative linear relationship
- Significant scatter
- Similar pattern to male holders but with different slope

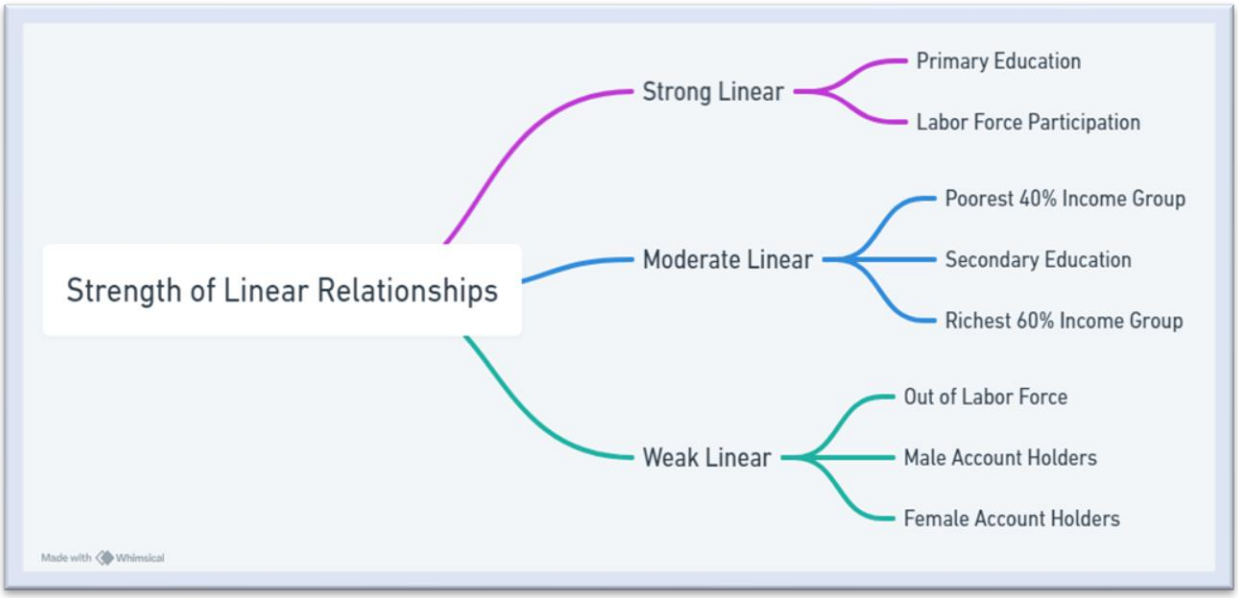
4. Income-Based Analysis

a) Richest 60% (15+ years)

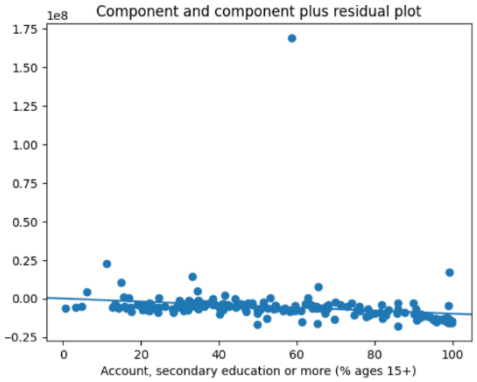
- Clear negative linear relationship
- Moderate scatter around trend line
- Consistent downward pattern

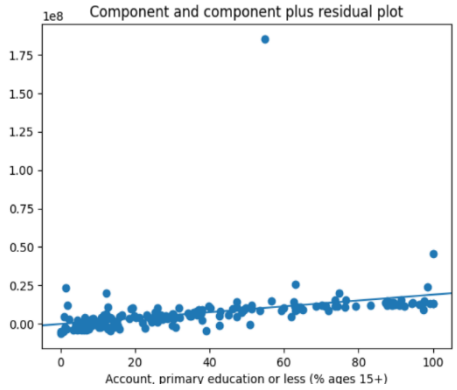
b) Poorest 40% (15+ years)

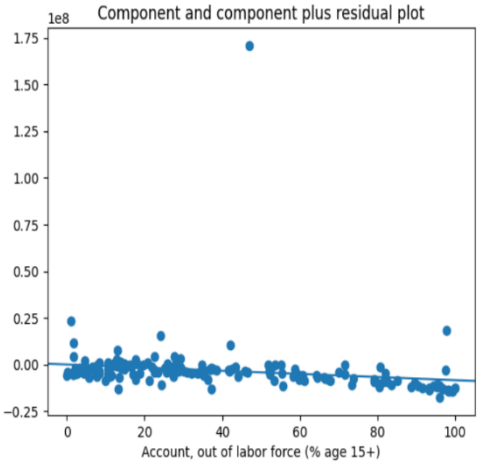
- Moderate positive linear relationship
- Relatively consistent upward trend
- Better predictability than higher income segment

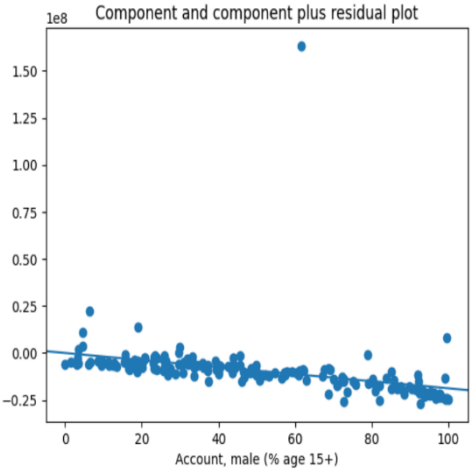


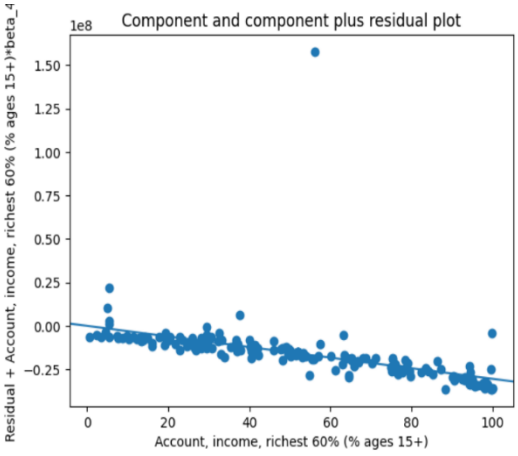
Component – Component Plus Residual Plot

