A Study for Determining the Impacts of Socioeconomic Factors on the Female Employment Rate

Md Muhtasim Billah

Abstract

Women, occupuying roughly half of the population of a country, play cruicial role in every sector. Despite being equally important for advancement of a demographic, their participation in the workforce has been overlooked for ages. Though developed countries are acknowledging this more and more, the underdeveloped and developing countries are still struggling in this regard. Nonetheless, the positive thing is that the percentage of existing women population who directly engage in the workforce, though not utterly satisfactory, is gradually increasing for which many socio-economic factors seem to be responsible. To address the effects of these factors on the percentage of working women, data that were collected individually on each of these factors by the World Bank, have been compiled into one primary dataset. Analyzing this dataset using multiple linear regression method, it was found that these socioeconomic factors significantly affect the percentage of working women and putting more emphasis on these, the participation of women in the total workforce can be enhanced remarkably.

Keywords: working women percentage, female employment, 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 were 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 goverment of Bangladesh 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.

Driven by this positive trend in the education sector, advancements in multilple other sectors are being observed. The Bangladeshi women have been participating more and more in the workforce by engaging themselves in every sector such as industry, economy, and services etc. This has also given rise to the number of female employers and entrepreneurs. While multiple socioecomonic factors may be held responsible, it is hard to narrow those down that sirectly influence this rise in female employment rate. In Bangladesh, the percentage of the total women pouplation who directly participate in the total workforce has increased by almost 10% in the last two decades. It remains questionable whether this increase is satisfactory or not but if the factors contributing to this increase can be determined, it would be possible to focus more on those to make further improvements. 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 selected predictors.

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 (80.55 million) and 51.6% are male (82.5 million). [3] Of these 80.55 million women, 67.38% are between the age of 15 to 64 who are considered to be the working age women. But only 33.44% are engaged in the workforce who are of age 15 or above [4] which indicates that only half of the available female workforce are currently being utilized and it used to be worse two decades back when it used to be only 24.30% in 1995.

The datasets selected for this study span over 25 years (from 1995 to 2019). Data has been collected separately from multiple secondary datasets from the World Bank databank for the employed women percentage and the related predictor variables. These datasets were compiled into one primary dataset and it corresponds to the 25 data points for the variables. There is one response variable which is the percentage of the employed women and 10 exlnanatory variables of predictors. Brief descriptions of these variables are given below.

PerFemEmploy

Employment to population ratio (%) of women who are of age 15 or older. Employment to population ratio is the proportion of a country's population that is employed. Employment is defined as persons of working age who, during a short reference period, were engaged in any activity to produce goods or provide services for pay or profit, whether at work during the reference period (i.e. who worked in a job for at least one hour) or not at work due to temporary absence from a job, or to working-time arrangements. Ages 15 and older are generally considered the working-age population.

FertilityRate

Fertility rate (birth per women). Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.

RatioMaletoFemale

Ratio of female to male labor force participation rate. Labor force participation rate is the proportion of the population ages 15 and older that is economically active: all people who supply labor for the production of goods and services during a specified period. Ratio of female to male labor force participation rate is calculated by dividing female labor force participation rate by male labor force participation rate and multiplying by 100.

PerFemEmployers

Employers, female (% of female employment). Employers are those workers who, working on their own account or with one or a few partners, hold the type of jobs defined as a "self-employment jobs" i.e. jobs where the remuneration is directly dependent upon the profits derived from the goods and services produced), and, in this capacity, have engaged, on a continuous basis, one or more persons to work for them as employee(s).

Agriculture

Employment in agriculture, female (% of female employment). Employment is defined as persons of working age who were engaged in any activity to produce goods or provide services for pay or profit, whether at work during the reference period or not at work due to temporary absence from a job, or to working-time arrangement. The agriculture sector consists of activities in agriculture, hunting, forestry and fishing, in accordance with division 1 (ISIC 2) or categories A-B (ISIC 3) or category A (ISIC 4).

Industry

Employment in industry, female (% of female employment). The industry sector consists of mining and quarrying, manufacturing, construction, and public utilities (electricity, gas, and water), in accordance with divisions 2-5 (ISIC 2) or categories C-F (ISIC 3) or categories B-F (ISIC 4).

Services

Employment in services, female (% of female employment). The services sector consists of wholesale and retail trade and restaurants and hotels; transport, storage, and communications; financing, insurance, real estate, and business services; and community, social, and personal services, in accordance with divisions 6-9 (ISIC 2) or categories G-Q (ISIC 3) or categories G-U (ISIC 4).

