The Effects of Education and Fertility on Female Labor

Jonathan Li and Chris Whall

2025-11-07

Abstract:

Using data from the World Bank, we ran multiple different regression models to explore the effects that education expenditure and fertility rates have on female labor force participation. By using other variables like parental leave and tertiary school enrollment as control variables we are able to reduce omitted variable bias. After running multiple different models we found that higher education expenditure associates with higher female labor force participation. Fertility rates however, displayed a non-linear, U-shaped relationship with female labor force participation. Understanding of these relationships can assist policy makers with improving labor market outcomes and overall economic performance.

Introduction:

The purpose of this analysis is the examine the relationships between specific variables and the female labor participation rate. We focus mainly on how government expenditure on education affects the female labor participation rate and on how the fertility rate affects the female labor participation rate. We utilize regression modeling tools to measure the effect that these variables have on the female labor force participation rate. By exploring how the female labor force participation rate is affected numerous variables, we can understand how the female labor force participation rate affects the overall economy. Having this information is important for policy makers in both developing and developed countries. Increased female labor participation leads to lower poverty rates and more stable households. This information can be used to bring about economic stability and growth, better access to education, and gender equality.

Data:

Our analysis uses global datasets from the World Bank on female labor participation rates, government expenditure on education, birth rates, fertility rates, parental leave, tertiary school enrollment, and GDP. After removing unnecessary columns and creating consistent column names, we merged all the datasets into one dataframe. We then sorted the countries by the amount of missing values they had and removed any country with over 25% missing values.

Female Labor Participation Rate:

The female labor participation rate is represented by female_labor_rate in our data. It is a quantitative variable that measures the female labor force participation rate in terms of % of female population ages 15+. It contains data for almost all countries in the world and spans from 1990 to 2024. Data can be found at https://data360.worldbank.org/en/indicator/WB_WDI_SL_TLF_CACT_FE_ZS.

Government Expenditure on Education

Government expenditure on education is represented by edu_exp in our data. It is a quantitative variable that measures the government expenditure on education in terms of total % of GDP. It contains data for almost all countries in the world and spans from 1970 to 2023. Data can be found at https://data360.worldbank.org/en/indicator/WB_WDI_SE_XPD_TOTL_GD_ZS.

Birth Rate

Birth rate is represented by birth_rate in our data. It is a quantitative variable that measures the crude birth rate per 1000 people. It contains data globally and spans from 1960 to 2023. Data can be found at https://data360.worldbank.org/en/indicator/WB_WDI_SP_DYN_CBRT_IN.

Fertility Rate

Fertility rate is represented by fertility_rate in our data. It is a quantitative variable that measures the total total fertility rates in terms of births per woman. It contains data globally and spans from 1960 to 2023. Data can be found at https://data360.worldbank.org/en/indicator/WB_WDI_SP_DYN_TFRT_IN.

Maternal Leave

Maternal leave is represented by days_female in our data. It is a quantitative variable that measures the total length of paid maternal leave in terms of calendar days. It containts data globally and spans from 1970 to 2023. Data can be found at https://genderdata.worldbank.org/en/indicator/sh-par-leve?.

Female Tertiary School Erollment

Female tertiary school enrollment is represented by enrolled_female in our data. It is a quantitative variable that measures female tertiary school enrollment in terms of % gross. It contains data globally and spans from 1970 to 2024. Data can be found at http://genderdata.worldbank.org/en/indicator/se-ter-enrrl.

GDP

GDP is represented by gdp in our data. It is a quantitative variable that measures a country's GDP in terms of US Dollars. It contains data globally that spans from 1960 to 2024. Data can be found at https://data360.worldbank.org/en/indicator/WB_WDI_NY_GDP_MKTP_CD.

