PORTFOLIO ASSIGNMENT 2

Andrew Guo

PROBLEM

Context

Women roughly occupy half of the world's population but when it comes to the total workforce of a country, the percentage of male and female workers 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 infrastructures are responsible for this discrepancy, it goes way beyond these. And to show the effects of multiple socioeconomic factors on the participation of women in the total workforce, percentage of female employment in the total labor force has been considered. Using multiple linear regression model, the relationship between these factors can be analyzed.

Content

For the current study, the data set has been chosen from a survey performed on the population of Bangladesh. The datasets selected for this study span over 25 years (from 1995 to 2019). Data has been collected separately from multiple datasets from the World Bank databank for the employed women percentage and the related predictor variables. These datasets were compiled into one 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.
- `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).
- `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.
- `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.

(a) Read in the data and use the `lm()` function to perform a multiple linear regression with `PerFemEmploy` as the response and all of the predictors. Use the `summary()` function to print the results.

```
femploy = read.csv('FemEmploy.csv')
attach(femploy)

lm.fit.femploy1 = lm(PerFemEmploy~FertilityRate + RatioMaletoFemale + PerFemEmployers +

summary(lm.fit.femploy1)
```

```
##
## Call:
## lm(formula = PerFemEmploy ~ FertilityRate + RatioMaletoFemale +
##     PerFemEmployers + Agriculture + Industry + Services + Wage.Salaried +
##     ContrFamWorkers + OwnAccount + Vulnerable)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.274518 -0.079047 -0.009186  0.063606  0.315424
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2351.7930  1359.7895  -1.730   0.1093
## FertilityRate         1.7000     1.8535   0.917   0.3771
## RatioMaletoFemale     1.1454     0.4183   2.738   0.0180 *
## PerFemEmployers      -4.7301     8.2105  -0.576   0.5752
## Agriculture          31.1658    15.1784   2.053   0.0625 .
## Industry             30.7485    15.1055   2.036   0.0645 .
## Services             31.2554    15.1832   2.059   0.0619 .
## Wage.Salaried        -7.7778     7.9012  -0.984   0.3444
## ContrFamWorkers     -26.6192    12.0221  -2.214   0.0469 *
## OwnAccount          -26.8093    11.9876  -2.236   0.0451 *
## Vulnerable           18.8815    11.7465   1.607   0.1339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1734 on 12 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.9968, Adjusted R-squared:  0.9942
## F-statistic: 378.9 on 10 and 12 DF,  p-value: 2.056e-13
```

(b) Is there a relationship between the predictors and the response? Briefly explain.

The global omnibus F-test yields a p- value of less than 0.05, indicating that in our model, we have at least one statistically significant predictor. If we examine each of the predictors within the model, we see that only the variables 'RatioMaletoFemale', 'ContrFamWorkers', and 'OwnAccount' are statistically significant, as they are the only ones with p-values less than 0.05.

(c) Which predictors appear to have a statistically significant relationship to the response at the $\alpha = .05$ level?

