Where are our women?

Exploring the shocking decrease in Female Labour Force Participation Rates

Course: EPA1315 – Data Analytics and Visualization

Masters Program: Engineering and Policy Analysis

University: Technical University of Delft

Authors: Aarthi Meenakshi Sundaram (4995813), Madhumita Naik (5043646)

Code: https://github.com/aarsundaram/dav_final/blob/master/DavFinalProjMD.Rmd

Summary

Gender equality has come a long way over the past few centuries. With the inclusion of gender equality as one of the sustainable development goals, more countries are paying attention to this goal. However, despite several international and national policies implemented to bolster gender equality, women are still a long way away from achieving complete equality in all areas – for example, political participation, labour force participation, property rights, etc.

In this report, we consider the matter of the working woman – why is it important that more woman join the work force of a country, and what are the factors that influence a woman to find profitable employment? It would be apt to include the inspiration behind this report - a quote by Annabel Crabb from her book "The Wife Drought", at this point:

"The obligation for working mothers is a precise one: the feeling that one ought to work as if one did not have children, while raising one's children as if one did not have a job."

Since we are considering the country-wise participation of women in the economic arena, the report focuses on the interaction of the female labour force participation rate (FLFPR) with Gross Domestic Product (GDP), among other indicators. Childbirth and child-rearing are important determining factor for a woman to work, hence fertility is also considered. The level of education will also be considered (since common sense dictates that an educated woman is more likely to take up an earning role) – this will be considered through gross enrolment ratio (GER) in secondary and tertiary schools. We will also be considering the proportion of seats held by women in the parliament – as a higher number of women in a position to influence policy implies that there would be more policies to support the inclusion of women in the workforce.

We have carried out the above analysis dividing countries over national income (high-income vs low-income). This division is required as high-income countries are mostly western, while low-income countries are developing countries from the east. This division highlights the cultural and economic differences that have developed in these countries throughout history. For example, the suffrage movement started in western countries in the late-19th century, while low-income countries are going through slow but steady cultural changes moving towards gender equality starting from just a few decades ago.

With the help of the collider model, we have tested the hypothesis that the labour force participation rate for women over the age of 15 is determined by GDP, fertility, GER_Secondary, GER_Tertiary, and proportion of women holding parliamentary seats.

The collider used

i. For high-income countries - 4 betas and 4 sigma's
ii. For low-income countries – 5 betas and 4 sigma's

which had a prior of a normal distribution between 0 and 0.01 for the different betas, and a uniform distribution between 0.01 and 0.99 for the variances, which when squared gave the sigma's.

From this model it became clear that correlation between different variables exists, but not a linear relation. The model is only designed to identify correlations between different variables, but this also means that a definitive conclusion cannot be reached on the exact influencing factors for FLFPS. Further research could focus on developing a model to check possible second-degree correlation between the aforementioned variables.

Table of Contents

1. Introduction

According to data from the World Bank, FLFPR over the world has declined over the past two decades:

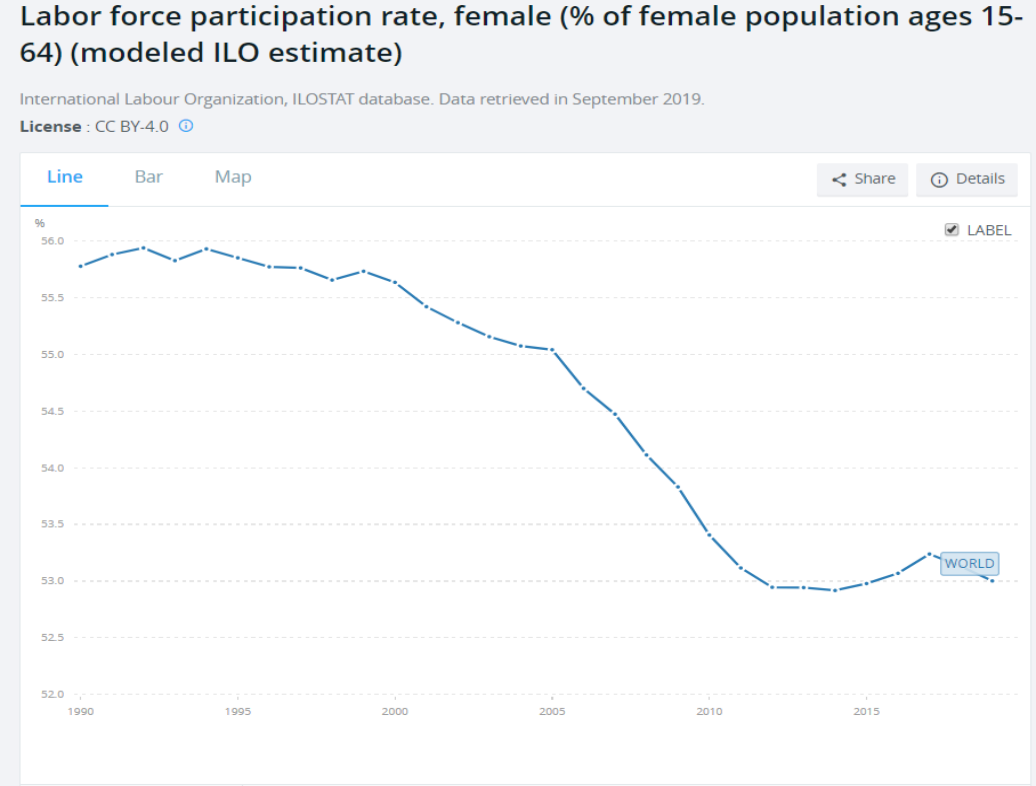


Fig. 1. Female Labor Force Participation Rate over time
Picture Credit: https://data.worldbank.org/indicator/sl.tlf.acti.fe.zs?end=2019&start=1990&view=chart

However, carrying out our own analysis over high-income and low-income countries, this is how the change in FLFPR differs between the two (with respect to GDP):

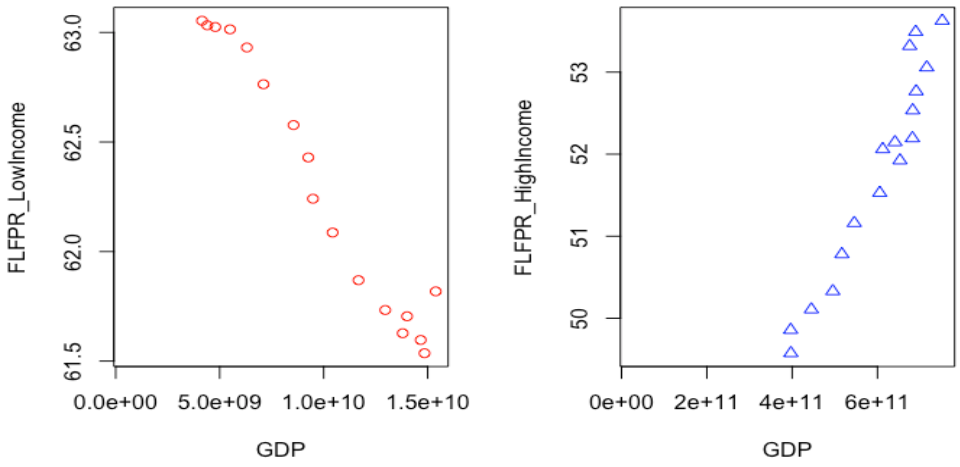


Fig.2 FLFPR vs GDP in low-income and high-income countries

As per the outline for classification of countries by the UN, countries are classified by their level of development, into high-income, upper-middle income, lower-middle income and low-income on the basis of per capita gross national income (GNI). Low income countries are those with GNI < $1,035 and high-income countries are those with GNI > $12,615.

This evidence merits an analysis into factors that govern the change of FLFPR in a different manner in high-income vs low-income countries. The above graphic shows that analysis of FLFPR on a global scale hides nuances of governing factors over countries, and paints an incorrect picture that fewer women all over the world are joining the workforce.

To further this question, we used data from UN Gender Stats and World Bank Open Data sources. The World Bank has freely available data over a wide range of indicators. The choice of different indicators that influence FLFPR has been discussed in a later chapter.

The research question guiding this paper is

"*What factors influence the participation of women in the labour force of a country?*"

The report follows the CRISP-DM process. The report starts with a Business Understanding section in chapter 2. Following this are the Data Collection, Data Exploration, and Data Quality sections in chapter 3. Chapter 4 is devoted to Data Preparation. Chapter 5 deals with the model created – Test Design, Model Options, Parameter Setting and Final Model Description, ending with an Overall assessment of the report.

2. Business Understanding

As per the International Labour Organization, the labour force participation rate is calculated as the proportion of country's working-age population that engages actively in the labour market, either by working or by looking for work. This provides an indicator to the size of the labour supply available to engage in economic activities such as production of goods and services.