Visualization:

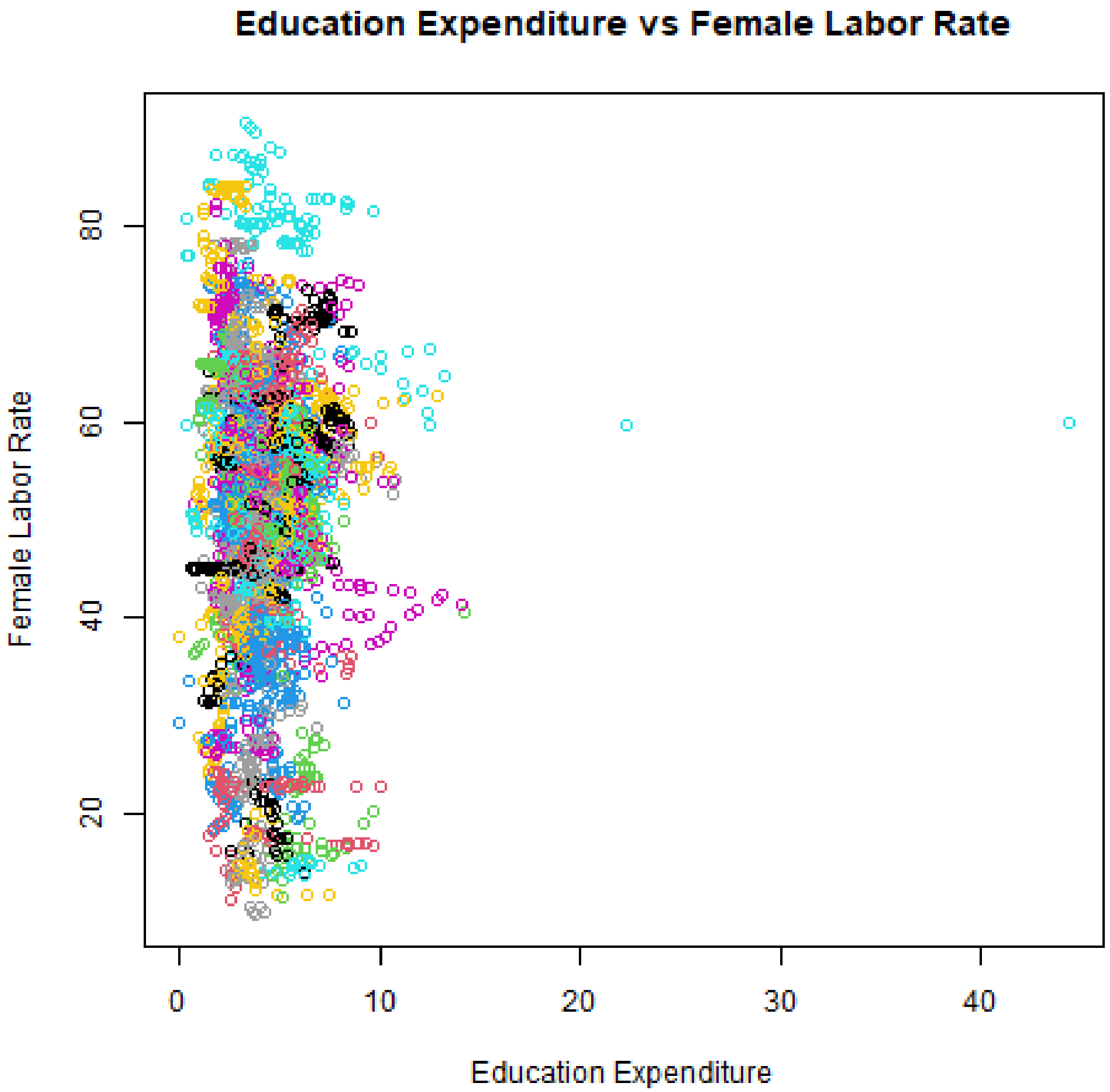

Figure 1: Scatter Plot of Education Expenditure and Female Labor Rate for each country.

From Figure 1 we can see that there is no clear relationship between education expenditure and female labor rates. However, we can see that for education expenditure there is a slight right skewed distribution. There are a handful of outliers that may cause issues with future modeling. It would be wise to test a log transformed variable for education expenditure to combat this.

