**ISyE 312  
Data Analytics Project Report  
Friday, April 23rd, 2021**

**Project Report**

The Effect of a Country’s GDP on the respective Female Labor Force and Fertility Rate in the Years 1990, 2000, 2010, and 2018

**Saranya Putrevu**

**Nivedha Gautham Raj**

**Maxwell Wind**

**Satvik Yagnamurthy**

1. **Introduction**  
     
   The development of countries throughout the world can be attributed to numerous characteristics including culture, economics, government, opportunities and more. In order to evaluate the strength of a country, a common metric that is utilized is the Gross Domestic Product per capita monetary measure. A country can calculate respective GDP through the summation of money spent by consumers, business, and government in an allotted time period. Countries with efficient workforces tend to translate into higher GDP values which can break down to educational opportunities and training.

To generate an efficient workforce, both men and women necessitate equal opportunities in the industry, government, and various entities. However, in history, large number of countries would focus on educating men and ensuring that these individuals would receive comparatively more focus to eventually act as economic, political, etc. leaders. Based on the current Fortune 500 CEO list updated in 2020, women CEOs make up of 7.4% of the total count. Women thus require additional opportunities in order to move up the ladder and receive the chance to make a powerful impact.

A potential reason as to why this number is lower than desired is due to the societal expectation of raising a child. Companies provide maternity leave opportunities though the policies and overall stigma is that these responsibilities fall under the women rather than the men. Potentially as a result of this, Japanese women for example on average have their first child at the age of 30.7 as per 2017 compared to 25.7 years old in 1975 according to Nippon Japanese Scholarly Analysis. Societal expectations firstly need to change, and women should be pushed towards receiving an education and providing for themselves.

According to the Washington Post, fifty years ago, 58% of United States College Students were men. As of 2019, 56% of United States College Students are women which is an incredible growth traced by the Educational Department of the United States Government. Therefore, our team is passionate to understand the relationship between a country’s GDP and the respective female labor force. In addition, the relationship between a country’s GDP and the respective fertility rate. Our team’s hypothesis is that there is a strong relationship between the female labor force and an inverse relationship between fertility rate to a country’s GDP. The data analysis will be focusing on exclusively the years 1990, 2000, 2010, and 2018. R will be used to analyze and visualize the datasets; the conclusion will be drawn from visualizations of the datasets.

1. **Objectives and Methodology**

**Question 1:** *What is the relationship between a country’s GDP and the number of women in the country’s labor force?*

Our team would expect that as a country’s GDP increases, the number of women in the country’s labor force would also increase.

**Question 2:** *What is the relationship between a country’s GDP and the fertility rate?*

Our team expects to see an inverse relationship, i.e., the higher a country’s GDP, the lower its fertility rate would be.

**Question 3:** *What is the relationship between the number of women in the labor force and the fertility rate of each country?*

Our team predicts that more women in the labor force would result in a lower fertility rate.

1. **Dataset**

**Dataset:** Labor Force, Female (% of total labor force) and GDP Growth (annual %), Fertility Rate, total(births per woman) - data sets drawn from the World Bank Indicators

<https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?view=chart>

<https://data.worldbank.org/indicator/SL.TLF.TOTL.FE.ZS?view=chart>

<https://data.worldbank.org/indicator/SP.DYN.TFRT.IN?view=chart>

**Dataset Description:**

* Annual % GDP growth rate of each country/region
* Female participation rate in each country/region’s labor force
* Fertility rate of each country/region
* Team focused on the years 1990, 2000, 2010, and 2018 on the data analysis.

1. **Analysis**
   1. **Simple Linear Regression:**

The three variables assessed in this case include Gross Domestic Product of each country, Female participation rate in each country, and fertility rate of each country. Through the simple linear regression, our team evaluated the relationship between GDP and each of the remaining variables.

