Journal of Quantitative Finance and Economics
Volume 5; Number 1; 2023; pp : 1-23
https://DOI:10.47509/JQFE.2023.v05i01.01

ARF INDIA
Academic Open Access Publishing
www. arfjournals. com

A Multi-Level Quantitative Analysis of the Relation of Socio-Economic Factors with the Gender Ratio in India

Khyati Khandelwal

Oxford Internet Institute, University of Oxford

To cite this paper:

Khyati Khandelwal (2023). A Multi-Level Quantitative Analysis of the Relation of Socio-Economic Factors with the Gender Ratio in India. *Journal of Quantitative Finance and Economics.* 5(1), 1-23. https://DOI:10.47509/JQFE.2023.v05i01.01

Abstract: *For many years, India has been one of the top countries with an unnaturally skewed gender ratio towards the male gender. This implies the clear evidence of pre-natal selection and/or selective abortion in the country. While a lot of research has been conducted encapsulating the probable cultural factors (on a country-wide level) that adversely affect the gender ratio, only a few have used regression models to model the relationship between factors such as literacy rate and sex ratio. Hence, in this paper, I have utilised data that spans over 30 years from the Reserve Bank of India to model the relationship of various socio-economic factors such as poverty rate and literacy rate with the sex ratio in various Indian states. This was done through a multi-level model with the Indian states as the second level, in order to model the variation in sex-ratio due to the cultural and historical diversity in Indian states as well. It was found that over 90 per cent of variance in sex ratio could be explained through the random effects of Indian states. Findings also indicate a positive relationship of sex ratio with literacy rate and poverty rate. A negative relationship between the percentage of rural population in an Indian state and its sex ratio was also found.*

Introduction

In a developing country like India, where culture plays a pivotal role in the socio-economic activities, it must be noted that it stands just 13th on the list of countries with the lowest gender ratios (number of females per 1000 males) in the world (World Bank Data Catalogue 2022). Making up 18 per cent of the world's population (World Bank Data Catalogue 2022), it becomes interesting to study India's gender ratio and the factors affecting it.

The field of gender inequality is highly researched, although most of the claims have remained qualitative. While papers such as the one regarding imbalanced sex ratio at birth by Li Shuzhou (2007) mentioned that rural areas in developing countries were the main hub of imbalanced sex ratio,

Article Info.: Received: 16 January 2023; Revised: 14 February 2023; Accepted: 27 February 2023; Online: 30 June 2023

there were others (Jaychandaran 2015) that had found that rural/ underdeveloped areas within a geography may not have the technology to determine the sex of a child prior to birth, and hence, those areas had lesser chances of female foeticide/infanticide. Gu, in her paper (1995) even discussed the impact of socio-economic development of a particular geographic locality on the sex ratio, especially in developing countries. She also discussed that other factors such as poverty rate might not be so important in determining the gender ratio in a particular sub-region as the cultural aspects in terms of want for a male child and patriarchal mindset, would be.

Bhattacharya (2015) in his paper revealed the inequality and prejudice a fraction of the society bears based on gender, in Russia, India, and China. The paper compared the differences amongst these countries related to gender inequalities and threw light on the policy issues faced by them. The study revealed that in a few north-eastern states, female children generally tend to possess a higher chance of survival than in other parts of India. It is to be noted that in a diverse country like India, it could prove to be interesting to capture these differences due to cultures and geographies.

Within the context of India, Rebeca (2010) proposed statistical methods to quantify the dependence of literacy rate and sex ratio in India. However, the study failed to capture other similar indicators such as poverty rate, etc., which this study aims to capture along with literacy rate.

Further, much research has been done in understanding the various population performance indicators such as gross domestic product, population density, literacy rate or poverty rate to understand their relationship with the sex ratio. In an international study, including both India and China, Seema Jaychandran (2015) mentioned the strong relationship between a high national GDP and an improved gender ratio. I will further build upon such studies, to explore whether similar or different findings are brought to light when considering data from within a country, segregated by states.

In order to capture group level, as well as individual level variances, multi-level models were employed in the past for economic data gathered from repeated measures. Multilevel modelling approach was employed by Novak (2017) to explain the influence of economic development on the subjective well-being of individuals (Novak & Pahor 2017) We will try to use a similar approach to capture within-level as well as between-level variables.

Overall, this paper aims to model the impact of socio-economic factors through numerical means in order to quantify their impact on the gender-ratio in India. It will also consider the variation that might be explained through geography (which is Indian states in our case). In this paper, I will explore the extent to which the difference in states impacts the sex-ratio, as it may imply that it is not only the economic variables, but also social and cultural differences that help in determining the improvement in sex ratio even within a particular country. Further, this study aims to use these findings to add context to previous studies, which were largely qualitative, through quantitative measure and analysis.

Methods

Data Collection and Pre-processing

The data used in this study was obtained from the official national statistics archives of the Reserve Bank of India [10]. Since some of the data was present in pdf (.pdf) format while the other was available in excel (.xlsx) format, it was all first collated by indicator, year and state, in an excel sheet. Initially, data was obtained for past 70 years (one measurement every 10 years). While my proposed model could deal with missing values within a column, there were a few years where the entire column for a particular indicator was missing. Hence, it was narrowed down to past 30 years of data, with repeated measurements taken every 10 years.

The indicators (independent variables) used in this data set are socio-economic variables that were available through Reserve Bank of India's primary collection methods through bodies such as National Statistics Survey of India, Office of the Registrar General and Census Commissioner, Ministry of Home Affairs, Government of India as well as NITI Aayog. These indicators included:

Rural Population: Number of people, in hundred thousands that live in rural areas in a particular state.

Literacy Rate: Number of people educated till the age of 14, as a per cent of the total population of a particular state.

Poverty Rate: Number of people in hundred thousands, living below poverty line. Here the poverty line is defined as the monthly consumption expenditure of below INR 972 per month in rural areas of a state, and that below INR 1407 in urban areas of that state.

Population Density: The number of people residing permanently per square kilometre in a state.

GDP: The gross domestic product[1] of the state.

Finally, as the response variable, I had taken the gender ratio (also referred to as the sex ratio in this paper, interchangeably). Here:

Sex-Wise Ratio: Number of females per thousand males in a particular state.

For each of the above columns, I had repeated measurements for the years 1991, 2001 and 2011. While new data was available for some indicators, it was not available for all the variables of interest since the last national census for many indicators was taken in 2011 in India.

A total of thirty-five States and Union Territories formed the groups under which each indicator was measured. This number excludes the state of Telangana which was created after 2011. For the Union Territories and States that were not present in a specific time period, the values were interpolated using moving averages from the data for the remaining time periods.

