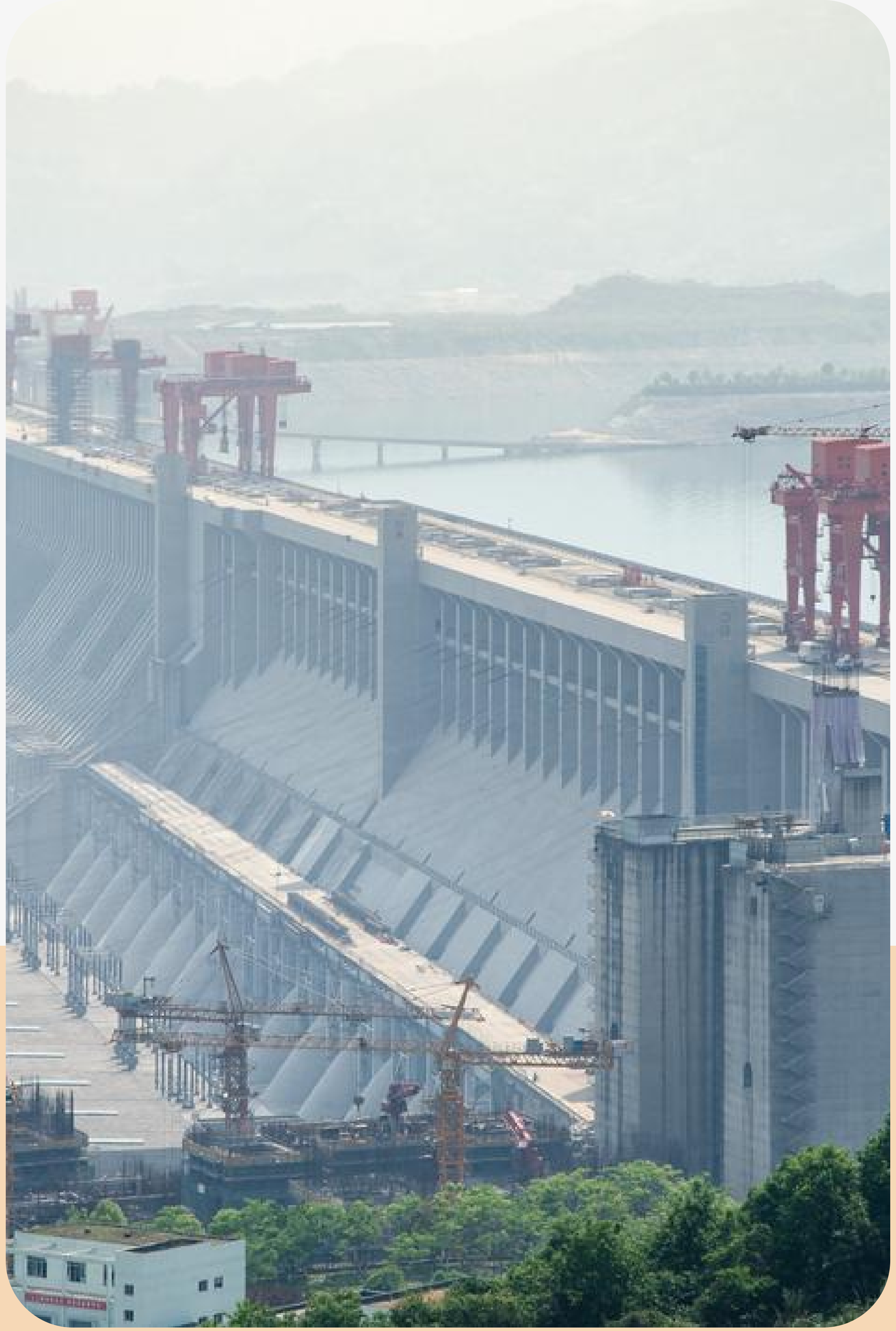


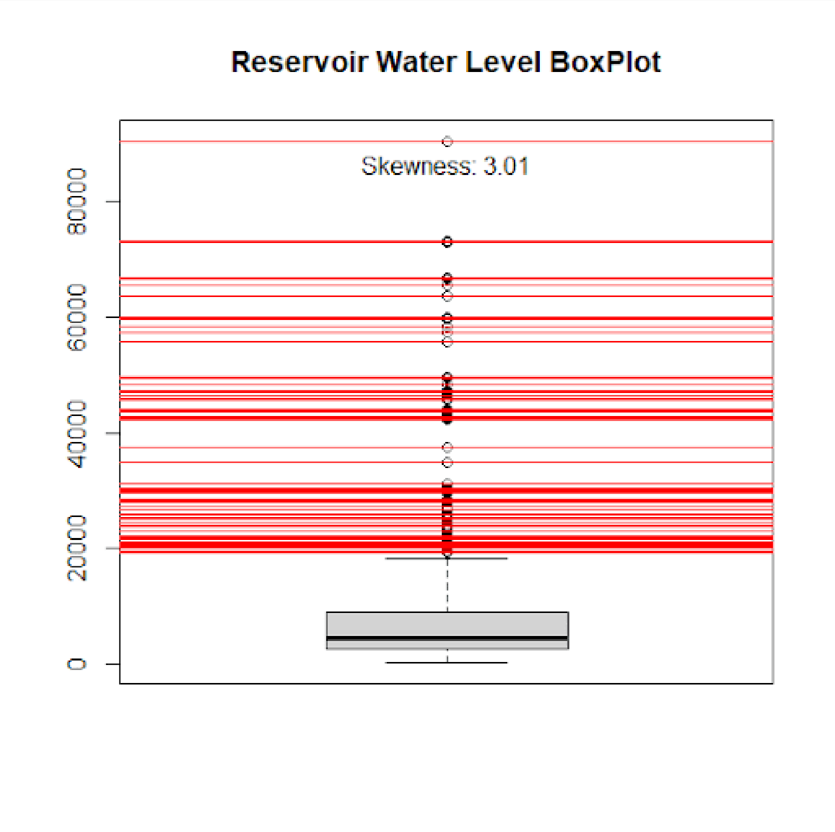
Econometrics - 1

