IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

FIXED EFFECT PANEL REGRESSION FOR THE QUANTIFICATION OF WATER QUALITY PARAMETERS AND ITS SIGNIFICANCE OVER LEAST SQUARE REGRESSION

[1]Arshita Srivastava, [2]Dr Rajeev Pandey, [3]Dr S.M.H.Zaidi

[1]Research Scholar, [2]Professor and Head, [3]Associate Professor
[1]Department of Statistics,
[1]University of Lucknow, Lucknow, India

Abstract: The present paper provides the utility of Fixed Effect Panel Regression Model to analyze the water quality parameters of River *Ganga* of India and aims to compare its effectiveness to Least Square Regression model. The statistical significance of both the models have been studied by suitable yearly observations for three different places as observed by Central Pollution Control board (CPCB), India. The results of the present paper show the evidence of better fit of goodness of the Fixed Effect Panel Regression Model.

Keywords- CPCB, Ganga, Water Quality Index, Panel Data, Ordinary Least Square Regression, Panel Regression

I. INTRODUCTION

Water resources are very rich in India but the continuous rapid increase in population is also leading to an increase in demand of irrigation, industrial and individual consumption thereby leading to a depletion of available water resources. **Vorosmarty** et al.(2000)[1], **Wagener et.al** (2010)[2] and **Vogel**(2011)[3] suggested that there is a growing need to improve our understanding of and ability to predict the effects of human activities on the hydrological cycle. Water quality parameters of water resources of various rivers in India are monitored by the monitoring stations established across rivers by the Central Pollution Control Board(CPCB)[4] and these stations are monitoring the real time water quality of water of rivers. Presently, the monitoring network of CPCB comprises of 870 stations in 26 States and 5 Union territories of the country.

Several studies have been made to analyze the water quality level of ground water and surface water located at different centers across the Globe over the time and regression techniques have been the key for predicting the water quality status of water for upcoming years. The present paper is devoted for the comparative study of two regressions viz. 1. The ordinary least square and 2. Panel regression models describing the methodology of both the regressions performed on panel data of water quality parameters of river Ganga. In the present study, the data of water quality has been observed from the web portal of the CPCB which are cross sectional observations of different water quality parameters of river Ganga across years commencing from 2006 to 2019 at three different monitoring Stations situated at (a) **Assi ghat, Varanasi (b) Ranighat, Kanpur and (C) Sangam, Prayagraj** of the **Uttar Pradesh** state of **India**.

Regression analysis establishes a relationship between a dependent or outcome variable with one or more independent or predictor variable. Parameters that are usually sampled or monitored for water quality are temperature, pH level, conductivity, turbidity, Total dissolved solid, Total suspended solid, Dissolved oxygen, Bio-chemical oxygen demand etc. which are directly observed whereas some of the key parameters such as Fluorine, Chlorine, Magnesium, Sulphur, Nitrate, Nitrite etc. are analyzed in laboratories of the CPCB. In the present study, the parameters responsible for water quality of river Ganga **temperature, pH level, dissolved oxygen(DO), bio- chemical oxygen demand(BOD), nitrate, fecal coliform and total coliform** are considered as the independent variable and **Water Quality Index** as the dependent variable . **Horton, (1965)[5]** suggested that the numerous water quality data could be combined into an overall index **Water quality index (WQI)** i.e. WQI comprises of a number describing the overall water quality at certain location and time based on water quality parameters and serves a useful indicator of water quality. The WQI ranges from 1 to 100, the value between 90-100 describing the excellent water quality, 70-89 stating the good water quality, 50-69 stating the medium water quality, 25-49 stating the bad water quality and 0-24 stating the worst water quality. The Weighted Arithmetic Mean Water Quality Index method for WQI is used for the analysis.