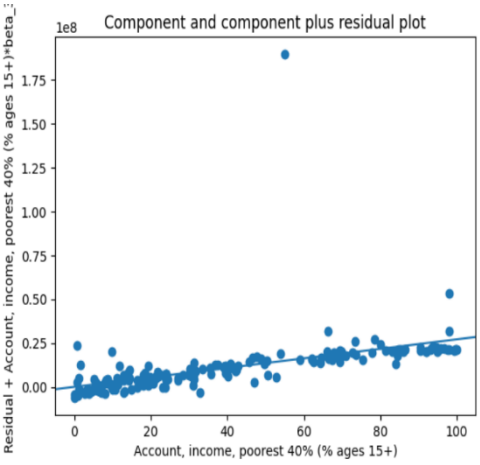


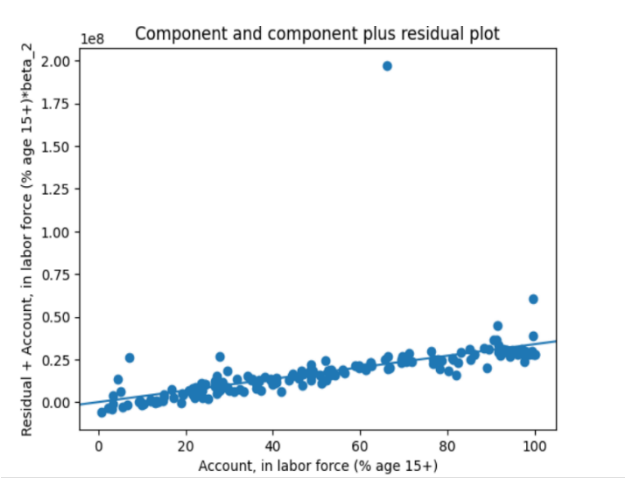


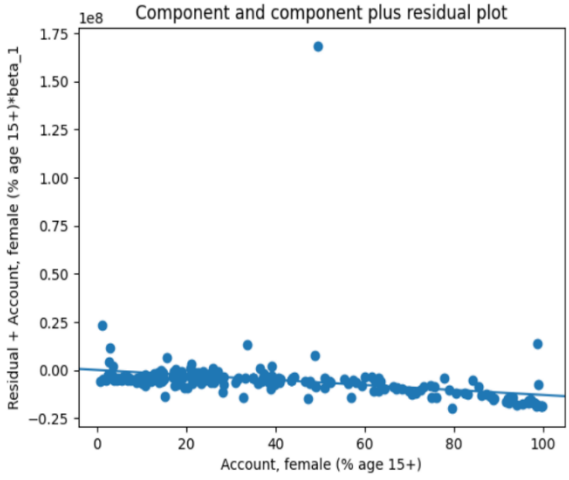




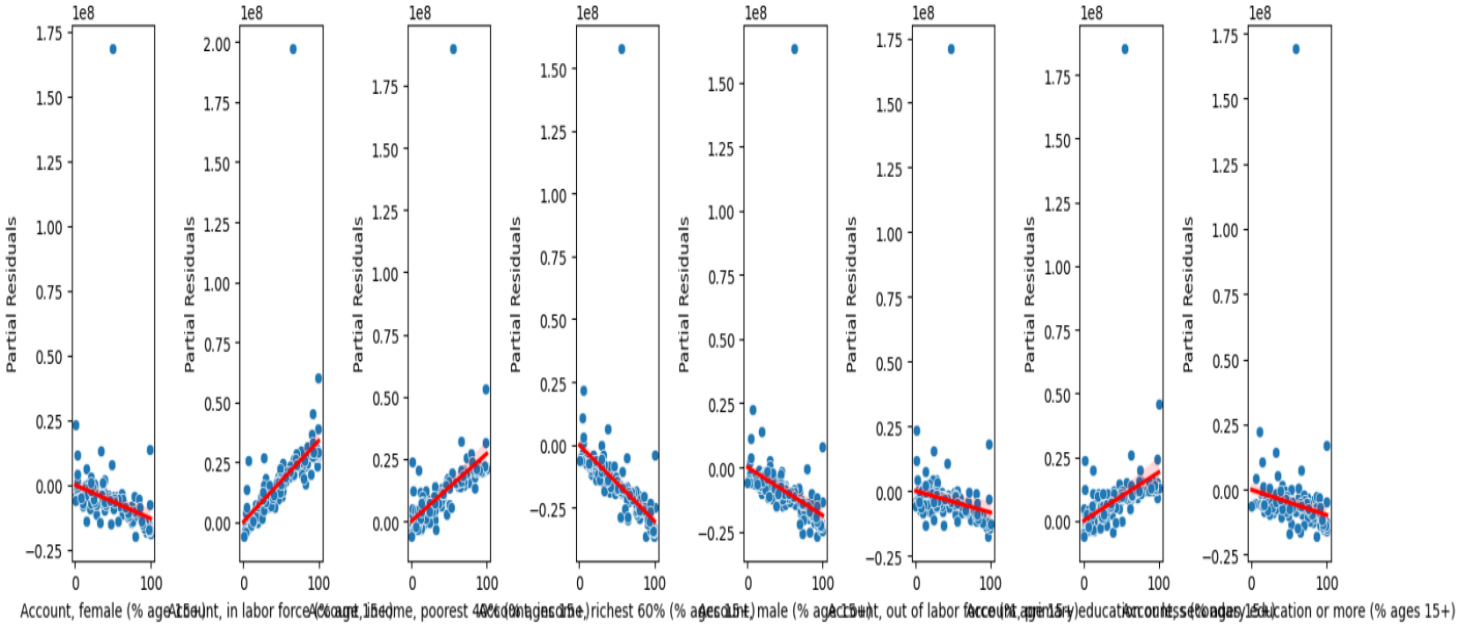








Partial Residual Plot -



(For clearer view visit the ipynb file)

Detailed analysis of heteroskedasticity patterns observed across eight financial inclusion indicators.

