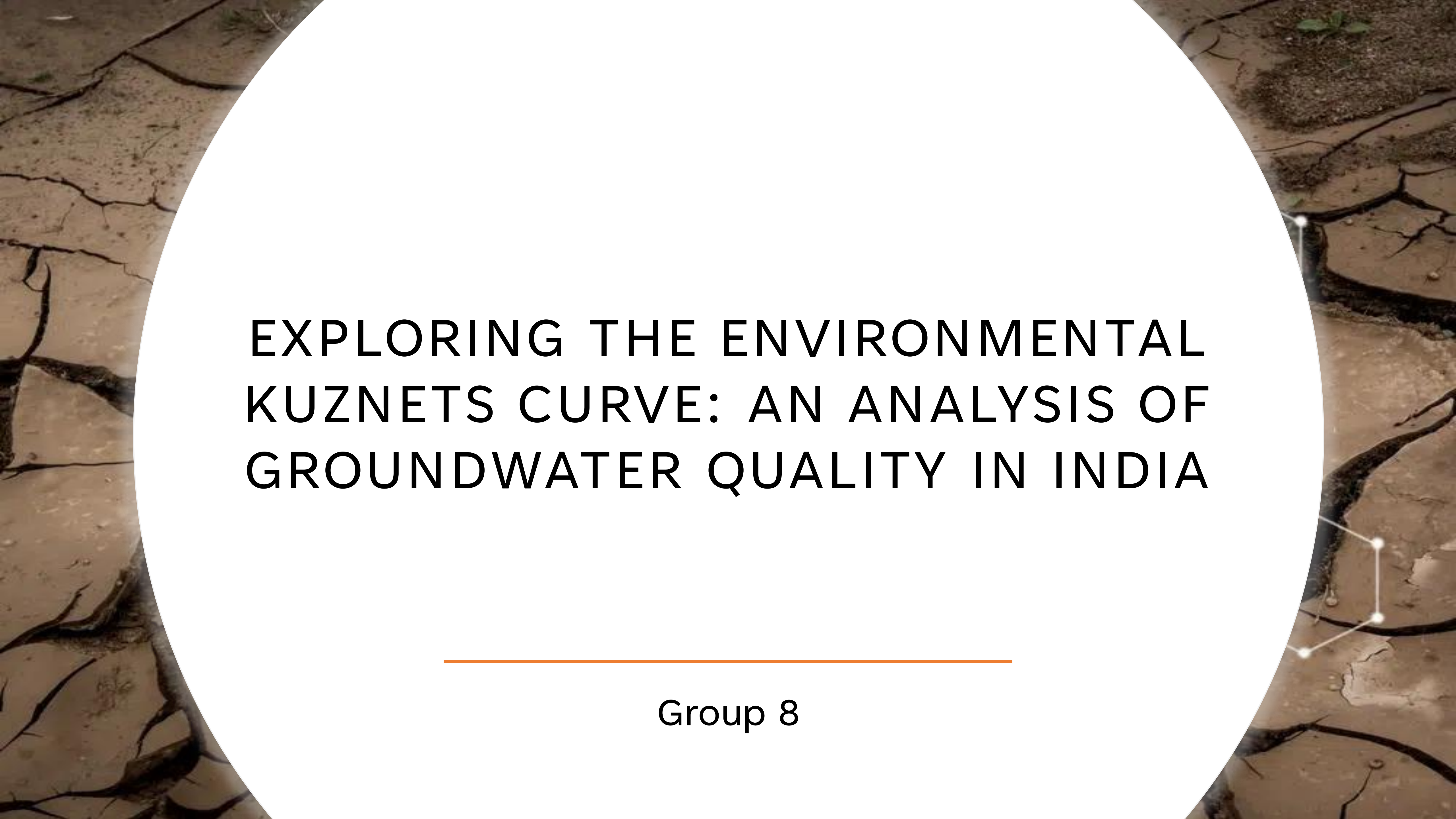


EXPLORING THE ENVIRONMENTAL KUZNETS CURVE: AN ANALYSIS OF GROUNDWATER QUALITY IN INDIA

Group 8

INTRODUCTION

The Environmental Kuznets Curve (EKC) is a theory that suggests a potential relationship between economic growth and environmental quality, it suggests that pollution other environmental issues initially increase with economic development but eventually decline as countries reach a certain level of wealth.