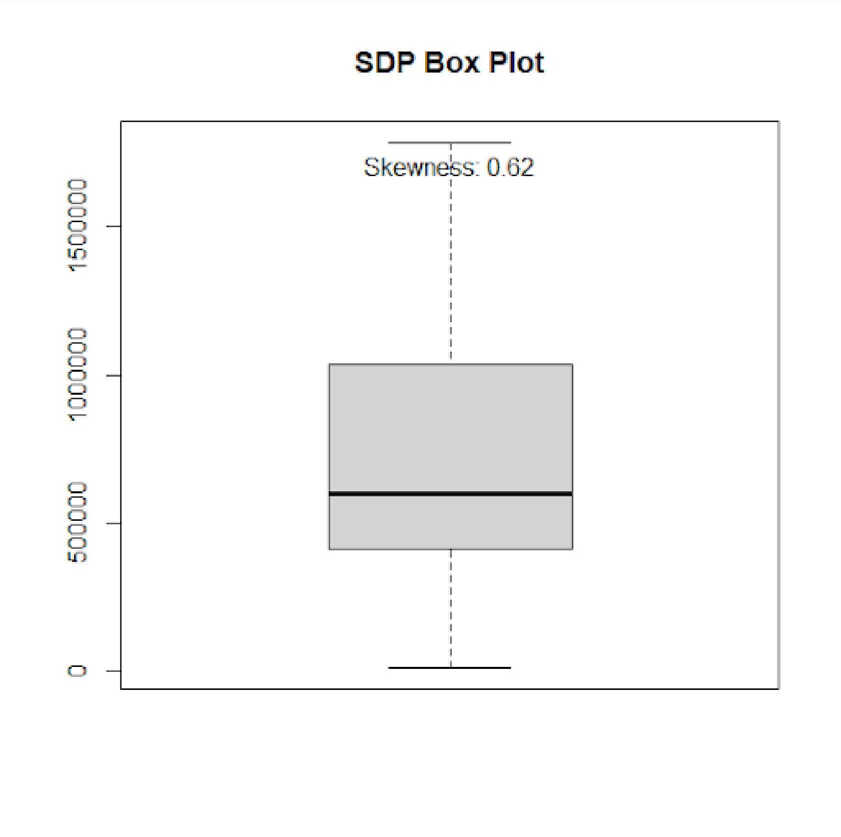
Analysis of Relationship Between Reservoir Water Level and Power Inequalities Across States in India