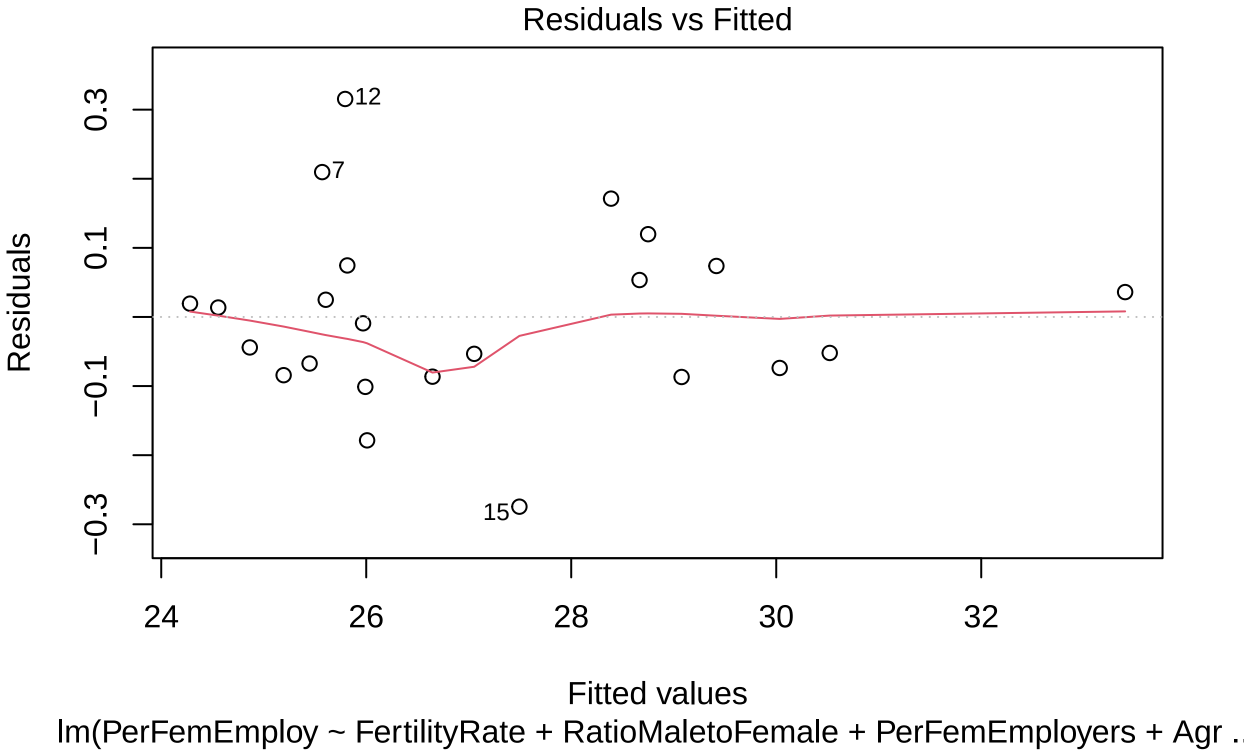
See answer above.

(d) What is the interpretation of the coefficient for the `RatioMaletoFemale` variable? Briefly explain.

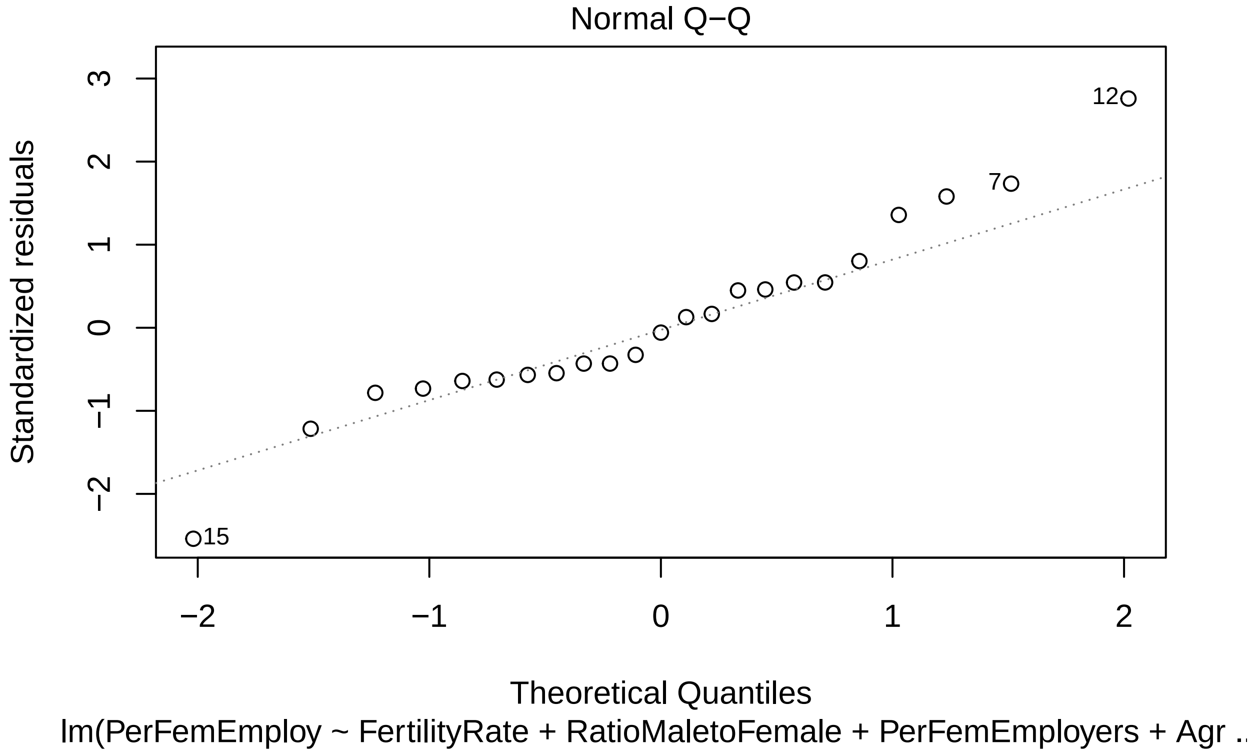
While holding all other predictors constant, the employment to population ratio of women who are of age 15 or older increases by a factor of 1.1454 for a one unit increase in the ratio of the female to male labor force participation rate.

