ECONOMETRICS PROJECT ON

"DETERMINANTS OF REGIONAL INCOME INEQUALITY ACROSS INDIAN STATES: A CROSS-SECTIONAL ANALYSIS"

NAME-ANWESHA DAS
ROLL NO.-07
COURSE- M.A ECONOMICS
SUBJECT- BASIC ECONOMETRICS
SUBMISSION TO- DR. NACHIKETA CHATTOPADHYAY

1. PROBLEM STATEMENT

Regional inequality has become one of the most persistent structural challenges in India's growth trajectory. Although national GDP has shown steady expansion, substantial disparities across states in per capita income, industrial development, infrastructure availability, and public expenditure continue to widen. These inter-state differences raise critical concerns about the inclusiveness and long-term sustainability of India's economic progress.

Traditional national-level indicators often mask deep variations in economic structure and developmental capacity at the state level. Variations in industrial output, social sector investment, access to reliable infrastructure such as power supply, and the availability of institutional credit—particularly to sectors like agriculture create uneven opportunities for income generation and productivity improvement. As a result, states differ significantly in their ability to convert economic growth into higher living standards.

Against this backdrop, the central question arises: To what extent do structural economic factors, public expenditure patterns, infrastructure access, and financial inclusion explain disparities in State Per Capita GDP across India?

This study seeks to address this question using comprehensive state-level data from the Reserve Bank of India (RBI), including Gross State Value Added (GVA) by Industry, Social Sector Expenditure, Per Capita Availability of Power, and Credit to Agriculture by Scheduled Commercial Banks. By empirically examining the relationship between these variables and State Per Capita GDP, the research aims to identify the economic and policy-driven determinants of regional inequality. The insights generated are intended to contribute to the broader discourse on fostering balanced and inclusive regional development in India.

OBJECTIVES :

- Assess the strength and direction of relationships between state GDP per capita and selected independent variables.
- Identify which factors most significantly contribute to or hinder economic prosperity at the state level.
- Examine whether states with higher industrial output, social spending, power availability, and agricultural credit also exhibit higher per capita incomes, or if imbalances persist.
- To offer insights on whether to prioritize industrialization, energy infrastructure, or social spending to maximize economic growth.

Hypotheses

Null Hypothesis (H_0):
There is no significant relationship between state per capita GDP and the selected predictors (industrial output, social expenditure, power availability, agricultural credit).

Alternative Hypothesis (H_1):
There is a significant relationship between state per capita GDP and one or more of the predictors, indicating that regional inequality is influenced by disparities in industrial activity, social spending, infrastructure, and agricultural financing.

2. DATA

2.1 Source

The dataset for this analysis consists entirely of state-level indicators obtained from the Reserve Bank of India's Handbook of Statistics on Indian States (HSIS), FY 2022–23.
All variable definitions strictly follow RBI's official classifications. The HSIS provides comprehensive, consistent and comparable data across states, ensuring the reliability of the empirical results.

2.2 Varibale Description

Variable Name	Role in Analysis	Definition	Units of Measurement
Per Capita Net State Domestic Product (at constant prices)	Independent Variable	Per Capita NSDP is defined as the the average income or economic output generated per person in a state, adjusted for inflation(base year 2011-12) calculated by dividing the state's total Net State Domestic Product (NSDP) by its total population., where NSDP is obtained by deducting consumption of fixed capital from GSDP.	₹ per person
Gross State Value Added (GSVA) by Industry	Dependent Variable	GSVA is the value of goods and services produced in the Industrial sector after subtracting intermediate consumption. It measures the contribution of the industrial sector to a state's economy.	₹ Crore
State-Wise Social Sector Expenditure	Dependent Variable	Social Sector Expenditure includes government spending on human development, covering education, medical & public health, water supply & sanitation, housing, urban development, welfare of SC/ST/OBC, labour welfare, social security, nutrition, rural development, etc.	₹ Crore
Per Capita Availability of Power	Dependent Variable	Measures the average amount of electricity available for consumption per person in the state annually.Per capita availability of power is defined as the total electricity available for consumption in a state (in million units) divided by the state population, Indicates power infrastructure quality and availability.	kWh per person
Credit to Agriculture by Scheduled Commercial Banks	Dependent Variable	Represents the total outstanding credit provided by Scheduled Commercial Banks to Agriculture and Allied Activities, including crop loans, term loans, dairy, fisheries, horticulture, and other priority sector agricultural lending.	₹ Crore

2.3 Descriptive statistics

```
    state              nsdp_pc        gva_industry       social_exp
Length:31          Min.   : 29909   Min.   :  2.305   Min.   :  4246
Class :character   1st Qu.: 74141   1st Qu.: 11.437   1st Qu.: 13267
Mode  :character   Median :123874   Median :107.929   Median : 54988
                   Mean   :126284   Mean   :145.232   Mean   : 70926
                   3rd Qu.:160127   3rd Qu.:189.652   3rd Qu.:112608
                   Max.   :295114   Max.   :611.830   Max.   :235607
    power_pc       credit_agri
Min.   : 366.4   Min.   :   303
1st Qu.: 657.1   1st Qu.:  3398
Median :1260.4   Median : 32894
Mean   :1314.8   Mean   : 61455
3rd Qu.:1744.1   3rd Qu.:104398
Max.   :3197.0   Max.   :288990
```

Variable	Min	Q1	Median	Mean	Q3	Max
nsdp_pc	29909	74141	123874	126284	160127	295114
gva_industry	2.305	11.437	107.929	145.232	189.652	611.830
social_exp	4246	13267	54988	70926	112608	235607
power_pc	366.4	657.1	1260.4	1314.8	1744.1	3197.0
credit_agri	303	3398	32894	61455	104398	288990

Key insights:

Per Capita NSDP(nsdp_pc):

- Range: ₹29,909 to ₹295,114
- High standard deviation (₹71,432) indicates substantial economic inequality among states.
- Median (₹123,874) is close to the mean (₹126,284), suggesting a moderately symmetric distribution.

GVA by Industry (gva_industry):

- Extremely wide range: ₹2.3 crore to ₹611,830 crore.
- Mean (₹145,232 crore) is higher than the median (₹107,929 crore), indicating right-skewed distribution with some highly industrialized states as outliers. It indicates a few highly industrialised states contributing disproportionately to the total industrial output.