Preliminary Data Analysis

In order to visualise the data first, each indicator was grouped by year, and plotted in a box plot. Here, each point denoted the measurement of that indicator per state, plotted over the years. A box plot is used for presenting the '5-number summary' which consists of the minimum and maximum range values, the upper and lower quartiles, and the median (Potter 2006). This plot of values provides a convenient way to visualise the distribution of a dataset.

Second, a correlation heat map plot was used to check for the presence of correlation between various variables. This plot visualises the correlation between every combination of variables present in the dataset. Since all the variables were continuous values (as opposed to being categorical or ranked), I used the Pearson's correlation coefficient (*r*) defined as:

$$r = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma}(x_i - \tilde{x})^2 \Sigma(y_i - \tilde{y})^2}$$

Where,
r = Pearson's correlation coefficient,
x_i = values of x-variable in the sample,
y_i = values of y-variable in the sample,
$\overline{x}$ = mean of the values of x-variable,
$\overline{y}$ = mean of the values of y-variable.

In case of the presence of correlation, the Variance Inflation Factor (VIF) was calculated for each of the independent variables in order to determine the limitations of this model. It is calculated as:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{x_j|x_{-j}}}$$

Where $R^2_{x_j|x_{-j}}$ is the R^2 from a regression of xj onto all the other x's [15].

The next step was to check for linear/quadratic/etc. relationships between 'Sex Ratio' and the explanatory variables for the model. This was done through plotting of scatter plots. In case no linear relationship was found, some transformations were performed in order to determine the relationship.

Taking the logarithm (log(x)) of a value affects larger values more than smaller values. This can be helpful if the distribution has a positive skew because it brings larger values closer to the centre. It must be noted that I couldn't take the log of 0, but since my data did not contain zeros so it did not cause a problem.

Multi-Level Model

Mechanism of a multi-level model and reasons for using it over linear regression:

While regression models can be quite effective in studying the behaviour of certain patterns in longitudinal data with repeated measures as they capture fixed effects, in case there is nesting or hierarchy present in the data, mixed effect models may be preferred, especially when it is hypothesised that the grouping in data may have a large impact on the independent variable in consideration.

West, Welch, and Gatecki (2014, p.9) provided a structured definition of fixed-effects and random-effects, "Fixed-effect parameters describe the relationships of the covariates to the dependent variable for an entire population, random effects are specific to clusters of subjects within a population." When West, Welch, and Gatecki (2014) talked about "relationships of the covariates to the dependent variable", the covariates were the independent variables in the model (West 2014).

Mixed effects models (or multi-level models) can, not only model the random effects of a clustering variable, but also capture and model variation around the intercept (random intercept model), around the slope (random slope model), and around the slope and intercept both (random intercept and slope model).

Mathematically, a generalised version of the equation of a regression model is:

$$y = \beta X + \epsilon$$

Where, y = dependent variable, βX = fixed effects, ϵ = error term.

My mixed effects model used the following general equation:

$$y = X\beta + Zu + \epsilon$$

Where, y = dependent variable, βX = fixed effects, Zu = random effects, ϵ = error term.

The model is based on the assumptions of independence of errors, equal variance of errors, as well as normality of errors.

Model and method specific to this study: In order to validate whether a multi-level model was required in the first place, I used the approach as discussed and put forth by Hair Jr. (2019). I used a variance components model (null model) to provide a baseline for subsequent analysis.

$$y_{ij} = \beta_0 + u_j + \epsilon_{ij}$$

Where, y_{ij} denotes the outcome for the jth member of the ith group, where β_0 is an overall mean, $u_j \sim N\,(0, \sigma_u^2)$ is a random effect for group i and $\epsilon_{ij} \sim N\,(0, \sigma_e^2)$ is the usual individual error term.

In this model, there were a total of two levels in the data. Level 1 was the lowest level of the hierarchy which was the unit of analysis, i.e. the measurements of indicator for each year (1991, 2001 and 2011). In the mixed effects model equation above, this is the "i" in the equation. Level 2 was the next level of the data hierarchy where all units of analysis from level 1 were clustered into groups. For my model, these groups were the respective states in which the units were analysed. In the mixed effects model equation above, this is the "j" in the equation.

Preliminary results from the above analysis were used to calculate the inter-class correlation coefficient (rstate) (also referred to as the variation partition coefficient) (Mahdi 2022).

$$r_{state} = \frac{\sigma_u}{\sigma_u + \sigma_e}$$

After obtaining these results, a Random Intercept Model without Interaction Term was employed. The equation used in my model was:

$$y_{ij} = \beta_0 + \beta_1\, RP_{1ij} + \beta_2\, LR_{2ij} + \beta_3\, PR_{3ij} + \beta_4\, PD_{4ij} + \beta_5\, GDP_{5ij} + u_j + \epsilon_{ij}$$

where, y_{ij} is the sex ratio (dependent variable) for each measurement in each group, β_0 is the global intercept for the model, RP_{1ij} represents measurements of rural population for 30 years for each state, LR_{2ij} represents measurements of literacy rate for 30 years for each state, PR_{3ij} represents measurements of poverty rate for 30 years for each state, PD_{4ij} represents measurements of population density for 30 years for each state, GDP_{5ij} represents measurements of population density for 30 years for each state. All these variables were measured at time intervals of 10 years.

My aim was to model the relationship between these socio-economic factors and sex ratio as a 1 level regression model to determine the significant explanatory variables and their corresponding coefficients, while also capturing the effect of variation due to various Indian states in level 2.

Advantages of Multi-Level Model: Individual level data may vary in their number of measurements by design or due to attrition; this does not have an effect on the model.

Similarly, individuals with missing dependent variable value were included under the missing at random assumption[2] in these models.

There is flexibility in the specification of dependency of the independent variable on random effects (z), for example, polynomial, spline, step functions.

It allows for clustering at higher levels, such as geography to understand the effect on individuals for belonging to a particular group, of our concerned variable.

Libraries and packages used: For the purpose of this study I used Python for data analysis. The libraries and packages used are Pandas, NumPy (popular packages for the analysis of data), Matplotlib and Seaborn (for data visualisation). For my statistics model, I used Statsmodels 'mixedlm'. The code can be found in the appendix.

Analysis and Results

Visualising the Data

First, let us look at all the indicators and their variation with time.