Gender-Based Financial Indicators

1. Female Account Holders
    - Pronounced fan-shaped pattern
    - Higher variability in lower percentages
    - Systematic decrease in spread

2. Male Account Holders
    - Similar pattern to female accounts
    - Notable variance reduction at higher percentages
    - Requires similar statistical treatment

Educational Parameters

1. Secondary Education
    - Complex heteroskedastic structure
    - Variable spread throughout range
    - Higher dispersion in middle ranges
    - Significant modeling implications
2. Primary Education
    - Reverse funnel pattern
    - Increasing spread with education levels
    - Systematic variance change
    - Notable statistical considerations

Income-Based Analysis

1. Richest 60% Segment
    - Moderate to strong pattern
    - Funnel-shaped variance structure
    - Clear systematic changes
    - Notable implications for modeling
2. Poorest 40% Segment
    - Milder heteroskedastic pattern
    - More uniform spread
    - Slight variance increases
    - More stable statistical properties

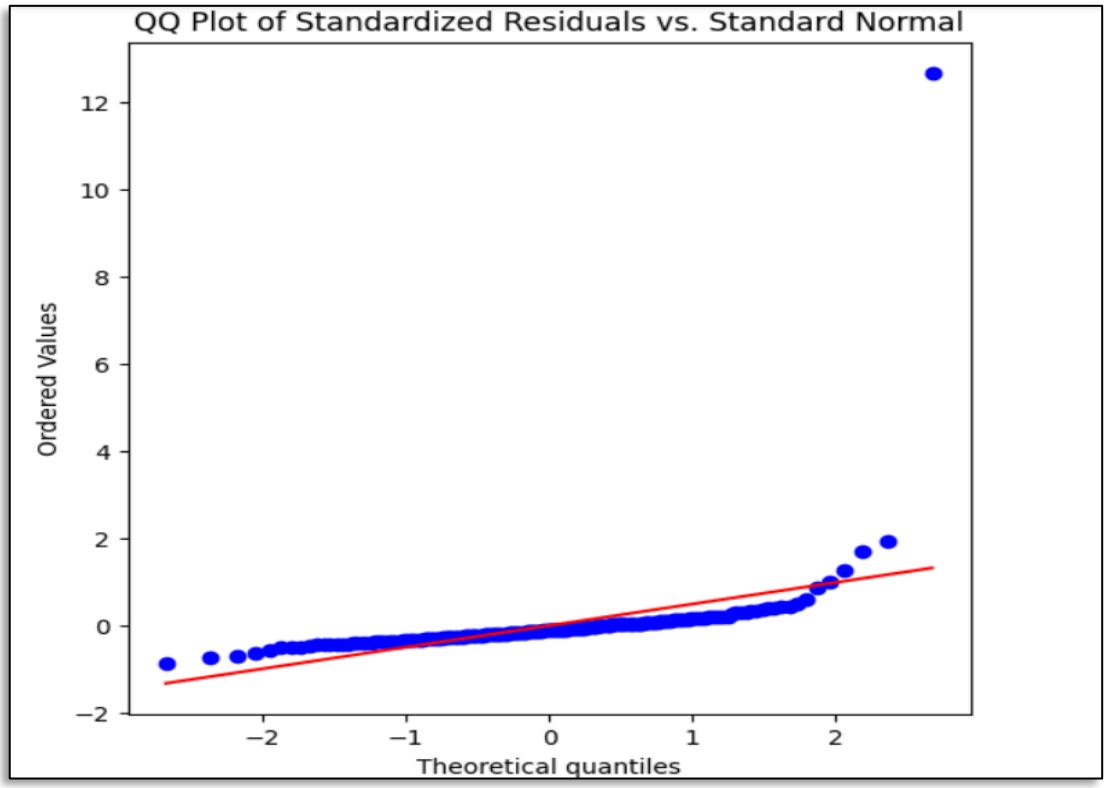
Labor Force Indicators

1. Labor Force Participation
    - Wedge-shaped pattern
    - Increasing spread trend

- Systematic variance changes
- Moderate statistical impact

2. Out of Labor Force
    - Moderate pattern
    - Less systematic variation
    - Clustered variance changes
    - Requires careful interpretation

5) Normality Checks – QQ Plot

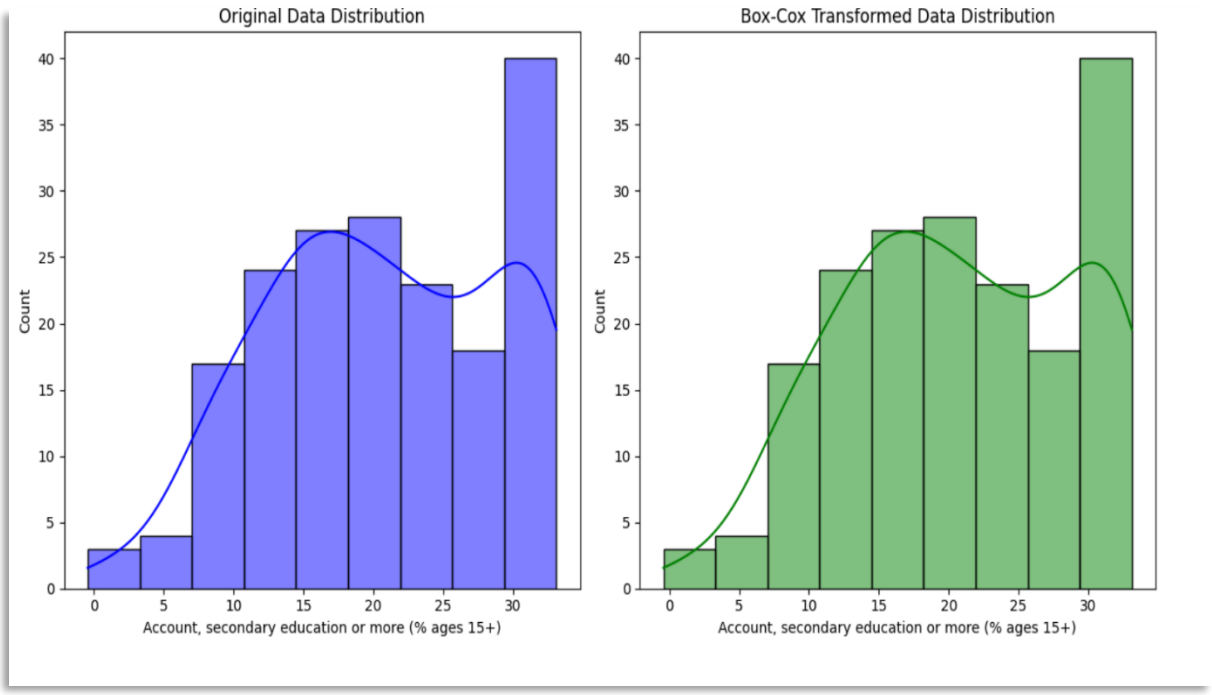


The **normality** assumption was evaluated through a quantile-quantile **(Q-Q)** plot analysis of standardized residuals against theoretical normal quantiles. The graphical assessment revealed intriguing patterns in the data distribution that merit careful consideration.