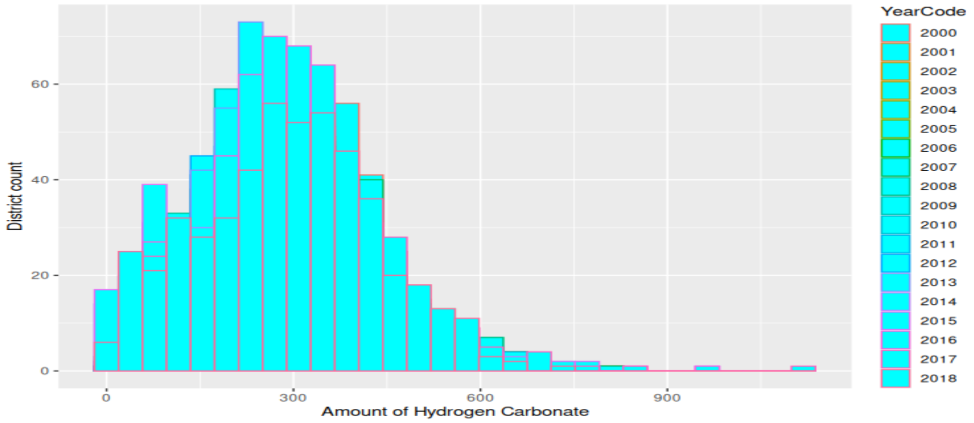
In our study:

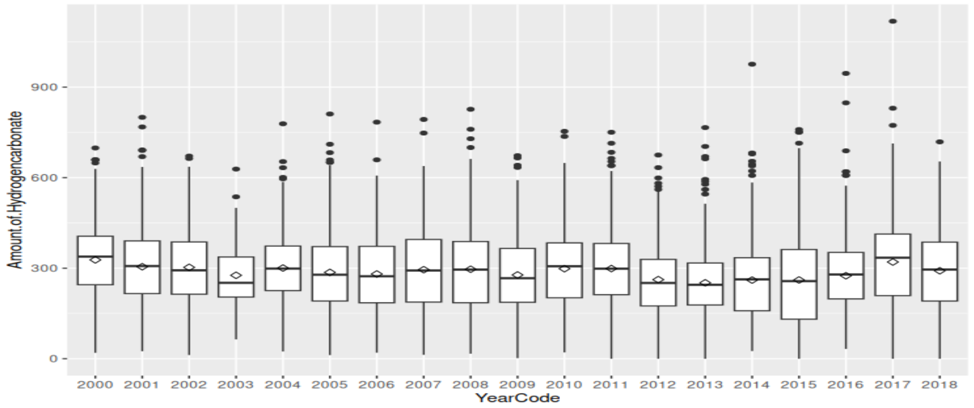
- We aim to explore the EKC hypothesis in the context of groundwater quality in India, a country where economic growth has been rapid in recent years and where groundwater is a vital resource for agriculture, drinking water, and industry.

- We investigate the link(s) between groundwater quality [as measured by the **concentration of Hydrogen carbonate (HCO3-)**] and various economic, social, and environmental factors, such as **state development, income inequality, education, and poverty**.

- We had 22 variables as groundwater quality indicators, among them we chose **concentration of HCO3-** as our indicator because its distribution was **closest to normal distribution** (shown in right) which would lead our **OLS estimators** to be more **accurate.** Furthermore, based on its box-plot, it had least number of outliers which made it an ideal candidate for our study.

- We have a **district-level time series** dataset having *district_year* as a unique identifier for each entry on which we will be performing our analysis.





VARIABLE DEFINITIONS

EQI

Environment quality indicator; contains the value of amount of HCO3- for each datapoint.

SDP

The *State Domestic Product* at constant prices. "sdp2" denotes it squared. "sdp3" denotes it cubed. "log_SDP" denotes log of SDP.

Ginni.Index

It is a measure of income or wealth inequality, with values ranging from 0 (perfect equality) to 1 (perfect inequality).

