Analysing the impact of Inequalities on Ground Water Level

Arnav Agarwal - 2021235
Hitesh Vatsayan - 2021392
Manas Gupta - 2019368
Siddharth - 2021424

Introduction

- India is one of the World's fastest growing economies. Such development often comes at a cost to the environment.
- According to the Kuznets curve there is an inverted U relationship between environmental degradation and GDP growth. This implies there is a point at which the level of degradation in the country comes even as the aggregate output keeps increasing.
- **Ground Water Level** is an important indicator of the quality of the environment; it indicates availability of water,prevalence of droughts and quality of water.
- Thus ground water level is a good proxy for quality of the environment and can be used to model the relationship between the level of environmental degradation and economic growth. Further it is also important to analyse whether social and economic inequalities also have an impact on environmental degradation.
- This is exactly what we aim to do in our project. We aim to first test the impact of aggregate output and SDP on level of groundwater and then see how inequalities tend to impact the quality of the environment.

Variables of Interest

Data Variable	Description	Symbol
Ground Water Level	Level of Ground Water-acts as an indicator for the quality of the environment	GWL
State Domestic Product	Aggregate economic output of the state	SDP
Gini Index	Indicator of income of wealth distribution that ranges between 0 and 1 0=Perfect Equality 1=Perfect Inequality	gini
Capital Expenditure	Spending on infrastructure projects and investments in technology and productivity	cap_exp
Aggregate Expenditure	Total Amount of Spending on goods and services in the country. Closely linked to the economic activity of the state. Indicates Economic Inequality between states.	agg_exp
Election Margin	Indicative of the margin with which a political party won the last elections. Indicator of Power Inequality	electionMargin

Baseline Model

$$GWL_{i,t}=\alpha_0 + \alpha_1 SDP_{i,t}+\alpha_2 SDP_{i,t}^{2}+\alpha_3 SDP_{i,t}^{3}+\alpha_4 GINI_i+\gamma_{i,t}$$

- We found that all parameters were significant at the 95% level.
- The value of R-squared indicates how much of the variation in the independent variable is explained by the dependent variable. This value is only 0.02008 which indicates that the model only explains about 2.08% variation in ground water level.
- F-statistic of 562.9 indicates that there is a significant relationship between the predictors and the outcome variable.
- While values $\alpha_1,\alpha_2,\alpha_3$ were found to be statistically significant they were all too small
  α_1= 1.665e-05 α_2=-1.209e-11 α_3=1.568e-18
  Such values hold no practical significance,showing that GWL is independent of SDP.
- α_4=8.675e+00.This shows that a unit increase in gini increases the GWL by 8.6m. However it must be remembered that gini index can only have a maximum value of 1 so a unit increase of gini index is not practical. With increasing inequality(increase in gini index) the ground water level increases.

```
Call:
lm(formula = GWL ~ SDP + SDP_2 + SDP_3 + gini)

Residuals:
    Min      1Q  Median      3Q     Max
-11.098  -5.337  -2.555   0.653 149.859

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.857e+00  2.704e-01  10.564  < 2e-16 ***
SDP          1.655e-05  9.462e-07  17.489  < 2e-16 ***
SDP_2       -1.209e-11  1.200e-12 -10.078  < 2e-16 ***
SDP_3        1.568e-18  4.340e-19   3.614 0.000302 ***
gini         8.675e+00  6.292e-01  13.786  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.46 on 109715 degrees of freedom
  (2678 observations deleted due to missingness)
Multiple R-squared:  0.02008,   Adjusted R-squared:  0.02005
F-statistic: 562.2 on 4 and 109715 DF,  p-value: < 2.2e-16
```

Enhanced Model

To enhance our model we included new explanatory variables margin of election victory, aggregate expenditure,capital expenditure.

$$GWL_{i,t}=\alpha_0 + \alpha_1 SDP_{i,t}+\alpha_2 SDP_{i,t}^{2}+\alpha_3 SDP_{i,t}^{3}+\alpha_4 GINI_i+\gamma_{i,t} + \alpha_5 CAP\_EXP + \alpha_6 AGG\_EXP + \alpha_7 ELEC\_MARGIN$$

We test for independence of all variables involved using the variance inflation factor. All values are close to 1 except SDP,SDP^2,SDP^3 which are expected to be closely related.

Comparing the two summaries, we can observe that adding three more regressors (capital expenditure, aggregate expenditure, and margin of election victory) and changing the level of measurement of the Gini index from district to state level did not significantly affect the R-squared value of the model. Both models have an adjusted R-squared value of around 0.04, which means that the model explains only a small portion of the total variation in the dependent variable.

