Impact of Female Fertility Rate on the Pariticipation of Women in Total Workforce

Md Muhtasim Billah

Abstract

Women roughly occupy half of the population of a country, whether it's a developed, developing or an underdeveloped one. But when it's about the total workforce of a country, the male and female percentages are rarely similar. This is even more prominent for the developing and underdeveloped countries. While several reasons such as the insufficient access to education, religious superstitions, lack of adequate infrastrucutres are responsible for this discrepancy, it goes way beyond these. One significant factor is the fertility rate of women which is a count for the total number of births per an individual woman. And to show its effects on the participation of women in the total workforce, percentage of female workers in the labor force has been considered. Fitting the data from two different surveys in a linear regression model, it has been found that there lies an inverse linear relationship between these two factors and with decreasing fertility rate, the percentage of female participation in the total labor force for a country increases remarkably.

Keywords: fertility rate, labor force percentage, regression, Bangladesh.

1. Introduction

Bangladesh is one of the third world Asian countries that has recently been declaired as a developing country because of its increasing GDP in the recent years.[1] Along with noticeable economic advancements in multiple sectors, parallel acievements in trade, agriculture and education has been promising. While the government keeps focusing more and more on ensuring primary, secondary and tertiary education for its entire population, decades-old religious dogma, societal superstitions and lack of access to proper education in the rural areas are proving to be major roadblocks towards the ultimate goal. Tackling one issue leads to several other conjoint issues which makes the problem even harder to solve. The Bangladesh goverment has increased its budget for education to a great extent over the last decade and as a result, the number of school-going children in the rural areas has increased and the number of dropouts has been reduced. Only making primary education more accessible to people from poor households has increasd the total literacy rate as well as the literacy rate among female. This also paved the way for secondary, tertiary and higher education for them.

But this achievement in the education sector is not being well reflected when it comes to the total workforce scenario of the country. Though literacy percentage of female has increased by 10% in the last decade [2], the percentage of female workers in the total workforce has not improved accordingly. One viable reason is the increase in dropout rates between different education levels (primary, secondary, tertiary). Though education is the most significant factor for a highly skilled workforce, it might not be very true for jobs that do not require education beyond a primary level. Specifically for such labor criterion, one important factor come in play for women which is the fertility rate or the total number of children per individual female. Many rural areas which are far from the light of proper education, bearing a child and being a caregiver to one's family is the primary expectation from women. Religious superstition is also a prominent factor that abstains these women from going beyond there households and contributing towards the labor force of the entire population. Women are bound to maitain their household which, very often, consists of not only several children but also other family members. The higher the number of children, more difficult it is to find a scope to contribute to the family economy.

A positive thing is, the fertility rate is seen to be gradually decreasing in the recent decades found from the data collected by UNESCO. Where it is arguable what factors lie responsible for this decreasing trend, its impact on other areas are certain. While its positive effects can be studied on many other factors, the one chosen for this study is the change of percentage of female workers in the toal labor force. It's only obvious that this percentage value will be dependent on multiple other parameters, but for this specific project, the point of interest will be its relation to the fertility rate change.

2. Dataset

For the current study, the data set has been chosen from a survey performed on the population of Bangladesh. Banagladesh has a total population of 163.05 million as of 2019, of which 49.4% are female and 51.6% are male. [3] The current labor force (of age 15+) in total is 95.71 million which is 58.74% of the total population. [4] Of this total workforce, 44.5% are women (as of 2019). [5] This percentage has significantly increased from 1995 when it used to be only 28.33%. The fertility rate on the same year was 3.71 per woman which, as of 2019, has been decreased to 2.028. [6] Here, the fractional value, though doesn't have any physical meaning, is an estimate for the average number of child born to a female individual.