MarginOfVictory

It contains the values of margin of victory (in terms of %age / 100) of the winning political party

percentagePoverty

It contains the %age (on a scale of 0 to 1) of population which is in poverty

HDI

It takes into account factors such as life expectancy, education, and per capita income to provide a measure of overall well-being and development in a country.

literacyRate

It refers to the percentage of the population who can read and write in a particular language, typically over the age of 15

Poverty_Literacy

The interaction term between percentagePoverty and Literacy Rate; _percentagePoverty * literacyRate._

Literacy_HDI

The interaction term between HDI and Literacy Rate; HDI * literacyRate.

Unlocking Insights: Regression Analysis

We run our baseline regression model which is,
(EQI)i,t = β0 + β1*SDP(i,t) + u(i,t); u(i, t): random error.

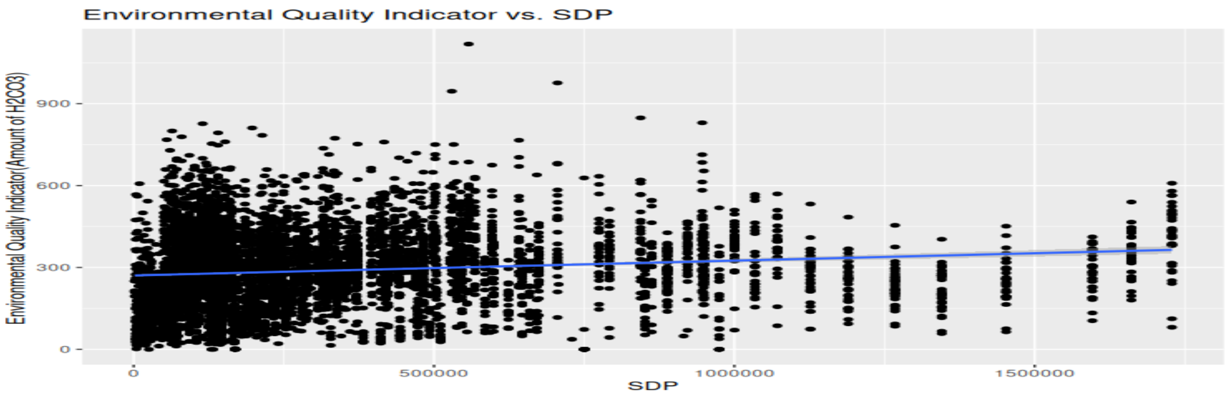
Dependent Variable : EQI (i.e. Amount of Hydrogen Carbonate) , N = 7179, R2 = 0.0161

Explanatory Variables	**Coefficient**
SDP	0.00005 ***
Intercept	270.8 ***

The **average value** of our EQI when SDP is 0 comes out to be 270.8 and the **average increase** in our EQI with a unit increase in SDP, holding other factors constant, comes out to be 0.00005. The standard error is 136.2 which is **huge.** Value of the goodness-of-fit statistic is 0.016.
Based on our above analysis, we can conclude that our current model **is highly inefficient and not a good starting point for our study.**

We visualize the model residuals to analyse a **potential non-linear relationship** between SDP and EQI:

Suggests a **non-linear relationship between economic development and environmental quality**. The negative sign implies that the relationship between SDP and EQI is **inverted U-shaped**, with the maximum EQI occurring at some intermediate level of economic development.



Based on the above plot, we find that there is a possibility that SDP and EQI might be **non-linearly correlated.**
We include higher order terms of SDP to account for this relationship. Further, adding the **gini-coefficient** as another explanatory variable, we get our **enhanced model** as follows:

EQI(i,t) = α0 + α1SDP(i,t) + α2SDP(i.t)^2 + α3SDP(i.t)^3 + α4GINI(I) + γ(i,t);

Dependent Variable : EQI (i.e. Amount of Hydrogen carbonate) , N = 6034, R2 = 0.049

Explanatory Variables	Coefficient
sdp	0.0005***
sdp2	-0.0000000007***
sdp3	0.0000000000000002 ***
ginni	38.01
Intercept	215.00 ***

*** p-value is less than 0.001

REGRESSION ANALYSIS (CONTD)