Wage.Salaried

Wage and salaried workers, female (% of female employment). Wage and salaried workers (employees) are those workers who hold the type of jobs defined as "paid employment jobs," where the incumbents hold explicit (written or oral) or implicit employment contracts that give them a basic remuneration that is not directly dependent upon the revenue of the unit for which they work.

ContrFamWorkers

Contributing family workers, female (% of female employment). Contributing family workers are those workers who hold "self-employment jobs" as own-account workers in a market-oriented establishment operated by a related person living in the same household.

OwnAccount

Own-account female workers (% of employment). Own-account workers are workers who, working on their own account or with one or more partners, hold the types of jobs defined as "self-employment jobs" and have not engaged on a continuous basis any employees to work for them. Own account workers are a subcategory of "self-employed".

Vulnerable

Vulnerable employment, female (% of female employment). Vulnerable employment is contributing family workers and own-account workers as a percentage of total employment.

Based on the available predictors, best model will be selected and analysis will be done on that model. The dataset will also be checked for any required data transformation.

3. Method

Since, there are multiple explanatory variables against the response, multiple linear regression (MLR) model has been chosen as the statistical method for analyzing this dataset. Before jumping to the conclusion, exploratory analysis will be done on the dataset, then the best model will be selected. The dataset will also be checked for any required data transformation which will be done before the actual analysis. The tests for partial slopes will be carried out and finally the diagnostics will be done using both graphical and statistical testing meeasures.

4. Results and Discussions

4.1 Exploratory Analysis

Two missing values were found for the variable "FertilityRate". Those two data points have been omitted and now we have 23 data points for the years 1995 to 2017.

4.2 Model Selection

The full model is as follows,

PerFemEmploy = β_0 + β_1 FertilityRate + β_2 RatioMaletoFemale + β_3 PerFemEmployers + β_4 Agriculture + β_5 Industry+ β_6 Services + β_7 Wage.Salaried + β_8 ContrFamWorkers + β_9 OwnAccount + β_{10} Vulnerable + ϵ_i.

Since the number of predictors is large for this dataset, we choose the stepwise regression method that uses the AIC (Akaike Information Criteria). The method suggested to drop two of the predictors- "FertilityRate" and "PerFemEmployers" and for this project, the suggested model will be used. So, the reduced model is,

PerFemEmploy = β_0 + β_1 RatioMaletoFemale + β_2 Agriculture + β_3 Industry+ β_4 Services + β_5 Wage.Salaried + β_6 ContrFamWorkers + β_7 OwnAccount + β_8 Vulnerable + ϵ_i.

4.3 Data Trasnformation

To see, if any data transformation is required, the BoxCox transformation method is used.

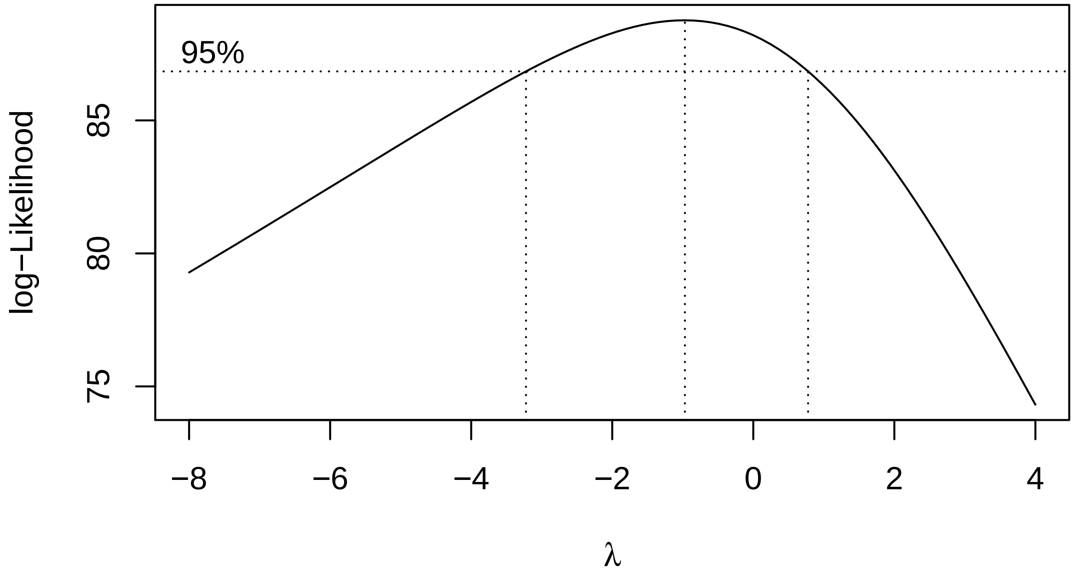

Figure 1: BoxCox transformation of the data.