The datasets selected for this study span over 23 years (from 1995 to 2017). Data has been collected separately from two separate secondary datasets from the World Bank databank for both the fertility rate and the percentage of female in the total workforce. Just by looking at the datasets, it is visible that the fertility rate has reduced to 2.028 in 2017 from 3.71 in 1995 and the percentage of female workers in the total workforce has been increased to 44.5% in 2017 from 28.33% in 1995. These two datasets were compiled into one primary dataset and it corresponds to the 23 data points for these two variables. Statiscal linear model will be applied on this primary dataset to analye their relation.

3. Methods

Since, there is only explanatory variable against the response, simple linear regression (SLR) model has been chosen as the statistical method for analyzing this dataset. The model can be expressed in mathematical terms as below.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where, β_0 is the intercept of the y-axis and β_1 is the slope of the fitted line, the estimates for which are given by following expressions.

$$\hat{\beta_1} = \frac{\sum(y-\bar{y})(x-\bar{x})}{\sum(x-\bar{x})^2}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

Y is the reponse varible, X is the explanatory variable or predictor and ε is the error. For our study, the response variable Y is the percentage of female workers in the total workforce and the explanatory variable is the fertility rate which is the number of average children per individual woman. The dataset was loaded in Rstudio for performing the simple linear regression.

4. Results and Discussion

4.1 Descriptive Analysis

For a first glance at the dataset, a scatter plot is generated in R which is given in Figure 1.

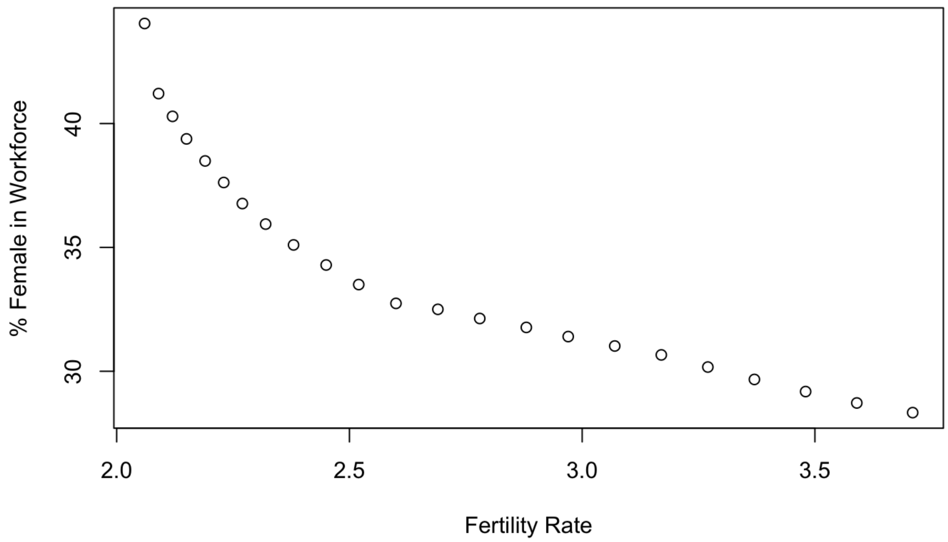

Figure 1: Scatter plot for female percentage in the total workforce vs the fertility rate.

From Figure 1, it is apparent that the data points follow a linear trend. This indicates that with decreasing fertility rate, the percentage of female workers in the workforce increases. And within the fertility range of 2.5 to 2.0, the % values increase even faster. This is understandable because, with the addition of one child, the workload and responsibilities in a household increases significantly. The data indicates that, up to 2 children per household, the female caregiver can find adequate time to set foot out of her household and find time to contribute to the national labor force. But if the number child goes beyond 3, the possibility of that happening is significantly compromised. Before 2000, when the fertility rate was higher, the participation of women in the total workforce was below 30%. But after that, upto a fertility rate of 2.5, the participation increases almost linearly. And when it goes below 2.5, the increase in female participation in workforce is remarkable.