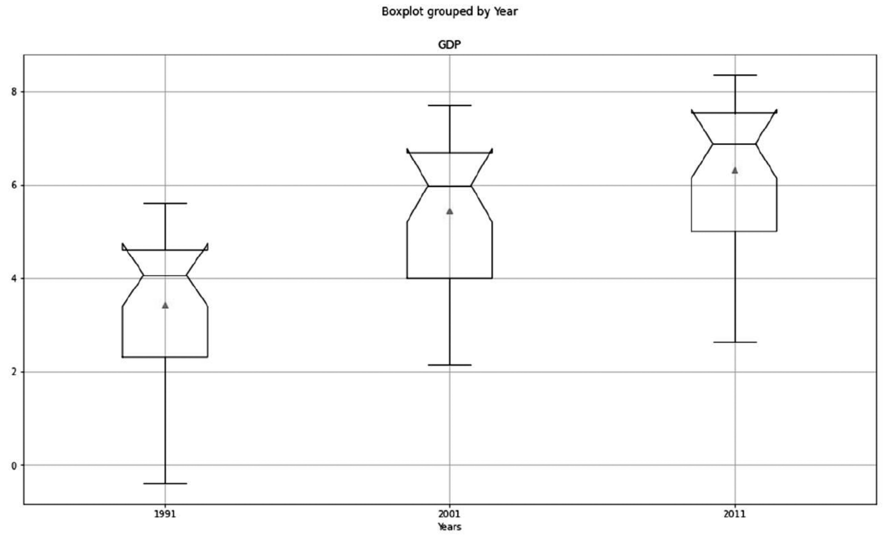


Figure 1: Above is the GDP plotted against the three decades. Here, the box plot is plotting the data points from each state, for the respective year as present. I observed that there is an upward trend in GDP of states over the years, with an increasing mean and no outliers

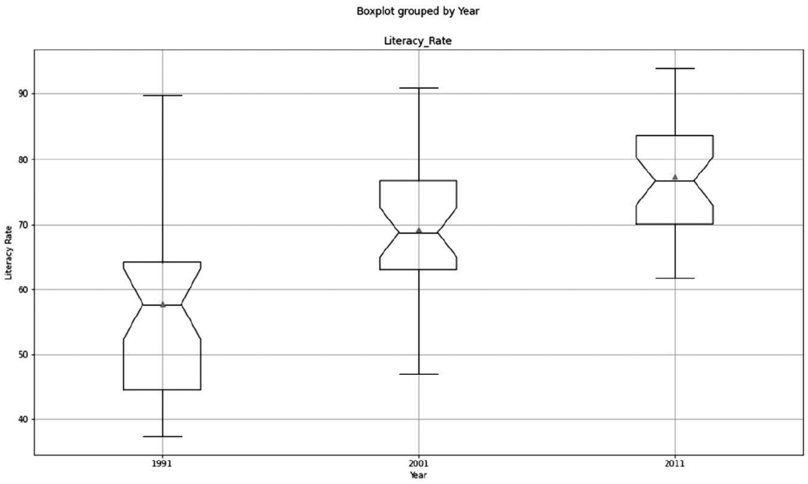


Figure 2: Above is the box plot of literacy rate measured for each state (represented by the boxes) for the respective years. I observed that in 2001, the spread of data is high, implying high differences in literacy rates between different states. However, in 2001 and 2011 not only is the literacy rate observed to increase, but also there is a decrease in spread of the data

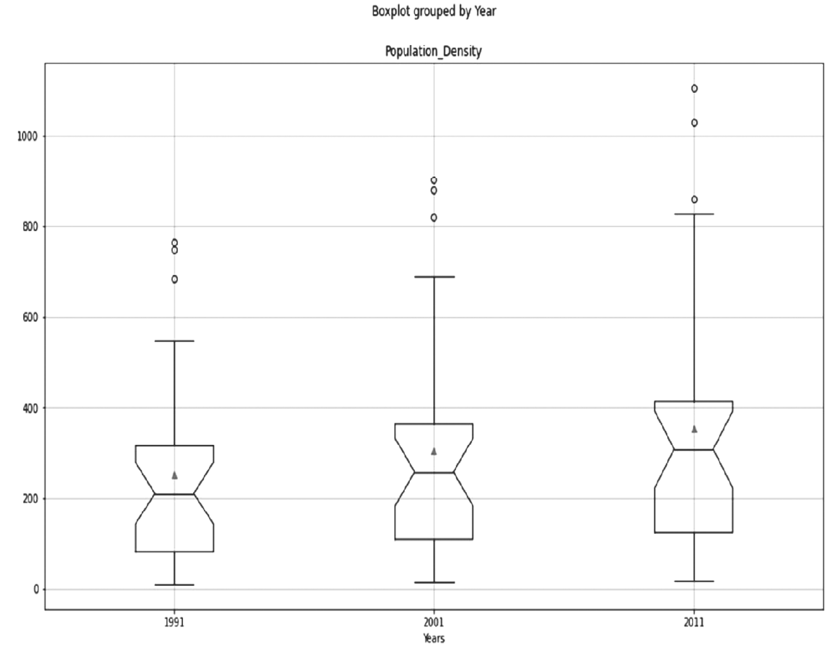


Figure 3: Above I have plotted the population density measured for each state, where each box represents measurements from the years 1991, 2001, 2011 respectively. There is presence of a few outliers throughout the three decades. Increase in overall population density over time is also observed through this plot

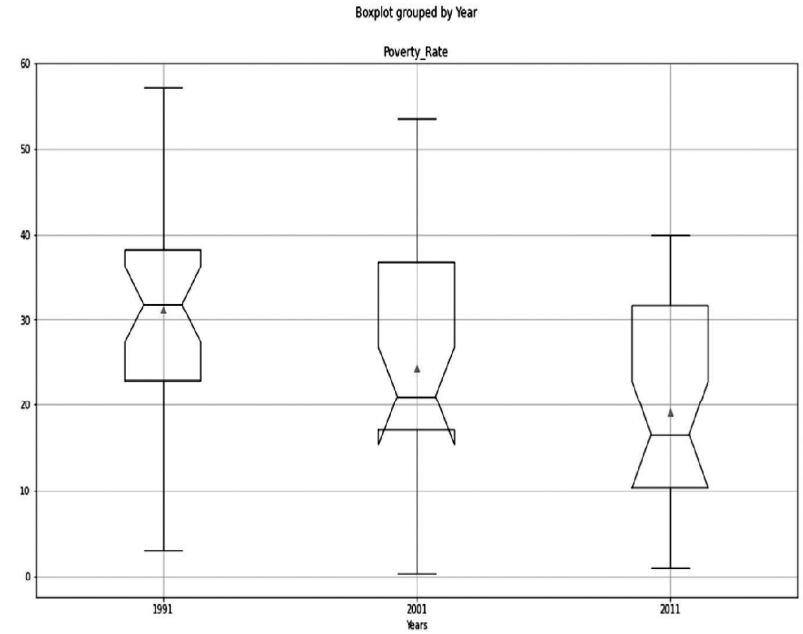


Figure 4: From the plot above, I observed poverty rate over the years, as measured for each state. While it is clear from the mean marks in the plot that average poverty rate is decreasing, it is also observed that the spread in poverty rate among the states has increased