The breakdown of this indicator by age and sex gives a more detailed profile of the distribution of the labour force of a country.

The female labour force participation rate, hence, describes the proportion of a country's working-age population, that is female, and above the age of 15, that is actively engaged in the labour market.

2.1 Issues

Until now, gender statistics have been region-neutral – they show LFPR (Labour Force Participation Rate), wage gaps, etc. country-wise or region-wise. There is little attention paid to causal relationships behind these trends.

If there is a huge difference in the manner in which fertility rates and labor force participation rates are related in the Sub-Saharan Africa, little effort is spent in understanding the reason for the same. Exploring causal relationships, if any, into how women belonging to different regions are responding to country-wide policies that affect economic growth, maternity leave, etc. can offer insights on how to improve labor force participation rates by region.

2.2 Limitations

Labour force data is usually obtained from population censuses that are based on a limited set of questions. The scope is limited and offers little opportunity for probing. Hence this data might not be consistent with surveys specific to labour force data. Consequently, this data will vary over countries depending on the number and type of questions asked in surveys that generate the data. Many times, this data also leaves out people that are employed in small businesses or in the informal economy.

There are also differences between countries on what constitutes the labour force – for example, inclusion or exclusion of military conscripts, workers that are part of the family, and unemployed people who are looking for work.

Even the limits of age for the working force differ between countries – where some opt a non-standard upper age limit for employed people. As an extension of this, there will always be a certain portion of the population beyond the upper age limit that is still working too.

2.3 Trends

Below is the analysis of the change of the various indicators used with respect to FLFPR. The time frame chosen is 2000-present, because this time frame has the most consistent set of data for all the concerned variables, over 110 countries (84 high-income countries, and 36 low-income countries). This also overlaps with the worldwide drop of FLFPR.

a. FLFPR over the last two decades

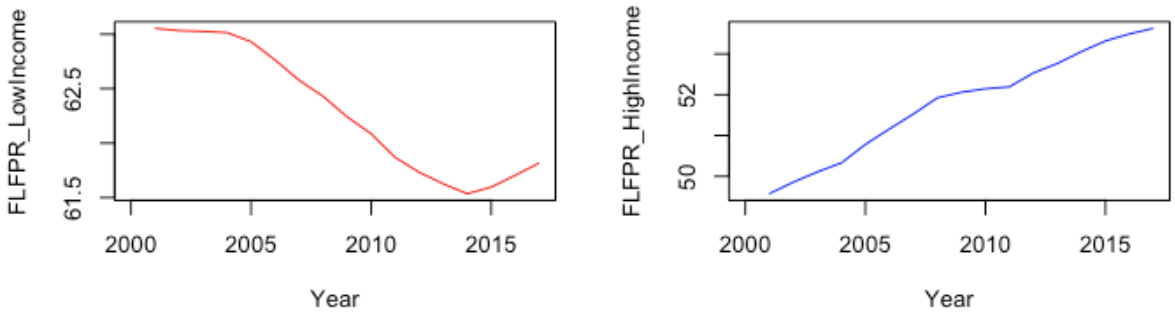


Fig.3 FLFPR changes over time in low-income and high-income countries

Over the last two decades, FLFPR has consistently dropped in low income countries, while it has increased consistently in high income countries.

b. FLFPR vs Gross Domestic Product

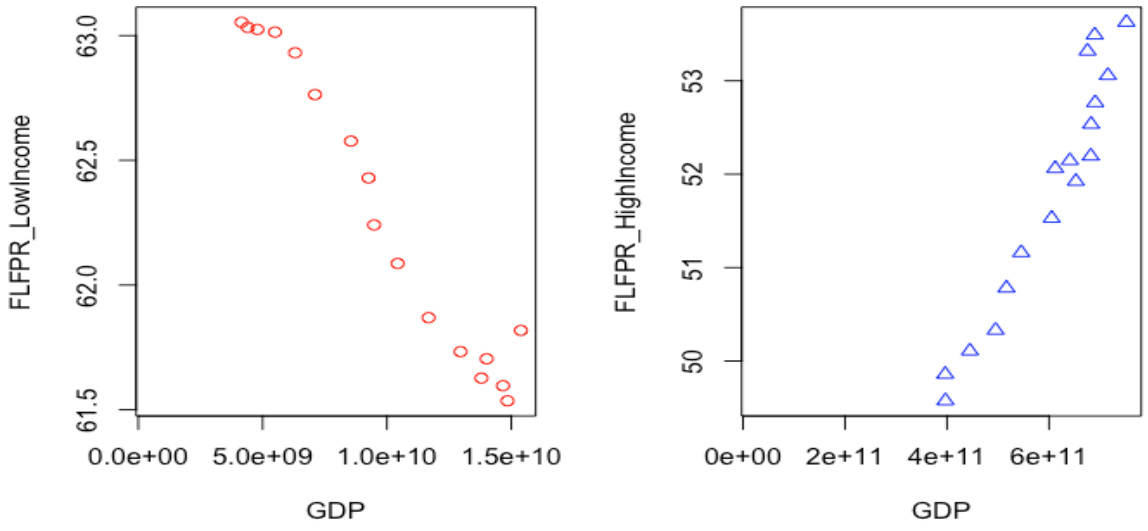


Fig. 4 FLFPR vs GDP in low-income and high-income countries

Mapping the change of FLFPR with the average GDP of low-income and high-income countries shows that

a) while GDP has increased in low-income countries, FLFPR has decreased
b) while GDP has increased in high-income countries, FLFPR has increased

c. FLFPR vs Female Seat Proportion in the National Government

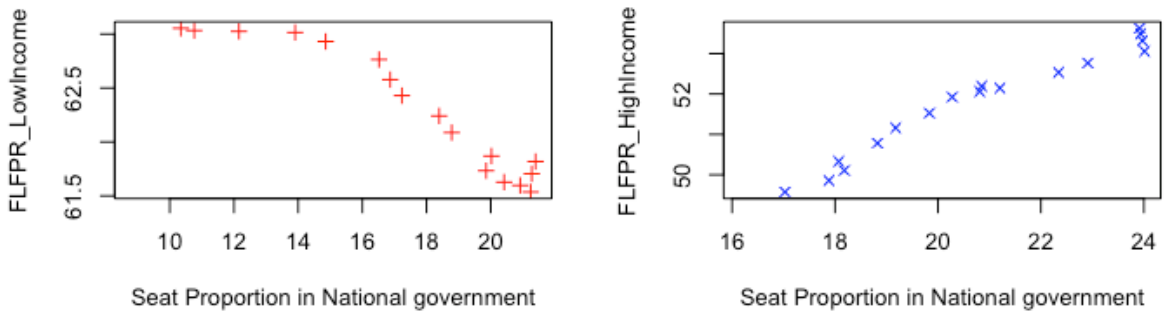


Fig.5 FLFPR vs Female Seat Proportion in the National Government for low-income and high-income countries

From the above graphic, it is evident that while the number of seats allotted in the national government has increased in both low-income and high-income countries, this rise has been accompanied by a drop in FLFPR in low-income countries but an increase in FLFPR for high-income countries.

d. FLFPR vs Gross Enrolment Ratio in Secondary Schools

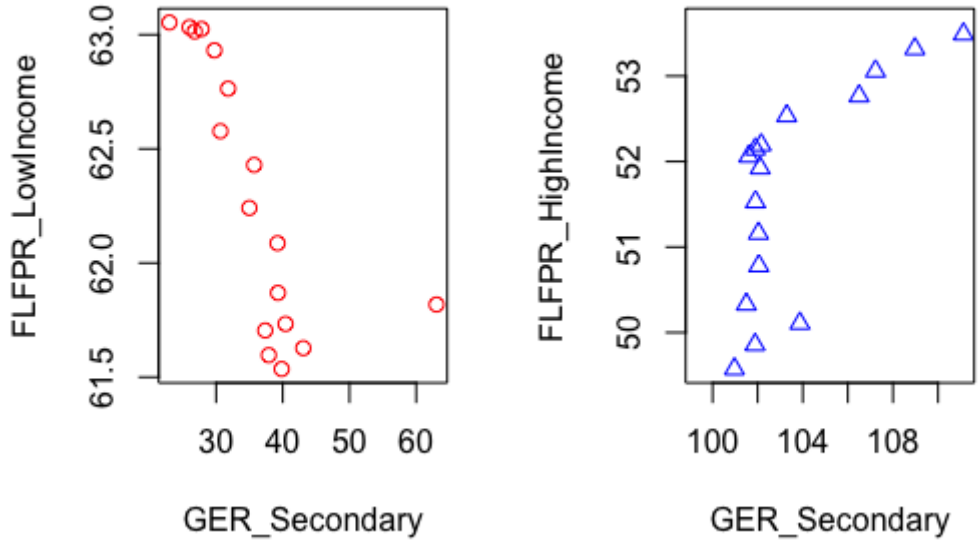


Fig.6 FLFPR vs Gross Enrolment Ratio in Secondary schools for low-income and high-income countries