4.2 Model Parameters Estimation

Now, for a more quantitative measure of how this data behaves if fit into a linear model, a simple linear regression is perfromed over this dataset. The regression line is given in Figure 2. From the figure, it is clear that the regression line has a negative slope, β_1 and a positive intercept, β_0. The line doesn't fit the data points very well becuase there is not much linearity among the data points. The summary of the simple linear regression model is given in Table 1.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept, β_0	54.9186	1.9550	28.09	< 2e-16
Slope, β_1	-7.6685	0.7084	-10.82	4.75e-10

Table 1: Summary of the coefficients of the simple linear regression model fitted on to the dataset.

From Table 1, it is found that the values for the parameters are $\beta_0 = 54.9186$ and $\beta_1 = -7.6685$ with an estimated error of 1.9550 and $ 0.7084$ respectively. With 95% confidence, the confidence intervals for β_0 is $(50.85, 58.98)$ and for β_1 is $(-9.14, -6.20)$. The value of β_0 indicates that for a fertility rate of 0, the percentage of female labor force will go as high as 54.92%. Though it is not seen from the regression line fitted on the scatter plot, it would be a good estimate if the fertility rate really goes to 0 whether it has a physical meaning or not. From the fitted model, we can predict the female participation parcentage for any given fertility rate since both are continuous variables. For example, for $X = 2.0$, the fitted value of response variable would be $Y = 39.58$ and for $X = 3.2$, we will get $Y = 30.38$. The cofindence interval for the fitted line is given in the Figure 3. The value of the slope of, $\beta_1 = -7.67$ means that for a unit increase in fertility rate, the female percentage will go down 7.67%. This negative slope exhibits the negative relation among the explanatory and response variables.

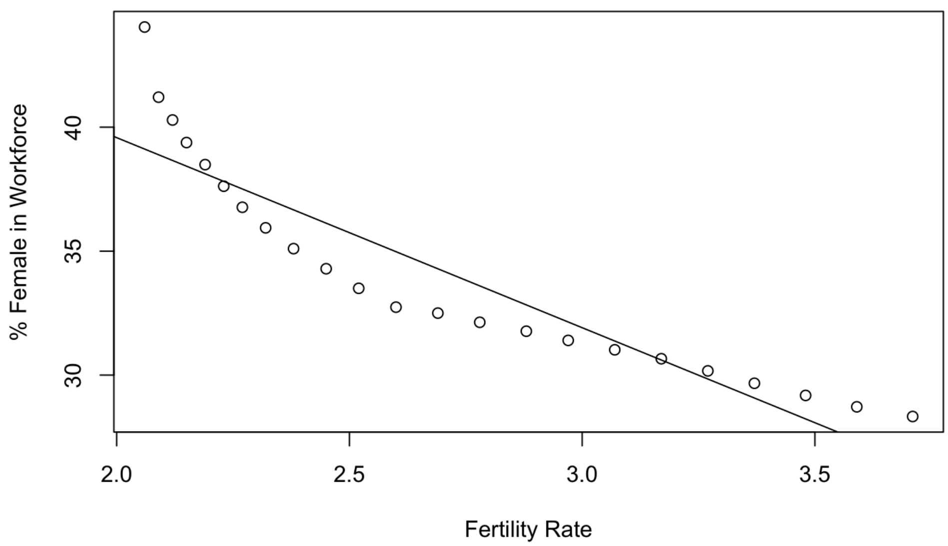

Figure 2: Regression line fitted in the scatter plot.