The value of λ seems to be close to -1. But the data weren't transformed since it didn't improve the results (overall p-value or the adjusted R^2). Also, the data comes from a normal distribution as found from the tests and there is linearity among the response and the predictors. This also indicates that there is no necessity to transform the data.

4.4 Model Parameters Estimation and Testing

For testing the significance of the partial slopes,

The null hypothesis is, $H_0 : \beta_i = 0$.
The alternative hypothesis is, $H_a : \beta_i \neq 0$.
Level of significane, $\alpha = 0.05$.

Multiple linear regression was performed on the dataset and important findings about the coefficients such as their estimates, confidence intervals, standard errors, t-statistics and the p-values are given in Table 1. Based on these attributes, important ideas can be gathered about the predictors and the response such as how they are related and how changing a unit of one can affect the response.

Coefficients	Confidence Interval	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	(-5395.749, 264.184)	-2565.782	1319.463	-1.945	0.072
RatioMaletoFemale	(0.431, 1.198)	0.815	0.179	4.561	<2e-16
Agriculture	(0.991, 57.052)	29.021	13.069	2.221	0.043
Industry	(0.648, 56.464)	28.556	13.012	2.195	0.046
Services	(1.037 57.110)	29.074	13.072	2.224	0.043
Wage.Salaried	(-5.483, -1.184)	-3.333	1.002	-3.326	0.005
ContrFamWorkers	(-49.413, -3.709)	-26.561	10.655	-2.493	0.026
OwnAccount	(-49.495, -3.755)	-26.625	10.663	-2.497	0.026
Vulnerable	(0.130, 46.366)	23.248	10.779	2.157	0.049

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

Based on Table 1, following comments can be made about the model.

Intercept

The intecept is found not statistically significant (p-value = 0.072). This means, there is not enough statistical evidence to show that the coefficient, β_0 differs from zero.

RatioMaletoFemale

RatioMaletoFemale has positive slope and the p-value (<2e-16) is implying that it highly statistically significant. This means, with an increase in the RatioMaletoFemale, there is significant increase in the PerFemEmploy. But from the estimate of the partial slope, it seems that unit change in this predictor value will only slightly affect the response.

Agriculture, Industry and Services

The p-values for Agriculture (0.043), Industry (0.046) and Services (<0.043) are indicating that there is sufficient statistical evidence to conclude that they employ significant effects on the PerFemEmploy and since the coefficients are positive, the effects are also positive. A unit change of these variables will change the response pretty largely.

Wage.Salaried, ContrFamWorkers and OwnAccount

The p-values for Wage.Salaried (0.005), ContrFamWorkers (0.026) and OwnAccount (0.026) are indicating that there is sufficient statistical evidence to conclude that they also have significant effects on the PerFemEmploy and since the coefficients are negative, the effects are also negative.

Vulnerable

It has a postive effect which is statistically significant (p-value = 0.049).

4.5 Model Diagnostics

4.5.1 Graphical Measures

It is apparent from the QQ plot (left) in Figure 2 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. But to have a more rigrous measure, only looking at these plots might not be enough and statistical tests will be required to be performed.

