COMPREHENSIVE STUDY OF QUANTIFICATION OF WATER QUALITY PARAMETERS AND THEIR SIGNIFICANCES

Dr. Rajeev Pandey1), Arshita Srivastava2), and Dr. S.M.H.Zaidi3)

1)Arshita Srivastava , Research Student, Department of Statistics, University of Lucknow, Lucknow, India

2)Dr. Rajeev Pandey, Professor, Department of Statistics, University of Lucknow

3) Dr S.M.H. Zaidi, Associate Professor, Department of Statistics, Shia PG College, Lucknow

ABSTRACT

The present study is devoted to elaborately study the use of panel regression models in modelling the water quality parameters of river Subarnarekha. The study takes into consideration three modelling techniques of fixed effect panel regression models viz. within-group estimation, first difference estimation and least square dummy variable estimation techniques. The statistical significance of all the models have been compared using the best goodness of fit measures viz. Root mean square error and R-square by considering the panel data of water quality parameters extracted from the Water Year Book issued by Central Pollution Control Board.

INTRODUCTION

The vast evoking field of modern research has led to the development of multidimensional data capturing several aspects into one. Panel data setup is one such multidimensional data setup comprising of two-dimensional data into one. In context to modelling the panel data, panel regression models have gained popularity in the recent years in comparison with cross-sectional and time-series models because of their ability to map the individual and time effect into a single model framework. Panel regression models can be seen as the technique of estimating the complex relationship between the predictors and the outcome variable by adding the time dimension to cross-sectional units in the data. It increases the efficiency of parameter estimation as it has a higher degree of freedom in the model by taking into consideration the data across catchments and through time simultaneously.

S. Steinschneider et. al (2013) [1] defined panel regression as a statistical technique that pools the multidimensional data recorded across individuals and through time in order to identify the characteristics of response that are unique to each cross-sectional unit and common across time. Panel regression can be defined as the powerful statistical modelling technique to model the panel data which controls the dependencies of unobserved individual factors on the outcome variable. Panel regression provides information on the individual behavior both across individuals and time. A standard panel regression stacks up the values if independent and dependent variable over time and different cross-sections. Panel regression allows the researcher to control the characteristics of the cross-sectional units that do not vary over time and cannot be observed as variables in the data.

The panel data being a longitudinal data set comprises of both time-series and cross-sectional data, the model will contain both the dimensions, *i* representing the cross-sectional dimension of the data and *t* representing the time dimension of the data. As explained in the previous chapter, the general panel data regression model for k explanatory variable, with N cross-sectional units and T time period can be written as below:

$$Y_{it} = \alpha + X_{it}'\beta + \varepsilon_{it} \quad (1.1)$$

Where i = 1,2,3….N and t= 1, 2, 3…..T α is the intercept, *Xit* is the explanatory variable *i* at time *t* , β is a vector of regression coefficients, and ε_{it} is the idiosyncratic error term of individual *i* at time *t* .

The panel data model for *k* explanatory variables considering the decomposition of the error terms into individual and time effect, with the varying intercept can be elaborated as follows:

$$y_{it} = \alpha + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \beta_3 x_{it,3} \dots\dots\dots\dots\dots \beta_k x_{it,k} + u_i + v_{it} \quad (1.2)$$

Where α is the intercept, β is the regression coefficients, u_i is the time-invariant individual effect and v_{it} is the random effect which varies with time and across cross-sections.

