Problem Statement

In the modern global economy, the debate persists on the efficacy of traditional economic metrics for assessing societal well-being. While GDP and GNI have historically been crucial economic indicators, their effectiveness in representing social welfare is debatable. This research explores the link between economic prosperity, as indicated by GNI per capita, and broader social well-being measures across nations. The study's core inquiry is whether aggregate economic metrics can adequately reflect the complex nature of societal development and human welfare. Traditional economic metrics exhibit notable limitations:

They neglect environmental externalities and sustainability. They exclude non-market activities and informal economies. They do not account for income and wealth distribution disparities. They overlook quality-of-life factors like leisure and work-life balance. They fail to encompass critical development indicators such as health and education quality.

Employing comprehensive World Bank data, this study investigates the relationship between GNI per capita and various social welfare indices, including the HDI and GDI. The research seeks to validate the hypothesis that aggregate economic metrics are inadequate indicators of societal well-being.

The main objectives of this analysis are:

- To evaluate the strength and nature of the relationships between economic and social indicators.
- To determine if higher GNI per capita correlates with enhanced social outcomes.
- To identify gaps between economic growth and social development.
- To contribute to the broader discourse on measuring national progress and well-being.

This investigation holds particular significance in the current global context, where nations increasingly acknowledge the necessity for holistic methods of assessing societal progress and development. The findings are intended to offer valuable insights for policymakers and researchers aiming to develop more comprehensive frameworks for evaluating national well-being.

Null Hypothesis (H_0):

There is no significant relationship between GNI per capita and social well-being indicators.

Mathematically expressed as: $\beta = 0$
(where β represents the coefficient of correlation between GNI per capita and social indicators)

Alternative Hypothesis (H_1):

There exists a significant positive relationship between GNI per capita and social well-being indicators.

Mathematically expressed as: $\beta \neq 0$

Expected Outcomes

If the null hypotheses are rejected, we expect to find:

1. Higher GNI per capita correlates with higher Social well-being indicators

These hypotheses will be tested at a significance level of $\alpha = 0.05$ (95% confidence level) using appropriate statistical methods including regression analysis and correlation tests.

Data Set

2.1 Source

The data for this analysis was sourced from the World Bank's World Development Indicators (WDI) database. The dataset includes various economic and social indicators for different countries.

Data Processing Steps:

1) Initial data extraction from World Bank database.

2) Data cleaning and Standard Scaling
   a) Applied StandardScaler from sklearn.preprocessing
   b) Scaling was essential due to vastly different ranges across variables:
      - GNI per capita: Large values in thousands (USD)
      - HDI and GDI: Small values between 0-1
      - Life expectancy: Medium range (40-90 years)
      - Schooling variables: Small range (0-20 years)
   c) Standard Scaling transformed all features to standardized form:
      - $x\_scaled = (x - \mu)/(\sigma)$
   d) This preserves zero values and handles sparse data effectively
   e) Maintains relationships between the original data points

2.2 Variable Description

Variable Name	Type	Units of Measurement	Description	Role in Analysis
Carbon Dioxide Emissions per Capita (production)	Float64	Tonnes	Measures the average carbon dioxide emissions produced per person in a country	Independent Variable
Expected Years of Schooling	Float64	Years	Represents the total number of years of schooling a child entering school can expect to receive	Independent Variable
Gender Development Index	Float64	Index (0-1)	Reflects gender inequalities in achievement in three basic dimensions of human development	Independent Variable
Gross National Income Per Capita (2017 PPP$)	Float64	Current US$	Average income per person, adjusted for purchasing power parity, in 2017 dollars	Dependent Variable
Human Development Index	Float64	Index (0-1)	Composite index measuring average achievement in key dimensions of human development	Independent Variable

Labour Force Participation Rate, Female	Float64	Percentage	Percentage of females aged 15 and older who are economically active	Independent Variable
Labour Force Participation Rate, Male	Float64	Percentage	Percentage of males aged 15 and older who are economically active	Independent Variable
Life Expectancy at Birth	Float64	Years	Average number of years a newborn is expected to live if current mortality rates continue	Independent Variable
Mean Years of Schooling	Float64	Years	Average number of years of education received by people ages 25 and older	Independent Variable
Share of Seats in Parliament, Female	Float64	Percentage	Percentage of parliamentary seats held by women	Independent Variable

- Total number of observations: 195 countries
- The data represents measurements from 2021

Data Exploration

3.1 Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
Carbon dioxide emissions per capita (production) (tonnes)	195.0	4.373053	5.546605	0.038	0.792500	2.5390	5.8505	39.884
Expected Years of Schooling (years)	195.0	13.561890	3.035598	5.635	11.673000	13.3810	15.6765	21.080
Gender Development Index (value)	195.0	0.948418	0.070363	0.456	0.926000	0.9700	0.9910	1.064
Gross National Income Per Capita (2017 PPP$)	195.0	21087.447118	22439.108954	690.661	4794.453000	12467.8550	32353.0670	146673.242
Human Development Index (value)	195.0	0.724526	0.154716	0.380	0.605000	0.7400	0.8480	0.967
Labour force participation rate, female (% ages 15 and older)	195.0	49.824304	14.692105	5.840	42.693571	52.1100	59.0800	82.940
Labour force participation rate, male (% ages 15 and older)	195.0	69.915473	9.982856	29.630	65.380000	69.9300	75.8500	98.580
Life Expectancy at Birth (years)	195.0	71.928241	7.897440	52.997	65.935500	72.3000	77.9525	86.895
Mean Years of Schooling (years)	195.0	9.028552	3.210252	1.341	6.527500	9.4240	11.6350	14.256
Share of seats in parliament, female (% held by women)	195.0	25.266372	12.173021	0.294	16.850500	25.1725	33.3330	54.717

(for clearer view please visit the .ipynb file)

Key insights:

1. **Economic Indicators**

- **GNI Per Capita (2017 PPP$)**:

- Extremely wide range: minimum of 690.66 to maximum of 146,673.24
- High standard deviation (22,439.11) indicates significant economic inequality between countries
- Median (12,467.86) much lower than mean (21,087.45) suggesting right-skewed distribution with some very wealthy outliers

2. Development Indices

- **Human Development Index (HDI)**:
  - Ranges from 0.380 to 0.967
  - Mean of 0.725 indicates moderate global development level
  - Relatively small standard deviation (0.155) suggests clustered development levels
- **Gender Development Index (GDI)**:
  - High mean of 0.948 indicates generally good gender parity
  - Small standard deviation (0.074) shows consistency across countries
  - Maximum value above 1 (1.064) indicates some countries where women's development exceeds men's

3. Education Metrics

- **Expected Years of Schooling**:
  - Global mean of 13.56 years
  - Range from 5.64 to 21.08 years shows significant educational disparity
  - 75% of countries achieve at least 11.67 years
- **Mean Years of Schooling**:
  - Average of 9.03 years globally
  - Wide range from 1.34 to 14.26 years indicates substantial educational inequality

4. Gender and Labor Statistics

- **Labor Force Participation**:
  - Significant gender gap: Male participation (mean 69.92%) substantially higher than female (mean 49.82%)
  - Female participation shows higher variability (std dev 14.69) compared to male (std dev 9.98)
- **Female Parliamentary Representation**:
  - Global mean of 25.27% female representation
  - Wide range from 0.29% to 54.72%
  - Median of 25.17% indicates balanced distribution

5. Environmental and Health Indicators

- **Carbon Dioxide Emissions**:

- High variability (std dev 5.55 tonnes per capita)
- Extremely skewed distribution with maximum at 39.88 tonnes
- Median of 2.54 tonnes much lower than mean of 4.37 tonnes