The above graphic highlights the below:

a. While Gross Enrolment Ratio in secondary schools has increased in low-income countries, FLFPR has decreased in the same time frame
b. While Gross Enrolment Ratio in secondary schools has increased in high-income countries, FLFPR has decreased in the same time frame

e. FLFPR vs GER_Tertiary

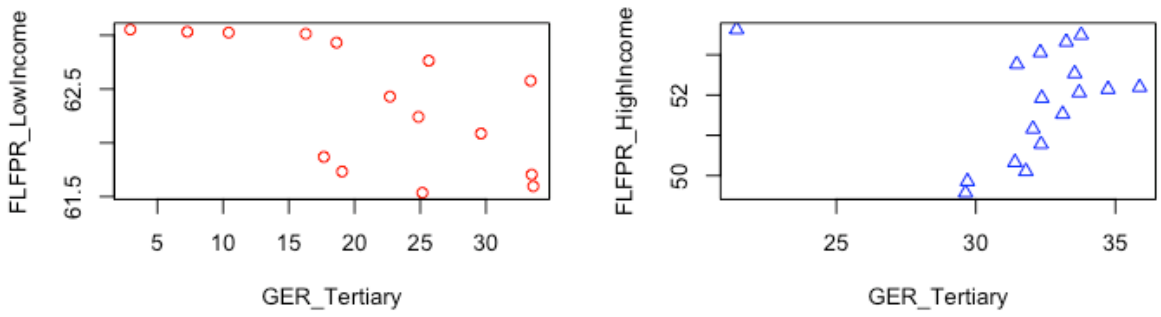


Fig.7 FLFPR vs Gross Enrolment Ratio in Tertiary Schools for low-income and high-income countries

The above graphic highlights the below:

a. While Gross Enrolment Ratio in tertiary schools has increased in low-income countries, FLFPR has decreased in the same time frame
b. While Gross Enrolment Ratio in tertiary schools has increased in high-income countries, FLFPR has decreased in the same time frame

f. FLFPR vs Fertility Rate

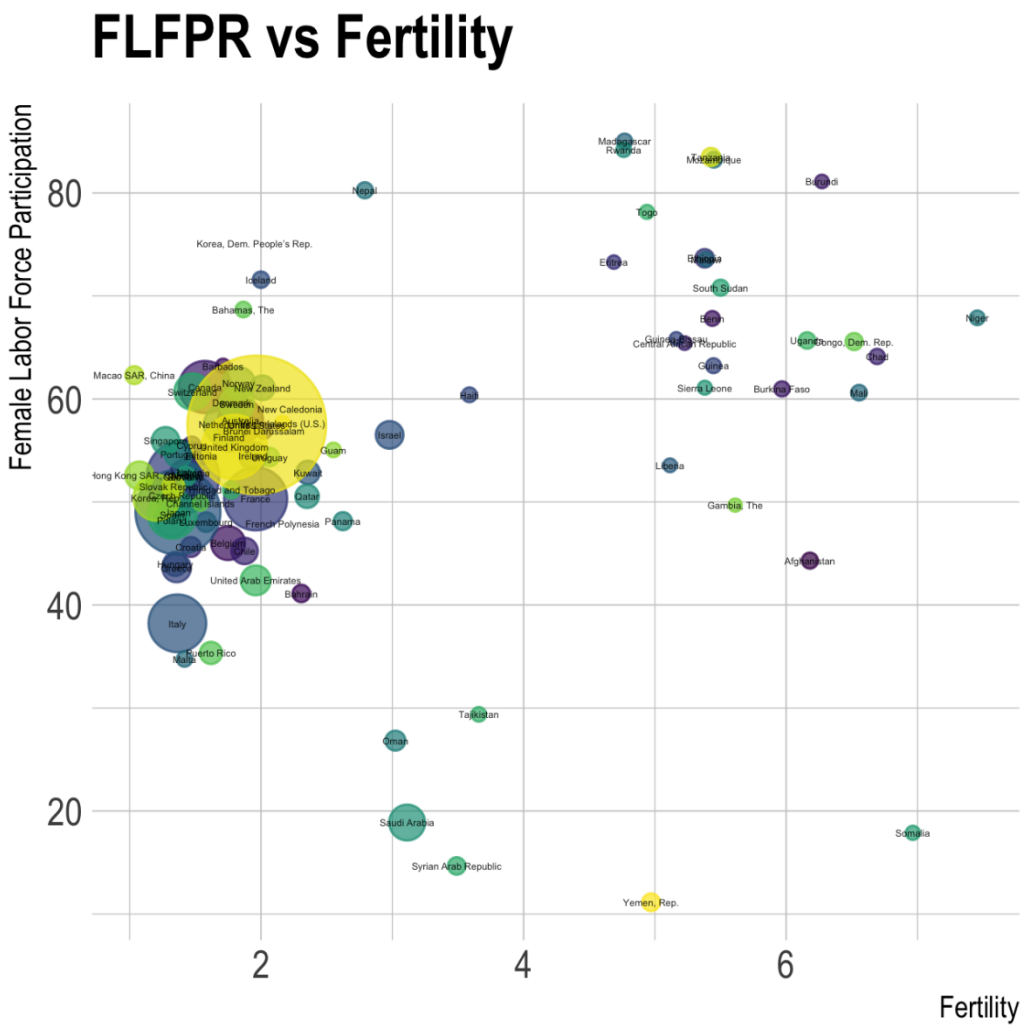


Fig. 8 FLFPR vs Fertility rate for both low-income and high-income countries

From the above bubble plot, we can see that high GDP countries correspond to low fertility rates and high FLFPR (high GDP corresponds to larger circular area).

Similarly, low-GDP countries correspond to high fertility rates and high FLFPR (low GDP corresponds to smaller circular area).

Hence, we can conclude that there is a relatively high correlation between FLFPR and fertility rate, the pattern being heavily influenced by the relative income of the country.

2.3.1 Conclusions regarding Trends:

Overall, it is evident that while all the contributing factors to labour force participation rate have increased (i.e GDP, GER_Tertiary, GER_Secondary, Female Seat Proportion in National Government), FLFPR has consistently decreased in low-income countries, and consistently increased in high-income countries.

2.4 Relevance for Organization for Economic Co-operation and Development (OECD)

This subject is of importance to the strategic counsellor of the OECD as direct consequence of Sustainability Development Goal #5. SDG #5 aims to achieve gender equality and empower all women and girls. This is in direct consequence to the above goal.

It is especially relevant to address the topic of gender equality over countries with differing incomes, as we can see drastically different trends between low-income and high-income countries in the change of FLFPR over the last two decades.

This also counts as a step closer to SDG #8 – full and productive employment and decent work for all, SDG #1 – ending poverty, SDG #2 – food security and SDG #10 – reducing inequalities.

According to UN Women, increasing employment rates of women in just the OECD countries to match that of Sweden, would increase the worldwide GDP by over 6 trillion USD, which is directly relevant to the core interest of OECD – economic development.

If we can identify the contributing factors to the above trends, the OECD can better target the root cause of these patterns and improve gender equality with the FLFPR indicator.

3. Data Understanding

3.1 Initial Data Collection

For the research behind this project, open data was collected from UN Gender Stats and the World Bank open database.

From UN Gender Stats the following data was collected – Seat Proportion of Women in the National Government, Gross Enrolment Ratio in Secondary and Tertiary Schools, GDP, Maternity Leave Length and Fertility Rate.

FLFPR over countries of different Gross National Incomes was collected from the World Bank Open Database.

More data was collected than was eventually used in the model – for example, Gross National Income for low- and high-income countries was collected. However, this corresponds closely with GDP, and was eventually dropped as it added no more value to the model than just inclusion on GDP.

3.2 Data Description

The data obtained shows the following parameters:

a. **Female Labour Force Participation Rate**
   This indicator shows the female labour force participation percentage (with respect to working population). The women considered are above the age of 15.

b. **Seat Proportion in National Government**
   This indicator shows the percentage of parliamentary seats in a single or lower chamber held by women. Based on the data collected, this indicator is directly correlated to FLFPR.

c. **Gross Enrolment Ratio in Secondary Schools**
   As per UNESCO, this indicator refers to the number of students enrolled in secondary school, regardless of age, expressed as a percentage of the official number of students attending secondary school.
   For low-income countries, there is a weak correlation between this indicator and FLFPR.
   For high-income countries, there is very low correlation between this indicator and FLFPR.

d. **Gross Enrolment Ratio in Tertiary Schools**
   As per UNESCO, this indicator refers to the number of students enrolled in tertiary school, regardless of age, expressed as a percentage of the official number of students attending tertiary school.
   For low-income countries, there is a weak correlation between this indicator and FLFPR.
   For high-income countries, there is no correlation between this indicator and FLFPR.

e. **Gross Domestic Product**
   GDP is the monetary value of all finished goods and services produced within a country

during a specific period – here, annually. GDP is strong indicator as to the economic health of a country.
There is a high correlation between GDP and FLFPR for all countries, hence this indicator has been included.

f. Fertility Rate

As per OECD, "the total fertility rate in a specific year is defined as the total number of children that would be born to each woman if she were to live to the end of her child-bearing years and give birth to children in alignment with the prevailing age-specific fertility rates."
There is a high correlation between fertility rate and FLFPR for all countries. Hence this indicator has been included.

g. Maternity Leave Length

Originally, this indicator was considered was considered. However, this indicator shows absolutely no correlation with FLFPR (detailed further in the Data Exploration section). Hence this was indicator was not included in the final model.