Upon running our enhanced model, we observe **statistically significant p-values** for our explanatory variables which suggests us that there is a relationship b/w our EQI and economic growth. Therefore, we choose to include more regressors (primarily based on the **power inequality**) to enhance our model. Upon doing research, using our domain knowledge and data mining, we choose to include the following 4 **new** regressors:

- MarginOfVictory
- PercentagePoverty
- HDI
- Literacy rate

Our **new enhanced model** is:

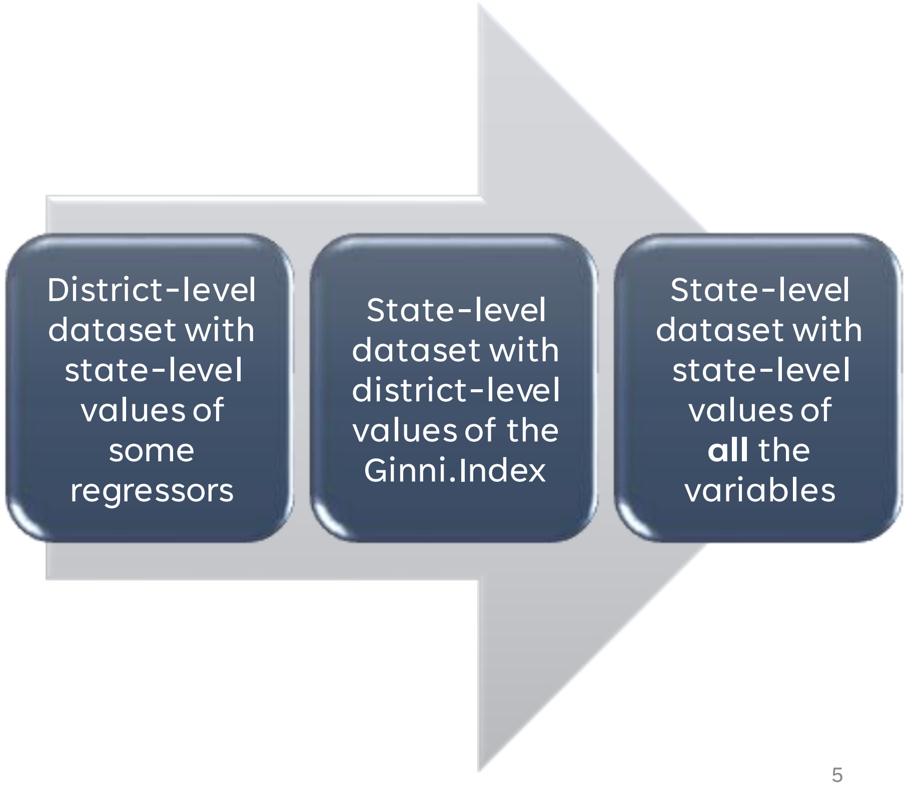
$$EQI(i,t) = \alpha0 + \alpha1SDP(i,t) + \alpha2SDP(i.t)^2 + \alpha3SDP(i.t)^3 + \alpha4GINI(I) + \alpha5MarginOfVictory(i,t) + \alpha6percentagePoverty(i,t) + \alpha7HDI(i,t) + \alpha8literacyRate(i,t) + \gamma(i,t);$$

Estimating our parameters using this model, we fine our *goodness-of-fit* statistic to be **0.17** with **3367 degrees of freedom**. Thus, we have **improved a lot in our model** in terms of estimation, and we will try to improve further.

Our original data was at a **district, year level granularity**. While finding datasets for our new regressors, we were able to find the MarginOfVictory data in same granularity but other three datasets were only available at **state-level granularity,** that too only for a few years.

Thus, we try to **compress and modify** our data to maximize the utility of all the datasets that are available to us by:

- For each state,year we assign the value of our variables to be the **average** of the values of all district,year for that state.
- For all years for which data isn't available, we assign the value which we have of the year which is closest to it
- We change the district-level Gini-coefficient we had, with the latest state-level gini-coefficient data.