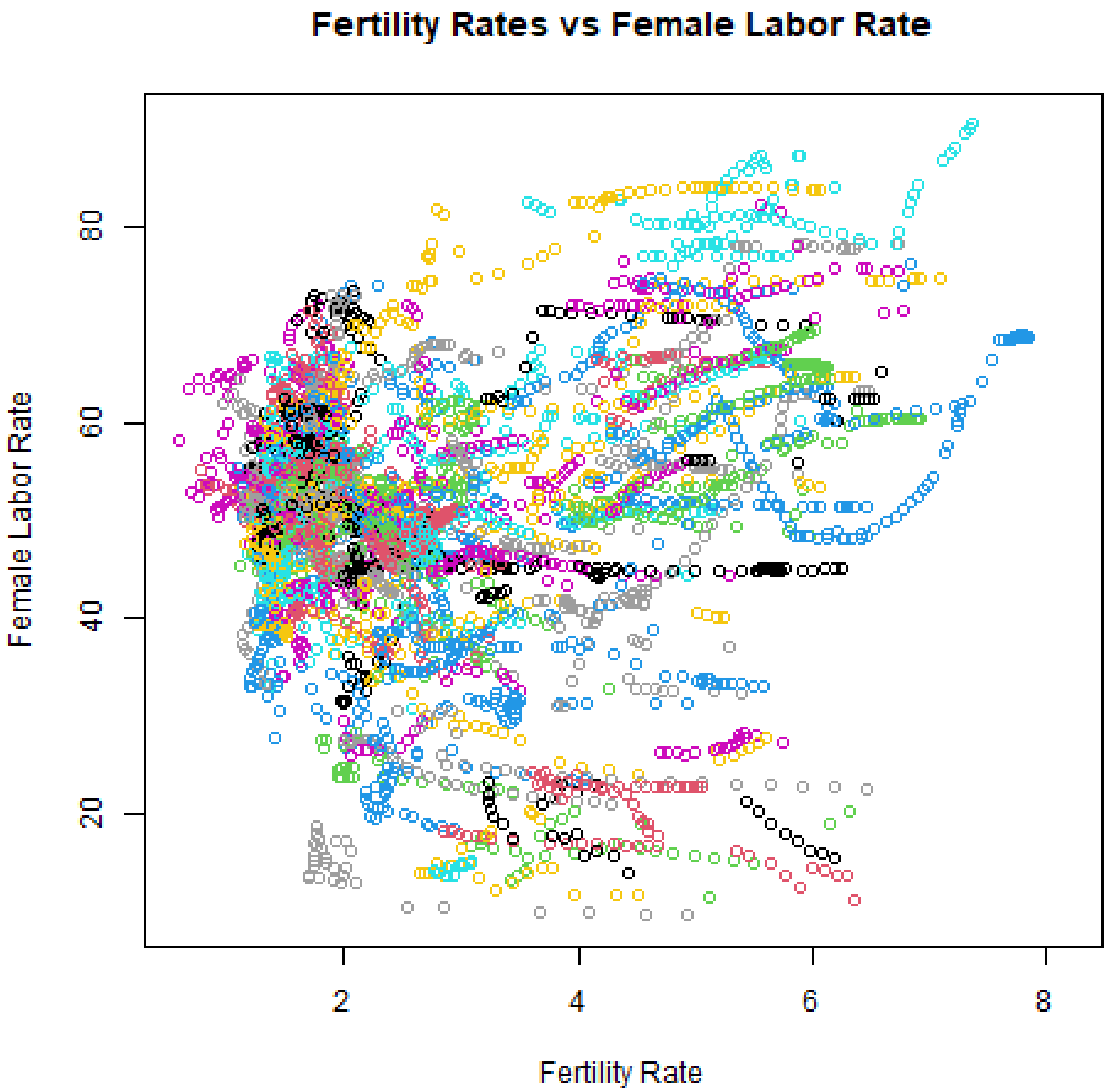

Figure 2: Scatter Plot of Fertility Rate and Female Labor Rate across countries.

In Figure 2, we see that there is no obvious relationship between fertility rate and female labor rate. When we look at individual countries, represented by different colors, we can see that each country has its own unique pattern. Fertility rate seems to affect the female labor force pariticipation rate differently for each country. This makes sense and we should see a clearer and more significant relationship between the two variables once we introduce other variables.