- Large standard deviation (₹155,421 crore) reflects significant structural variation in state economies.

Social Sector Expenditure(social_exp):

- Range: ₹4,246 crore to ₹235,607 crore.
- Mean (₹70,926 crore) > Median (₹54,988 crore), suggesting skew toward higher-spending states.
- High variability (std dev ₹64,781 crore) highlights unequal fiscal priorities across states. Social expenditure levels vary significantly across states, suggesting wide differences in government prioritisation of welfare and human development.

Per Capita Power Availability (power_pc):

- Range: 366.4 kWh to 3,197.0 kWh.
- Median (1,260.4 kWh) is slightly below the mean (1,314.8 kWh), indicating mild right skew.
- Significant disparity evident, with some states having less than half the power access of others.

Credit to Agriculture (credit_agri):

- Range: ₹303 crore to ₹288,990 crore.
- Extreme right skew: Mean (₹61,455 crore) >> Median (₹32,894 crore),driven by a few states receiving very large agricultural credit flows.
- High standard deviation (₹83,456 crore) points to severe inequalities in agricultural credit access. Agricultural credit distribution is extremely uneven.

3. LINEARTY

3.1 Linearity check

An initial ordinary least squares (OLS) regression was estimated using the original (untransformed) independent variables:

Gross State Value Added by Industry (gva_industry)
State-Wise Social Sector Expenditure (social_exp)
Per Capita Availability of Power (power_pc)
Credit to Agriculture (credit_agri)

Model:

$$nsdp\_pc=\beta0+\beta1(gva\_industry)+\beta2(social\_exp)+\beta3(power\_pc)+\beta4(credit\_agri)+\epsilon$$

Component + Residual (CR) Plots were generated to assess linearity between each predictor and the dependent variable

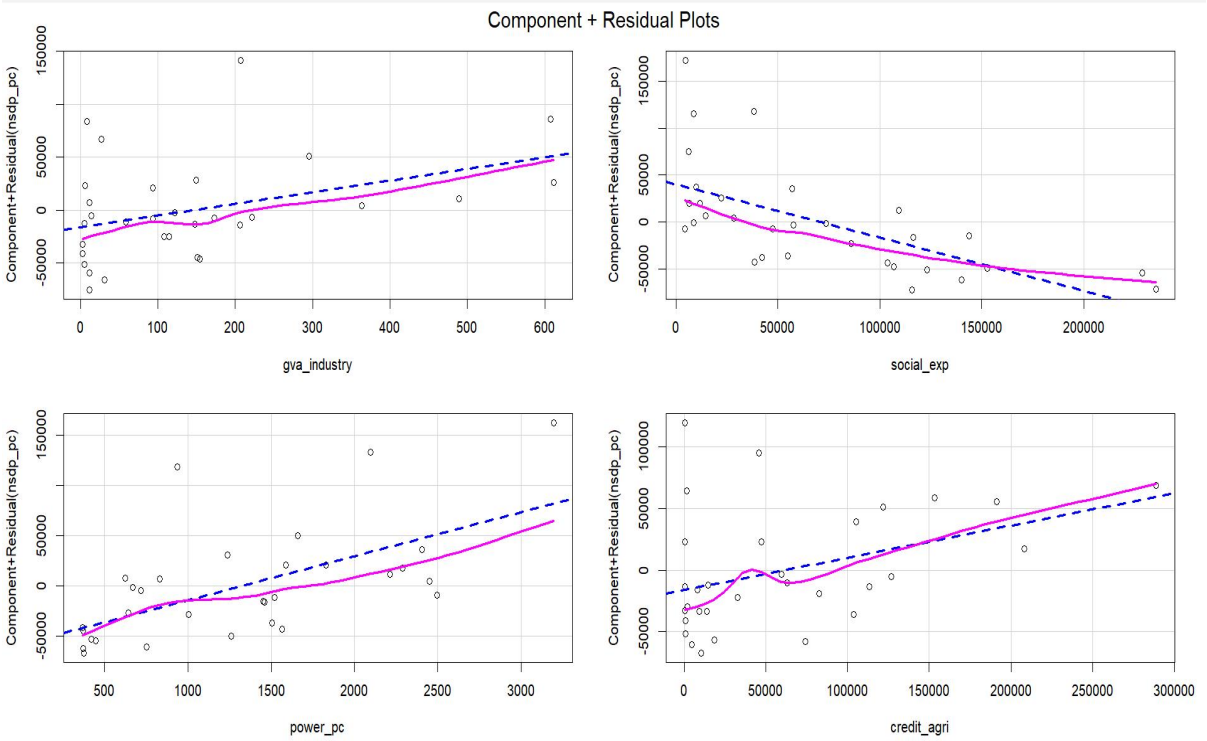

FIGURE 1: crPlots of original model

Gross State Value Added by Industry (gva_industry)

- Non-linear curvature across the entire range
- The majority of the data points are clustered near the origin.

Social Sector Expenditure (social_exp)

- Relationship clearly non-linear
- Distinct bending at both low and high expenditure levels
- The vertical spread (residuals) increases significantly at higher levels of expenditure

Per Capita Availability of Power (power_pc)

- Curvature present but less severe than others
- Partial alignment with the linear trend.
- At high levels of power consumption, the curve diverges more from the linear trend

Credit to Agriculture (credit_agri)

- Non-linear curvature.
- The vast majority of the scatter is highly concentrated at the extreme low end of the credit axis. Residuals show a structured pattern.
- The smoothed line increases sharply at low values, flattens in the mid-range, and then increases again at high values.

3.2 Linearity After Transforming Independent Variables

To address the non-linearity observed in the original model, appropriate mathematical transformations were applied on independent variables.

Independent Variable	Transformation Applied	Rationale
GSVA Industry	Logarithm	Corrects the convex functional form and stabilizes the variance.
Social Sector Expenditure	Logarithm	Converts the diminishing returns relationship into a proportional linear one.
Per Capita Availability of Power	Logarithm	Reduces variability and improves linearity.
Credit to Agriculture	Square Root	Too spread out the clustered data and create a straight line.