- **Life Expectancy**:
  - Global mean of 71.93 years
  - Range from 52.99 to 86.90 years shows significant health inequality
  - Relatively small standard deviation (7.90) suggests clustering around mean

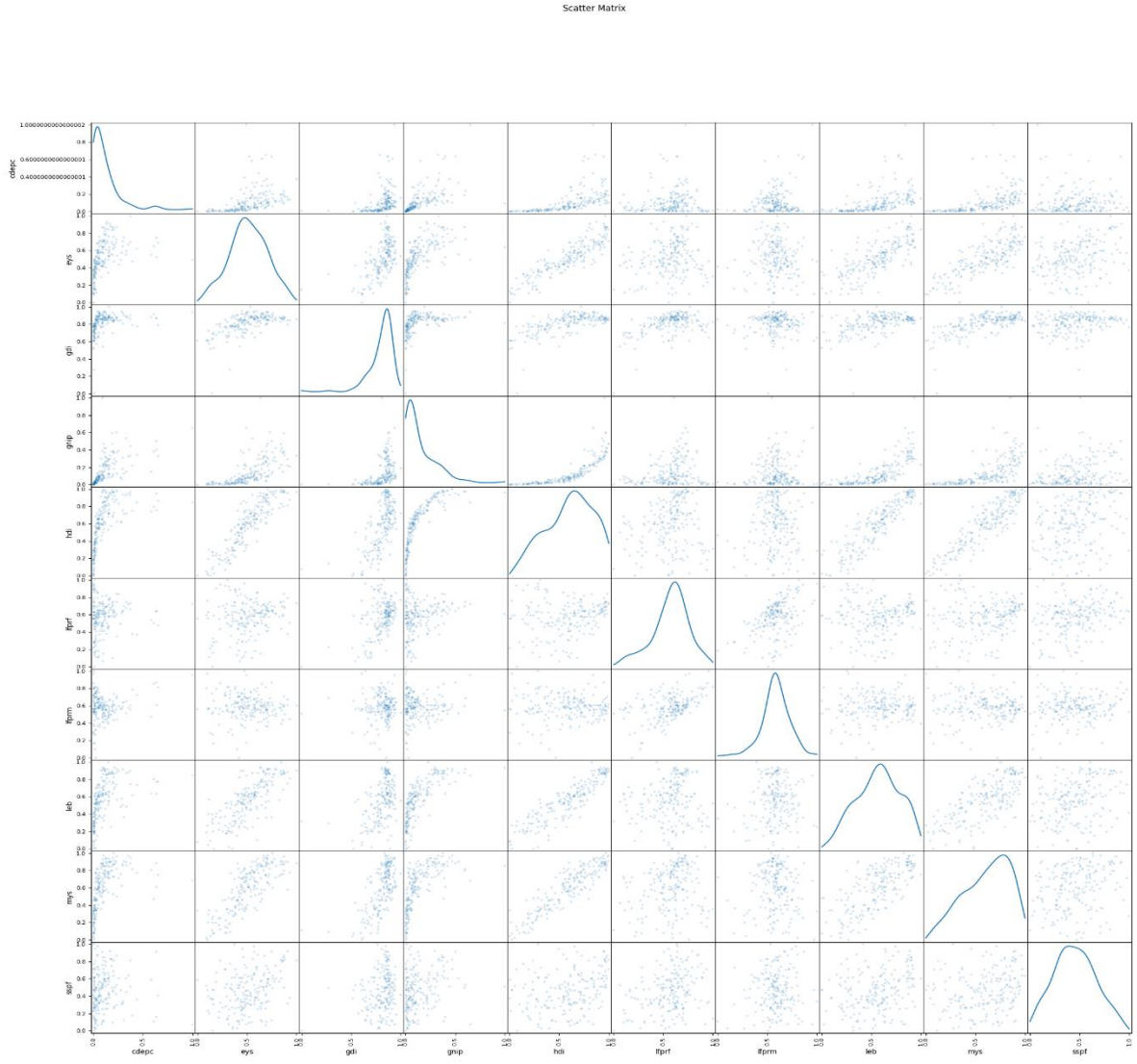
These statistics reveal significant global disparities in economic, social, and environmental indicators, with particularly notable gaps in economic measures and gender-related metrics.

3.2 Graphical Analysis

Distribution Analysis

Key Patterns Observed from Scatter Matrix

- Most variables show non-normal distributions as evident from the diagonal plots
- Several variables display right-skewed distributions, particularly economic indicators
- Some variables show more symmetric distributions, especially the index measures (HDI, GDI)



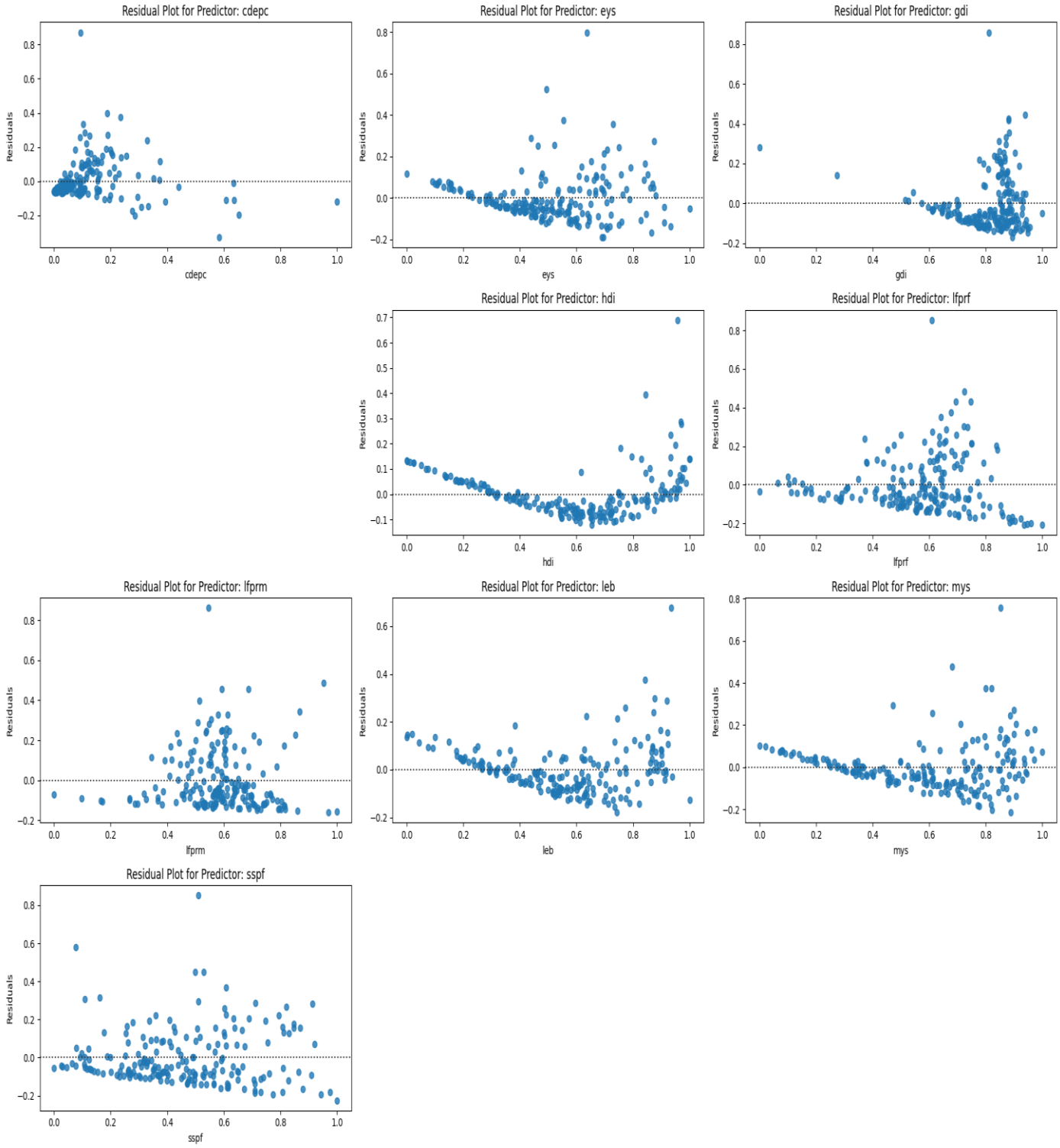
Residual Plot

Key Patterns Observed from Residual Plots

Heteroscedasticity:

1. GDI (Gender Development Index):
    - Residual spread increases at higher values
    - Fan-shaped pattern indicates heteroscedasticity
2. Labor Force Participation (lfprf and lfprm):
    - Variable spread of residuals across predictor values
    - Greater variance at middle ranges

We will later try to confirm heteroscedasticity using hypothesis testing.



Linearity checks: Partial Regression Plot and CCPR Plot

Based on the Component-Component Plus Residual Plot and Partial Regression Plots, here's the linearity analysis:

Linearity Assessment

Strong Linear Relationships:

1. **HDI (Human Development Index):**
    a) Shows strong positive linear relationship
    b) Relatively consistent scatter around the fitted line
    c) Few outliers at extreme values

Moderate Linear Relationships:

1. **Female Labor Force Participation (lfprf):**
    a) Moderate positive linear trend
    b) Some scatter but generally follows linear pattern
    c) Few influential points at extremes

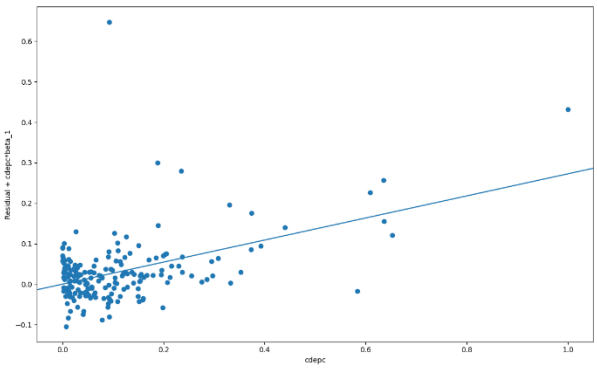
2. **Expected Years of Schooling (eys):**
    a) Moderate positive linear relationship
    b) Some deviation from linearity at higher values
    c) Generally acceptable linear fit

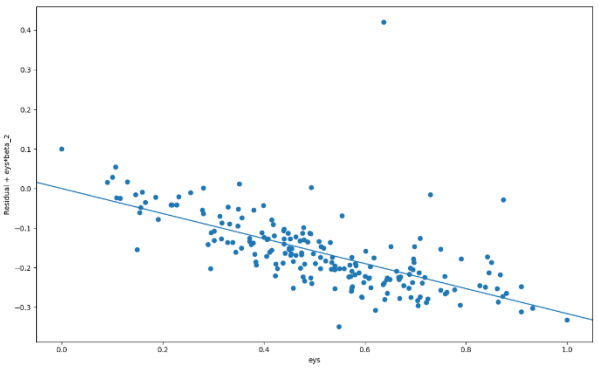
3. **GDI (Gender Development Index):**
    a) Slight non-linear pattern
    b) Data points at higher values
    c) Suggests possible polynomial relationship

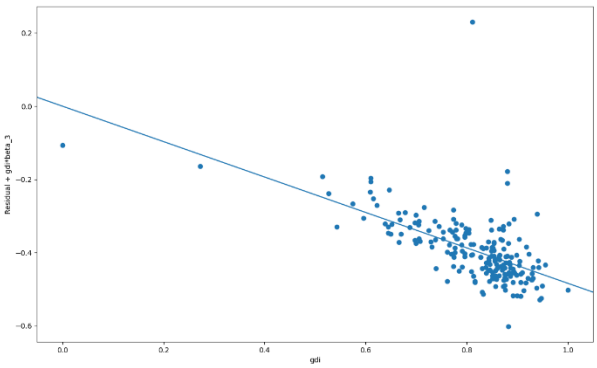
4. **Mean Years of Schooling (mys):**
    a) Slight non-linear trend
    b) Scattered pattern around fitted line
    c) May benefit from transformation

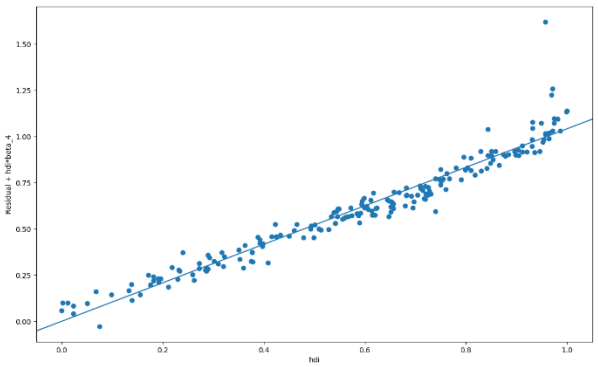
5. **Share of Seats in Parliament (sspf):**
    a) Weak linear relationship
    b) Scattered pattern with outliers
    c) May not be a significant predictor