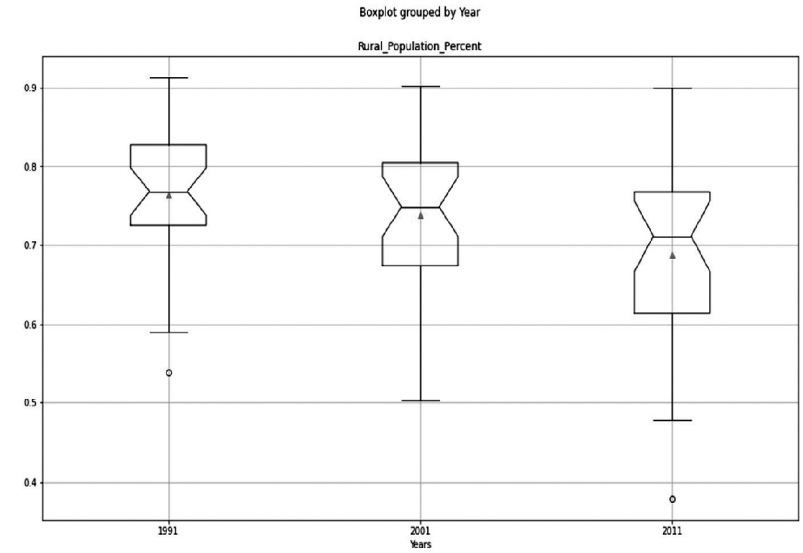


Figure 5: The above box plots indicate the decadal measurements of rural population, as a per cent of total population. The spread in box plots indicates the spread in poverty rates among various states of India. Poverty rate is observed to be strictly decreasing over the years

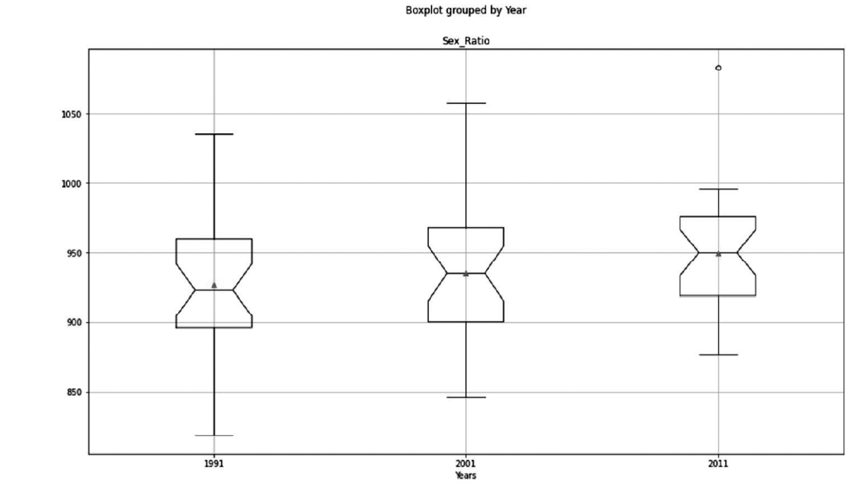


Figure 6: In the above graph, each box plot represents the measurements of sex ratio of each state in the particular year as indicated in the categorical y-axis. Not much linear growth in sex ratio is visible over time. Further, the length of the whiskers for the 2011 period is observed to be considerably lesser than that of the previous two decadal measurements

Next, I plotted a scatter plot for all independent variables with the dependent variable (sex ratio) to determine if there was a linear relationship present for my model.

For GDP, I had chosen to log-transform the GDP, as it would bring down the extremely large values, to have a better fit for the model.

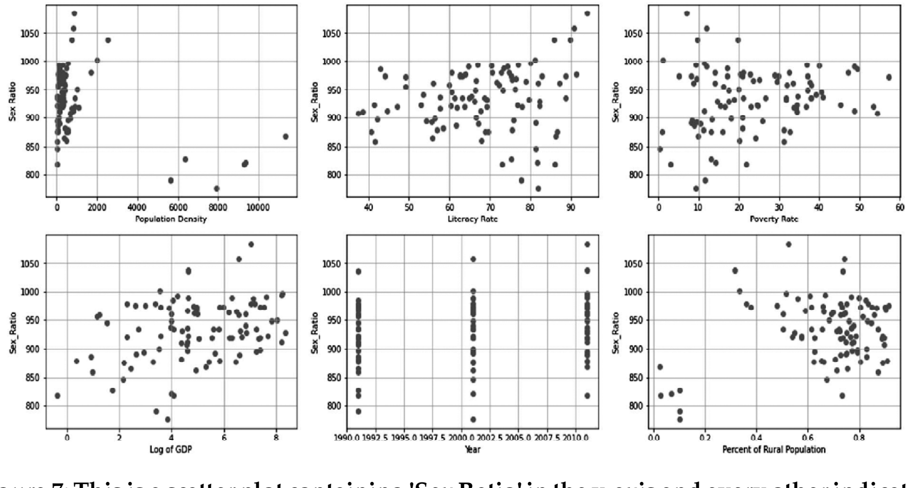


Figure 7: This is a scatter plot containing 'Sex Ratio' in the y-axis and every other indicator, plus year, in the x-axis. It is observed that excluding a few outliers, a linear relationship is visible for the various indicators

Finally, the correlation plot among the various factors was visualised in order to contextualise the findings of the model, and check for correlation between different factors.

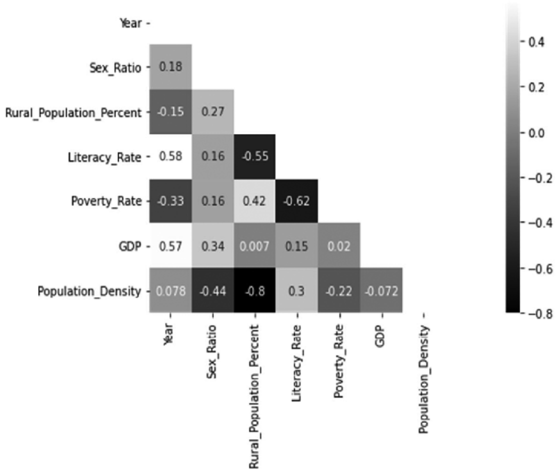


Figure 8: This is a correlation heat map where lighter shades of grey represent higher positive correlation and darker shades represent higher negative correlation between the variables. Moderate negative correlation is observed between Literacy Rate and Poverty Rate (-0.62) as well as Literacy Rate and Per cent of Rural Population (-0.55), while Per cent of Rural Population and Population Density show high negative correlation (-0.8). Literacy Rate also shows moderate positive correlation with Year (+0.58)

Further analysis of the presence of co-linearity between variables is shown in the appendix. I found that the Variance Inflation Factor was not problematic (was less than 5) for my variables (Mahdi 2022). Hence, I could go ahead with my model.