Panel regression models can be structured into three types: Pooled Ordinary Least Square Regression, Fixed Effect Panel Regression Models and Random Effect Panel Regression Models. The POLS model does not take into consideration the time and cross-sectional component of the panel data and assumes that the behavior of the data is independent of time and space and does not vary across time periods and cross-sections. To overcome this drawback of POLS, fixed effect panel regression models are employed. As explained by **W. M. Mason (2001)** [2] fixed effect models are the modelling techniques which control the individual specific effects which do not vary over time. Fixed Effect Models considering the time and cross-sectional framework of panel data, provide means to control the individual-specific effect caused by the presence of correlation of extraneous factors with the explanatory variables, also named as the omitted variables. **Stock and Watson (2003)** [3] indicated that it can be assumed that if the unobserved variables do not change over time, then any variation in the dependent variable can be accounted for by the fixed characteristics taken into consideration by the fixed effect models. **Allison (2009)** [4] described fixed effect models as the regression models that make it possible to control the variables that have not or cannot be measured by using each individual as its own control. Fixed effect panel regression models can be seen as the method of controlling the variables that have or have not been observed as long as they stay constant within some larger category. The fixed effect models allow the unobserved variables to have association with the observed variables and allow the intercept of the regression model to vary freely across individuals or groups i.e., the intercept of the model is not constant across all the cross-sectional units. The fixed effect models are implied on the panel data to control the individual-specific time- invariant attributes that do not vary across time and assumes that the individual-specific effects are correlated with the independent variables. The fixed effect panel regression model assumes that the unobserved effects are fixed for each cross-sectional units and do not vary across cross-sections and time while the random effect model assumes that the unobserved heterogeneity in the data is not fixed for individual *i* rather, considers the individual-specific effect as random. The present study focuses only on the modelling techniques of fixed effect panel regression models on the panel data of water quality parameters of river Subarnarekha.

The fixed effect panel regression model can be stated as follows:

$$Y_{it} = X_{it}\beta + u_i + \varepsilon_{it} \quad (1.3)$$

The fixed models allow u_i to be correlated with the regressors X_{it}. i.e., Cov (X_{it}, u_i) ≠ 0, which implies that E [u_i | $x_{i1}, x_{i2}, \dots\dots, x_{iT}$] = E [u_i | X_i] = h (X_i) with constant Var [u_i | X_i]. But unlike X_{it}, u_i cannot be directly observed.

The next step of fixed effect modelling is the estimation of model parameters β and u_i for each of the N cross-sectional units of the panel data. To achieve this, the fixed effect model provides three estimation techniques, viz: Within-Group Estimation, First Difference Estimation And Least Square Dummy Variable (LSDV) Estimation Technique. The first two techniques of parameter estimation focus on eliminating the individual-

specific effect before estimation while the last technique LSDV incorporates the individual-specific effect using dummy variable.

The within-group estimator eliminates u_i by ***demeaning*** the variables using the within transformation. The term demeaning implies the process of subtracting each observation by its entity mean value i.e., the mean values of all the observed variables is calculated and then the same is subtracted from the observation of each individual and in this way the effect of individual-unique-time-invariant effect is wiped out from the outcome variable. Other than within-group estimation, an alternative way to eliminate the individual effect u_i from the panel regression model is by taking first difference of the fixed effect model with respect to time. The first difference method of estimation eliminates the unobserved effect by subtracting the observation of the previous time period from the observation of the current time period. The method is also known as difference-in-difference as it excludes the effect of change strictly over time and across units. In the third approach of fixed effect parameter estimation, the unobserved individual-effect is explicitly brought into the model. In this method, dummy variables are created for each subject and are included into the model and thus a matrix of dummy variables are included in the model.

The study aims to compare the efficacy of three fixed effect panel regression models on the panel data of water quality parameters. Water quality parameters of various rivers of India is being monitored by ***Central Pollution Control Board (CPCB)*** through real time monitoring technique by establishing various monitoring stations. For the present study, the data of water quality parameters of river *Subarnarekha* has been observed cross-sectionally over four monitoring stations (a) Jamshedpur (b) Ghatsila road bridge (c) Ghatsila and (d) Baridhinala across years commencing from 2005 to 2017, extracted from the ***Water Year Book, 2018*** of Central water commission. The study considers a total of twelve water quality parameters cross-sectionally observed over the monitoring stations across years, namely, p**H, Bio-chemical Oxygen Demand(BOD), Dissolved Oxygen(DO),Calcium(Ca), Chlorine(Cl), Fluorine(F), Iron(Fe), Potassium(K), Magnesium(Mg), Sodium(Na), Nitrate(NO_3) and Nitrite(NO_2)**, serving as the independent variables for the regression models. As suggested by ***Horton (1965)***[5], the water quality parameters can be combined into an overall index known as **Water Quality Index (WQI).** Water quality index gives the number which describes the overall quality of the river water at a particular location and time based on the parameters taken into the consideration. Ranging from 1-100, the value of WQI between 90-100 describes excellent water quality, 70-89 describes good water quality, 50-69 describes medium water quality, 25-49 describes bad water quality and 0-24 describes worst water quality.