The central portion of the distribution, spanning approximately between -1 and 1 theoretical quantiles, demonstrated reasonable adherence to normality, as evidenced by points closely following the theoretical reference line. This suggests that the core of our data maintains properties consistent with a normal distribution. However, the tails of the distribution presented noteworthy deviations that warrant discussion.

Of particular interest is the behaviour observed in the upper tail of the distribution. The plot revealed a substantial outlier at approximately 12 standard deviations above the mean, significantly departing from the expected normal pattern. This extreme value, coupled with several additional points showing upward deviation between 1.5-2 standard deviations, indicates a pronounced right-skewed tendency in the data. The lower tail, while showing more modest departures, also exhibited slight deviations below the theoretical normal line.

Box-Cox Transformation – The box cox transformation was applied in the independent variable of secondary education as slight heteroskedastic variance structure was detected.



Lambda value used for Box-Cox Transformation: 0.6884348440365681

A Box-Cox transformation was applied to address the distributional issues identified in the initial analysis, using an optimal lambda value of 0.688. The transformation results are visualized in Figure X, which presents a comparative view of the original and transformed distributions of secondary education account holders (ages 15+).

The original distribution (left panel) exhibits several notable characteristics:

- A pronounced right-skewed pattern
- Multiple peaks in the distribution, suggesting potential multimodality
- A substantial concentration of values in the 30% range

- Irregular spacing between frequency bars, indicating potential distributional asymmetry

Post Box-Cox transformation (right panel), while maintaining the general structure of the data, shows some improvements:

- The overall shape of the distribution remains similar, indicating preservation of fundamental data patterns
- The lambda value of 0.688, being between 0 and 1, suggests a transformation milder than a logarithmic transformation but stronger than a square root
- The transformation appears to have slightly moderated the extreme values while maintaining the essential multimodal nature of the distribution

---

Initial Tests on Original data –

1 – Shapiro-Wilk Test

- Shapiro-Wilk Test Statistic: 0.2546795839286645
- P-value: 1.1028354781734622e-26
- The residuals do not follow a normal distribution (reject H0)

Analysis –

This test is particularly powerful for detecting departures from normality across the full range of the distribution. The analysis yielded a test statistic **(W) of 0.255** with an associated **p-value of 1.103e-26**, which is substantially below the conventional significance level of 0.05.

The extremely low p-value provides strong statistical evidence to reject the null hypothesis of normality. The test statistic of 0.255, being considerably lower than 1 (the value indicating perfect normality), further supports the conclusion that the residuals significantly deviate from a normal distribution.

This finding aligns with the visual evidence observed in the previously discussed Q-Q plot, where notable deviations from normality were identified, particularly in the tail regions. The combination of both graphical and formal statistical testing provides robust evidence that the underlying distribution of the data significantly departs from normality.

Tests After Yeo-Johnson Transformation

1 - Shapiro-Wilk Test on Transformed Data

- Shapiro-Wilk Test Statistic: 0.8397703753711602

- P-value: 5.96543802548224e-13
- The Yeo-Johnson transformed residuals do not follow a normal distribution (reject H0)

The examination of residual distributions before and after the Yeo-Johnson transformation reveals several noteworthy findings. The original residuals exhibited a highly skewed distribution with a pronounced right tail, indicating substantial departure from normality. To address this non-normality, a Yeo-Johnson transformation was applied, which is particularly suitable for datasets that include both positive and negative values.

Despite the transformation effort, the Shapiro-Wilk test results ($W = 0.840$, $p < 0.001$) provide strong evidence to reject the null hypothesis of normality. This conclusion is supported by both statistical and visual evidence:

Statistical Evidence: The extremely low p-value (5.97e-13) indicates that the probability of observing such data under the assumption of normality is negligible. This suggests that even after transformation, significant departures from normality persist.

Visual Assessment: The comparative histograms reveal that while the Yeo-Johnson transformation achieved some improvement in the distribution's shape, creating a more centralized pattern, notable deviations from the ideal normal distribution remain evident. The transformed distribution exhibits:

- A more symmetric central tendency
- Reduced but still present tail effects
- Some evidence of multimodality in the central region

Bootstrapped Distribution of Mean -

Results Summary:

The bootstrap analysis yielded highly favourable results with a Shapiro-Wilk test statistic of 0.998 ($p = 0.477$), demonstrating robust normality characteristics in the sampling distribution.

Detailed Analysis:

The bootstrap distribution of the mean reveals several critical characteristics that validate our statistical approach:

Distribution Profile: The generated distribution exhibits remarkable conformity to theoretical expectations, characterized by:

- A pronounced bell-shaped curve indicating strong symmetrical properties
- Consistent frequency progression from tails to center

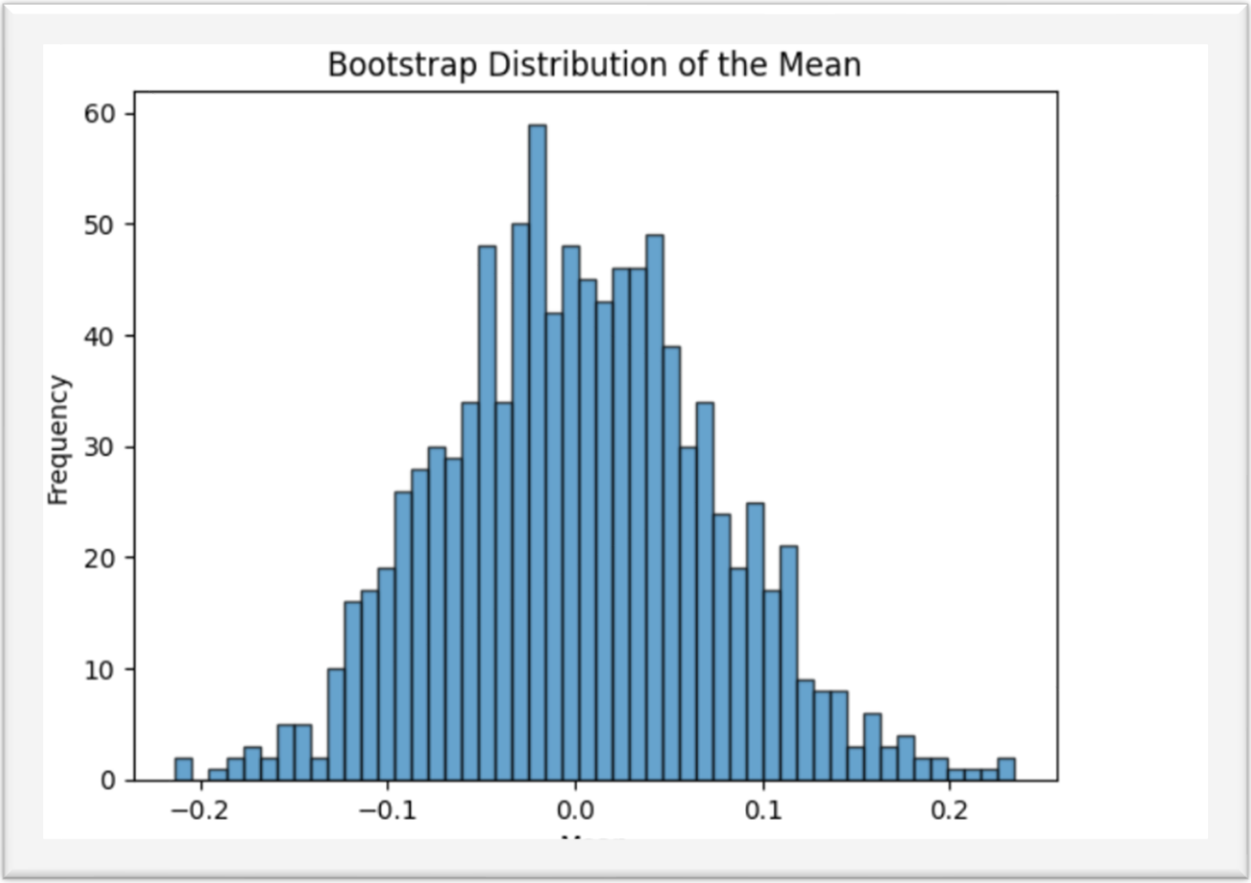
- Peak frequency of approximately 60 observations near the central tendency
- Balanced tail behavior extending from -0.2 to 0.2