(e) Use the `plot()` function to produce diagnostic plots of the regression fit. Comment on any problems you see with the fit in terms of the appropriateness of a linear model, normality, heteroscedasticity, outliers, and high leverage observations.

```r
par(mfrow=c(1,1))
plot(lm.fit.femploy1, which = c(1))
```

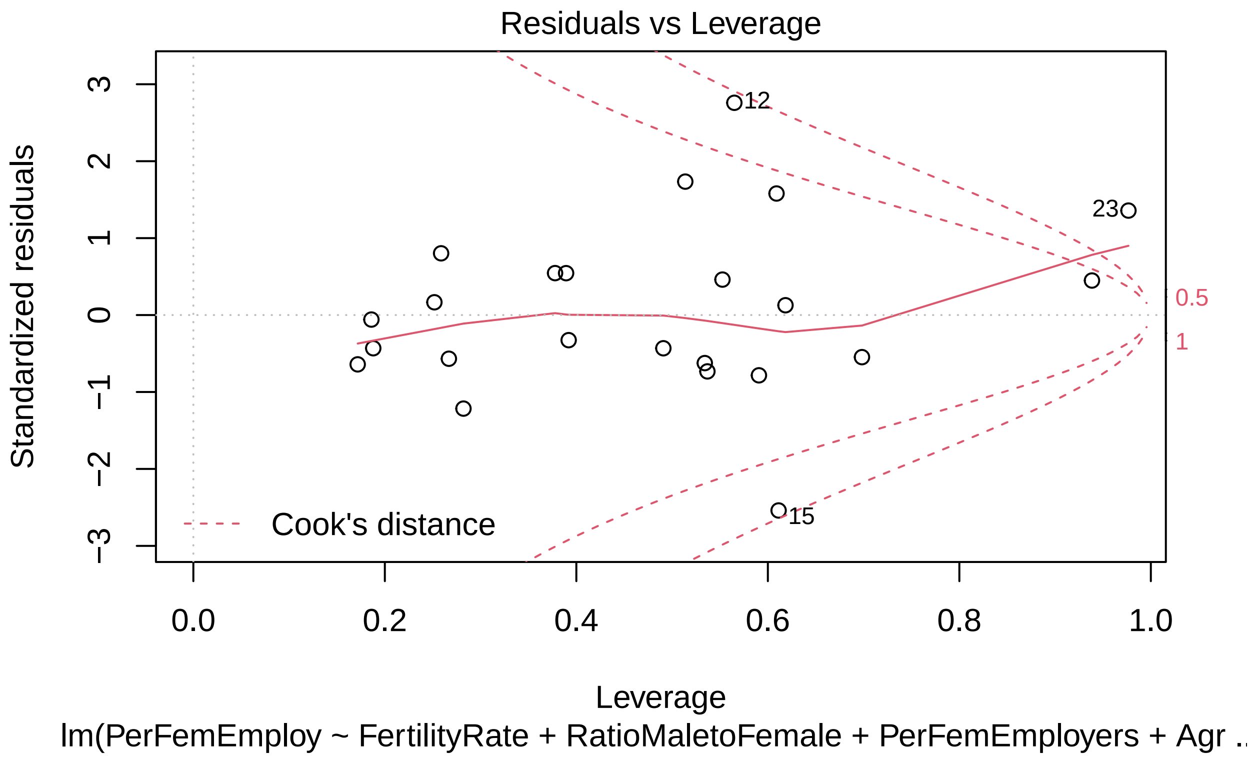


```r
plot(lm.fit.femploy1, which = c(2))
```