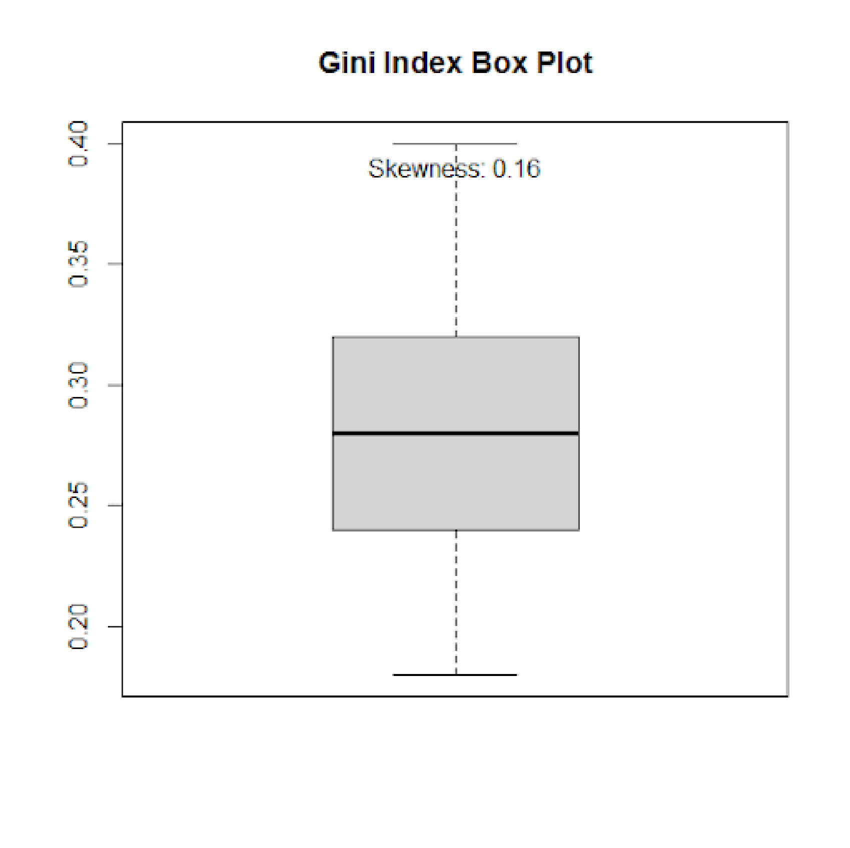
Sarthak Daksh(2020403)
Aditi Saxena (2021371)
Twisha Kacker (2021432)
Manish Sehrawat- 2021400
Sparsh Sehgal - 2021425

Outliers Analysis

We had chosen data set from the year 2011-2019. Indicators initially used were district and years-wise water reservoir levels, Gini index, and SDP. After combining our dataset, we did pre-processing of the data, which involved creating various plots and finding if there were any outliers in the data.







Reservoir Water Level has a significant positive skewness with outliers on the right side of the distribution.

SDP has a modest positive skewness with outliers on the right side, but not as obvious as Reservoir Water Level.

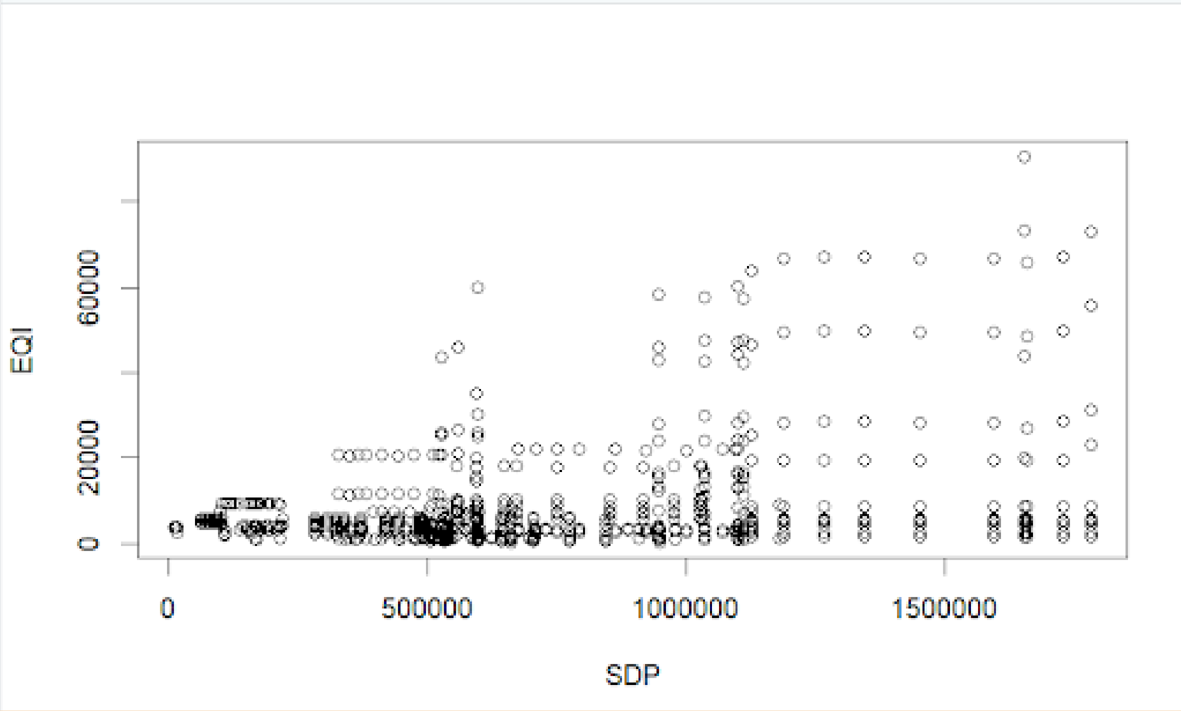
Gini has an approximately symmetrical distribution with a slight tendency towards a longer tail on the right side.

