A Study on Female Participation in Workforce in World

Sonal G., Sushma Y., Shamima H.

13 August, 2022

Introduction

Half of the world's population roughly comprises of women but when compared to a country's total workforce the male and female workers percentage is rarely similar. If you look at the developing and underdeveloped countries, it's even more prominent. Insufficient access to education, religious superstitions, lack of adequate infrastructures are some of the reasons responsible for this discrepancy, also it goes way beyond these. The total labor force has been considered to show the effects of multiple socioeconomic factors on the women participation in the total workforce ad percentage of female employment. The relationship between these factors can be analyzed using multiple linear regression model.

Problem

Our original data comes from World Bank database where they have a collection of development indicators to estimate various socio-economic factors for all nations in the world. The data was collected using the Data Bank online resource which allows users to form custom time-series data sets based on chosen filters like countries, years and development indicators. We gathered data for each of the 11 development indicators (including the employed women percentage and related predictor variables) across 217 countries for the most recent year, 2019. After performing data preprocessing like removing missing values and labeling our indicators to more simple variable names .etc , we get our final data having 187 data points (countries). There is one response variable which is the percentage of the employed women explanatory variables of predictors. Brief descriptions of these variables are given below.

1. 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.

2. 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.

3. 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.

4. 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).

5. 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).

6. 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).

7. 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.

8. 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.

9. 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".

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

Purpose

We can apply Linear Regression Model and other statistical methods on this dataset to analyze if there are any viable relationship between the response of the variables and the predictor.

Methodology

A. Data Preprocessing

A.1. Labelling Variable Names

We converted the indicator names to more simple and appropriate variable names.

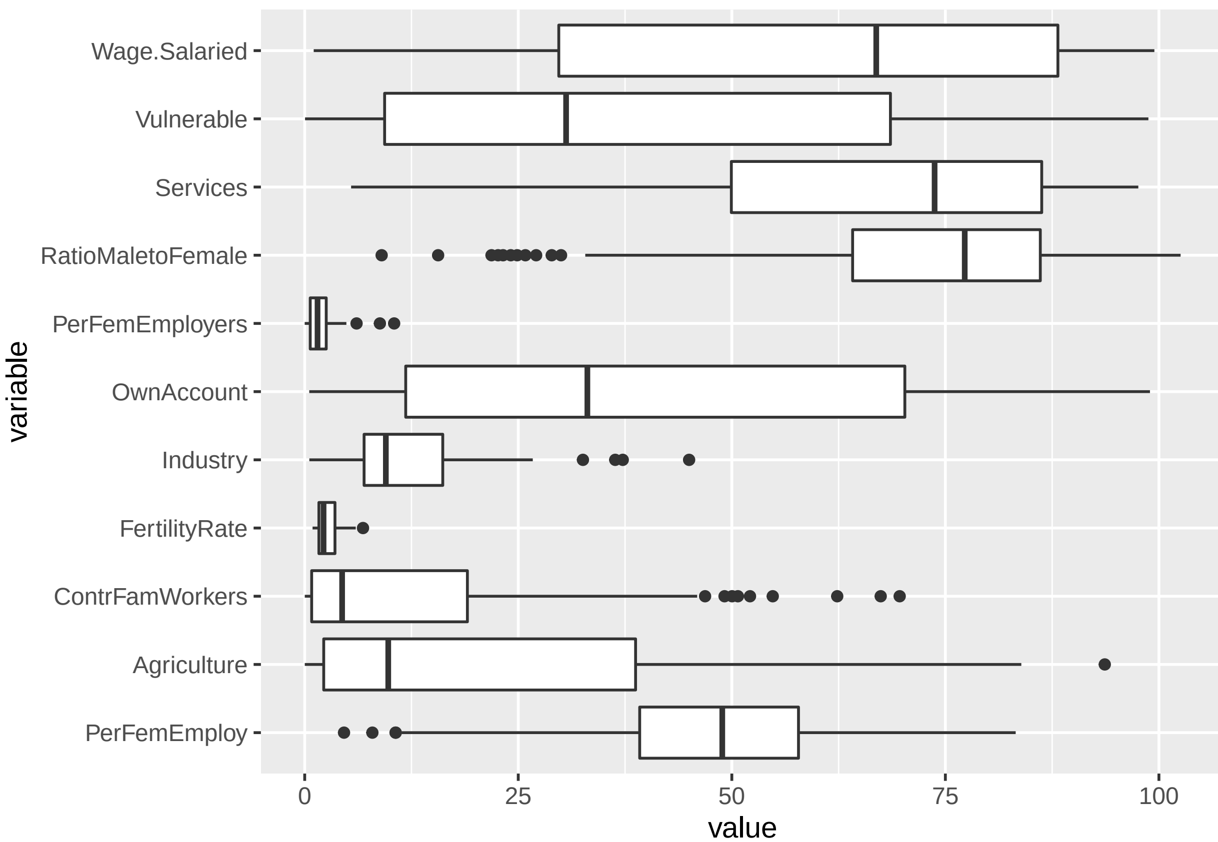
A.2. Missing Values

```
##                id       Country Code       Country Name       PerFemEmploy
##                 0                  0                  0                 30
##       Agriculture    ContrFamWorkers      FertilityRate           Industry
```

```
##                30                30                17                30
##        OwnAccount   PerFemEmployers RatioMaletoFemale          Services
##                30                30                30                30
##        Vulnerable     Wage.Salaried
##                30                30
```