A new regression model was estimated using the transformed variables:

Model 2:

$$nsdp\_pc=\beta0+\beta1(ln\_gva\_industry)+\beta2(ln\_social\_exp)+\beta3(ln\_power\_pc)+\beta4(credit\_agri\_sqrt)+\epsilon$$

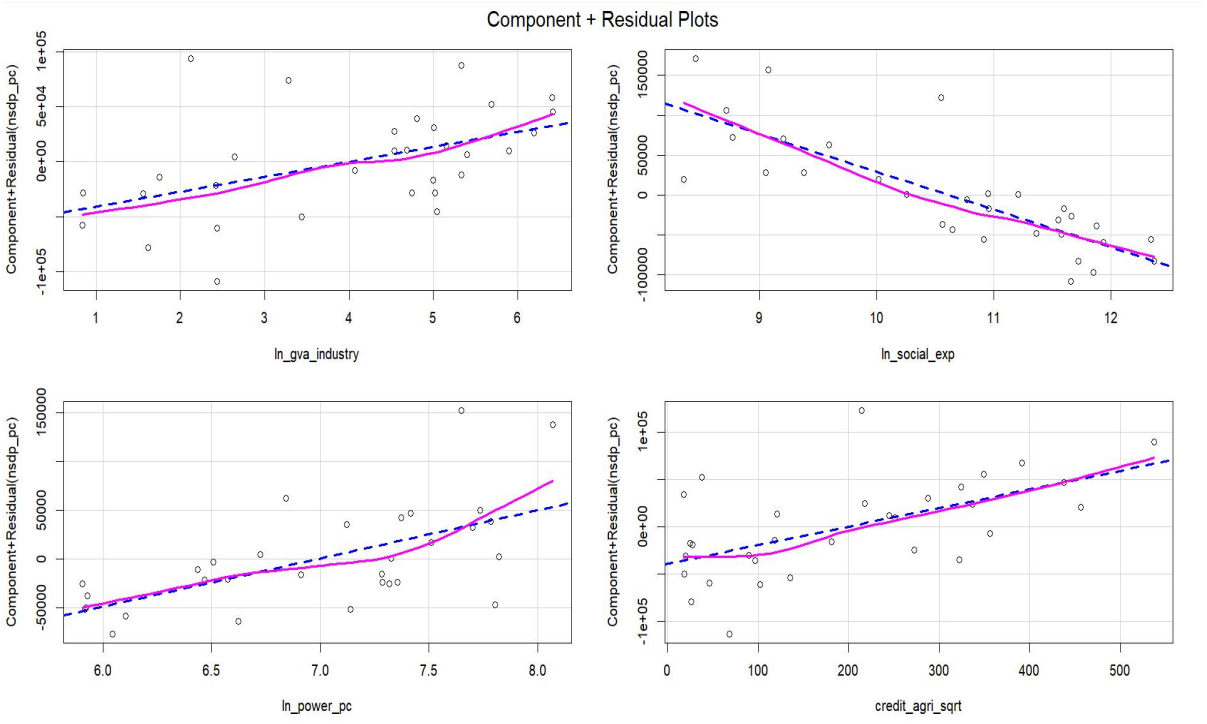

FIG 2. crPlots of model 2

The CR plots indicate substantial improvement in linearity compared to the original model. After applying logarithmic and square-root transformations, all independent variables now exhibit strong linear relationships with per capita NSDP.Therefore, the transformed model (Model 2) satisfies the linearity assumption of the classical linear regression model.

4. HOMOSCEDASTICITY

4.1 Initial Detection of Heteroscedasticity

After confirming linearity in Model 2, heteroscedasticity was assessed using:

Visual Inspection: Plot of standardized residuals squared vs. yhat
Formal Test: Studentized Breusch–Pagan (BP) test

Results for Model 2:

Breusch–Pagan Test:
BP = 9.62, df = 4, p-value = 0.04734

Visual Plot :: The plot of exhibited a discernible pattern (e.g., a funnel or cone shape) , indicating that the variance of the errors was increasing as the predicted NSDP per capita increased.

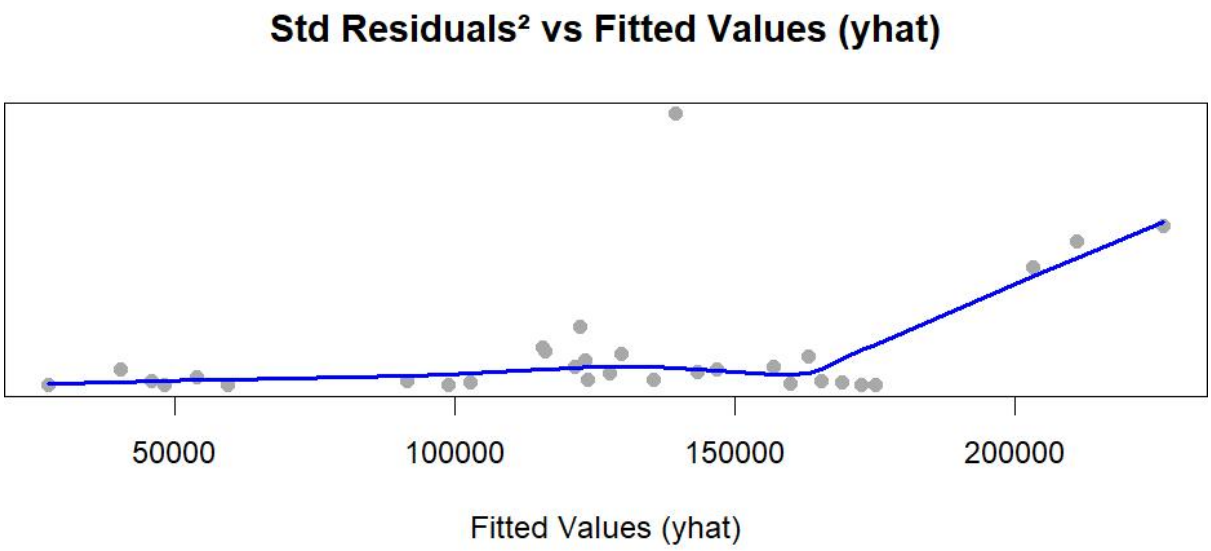

FIG.3 standardized residuals square vs yhat plot for model2

Interpretation:
Since p-value<0.05, the null hypothesis of homoscedasticity was rejected. This indicated the presence of heteroscedasticity in Model 2

4.2 Remedial by Dependent Variable Transformation

To address heteroscedasticity, a log transformation was applied to the dependent variable(per capita NSDP).A new model was estimated with log-transformed NSDP per capita:

Model 3:
$ln\_nsdp\_pc=\beta0+\beta1(ln\_gva\_industry)+\beta2(ln\_social\_exp)+\beta3(ln\_power\_pc)+\beta4(credit\_agri\_sqrt)+\epsilon$

4.3 Post-Transformation Heteroscedasticity Check

This model was re-tested using the Breusch-Pagan and Goldfeld-Quandt (GQ) tests.