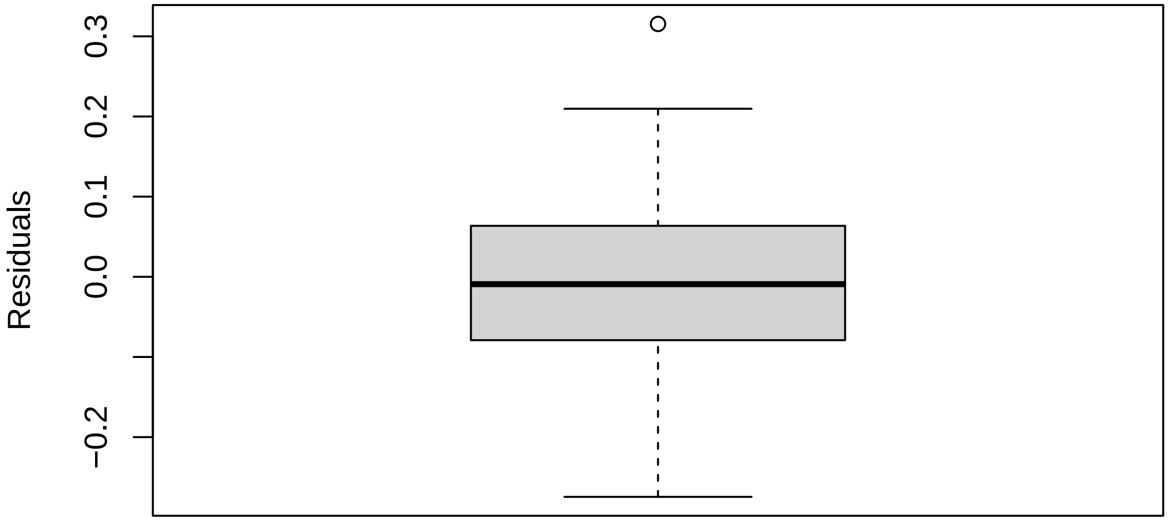


```r
plot(lm.fit.femploy1, which = c(5))
```