MATERIALS AND METHODS

The panel data considered for the study is as follows:

Place	Period	Dissolved Oxygen	pH	Bio-Chemical Oxygen Demand (mg/L)	……	…….	Nitrate	Water quality Index
Jamshedpur	2005	6.0	7.7	1.7	.	.	2.2	24
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	2017	3.4	7.4	1.0	.	.	0	20
Ghatsila road bridge	2005	6.0	7.7	1.4	.	.	2.26	26
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	2017	4.5	7.4	0.8	.	.	0	20
Ghatsila	2005	0.3	7.7	0.9	.	.	.	27
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	2017	5.2	7.4	1.5	.	.	.	26
Baridhinala	2005	0.3	7.4	1.5	.	.	1.68	32
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	2017	0.7	7.4	3.5	.	.	0	26

The secondary data extracted from the Water Year Book 2018 of Central Pollution Control Board(CPCB) has been converted into the panel data by taking the observations of water quality parameters cross sectionally over the monitoring stations across years.

The Weighted Arithmetic Mean Method as proposed by ***Brown et.al(1972)***[6] has been used to calculate the water quality index, the formula of the same is as undernoted:

$$\boldsymbol{WQI = \frac{\sum_{i=1}^{n} q_i w_i}{\sum_{i=1}^{n} w_i}} \qquad \textbf{(1.4)}$$

where,

q_i=quality rating (sub index) of ith water quality parameter

w_i= unit weight of ith 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:

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

Where,

v_i= estimated value of the ith parameter

v_{io} = ideal value of the ith parameter

s_i= standard permissible value of the ith 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:

$$\boldsymbol{w_i = \frac{k}{s_i}}$$

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

Here, the Water Quality Index(WQI) is represented as dependent variable Yi, Sodium as x_1, pH as x_2, Dissolved Oxygen(DO) as x_3, Bio-chemical oxygen demand as x_4, Calcium as x_5, Iron as x_6, Nitrate as x_7, Nitrite as x_8, Potassium as x_9, Magnesium as x_{10}, Chlorine as x_{11} and Fluorine as x_{12}.

The fixed effect models have been applied on the panel data using the statistical software STATA. The statistical significance of the models have been compared for checking the goodness of fit of the model.

RESULTS

Fixed effect regression model using the method of Least Square Dummy Variable

As mentioned above, the data contains 4 cross-sectional units, hence, three dummy variables were created, taking the Jamshedpur as the reference cross-section such that they have value 1 for the respective cross-section otherwise 0 i.e., $Dummy_{Ghatsila}$ = 1 if Place = Ghatsila otherwise 0. The OLS model was then fit on the data including the three dummy variables created.

Table 1: ***ANOVA and Model fit***

Source | SS	df	MS	Number of obs = 60
----------+----------			F(15, 44) = 562.69
Model | 113796.42	15	7586.42801	Prob > F = 0.0000
Residual | 593.226654	44	13.482424	R-squared = 0.9948
----------+----------			Adj R-squared = 0.9930
Total | 114389.647	59	1938.80757	Root MSE = 3.6718