The formula for the Water Quality Index (WQI) was proposed by **Brown et. al (1972)[6]** as under noted:

$$WQI = \frac{\sum_{i=1}^{n} q_i w_i}{\sum_{i=1}^{n} w_i} \qquad (1.1.1)$$

where,
q_i=quality rating (sub index) of i^{th} water quality parameter
w_i= unit weight of i^{th} water quality parameter; $\sum_{i=1}^{n} w_i = 1$
q_i, relates the value of the parameter in polluted water to the standard permissible value is obtained as follows:

$$q_i = 100 * \left(\frac{v_i - v_{io}}{s_i - v_{io}}\right) \qquad (1.1.2)$$

Where,
v_i= estimated value of the i^{th} parameter
v_{io} = ideal value of the i^{th} parameter
s_i= standard permissible value of the i^{th} parameter
(In most cases, v_{io}=0 except for pH and Dissolved Oxygen)
The unit weight (w_i), is inversely proportional to the values of the recommended standards is obtained by:

$$w_i = \frac{k}{s_i} \qquad (1.1.3)$$

Where k=$\frac{1}{\sum_{i=1}^{n} \frac{1}{s_i}}$

The regression techniques, i.e. Ordinary least square regression and Fixed effect panel regression have been compared in the present study to check for the fitting of best regression model on panel data that can be used for future prediction of water quality index with 95% of Confidence Interval. Ordinary Least Square regression is a method to find linear regression in a set of data whereas the Panel regression is a modeling method adapted for panel data i.e longitudinal data or cross-sectional data as suggested by **Erica(2019)[7]**. It is widely used where the behavior of statistical units (i.e. panel units) is followed across time. **Roberta et.al(2011)[8], Scott Steinschneider(2013)[9] and Bernhard Brugger[10]** suggested that panel data regression is a powerful way to control dependencies of unobserved, independent variables on a dependent variable, which can lead to biased estimators in traditional linear regression models. Fixed effect panel regression technique is used in analyzing the impact of variables that vary over time controlling all the time- invariant differences between the individuals, so the estimated coefficients of the fixed effect models cannot be biased because of the omitted time-invariant characteristics. Mixed effect model comprises of a statistical model containing both fixed and random effects.
Ordinary Least Square model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots\ldots\ldots\ldots + \beta_n x_{ni} + \varepsilon_i \qquad (1.2.1)$$

i.e $$Y_i = \beta_0 + \sum_{i=1}^{n} \beta_i x_{ni} + \varepsilon_i$$

As suggested by **J.A. Kupolusi et.al (2015)[11]** Panel Regression model is undernoted as:

$$Y_{it} = \alpha_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \cdots + \beta_n x_{nit} + \varepsilon_{it} \qquad (1.2.2)$$

Where
Y_{it} = The dependent variable, i= entity and t= time
$X_{n,it}$ = The independent variables
β_n = The coefficients of independent variable
α_i = Individual effect
ε_{it} = The error term.
Mixed effect model:

$$Y = X\beta + Zu + \varepsilon \qquad (1.2.3)$$

Y is a known vector of observations
β is an unknown vector of fixed effects
u is an unknown vector of random effects, with mean E(u)=0and variance- covariance matrix V(u)=G
ε is an unknown vector of random errors, with mean E(ε)=0 and variance V(ε)= R
X and Z are known design matrices relating the observations Y to β and u respectively.

II. METHEDOLOGY

Table 2.1: Data

Place	Period	Temprature	Dissolved Oxygen	pH	Bio-Chemical Oxygen Demand (mg/L)	Nitrate (mg/ L)	Faecal Coliform (MPN/ 100 mL)	Total Coliform (MPN/ 100 mL)	Water quality Index
Assi Ghat Varanasi	2006	30.0	9	8.2	3.8	0.6	13000	17000	38
	2007	30.0	9	7.4	11.2	3.8	13000	13000	33
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	2019	31.5	10	8.4	3.3	0.24	1700	3400	0
Sangam, Prayagraj	2006	33	8.4	8.6	5.8	2.5	13000	50000	0
	2007	31	9.3	8.7	4.7	2.8	9000	14000	0
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	2019	32.8	11.5	8.4	3.4	1.8	13000	27000	42
Ranighat, Kanpur	2006	29.5	8.8	8.6	4.4	2.0	9000	21000	40
	2007	30.0	11.0	8.6	6.4	2.9	7500	15000	38
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	2019	32.0	10.3	8.7	4.0	0	3400	5800	41

In the present study, the secondary data set extracted from Central Pollution Control Board (CPCB) portal is used, for different monitoring stations of river serve as cross sectional units and the years of monitoring as the time period.