Simple linear regression between GDP and fertility rate:

1990:

Text

Description automatically generated

2000:

Text

Description automatically generated

2010:

Text

Description automatically generated

2018:

Text

Description automatically generated

Of the four years our team analyzed, the R-Squared values for the simple linear regression models were 0.1257, -0.002476, -0.001773, and 0.01762 respectively. The two years with negative R-Squared values (2000 and 2010) illustrate that the two models' predictions tend to be less accurate than the average value of the data set. In the year 1990 the R-Squared value for our model was 0.1257, this tells us that 12.57% of the variation in the outcome variable (GDP) is explained by the predictor variable (fertility rate). Similarly, in the year 2018, 1.7% of the variation in GDP is explained by fertility rate. These R-Squared values are extremely low indicating that there is a very poor fit. The simple linear regression model output also provided us with coefficients.

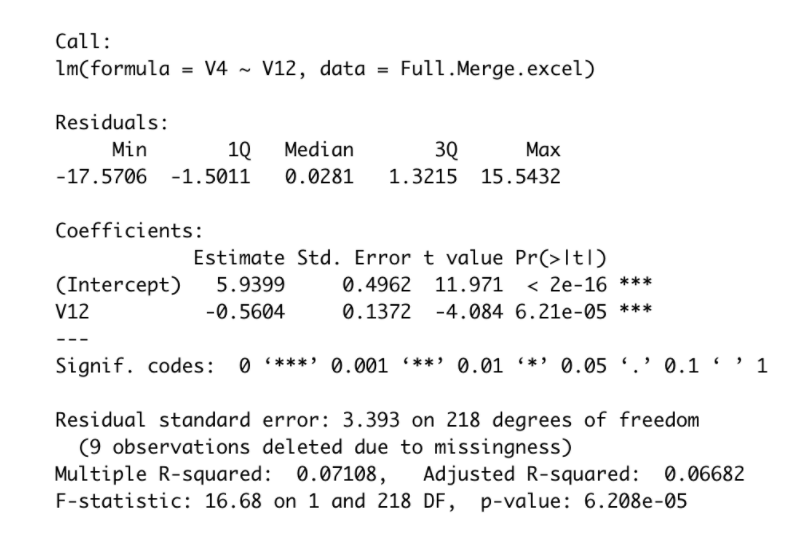
Simple linear regression between GDP and female participation rate:

1990:

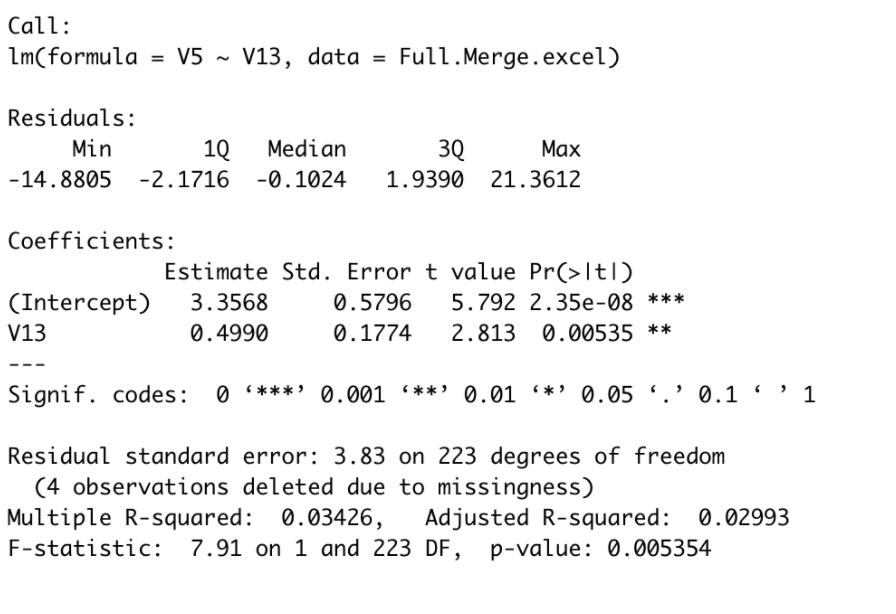
Text

Description automatically generated

2000:



2010:



2018:

Text, letter

Description automatically generated

* 1. **Multiple Linear Regression:**

A multiple linear regression model was created for each of the four years we analyzed – 1990, 2000, 2010, and 2018.