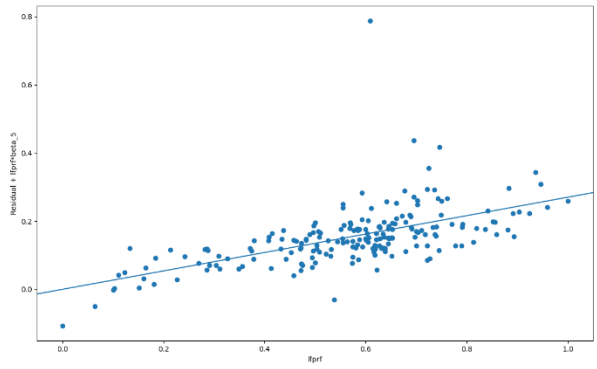
Component-Component Plus Residual Plot

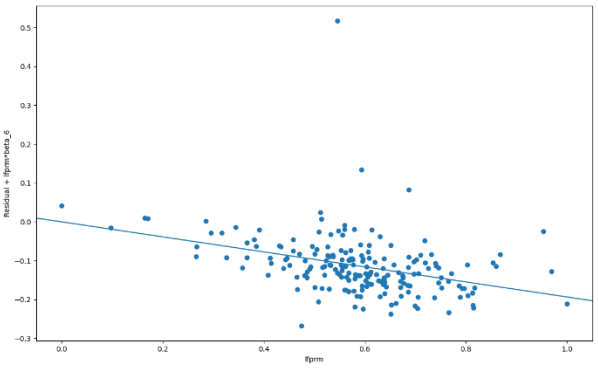


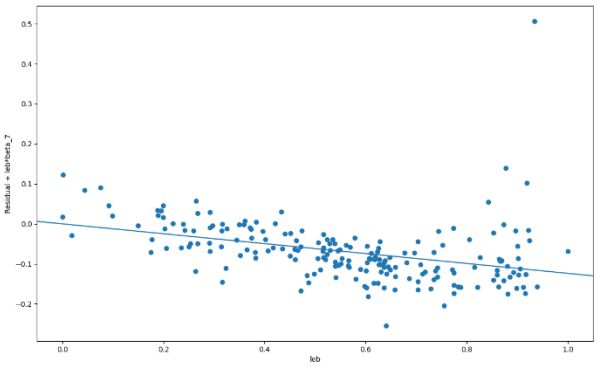


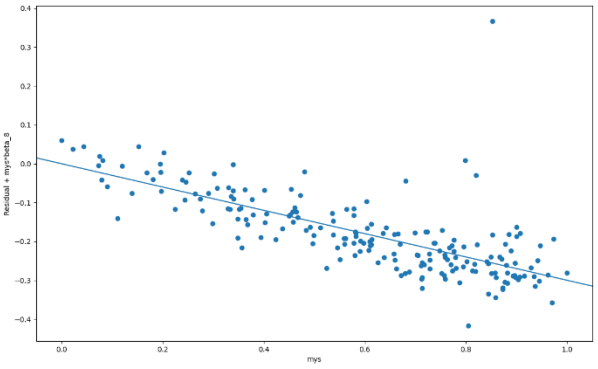


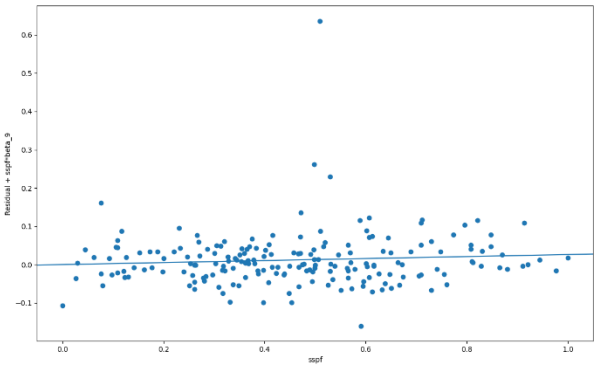




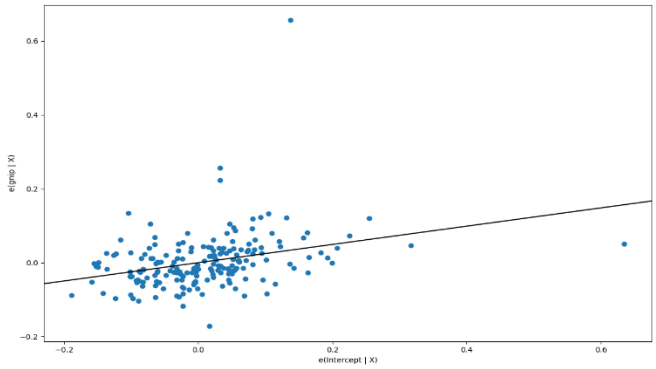


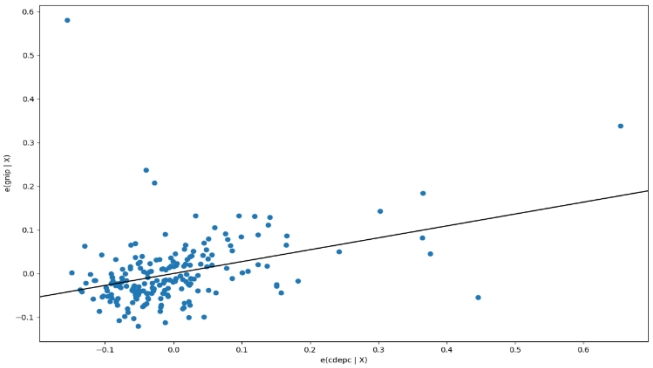


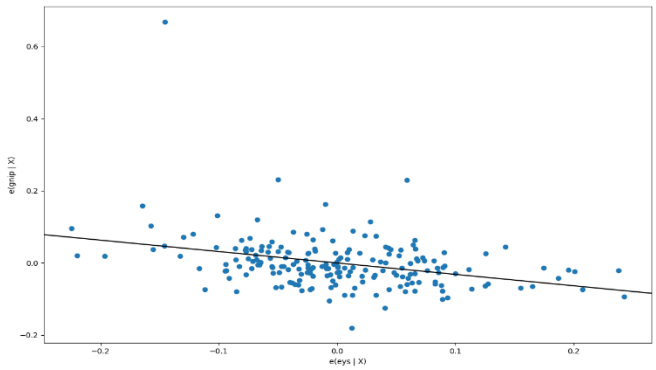


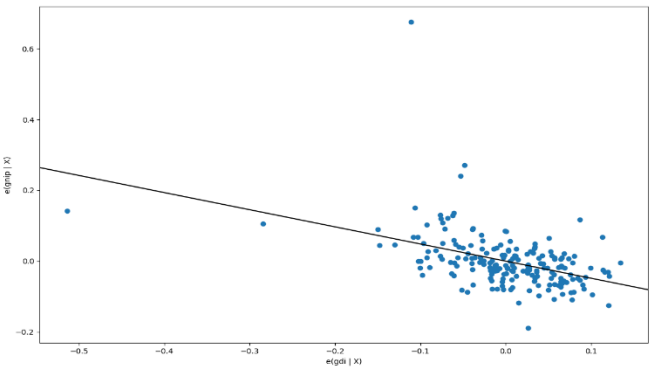


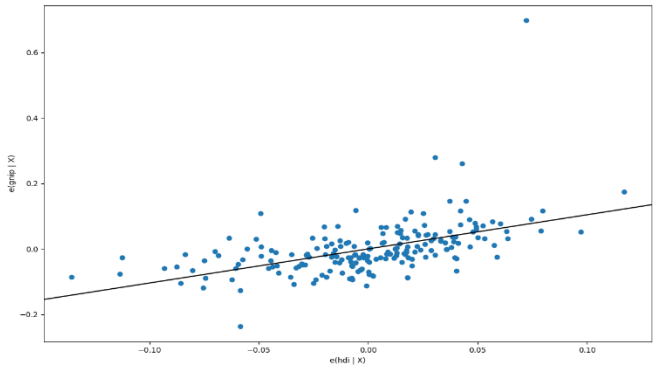
Partial Regression Plot

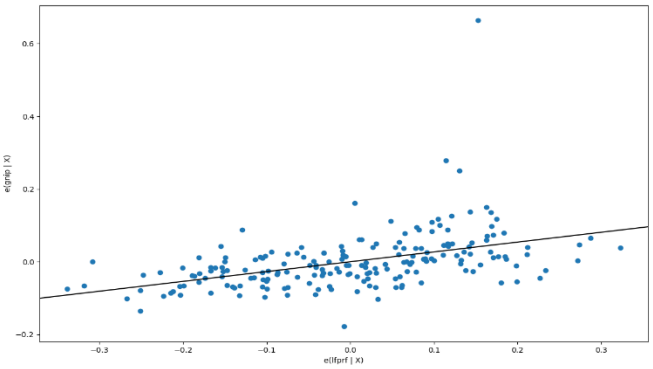


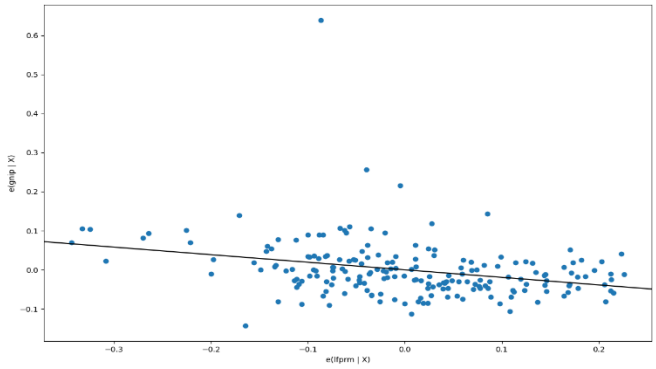