3.3 Data Quality

The quality of data depends on the number of missing data points. There were several instances where data points were not available for many countries. The final amount of data available depends on the year, and the specific country.

To account for this, data from the past two decades, i.e., 2000-present was chosen. This time frame has the most consistent data for all the low-income and high-income countries. This also overlaps with the considerations of FLFPR's drop worldwide.

3.4 Exploratory Analysis

For the purposes of data exploration, a correlation matrix of the normalized variables was created. This makes any correlation between the variables clear.

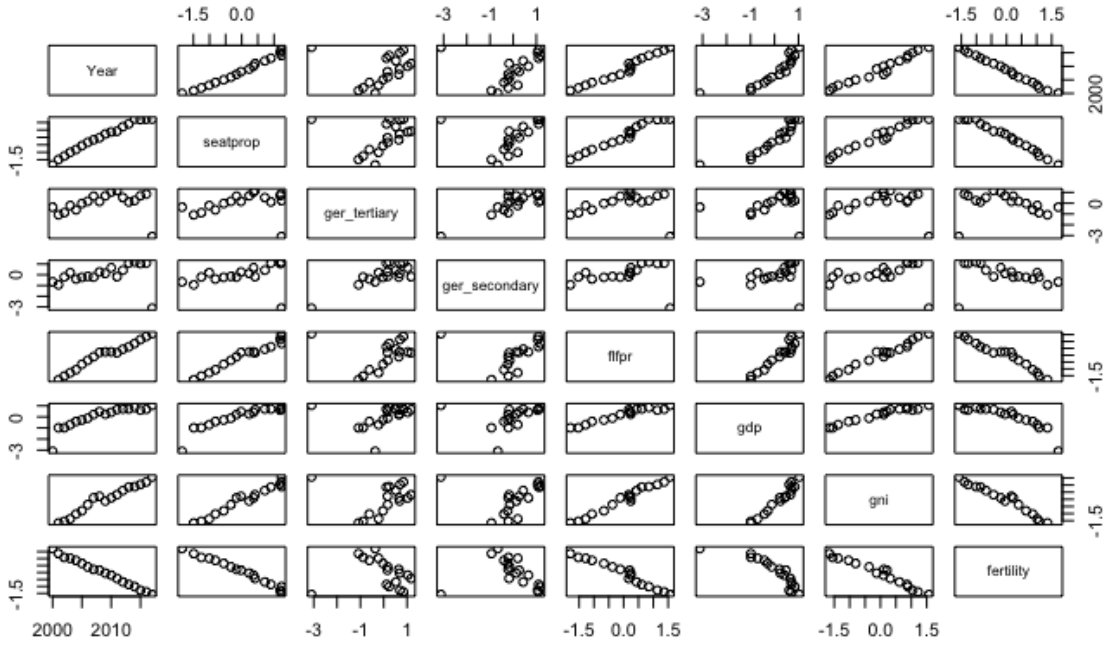


Fig. 9 Correlation Plot for all relevant variables (normalized by population)

To make the extent of the correlation evident, the below correlation plots were created. Dark colors show high levels of correlation. The extent of blue shows a positive correlation (i.e., a is directly proportional to b); the extent of red shows a negative correlation (i.e., a is inversely proportional to b).

3.4.1 Correlation Matrix for High Income Countries

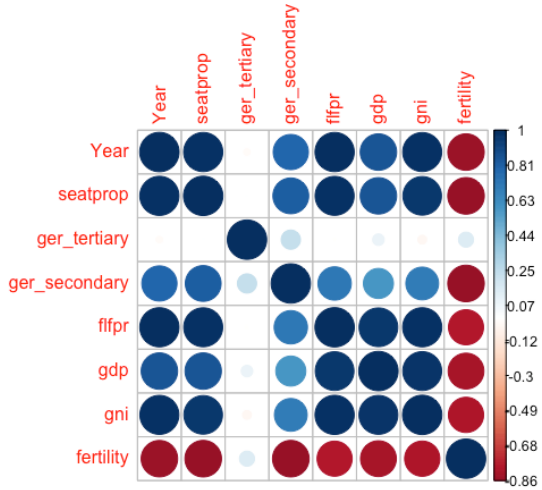


Fig. 10 Correlation plot for High-Income Countries

From the above figure, it is clear that for high income countries, there is a positive correlation between FLFPR and the following variables: Female Seat Proportions in the National Government, GDP, and GNI. The correlation between GER (both secondary and tertiary) is low or non-existent. There exists a negative correlation between Fertility Rate and all other variables.

For further clarity, below is the table that shows the actual correlation values for all the variables:

```
Correlation for High Income
                Year seatprop ger_tertiary ger_secondary flfpr  gdp   gni
Year            1.00     0.99        -0.03          0.80  0.99 0.85  0.98
seatprop        0.99     1.00         0.00          0.83  0.98 0.85  0.96
ger_tertiary   -0.03     0.00         1.00          0.24  0.00 0.08 -0.05
ger_secondary   0.80     0.83         0.24          1.00  0.72 0.58  0.69
flfpr           0.99     0.98         0.00          0.72  1.00 0.97  0.99
gdp             0.85     0.85         0.08          0.58  0.97 1.00  0.97
gni             0.98     0.96        -0.05          0.69  0.99 0.97  1.00
```

3.4.2 Correlation Matrix for Low Income Countries

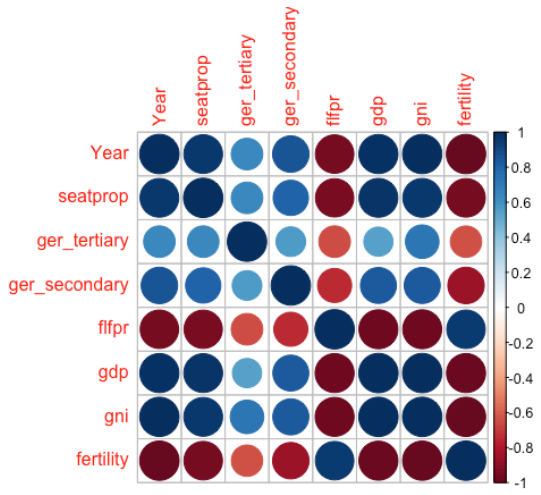


Fig. 11 Correlation plot for Low-Income Countries

From the above figure, it is clear that for low-income countries, there is a negative correlation between FLFPR and the following variables: Female Seat Proportions in the National Government, Gross Enrolment Ratio at Secondary and Tertiary Levels, GDP and GNI. In contrast with high-income countries, FLFPR is positively correlated with Fertility Rate.

The extent of correlation between FLFPR and GER (at both Secondary and Tertiary levels) is weak in comparison with the rest of the variables. The extent of correlation is more than in High-income countries. For further clarity, below is the table that shows the actual correlation values for all the variables:

```
Correlation for Low Income
                Year seatprop ger_tertiary ger_secondary flfpr   gdp   gni
Year            1.00     0.97         0.64          0.86 -0.95  0.98  0.99
seatprop        0.97     1.00         0.64          0.80 -0.94  0.97  0.97
ger_tertiary    0.64     0.64         1.00          0.57 -0.64  0.55  0.73
ger_secondary   0.86     0.80         0.57          1.00 -0.75  0.83  0.84
flfpr          -0.95    -0.94        -0.64         -0.75  1.00 -0.98 -0.98
gdp             0.98     0.97         0.55          0.83 -0.98  1.00  1.00
gni             0.99     0.97         0.73          0.84 -0.98  1.00  1.00
```

3.4.3 Other Considerations/Observations

Considering the above matrices, the rows for both GDP and GNI show similar correlations with all variables. Hence the inclusion of one of these is sufficient. For further analyses, GDP was chosen.

a. **Relationship between FLFPR and Maternity Leave Length**
   Originally, Maternity Leave length was considered for possible correlation with FLFPR. The below graphic, however, shows that this indicator has no correlation with FLFPR, and hence this indicator was no longer considered.