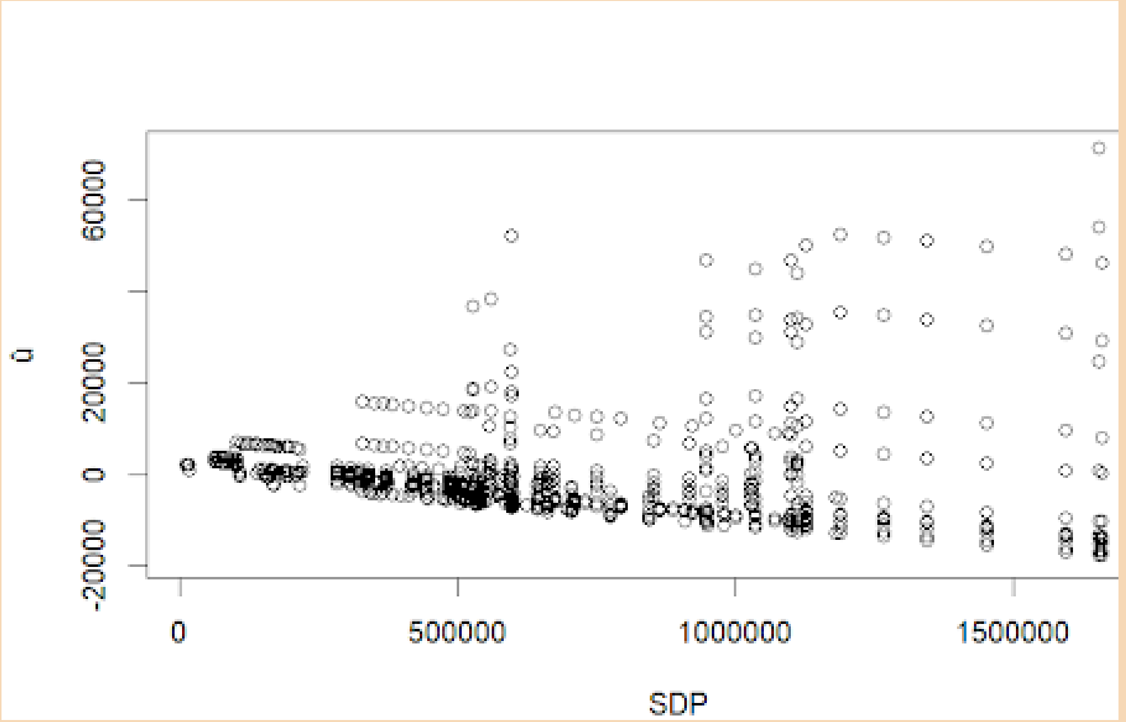
Model Summary

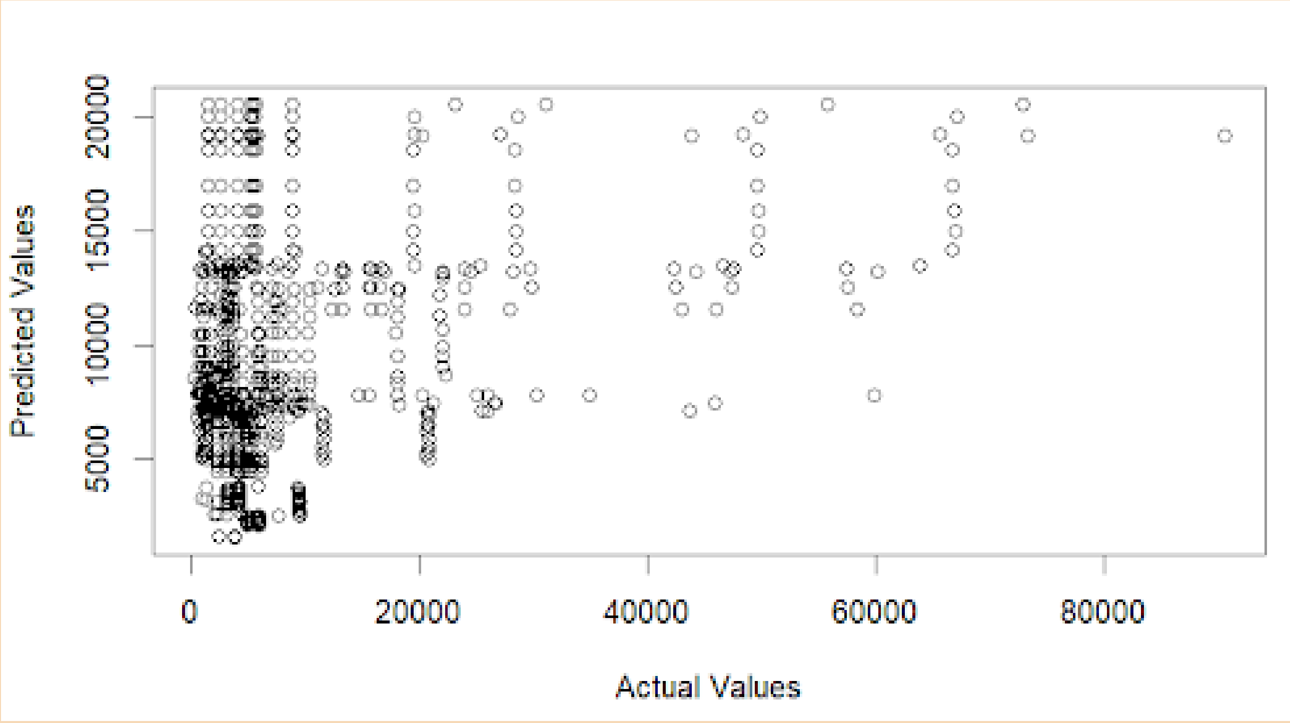
COEFFICIENTS	Estimate Std.	Error t	value	Pr(>|t|)
Intercept	1.432e+03	8.089e+02	1.771	0.077 .
SDP	1.072e-02	9.558e-04	11.212	<2e-16 ***