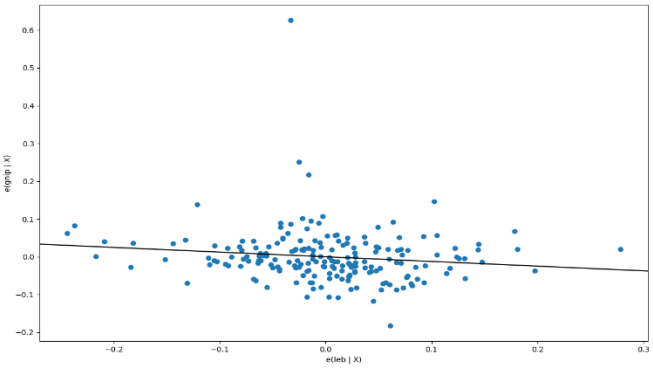


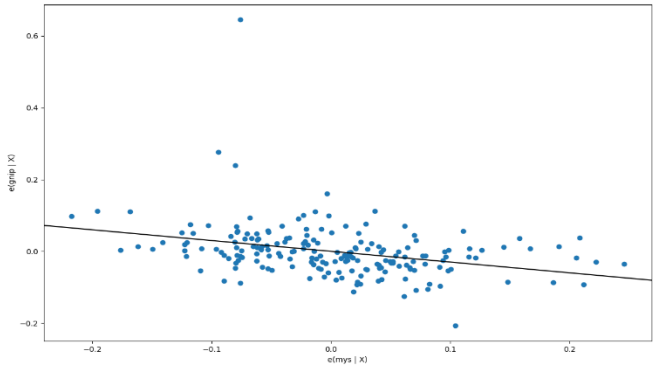


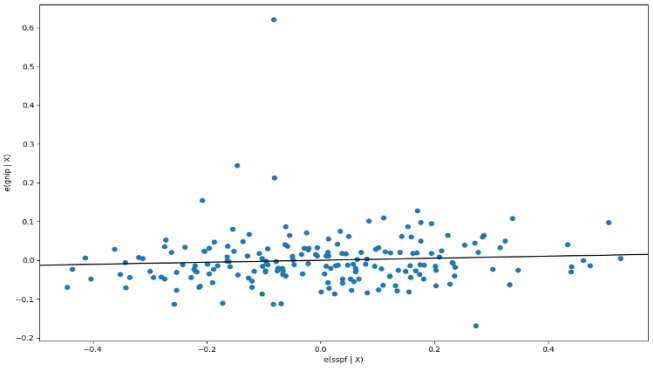






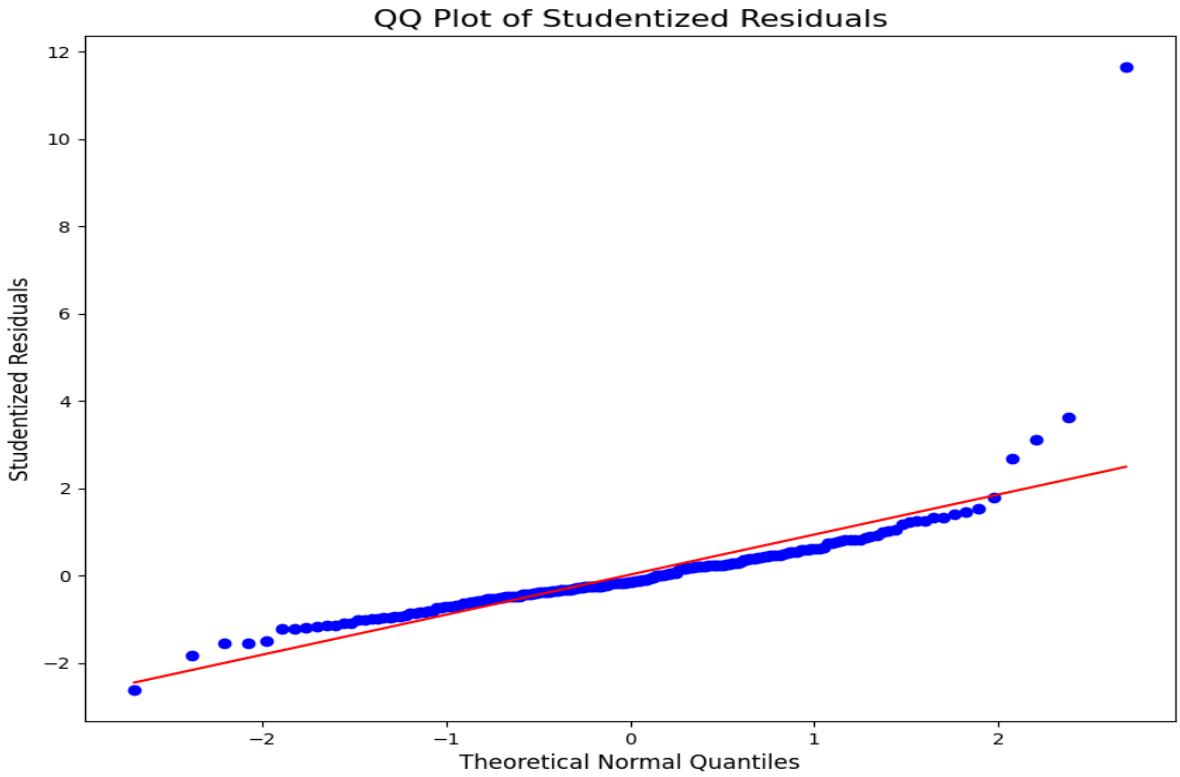






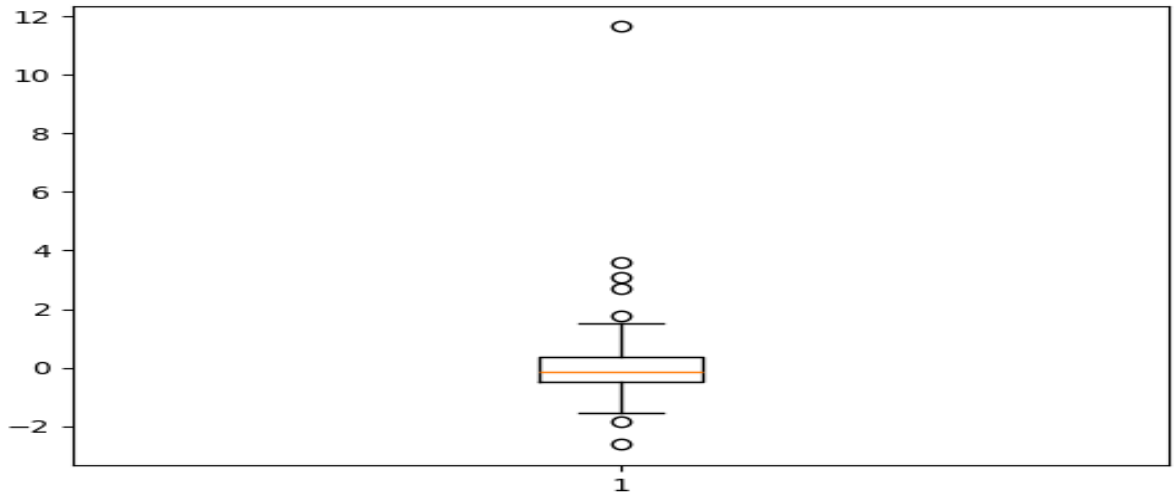
Normality tests

QQ plot



The Q-Q plot suggests that our response variable is heavily skewed or contains extreme outliers. This is why the data points deviate significantly from the 45-degree line, especially for larger values. This implies data might not be normal. Data contains extreme Outliers. Data is affected by very high numeric values.