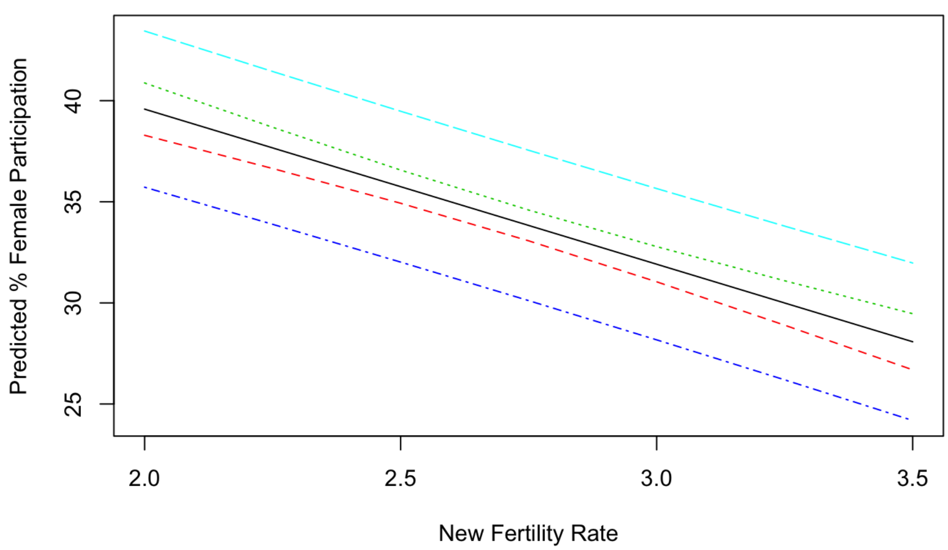

Figure 3: Prediction and Cofidence Intervals.

The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. For this dataset, the t-statistic values (Table 1) are relatively far away from zero and are large relative to the standard error, which indicates a relationship between the fertility rate and the female participation exists. These t-values are also used to compute p-values. The $Pr\ (> t)$ relates to the probability of observing any value equal or larger than t. A small p-value indicates that it is unlikely that we will observe a relationship between the predictor (fertility) and response (female percentage) variables only by chance. In our model, the p-values are extremely small and the null can be rejected for the given X and Y.

Parameters	Values	DF
Residual standard error	1.75	21
Multiple R-squared	0.848	
Adjusted R-squared	0.8408	
F-statistic	117.2	(1,21)
p-value	4.751e-10	

Table 2: Other parameters of the SLR model fitted on to the dataset.

Some other important parameters of the model are given in Table 2. The residual standard error is a measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term ε. Due to the presence of this error term, we are not capable of perfectly predicting our response variable from the predictor. The residual standard error is the average amount that the response will deviate from the true regression line. In our case, the female participation rate can deviate from the true regression line by approximately 1.75% on average. In other words, given that the mean participation of female is 54.92% and that the residual standard error is 1.75%, we can say that the percentage error is 3.57%. It's also worth noting that the residual standard error was calculated with 21 degrees of freedom. Simplistically, degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters (restriction). In our case, we had 23 data points and two parameters (intercept and slope).

The R-squared (R^2) statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. R^2 is a measure of the linear relationship between our predictor variable and our response/target variable. It always lies between 0 and 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). For our dataset, the R^2 we get is 0.848. Or roughly 85% of the variance found in the response variable (female participation) can be explained by the predictor variable (fertility rate).

F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H_0 : There is no relationship between X and Y). The reverse is true as if the number of data points is small, a large F-statistic is required to be able to ascertain that there may be a relationship between predictor and response variables. In our example the F-statistic is 117.2 on 1 and 21 DF which is very large and gives a p-value of $4.751e^{-10}$. So, the null is rejected and it indicates that there is strong relation between fertility rate and female participation percentage.

4.3 Model Diagnostics

4.3.1 Graphical Measures

It is apparent from the QQ plot (left) in Figure 4 that the data point tends to deviate at the upper tail which indicates non-normality of the dataset. From the residuals vs fitted plot (right) is is visible that the data have a pattern rather than a random distribution which indicates a non constant variance.