The linear regression model for **Reservoir Water Level i,t = β0 + β1SDPi,t + ui,t**
The coefficients show that the value of EQI at SDP = 0 is 1.432e+03, and on average, the Reservoir Water Level increases by 1.072e-02 each year per unit increase in SDP.

Relationship Analysis







Plot indicates weak relationship and wide scatter around the line of best fit, implying inaccurate predictions. A high SDP implies unreliable and inconsistent model predictions across a wide range of values.

Plot displays the relationship between residuals and SDP predictor. Residuals are widely dispersed, and SDP range is high, indicating the model may not capture all vital relationships in the data.

Plot displays predicted vs. observed outcome variable values. Scatter is high, and slope is <1, indicating the model's high accuracy.

MLRM Summary

This passage describes the results of a regression analysis with 4 predictor variables. The analysis found that two of the predictor variables (SDP^2 and SDP^3) have a curvilinear relationship with Reservoir Water Level, while Gini's coefficient has a positive but small impact on Reservoir Water Level. The model indicates significant but non-linear relationships between the predictors and the response variable. However, there is still unexplained variation in the data, suggesting that other factors may also influence Reservoir Water Level. When the number of regressors is increased, the model shows that the water level increases with an increase in SDP, indicating a heteroscedastic and nonlinear relationship. Overall, the model suggests that the water level increases by a certain amount every year with respect to the SDP.

COEFFICIENTS	Estimate Std.	Error t	value	Pr(>|t|)
Intercept	1140	2850	0.4	0.68932
SDP	-26731	14578	-1.834	0.06711
SDP2	98479	34714	2.837	0.00468
SDP3	-57246	23171	-2.471	0.01371
GINI	18366	7732	2.375	0.01778

$$\text{Reservoir Water Level } i,t = \alpha 0 + \alpha 1 SDPi,t + \alpha 2 SDP^2 i.t + \alpha 3 SDP^3 i.t + \alpha 4 GINIi + \gamma i,t$$

It indicates that the model is significant and explains 46.15% of the variance in the response variable. The model has a residual standard error of 3535 and includes 4 predictors. The adjusted R-squared is 0.4485, which indicates that the model is not overfitting the data.