Modeling the Data

First, I fitted the null model (variance components model) to find the variance partition coefficient for my 2-level model in statsmodels in Python. The following results were obtained for the same:

```
                     Mixed Linear Model Regression Results
-------------------------------------------------------------------------------
Model:                    MixedLM Dependent Variable: Sex Ratio
-------------------------------------------------------------------------------
No. Observations:                 96     Method:                REML
No. Groups:                       32     Scale:                 287.8648
Min. group size:                   3     Log-Likelihood:        -458.8348
Max. group size:                   3     Converged:             Yes
Mean group size:                 3.0
-------------------------------------------------------------------------------
                                 Coef.        Std.Err.    z      P>|z|   [0.025 0.975]
-------------------------------------------------------------------------------
Intercept         931.771        9.505        98.029      0.000  913.141  950.400
State Var                        2795.090                 52.731
-------------------------------------------------------------------------------
```

Here, I calculated the VPC (Level 2 Inter-Class Correlation Coefficient) as 2795.090/(2795.090 +287.8648) = 0.91178. In this last case, I had estimated that States random effects represented approximately 92 per cent of the total variance of the residuals (Grech, Mamo & Xjenza 2014). The intercept (931.771) of this null model indicated that the average decadal sex ratio of the States was expected to be 931.771 females per 1000 males. This provided a good reason to choose multi-level model over a simple regression model.

Next, I incorporated all of my independent variables into the model, with the group as States. The following results were obtained:

Mixed Linear Model Regression Results

Model:	*MixedLM*	*Dependent Variable:*	*Sex_Ratio*
No. Observations:	87	Method:	REML
No. Groups:	29	Scale:	105.4760
Min. group size:	3	Log-Likelihood:	-368.7136
Max. group size:	3	Converged:	Yes
Mean group size:	3.0		

	Coef.	*Std.Err.*	*z*	*P>|z|*	*[0.025*	*0.975]*
Intercept	872.804	48.541	17.981	0.000	777.665	967.943
Poverty Rate	1.051	0.249	4.220	0.000	0.563	1.540
Literacy Rate	0.859	0.303	2.832	0.005	0.264	1.453
GDP	1.453	2.232	0.651	0.515	-2.922	5.828
Population Density (log)	10.092	8.308	1.215	0.225	-6.193	26.376
Rural Population Percent	-109.735	33.826	-3.244	0.001	-176.032	-43.438
State Var	1626.390	55.125				

From amongst the p-values obtained in the above results, I discarded the variables with a p-value higher than the accepted standard of 0.05 from the model. Hence, I obtained the following results after discarding the insignificant variables:

Mixed Linear Model Regression Results

Model:	*MixedLM*	*Dependent Variable:*	*Sex Ratio*
No. Observations:	87	Method:	REML
No. Groups:	29	Scale:	107.5023
Min. group size:	3	Log-Likelihood:	-375.0735
Max. group size:	3	Converged:	Yes
Mean group size:	3.0		

	Coef.	*Std.Err.*	*z*	*P>|z|*	*[0.025*	*0.975]*
Intercept	916.365	34.745	26.374	0.000	848.267	984.464
Poverty Rate	1.098	0.247	4.452	0.000	0.615	1.582
Literacy Rate	1.204	0.206	5.843	0.000	0.800	1.607
Rural Population Percent	-118.896	33.286	-3.572	0.000	-184.134	-53.657
State Var	1652.202	54.420				

The model above was finally interpreted. Firstly, the inter-class correlation coefficient for the model was calculated as 1652.202/(1652.202+107.5023) which was equal to 0.9389. Thus, it can be said that the states random effects represent approximately 93.89 per cent of the variance of the residuals. Next, from this model, I can determine that a unit increase in poverty rate results in 1.098 units increase in the sex ratio; a unit increase in literacy rate results in 1.204 units increase in the sex ratio; a unit increase in the percentage of rural population results in 118.896 units decrease (due to negative sign in coefficient) in sex ratio.

Discussion

Findings

From the data visualisation, a few interesting findings were generated. Firstly, in Figure 1, I observed a growth in GDP for almost all of the states over each decade. India, being one of the fastest developing countries, shows this consistent growth over the years.

Similarly, a considerable growth in literacy rate over the years was observed in Figure 2. The reason for this steady growth is considered to be the effective steps taken by the government of India such as making elementary education a fundamental right for an Indian citizen in 2002, free educational programs, and forming newer educational institutes country-wide (Shah 2013).

Population density in Figure 3 was also observed to be growing over the years, with a few outliers on the higher side. The presence of a few outliers was likely due to flock of people from smaller towns across the country to business and development hubs like Delhi and Mumbai.

In the graphs of percentage of rural population, as well as poverty rates over the years in Figure 4 and Figure 5, an interesting phenomenon was captured. While both the variables continued to decrease, the variance went up over time. Das (2012) talked about this widening urban-rural gap in India, where some places continue to develop drastically more than others, both economically as well as socially.

Finally, upon analysing Figure 6, I was able to observe that the sex ratio, although improved with time (with higher females per thousand males each decade), this growth curve was not as steep as that of literacy rate or GDP. This lends to the notion that while other factors that indicate social development, such as literacy rate are steadily increasing, the gender ratio fails to catch up. Hence, this study proposes that there are certain unaddressed cultural factors which affect the improvement of gender ratio and lead to bias against female child in India.

In Figure 8, the correlation matrix indicates the high negative correlation between population density and percentage of rural population among the states[3]. Moderate negative correlation between literacy rate and poverty rate is also observed, which corroborates the idea that education can lead to better economic conditions.

In terms of the dependent variable, i.e., gender ratio, while there isn't a significant correlation between any of the factors, there is some moderate positive correlation with population density. This could either be due to random chance, or the correlation coefficient may have been impacted due to the presence of higher proportion of females in densely populated urban areas. Although, the second reason can be questioned, as many papers in the past have suggested that urbanisation in India may even sometimes lead to a negative impact on social indicators such as gender ratio (Das, & Pathak 2012) due to better availability of pre-natal sex detection equipment (Echávarri & Ezcurra 2010), combined with no social or cultural changes in bias against females.