We found some missing value for 30 countries. Since there was no data for these countries, we chose to omit them from our analysis. So, instead 217 data points, we'll be dealing with 187 countries as observations.

A.3. Outlier Analysis



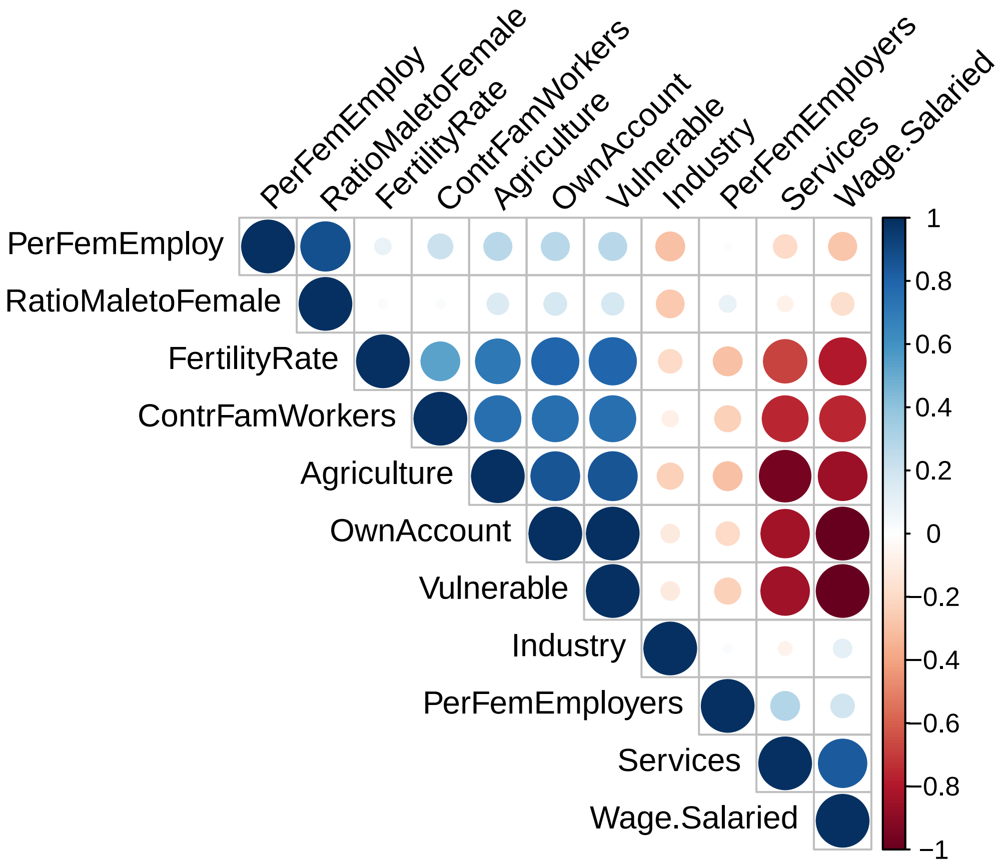
Outliers were detected and analyzed using the Outlier Box plots. From the outliers box plot we inferred that the data consists of many outliers for the target variable. However, the outliers for variable corresponded to outliers RatioMaletoFemale, PerFemEmployers, Industry, FertilityRate, ContrFamWorkers, Agriculture, and PerFemEmploy. Hence, We conclude that these outliers are legitimate outliers and we decided to retain to retain them in the data.