However, the coefficients of the Gini index regressor differ between the two models. In the first model (with district-level Gini), the coefficient of the Gini index is 20.78, while in the second model (with state-level Gini), the coefficient is 11.49. This indicates that the effect of the Gini index on the dependent variable (which is not mentioned in the question) is weaker when using state-level data compared to district-level data.

```
Call:
lm(formula = GWL ~ SDP + SDP_2 + SDP_3 + electionMargin + capital_exp +
    agg_exp + newgini)

Residuals:
     Min       1Q   Median       3Q      Max
-14.329  -5.218  -2.588   0.814 150.113

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.487e-01  3.657e-01   1.227  0.21986
SDP             3.341e-05  1.018e-06  32.830  < 2e-16 ***
SDP_2          -2.440e-11  1.210e-12 -20.169  < 2e-16 ***
SDP_3           4.788e-18  4.323e-19  11.077  < 2e-16 ***
electionMargin  3.073e-02  5.350e-03   5.745 9.24e-09 ***
capital_exp     7.517e-05  2.028e-05   3.707  0.00021 ***
agg_exp        -1.577e-04  3.694e-06 -42.685  < 2e-16 ***
newgini         1.149e+01  8.484e-01  13.546  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.32 on 109712 degrees of freedom
  (2678 observations deleted due to missingness)
Multiple R-squared:  0.04383,   Adjusted R-squared:  0.04377
F-statistic: 718.4 on 7 and 109712 DF,  p-value: < 2.2e-16
```

State Gini

```
Call:
lm(formula = GWL ~ SDP + SDP_2 + SDP_3 + gini)

Residuals:
    Min      1Q  Median      3Q     Max
-12.090  -4.798  -2.493   0.773 148.927

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.351e+00  9.140e-01  -1.478   0.1394
SDP          2.282e-05  3.071e-06   7.429 1.29e-13 ***
SDP_2       -2.104e-11  4.521e-12  -4.654 3.35e-06 ***
SDP_3        4.820e-18  1.811e-18   2.662   0.0078 **
gini         2.078e+01  2.865e+00   7.255 4.71e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.21 on 4609 degrees of freedom
Multiple R-squared:  0.04377, Adjusted R-squared:  0.04294
F-statistic: 52.74 on 4 and 4609 DF,  p-value: < 2.2e-16
```

District Gini

Standard errors play a crucial role in estimating the precision of the regression coefficients or parameters.
The smaller the standard error the more precise the value is.
As the standard error is low in our regression model analysis, the values of our estimated coefficients are decently precise.

T-test for GWL across different state groups

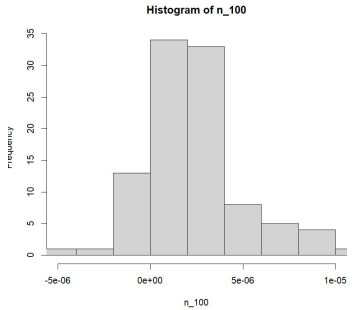
Zone Pair	t-value	Degrees of Freedom	p-value	95% Confidence Interval	Mean of Zone 1	Mean of Zone 2
North vs South	17.269	2077.6	< 2.2e-16	6.546 to 8.224	15.004	7.619
East vs West	-10.55	489.66	< 2.2e-16	-6.249 to -4.287	3.825	9.093
North vs Central	20.721	1821.6	< 2.2e-16	7.683 to 9.290	15.004	6.517
Central vs South	-6.228	1212.5	6.50E-10	-1.448 to -0.754	6.517	7.619
North vs East	27.738	1715.4	< 2.2e-16	10.388 to 11.969	15.004	3.825
North vs West	9.2737	1159.8	< 2.2e-16	4.660 to 7.161	15.004	9.093
South vs East	23.548	910.24	< 2.2e-16	3.477 to 4.110	7.619	3.825
South vs West	-2.838	570.16	0.004701	-2.494 to -0.454	7.619	9.093

Chow-test for GWL across different state groups

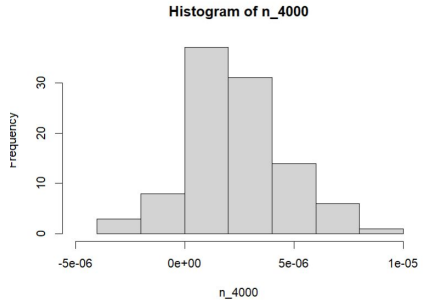
Zone Pair	F-statistic	p-value
North-South	53.859	< 2.2e-16
East-West	15.938	2.89E-15
North-Central	46.515	< 2.2e-16
South-Central	31.487	< 2.2e-16
North-East	52.896	< 2.2e-16
South-West	15.615	6.44E-15
East-Central	120.84	< 2.2e-16
West-Central	6.2109	1.06E-05

The p-value for all the pairs is extremely small and are thus significant.
We can see that the mean of North zone is 15.004 which is the highest value and for the west zone it is 3.825 which is the lowest.