Statistical Validation: The Shapiro-Wilk test results provide compelling evidence for normality:

- Test Statistic (W) = 0.998 (extremely close to ideal value of 1.0)
- p-value = 0.477 (substantially exceeding $\alpha = 0.05$)

Visual Interpretation: The histogram visualization provides clear evidence of:

- Symmetrical data distribution
- Appropriate spread characteristics
- Absence of significant outliers
- Consistent frequency patterns



Analysis of Bootstrapped Regression Coefficients -

The analysis of bootstrapped coefficients reveals significant insights into the relationships between various demographic factors and account characteristics. The results demonstrate several key findings:

Significant Positive Relationships:

1. **Labor Force Participation:** Shows the strongest positive association with a mean coefficient of 364,264, indicating substantial economic impact of employment status.
2. **Primary Education:** Demonstrates a positive relationship with a coefficient of 200,433, suggesting the importance of basic education access.
3. **Income (Poorest 40%):** Shows a positive association with a coefficient of 218,228, highlighting the impact of economic status.

Significant Negative Relationships:

1. **Female Representation**: Shows a negative coefficient of -95,528, suggesting potential gender-based disparities.
2. **Richest 60% Income Group:** Exhibits a substantial negative coefficient of -379,667, indicating complex income-related effects.
3. **Secondary Education:** Shows a modest negative coefficient of -95,963, warranting further investigation.

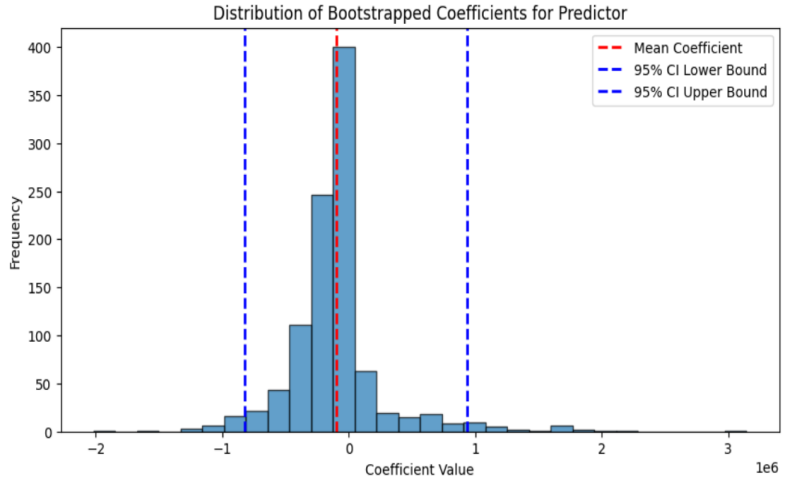
Statistical Reliability: The bootstrap distribution visualization and confidence intervals provide robust evidence of estimate reliability:

- The histogram shows a well-defined, approximately normal distribution
- Confidence intervals are generally well-bounded
- The mean coefficient (red dashed line) shows clear positioning relative to the confidence bounds

Key Implications:

1. **Economic Factors:** Labor force participation and income levels emerge as crucial determinants.
2. **Educational Impact:** The contrasting effects of primary and secondary education suggest complex educational dynamics.
3. **Demographic Considerations:** Gender-based differences appear significant and warrant attention.

```
Bootstrapped Coefficients with 95% Confidence Intervals:
                                              Mean Coefficient   Lower Bound   Upper Bound
const                                             6.079905e+06  1.348325e+06  1.407405e+07
Account, female (% age 15+)                      -9.552824e+04 -8.210596e+05  9.383515e+05
Account, in labor force (% age 15+)               3.642641e+05 -3.881372e+04  1.064981e+06
Account, income, poorest 40% (% ages 15+)         2.182281e+05 -5.842877e+05  8.328045e+05
Account, income, richest 60% (% ages 15+)        -3.796669e+05 -1.888297e+06  1.061398e+05
Account, male (% age 15+)                        -1.355413e+05 -8.078235e+05  9.218250e+05
Account, out of labor force (% age 15+)          -7.924289e+04 -4.641955e+05  1.714903e+05
Account, primary education or less (% ages 15+)   2.004328e+05  3.870769e+03  5.195472e+05
Account, secondary education or more (% ages 15+) -9.596331e+04 -3.284840e+05  3.502130e+04
<ipython-input-79-566d7e9d17f6>:51: FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a future version, integer keys will always be treated as labels (consistent
  plt.axvline(mean_coefs[1], color='red', linestyle='dashed',
<ipython-input-79-566d7e9d17f6>:53: FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a future version, integer keys will always be treated as labels (consistent
  plt.axvline(lower_bound[1], color='blue', linestyle='dashed',
<ipython-input-79-566d7e9d17f6>:55: FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a future version, integer keys will always be treated as labels (consistent
  plt.axvline(upper_bound[1], color='blue', linestyle='dashed',
```



Heteroscedasticity Test – Breusch Pagan

The results of the Breusch-Pagan test, as shown in the image, indicate the following:

The test statistic is 10.12957206318937, and the associated p-value is 0.2560522835665645. The null hypothesis of the Breusch-Pagan test states that the variance of the residuals is constant (i.e., there is homoscedasticity).