Breusch–Pagan Test for Model 3:

BP = 3.9945, df = 4, p-value = 0.4067

Interpretation: Since the p-value > 0.05, we fail to reject the null hypothesis of homoscedasticity. The transformation was successful in stabilizing the variance.

Goldfeld–Quandt Test for Model 3:

GQ = 1.9924, df1 = 11, df2 = 10, p-value = 0.1438

Interpretation: Since the p-value > 0.05, the test also confirms the absence of heteroscedasticity

Visual Inspection: Plot of standardized residuals² vs. fitted values now showed random scatter with no clear pattern.

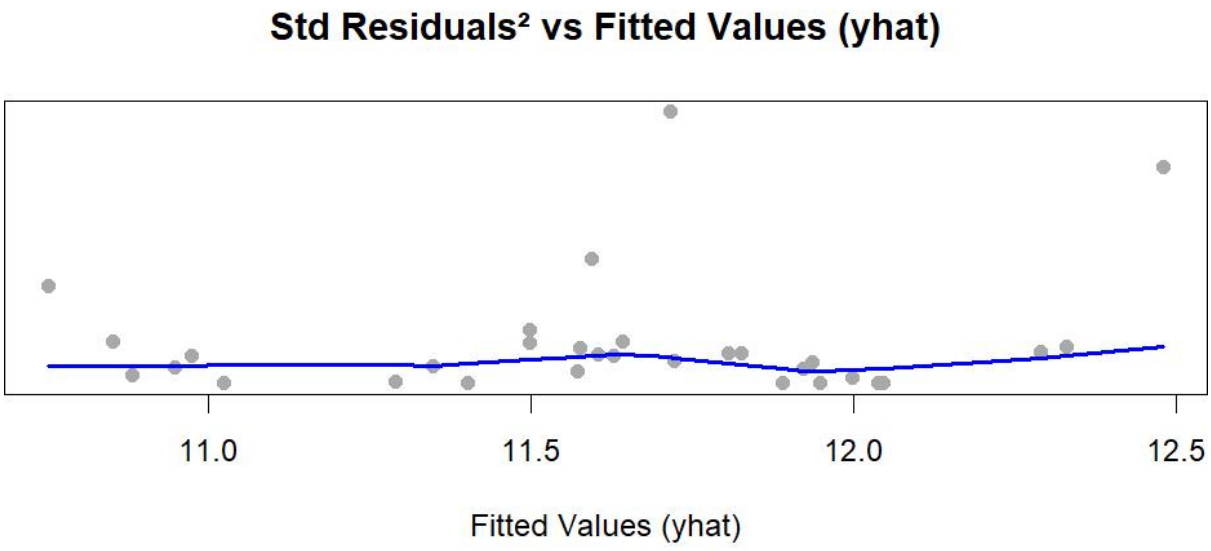

FIG.4 standardized residuals square vs yhat plot for model3

The log transformation of the dependent variable successfully eliminated heteroscedasticity, and Model 3 now satisfies the homoscedasticity assumption. The model adheres to this crucial requirement of linear regression, indicating that it is correctly specified with respect to the variance structure. Consequently, standard inferential procedures remain valid and can be applied.

5. NORMALITY

5.1 Q–Q Plot

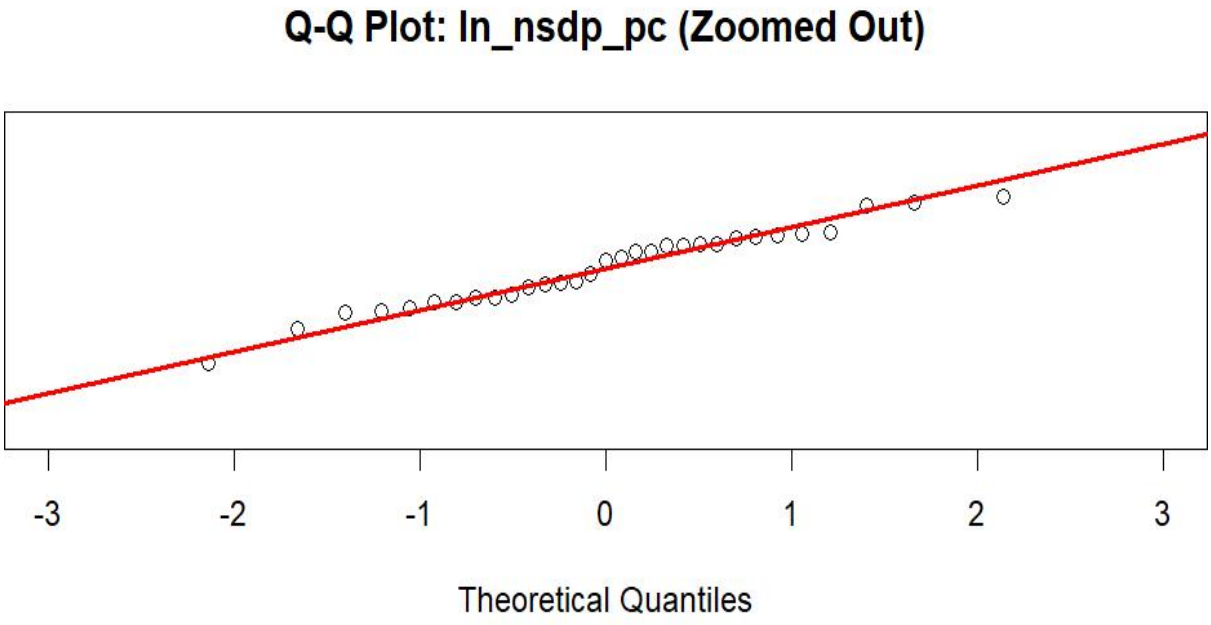

FIG.5 Q-Q Plot for normality

The Q-Q Plot of the dependent variable (ln_nsdp_pc) shows that the majority of the data points lie very close to the diagonal line , with minimal deviation at the tails. Overall, the visual evidence suggests that the distribution does not depart significantly from normality.