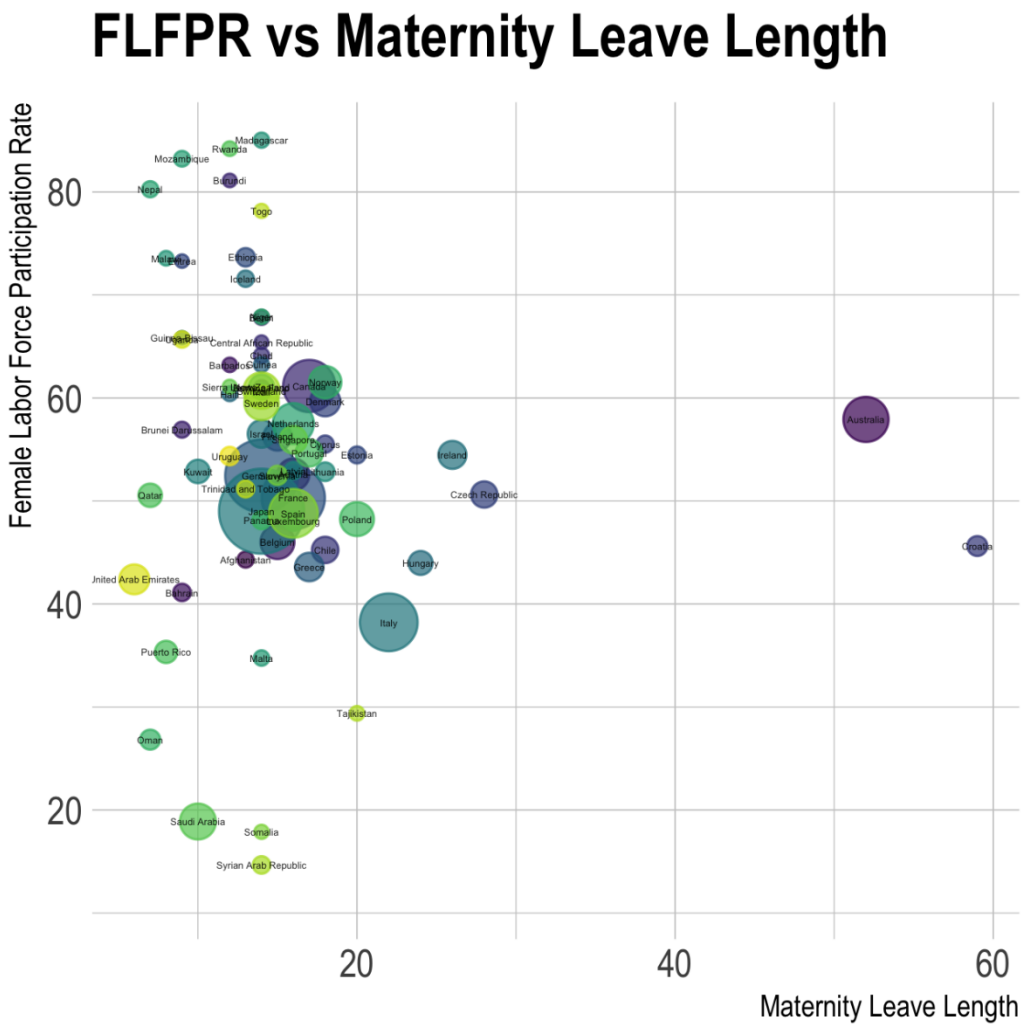


Fig. 12 FLFPR vs Maternity Leave Length for Low-Income and High-Income countries

b. **Relationship between Income and Fertility Rates**

As an interesting observation, the below graphic shows the link between fertility and income. High-income countries(blue) show low fertility rates, while low-income countries(red) show high-fertility rates.

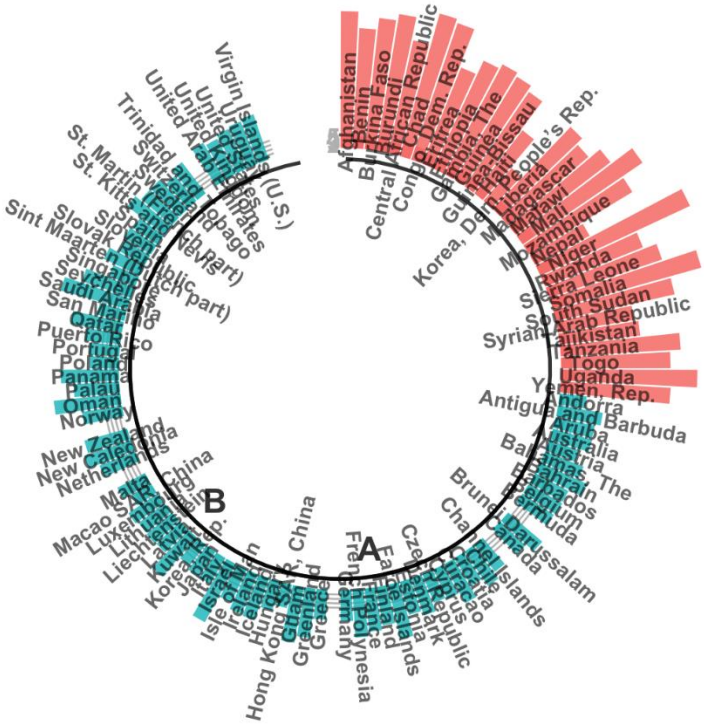


Fig.13 Relationship between Fertility and Income

3.4.4 Causal Diagram

With the help of the correlation matrices from 3.4.1 and 3.4.2, the below causal diagram can be created:

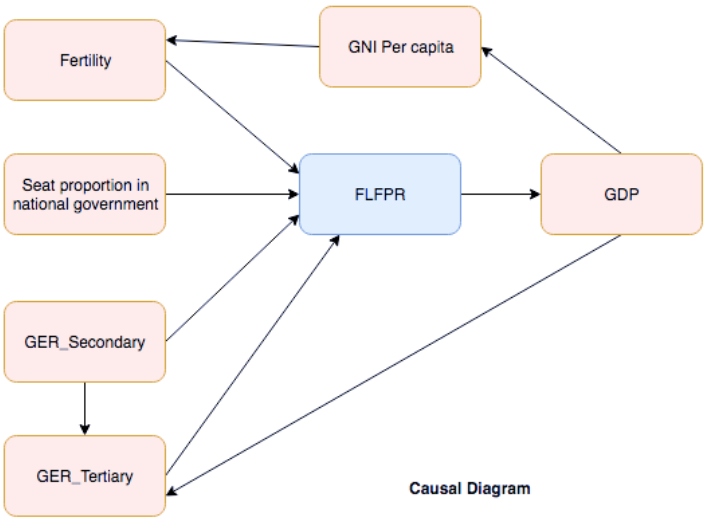


Fig. 14 Causal Diagram for relevant indicators

4. Data Preparation

To prepare data for further analysis, it should be cleaned and standardised. This has been done by renaming columns, dropping columns, reshaping data using reshape2 packages' melt and dcast functions. Challenges faced with handling different datasets:

Each of the variables we looked at for our analysis- Seat Proportion, GER_secondary, GER_Tertiary, FLFPR, GDP, GNI, Population, Maternity Leave Length, Proportion of wages paid during maternity, Fertility came from two different sources

1) **UNDP data:** fertility, maternity leave, wages paid, FLFPR. (Sample header table for Maternity Leave Length shown below)

	Indicator.Name <fctr>	Region <fctr>	Country.Code <int>	Country <fctr>	Measure.Name <fctr>	Measure.Value <fctr>
1	Length of maternity leave	Southern Asia	4	Afghanistan	Maternity leave duration	Less than 14 weeks
2	Length of maternity leave	Southern Asia	50	Bangladesh	Maternity leave duration	At least 14 weeks
3	Length of maternity leave	Southern Asia	144	Sri Lanka	Maternity leave duration	Less than 14 weeks
4	Length of maternity leave	Southern Asia	356	India	Maternity leave duration	Less than 14 weeks
5	Length of maternity leave	Southern Asia	364	Iran (Islamic Republic of)	Maternity leave duration	Less than 14 weeks
6	Length of maternity leave	Southern Asia	524	Nepal	Maternity leave duration	Less than 14 weeks

6 rows | 1-7 of 9 columns

It contains text in footnotes and even within the Measure.Value on some occasions:

Footnote <fctr>
Entitlement: A female employee is entitled to 90 days of paid maternity leave. Thirty days of the leave is granted before delivery and the rest (60 days) is granted to her after deli...
Compulsory: Employers are prohibited from knowingly employing a woman in his establishment and women are prohibited from working in any establishment during the eight ...
Compulsory: 2 weeks up to including the day of her confinement and 10 weeks immediately following that day if the confinement results in the issue of a live child and such wo...
Compulsory: No woman shall work in any establishment during the 6 weeks immediately following the day of her delivery or miscarriage. Entitlement: 12 weeks being the 6 wee...
Compulsory: No compulsory leave requirements identified. However the Labour Code stipulates that 45 days of the 90 day maternity leave entitlement shall be taken if possible ...
Compulsory: No compulsory leave provisions identified. Entitlement: 52 days.