Inspection of Outliers using Box-plot:



Initial Tests on Original Data(Normalized/Scaled) (GNI per capita)

1. **Kolmogorov-Smirnov Test**

- Statistic = 0.1902
- p-value = 1.181e-06 (< 0.05)
- Result: Reject null hypothesis of normality

2. **Shapiro-Wilk Test**

- Statistic = 0.8023
- p-value = 5.569e-15 (< 0.05)
- Result: Reject null hypothesis of normality

Justification:

1. Initial Data Characteristics The data exhibited marked non-normality. Variables displayed heterogeneous scales and ranges. Both positive and negative values were present post-standardization.

2. Standard Scaling Implementation Standard Scaling was imperative due to: Diverse scales among variables (GNI in thousands contrasted with HDI in 0-1). It ensures uniform contribution of all features to the model. Zero values and data structure are maintained. Formula: x_scaled = (x - μ)/σ. This resulted in features being comparable with mean=0 and std=1.

3. Transformation Attempts Unsuccessful Transformations: Box-Cox Transformation was inapplicable due to negative values post-standardization. It necessitates strictly positive values. Log Transformation failed owing to negative values. This would compromise data structure. Square Root Transformation was not feasible due to negative values. It would complicate interpretation. Successful Approach: Yeo-Johnson Transformation

Chosen because: It accommodates both positive and negative values. Zero values are preserved. It is more adaptable than Box-Cox. It is appropriate for standardized data.

Tests After Yeo-Johnson Transformation

1. **Kolmogorov-Smirnov Test on Transformed Data**

- Statistic = 0.1089
- p-value = 0.0180 (< 0.05)
- Result: Still reject normality at 5% significance

2. **Shapiro-Wilk Test on Transformed Data**

- Statistic = 0.9342
- p-value = 1.007e-07 (< 0.05)
- Result: Still reject normality

Conclusion and Justification

Notwithstanding the implications of formal tests indicating non-normality, we are able to advance with the analytical process due to the following justifications:

- Central Limit Theorem Sample size exceeds 30 (n = 195).
- The weak law of large numbers is applicable: The sampling distribution of means will tend to approximate normality
- Visual Evidence: The Q-Q plot illustrates a satisfactory level of normality Transformations have enhanced the distributional shape Residual deviations are primarily observed in the tails.
- Practical Considerations: Economic datasets frequently exhibit some degree of non-normality
- Regression methodologies are generally robust Bootstrap techniques have been employed for the estimation of standard error

Comparison of Standard Error estimates with bootstrapped estimates:

	Predictor	Mean Coefficient	Standard Error	2.5th Percentile	97.5th Percentile
0	Intercept	0.000000	0.000000	0.000000	0.000000
1	cdepc	0.234950	0.082266	0.053210	0.375478
2	eys	-0.399780	0.107123	-0.636864	-0.215341
3	gdi	-0.379958	0.081057	-0.588628	-0.260557
4	hdi	1.792659	0.303347	1.275095	2.456072
5	lfprf	0.336594	0.051260	0.250481	0.447308
6	lfprm	-0.182882	0.034505	-0.255517	-0.121888
7	leb	-0.193845	0.077207	-0.373645	-0.046580
8	mys	-0.477073	0.103176	-0.708822	-0.292973
9	sspf	0.038224	0.031260	-0.026014	0.093807

(bootstrapped estimates)

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.939e-18	0.034	-2.06e-16	1.000	-0.067	0.067
cdepc	0.2468	0.045	5.478	0.000	0.158	0.336
eys	-0.4044	0.083	-4.898	0.000	-0.567	-0.242
gdi	-0.3644	0.054	-6.773	0.000	-0.471	-0.258
hdi	1.7865	0.220	8.125	0.000	1.353	2.220
lfprf	0.3351	0.049	6.887	0.000	0.239	0.431
lfprm	-0.1822	0.042	-4.339	0.000	-0.265	-0.099
leb	-0.1876	0.102	-1.847	0.066	-0.388	0.013
mys	-0.4849	0.102	-4.734	0.000	-0.687	-0.283
sspf	0.0381	0.038	1.004	0.317	-0.037	0.113

(estimates when normality assumed)

Key takeaways:

- Strong consistency between bootstrapped and normal estimates
- Most variables are highly significant ($p < 0.05$)
- Bootstrapped standard errors are generally larger, indicating more conservative estimates
- Bootstrapped intervals are generally wider, suggesting more conservative inference
- Both methods agree on direction and approximate magnitude of effects
- Bootstrapped estimates provide more robust uncertainty quantification
- Results are robust across both methods
- Bootstrapping provides more conservative estimates

- Key relationships maintain significance under both approaches
- Confidence in results due to consistency across methods

Heteroscedasticity analysis

Breusch-Pagan and White tests:

[97]:

	Metric	Value
0	LM Statistic	11.187098
1	LM-Test p-value	0.263102
2	F-Statistic	1.251038
3	F-Test p-value	0.266584

Decision:

- Both p-values > 0.05 significance level
    - LM-Test p-value = 0.263 > 0.05
    - F-Test p-value = 0.267 > 0.05
- Therefore, fail to reject the null hypothesis

There exists a lack of substantial evidence indicating the presence of heteroscedasticity. The postulate of constant variance is duly fulfilled. The residuals of the model seem to exhibit a uniform variance. The ordinary least squares (OLS) estimations demonstrate reliability. There is no necessity for employing robust standard errors or weighted least squares techniques. The model adheres to this crucial assumption pertinent to linear regression.

This implies that our model is appropriately specified concerning the variance structure, thereby rendering standard inferential methodologies suitable.

Model Selection

Comparison of Different Subsets

- Various combinations of predictors were tested using a systematic approach.
- The best model was identified through Mallows' Cp criterion, which balances model fit and complexity.
- The ANOVA test was performed to compare restricted and unrestricted models, confirming that the selected model significantly improved fit without unnecessary complexity.

Best Mallow's CP = 16.05643215600233
Best Subset of Variables = ['cdepc', 'eys', 'gdi', 'hdi', 'lfprf', 'lfprm', 'leb', 'mys']

Regression Model Comparisons:

OLS Regression Results

Dep. Variable:	gnip	R-squared:	0.790
Model:	OLS	Adj. R-squared:	0.780
Method:	Least Squares	F-statistic:	77.21
Date:	Thu, 26 Dec 2024	Prob (F-statistic):	7.12e-58
Time:	17:39:15	Log-Likelihood:	-124.65
No. Observations:	195	AIC:	269.3
Df Residuals:	185	BIC:	302.0
Df Model:	9		
Covariance Type:	nonrobust		

OLS Regression Results

Dep. Variable:	gnip	R-squared:	0.789
Model:	OLS	Adj. R-squared:	0.780
Method:	Least Squares	F-statistic:	86.73
Date:	Thu, 26 Dec 2024	Prob (F-statistic):	1.22e-58
Time:	17:35:49	Log-Likelihood:	-125.18
No. Observations:	195	AIC:	268.4
Df Residuals:	186	BIC:	297.8
Df Model:	8		
Covariance Type:	nonrobust		

- Both models show similar R-squared and adjusted R-squared values, indicating consistent explanatory power.
- The second model has slightly better AIC and BIC values, suggesting a more efficient model with fewer predictors.