5.2 Statistical Tests for Normality

To statistically confirm normality, two formal tests were conducted on the dependent variable (ln_nsdp_pc):

Shapiro–Wilk Test

W = 0.97131, p-value = 0.5557

Interpretation: Since the p-value > 0.05, we fail to reject the null hypothesis that th e data arenormally distributed.

Anderson–Darling Test

A = 0.3906, p-value = 0.3606

Interpretation: Since the p-value > 0.05, we fail to reject the null hypothesis that the data are normally distributed.

Both the graphical (Q–Q plot) and formal statistical tests consistently indicate that the dependent variable (ln_nsdp_pc) follows a normal distribution. Therefore, the normality assumption is satisfied,supporting the validity of hypothesis testing, confidence intervals, and prediction intervals using standard OLS inference.

6. MODEL SELECTION

After validating the OLS assumptions (Linearity, Homoscedasticity, and Normality), the next step is to use a model selection criterion to identify the optimal subset of predictors that offers the best trade-off between model fit and complexity. Mallows' Cp criterion is used for this purpose.

Number of Predictors	**Mallows' Cp**
1	19.44
2	12.12
3	8.90
4	5.00

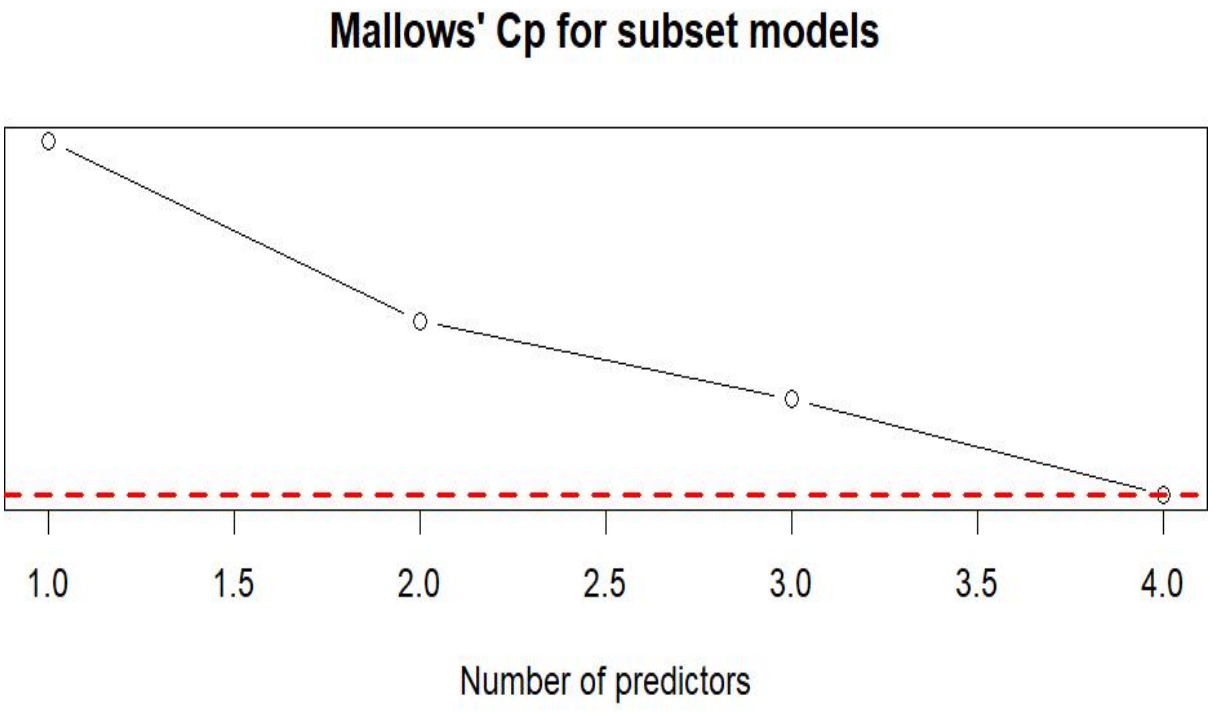

FIG.6 Mallows' Cp

Interpretation:

- The 1–3 predictor models have Cp values well above the reference line, indicating high bias or insufficient explanatory power.
- The 4-predictor model, which includes:- ln_gva_industry, ln_social_exp, ln_power_pc, credit_agri_sqrt ; achieves Cp = 5, which matches the ideal benchmar.
- The Mallows' Cp analysis clearly indicates that the full model is the most appropriate and efficient specification.It provides the best balance between goodness-of-fit and model complexity. Therefore, excluding any variable would likely omit relevant explanatory information , significantly increase the Mallows' Cp value and increase the prediction error.

7. MULTICOLLINEARITY

To assess whether multicollinearity affects the stability of coefficient estimates in Model 3, multiple diagnostic tools were employed, including Variance Inflation Factors (VIF), eigenvalue decomposition of the correlation matrix, condition number and condition indices .

7.1 Variance Inflation Factors (VIF)

Predictor	VIF
ln_gva_industry	1.627678
ln_social_exp	5.010269
ln_power_pc	1.351903
credit_agri_sqrt	4.385512

Interpretation:

• VIF values below 10 indicate no severe multicollinearity issues.
• Only ln_social_exp is close to the boundary (VIF ≈ 5), indicating moderate multicollinearity but still not severe enough to undermine model reliability.

7.2 Condition Number and Condition indices

Correlation matrix

	ln_gva_industry	ln_social_exp	ln_power_pc	credit_agri_sqrt
ln_gva_industry	1.0000000	0.5456035	0.2429006	0.3821149
ln_social_exp	0.5456035	1.0000000	0.1107352	0.8379196
ln_power_pc	0.2429006	0.1107352	1.0000000	0.3144386
credit_agri_sqrt	0.3821149	0.8379196	0.3144386	1.0000000

Key Insights:

The strongest link is between social expenditure and agricultural credit (0.838), suggesting these two policy areas are closely aligned or influenced by similar factors.
Industrial GVA shows moderate connection to social spending but weaker links to energy use and agricultural credit.
Power consumption per capita is relatively independent of the other variables, showing the weakest correlations overall.