6 rows | 10-10 of 9 columns

2) **World Bank Open Data :** Seat Proportion, GER_secondary, GER_Tertiary, FLFPR, GDP, GNI, Population. (Sample header table for GDP shown below)

	Country.Name <fctr>	Country.Code <fctr>	Indicator.Name <fctr>	Indicator.Code <fctr>	X1960 <dbl>	X1961 <dbl>	X1962 <dbl>	X1963 <dbl>	X1964 <dbl>
1	Aruba	ABW	GDP (current US$)	NY.GDP.MKTP.CD	NA	NA	NA	NA	NA
2	Afghanistan	AFG	GDP (current US$)	NY.GDP.MKTP.CD	537777811	548888896	546666678	751111191	8e+08
3	Angola	AGO	GDP (current US$)	NY.GDP.MKTP.CD	NA	NA	NA	NA	NA
4	Albania	ALB	GDP (current US$)	NY.GDP.MKTP.CD	NA	NA	NA	NA	NA
5	Andorra	AND	GDP (current US$)	NY.GDP.MKTP.CD	NA	NA	NA	NA	NA
6	Arab World	ARB	GDP (current US$)	NY.GDP.MKTP.CD	NA	NA	NA	NA	NA

6 rows | 1-10 of 65 columns

UNDP data had data accumulated till that year. Meaning the data was available for one year.

Whereas World Bank Open Data has all the indicators in multiple columns for years with countries by rows. Many of the data for many countries were missing, for several years.

We took two outputs of final cleaned data:

a) Country wise mean data for all 10 variables over span of 2000 to 2017 (Fig. 15)

b) Year Wise mean data for all 10 variables over all countries (aggregated to high-income and low-income) (Fig. 16)

Country	seatprop	ger_tertiary	ger_secondary	molten.flfpr	molten.gdp	molten.gni	molten.pop	ml_wagespaid1	ml_length1	fertility
Afghanistan	27.6062667	*NA*	38.735111	44.30929	1.258717e+10	1783.3333	27728613.72	100.0	13	6.180722
Andorra	28.9682556	13.584444	*NA*	*NA*	3.079125e+09	*NA*	75223.06	100.0	16	1.226000
Antigua and Barbuda	9.1617944	53.282500	94.866296	*NA*	1.163658e+09	20081.7647	84716.61	60.0	13	2.051222
Australia	25.4493833	31.467500	155.811667	57.92047	9.538835e+11	37512.9412	21141260.72	100.0	52	1.843444
Austria	30.2064556	25.770000	99.342353	52.57359	3.415687e+11	41512.3529	8295005.78	100.0	16	1.425000
Bahrain	4.3333333	47.560000	96.142708	41.12177	2.339184e+10	38088.8235	1063814.56	100.0	9	2.306833
Barbados	12.3809556	40.500000	105.518444	63.21353	4.231656e+09	14957.6471	277520.00	100.0	12	1.711111
Belgium	34.1481333	25.501667	157.695833	46.01041	4.139512e+11	39241.7647	10728826.50	75.0	15	1.748889
Benin	8.0321222	31.226667	38.526296	67.81700	6.247003e+09	1777.0588	8648546.78	100.0	14	5.437389
Brunei Darussalam	9.0909000	47.906429	95.325098	56.91953	1.218717e+10	77921.1765	367200.33	100.0	9	2.022778
Burkina Faso	12.6617000	17.185000	20.751146	60.97782	7.722524e+09	1340.5882	14719642.61	100.0	14	5.967722
Burundi	26.4067278	16.856667	23.800000	81.11329	1.762930e+09	685.2941	8175931.44	100.0	12	6.272222
Canada	22.5098611	*NA*	105.340000	61.16418	1.404504e+12	38795.8824	32799229.39	55.0	17	1.572794
Central African Republic	9.5374333	*NA*	14.632778	65.41706	1.633192e+09	814.1176	4089619.06	50.0	14	5.226611

Fig. 15 Structure Data by Country

	Year	seatprop	ger_tertiary	ger_secondary	flfpr	gdp	gni	fertility
1	2000	13.73136	29.95375	78.96486	NaN	17448311800	NaN	3.155015
2	2001	14.83150	27.72500	77.18401	54.26909	285355050338	19011.79	3.099015
3	2002	15.54243	28.33939	81.92219	54.44643	285648678117	19462.14	3.055683
4	2003	16.28663	30.41516	84.53749	54.60666	320454921647	20167.98	3.046062
5	2004	16.86382	29.17853	80.67869	54.74783	356778712430	21534.40	3.023098
6	2005	17.65109	31.00774	81.71865	55.01283	373698049822	22463.65	2.988700
7	2006	18.32762	31.63290	82.28339	55.20180	394837345646	24221.88	2.949722

Fig. 16 Structured Data by Year

To finalize our data preparation, the data is normalized which is necessary at the stage where we design and run our models. Overview of normalized table for one case showing first 7 entries out of 18 (years)

	Year	seatprop	ger_tertiary	ger_secondary	flfpr	gdp	gni	fertility
1	2000	-1.85133959	-0.36630568	-0.6616330	NaN	-3.1151418	NaN	1.72057920
2	2001	-1.48878739	-1.07959146	-0.9360393	-1.7935222	-0.9859043	-1.65898954	1.36029399
3	2002	-1.25450003	-0.88296175	-0.2059447	-1.5242469	-0.9835707	-1.54771326	1.08150737
4	2003	-1.00924947	-0.21863631	0.1970392	-1.2809385	-0.7069417	-1.37331276	1.01960926
5	2004	-0.81903754	-0.61440610	-0.3975532	-1.0665836	-0.4182518	-1.03568940	0.87186713
6	2005	-0.55958998	-0.02898769	-0.2373088	-0.6641983	-0.2837823	-0.80608800	0.65056234
7	2006	-0.33663881	0.17108798	-0.1502895	-0.3772644	-0.1157739	-0.37165526	0.39979240

Fig. 17 Normalized Data

Our final structured data contains the data for all variables meaned by year or country as required for data/plotting purposes and looked at 110 countries (84 high income and 36 low-income countries) over a span of 18 years.

Cleaning data was done using a combination of replacing by mean or deleting NA values when they filled an entire row. The following workflow was kept in mind.

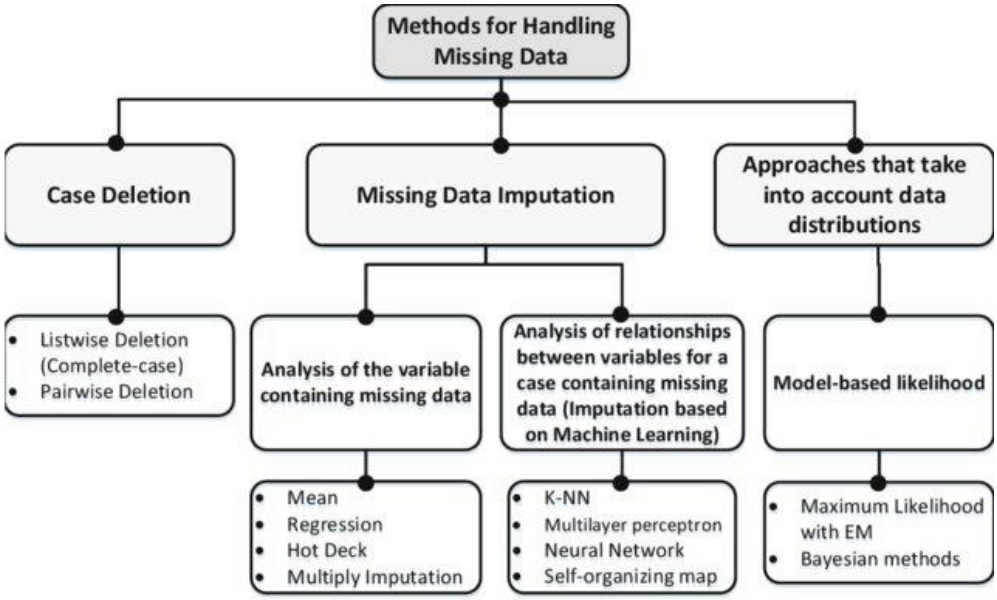


Fig. 18: Dealing with missing data

Picture Credits: (Skarga-Bandurova, I., Biloborodova, T., & Dyachenko, Y. (2018). Strategy to Managing Mixed Datasets with Missing Items. Communications in Computer and Information Science Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations, 608–620. doi: 10.1007/978-3-319-91476-3_50)

5. Modelling

5.1 Test Design

The model will be evaluated by looking at the different variables and the proposed correlations between the variables. As was visible in the correlation matrix, most correlations resemble linear connections, which is the connection which will be looked for by the model.