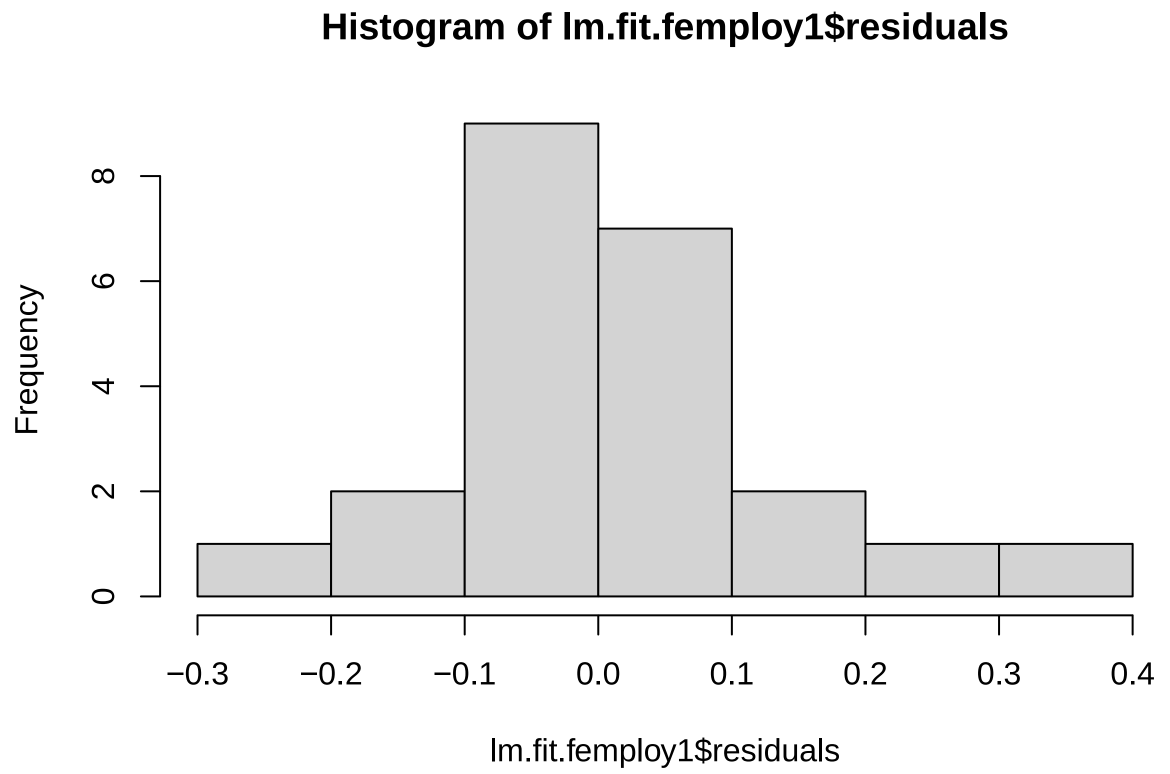
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



```
boxplot(lm.fit.femploy1$residuals, ylab="Residuals")
```



```
hist(lm.fit.femploy1$residuals)
```



Our boxplot indicates the data may be normally distributed, as the distribution does appear to be somewhat symmetrical. However, it does indicate that we have at least one outlier with a residual outside of 75% of all other residuals. The histogram reveals a level of asymmetry as our data appears to be slightly skewed to the right

The quantiles on the Q-Q plot follow a moderately strong linear relationship, indicating that our data appears to be approximately normal. However, the values at the tails of the plot (especially at the upper end) show that there may be some abnormality in the distributions, but because the overall shape of the plot is somewhat linear, we can make the initial claim that our data is approximately normal.

The standardized residuals versus leverage plot indicates both many outstanding outliers and high leverage points present. For instance, observations 12 and 15 both contain a leverage statistic of 0.6 while nearing +-3 standard deviations, indicating that they are both high leverage points and outliers. Observation 23 has a leverage statistic of nearly 1, showing that it is an extremely potent high leverage point. Overall, many of the points on this plot appear to hold high leverage, with many ranging from 0.2 to 1 with standard deviations between +-1, although there are certain observations nearing +-3 standard deviations.

Our residual plot does indicate that there is a minor fanning out as the residuals as the fitted values increase, although this fanning decreases around observation 29, indicating that our data may not homoscedastic. However, it would help if there were more observations to describe a more definite pattern; there are not that many observations to make an ideal, definite conclusion shape of the residuals. However, the fanning in and out is noticeable enough to discredit the homoscedasticity of our data.

(f) Based on the results in parts (b), (c), and (e), what actions might you take?

First, I would remove all predictors that are statistically insignificant. Then, I would create a new model only using those predictors that are significant, then using forward model selection, consider any additional/possible predictors that could have an effect on the response and include them only if they are statistically significant. Afterwards, I would perform model diagnostics to assess both normality and homoscedasticity. Based on these results, I will do the following:

If homoscedastic and approximately normal, I will make no adjustments or manipulations on X or Y.

If homoscedastic but not approximately normal, I would make transformations on X, perhaps trying log(x) or sqrt(x) and hope that normality is achieved without disturbing homoscedasticity.

If heteroscedastic but approximately normal, I would make transformations on Y, perhaps trying log(y) or sqrt(y) and hope to achieve homoscedasticity without disturbing normality.

If heteroscedastic and not approximately normal, I would first make transformations on Y to achieve homoscedasticity. Once this is achieved through some manipulation on Y, then I will transform on X in hopes to achieve normality through some manipulation on X.