District-level dataset with state-level values of some regressors
State-level dataset with district-level values of the Ginni.Index
State-level dataset with state-level values of **all** the variables

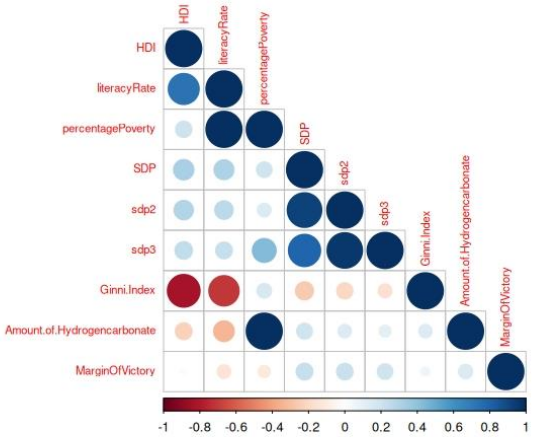
REGRESSION ANALYSIS(CONTD)

Now, we have our current regression model as:

$EQI(i,t) = \alpha0 + \alpha1SDP(i,t) + \alpha2SDP(i.t)^2 + \alpha3SDP(i.t)^3 + \alpha4GINI(I) + \alpha5MarginOfVictory(i,t) + \alpha6percentagePoverty(i,t)+ \alpha7HDI(i,t) + \alpha8literacyRate(i,t) + \gamma(i,t);$

Goodness-of-fit = 0.31 ; degrees of freedom = 320;

We have significantly improved our model. To further improve, we will check for **potential collinearity b/w our explanatory variables.** This is because high correlations between independent variables can **affect the stability of the coefficients and the overall fit of the model.** We get the following **correlation plot :**



On the basis of above plot, as well as Variance Inflation factor (VIF) list, we chose to introduce **2 interaction terms** in our model and **remove the regressors "percentagePoverty" and "MarginOfVictory"** as they didn't effect our model statistically significantly.

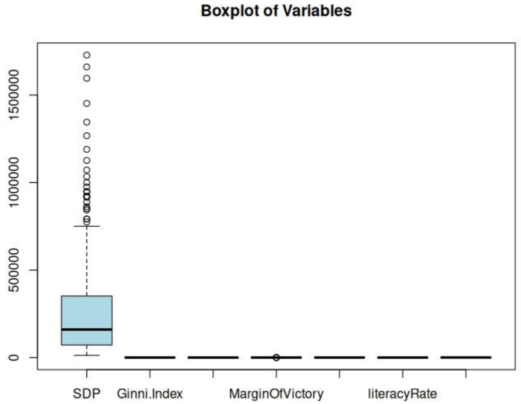
Furthermore, We used **splines** to capture the **non-linear** relationship between the response variable and the predictor variables and find that **R-squared = 0.61** in that case, which is another significant improvement.

We check for **outliers** and find out that **SDP** has many outliers. Incorporating all the conditions till now, we come to **final most enhanced model:**

$$EQI(i,t) = \alpha0 + \alpha1log\_SDP(i,t) + \alpha2SDP(i.t)^2 + \alpha3SDP(i.t)^3 + \alpha4GINI(I) + \alpha5Poverty\_Literacy(i,t) + \alpha6Literacy\_HDIi,t)+ \alpha7HDI(i,t) + \alpha8literacyRate(i,t) + \gamma(i,t);$$

Goodness-of-fit = 0.37 ; degrees of freedom = 320; (without splines)

Goodness-of-fit = 0.62 ; degrees of freedom = 320; (with splines)