B. Exploratory Data Analysis

Let's see if the data meets the first assumption for regression analysis i.e. linear relationship of response with at least one of the regressors. The exploratory analysis will help to reveal relationship between the response and the regressor variables. The obtained results can help us narrow down our search for potential predictors that have significant effect on determining the percentage of employed women for a nation.

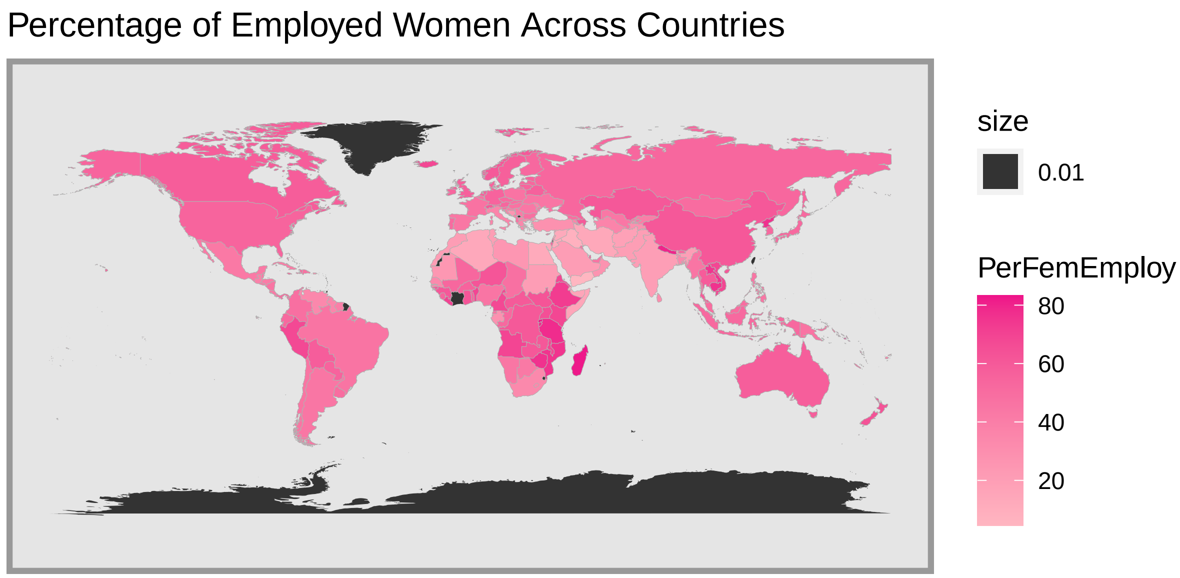
B.1. Correlation Visualization

We can see correlations between the variables by visualizing though correlation plot.



Our response *PerFemEmploy* is highly positively correlated with independent variable *RatioMaletoFemale*. Some independent variables seem to be highly correlated (for eg. *OwnAccount* is highly correlated with *Vulnerable* (+),*Agriculture* (+), *Services* (-) and *Wage.Salaried* (-) .etc). We need to have a closer look at these variables to determine if multicollinearity is present in data models.

B.2. Geographic Analysis



The above world map shows the percentage of employed women across different countries in the world.By looking at the above graph, majority of the countries have more than 60% or more women employed whereas a very countries lie less than 40% of the women employed.