The **regress** command of Stata handles the OLS model. The Ordinary Least Square Regression model for WQI(dependent variable) is considered as:

$$WQI_i = \beta_0 + \beta_1 temperature_i + \beta_2 pH_i + \beta_3 Dissolved\ Oxygen_i + \beta_4 Bio - chemical\ Oxygen\ Demand_i + \beta_5 Fecal\ Coliform_i + \beta_6 Total\ Coliform_i + \beta_7 Nitrate_i + \varepsilon_i \qquad (1.2.4)$$

Here, we represent Water Quality Index(WQI) as Yi, temperature as x_1, pH as x_2, Dissolved Oxygen(DO) as x_3, Bio-chemical oxygen demand as x_4, Fecal coliform as x_5 and Total Coliform as x_6.

The **xtreg** command of Stata has been used to perform fixed effect panel regression. The Fixed effect panel regression model for present study is:

$$WQI_i = \alpha_i + \beta_1 temperature_{it} + \beta_2 pH_{it} + \beta_3 Dissolved\ Oxygen_{it} + \beta_4 Bio - chemical\ Oxygen\ Demand_{it} + \beta_5 Fecal\ Coliform_{it} + \beta_6 Total\ Coliform_{it} + \beta_7 Nitrate_{it} + \varepsilon_{it} \qquad (1.2.5)$$

III. RESULTS AND DISCUSSION

3.1 Results of the Ordinary Least Square Regression

Table 3.1.1: Descriptive Statistics

Source	SS	df	MS	
				Number of obs = 42
				F(7, 34) = 1.21
Model	1827.70694	7	261.100991	Prob > F = 0.3245
Residual	7341.93592	34	215.939292	R-squared = 0.1993
				Adj R-squared = 0.0345
Total	9169.64286	41	223.649826	Root MSE = 14.695

The R value represents the correlation between the independent and dependent variables.

In Table 3.1, R-square shows the total variation in the dependent variable that could be explained by the independent variables. In this case, 19% of variation in Water Quality Index could be explained by the independent variables.

P-value/ Sig value: Generally, 95% confidence interval or 5% level of the significance level is chosen for the study. The p-value for the OLS model is 0.324, thereby showing that the result is insignificant.

F-ratio represents an improvement in the prediction of the variable by fitting the model after considering the inaccuracy present in the model. A value greater than 1 for F-ratio yield efficient model. Here, the F-ratio is 1.21.

Table 3.1.2: Coefficient Table

WaterqualityIndex	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
Temprature	-.5707505	1.271824	-0.45	0.656	-3.155408	2.013907
DissolvedOxygen	6.145891	2.680375	2.29	0.028	.6987143	11.59307
pH	-15.8997	7.792154	-2.04	0.049	-31.73526	-.0641383
BioChemicalOxygenDemandmgL	-.9722947	1.721446	-0.56	0.576	-4.470693	2.526103
NitratemgL	-1.07732	2.383363	-0.45	0.654	-5.920898	3.766257
FaecalColiformMPN100mL	.0004565	.0003478	1.31	0.198	-.0002503	.0011633
TotalColiformMPN100mL	-.0000687	.0000829	-0.83	0.413	-.0002372	.0000998
_cons	127.9577	59.30789	2.16	0.038	7.429521	248.4858

Table 3.1.2 shows that none of the value is below the tolerable level of significance for the study i.e. below 0.05 for 95% confidence interval in this study.