As evident, the p-value is extremely small and thus the means amongst the different zones are significantly different.
The high value of F-statistic implies the high difference in the regression coefficients.




```
> sd(n_100)
[1] 2.684609e-06
> mean(n_100)
[1] 1.971999e-06
```




```
> sd(n_4000)
[1] 2.092379e-06
> mean(n_4000)
[1] 2.032876e-06
```

The Monte Carlo Simulation was run on the data to verify the consistency of OLS estimates.We can see with increase in the value of n, the standard deviation of the OLS estimate does get smaller. However we don't see an exponential decrease possibly because the size of our data is too small to model for large values of n.

Maximum Likelihood Strategy

One assumption we make during OLS estimation is that the errors are normally distributed. However this assumption may not always hold.
For instance in this case large negative values of the error term may cause the ground water level to become negative which is not possible.
Therefore the assumption of normal errors is fallacious in our model.
In such a situation the Maximum Likelihood strategy comes to our rescue. We can now assume a different distribution of errors, one that we feel is more likely for our model.
For my model I tried running MLE using Poisson,Quasi and Gaussian distributions(similar to normal).

Poisson

```
Call:
glm(formula = GWL ~ SDP + SDP_2 + SDP_3 + gini + electionMargin +
    capital_exp + agg_exp, family = poisson(link = "identity"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2770  -1.7955  -0.8334   0.2423  23.5523

Coefficients:
                Estimate Std. Error  z value Pr(>|z|)
(Intercept)     2.009e+00  6.760e-02   29.716  < 2e-16 ***
SDP             3.310e-05  2.521e-07  131.269  < 2e-16 ***
SDP_2          -2.482e-11  3.097e-13  -80.120  < 2e-16 ***
SDP_3           4.857e-18  1.112e-19   43.681  < 2e-16 ***
gini            5.028e+00  1.736e-01   28.962  < 2e-16 ***
electionMargin  3.233e-02  1.461e-03   22.132  < 2e-16 ***
capital_exp     1.792e-05  3.504e-06    5.114 3.15e-07 ***
agg_exp        -1.069e-04  4.403e-07 -242.770  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 790768  on 109719  degrees of freedom
Residual deviance: 723086  on 109712  degrees of freedom
  (2678 observations deleted due to missingness)
AIC: Inf

Number of Fisher Scoring iterations: 25
```

Gaussian

```
Call:
glm(formula = GWL ~ SDP + SDP_2 + SDP_3 + gini + electionMargin +
    capital_exp + agg_exp, family = gaussian(link = "identity"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.407   -5.263   -2.489    0.875  148.905

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.844e+00  2.772e-01  10.260  < 2e-16 ***
SDP             2.992e-05  1.006e-06  29.753  < 2e-16 ***
SDP_2          -2.214e-11  1.210e-12 -18.304  < 2e-16 ***
SDP_3           4.225e-18  4.328e-19   9.764  < 2e-16 ***
gini            5.895e+00  6.300e-01   9.358  < 2e-16 ***
electionMargin  3.726e-02  5.344e-03   6.973 3.12e-12 ***
capital_exp     2.436e-04  1.692e-05  14.401  < 2e-16 ***
agg_exp        -1.573e-04  3.696e-06 -42.568  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 128.2609)

    Null deviance: 14703892  on 109719  degrees of freedom
Residual deviance: 14071757  on 109712  degrees of freedom
  (2678 observations deleted due to missingness)
AIC: 843970

Number of Fisher Scoring iterations: 2
```

Quasi

```
Call:
glm(formula = GWL ~ SDP + SDP_2 + SDP_3 + gini + electionMargin +
    capital_exp + agg_exp, family = quasi(link = "identity"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.407   -5.263   -2.489    0.875  148.905

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.844e+00  2.772e-01  10.260  < 2e-16 ***
SDP             2.992e-05  1.006e-06  29.753  < 2e-16 ***
SDP_2          -2.214e-11  1.210e-12 -18.304  < 2e-16 ***
SDP_3           4.225e-18  4.328e-19   9.764  < 2e-16 ***
gini            5.895e+00  6.300e-01   9.358  < 2e-16 ***
electionMargin  3.726e-02  5.344e-03   6.973 3.12e-12 ***
capital_exp     2.436e-04  1.692e-05  14.401  < 2e-16 ***
agg_exp        -1.573e-04  3.696e-06 -42.568  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 128.2609)

    Null deviance: 14703892  on 109719  degrees of freedom
Residual deviance: 14071757  on 109712  degrees of freedom
  (2678 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 2
```

Heteroskedasticity or Homoscedasticity?

```
        studentized Breusch-Pagan test

data:  model
BP = 83.192, df = 9, p-value = 3.743e-14
```

The linear regression models assume homoscedasticity i.e. same variances. If variances are different then this condition is violated and we have heteroskedasticity.