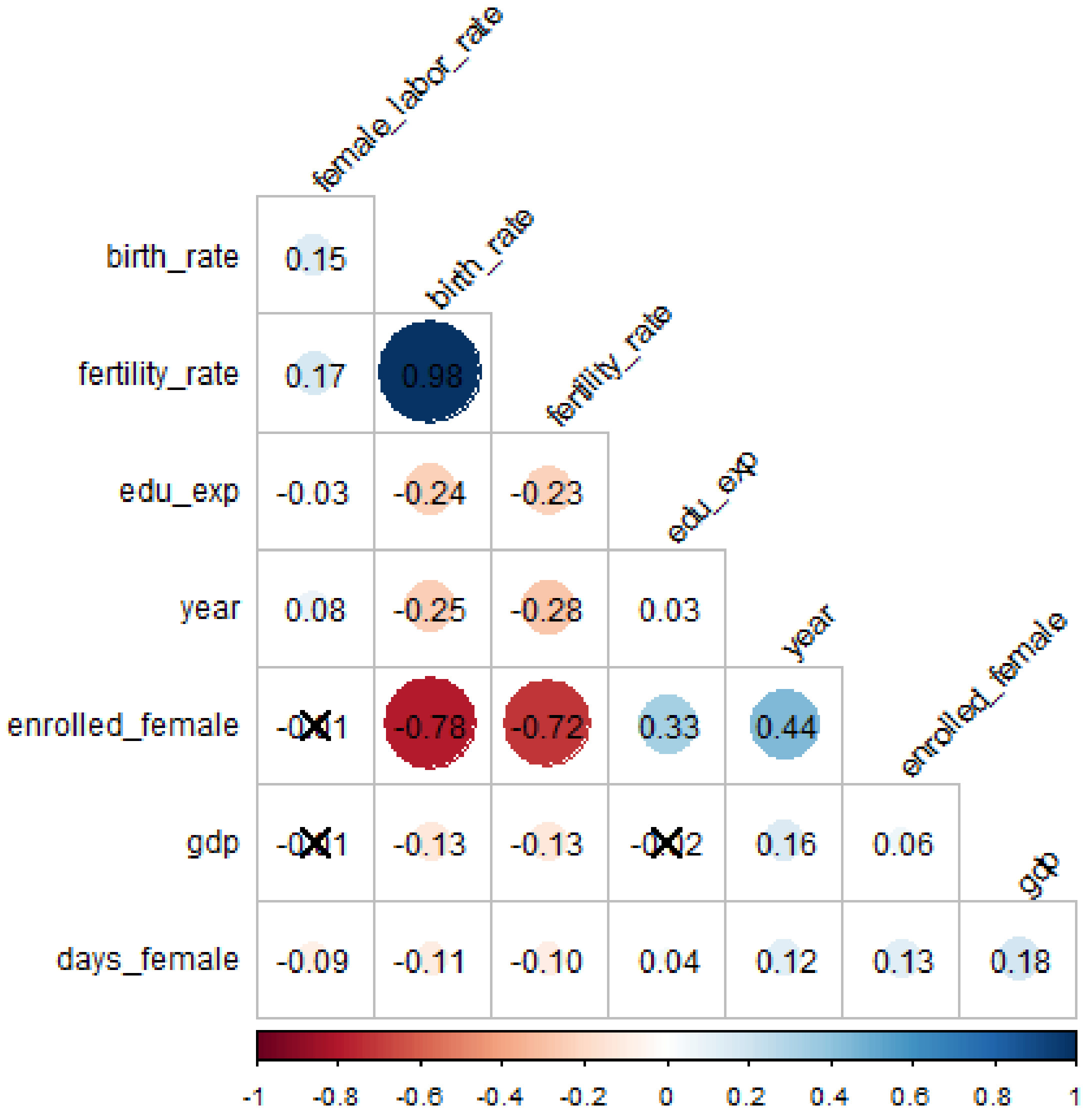

Figure 3: Correlation Matrix across all variables.

Figure 3 shows us the correlation between all variables in our data. Birth rate and fertility show extremely high positive correlation. This makes sense as they are fundamentally very similar statistics. We will have to be very careful when modeling both of these variables as we would risk collinearity. Female tertiary school enrollment has very high negative correlation with birth rate and fertility rate. It seems that as more women enter higher education, less children are born. There is a very weak negative significant relationship between education expenditure and female labor force participation. This is very strange as we expect there to be a stronger positive correlation between the two variables. Running a model with these two variables and other variables could provide more information on the behavior of this relationship. GDP seems to not have a significant effect on the female labor force participation rate. We can include it in initial models, but potentially look to omit if it seems to provide no value to the model.