3.2 Results of Fixed Effect Panel Regression

Table 3.2.1: Descriptive Statistics

Fixed-effects (within) regression	Number of obs = 42
Group variable: Placeid	Number of groups = 3
R-sq:	Obs per group:
within = 0.4572	min = 14
between = 0.7626	avg = 14.0
overall = 0.0569	max = 14
	F(7,32) = 3.85
corr(u_i, Xb) = -0.7056	Prob > F = 0.0038

In the fixed effect panel regression, as evident from the R-square value, **45%** of variation in Water quality Index is being explained by the independent variables.

Table 3.2.1shows that the p-value given by panel regression (**0.0038**) is less than the tolerable level of significance for the study and hence the result is significant.

F-ratio for Fixed effect panel regression is **3.85**, which is much higher than Ordinary Least Square regression.

Table 3.2.2: Coefficient table

WaterqualityIndex	Coef. Std. Err.	t P>t	[95% Conf. Interval]
Temprature	-2.284105 1.093394	-2.09 0.045	-4.511276 -.0569345
DissolvedOxygen	2.661008 2.601017	1.02 0.314	-2.63709 7.959107
pH	-29.4208 7.013059	-4.20 0.000	-43.70594 -15.13567
BioChemicalOxygenDemandmgL	-.905534 1.438973	-0.63 0.534	-3.836626 2.025558
NitratemgL	-3.294219 2.459796	-1.34 0.190	-8.304658 1.716221
FaecalColiformMPN100mL	.0001761 .0003222	0.55 0.588	-.0004802 .0008324
TotalColiformMPN100mL	-2.51e-06 .0000738	-0.03 0.973	-.0001528 .0001477
_cons	332.0111 65.18162	5.09 0.000	199.2405 464.7817
sigma_u	17.31603		
sigma_e	11.768248		
rho	.68405177 (fraction	of variance due	to u_i)

Table 3.2.2 shows that value for atleast one variable(fecal coliform) is below the tolerable level of significance for the study i.e. below 0.05 for 95% confidence interval in this study.

Discussion

From the Table 3.1.1 it is evident that the OLS method for the panel data gives insignificant result, thereby not providing a goodness of fit model. However, the Table 3.2.1, provides the evidence of Panel regression giving a significant result, thereby providing a better goodness of fit model for panel data.

Therefore, it is recommended to use panel regression instead of OLS for panel data.

REFERENCES

[1] Vorosmarty, C. J., P. Green, J. Salisbury, and R. B. Lammers (2000),Global water resources: Vulnerability from climate change and population growth, Science, 289(5477), 284–288.

[2] Wagener, T., M. Sivapalan, P. A. Troch, B. L. McGlynn, C. J. Harman, H.V. Gupta, P. Kumar, P. S. C. Rao, N. B. Basu, and J. S. Wilson (2010),The future of hydrology: An evolving science for a changing world,Water Resour. Res., 46,W05301, doi:10.1029/2009WR008906.

[3] Vogel, R. M. (2011), Hydromorphology, J. Water Resour. Plann. Manage.,137(2), 147–149.

[4] R.M.Bhardwaj (2005), Water quality monitoring in India- Achivements and constraints.

[5] Horton, R.K.,(1965), An index number system for rating water quality, J. Water Pollu. Cont. Fed., 37(3). 300-305.

[6] Brown RM, McClellan NI, Deininger RA, Tozer RG (1972) A water quality index—do we dare?–Water Sew Works 117 : 339—343.

[7] Erica(2019), Introduction to the fundamentals of Panel Data.

[8] Roberta Torre, Mikko Myrskyla (2011), Income inequality and population health: a panel data analysis.

[9] Scott Steinschneider, Yi-Chen E. Yang, Casey Brown (2013), Panel regression techniques for identifying impacts of anthropogeniclandscape change on hydrologic response.

[10] Bernhard Brugger : A guide to panel data regression: Theroy and Implementation.

[11] J.A. Kapulosi, R.A Adeleke, O. Akinyemi, B. Oguntuase,(2015), Comparative Analysis Of Least Square Regression And Fixed Effect Panel Data Regression Using Road Traffic Accident In Nigeria.