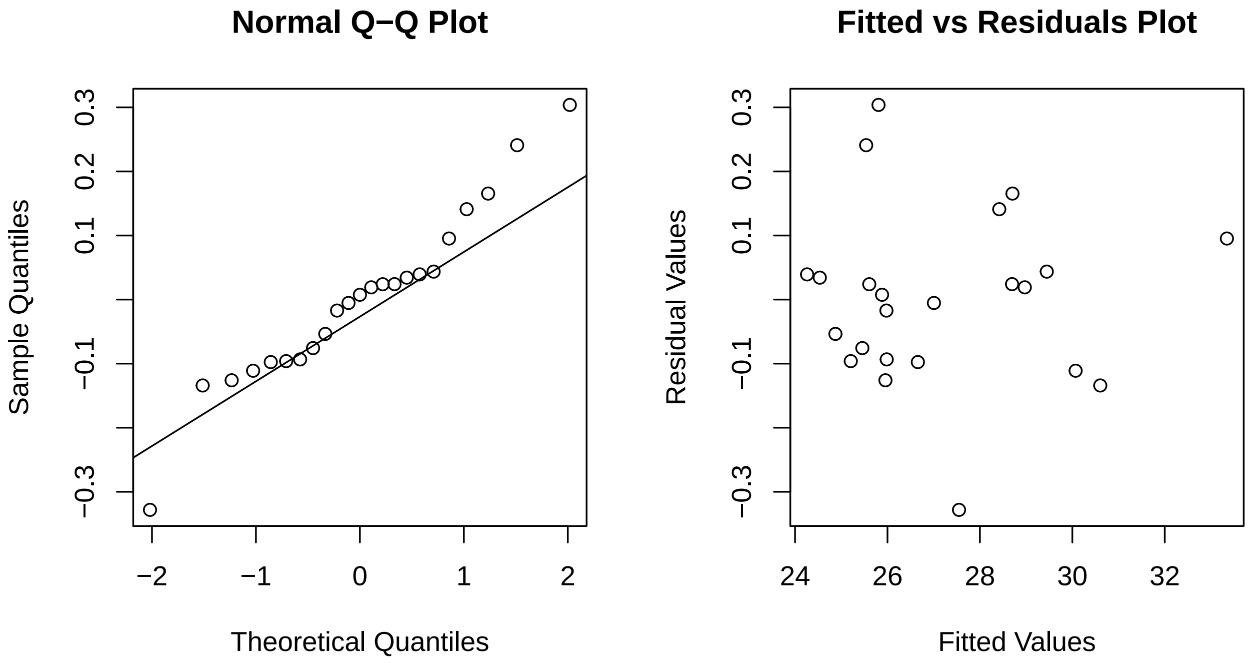

Figure 2: Diagnostic plots. On the left, QQ plots for checking normality of the residuals. On the right, Residuals vs fitted plot for checking the constancy of the variance.

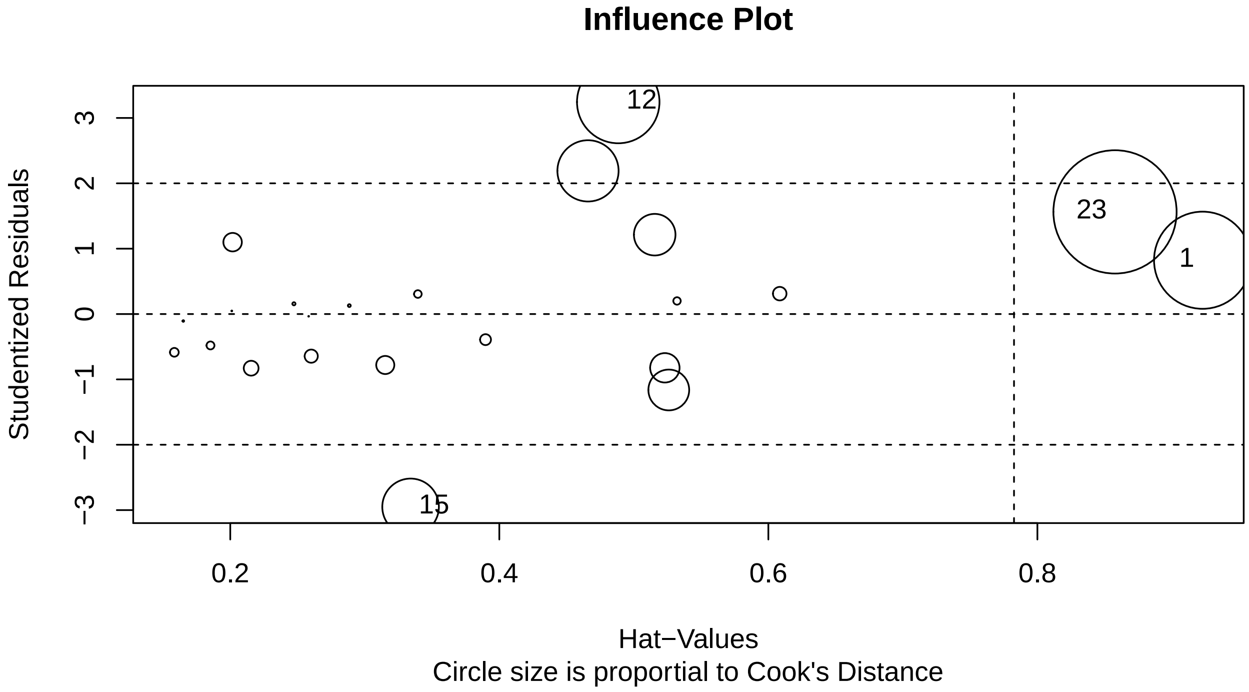

Figure 3: Influence plot is given for identifying potential influential points based on the Cook's distances, hat values and the studentized residuals.