Finally, through my model, I was able to establish that nearly 90 per cent of variance in sex ratio can be explained by the effect of Indian States. Research by Ritchie (2019) concluded that 'in countries such as Indonesia and India, where there is a clear son preference, the sex ratio at birth becomes increasingly skewed with birth order (the third or fourth born children are more likely to be boys than the first or second child).' Such reasoning may be applied to various states within India through the findings of my model. It could be that similar disposition against female child in a few states culturally and historically, is one of the major factors affecting sex ratio.

This model also factors in socio-economic variables and their impact on the sex ratio. While one would intrinsically reason the presence of a positive coefficient between literacy rate and sex ratio, the reasoning may not be so straightforward for the positive coefficient between poverty rate and sex ratio. My findings in this paper for poverty rate's coefficient sign are robust to the addition of new variables (including 'Year' as an explanatory variable). It is proposed that economically richer families' *willingness, ability, and readiness* (causes for pre-natal sex selectiveness, as explained by Ritchie 2019) to opt for a male child are all favourable. The willingness as put forth by Singh (2021), is that mothers belonging to middle-class to richer families in India have stronger desire for a son, to take care of them in their old age and take forward the family name. The ability and readiness to go for pre-natal selection hence follows due to the presence of ample resources to perform abortion of female child and opt for a son instead.

Finally, the findings on negative relationship between the percentage of rural population present in a state and the sex ratio in that state simply follow from better penetration of sex-determination technologies in urbanised areas over rural areas (Jaychandaran 2015). Many members of rural communities may be *willing* to opt for a male child but might not have the means to do so monetarily.

Limitations

There are a number of limitations in the study due to the kind of data employed for the model. Both wider (spanning over more years) and deeper (containing a greater number of measurements) forms of data would improve the accuracy of results and claims for this study. The latest year used in our model was 2011, and more recent measurements would likely have resulted in more current findings. In the multi-level model, in order to further contextualise the findings, more levels such as district or even household level data could prove to be more helpful in truly understanding the impact of belonging to a particular geography over another.

Conclusion and Future Study

To conclude, we can say that our preliminary findings are in accordance with past literature, while our model itself provides a new perspective to the pre-existing literature on gender ratio. It improves on the past regression models that fail to capture state-level variances, while providing a quantitative support to prior qualitative claims regarding causes of skewness in gender ratio in various geographies of the world.

Further research could possibly utilise the district or household level data, as well as data that spans over longer periods of time and has more recent measurements for better accuracy in results.

Notes

1. The GDP is calculated through calculation of the expenditure (at market prices) method. It involves summing the domestic expenditure on final goods and services across various streams during a particular time period (it is the net domestic product at factor cost for each state).
2. Missing at random means there might be systematic differences between the missing and observed values, but these can be entirely explained by other observed variables.
3. The proposed reasons for this are two factors. Firstly, the passive factor is that vast areas of Indian geographies consist of inhabitable land and climate conditions, where only rural areas are present with very low population densities. Secondly, the active factor is that a large number of people migrate from rural areas to urban areas, causing disbalance in population densities.

References

Bhattacharya, Prabir C., & Saxena, V. (2015). *Socio-economic determinants of child and juvenile sex ratios in India: A longitudinal analysis with district-level data*. Heriot-Watt University Economics Discussion Papers 1503.

Das, D., & Pathak, M. (2012). "The growing rural-urban disparity in India: Some issues." *International Journal of Advancements in Research & Technology* 1.5: 1-7.

Echávarri, R.A., & Ezcurra, R. (2010). Education and gender bias in the sex ratio at birth: Evidence from Potter, India. *Demography*; 47 (1): 249–268.

Goldstein, H., Healy, M.J.R., & Rasbash, J. (1994). "Multilevel time series models with applications to repeated measures data." *Statistics in Medicine 13.16*: 1643-1655.

Grech, V., Mamo, J., & Xjenza. (2014). Vol.2(1), p. 81-90. Retrieved from *https://www.xjenza.org/The-male-to-female-ratio-at-birth_Pold_33.html*

Gu, B., & Roy, K. (1995). "Sex ratio at birth in China, with reference to other areas in East Asia: what we know." *Asia Pacific Population Journal 10*: 17-42.

Hair Jr, Joseph, F., & Fávero, L.P. (2019). "Multilevel modelling for longitudinal data: Concepts and applications." *RAUSP Management Journal 54*: 459-489.

https://www.rbi.org.in/Scripts/AnnualPublications.aspx?head=Handbook%20of%20Statistics%20on%20Indian%20States

Jaychandaran, S. (2015). The roots of gender inequality in developing countries. *Annu. Rev. Econ. 7*:63–88.

Kristin *et al.* (2006). "Methods for presenting statistical information: The box plot." *VLUDS.*

Mahdi, A. (2022). *Lecture Notes – Applied analytical Statistics*. University of Oxford.

Novak, M., & Pahor, M. (2017). Using a multilevel modelling approach to explain the influence of economic development on the subjective well- being of individuals, *Economic Research-Ekonomska Istrazivanja*, 30:1, 705-720, DOI: 10.1080/1331677X.2017.1311229

Ritchie, H., & Roser, M. (2019). "Gender ratio." *Our World in Data.*

Shah, N. (2013). "Literacy rate in India." *International Journal of Research in All Subjects in Multi Languages* 1.7: 12-16.

Shuzhuo, Li. (2007)."Imbalanced sex ratio at birth and comprehensive intervention in China." 4th Asia Pacific Conference on Reproductive and Sexual Health Rights, Hyderabad, India.

Singh, A., Kumar, K., Yadav, A.K., James, K.S., McDougal, L., Atmavilas, Y., & Raj, A. (2021). Factors influencing the sex ratio at birth in india: A new analysis based on births occurring between 2005 and 2016. *Stud Fam Plann. 52*(1):41-58. doi: 10.1111/sifp.12147. Epub 2021 Feb 22.

West *et al.* (2014). "Linear mixed models: An overview." *Linear Mixed Models* : 9-58.