Role of Standard Error

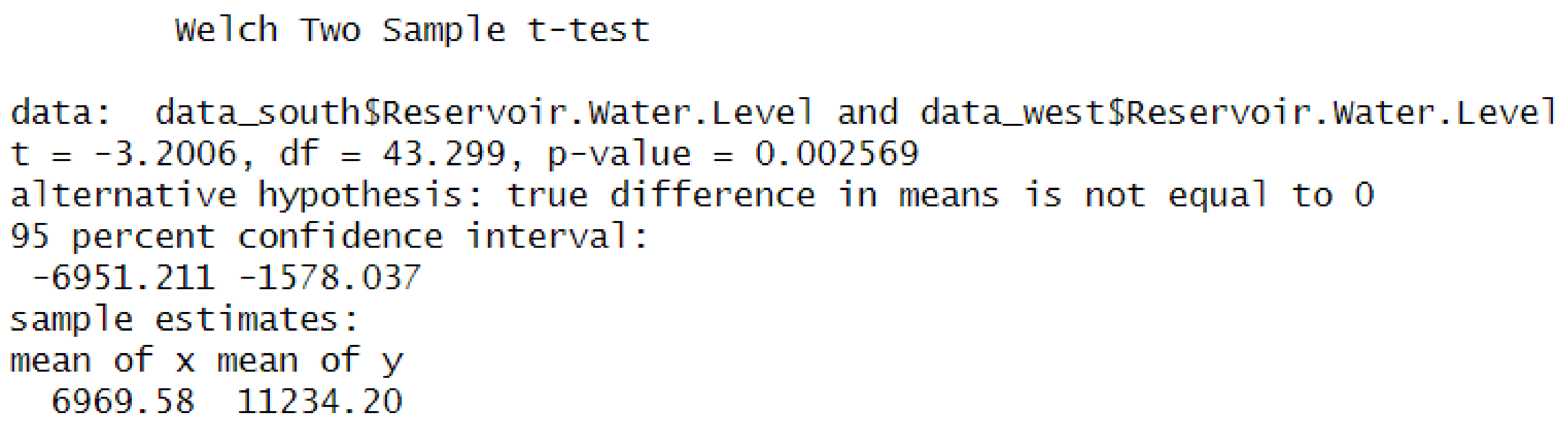
	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	-1.23E+04	7.17E+03	-1.716	0.0881
SDP	6.76E-03	1.12E-03	6.049	1.03E-08
GiniState	1.24E+04	5.69E+03	2.186	0.0303
Literacy	1.60E+02	6.58E+01	2.429	0.0163
Rainfall	-7.37E-01	4.06E-01	-1.818	0.071
HydroPower	2.94E-02	3.81E-02	0.77	0.4422
Per Capita Income	1.68E-02	7.70E-03	2.184	0.0305
Mortality Rate	3.86E+00	4.59E+01	0.084	0.933
Unemployment	-1.08E+01	1.32E+01	-0.815	0.416
EEP	-2.90E+00	1.12E+00	-2.593	0.0104

In a regression model, the standard errors of the coefficients are important in determining statistical significance. For instance, the standard error of the coefficient for SDP is 0.001117 and for GiniState is. Standard errors are used to calculate t-values.

The residual standard error, at the bottom of the output, measures the distance between observed and predicted values with a value of 3638. Standard errors are significant in regression analysis to evaluate coefficient reliability, independent variable significance, data variability, and prediction accuracy.

T- Test and Chow Test

Chow test and T-test were conducted to find any structural break in the data for mean environmental Quality. After conducting the Chow test, we got the values for state groups south and west for a structural break. To confirm this, we did T-test on all the states and acquired similar results from T-test.