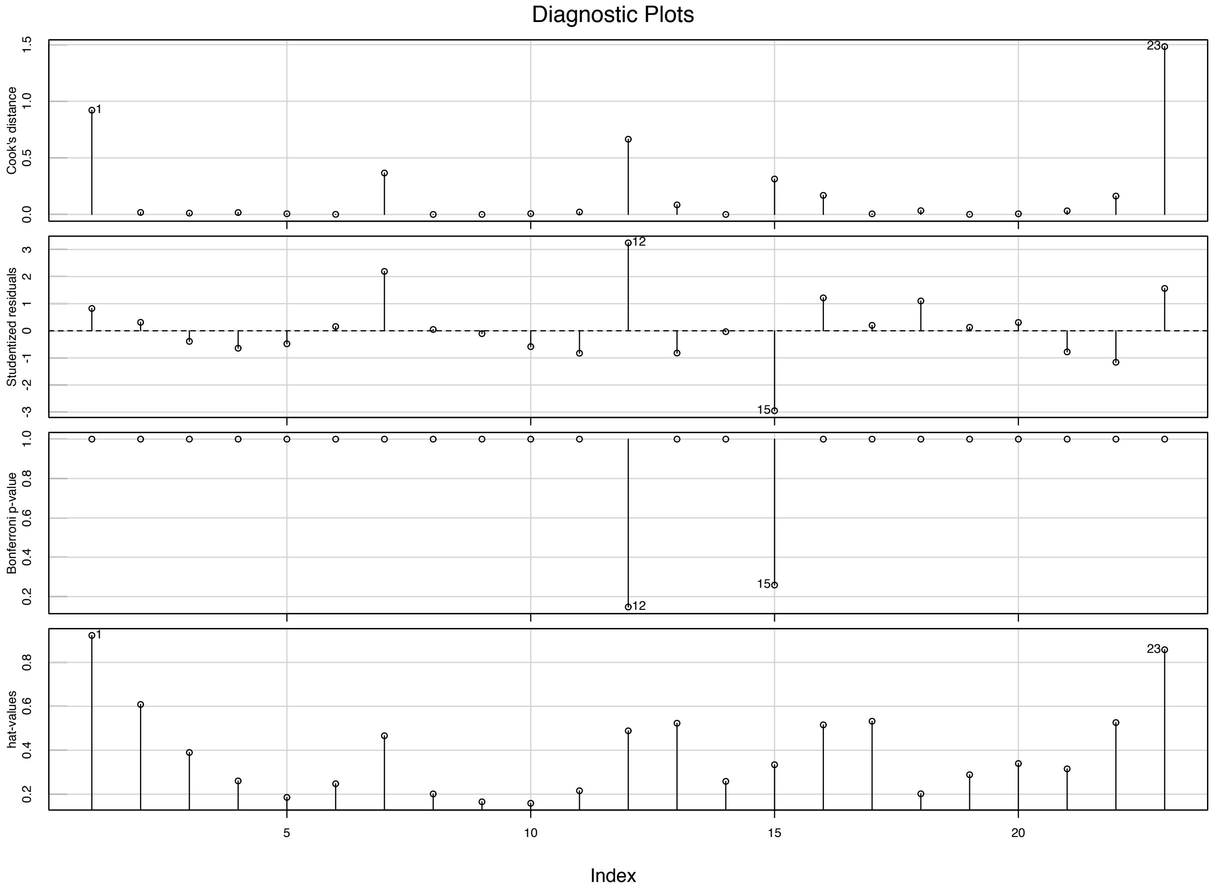

Figure 4: Diagnostic plots for checking outliers and influential points. The Cook's distances, studentized residuals, Bonferroni p-values and hat values are provided for each data point.

From Figure 3, we see that there is a huge outlier in the dataset (observation number 23) which can be an influencial point. For building a better model, observations number 1, 7, 12 and 15 can also be considered. Looking at Figure 4, we see that some of the points (1, 12, 15, 23) are far away from most of the points which also have very high value of studentized residuals as well as Cook's distance. All the potential influential points have been listed in Table 2. They can be considered as the outliers and possible influencial data points. This points can be dropped to fit a better model to this dataset. But, for this projects, since the sample size is not too large, these data points will be kept as omitting them didn't change the outcome to a great extent.

Data Points	Studentized Residuals	Hat Values	Cook's Distance
1	0.8231333	0.9227851	0.9209092
12	3.2442352	0.4884111	0.6644217
15	-2.9511063	0.3340548	0.3130359
23	1.5628230	0.8576862	1.4827616

Table 2: Studentized residuals, hat values and Cook's D for the potential influencial points.