C. Regression and Statistical Analysis

C.1. Full Model

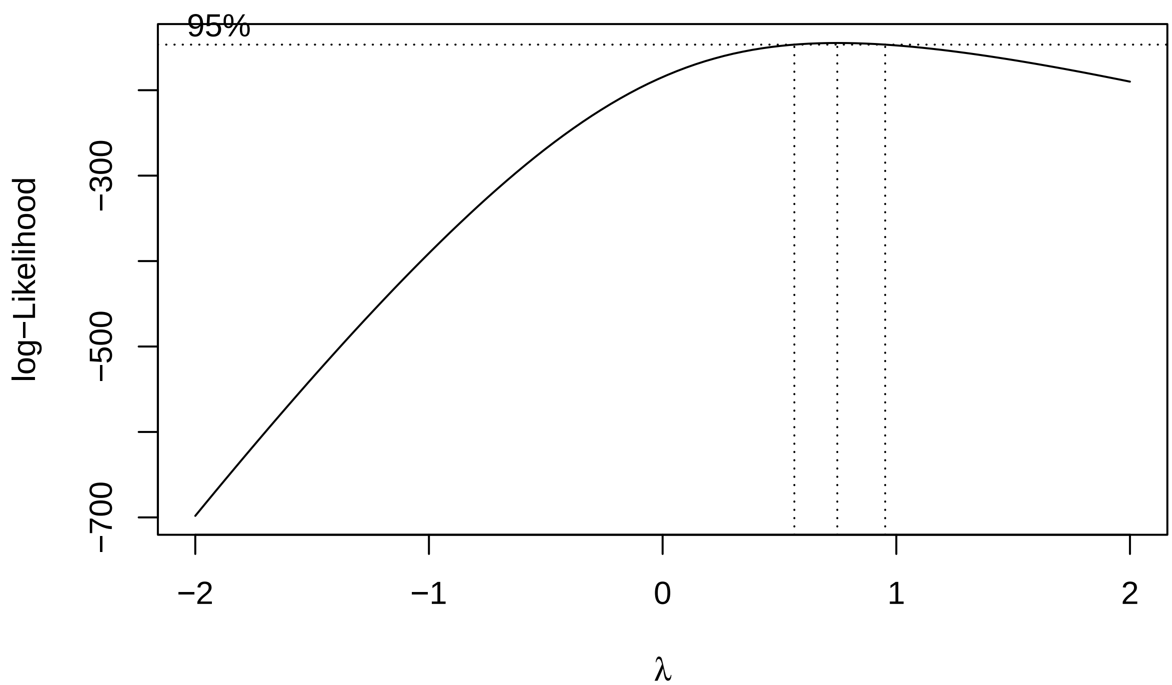
First, we prepare a linear model using all data in the dataset. This includes 1 response variable *PerFemEmploy* and 10 explanatory variables. A model summary is shown below:

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)    
## Agriculture          1   3620    3620  68.361 3.22e-14 ***
## ContrFamWorkers      1      7       7   0.125    0.724    
## FertilityRate        1   1067    1067  20.149 1.29e-05 ***
## Industry             1   2841    2841  53.652 8.27e-12 ***
## OwnAccount           1   2714    2714  51.263 2.11e-11 ***
## PerFemEmployers      1     12      12   0.225    0.636    
## RatioMaletoFemale    1  28576   28576 539.672  < 2e-16 ***
## Services             1    143     143   2.698    0.102    
## Vulnerable           1     36      36   0.672    0.413    
## Wage.Salaried        1     30      30   0.558    0.456    
## Residuals          176   9319      53                     
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The predictors *Agriculture* , *FertilityRate* , *Industry*, *OwnAccount*, *RatioMaletoFemale* are significant for this model (p-values less than 0.05 significance level).
- For our full model, we get p-value of 2.2e-16, which is very less than α. We reject the null hypothesis and say this regression model is significant to be considered.
- The R^2 (0.80) and R^2_{Adj} (0.79) values are good enough. But these high values may also be due to more variables used in the model. We need to simplify the model by removing redundant terms.
- The coefficient estimates are high and we might need to perform centering of the regressors.

C.2. Possible Transformations




```
## [1] "lambda = 0.75"
```

The Box-Cox transformation suggests that we transform our response before commencing. But let's check our model assumptions before so we can see if it meets constant variance assumption.

C.3. Diagnostics Residual Analysis

After fitting full we check if our full model meets the four regression assumptions:

1. We see the *Residuals Vs Fitted Plot* for seeing whether there is a linear relationship between the response and predictors. As the red fitted line is approximately close to the horizontal (residual=0) line, we conclude our model meets Linearity Assumption.