Interpretation of Restricted vs. Unrestricted ANOVA

[115]:
```
#anova testing for restricted and unrestricted model
anova_table = sm.stats.anova_lm(subset_model,rig_model)
anova_table
```

[115]:

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	186.0	41.223803	0.0	NaN	NaN	NaN
1	185.0	41.000534	1.0	0.223269	1.007422	0.316832

Decision:

- The p-value (0.316832) is greater than the typical significance level (0.05).
- Fail to reject the null hypothesis.

Conclusion:

- The additional parameters in the unrestricted model do not significantly improve the model fit.
- The restricted model is adequate for explaining the variability in the data.
- This suggests that the simpler model is preferable, as it provides a similar fit with fewer parameters, enhancing interpretability and reducing complexity.

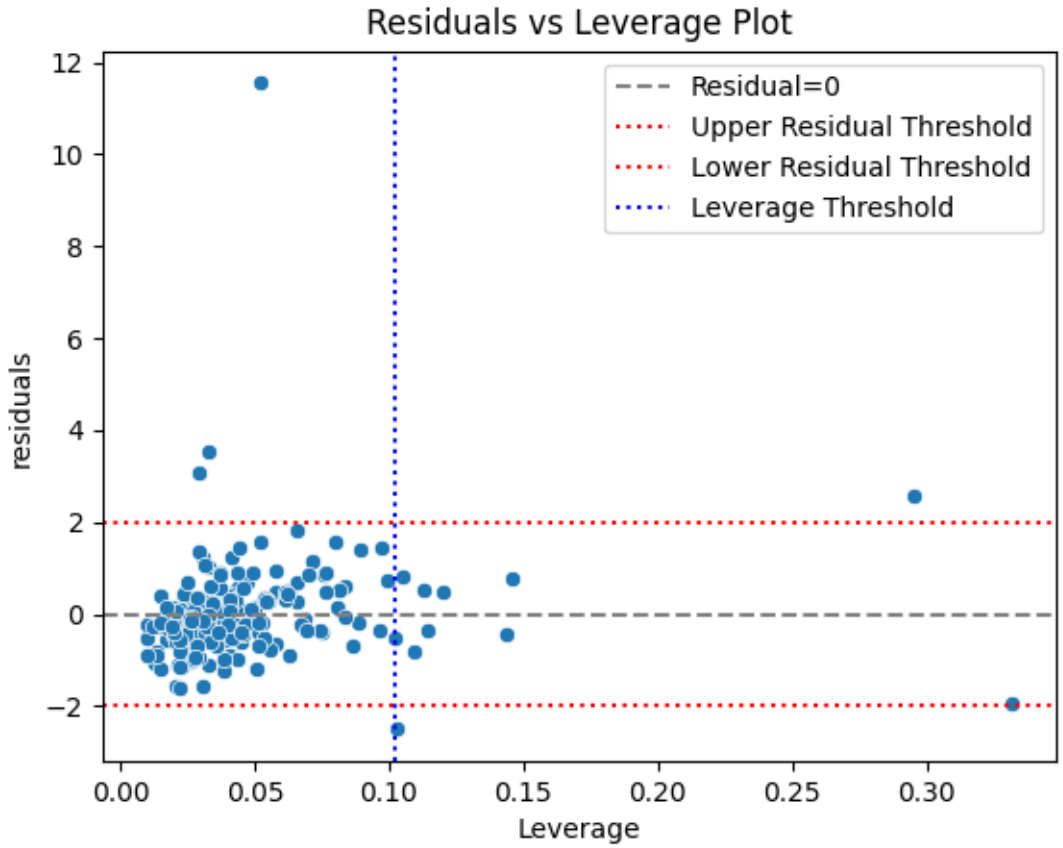
Influence Analysis

Cook's Distance Findings

- Cook's Distance identifies influential data points that significantly affect the model's predictions.
- Threshold for influence: Cook’s Distance>4nCook’s Distance>n4
- Influential points detected: [27, 81, 97, 100, 142, 157, 163, 177, 184, 192]
- These points have a substantial impact on the model's coefficients and predictions.

Leverage Points

- Leverage measures the influence of a data point based on its position in the predictor space.
- High leverage threshold: (2*(p+1))/n
- High leverage points detected: [0, 12, 24, 56, 92, 115, 142, 177, 181, 192]
- These points are far from the mean of the predictor variables and can disproportionately affect the model fit.



Impact on Model

OLS Regression Results

Dep. Variable:	gnip	R-squared:	0.789
Model:	OLS	Adj. R-squared:	0.780
Method:	Least Squares	F-statistic:	86.73
Date:	Thu, 26 Dec 2024	Prob (F-statistic):	1.22e-58
Time:	18:06:04	Log-Likelihood:	-125.18
No. Observations:	195	AIC:	268.4
Df Residuals:	186	BIC:	297.8
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.939e-18	0.034	-2.06e-16	1.000	-0.067	0.067
cdepc	0.2399	0.045	5.388	0.000	0.152	0.328
eys	-0.3950	0.082	-4.815	0.000	-0.557	-0.233
gdi	-0.3668	0.054	-6.826	0.000	-0.473	-0.261
hdi	1.7888	0.220	8.135	0.000	1.355	2.223
lfprf	0.3471	0.047	7.360	0.000	0.254	0.440
lfprm	-0.1837	0.042	-4.379	0.000	-0.267	-0.101
leb	-0.1781	0.101	-1.760	0.080	-0.378	0.021
mys	-0.4901	0.102	-4.792	0.000	-0.692	-0.288

Omnibus:	200.711	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7675.656
Skew:	3.805	Prob(JB):	0.00
Kurtosis:	32.779	Cond. No.	15.8

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

+

OLS Regression Results

Dep. Variable:	gnip	R-squared:	0.881
Model:	OLS	Adj. R-squared:	0.876
Method:	Least Squares	F-statistic:	167.1
Date:	Thu, 26 Dec 2024	Prob (F-statistic):	4.32e-79
Time:	18:06:13	Log-Likelihood:	-33.208
No. Observations:	189	AIC:	84.42
Df Residuals:	180	BIC:	113.6
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0347	0.022	-1.608	0.110	-0.077	0.008
cdepc	0.2744	0.034	8.000	0.000	0.207	0.342
eys	-0.2900	0.052	-5.560	0.000	-0.393	-0.187
gdi	-0.3534	0.040	-8.757	0.000	-0.433	-0.274
hdi	1.4316	0.142	10.075	0.000	1.151	1.712
lfprf	0.3069	0.030	10.200	0.000	0.248	0.366
lfprm	-0.1818	0.027	-6.806	0.000	-0.235	-0.129
leb	-0.1335	0.064	-2.079	0.039	-0.260	-0.007
mys	-0.3415	0.065	-5.239	0.000	-0.470	-0.213

Omnibus:	11.000	Durbin-Watson:	1.919
Prob(Omnibus):	0.004	Jarque-Bera (JB):	11.357
Skew:	0.529	Prob(JB):	0.00342
Kurtosis:	3.567	Cond. No.	15.5

- Removing influential and high leverage points resulted in a refined model:
  - **Original Model**: R-squared = 0.789, Adj. R-squared = 0.780
  - **Influence-less Model**: R-squared = 0.881, Adj. R-squared = 0.876
- The refined model shows improved fit and stability, indicating that the original model was affected by these influential points.
- The removal of these points led to a more reliable and robust model, enhancing the accuracy of predictions and interpretations.