Multicolinearity is another issue that needs to be considered while building regression model. If there is multicolinearity in the dataset, the error variance gets larger. One way to check for multicollinearity is the pairwise correlation test of the predictors which is given in Figure 5.

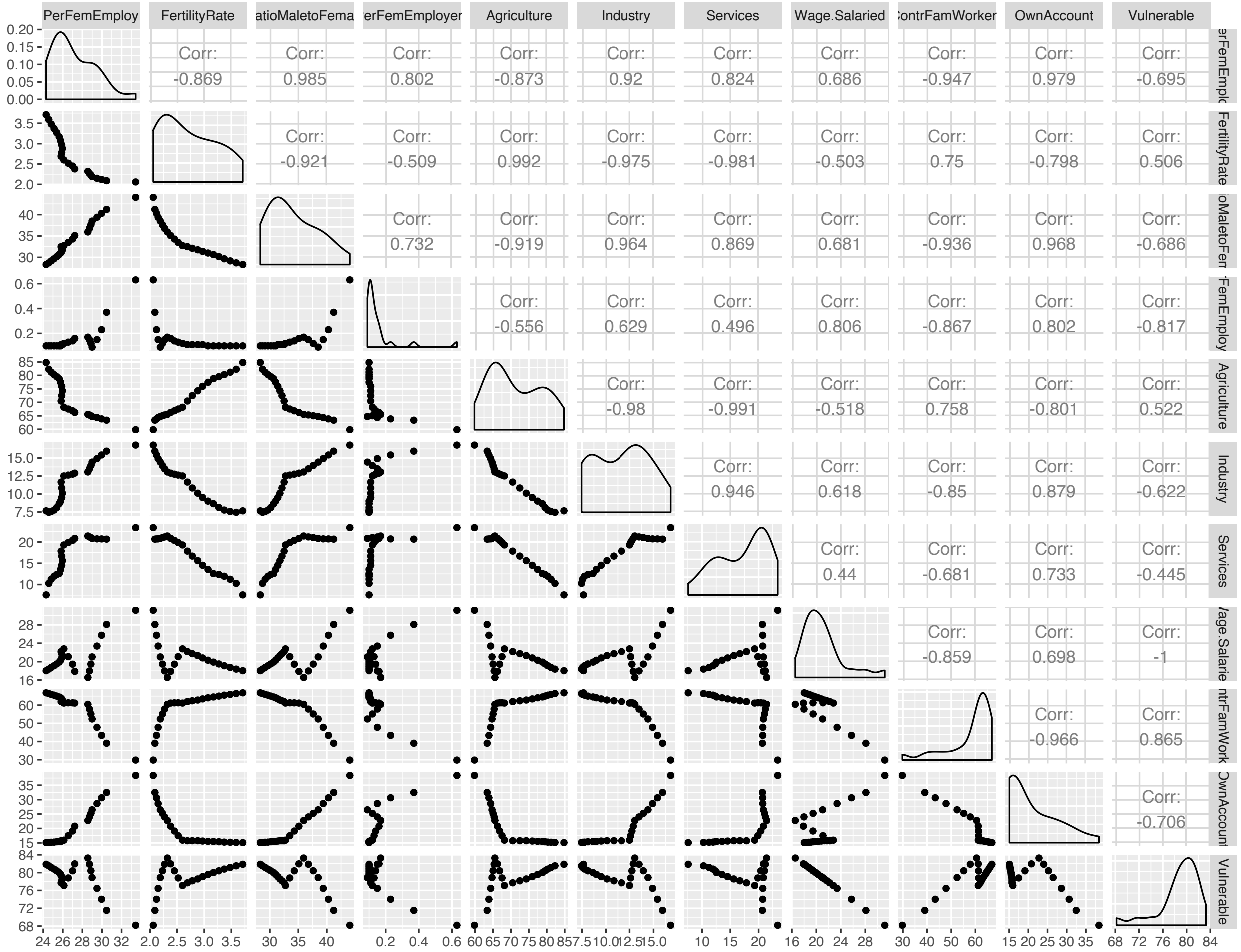

Figure 5: Pairwise correlation tests of the predictors.

It is seen from the pairwise tests that most of the predictors are highly correlated with each other. It can also be checked by looking at the variance inflation factors (VIFs) of the predictors provided in Table 3 which is another formal diagnostic method for checking multicollinearity.

Predictors	VIFs
RatioMaletoFemale	468.489
Agriculture	7076429.422
Industry	1136397.883
Services	2639537.266
Wage.Salaried	9016.619
ContrFamWorkers	8051078.169
OwnAccount	4038657.444
Vulnerable	1102937.897

Table 3: Diagnostic tests for checking the assumptions.