Table 2: ***Parameter Estimation***

Waterqualityindex	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
pH_GENpHunits	7.947616	2.742351	2.90	0.006	2.420771	13.47446
Biochemicaloxygende~L	1.690605	.0901777	18.75	0.000	1.508863	1.872346
DissolvedOxygenmgL	1.416922	.7468775	1.90	0.064	-.0883109	2.922154
CamgL	-.2788051	.1070492	-2.60	0.013	-.4945486	-.0630616
ClmgL	-.1901127	.0900868	-2.11	0.041	-.3716708	-.0085547
FmgL	17.33844	1.431606	12.11	0.000	14.45323	20.22366
FemgL	39.03916	1.174056	33.25	0.000	36.67301	41.40532
KmgL	.8397624	.2649937	3.17	0.003	.3057027	1.373822
MgmgL	.6466036	.2306117	2.80	0.007	.1818363	1.111371
NamgL	-.0759382	.1232161	-0.62	0.541	-.3242639	.1723875
NitriteNO2	46.47683	20.0046	2.32	0.025	6.160205	86.79345
NitrateNO3	.6055449	.4453931	1.36	0.181	-.292086	1.503176
Dummy_Ghatsilaroadb~e	-1.848189	1.406917	-1.31	0.196	-4.683644	.9872663
Dummy_Baridhinala	-1.414881	2.451734	-0.58	0.567	-6.356025	3.526264
Dummy_Ghatsila	-3.892269	1.546694	-2.52	0.016	-7.009425	-.7751126
Placeid						
2	0 (omitted)					
3	0 (omitted)					

4	0 (omitted)					
_cons	-51.55929	19.70876	-2.62	0.012	-91.27968	-11.83889

The table 1 provides the ANOVA table and Model fit table. The value of F-statistic test run on ANOVA depicted by the table 1 is quite high showing that the model is efficient for the present panel data of water quality parameters. The p-value depicted by the table is less than 0.05 at 95% confidence interval showing that the model is statistically significant for the model fitting.The R-square value given by the table of model fit shows that 99% of the variation in the Water quality index could be explained by the water quality parameters included in the model. The Adj R-square value shows that after adjusting the model for degrees of freedom, the model could explain 99% of the variation in the dependent variable by the independent variable.

The table 2 represents the parameter estimated by the model. The p-value of each independent variable given by the model shows that 9 out of 12 independent variables considered in the model were found significant as their p-value was found less than 0.05 for 95% confidence interval.

Fixed effect regression model using the method of Within-group estimation

The within-group estimator fits the OLS model on the difference variables created by subtracting the mean value of the variable from the observed value of the variable. The difference variables are created for all the independent and dependent variables.

Table 3: ***ANOVA and Model fit***

Source	SS	df	MS	Number of obs.	60
				F(12, 47)	425.07
Model	64382.2722	12	5365.18935	Prob > F	0.000
Residual	593.226694	47	12.6218446	R-squared	0.9909
				Adj R-squared	0.9885
Total	64975.4989	59	1101.27964	Root MSE	3.5527

Table 4*: Parameter Estimation*

Waterqualityindex_s~r	Coef.	Std. Err.	t	P>t	[95% Conf.	
pH_GENpHunits_star	7.948	2.653	3.000	0.004	2.610	13.286
BOD_star	1.691	0.087	19.380	0.000	1.515	1.866
DissolvedOxygenmgL_~r	1.417	0.723	1.960	0.056	-0.037	2.871
CamgL_star	-0.279	0.104	-2.690	0.010	-0.487	-0.070
ClmgL_star	-0.190	0.087	-2.180	0.034	-0.365	-0.015
FmgL_star	17.338	1.385	12.520	0.000	14.552	20.125
FemgL_star	39.039	1.136	34.370	0.000	36.754	41.324
KmgL_star	0.840	0.256	3.280	0.002	0.324	1.356
MgmgL_star	0.647	0.223	2.900	0.006	0.198	1.095
NamgL_star	-0.076	0.119	-0.640	0.527	-0.316	0.164
NitriteNO2_star	46.477	19.356	2.400	0.020	7.538	85.415
NitrateNO3_star	0.606	0.431	1.410	0.167	-0.261	1.472
_cons	0.000	0.459	0.000	1.000	-0.923	0.923

The table 3 provides the ANOVA table and Model fit table. The high value of F-statistic test shows that the model is efficient for the present panel data of water quality parameters. The p-value depicted by the table is less than 0.05 at 95% confidence interval showing that the model is statistically significant for the model fitting. The R-square value shows that 99% of the variation in the Water quality index could be explained by the water quality parameters included in the model. The Adj R-square value shows that after adjusting the model for degrees of freedom, the model could explain 98% of the variation in the dependent variable by the independent variable. The table 4 represents the parameter estimated by the model. The p-value of each independent variable given by the model shows that 9 out of 12 independent variables considered in the model were found significant as their p-value was found less than 0.05 for 95% confidence interval.