Overall, the influence analysis highlights the importance of identifying and addressing influential data points to ensure the validity and reliability of the regression model.

Here we are treating Influence points by removing them.

Multicollinearity

VIF

	feature	VIF
0	cdepc	1.320877
1	eys	2.433301
2	gdi	1.594046
3	hdi	6.522220
4	lfprf	1.398803
5	lfprm	1.244763
6	leb	3.000055
7	mys	3.033925

Interpretation:

- VIF values below 10 indicate no severe multicollinearity issues.
- **hdi** has a VIF of 6.52, suggesting moderate multicollinearity.

Condition Indices

Condition indices assess the sensitivity of the regression coefficients to small changes in the data:

- Condition number: 15.5

Interpretation:

- A condition number above 30 indicates strong multicollinearity.
- The condition number of 15.5 suggests mild multicollinearity, which is generally acceptable.

Eigenvalues:

- Large eigenvalues: [774.85, 275.42, 138.30]
- Small eigenvalues: [3.21, 79.62, 34.14, 44.61]

Small eigenvalues indicate potential multicollinearity, but the overall condition is manageable.

Model Results

Regression Coefficients

- Intercept: -0.0347
- cdepc (Carbon Dioxide Emissions per Capita): 0.2744
- eys (Expected Years of Schooling): -0.2900
- gdi (Gender Development Index): -0.3534
- hdi (Human Development Index): 1.4316
- lfprf (Labor Force Participation Rate, Female): 0.3069
- lfprm (Labor Force Participation Rate, Male): -0.1818

- leb (Life Expectancy at Birth): -0.1335
- mys (Mean Years of Schooling): -0.3415

Statistical Significance

- All predictors are statistically significant with p-values < 0.05, indicating strong evidence against the null hypothesis of no effect.
- The intercept is not statistically significant (p = 0.110).

R-squared and Adjusted R-squared

- R-squared: 0.881
    - Indicates that 88.1% of the variability in GNI per capita is explained by the model.
- Adjusted R-squared: 0.876
    - Adjusts for the number of predictors, confirming the model's robustness.

F-statistics

- F-statistic: 167.1
- Prob (F-statistic): 4.32e-79
    - The high F-statistic and low p-value indicate that the model is statistically significant overall.

Model Interpretation

The model proficiently elucidates the correlation between Gross National Income (GNI) per capita and the designated predictors.

Positive Effects:

The variables Carbon Dioxide Emissions per Capita (cdepc) and Human Development Index (hdi) exhibit positive coefficients, implying that elevated carbon emissions and a heightened human development index correlate with increased GNI per capita. The Labor Force Participation Rate, Female (lfprf) similarly exerts a positive influence on GNI per capita, signifying the economic contributions attributable to female labor force engagement.

Negative Effects:

The variables Expected Years of Schooling (eys), Gender Development Index (gdi), Labor Force Participation Rate, Male (lfprm), Life Expectancy at Birth (leb), and Mean Years of Schooling (mys) display negative coefficients, indicating that these variables are inversely related to GNI per capita.

The model furnishes a thorough comprehension of the manner in which diverse economic and social indicators affect national income, with salient predictors providing valuable insights into potential policy ramifications.

Hypothesis Testing

ANOVA Results

Anova Testing for Relation among parameters

	sum_sq	df	F	PR(>F)
cdepc	5.591652	1.0	64.004374	1.467449e-13
eys	2.700491	1.0	30.910948	9.624104e-08
gdi	6.700027	1.0	76.691297	1.439206e-15
hdi	8.867671	1.0	101.503054	3.251724e-19
lfprf	9.089192	1.0	104.038672	1.439823e-19
lfprm	4.047318	1.0	46.327285	1.437232e-10
leb	0.377723	1.0	4.323571	3.900536e-02
mys	2.397853	1.0	27.446824	4.480939e-07
Residual	15.725446	180.0	NaN	NaN

The findings from the ANOVA analysis yield a thorough assessment of the significance attributed to each predictor within the model framework. The notably low p-values associated with the majority of predictors serve as compelling evidence against the null hypothesis, which asserts that these predictors exert no influence on GNI per capita. This observed statistical significance implies that the variables incorporated into the analysis play an essential role in elucidating the variability observed in national income.

Significance of Predictors:

The markedly low p-values for variables such as hdi and lfprf underscore their considerable effect on GNI per capita. Furthermore, even predictors exhibiting relatively elevated p-values, such as leb, continue to demonstrate significance, indicating that their presence contributes meaningfully to the overall model.

Implications:

These findings emphasize the necessity of incorporating a varied array of economic and social metrics when undertaking an analysis of national income. Policymakers may utilize these insights to prioritize initiatives that yield the most substantial impact, particularly those aimed at enhancing human development and labor force participation.

Model Comparisons

The ANOVA test substantiates the integration of all significant predictors, reinforcing their critical role within the model. This inclusive methodology guarantees that the model captures the complex dynamics inherent in economic and social determinants influencing GNI per capita.

Comprehensive Model:

By incorporating all significant predictors, the model affords a comprehensive perspective on the determinants influencing national income. This methodology mitigates the risk of oversimplification and ensures that vital variables are retained, which could otherwise result in skewed findings.

Conclusions

Key Findings

Economic Indicators and Their Relationship with Social Metrics:

- Gross National Income (GNI) per capita demonstrates a noteworthy correlation with social welfare indicators, yet it does not consistently correspond with enhanced societal outcomes across all evaluative metrics.
- Human Development Index (HDI): A substantial positive correlation suggests that elevated GNI per capita frequently associates with enhanced human development outcomes, indicative of improved life expectancy, educational attainment, and quality of living.

- Gender Development Index (GDI): A marginally non-linear association indicates that nations with elevated GNI per capita typically display higher gender equity; however, exceptions exist that weaken this correlation. Life Expectancy at Birth: Counterintuitively, this variable exhibited a negative correlation with GNI per capita, implying that increased national income may not invariably relate to proportional advancements in health outcomes.
- Labor Force Participation Rate (Female): This variable exerts a positive impact, underscoring the economic advantages derived from increased female participation in the labor market.
- Labor Force Participation Rate (Male): A negative correlation in this instance suggests that dependence on male-centric labor markets may restrict diversified economic contributions.
- Mean Years of Schooling and Expected Years of Schooling: Both variables displayed unanticipated negative correlations, potentially indicative of disparities in educational quality or inefficiencies in converting educational years into productive outcomes within certain high-income nations.
- Carbon Dioxide Emissions per Capita: A positive correlation highlights the environmental costs frequently linked to elevated GNI, suggesting unsustainable practices associated with economic expansion.

Model Performance and Refinement:

The final regression model elucidates 88.1% of the variability in GNI per capita, accompanied by an adjusted R-squared of 87.6%, signifying substantial predictive efficacy.

The exclusion of influential and high-leverage observations markedly enhanced the reliability of the model, as evidenced by a significant increase in R-squared values from 78.9% to 88.1%.

The statistical significance of all predictors ($p < 0.05$) and the absence of heteroscedasticity or severe multicollinearity further validate the integrity of the model.

Complex Interactions Between Economic Growth and Social Welfare:

Elevated GNI per capita does not invariably assure proportional enhancements in social welfare, as illustrated by negative or non-linear correlations in variables such as life expectancy and educational attainment. Social development frequently lags behind economic growth in countries characterized by significant income disparities or environmental challenges.

For Reproducibility and Verification adding link to github where dataset and codes are available:

https://github.com/NamanDudhoria/Economic-Regression-Analysis-Project