2. We perform *Durbin-Watson Test* to check is error terms are independent or not. We get significant test statistic to prove that the error terms are independent.

3. To check if our data follows normal distribution, we check the *Normality Q-Q Plot*. We conclude error terms are normal.

4. We see the *Scale-Location Plot* and perform *Breusch-Pagan test* to check if model meets the constant variance assumption. As the error terms are randomly distributed and show no definite pattern, we conclude the error terms have equal variance. We don't need to transform our response now.

5. We check for outliers, leverages and influential points. We find some leverages and outliers but decide not to remove them for further analysis.

6. We found very high multicollinearity for our full model. It was mostly due to variables *OwnAccount* and *Wage.Salaried*.

C.3. Variable Selection

We perform variable selection using various regression methods such as best subsets, forward regression, backward regression and step wise regression. Most of our models suggested by methods using forward, backward and step wise regression methods were also found in our best subsets table. We check for model adequacy in terms of regression assumptions, multicollinearity, PRESS values, predictive R^2 value and also check if a model's fit can be improved by adding non-linear terms. If a model doesn't meet constant-variance assumption we used Box-Cox transformation on response to stabilize the variance The results for our best subsets model is shown below:

```
##                                                             Best Subsets Regression
## ----------------------------------------------------------------------------------------------------------------------
## Model Index    Predictors
## ----------------------------------------------------------------------------------------------------------------------
##      1         RatioMaletoFemale
##      2         ContrFamWorkers RatioMaletoFemale
##      3         ContrFamWorkers PerFemEmployers RatioMaletoFemale
##      4         ContrFamWorkers FertilityRate PerFemEmployers RatioMaletoFemale
##      5         ContrFamWorkers FertilityRate Industry PerFemEmployers RatioMaletoFemale
##      6         Agriculture ContrFamWorkers Industry PerFemEmployers RatioMaletoFemale Services
##      7         Agriculture ContrFamWorkers FertilityRate Industry PerFemEmployers RatioMaletoFemale S
##      8         Agriculture ContrFamWorkers FertilityRate Industry PerFemEmployers RatioMaletoFemale S
##      9         Agriculture ContrFamWorkers FertilityRate Industry PerFemEmployers RatioMaletoFemale S
##     10         Agriculture ContrFamWorkers FertilityRate Industry OwnAccount PerFemEmployers RatioMal
## ----------------------------------------------------------------------------------------------------------------------
##
##                                                       Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------------------------
##                       Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC         SBC          MSE
## ----------------------------------------------------------------------------------------------------------------------
##   1        0.7576      0.7563      0.7529    38.3624    1310.5002    779.1481    1320.1935    11848.
##   2        0.7943      0.7921      0.7871     6.8642    1281.8169    751.1062    1294.7413    10110.
##   3        0.7970      0.7936      0.7878     6.4363    1281.3844    750.7687    1297.5400    10034.
##   4        0.8001      0.7957      0.7888     5.5996    1280.5017    750.0585    1299.8883     9935.
```

```
##   5        0.8028       0.7973       0.7892      5.1201      1279.9450     749.7174     1302.5628      9854.
##   6        0.8032       0.7966       0.7854      6.7952      1281.6074     751.4822     1307.4563      9892.
##   7        0.8054       0.7978       0.7861      6.7767      1281.4962     751.6379     1310.5762      9835.
##   8        0.8060       0.7972       0.7832      8.2303      1282.9205     753.2237     1315.2316      9861.
##   9        0.8072       0.7974       0.7816      9.0619      1283.6836     754.2354     1319.2258      9851.
##  10        0.8073       0.7964       0.7787     11.0000      1285.6178     756.3020     1324.3912      9904.
## ------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
##
##                                    Stepwise Selection Summary
## -----------------------------------------------------------------------------------------------
##                                 Added/                     Adj.
## Step         Variable          Removed     R-Square     R-Square      C(p)         AIC        RMSE
## -----------------------------------------------------------------------------------------------
##    1    RatioMaletoFemale     addition       0.758        0.756     38.3620    1310.5002    7.9598
##    2     ContrFamWorkers      addition       0.794        0.792      6.8640    1281.8169    7.3528
## -----------------------------------------------------------------------------------------------
```