- Condition Number: 4.643675

Interpretation:

A condition number above 30 indicates strong multicollinearity.
The condition number of 4.64 is very low .This provides strong reassurance that the overall model does not suffer from multicollinearity.

- Condition indices are derived from the eigenvalues and measure the sensitivity of each predictor to small changes in the data.

Predictor	Condition Index
ln_gva_industry	1.000000
ln_social_exp	1.577279
ln_power_pc	1.874588
credit_agri_sqrt	4.643675

All condition indices are **< 5**, further supporting the absence of harmful multicollinearity.

8. INFLUENCE ANALYSIS

To identify influential observations that disproportionately affect regression estimates, leveraging diagnostics including:

Leverage
Studentized Residuals
DFBETAS
DFFITS
Cook's Distance

8.1 Leverages

- Leverage measures how far an observation's predictor values are from the mean of all predictors. High-leverage points can unduly influence the regression line.
- Rule of Thumb: Leverage value (hii) > 2(k+1)/n =2(4+1)/31 = 0.3226.
- High-leverage states: **23, 24**

State	Leverage
23	0.3374
24	0.3582

- These states have predictor values far from the mean of the predictor variables, potentially exerting high influence on coefficient estimates.

8.2 Outliers

- Studentized residuals identify outliers in the response variable.
- Rule of Thumb: ｜ri｜>2
- Outliers: observations: **30, 31**
- These states have actual NSDP per capita values far from what the model predicts.

8.3 DFBETAS

- DFBETAS measures how much each coefficient changes when an observation is removed.
- Rule of Thumb: ｜DFBETAS| >2/ sqrt(n) =0.359
- Observations exceeding the threshold for at least one coefficient: 4, 6, 23, 24, 30, 31
- These observations significantly affect one or more regression coefficients.

8.4 DFFITS

- DFFITS detects observations that significantly alter predicted values.
- Rule of Thumb: |DFFITS| > 2× sqrt((p+1)/n) =0.802
- Influential Observations: **4, 30, 31**

state	DFFITS
4	-0.747

state	DFFITS
30	1.296
31	-1.411

- These points substantially influence their own fitted values.

8.5 Cook's Distance

- Cook's Distance summarizes the influence of each observation on all fitted values. Provides an overall measure of an observation's influence, combining both leverage and residual size to assess how much the entire set of fitted values changes when the observation is removed.

- Rule of Thumb: Di >4/n = 0.129
- Influential Observations: 30, 31
- These two observations have the strongest joint influence on all regression coefficients and predictions.

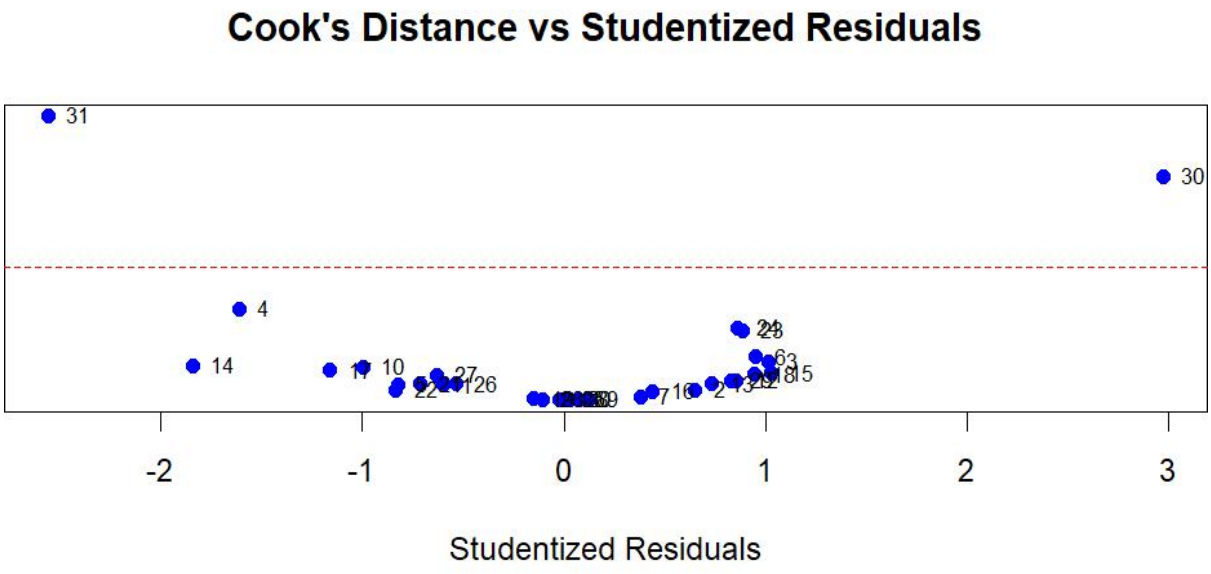

FIG.7 Cook's D vs residuals plot

Key Observations :

- Observations 30 and 31 are highly influential. Both lie well above the reference Cook's distance threshold (red dashed line).They also have large studentized residuals.
  These are the most problematic observations in the dataset.

- Observations labeled 4, 14, 17, 10, 22, 27 appear with moderate Cook's distance values, but none cross the influence threshold.
  They may have slightly unusual predictor combinations (leverage),but they do not exert strong influence on the regression coefficients.

- Majority of observations are clustered near zero.The model is stable for most of the dataset.