5.2 Models

There are three possible models that can be used to explore data – chain, collider, fork. These options are considered in an elementary fashion, as that is sufficient for the level of analysis in this report. In reality, it is probable that none of these models are followed exactly. However, the use of these models helps identify interconnections in variables.

5.2.1 Chain

A chain is a model of the type where there exists a chain-like correlation between variables –

A → B → C – that is, A influences B, which in turn influences C.

Fig. 19 Chain Model Simplified

5.2.2 Fork

In this type of model, a single variable influences multiple variables. An example of a fork model would be that marine pollution affects economies worldwide.

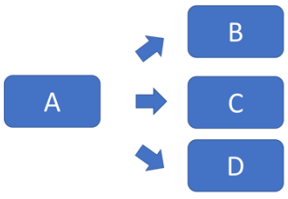


Fig. 20 Fork Model simplified

5.2.3 Collider (Chosen Model)

This is the model that we are considering for the report. In this case, several variables influence one primary variable (as is evident from the causal diagram).

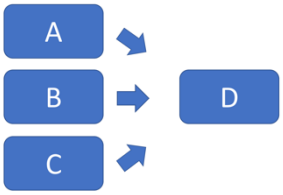


Fig. 21 Collider Model Simplified

Below are the collider models chosen for the problem statement:

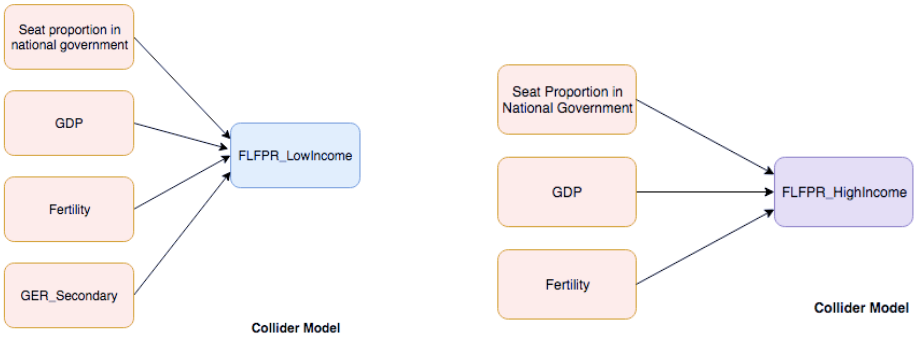


Fig. 22 Collider Models for Low-Income and High-Income countries

The below model was also considered, but rejected as we want to examine what affects FLFPR primarily. The impact of FLFPR on GDP is a different research question entirely.

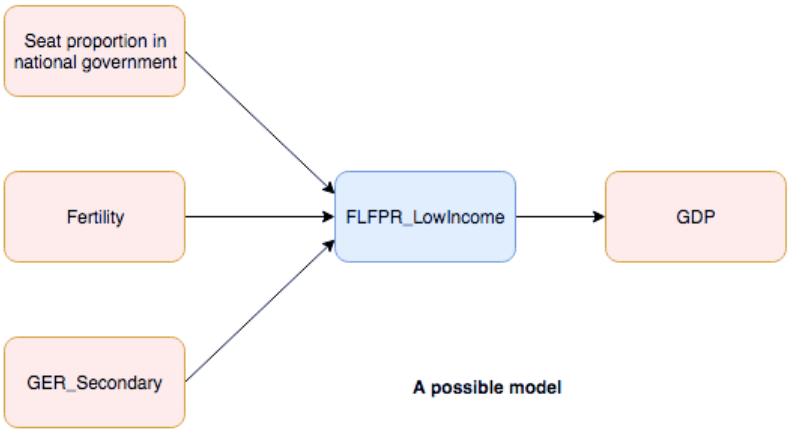


Fig. 23 Other model considerations

5.3 Parameter Settings

For a collider model, some specific parameters need to be set – prior for beta, prior for inverse variance, and likelihood. We created two models – one for high-income countries, and one for low-

income countries, as the latter includes one extra variable (Gross Enrolment Ratio in Secondary Schools).

5.3.1 Prior for beta

The prior for the beta will be taken from a normal distribution between 0 and 0.01.
For high income countries, we used 4 beta's and 4 sigma's. For low-income countries, we used 5 beta's and 4 sigma's. For each model, the number of normal distributions is equal to the number of beta's used.

5.3.2 Prior for the inverse variance

The inverse variance is used for the purpose of setting the sigma of variables in the model. The variable is set by taking a number from a uniform distribution between 0.01 and 0.99. Thus, a uniform distribution between 0 and 1 (excluding 0 and 1) is chosen. The (4 for high-income, and 5 for low-income countries) are squared to obtain sigma values, and then inverted to set upper limits for the variables in the likelihood.

5.3.3 Likelihood

The likelihood was calculated as shown below:

a. Low-Income countries

$$\mu_{2i} = \beta_1 + \beta_2 * seatprop_i + \beta_3 * gdp_i + \beta_4 * fertility_i + \beta_5 * gersecondary_i$$

b. High-Income Countries

$$flfpr_i \sim dnorm(\mu_{1i}, inv.var_1)$$
$$\mu_{1i} = \beta_1 + \beta_2 * seatprop_i + \beta_3 * gdp_i + \beta_4 * fertility_i$$

5.3.4 Other parameters

Other parameters that can be changed in the model include the burn in period which is now set at 10000 samples. The number of samples that the model needs to generate (after the burn in period) can also be modified - this number has been set to 20000.

5.4 Model Description

The below tests were used to evaluate model convergence:
a. Trace Plots – to assess coverage and consistency through the sampling space for all the relevant parameters that have been considered.

b. Gelman Plots – to assess convergence around 1 for all parameters to ascertain that variance within the chain is higher than between chains.

Using trace plots and Gelman plots, it appears that the model doesn't seem to have any problems. All the beta values for both low and high income runs show similar ranges and consistency.

5.4.1 High Income Model:

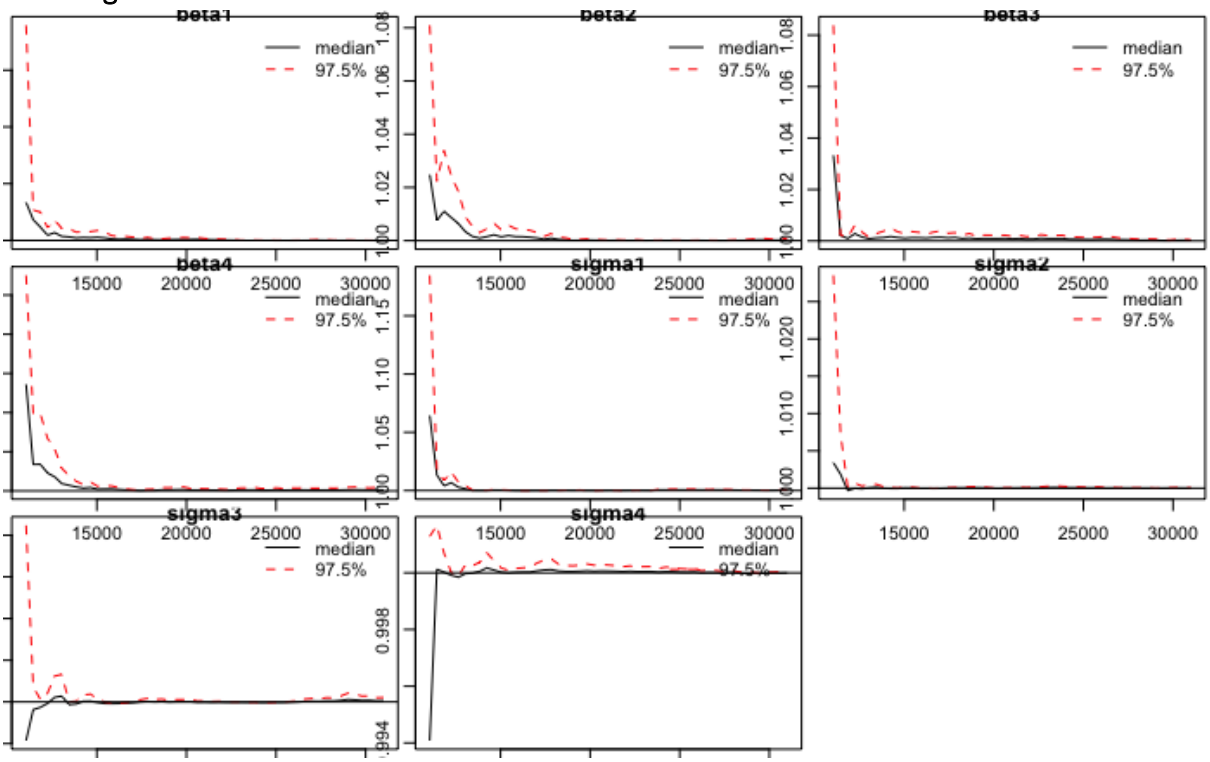


Fig. 24 Gelman Beta and Sigma convergence plots for High Income model

In order to test convergence, Gelman-Rubin statistics are plotted for all iterations. These plots unilaterally show that the statistics converge around 1, indicating that there may be convergence in our model.