World Bank Data Catalogue. (2022). Retrieved from *https://datacatalog.worldbank.org/home*

Appendix

Data

State/Union Territory	*Year*	*Sex Ratio (females per 1000 males)*	*Literacy Rate (percent)*	*Poverty Rate (percent)*	*Population Density*	*GDP (INR crore)*	*Rural Population (Percent)*
Andaman & Nicobar Islands	1991	818	73.02	3	34	0.68	73.31%
Andaman & Nicobar Islands	2001	846	81.3	0.4	43	8.57	67.42%
Andaman & Nicobar Islands	2011	876	86.63	1	46	26.97	62.20%
Andhra Pradesh	1991	972	44.08	29.9	242	143.42	73.11%
Andhra Pradesh	2001	978	60.47	21.1	277	1317.50	72.70%
Andhra Pradesh	2011	993	67.02	9.2	308	3638.35	66.64%
Arunachal Pradesh	1991	859	41.59	31.1	10	2.64	87.17%
Arunachal Pradesh	2001	893	54.34	25.9	13	18.64	79.23%
Arunachal Pradesh	2011	938	65.39	34.7	17	53.27	77.02%
Assam	1991	923	52.89	34.4	286	35.77	88.90%
Assam	2001	935	63.25	37.9	340	336.68	87.09%
Assam	2011	958	72.19	32	398	706.83	85.90%
Bihar	1991	907	37.49	54.4	685	96.73	89.60%
Bihar	2001	919	47	53.5	881	505.79	89.54%
Bihar	2011	918	61.8	33.7	1106	1503.98	88.70%
Chandigarh	1991	790	77.81	11.6	5632	30	10.28%
Chandigarh	2001	777	81.94	9.2	7900	46.49	10.21%
Chandigarh	2011	818	86.05	21.8	9258	59.40	2.75%
Chhattisgarh	1991	985	42.91	49.4	130	58.14	82.60%
Chhattisgarh	2001	989	64.66	48.7	154	552.22	79.91%
Chhattisgarh	2011	991	70.28	39.9	189	2016.53	76.76%
Delhi	1991	827	75.29	13.1	6352	5.66	10.07%

contd.

State/Union Territory	*Year*	*Sex Ratio (females per 1000 males)*	*Literacy Rate (percent)*	*Poverty Rate (percent)*	*Population Density*	*GDP (INR crore)*	*Rural Population (Percent)*
Delhi	2001	821	81.67	14.2	9340	54.92	6.82%
Delhi	2011	868	86.21	9.9	11320	202.16	2.50%
Goa	1991	967	75.51	25	316	99.44	58.97%
Goa	2001	961	82.01	8.7	364	934.55	50.22%
Goa	2011	973	88.7	5.1	394	1616.90	37.83%
Gujarat	1991	934	61.29	31.8	211	58.37	65.51%
Gujarat	2001	920	69.14	23	258	550.36	62.64%
Gujarat	2011	919	78.03	16.6	308	1619.37	57.40%
Haryana	1991	865	55.85	24.1	372	11.57	75.37%
Haryana	2001	861	67.91	20.1	478	139.38	71.08%
Haryana	2011	879	75.55	11.2	573	343.79	65.12%
Himachal Pradesh	1991	976	63.86	22.9	93	13.90	91.32%
Himachal Pradesh	2001	968	76.48	9.5	109	141.85	90.19%
Himachal Pradesh	2011	972	82.8	8.1	123	341.57	89.96%
Jammu & Kashmir	1991	896	42.23	13.2	77	100	77.11%
Jammu & Kashmir	2001	892	55.52	9.4	100	285.00	75.19%
Jammu & Kashmir	2011	889	67.16	10.4	124	726.60	72.63%
Jharkhand	1991	922	41.39	45.3	274	102.70	78.75%
Jharkhand	2001	941	53.56	39.1	338	927.88	77.76%
Jharkhand	2011	948	66.41	37	414	2483.54	75.95%
Karnataka	1991	960	56.04	33.4	235	53.65	69.08%
Karnataka	2001	965	66.6	23.6	276	660.52	66.01%
Karnataka	2011	973	75.37	20.9	319	1854.34	61.33%
Kerala	1991	1036	89.81	19.7	749	103.06	73.60%
Kerala	2001	1058	90.86	12	820	715.25	74.04%
Kerala	2011	1084	94	7.1	860	1125.40	52.30%

contd.

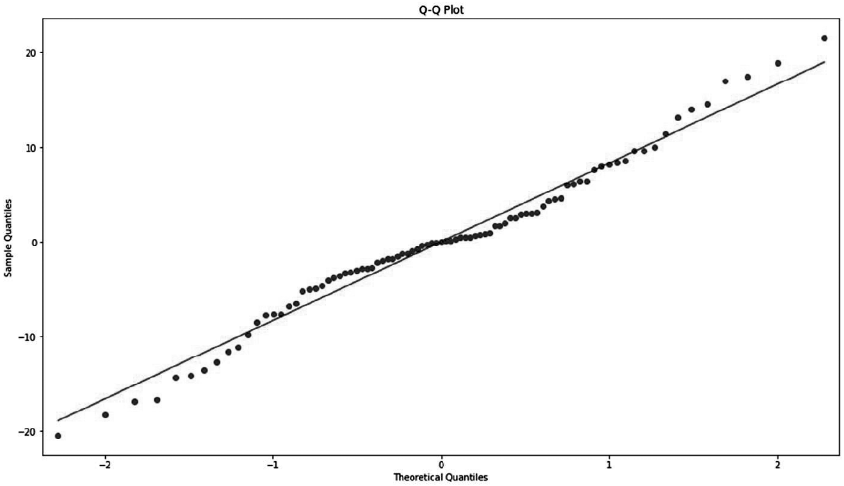
State/Union Territory	*Year*	*Sex Ratio (females per 1000 males)*	*Literacy Rate (percent)*	*Poverty Rate (percent)*	*Population Density*	*GDP (INR crore)*	*Rural Population (Percent)*
Madhya Pradesh	1991	912	44.67	48.6	158	100	74.73%
Madhya Pradesh	2001	919	63.74	36.7	196	255.03	73.54%
Madhya Pradesh	2011	931	69.32	31.7	236	755.70	72.37%
Maharashtra	1991	934	64.87	38.1	257	271.39	61.31%
Maharashtra	2001	922	76.88	24.5	315	2179.63	57.57%
Maharashtra	2011	929	82.34	17.4	365	4231.20	54.78%
Manipur	1991	958	59.89	38	82	3.43	72.51%
Manipur	2001	978	70.5	47.1	97	29.37	74.89%
Manipur	2011	992	76.9	36.9	115	68.68	60.78%
Meghalaya	1991	955	49.1	16.1	79	3.19	81.41%
Meghalaya	2001	972	62.56	17.1	103	36.51	80.42%
Meghalaya	2011	989	74.43	11.9	132	102.77	79.91%
Mizoram	1991	921	82.26	15.3	33	10	53.91%
Mizoram	2001	935	88.8	21.1	42	15.55	50.39%
Mizoram	2011	976	91.33	20.4	52	20.22	47.86%
Nagaland	1991	886	61.65	9	73	2.48	82.73%
Nagaland	2001	900	66.59	20.9	120	34.02	82.76%
Nagaland	2011	931	79.6	18.9	119	82.84	71.15%
Odisha	1991	971	49.09	57.2	203	48.97	86.62%
Odisha	2001	972	63.08	37	236	396.62	85.01%
Odisha	2011	979	72.89	32.6	270	1135.87	83.32%
Puducherry	1991	979	74.74	14.1	1683	10	36.01%
Puducherry	2001	1001	81.24	1.2	1989	35.29	33.47%
Puducherry	2011	1037	85.85	9.7	2547	102.79	31.65%
Punjab	1991	882	58.51	20.9	403	78.45	70.45%
Punjab	2001	876	69.65	15.9	484	639.95	66.08%
Punjab	2011	895	75.84	8.3	551	1380.61	62.52%

contd.