Analysis:

To better understand the true relationship between these variables and the effects these variables have on the female labor force participation rate, we ran multiple different regression models with the female labor force participation rate as our dependent variable.

Model 0:

First we want to run bivariate model with just education expenditure and female labor rate to get a better understanding of the relationship between just these two variables.

$$female\_labor\_rate_i = \beta_0 + \beta_1 edu\_exp + \varepsilon_i$$

```
##
## Call:
## lm(formula = female_labor_rate ~ edu_exp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.886  -5.901   0.480   8.458  38.805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  52.4011     0.5216 100.456   <2e-16 ***
## edu_exp      -0.2248     0.1137  -1.978    0.048 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13.64 on 4553 degrees of freedom
##   (1154 observations deleted due to missingness)
## Multiple R-squared:  0.0008582,  Adjusted R-squared:  0.0006388
## F-statistic: 3.911 on 1 and 4553 DF,  p-value: 0.04803
```

In this model, we see that education expenditure has a significant negative effect on the female labor rate similar to the correlation matrix in Figure 3. According to this model, if a country spends no money on education, the expected female labor rate is 52.4011. For every unit increase in education expenditure, the female labor rate is expected to decrease by -0.2248. Intuitively, this doesn’t make much sense. While the coefficient is significant at the 5% level, it is very close to being insignificant and at higher levels it does become insignificant. It is possible that adding more variables will reduce ommitted variable bias and provide accurate model.

Model 1:

Next, we can introduce basic control variables like year and GDP. Doing this will potentially create some changes in how the regression calculates the relationship between education expenditure and the female labor rate.

$$female\_labor\_rate_i = \beta_0 + \beta_1 edu\_exp + \beta_2 year + \beta_3 gdp + \varepsilon_i$$

```
##
## Call:
```

```
## lm(formula = female_labor_rate ~ edu_exp + year + gdp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.023  -5.288   0.448   8.267  41.272
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.099e+02  4.570e+01  -4.594 4.47e-06 ***
## edu_exp     -2.213e-01  1.137e-01  -1.946   0.0517 .
## year         1.306e-01  2.274e-02   5.742 9.99e-09 ***
## gdp         -4.833e-14  2.912e-14  -1.660   0.0970 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.6 on 4539 degrees of freedom
##   (1166 observations deleted due to missingness)
## Multiple R-squared:  0.008211,   Adjusted R-squared:  0.007555
## F-statistic: 12.53 on 3 and 4539 DF,  p-value: 3.736e-08
```

In this model, we see that education expenditure actually becomes statistically insignificant at the 5% level. GDP is also insignificant at the 5% level. Overall, this model doesn't provide much explanatory value as the only significant variable is the year. This model just tells us that as time passes, the female labor rate will increase. The R-squared for this model is 0.008211 which is very small telling us that this model is rather weak in explaining the effects of education expenditure, year, and gdp on the female labor rate.

Model 2

From Figure 1 we saw that education expenditure was slightly skewed and had a few very extreme outliers. This could be affecting the model and making it less accurate. To try and combat this we can apply a log transformation to education expenditure and testing a linear-log model.

$$female\_labor\_rate_i = \beta_0 + \beta_1 log\_edu\_exp + \beta_2 year + \beta_3 gdp + \varepsilon_i$$

```
##
## Call:
## lm(formula = female_labor_rate ~ log_edu_exp + year + gdp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.471  -5.405   0.548   8.247  41.170
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.093e+02  4.562e+01  -4.588 4.59e-06 ***
## log_edu_exp -2.078e+00  4.762e-01  -4.363 1.31e-05 ***
## year         1.312e-01  2.271e-02   5.779 8.03e-09 ***
## gdp         -4.522e-14  2.906e-14  -1.556     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.57 on 4538 degrees of freedom
##   (1166 observations deleted due to missingness)
## Multiple R-squared:  0.01146,    Adjusted R-squared:  0.01081
## F-statistic: 17.54 on 3 and 4538 DF,  p-value: 2.56e-11
```

After applying the log transformation, education expenditure becomes significant at the 5% level. This model tells us that for every 1% increase in education expenditure, we expect a decrease in the female labor rate by -0.02078 keeping all else constant. However intuitively, this doesn’t make much sense. We expect that as government expenditure in education increases, the female labor rate should increase instead of decrease. Adding more variables will further reduce omitted variable bias and possibly give us a more accurate model. Overall, Model 2 shows improvement over Model 1 as its R-squared is 0.01146 giving it slightly more explanatory power.