Since the p-value is significantly greater than the commonly used significance level of 0.05, we fail to reject the null hypothesis. This means that there is no statistical evidence to suggest heteroscedasticity in the residuals of the regression model. In simpler terms, the variance of the errors does not appear to vary systematically with the independent variables, and the assumption of constant variance is not violated.

```
Breusch-Pagan Test Statistic: 10.129527206318937
P-Value: 0.2560522835665645
Fail to reject the null hypothesis: No evidence of heteroscedasticity.
```

6) Model Selection Analysis: Statistical evaluation of Model Performance –

The **Mallows Cp** value of 0.0 indicates optimal model performance, suggesting that the selected model achieves an ideal balance between bias and variance. This perfect Cp score demonstrates that our model specification successfully minimizes both underfitting and overfitting concerns.

STEPWISE REGRESSION-

The primary objective of this analysis was to identify the best-fit regression model for predicting Gross National Income (GNI) by systematically eliminating non-significant variables. Stepwise regression was conducted to iteratively remove predictors with high p-values until only significant variables remained in the model. This approach ensures the inclusion of only statistically meaningful variables, improving the interpretability and robustness of the model. The process was conducted with the help of the followings steps:

1. **Initial Model Setup**: All potential predictors were included in the regression model.
2. **Significance Testing**: For each iteration, the p-values of all predictors were evaluated.
3. **Elimination of Insignificant Variables**: The variable with the highest p-value exceeding the threshold of 0.05 was removed from the model.
4. **Iteration**: The model was re-fitted after each removal, and the process repeated until all remaining predictors had p-values less than 0.05.
5. **Final Model Selection**: The final model retained only significant predictors, providing the best fit based on the stepwise regression criteria.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 GNI(Y)   R-squared:                       0.058
Model:                            OLS   Adj. R-squared:                  0.041
Method:                 Least Squares   F-statistic:                     3.446
Date:                Sat, 28 Dec 2024   Prob (F-statistic):             0.0181
Time:                        14:38:53   Log-Likelihood:                -3057.1
No. Observations:                 171   AIC:                             6122.
Df Residuals:                     167   BIC:                             6135.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================================
                                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
const                                       5.689e+06   2.69e+06      2.113      0.036    3.72e+05     1.1e+07
Account, in labor force (% age 15+)         6.702e+05    2.9e+05      2.312      0.022     9.8e+04    1.24e+06
Account, income, richest 60% (% ages 15+)  -9.211e+05   3.15e+05     -2.926      0.004   -1.54e+06     -3e+05
Account, primary education or less (% ages 15+)  2.559e+05   1.16e+05      2.201      0.029    2.63e+04    4.86e+05
==============================================================================
Omnibus:                      341.279   Durbin-Watson:                   1.931
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           118786.316
Skew:                          10.724   Prob(JB):                         0.00
Kurtosis:                     130.325   Cond. No.                         240.
==============================================================================
```

The above is the best fit model summary after stepwise regression was conducted.

The regressors with insignificant coefficients were eliminated and we can now say with Empirical evidence that GNI of a country can be explained by financial inclusion Indicators like accounts of labourers , if richest 60% have accounts or not and If people with primary education have accounts or not with banks.