We generate 7 models using results of variable selection methods. All our models follow regression assumptions. The only criteria left to pick the best performing model is the predictive R^2 value, complexity of the model (like whether the response was transformed and the no. of predictors used) and multicollinearity was present in the model. We can summarize the results of variable selection models in the table below:

```
## # A tibble: 7 x 5
##   Model `predrsq (in %)` multicollinearity `boxcox(response)` n_predictors
##   <chr>            <dbl> <chr>             <chr>                     <dbl>
## 1 m0                78.7 No                No                            2
## 2 m1                78.9 No                No                            5
## 3 m2                80.2 Very High         Yes                           7
## 4 m3                78.8 No                Yes                           1
## 5 m4                78.8 No                No                            3
## 6 m5                79.8 High              Yes                           6
## 7 m6                78.9 No                No                            4
```

Model *m2* had the highest predictive R^2 value (80.15) but had very high multicollinearity. If we discard the models where multicollinearity was found (m2 & m5), we have to see model good in terms of complexity and prediction power. Though *m3* is also a good choice, it has very simple fit (simple linear regression model with 1 predictor). We can validate our results by performing cross-validation on models *m0*, *m1*, *m4* and *m5*.

C.4. Cross-Validation

We create 10 folds divided in 1:1 ratio for training and validation sets from our data. Each of the 4 models are fit on these folds using cost = rtmspse .

```
##
## 2-fold CV results:
##    Fit       CV
## 1 Fit1 4.485589
## 2 Fit2 4.444861
## 3 Fit3 4.492098
## 4 Fit4 4.490718
##
## Best model:
##     CV
## "Fit2"
```

From the above CV fit summary, we find that *Fit 1* or Model *m0* performs the best on 10-fold cross validation with the least CV score . Model *m0* is good both in terms of predictive power (Pred. $R^2 = 78.70\%$) and model complexity (2 predictors used with no transformation on the response). Therefore, we can deduce *m0* as our final model in the next section.

Results

Summary of our final model:

```
##
## Call:
## lm(formula = PerFemEmploy ~ RatioMaletoFemale + ContrFamWorkers,
##     data = df[4:14])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5086  -4.9460   0.6985   4.7566  21.4354
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.87396    2.11376  -3.252  0.00136 **
## RatioMaletoFemale  0.71930    0.02780  25.871  < 2e-16 ***
## ContrFamWorkers    0.20123    0.03513   5.728 4.08e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.353 on 184 degrees of freedom
## Multiple R-squared:  0.7943, Adjusted R-squared:  0.7921
## F-statistic: 355.3 on 2 and 184 DF,  p-value: < 2.2e-16
```

Regression Equation for Estimated Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$PerFe\hat{m}Employ = -6.87 + 0.72 RatioMaletoFemale + 0.20 ContrFamWorkers$$

Discussion

- The intercept is -6.87. i.e. average % of female employed has been negatively impacted in that year.

- Our response is positively correlated with *RatioMaletoFemale*. Unit increase in *RatioMaletoFemale* increases our response by 0.72 units, keeping all other predictors constant. We can interpret it as a healthy ratio of Male:Female in a country promotes female employment opportunities in a country.
- Our response is positively correlated with *ContrFamWorkers*. Unit increase in *ContrFamWorkers* increases our response by 0.20 units, keeping all other predictors constant. The more *ContrFamWorkers* in a household increases the chances of females being employed.

Conclusion

From Exploratory Data Analysis and Regression and Statistical Analysis, we identified the most important and statistical significant attributes affecting the percentage of female employed in 2019. The variables RatioMaletoFemale and ContrFamWorkers had a significant effect on our response and account most part of the variance explained through our model. We are surprised that factors like Industry and FertilityRate have a low impact on the response.

We are confident that our model deals with multicollinearity and has low bias. We successfully reduced 10 variables into 2 significant ones.

References

1. https://genderdata.worldbank.org/data-stories/flfp-data-story/#:~:text=The%20global%20labor%20force%20particip
2. https://ourworldindata.org/female-labor-supply
3. https://www.kaggle.com/datasets/mdmuhtasimbillah/female-employment-vs-socioeconimic-factors
4. https://databank.worldbank.org/source/world-development-indicators/Type/TABLE/preview/on#