Model 3

We can now introduce the rest of our variables into our model. After some testing, we found that a model containing both birth rate and fertility rate showed extremely high levels of multicollinearity. Therefore, we decided to remove birth rates from our model and keep just fertility rates, which will give us a better understanding of individual household environments.

$$\begin{aligned} female\_labor\_rate_i = \beta_0 + \beta_1 edu\_exp + \beta_2 fertility\_rate + \beta_3 days\_female + \\ \beta_4 enrolled\_female + \beta_5 year + \beta_6 gdp + \varepsilon_i \end{aligned}$$

```
##
## Call:
## lm(formula = female_labor_rate ~ log_edu_exp + fertility_rate +
##     days_female + enrolled_female + year + gdp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.557  -5.780   1.823   8.567  35.365
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.716e+02  8.525e+01  -2.013   0.0442 *
## log_edu_exp      2.300e-01  8.032e-01   0.286   0.7746
## fertility_rate   3.698e+00  2.906e-01  12.724  < 2e-16 ***
## days_female     -2.301e-02  4.874e-03  -4.720  2.5e-06 ***
## enrolled_female  1.296e-01  1.359e-02   9.536  < 2e-16 ***
## year             1.027e-01  4.239e-02   2.423   0.0155 *
## gdp              1.135e-13  2.321e-13   0.489   0.6249
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.36 on 2288 degrees of freedom
##   (3413 observations deleted due to missingness)
## Multiple R-squared:  0.0785, Adjusted R-squared:  0.07609
## F-statistic: 32.49 on 6 and 2288 DF,  p-value: < 2.2e-16
```

After introducing other variables, the coefficient for logged education expenditure becomes positive, but become insignificant. All other variables besides GDP are significant at the 5% level. This model tells us that for every one unit increase in the fertility rate, we expect on average a 3.689 increase in the female labor rate keeping all else constant. However, we would normally expect the fertility rate to have a negative relationship with the female labor rate, as if more women are working, they will have less time to have children. It is possible that there is a non-linear relationship between the fertility rate and the female labor rate. Overall, this model shows improvements over Model 2 in terms of the R-squared value.

Model 4 :

To explore this potential this potential non-linear relationship between fertility rate and female labor rate, we can run bivariate regression with just fertility rate and female labor rate.

$$female\_labor\_rate_i = \beta_0 + \beta_1 fertility\_rate + \varepsilon_i$$

```
##
## Call:
## lm(formula = female_labor_rate ~ fertility_rate, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.690  -5.510   1.320   8.514  31.999
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     46.9243     0.4277  109.72   <2e-16 ***
## fertility_rate   1.5651     0.1308   11.96   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13.44 on 4531 degrees of freedom
##   (1175 observations deleted due to missingness)
## Multiple R-squared:  0.03062,    Adjusted R-squared:  0.03041
## F-statistic: 143.1 on 1 and 4531 DF,  p-value: < 2.2e-16
```

Fertility remains both positive and significant in the bivariate simple regression. We would expect the opposite for more developed countries. However, less developed countries will high fertility rates and high female labor participation because most labor would be agricultural in these countries. We can assume that the relationship between fertility rate and female labor force participation across countries is not linear. To capture this behavior we can introduce a quadratic term for fertility rate.

$$female\_labor\_rate_i = \beta_0 + \beta_1 fertility\_rate + \beta_2 fertility\_rate^2 + \varepsilon_i$$

```
##
## Call:
## lm(formula = female_labor_rate ~ fertility_rate + I(fertility_rate^2),
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.810  -5.626   0.922   7.511  34.321
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         62.67251    0.95738   65.46   <2e-16 ***
## fertility_rate      -9.55044    0.62263  -15.34   <2e-16 ***
## I(fertility_rate^2)  1.53359    0.08412   18.23   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 12.97 on 4530 degrees of freedom
##   (1175 observations deleted due to missingness)
## Multiple R-squared:  0.09689,    Adjusted R-squared:  0.09649
## F-statistic:   243 on 2 and 4530 DF,  p-value: < 2.2e-16
```