Text

Description automatically generated

This model has a R squared value of 0.13, implying a poor fit of model. Looking at the coefficients, we see that the female % of the total labor force has a negative effect on the GDP, while fertility rate has a 19% effect on GDP. Thus, we see that the female % of the total labor force has an inverse relationship with the GDP while fertility rate has a minimal effect on GDP. However, the model has an extraordinarily low p-value, implying the variables have a strong effect on the GDP.

A picture containing text, receipt, screenshot

Description automatically generated

The 2000 data, when modelled, gives a R-squared value of 0.07. This implies a very poor fit of model. Analyzing the coefficients, we see that both, the female % of total labor force and fertility rate are inversely related to GDP. However, the very low p-value implies that changes in the predictor variables will have a strong effect on the response variable.

Text

Description automatically generated

The 2010 model has a R-squared value of 0.04, implying another poor fit. The female % of the total labor force has an inverse relationship with GDP. The fertility rate has a 50% effect on the GDP. Therefore, we can conclude that the fertility rate has a significant relationship on GDP while % of women working in the labor force does not. A p-value of 0.01 implies a significant relationship between the predictor variables and the response variable.

Text, table

Description automatically generated

The R-squared value for the 2018 model is 0.02, implying a weak fit. According to the 2018 model, both the female % of total labor force and fertility rate have a weak relationship with the GDP. Their effect is almost negligible. Additionally, a p-value > 0.05 implies the predictor variables do not have a strong effect on the response variable.

From the R-squared values of each of the different multiple regression models, it can be concluded that the models are not a good fit. However, the coefficients and the p-values indicate otherwise in some cases. Therefore, we check the assumptions and other influential factors in the model.

* + 1. **Outliers and Influential Cases**

Outliers can have a significant effect on the data and largely affect the data analysis. In the analysis that was conducted for this project, a few countries did not have data entries in one of the three databases – GDP growth, female % of total labor force and fertility rate. These countries were omitted from the databases to facilitate calculations.

Furthermore, R Studio omitted some countries, as shown in the screenshots of the multiple linear regression models. These countries were omitted by the software as they would have data entries missing in one of the years (columns), that would make calculations more difficult. Thus, in this study, the outliers were removed as such.

However, as the following residual plots will exhibit, there are still a few remaining outliers and influential cases. Despite those data points lying outside the range of most of the data, they cannot be omitted as they still have a considerable influence on the regression models. If they are omitted from the analysis, the model changes, but not for the better. Therefore, none of the data points are eliminated.

* + 1. **Residual Analysis**

The following residual plots were obtained for the different multiple linear regression models: Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart

Description automatically generatedChart, line chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

The different residual plots are presented in the order of 1990, 2000, 2010 and 2018. All the residual vs fitted plots look similar – they are all funnel shaped. Meanwhile, all the residual Q-Q plots show a light-tailed distribution. Therefore, the residual plots do not pass the test for normality, i.e., they do not have a normal distribution.

* + 1. **Checking Assumption of Independence:**

The independence of the data can be checked by checking the collinearity of the variables. This is done as follows: Text

Description automatically generatedText

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

As the coefficients of the multiple regression model suggested, there is not a strong correlation between the different variables. Neither of the predictor variables have a strong effect on the response variable until the 2018 model. Furthermore, the female % of total labor force does not have a positive correlation with the GDP until the 2018 model. Thus, it can be concluded from the correlation matrix that neither the female % of total labor force nor the fertility rate has a strong correlation, i.e., effect on the GDP of a country.

* + 1. **Checking the Assumption of Multicollinearity:**

The multicollinearity – a strong correlation between any two of the variables – is checked by performing a variance inflation factor (VIF) test. The results were as follows:

Text

Description automatically generated

The VIF test indicates strong collinearity if the model returns a value of larger than 15. In the case of these multiple linear regression models, all of the values are at 1. Therefore, it is easy to conclude that none of the variables are strongly correlated. Furthermore, as the mean of the VIF values also equate to 1, it can be concluded that there are no biases in the data.

1. **Conclusion**