State/Union Territory	*Year*	*Sex Ratio (females per 1000 males)*	*Literacy Rate (percent)*	*Poverty Rate (percent)*	*Population Density*	*GDP (INR crore)*	*Rural Population (Percent)*
Rajasthan	1991	910	38.55	34.4	129	78.23	77.12%
Rajasthan	2001	921	60.41	24.8	165	799.36	76.62%
Rajasthan	2011	928	66.11	14.7	200	958.89	75.13%
Sikkim	1991	878	56.94	31.1	57	1.44	90.89%
Sikkim	2001	875	68.81	13.1	76	8.81	88.91%
Sikkim	2011	890	81.42	8.2	86	13.92	74.80%
Tamil Nadu	1991	974	62.66	28.9	429	127.55	65.85%
Tamil Nadu	2001	987	73.45	17.1	480	1239.01	55.96%
Tamil Nadu	2011	996	80.09	11.3	555	3822.29	51.60%
Tripura	1991	945	60.44	40.6	263	4.59	84.69%
Tripura	2001	948	73.19	17.4	305	54.33	82.93%
Tripura	2011	960	87.22	14.1	350	147.14	73.82%
Uttar Pradesh	1991	876	40.71	32.7	548	228.73	80.33%
Uttar Pradesh	2001	898	56.27	18	690	1629.26	79.22%
Uttar Pradesh	2011	912	67.68	11.3	829	3671.85	77.73%
Uttarakhand	1991	936	57.75	40.9	133	100.00	76.83%
Uttarakhand	2001	962	71.62	37.7	159	131.79	74.33%
Uttarakhand	2011	963	78.82	29.4	189	479.83	69.77%
West Bengal	1991	917	57.7	34.3	767	155.90	72.52%
West Bengal	2001	934	68.64	26.7	903	1390.57	72.03%
West Bengal	2011	950	76.26	20	1028	3080.18	68.13%

Diagnostics

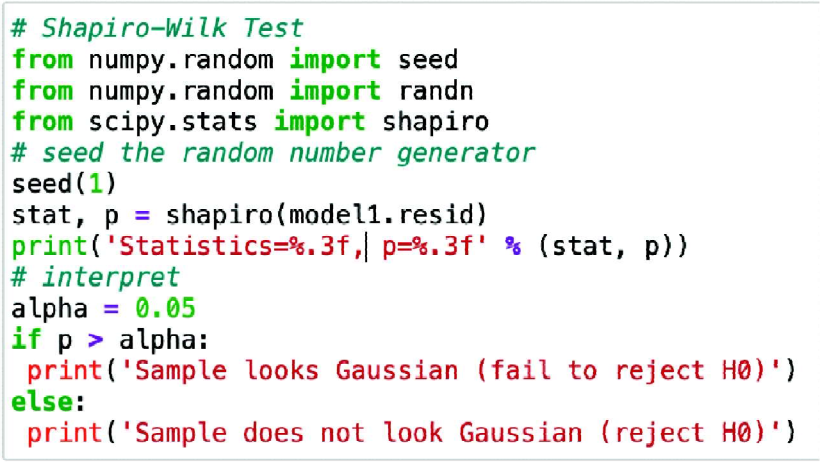
Checking for normality with a Q-Q plot:



The Q-Q plots show that some extreme values are present in Y and X. However, this is not likely to influence my model much. I checked further with a Shapiro-Wilk test.

Shapiro-Wilk Test:

The following code was used on Python to generate results for our model.



```
# Shapiro-Wilk Test
from numpy.random import seed
from numpy.random import randn
from scipy.stats import shapiro
# seed the random number generator
seed(1)
stat, p = shapiro(model1.resid)
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
 print('Sample looks Gaussian (fail to reject H0)')
else:
 print('Sample does not look Gaussian (reject H0)')
```

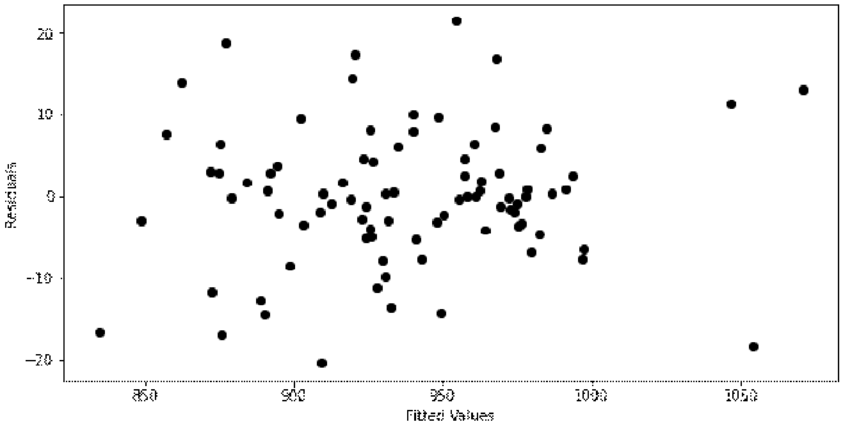
The output for the above code was:

Statistics=0.983, p=0.316

Sample looks Gaussian (fail to reject H0)
Hence, my assumption for normality was met.

Non-Linearity and Non-Constant Variance:

I had plotted residuals vs fitted values (Tukey-Anscombe Plot) to check for the above two assumptions.



The above plot looks satisfactory.

Collinearity:

The test for collinearity was done by calculating the Variance Inflation Factor (VIF) as discussed above in the paper. No values above 5 were found.

VIF (Rural Population) = 1/ (1-0.736) = 3.7878

VIF(Literacy Rate) = 1/(1-0.792) = 4.8077

VIF(Poverty Rate) = 1/(1-0.556) = 2.2522