After introducing the quadratic term for fertility rate, the coefficient for the linear term becomes negative. This confirms our hypothesis that there is a non-linear relationship between fertility rate and female labor rate. We can finalize this model by introducing the quadratic term back to the original model with all of the other variables.

$$\begin{aligned} female\_labor\_rate_i = \beta_0 + \beta_1 edu\_exp + \beta_2 fertility\_rate + \beta_3 fertility\_rate^2 \\ + \beta_4 days\_female + \beta_5 enrolled\_female + \beta_6 year + \varepsilon_i \end{aligned}$$

```
##
## Call:
## lm(formula = female_labor_rate ~ edu_exp + fertility_rate + I(fertility_rate^2) +
##     days_female + enrolled_female + year, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.573  -6.120   1.557   7.843  38.791
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.025e+02  8.288e+01  -3.650 0.000268 ***
## edu_exp              6.133e-01  1.852e-01   3.311 0.000944 ***
## fertility_rate      -9.906e+00  1.095e+00  -9.042  < 2e-16 ***
## I(fertility_rate^2)  1.758e+00  1.369e-01  12.843  < 2e-16 ***
## days_female         -2.323e-02  4.667e-03  -4.978 6.91e-07 ***
## enrolled_female      5.927e-02  1.407e-02   4.211 2.64e-05 ***
## year                 1.781e-01  4.132e-02   4.311 1.69e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13.87 on 2288 degrees of freedom
##   (3413 observations deleted due to missingness)
## Multiple R-squared:  0.1404, Adjusted R-squared:  0.1382
## F-statistic:  62.3 on 6 and 2288 DF,  p-value: < 2.2e-16
```

After running some tests, we found that with the quadratic term introduced, the normal education expenditure variable performed better in the model. This means that the previous insignificance was simply caused by omitted variable bias. We also removed GDP from the model due to its consistent insignificance. The R-squared for this model is 0.1404 which is an improvement from Model 3.

According to Model 4, for every one unit increase in government expenditure on education, we expect on average a 0.6133 increase in the female labor rate keeping all else constant. The more governments spend on education, the more woman are willing to participate in the labor force. This is important to the economy as this means higher employment and a stronger more stable economy.

Fertility rate and female labor rates have a non-linear U-shaped relationship meaning that for more developed countries that already have low fertility rates, increasing the fertility rate will decrease the female labor rate. However, for less developed countries that already have high fertility rates, increasing the fertility rate will actually increase the female labor rate. This information can help governments decide on certain policies to improve their country's economy. This being said, it is important to keep in mind that fertility rate and female labor rate may not have a completely causal relationship. As countries become more developed and more urbanized, the fertility rate will naturally decrease as a side effect. There may be confounding variables between these two variables or even reverse causality.

For every additional day of maternal leave given, we expect on average a decrease of 0.02323 in the female labor rate. Maternal leave, while significant at the 5% level, may similarly not have a clear causal effect on the female labor rate. While we assume that having more maternal leave may encourage more women to enter the labor force, our model tells us otherwise. It is important to consider possible confounders. For example, it is possible that countries with high amounts of maternal leave are countries that are very progressive and already have very strong economies that do not require women to provide for their households. This is just one possible explanation for this interesting behavior that can be futher explored.

For every additional unit of female tertiary school enrollment, we expect on average an increase of 0.05927 in the female labor rate. This further supports the idea that investing in education will increase the female labor rate. If more women have higher education credentials, more women will be willing to enter the labor force.

Conclusions

Based on the analysis of our models, we find that higher government expenditure on education is associated with higher levels of female labor force participation. We also found that the fertility rate has a non-linear U-shaped relationship with the female labor force participation rate. This means the the potential effects of fertility rate on the female labor force participation rate will vary for different countries. We also found that higher amounts of maternal leave is associated with lower amounts of female labor force participation and higher amounts of female tertiary school enrollment is associated with higher amounts of female labor force participation. There is a lot of additional work that can be done to further explore the relationships between these variables. This includes and is not limited to, introducing interaction terms, utilizing time-fixed effects and entity-fixed effects, as well as exploring additional variables that may provide more information on how the female labor force participation rate behaves. Overall, we were able to establish relatioships between select variables and the female labor force participation rate. Policy makers can use this information to make decisions that promote development and stability, increase access to education, and improve gender equality.