However, since for this project, the regression model is used only for prediction, the presence of multicolinearity can be overlooked.

A linearity of the predictors in terms of their residuals can be checked by plotting the residuals vs the predictors plots which is given in Figure 6. From the plots, it is evident that for all the predictors the residuals are very linear whether they have a positive or negative trend.

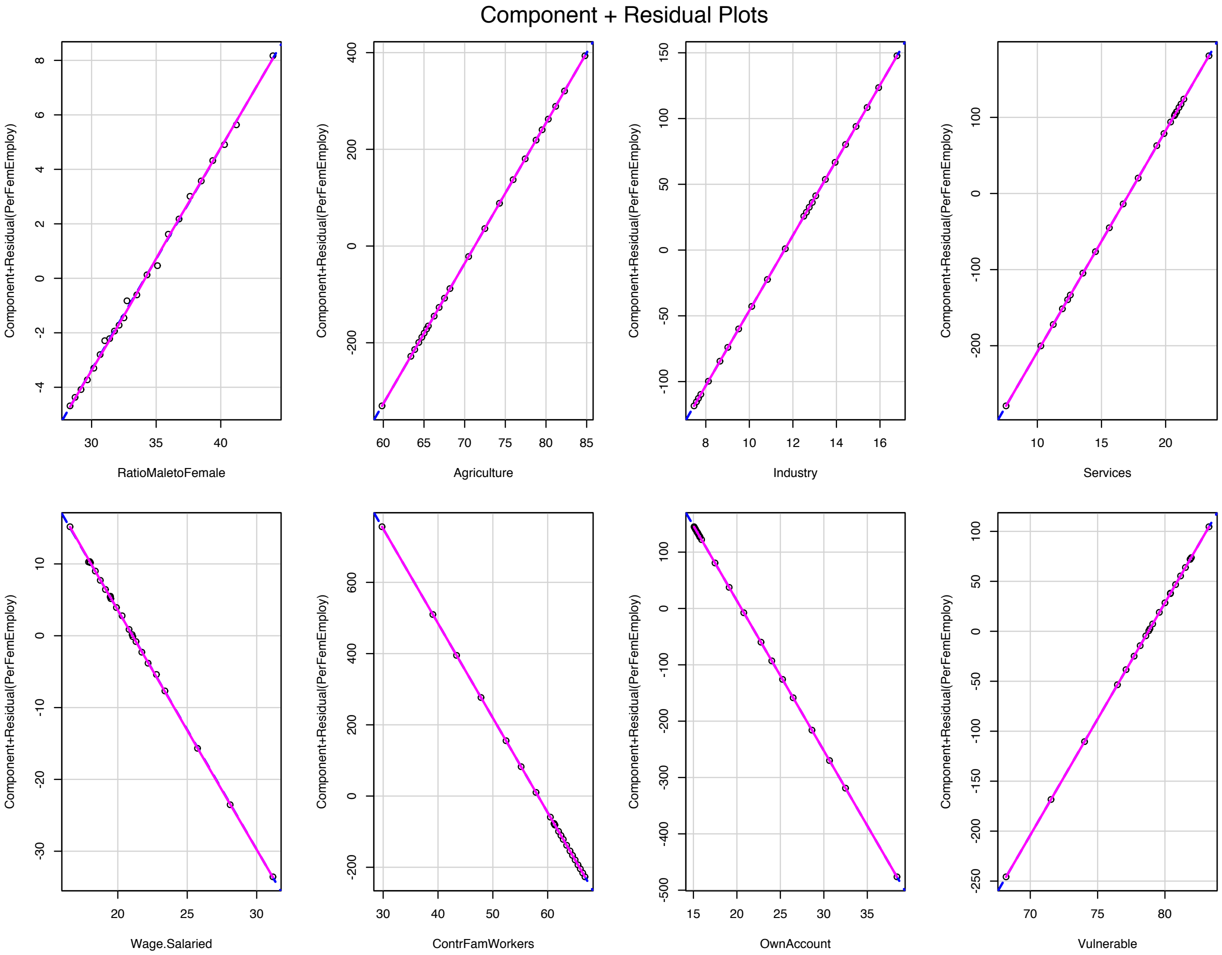

Figure 6: Residuals plots of the predictors.

4.5.2 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. The Breusch-Pagan test also confirms that the residuals have constant variance. The Durbin Watson statistic found from the test is very close to 2 which indicates negligible autocorrelations in the dataset. Since the p-value (0.2663) is less than $\alpha = 0.05$, it also indicates that there is no autocorrelation in the dataset.