Fixed effect regression model using the method of First Difference

The first difference estimation creates the difference variable by subtracting the value of the variable of the previous year from the value of the current year and fits the OLS model on the difference variable thus created.

Table 5: ***ANOVA and Model fit***

Source	SS	df	MS	Number of obs	56
				F(12, 43)	506.26
Model	85520.7726	12	7126.73105	Prob > F	0
Residual	605.319426	43	14.077196	R-squared	0.993
				Adj R-squared	0.991
Total	86126.092	55	1565.92895	Root MSE	3.752

Table 6: ***Parameter Estimation***

Waterqualityindex_FD	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
pH_GENpHunits_FD	5.257	2.284	2.300	0.026	0.650	9.864
Biochemicaloxygende~D	1.612	0.065	24.960	0.000	1.482	1.742
DissolvedOxygenmgL_FD	-0.020	0.583	-0.030	0.972	-1.197	1.156
CamgL_FD	-0.214	0.080	-2.690	0.010	-0.375	-0.054
ClmgL_FD	-0.142	0.067	-2.130	0.039	-0.276	-0.008
FmgL_FD	20.327	1.071	18.980	0.000	18.167	22.487
FemgL_FD	38.644	0.762	50.730	0.000	37.108	40.180
KmgL_FD	0.394	0.210	1.880	0.067	-0.029	0.817
MgmgL_FD	0.757	0.192	3.950	0.000	0.370	1.145
NamgL_FD	-0.042	0.081	-0.520	0.604	-0.205	0.120
NitriteNO2_FD	46.930	14.531	3.230	0.002	17.625	76.235
NitrateNO3_FD	-0.549	0.401	-1.370	0.179	-1.357	0.260
_cons	-1.024	0.517	-1.980	0.054	-2.067	0.018

The table 5 provides the ANOVA table and Model fit table. As discussed, the number of observations for present panel data is 60 but the first difference model considers 56 observations as during creating the difference variables the variables for the year 2005 have been removed from the data as there is no previous year for 2005 and the difference could not be created for the year 2005.The high value of F-statistic test and the p-value depicted by the table shows that the model is statistically significant for the model fitting. The R-square value shows that 99% of the variation in the Water quality index could be explained by the water quality parameters included in the model. The Adj R-square value shows that after adjusting the model for degrees of freedom, the model could explain 99% of the variation in the dependent variable by the independent variable. The table 6 shows that the p-value of 8 out of 12 independent variables considered in the model were found significant as their p-value was found less than 0.05 for 95% confidence interval.

Model Comparison

Table 7: ***Table of model comparison***

Model	Root Mean squared error	R-square	No. of non-zero coefficients
Fixed effect: within-groups	3.55	0.990	9
Fixed effect: Least square dummy variable	3.67	0.994	9
Fixed effect: First difference	3.75	0.993	8

Upon comparing the three fixed effect panel regression models, it is evident that where all the models nearly explained the same amount of variation in the dependent variable, the **Fixed effect within group estimation model** provided the least root mean squared error and thus is considered as the best goodness of fit model.

DISCUSSION

The present work is devoted to elaborate the use of fixed effect panel regression models in modelling the water quality parameters of river Subarnarekha observed for years 2005- 2017 commencing 4 different monitoring stations, taking 12 water quality parameters into consideration, extracted from the Central Pollution Control Board Portal. The study employed three methods of parameter estimation for the fixed effect panel regression models. Upon comparing the models for best goodness of fit results, it was found that the **fixed effect within-group estimation method** provided the least value of root mean squared error. Hence, the study recommends the use of **Fixed Effect Within-Group Estimation** technique for modelling the water quality parameters.

REFERENCES

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

2. W. M. Mason, Statistical Analysis: Multilevel Methods, International Encyclopedia of the social sciences and behavioral sciences, 2001.

3. James H. Stock, Mark W. Watson, Introduction to econometrics Boston: Pearson Addison Wesley, 2003.

4. Paul D. Allison, Fixed Effects Regression Models, Sage Publications, 2009.

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.