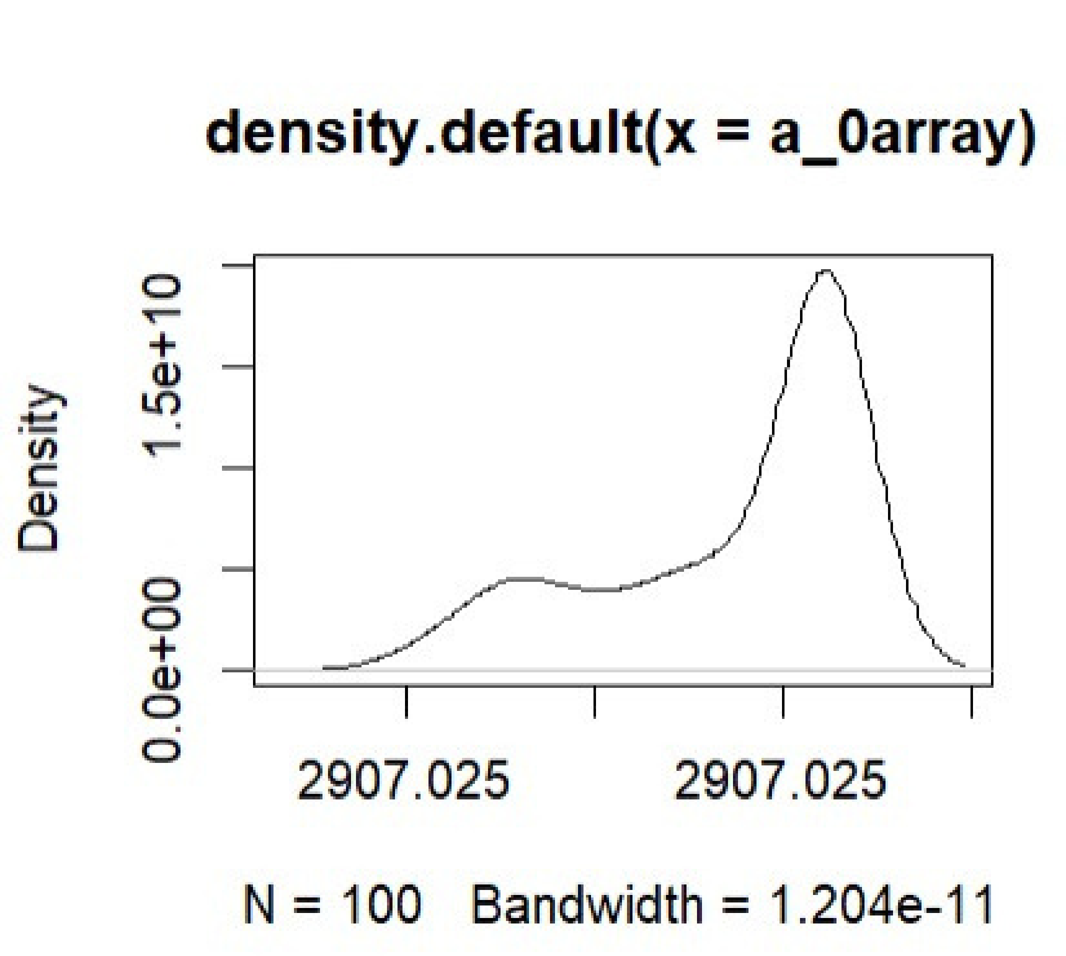


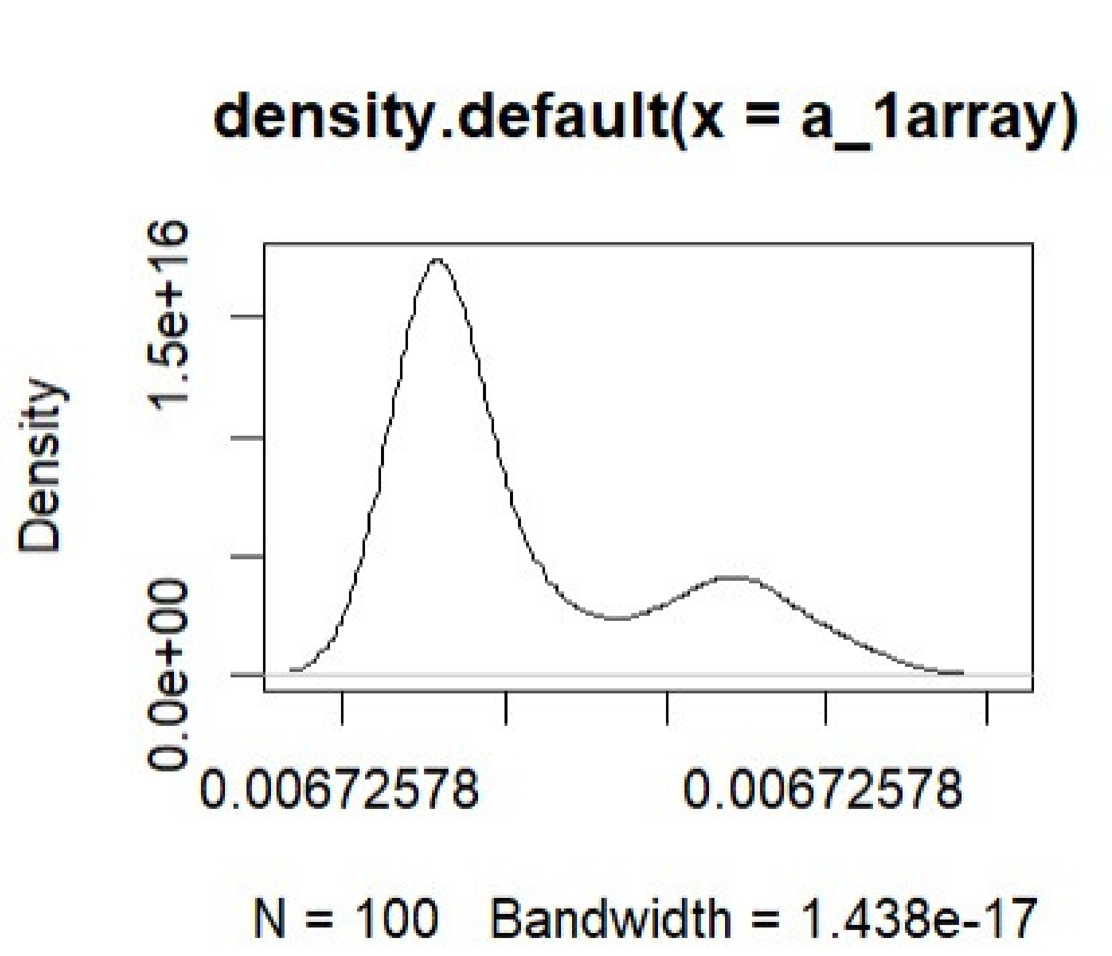
```
        Welch Two Sample t-test

data:  data_south$Reservoir.Water.Level and data_west$Reservoir.Water.Level
t = -3.2006, df = 43.299, p-value = 0.002569
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6951.211 -1578.037
sample estimates:
mean of x mean of y 
  6969.58  11234.20 
```