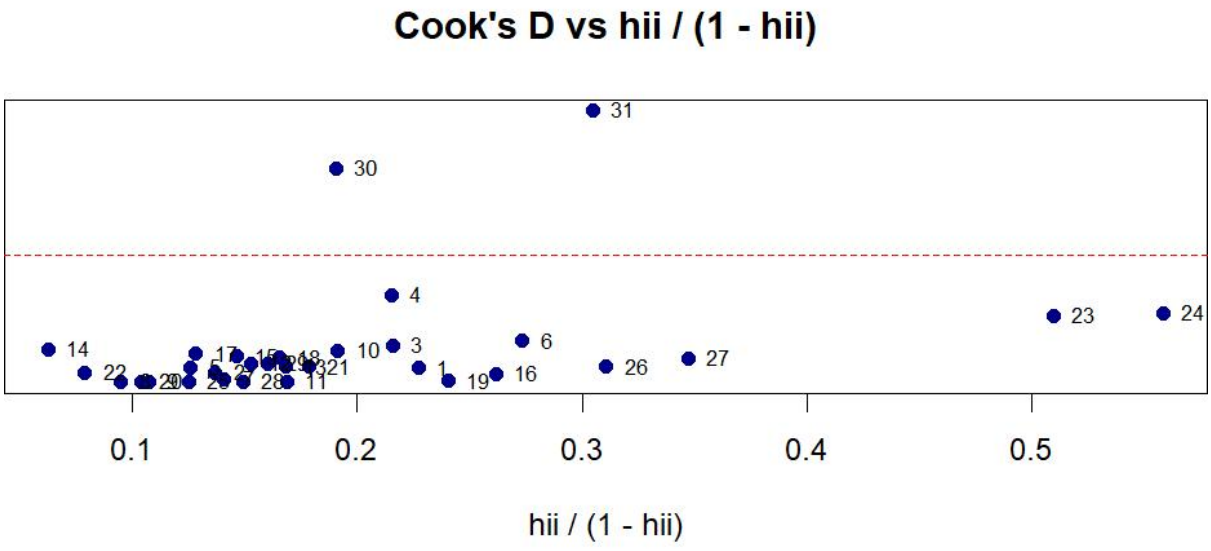

FIG.8 Cook's D vs hii/(1-hii)

Key Observations :

- This plot jointly evaluates: Influence (Cook's D on the vertical axis)
  Leverage-adjusted distance (hii/(1−hii) on horizontal axis)

- Observations 30 and 31 are the most influential points. Both lie well above the Cook's distance cutoff and also have moderately large values of hii/(1−hii).

- These two observations have an outsized impact on the regression model**.**

- Observations 23 and 24 have high leverage but low influence.
  They appear on the far right, they have extreme predictor values
  But remain below the Cook's distance threshold, their residuals are small enough that they do not distort the regression coefficients.

- They are leverage points, not influential points. Thus, they should be monitored, but not removed.

- The model is stable for the majority of the data.

8.6 Covariance Ratio

- The covariance ratio measures the impact of each observation on the precision of the regression estimates.
- Rule of Thumb: $|COVRATIOi-1| \geq 3p/n = 3\times4/31\approx0.3873\times4/31\approx0.387$

  Lower bound = 0.613
  Upper bound =1.387
- Influential Observations: Obs 30: COVRATIO = 0.319 (< 0.613) Obs 31: COVRATIO = 0.497 (< 0.613)
- Observations 30 and 31 significantly reduce the precision of coefficient estimates

8.6 Impact on Model

To assess the stability of the regression estimates, observations 30 and 31 were removed, and the model was re-estimated. The resulting specification (Model 4) shows notable improvements in both coefficient precision and explanatory power.

Model 3 Regression Results:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      12.3204307  1.3247697   9.300  9.4e-10 ***
ln_gva_industry   0.1242121  0.0430972   2.882 0.007820 **
ln_social_exp    -0.4278713  0.1047868  -4.083 0.000376 ***
ln_power_pc       0.4229550  0.1000434   4.228 0.000258 ***
credit_agri_sqrt  0.0019181  0.0007895   2.430 0.022326 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3139 on 26 degrees of freedom
Multiple R-squared:  0.7024,    Adjusted R-squared:  0.6566
F-statistic: 15.34 on 4 and 26 DF,  p-value: 1.456e-06
```

Model 4 (Excluding Influential Observations 30 & 31) Regression Results:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      13.4504780  1.0803566  12.450 5.81e-12 ***
ln_gva_industry   0.1581986  0.0358482   4.413 0.000185 ***
ln_social_exp    -0.5379941  0.0871287  -6.175 2.22e-06 ***
ln_power_pc       0.3950522  0.0841697   4.694 9.05e-05 ***
credit_agri_sqrt  0.0023961  0.0006331   3.785 0.000906 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2476 on 24 degrees of freedom
Multiple R-squared:  0.8112,    Adjusted R-squared:  0.7798
F-statistic: 25.78 on 4 and 24 DF,  p-value: 2.198e-08
```

Model Fit:

Model 3 : $R^2 = 0.7024$, Adjusted $R^2 = 0.6566$
Model 4 : $R^2 = 0.8112$, Adjusted $R^2 = 0.7798$

- Removing influential observations significantly improved model fit.

- There is no substantial change in the magnitude or direction of the coefficients,confirming the stability of the model.
- The signs remain consistent and effect sizes are stable, indicating that the model is robust to the exclusion of influential points.

Final Model Selection

The model 4 (without observations 30 and 31) is preferred because:

Higher explanatory power ($R^2 = 0.8112$)
Lower residual standard error
Stronger statistical significance across all predictors
Maintains coefficient stability

Final Model Equation:

ln_nsdp_pc=13.4505+0.1582(ln_gva_industry)−0.5380(ln_social_exp)+0.3951(ln_power_pc) +0.0024(credit_agri_sqrt)

9. MODEL RESULTS

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.50917 -0.20420  0.02046  0.19628  0.35228