Instead of linear regression, one might use weighted regression (i.e. it assigns weight to each sample point depending on the variance of the fitted value). So points with high variances get lower weights and vice versa.
However, the best strategy would be to try Feasible Generalised Least Squares(FGLS) as discussed in class. FGLS accounts for heteroskedasticity by using weights that are inversely proportional to the variance of the errors.

Since the p-value is extremely small, the result is significant and hence the different state groups show heteroscedasticity. Also the high value of BP suggests that the difference in the variances is significant.

New work in our project

We added these regressors to enhance our new model:

Capital Expenditure (cap_exp): This is a new regressor added to the model which represents the amount of money a government spends on long-term assets such as buildings, machinery, and equipment. It is expected to have a positive effect on groundwater level as it can improve infrastructure and water management systems.

Aggregate Expenditure (agg_exp): This regressor represents the total amount of government spending on all goods and services in the economy. It is expected to have a negative effect on groundwater level as it can increase water demand and lead to over-extraction of groundwater resources.

Margin of Election Victory (electionMargin): This regressor represents the difference in percentage points between the winning candidate and the closest opponent in a state election. It is expected to have a positive effect on groundwater level as a larger margin of victory can indicate a more stable government, which can lead to better implementation of water management policies.

Conclusion

- We run the regression on the enhanced model and obtain the results given on the left.
- The value of all coefficients is extremely small except the Intercept and α_4 and thus practically insignificant.However all independent variables except for SDP_3 are statistically significant since there p-values are extremely small.
- The value of R squared at 0.04299 is extremely small which means that the model manages to explain only about 4.3% of variation in GWL.
- The value of F-stat however is 704.1 which is very large along with its p-value indicates that the model is statistically significant.
- In districts with higher margins of victory the groundwater level increases more. However since the estimate is small(0.037) it is clear that the margin of victory does not have a significant impact on groundwater.
- Capital Expenditure also has a positive relationship with groundwater level though the estimate is once again small. This could be because groundwater levels increase with capital investments such as dams,canals and irrigation systems. This relationship is as per expectations.
- Aggregate Expenditure is negatively related to groundwater level though the estimate is small. This could be because the aggregate expenditure is an indicator of increased wealth and aggregate output which leads further exploitation of the environment. This is as per our expectations.
- **Kuznets Curve**: Given the fact that the estimates are extremely small it is tough to say whether our data follows the EKC. It is easiest to say that the impact of SDP_2 and SDP_3 on GWL is practically insignificant. While SDP coefficient is positive which indicates an increase in GWL with SDP.

```
Call:
lm(formula = GWL ~ SDP + SDP_2 + SDP_3 + gini + electionMargin +
    capital_exp + agg_exp)

Residuals:
    Min      1Q  Median      3Q     Max
-14.407  -5.263  -2.489   0.875 148.905

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.844e+00  2.772e-01  10.260  < 2e-16 ***
SDP             2.992e-05  1.006e-06  29.753  < 2e-16 ***
SDP_2          -2.214e-11  1.210e-12 -18.304  < 2e-16 ***
SDP_3           4.225e-18  4.328e-19   9.764  < 2e-16 ***
gini            5.895e+00  6.300e-01   9.358  < 2e-16 ***
electionMargin  3.726e-02  5.344e-03   6.973 3.12e-12 ***
capital_exp     2.436e-04  1.692e-05  14.401  < 2e-16 ***
agg_exp        -1.573e-04  3.696e-06 -42.568  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.33 on 109712 degrees of freedom
  (2678 observations deleted due to missingness)
Multiple R-squared:  0.04299,   Adjusted R-squared:  0.04293
F-statistic: 704.1 on 7 and 109712 DF,  p-value: < 2.2e-16
```

- As observed before, the p-value is small implying the difference between the mean ground water levels among different regions is significant.
- Also from the Breusch-Pagan test, we get a really small p-value indicating heteroscedasticity among the different state regions.