Test Name	Assumptions	Statistic	P-value	Decision
Shaphiro-Wilk	Normality of the residuals	W = 0.95996	0.4625	Normal Residuals
Breusch-Pagan	Homogeneity of residual variance	BP = 11.024	0.2003	Constant variance
Durbin-Watson	Autocorrelation of the residuals	DW = 2.3327	0.2663	No Autocorrelation

Table 3: Diagnostic tests for checking the assumptions.

4.6 Model Evaluation

From the Table 4, looking the p-value of the fitted model (8.11e-16), it can be said that it is statistically significant with an $\alpha = 0.05$. It means that the overall model is significant to study the functional relationship between the PerFemEmploy and the predictor variables. Also, from the high R-squared (0.9965) and Adjusted R-squared value (0.9945), it can be concluded that a high percentage of variance (~99.5%) was explained by this model which also indicates the model's goodness of fit.

Attributes	Values	DF
Residual standard error	0.1697	14
Multiple R-squared	0.9965	
Adjusted R-squared	0.9945	
F-statistic	494.1	(8,14)
p-value	8.11e-16	

Table 4: Other attributes of the MLR model fitted on to the dataset.

5. Conclusion

The current study was primarily focused on finding the effect of various socio-economic factors that have influence on the percentage of women participating in the overall national workforce of Bangladesh. Several relavant factors were analyzed based on the Worldbank dataset and it was found that some of these factors significantly affect the female worker percentage of the labor force. But it is only obvious that this percentage depends on multiple other factors. So, the results found here is only the marginal effect of the oredictors under onsideration on the 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.

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 PREPARATION
#setting working directory
setwd("/Users/muhtasim/Desktop/STAT530/Projects/Project2_MLR")
#importing data
mydata=read.csv("MLR2.csv", header=T)
#data summary
summary(mydata)
```

```
dim(mydata)
#checking for NA values
NA_values=data.frame(no_of_na_values=colSums(is.na(mydata)))
#head(NA_values,28)
#removing NA values
ix=apply(mydata,1,function(x) !any(is.na(x))) #for removing any missing data for any x
#new dataset with removed NA values
newdata=mydata[ix,]
#dim of the new data set is now 460, after deleting 4 NA values
dim(newdata)


###APPLYING STATISTICAL METHOD
#fitting the data in a MLR model
fit = lm(PerFemEmploy ~ FertilityRate + RatioMaletoFemale +
           PerFemEmployers + Agriculture + Industry + Services
         + Wage.Salaried + ContrFamWorkers + OwnAccount + Vulnerable, data = newdata)

#Model Selection
#Stepwise Regression using AIC criteria
library(MASS)
slm=stepAIC(fit,direction="both")
slm
summary(slm)
#new model
fit1=lm(PerFemEmploy ~ RatioMaletoFemale
        + Agriculture + Industry + Services
        + Wage.Salaried + ContrFamWorkers + OwnAccount + Vulnerable, data = newdata)
#boxcox transformation
boxcox(PerFemEmploy ~ RatioMaletoFemale
          + Agriculture + Industry + Services
       + Wage.Salaried + ContrFamWorkers + OwnAccount + Vulnerable, data = newdata,
       lambda = seq(-8.0, 4.0, length = 10))
#model parameters estimation
#partial slopes for the predictors
round(fit1$coefficients, 3)
#Test for the significance of the slopes
library(vars)
round(coeftest(fit1), 3)


###MODEL DIAGNOSTICS
##normality test
#Shapiro_Wilk test for normality
shapiro.test(fit1$residuals)
#QQ plot for checking normality
qqnorm(fit1$res)
qqline(fit1$res)
hist(fit1$res)
##constant variance test
#residuals vs fitted plot
plot(fit1$fitted.values, fit1$residuals)
#Breusch-Pagan test
```

```
lmtest::bptest(fit1)
##test for autocorrelation
#Durbin-Watson test
dwtest(fit1)
##linearity tests (residuals vs X's plots)
crPlots(model = fit1, id.n = 5, layout=c(2,4))
##multicollinearity check
#load the package car
library(car)
#paiwise correlation check
#pairwis relation plots
X=newdata[,2:11]
library(GGally)
ggpairs(X)
#variance inflation factors
vif(fit1)
##influencial points check
##potential outlier detection
influenceIndexPlot(model = fit1, id.n = 5)
#Influence Plot
influencePlot(fit1, main="Influence Plot",
              sub="Circle size is proportial to Cook's Distance")
```