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      13.4504780  1.0803566  12.450 5.81e-12 ***
ln_gva_industry   0.1581986  0.0358482   4.413 0.000185 ***
ln_social_exp    -0.5379941  0.0871287  -6.175 2.22e-06 ***
ln_power_pc       0.3950522  0.0841697   4.694 9.05e-05 ***
credit_agri_sqrt  0.0023961  0.0006331   3.785 0.000906 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2476 on 24 degrees of freedom
Multiple R-squared:  0.8112,    Adjusted R-squared:  0.7798
F-statistic: 25.78 on 4 and 24 DF,  p-value: 2.198e-08
```

Statistical Significance of Coefficients

- All predictors are statistically significant with p-values < 0.05.
- The null hypothesis that each coefficient equals zero is rejected, confirming that all selected variables significantly influence per capita state income.

R-squared and Adjusted R-squared

$R^2 = 0.8112$

Adjusted $R^2 = 0.7798$

- This means that 81.12% of the variation in log per capita NSDP across states is explained by the model.
- Adjusted for number of predictors, the model explains 77.98% of the variation, indicating strong explanatory power.

F-Statistics

F-statistic = 25.78

Prob (F-statistic) = 2.20e–08

- The high F-statistic and low p-value indicate that the model is statistically significant overall.

Model Interpretation

The model effectively captures and explains the relationship between Net State Domestic Product (NSDP) per capita and the selected economic and developmental determinants across Indian states. The refined specification (Model 4), estimated after removing influential observations, demonstrates strong explanatory capability and yields economically meaningful results.

Positive Effects

The variables Gross State Value Added (Industry), Per Capita Power Availability, and Agricultural Credit exhibit positive and statistically significant coefficients.
This suggests that:

Industrial GSVA (ln_gva_industry): Higher industrial output is associated with increased per capita NSDP, underscoring the pivotal role of industrialization in state-level income generation and structural transformation.

Power Availability (ln_power_pc):This shows the strongest positive correlation, signifying that greater access to electricity positively contributes to economic productivity.

Agricultural Credit (credit_agri_sqrt): Enhanced credit disbursement to agriculture and allied sectors supports higher rural productivity and income, thereby contributing positively to overall state income levels.

These positive coefficients collectively highlight the significance of productive capacity, infrastructure, and financial support to primary sectors in elevating economic well-being.

Negative Effects

The variable Social Sector Expenditure (ln_social_exp) carries a negative and highly significant coefficient, indicating an inverse relationship with per capita NSDP.
This suggests potential inefficiencies, lagged returns, or the possibility that states with lower incomes allocate proportionally larger shares of their budgets to social services.

Panda and Sahay (2020) documents substantial variation in social sector expenditure across Indian states and finds that the share of social sector expenditure remains systematically lower in high-income states compared to middle- and low-income states. This is a policy response where lower economic development and thus lower NSDP per capita dictates a greater need for social investment.

10. HYPOTHESIS TESTING

To formally assess the statistical significance of the explanatory variables in the final specification (Model 4), an Analysis of Variance (ANOVA) was conducted.

```
Analysis of Variance Table

Response: ln_nsdp_pc
                Df  Sum Sq Mean Sq F value    Pr(>F)
ln_gva_industry  1 0.69766 0.69766  9.5663  0.004971 **
ln_social_exp    1 1.77744 1.77744 24.3721 4.879e-05 ***
ln_power_pc      1 3.13365 3.13365 42.9683 8.848e-07 ***
credit_agri_sqrt 1 0.79391 0.79391 10.8861  0.003016 **
Residuals       24 1.75030 0.07293
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The findings from the ANOVA analysis provide a comprehensive assessment of the statistical significance of each predictor within the model framework explaining per capita Net State Domestic Product (NSDP). The consistently low p-values associated with all explanatory variables constitute strong evidence against the null hypothesis. The observed statistical significance indicates that the selected variables play a crucial role in explaining inter-state income disparities.

Significance of Predictors

The ANOVA results reveal that Per Capita Availability of Power (ln_power_pc) exhibits the highest F-statistic, underscoring its dominant role in influencing per capita NSDP. This highlights the importance of infrastructure availability in supporting economic productivity and income generation at the state level.

Similarly, Social Sector Expenditure (ln_social_exp) demonstrates a highly significant F-statistic, confirming its substantial explanatory power within the model. Although its estimated coefficient is negative, the strong statistical significance suggests that social expenditure is systematically associated with state income levels, likely reflecting compensatory fiscal behavior among lower-income states.

Gross State Value Added by Industry (ln_gva_industry) and Agricultural Credit (credit_agri_sqrt) are also statistically significant, indicating that industrial activity and access to agricultural finance contribute meaningfully to variations in per capita NSDP across states.

Implications

The ANOVA outcomes validate the inclusion of all four predictors in the final model. Removing any variable would result in a less complete representation of the drivers of regional inequality. These findings emphasize the necessity of incorporating a broad set of economic, infrastructural, and fiscal indicators when analysing regional income disparities. The statistical significance of all predictors suggests that state income outcomes are shaped by multiple interrelated factors rather than a single dominant driver.This approach ensures the model captures multiple dimensions of development.
Policymakers may draw on these insights to prioritize infrastructure development, industrial expansion, and targeted financial support.

Model Validation and Completeness

The ANOVA results substantiate the inclusion of all explanatory variables in the final specification, reinforcing their relevance within the model. This inclusive approach ensures that the regression captures the multifaceted dynamics underlying regional economic performance

11. CONCLUSION

Key Findings

This study examined the determinants of regional economic inequality across Indian states using a cross-sectional regression framework. The final model refined through rigorous diagnostic testing reveals several statistically and economically significant relationships.

Industrial Output exerts a positive and significant influence on per capita Net State Domestic Product (NSDP), underscoring the central role of industrialisation in driving state-level economic prosperity.

Power Availability emerges as the strongest positive predictor, highlighting the critical importance of infrastructure provision as a key enabler of productivity and income growth across states.

Agricultural Credit displays a positive association with per capita NSDP, indicating that improved access to formal credit in agriculture and allied sectors contributes meaningfully to broader economic performance.

In contrast, Social Sector Expenditure exhibits a negative and statistically significant coefficient, suggesting that higher social spending may coincide with lower income levels in the short run. This relationship is likely reflective of redistributive fiscal behaviour, efficiency differentials across states, or lagged returns to social investment rather than a direct adverse impact of social expenditure on income levels.

Overall, the model explains 81.12 per cent of the variation in log per capita NSDP, confirming strong explanatory power and robustness after addressing influential observations.

The results highlight that economic prosperity, measured by per capita NSDP, does not uniformly align with all development-related expenditures. While industrial activity and infrastructure investment exhibit clear positive returns, social sector expenditure displays a more nuanced inverse relationship in a cross-sectional context. This finding suggests that economic growth and social development do not automatically progress in tandem and require deliberate policy coordination, improved expenditure efficiency, and long-term planning to ensure mutually reinforcing outcomes.

Limitations

Despite its contributions, the study is subject to several limitations. First, the cross-sectional nature of the analysis captures relationships at a single point in time, limiting insights into dynamic and long-term effects. Second, important factors such as governance quality, urbanisation, institutional efficiency, and geographical characteristics were not explicitly incorporated and may influence regional income outcomes.

Finally, although variable transformations were necessary to satisfy econometric assumptions, they complicate the direct interpretation of coefficients in policy-oriented terms. Future research could address these limitations by employing panel data techniques, incorporating institutional and governance indicators, or using spatial econometric models to account for inter-state spillovers and regional interdependence.

For Reproducibility and Verification adding link to github where dataset and codes are available:
https://github.com/Ann-oying/regional-income-inequality-india