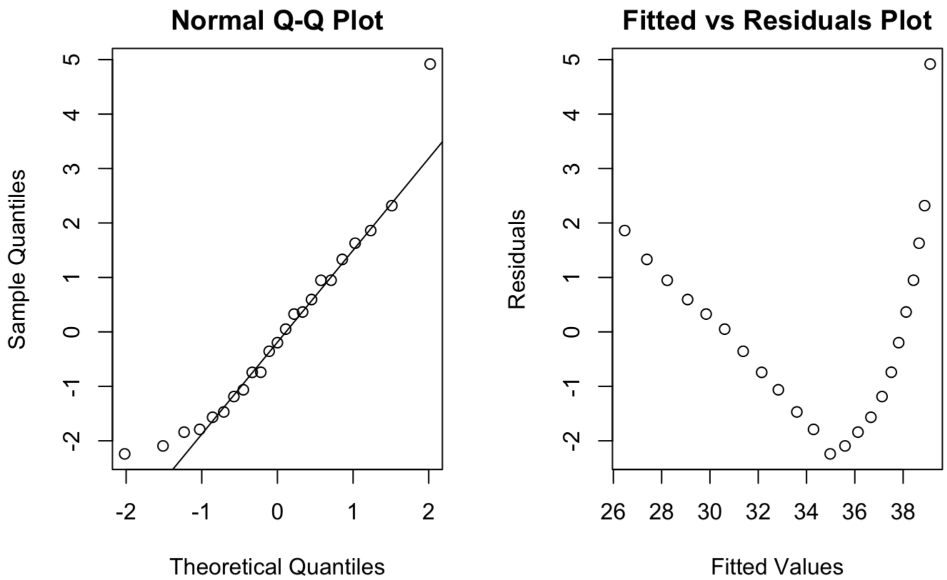

Figure 4: Diagnostic plots. QQ plots for checking normality (left) and residuals vs fitted plot (right).

4.3.2 Diagnostic Tests

To get a better insight about the assumptions, a test for each has been performed and tabulated in Table 3. Though from the QQ plot the normality of the residuals was not confirmed, the Shapiro-Wilk test confirms it. And though the residuals vs fitted plot showed a pattern, the Breusch-Pagan test confirms that the residuals have constant variance. The Durbin Watson statistic found from the test is very far from 2 which indicates a high correlations.

Test Name	Assumptions	Statistic	P-value	Decision
Shaphiro-Wilk	Normality of the residuals	W = 0.93043	0.1117	Normal Residuals
Breusch-Pagan	Homogeneity of residual variance	BP = 2.4423	0.1181	Constant variance
Durbin-Watson	Autocorrelation of the residuals	DW = 0.16715	2.011e-13	Non-zero Autocorrelation

Table 3: Diagnostic tests for checking the assumptions.

4.4 Data Transformation

To see if the model can be a better fit with transformed data, we can perform data terformation. To find the transformation suitable for the current data, BoxCox transformation has been done which is given in Figure 5. The value for λ was found to be closest to -6.