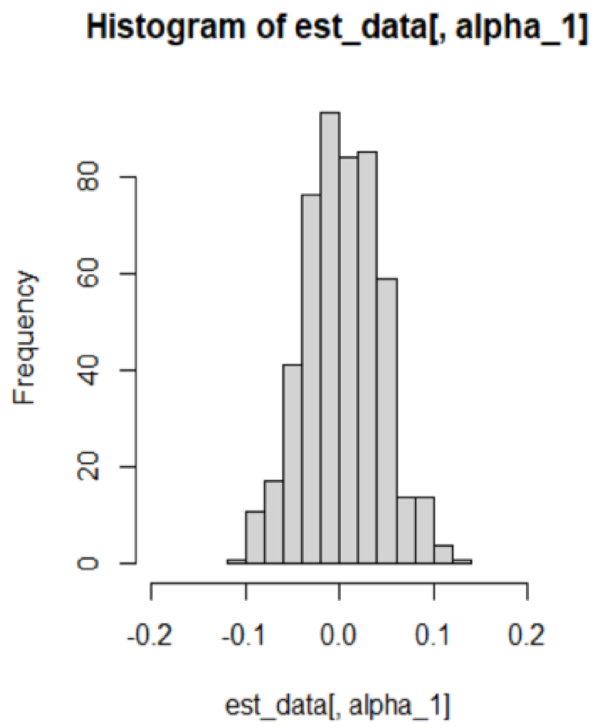


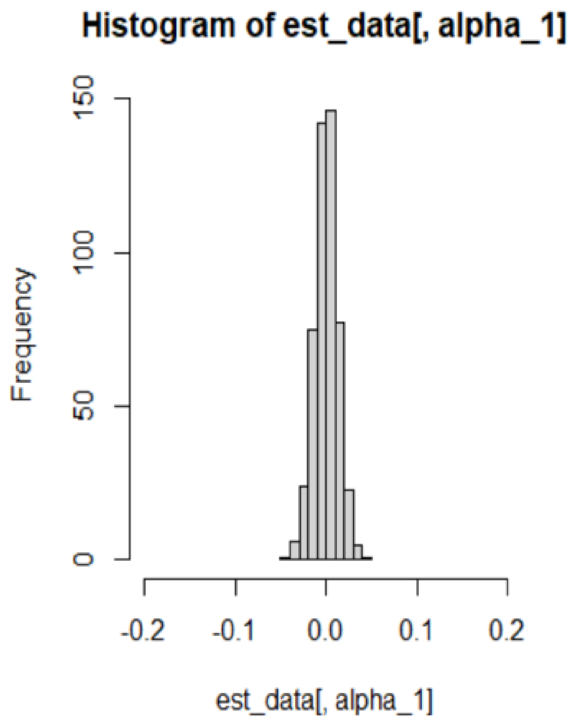
Role of standard errors: For our study, the standard errors help to determine whether the estimated coefficients are statistically significant or not. A small standard error implies that the estimated coefficient is likely to be closer to the true population value, while a large standard error implies that the estimated coefficient may be less precise and less reliable.

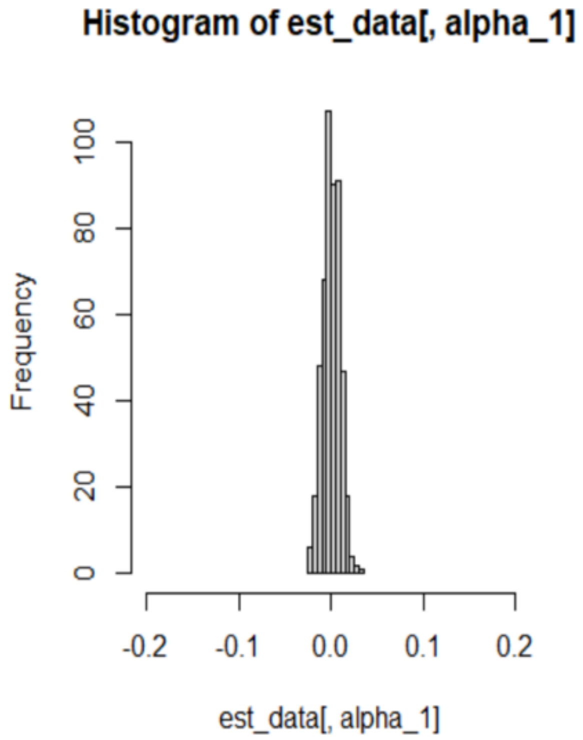
Furthermore, standard errors also help in assessing the precision of our estimated coefficients. If the standard errors are low, then we can be confident that our estimates are precise, and vice versa. Standard errors also help us to identify influential observations, which are observations that have a large effect on the estimated coefficients.

MONTE-CARLO SIMULATION

- We run Monte- Carlo simulation for sample sizes 500, 1000 and 2000.
- With increase in number of sample sizes, standard deviation decreased and estimated value came close to true value.
- With the increase in sample size, histogram gets skinnier and more concentrated around true value.







TESTING FOR BREAK (T-TEST AND CHOW TEST)

We have divided states into five groups, i.e., Northern, Southern, Central, Western, and Eastern Region. And then, we performed the t-test and Chow test by taking two state groups at a time.