The test was conducted on two sets of data, one from the South and the other from the West, and it found a statistically significant difference in the means of the Reservoir Water Level between the two regions. The calculated t-value was -3.2006 with a p-value of 0.002569, which indicates that the difference in the means is unlikely to be due to chance.

The sample means for Reservoir Water Level were found to be 6969.58 for the South and 11234.20 for the West. The 95 percent confidence interval for the difference in means was -6951.211 to -1578.037, which suggests that the true difference in means between the two regions is likely to lie within this range.

Monte Carlo Simulation





Estimates from the complete dataset:-

a0 - 2942.003
a1 - 0.006865

Estimates from the 80% dataset:-

a0 - 2907.25
a1 - 0.006725

Regression Model : v43(i,t) = a0 + a1SDP(i,t)

Monte Carlo simulations were performed on 80% of the dataset with water reservoir level as the dependent variable and SDP(state domestic product) as the independent variable.

The estimate a0 is within the error of 1.18%
The estimate a1 is within the error of 2.04%

Maximum Likelihood Estimation

Maximum likelihood estimation can improve the estimation of the coefficients of a regression model, including Reservoir.Water.Level ~ SDP + GiniState + Literacy + Rainfall + HydroPower + Per Capita Income + Mortality Rate + Unemployment + EEP, by maximizing the likelihood of the observed data. By using this method, the model can better capture the relationship between the dependent variable and the independent variables, resulting in more accurate coefficient estimates and predictions of the reservoir water level. MLE provides a method for estimating the model parameters accurately and handling missing data. By choosing an appropriate likelihood function, MLE can also provide insight into the distribution of the response variable and the assumptions made about the errors in the model. Additionally, MLE can be used to compare different models and select the one that provides the best fit for the data and maximum likelihood estimation provides a framework for hypothesis testing and model selection. Overall, MLE can improve the accuracy and precision of my regression analysis by accounting for the distribution of the response variable and the error structure of the model.

Analysis of Variance

	RSS	Res.Df	F Statistic		Sum of Sq F
Model 1:	DSouth	157	NA	1578355465	
Model 2:	DNorth	157	1.00	1615980585	-37625120
Model 3:	DNortheast	157	0.10	1629102280	-13121694
Model 4:	DEast	157	4.98	1588804044	40298235
Model 5:	DWest	157	4.95	1566808495	21995549
Model 6:	DCentre	157	2.00	1604764697	-37956202

Looking at the ANOVA table, we can see that none of the F-statistics for the six models are significant at the 0.05 level of significance, as all of the p-values are greater than 0.05. This indicates that we do not have enough evidence to reject the null hypothesis, and we cannot conclude that there are significant differences in environmental quality across the state-groups based on the predictors included in the models.

If the variance differs significantly across state groups, then the assumption of homoscedasticity, also known as the assumption of constant variance, is violated in ordinary least squares (OLS) regression.
In our data we can see that there is a consistency in the residual data, hence the data is homoscedastic.

To account for a variance structure where the variance of the residuals differs significantly across state groups, we can use a linear mixed-effects model (LMM), also known as a mixed-effects regression model or hierarchical linear model. We can add a random intercept for each state group, which allows us to model the differences in the residual variances across groups.

Conclusion

- Initially, we found factors SDP, Gini, Literacy, Rainfall, HydroPower, PerCapita Income, Mortality Rate, Unemployment, and EEP as affecting the water reservoir level. We chose an effective political party as our power inequality for the topic.

- While working with the model, we realized that Mortality Rate and rainfall don't have any relation with the Reservoir Water Level. Hence, to improve our model, we eventually reduced these two regressors to include only 4 regressors in the model.

- From our project, we have found that the model indicates significant but non-linear relationships between the predictors and the response variable. Through this, we found that SDP, Hydro power, Per Capita Income, unemployment and EEP are the factors that influence water reservoirs the most.

- We performed Chow Test and T-test to help us know the collinearity of the data and the structural breaks in the data. This helped us Know about the most significant values in our data with respect to the state groups. We found the regions south and west most affecting the water reservoir levels.

- We saw from our residual data that there is a pattern in a residual graph. We can see from the graph that we are omitting relevant datasets, which rejects our homoscedastic assumption taken in the first assignment.

Thank You