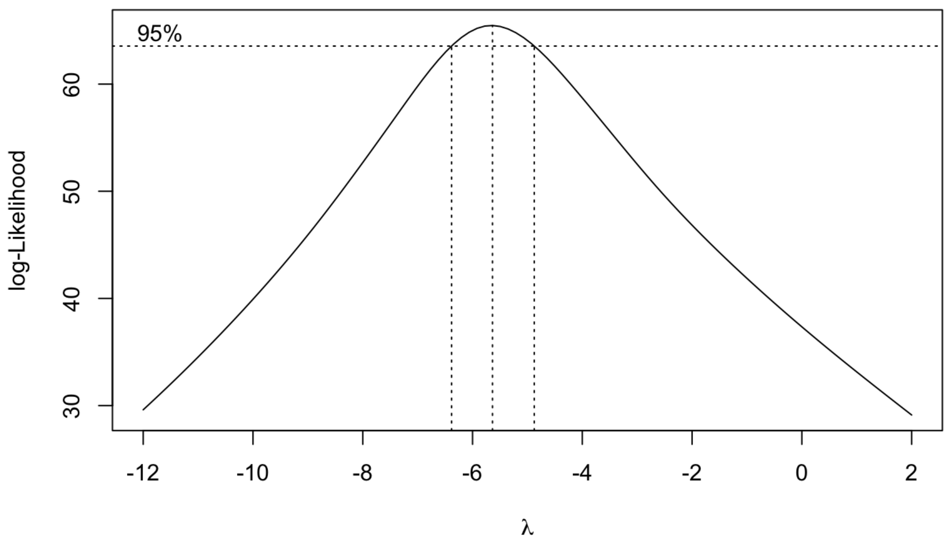

Figure 5: BoxCox transformation for the datasets.

The transformed data was seen to give a slightly better fit for the model. The standard residual error was reduced and the QQ and residual plots were better fit for the assumptions which can also be seen from the increased significane of the tests. But since the final outcome is not influnced by a great measure, the results are not provided here.

5. Conclusion

The current study was primarily focused on finding the effect of fertility rate of women on the percentage of women participating in the national workforce of Bangladesh. It was found that the fertility rate has significant effect on the female worker percentage of the labor force. But it is only obvious that female percentage depends on several other factors. So, the results found here is only the marginal effect of fertility rate on female percentage given that other situations remain constant. If more predictors are taken into account, then more insights about the factors that control the female participation could be achieved. The provided diagnostics measures of the fitted linear model confirm the general aspect of the model. However, inclusion of more relevant predictors is necessary in order to study the female participation percentage variability more precisely.

Reference

[1] World Bank national accounts data, and OECD National Accounts data files. ID: NY.GDP.MKTP.KD.ZG
[2] UNESCO Institute for Statistics. ID: SE.ADT.1524.LT.FE.ZS
[3] United Nations Population Division. World Population Prospects: 2019 Revision. ID: SP.POP.TOTL
[4] ILOSTAT database and World Bank population estimates, September 2019. ID: SL.TLF.TOTL.IN
[5] ILOSTAT database and World Bank population estimates, September 2019. ID: SL.TLF.TOTL.FE.ZS
[6] United Nations Population Division. World Population Prospects: 2019 Revision. ID: SP.DYN.TFRT.IN

Appendix

R code used for analyzing the data.

```r
#data setup
setwd("/Users/muhtasim/Desktop/STAT530/Project1")
mydata=read.csv("SLR2.csv",header=T)
rate=mydata[,1]^(-6) #add the power only for data transformation
worker=mydata[,2]
#scatter plot
plot(rate,worker,ylab="% Female in Workforce",xlab="Fertility Rate")
#SLR
model=lm(worker~rate,data=mydata)
abline(model)
#model parameters
s=summary(model)
coefficients(model)
cf=confint(model,level=0.95)
#prediction intervals
y=worker
x=rate
predict(lm(y ~ x))
#creating new values for prediction
new <- data.frame(x = seq(2.0, 3.5, 0.25))
predict(lm(y ~ x), new, se.fit = TRUE)
pred.w.plim <- predict(lm(y ~ x), new, interval = "prediction")
pred.w.clim <- predict(lm(y ~ x), new, interval = "confidence")
#prediction plots
matplot(new$x, cbind(pred.w.clim, pred.w.plim[,-1]),  type = "l",
        ylab = "Predicted % Female Participation", xlab = "New Fertility Rate")
#Diagnostic plots
#plot of residual vs x and fitted values
par(mfrow=c(1,2))
plot(rate, model$res)
plot(model$fitted, model$res)
# Normal probability plot of Residuals
qqnorm(model$res)
qqline(model$res)
hist(model$res)
#Diagnostic tests
#test for normality
library(stats)
shapiro.test(model$res)
#breusch pagan test
#might need to install lmtest
library(lmtest)
bptest(model)
#durbin-watson test
dwtest(model)
#data transformation
library(MASS)
boxcox(worker ~ rate, data = mydata,
       lambda = seq(-12.0, 2.0, length = 10))
```