In the case of the t-test we have rejected the null hypothesis (mean EQI of region A = mean EQI of region B) while testing the hypothesis for (Northern, Southern) , (Eastern, Western), (Northern, Central), (Northern , Eastern) , (Southern, Eastern) as the p-value is less than significance level (5%) in all the cases indicating structural break in mean EQI. We have accepted the null hypothesis for (Central ,Southern), (Northern, Western) , (Southern, Western) as the p-value is more than significance level.

For chow test, we have got p-value less than significance level in all the cases which concludes that there are structural breaks in the data, implying that coefficient of regression models will differ for all the region.

EQI(I,t) = α_0+α_1log_SDP(i,t) + α_2Gini.Index(i,t) + α_3sdp2(i,t) + α_4sdp3(i,t) + α_5 HDI(i,t) + α_6 literacyRate(i,t) + α_7 Poverty_literacy(i,t) + α_8 Literacy_HDI(i,t) + ui,t

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

- Post running our linear regression model we went ahead to check how close the maximum likelihood estimators of the coefficients of our regressors are.
- Upon running the MLE we saw that the coefficients are the similar as that of the OLS estimates, but, there is indeed a difference in the standard deviation value which here is less indicating that the MLE is more precise than the OLS estimators.
- This preciseness is induced because of not accounting for the loss of the degrees of freedom in the variance calculation of our model while using MLE.
- The results arrived upon by using MLE did not deviate much from the previous results since we used Amount of Hydrogencarbonate as our EQI which was indeed the closest to the Gaussian Normal Distribution assuming which we have carried out the MLE procedure.

Explanatory Variable	Coefficient Estimates using MLE
SDP	63.55
Ginni.Index	-110.53
MarginOfVictory	-2.10e-10
percentagePoverty	1.29e-16
HDI	-218.7
LiteracyRate	-658.96
Poverty_Literacy	-265.03
Literacy_HDI	69.51
Standard Error	98.65

NEW FINDINGS / CONCLUSION

- Our study examined the Environmental Kuznets Curve (EKC) hypothesis in the context of groundwater quality in India, a rapidly growing economy. We initially used a baseline regression model, but found it to be inefficient with low goodness-of-fit and high standard error. To improve our model, we included additional regressors based on power inequality, adjusted data granularity, and considered variables such as log_SDP, GINI, Poverty_Literacy, Literacy_HDI, HDI, and literacyRate. Our final enhanced model showed significant improvement with higher goodness-of-fit statistics of 0.37 (without splines) and 0.62 (with splines) and a total of 320 degrees of freedom. We also checked for collinearity between variables and used splines to capture non-linear relationships. Overall, our findings support the existence of a relationship between economic growth and groundwater quality in India, aligning with the EKC hypothesis.

- Furthermore We have also tested for structural breakdown and we found that structural breakdown exist and level of hydrogen carbonate in groundwater impacted based on which part of India it lies.

- We also conducted a maximum likelihood estimation to estimate the parameters of our model. This approach allowed us to obtain the most likely values for our model's parameters based on the observed data, and helped us make statistically sound inferences about the relationship between our explanatory variables and the response variable.

- Our analysis reveals that there is a significant difference in the average environmental quality across state groups. Further investigation through the Bartlett test of homogeneity of variances suggests that the variance in environmental quality is not equal across state groups (Bartlett's K-squared = 38.155, df = 4, p-value = 1.041e-07).

- If the variance significantly differs across state groups, the assumption of homoscedasticity of the residuals in OLS is violated. OLS assumes that the variance of the errors is constant across all levels of the independent variables.

- To account for this heteroscedasticity in the data, we have used a groupwise Feasible Generalized Least Squares (FGLS) estimation strategy. This estimation method takes into account the variance structure of the data by estimating separate error variances for each group and allows for unbiased parameter estimates even when the variances differ across groups. This showed the **R-squared of 0.38** which is an improvement over our previous OLS and MLE parameter estimations :-)

THANK YOU

Group 8:

- Ashutosh Gera (2021026)
- Anubhav Patel (2019148)
- Dev Mann (2021382)
- Razafiamiadana Jacqueline Precilia (2021411)