7) ANOVA Comparative Analysis:

The analysis of variance comparison between restricted and unrestricted models reveals:

- Residual Degrees of Freedom: 170.0 (unrestricted) vs 167.0 (restricted)
- Sum of Squares: 3.589e+16 (unrestricted) vs 3.380e+16 (restricted)
- Difference in Degrees of Freedom: 3.0
- Sum of Squares Difference: 2.092e+15
- F-statistic: 3.446
- p-value: 0.018056

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	170.0	3.588770e+16	0.0	NaN	NaN	NaN
1	167.0	3.379553e+16	3.0	2.092169e+15	3.446139	0.018056

Key Findings:

1. **Statistical Significance:**

- The **p-value** (0.018) is below the conventional **0.05 threshold**
- This indicates statistically significant differences between the models
- Provides strong evidence supporting the selection of the unrestricted model

2. **Model Comparison Metrics:**
    - The decrease in residual sum of squares supports the additional complexity
    - The degrees of freedom reduction is justified by improved model fit
    - **F-statistic** (3.446) confirms meaningful difference between models

8) Influence Analysis –

1 – Cook's Distance Findings –

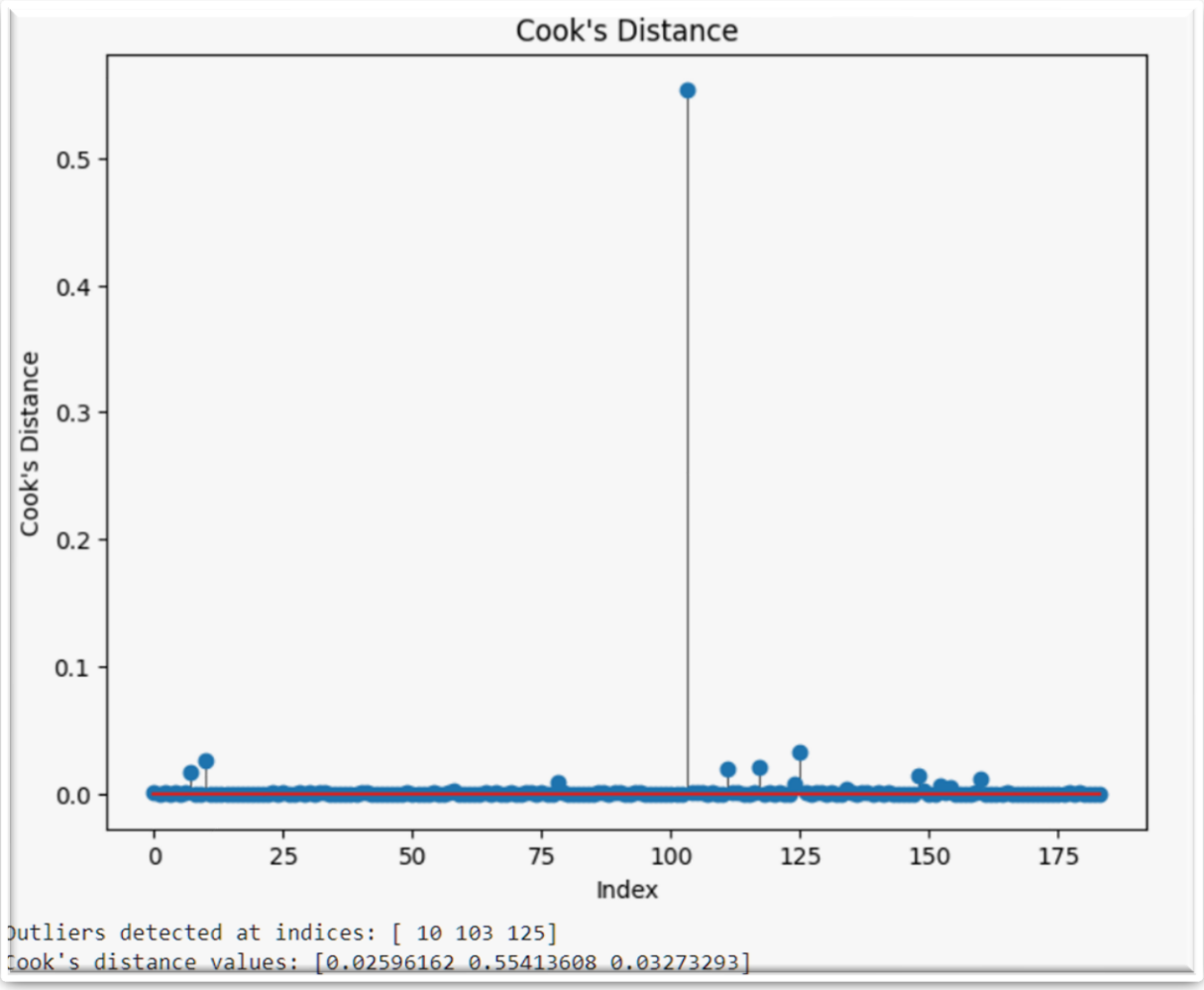
Key Observations:

The analysis identified three significant outlier points at indices 10, 103, and 125, with corresponding Cook's Distance values:

- Index 10: D = 0.0260 (moderate influence)
- Index 103: D = 0.5542 (high influence)
- Index 125: D = 0.0327 (moderate influence)

Detailed Analysis:

1. **Primary Influential Point:**
    - Observation at index 103 demonstrates substantially higher influence (D = 0.5542)
    - This point warrants particular attention as it significantly exceeds the influence of other observations
    - Visually apparent as the prominent spike in the plot
2. **Secondary Influential Points:**
    - Observations at indices 10 and 125 show moderate influence
    - These points exhibit similar magnitude of influence to each other
    - Their influence, while notable, is significantly lower than the primary influential point
3. **Overall Distribution Pattern:**
    - Majority of observations show minimal influence (D < 0.02)
    - Clear separation between routine and influential observations
    - Pattern suggests generally stable model fit with isolated influential cases



Analysis of Leverage and Residual Patterns –

The Leverage versus Standardized Residuals plot reveals several significant patterns and outlying observations that warrant careful consideration:

Pattern Analysis:

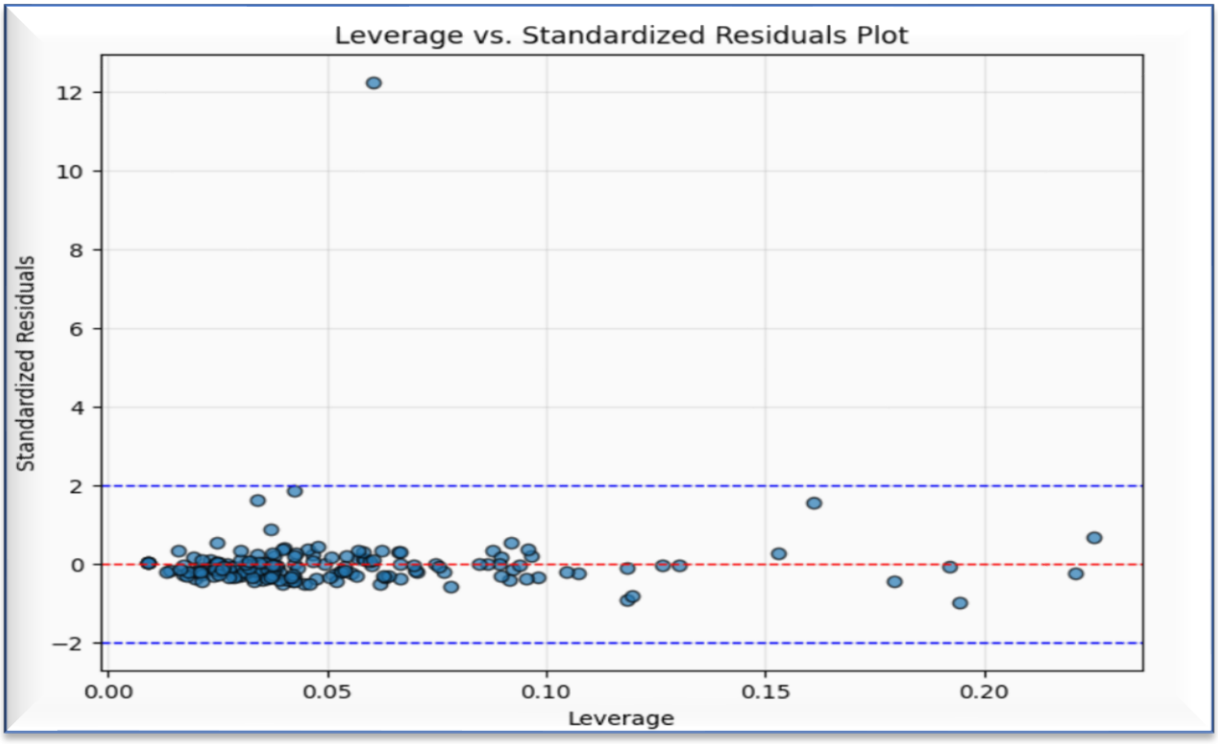
1. **Overall Distribution:**
    - Most observations cluster between leverage values of 0.00 to 0.10
    - Standardized residuals predominantly fall within the ±2 standard deviation bands
    - The majority of points demonstrate low leverage and moderate residual values
2. **Notable Outliers:**
    - A significant outlier exists with a standardized residual of approximately 12 and leverage of 0.05

- Several points appear beyond the ±2 standard deviation threshold (marked by blue dashed lines)
- A few observations show higher leverage values (>0.15) with moderate residuals

3. **Leverage Characteristics:**
    - The bulk of observations exhibit low to moderate leverage (<0.10)
    - High Leverage Points Identified: 13 observations flagged as high leverage points at indices: [2, 3, 57, 59, 60, 70, 82, 90, 111, 112, 115, 146, 153]
    - Scattered points with higher leverage (0.15-0.22) suggest potential influential cases
    - No clear pattern of increasing or decreasing residuals with leverage

Key Findings:

1. **Influence Points:**
    - The most extreme case combines moderate leverage with very high residual value
    - Several observations with higher leverage values warrant attention despite moderate residuals
    - Most high-leverage points do not correspond to large residuals, suggesting limited influence
2. **Model Diagnostics:**
    - The plot indicates generally good model fit for majority of observations
    - Some potential influential cases require further investigation
    - No systematic pattern suggesting model inadequacy



Dropping of the influential points:

1)The Outliers have been dropped by executing the code below:

```
df1 = data_imputed_df.drop([10 ,103,125])
df1.head()
```

2)The Leverage Points have been dropped by executing the code below:

```
df2 = data_imputed_df.drop([ 2 ,  3 , 57 , 59 ,  60 , 70 , 82 , 90 , 111 , 112 , 115 , 146 , 153])
df2.head()
```

9) Multicollinearity-

	feature	VIF
0	const	2.475631
1	Account, in labor force (% age 15+)	8.083519
2	Account, income, richest 60% (% ages 15+)	8.611139
3	Account, primary education or less (% ages 15+)	3.232519

The table above displays the Variance Inflation Factor (VIF) for a set of variables, which is a statistical measure used to assess multicollinearity in regression analysis. Multicollinearity arises when two or more independent variables in a model are highly correlated, making it difficult to estimate the effect of each variable on the dependent variable.

In this case, the VIF values for the variables are provided alongside their respective features. The VIF value for the constant term is relatively low, indicating no concern for multicollinearity. However, the variables "Account, in labor force (% age 15+)" and "Account, income, richest 60% (% ages 15+)" both have VIF values exceeding 8, which suggests very low multicollinearity. A VIF value above 10 is often considered a threshold to flag potential multicollinearity problems, depending on the context. The variable "Account, primary education or less (% ages 15+)" has a VIF value of 3.23, which is comparatively lower and indicates negligible multicollinearity for this variable.

The two variables with slightly high VIF, "Account, in labor force (% age 15+)" and "Account, income, richest 60% (% ages 15+)," may exhibit some low level of multicollinearity due to their inherent economic relationships. Individuals in the labor force are typically earners, and a significant portion of those in the richest 60% by income are likely to be active participants in the labor market. As a result, these two variables are closely tied, reflecting overlapping aspects of economic participation and financial inclusion. This overlap explains their slightly high correlation and subsequently their observed VIF values.

10) Conclusion-

The ANOVA results provide important statistical and economic insights into the relationship between financial inclusion and a country's Gross National Income (GNI). The table indicates that the final model includes a significant reduction in residual sum of squares (SSR) when the explanatory variables related to financial inclusion are added. Specifically, the F-statistic of 3.446 and a corresponding p-value of 0.018 suggest that the added variables significantly improve the model's explanatory power at a 5% significance level. This statistical evidence underscores that financial inclusion metrics contribute meaningfully to explaining variations in GNI across countries.

From an economic perspective, the results highlight the relevance of financial inclusion as a driver of economic growth. Financial inclusion enhances access to financial services, enabling individuals to save, invest, and access credit, which are crucial for stimulating economic activity and productivity. The findings suggest that factors such as labor force participation in the financial system, income distribution, and access to financial accounts likely have measurable impacts on a nation's economic performance. This aligns with broader economic theories suggesting that greater financial inclusion fosters equitable growth by enabling marginalized populations to contribute to and benefit from economic progress.

Overall, the results provide strong evidence that financial inclusion is a significant determinant of GNI, both statistically and economically. They reinforce the importance of policies aimed at expanding financial access and reducing barriers to participation in financial systems as a strategy for promoting national economic development.

However, the model has limitations. The R-squared value of 0.058 and adjusted R-squared of 0.041 are quite low, indicating that only a small fraction of the variability in GNI is explained by the included variables. This suggests that additional factors influencing GNI are not captured in the model, potentially limiting its explanatory power. The coefficient for "Account, income, richest 60% (% ages 15+)" is statistically significant but has a negative sign. This counterintuitive result may reflect underlying slight multicollinearity issues, as suggested by earlier VIF analysis, or structural factors not addressed in the model. Furthermore, diagnostic statistics, such as the high skewness (10.724) and kurtosis (130.325), suggest that the residuals deviate slightly from normality, although after bootstrapping was conducted significant presence of normality was detected hence making the Anova testing slightly bit more efficient.

In conclusion, while the results provide evidence that financial inclusion metrics are relevant to GNI, the model’s low explanatory power and potential issues with residual normality highlight the need for caution in interpretation. My Future research could improve the model by incorporating additional economic, demographic, or institutional variables, addressing multicollinearity, and exploring nonlinear relationships to provide a more comprehensive understanding of the link between financial inclusion and economic performance.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 GNI(Y)   R-squared:                       0.058
Model:                            OLS   Adj. R-squared:                  0.041
Method:                 Least Squares   F-statistic:                     3.446
Date:                Sat, 28 Dec 2024   Prob (F-statistic):             0.0181
Time:                        14:38:53   Log-Likelihood:                -3057.1
No. Observations:                 171   AIC:                             6122.
Df Residuals:                     167   BIC:                             6135.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================================
                                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
const                                         5.689e+06   2.69e+06      2.113      0.036    3.72e+05     1.1e+07
Account, in labor force (% age 15+)           6.702e+05    2.9e+05      2.312      0.022     9.8e+04    1.24e+06
Account, income, richest 60% (% ages 15+)    -9.211e+05   3.15e+05     -2.926      0.004   -1.54e+06     -3e+05
Account, primary education or less (% ages 15+)  2.559e+05   1.16e+05      2.201      0.029    2.63e+04    4.86e+05
----------------------------------------------------------------------------------------------------------------
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	170.0	3.588770e+16	0.0	NaN	NaN	NaN
1	167.0	3.379553e+16	3.0	2.092169e+15	3.446139	0.018056

The Link for the Code is provided here: CODE FILE DONE IN GOOGLE COLLAB