The autocorrelation plots in figure 2 show that there is no autocorrelation for beta or sigma. Finally, sample sizes were inspected for all parameters as can be seen in table, below. The sample sizes well exceed any number data observations and easily cross the threshold of 10,000 iterations.

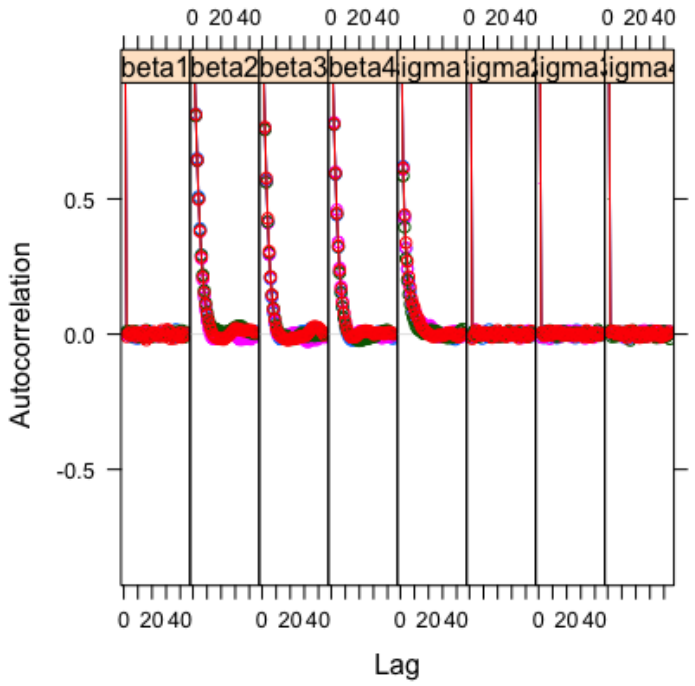


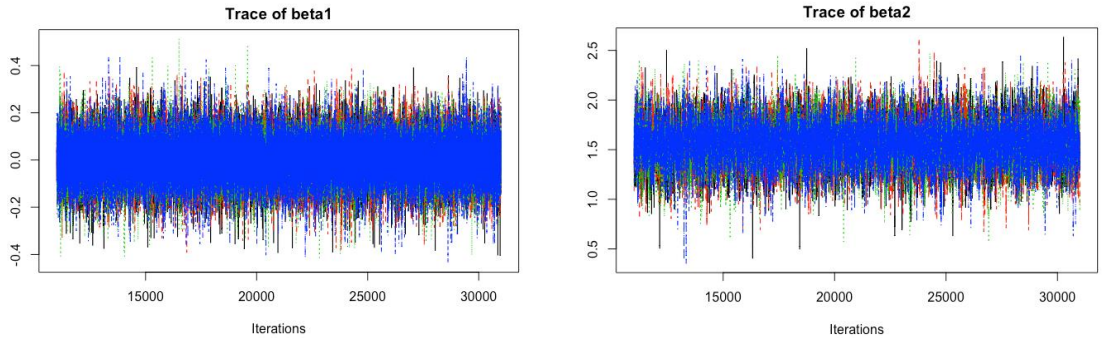
Fig. 25 Autocorrelation plots for Beta and sigma values for High Income countries

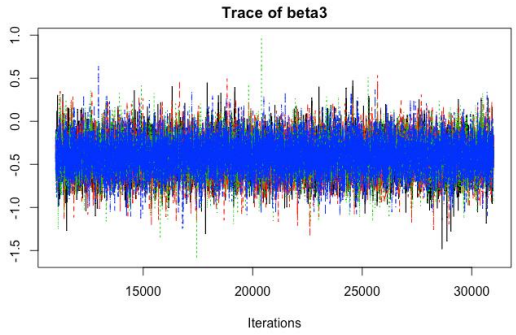
Table1: Effective Sample Sizes [High Income]

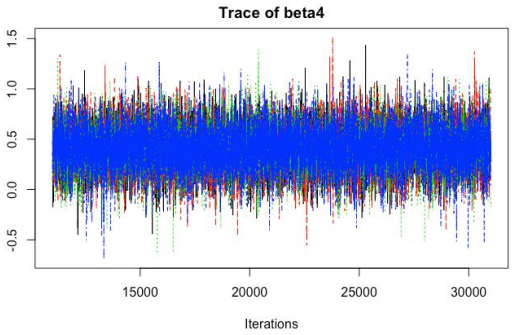
beta1	beta2	beta3	beta4	sigma1	sigma2	sigma3	sigma4
80000.00	10466.35	12932.40	11856.76	13722.71	80470.36	80495.41	79853.16

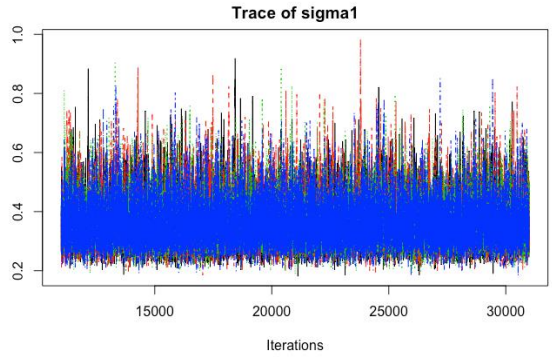
Based on the model diagnostics discussed above, the collider model requires no changes in burn-in period, the use of thinning, or additional runs/chains.

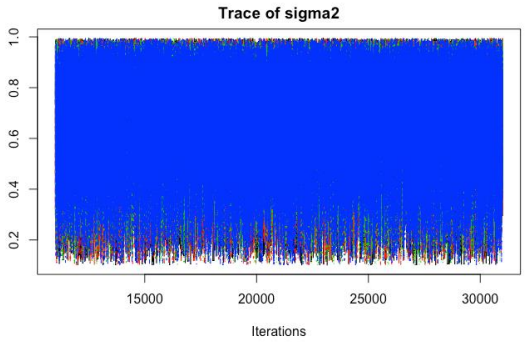
5.4.2 High-Income Model Output Images

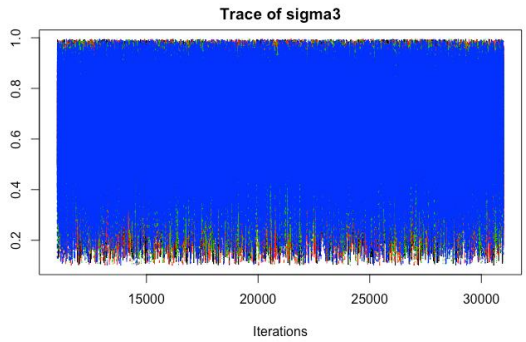


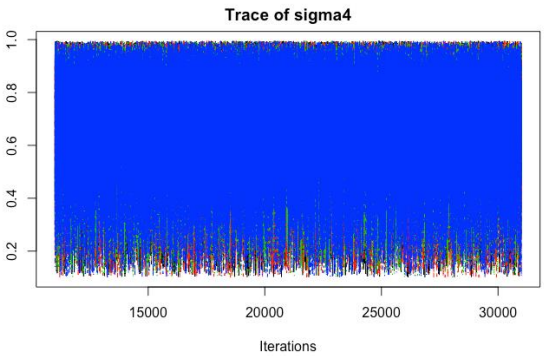

Trace of beta3
1.0
0.5
0.0
-0.5
-1.0
-1.5
15000
20000
25000
30000
Iterations


Trace of beta4
1.5
1.0
0.5
0.0
-0.5
15000
20000
25000
30000
Iterations


Trace of sigma1
1.0
0.8
0.6
0.4
0.2
15000
20000
25000
30000
Iterations


Trace of sigma2
1.0
0.8
0.6
0.4
0.2
15000
20000
25000
30000
Iterations


Trace of sigma3
1.0
0.8
0.6
0.4
0.2
15000
20000
25000
30000
Iterations


Trace of sigma4
1.0
0.8
0.6
0.4
0.2
15000
20000
25000
30000
Iterations

5.4.5 Low Income Model

In order to test convergence, Gelman-Rubin statistics are plotted for all iterations. These plots unilaterally show that the statistics converge around 1, indicating that there may be convergence in our model.

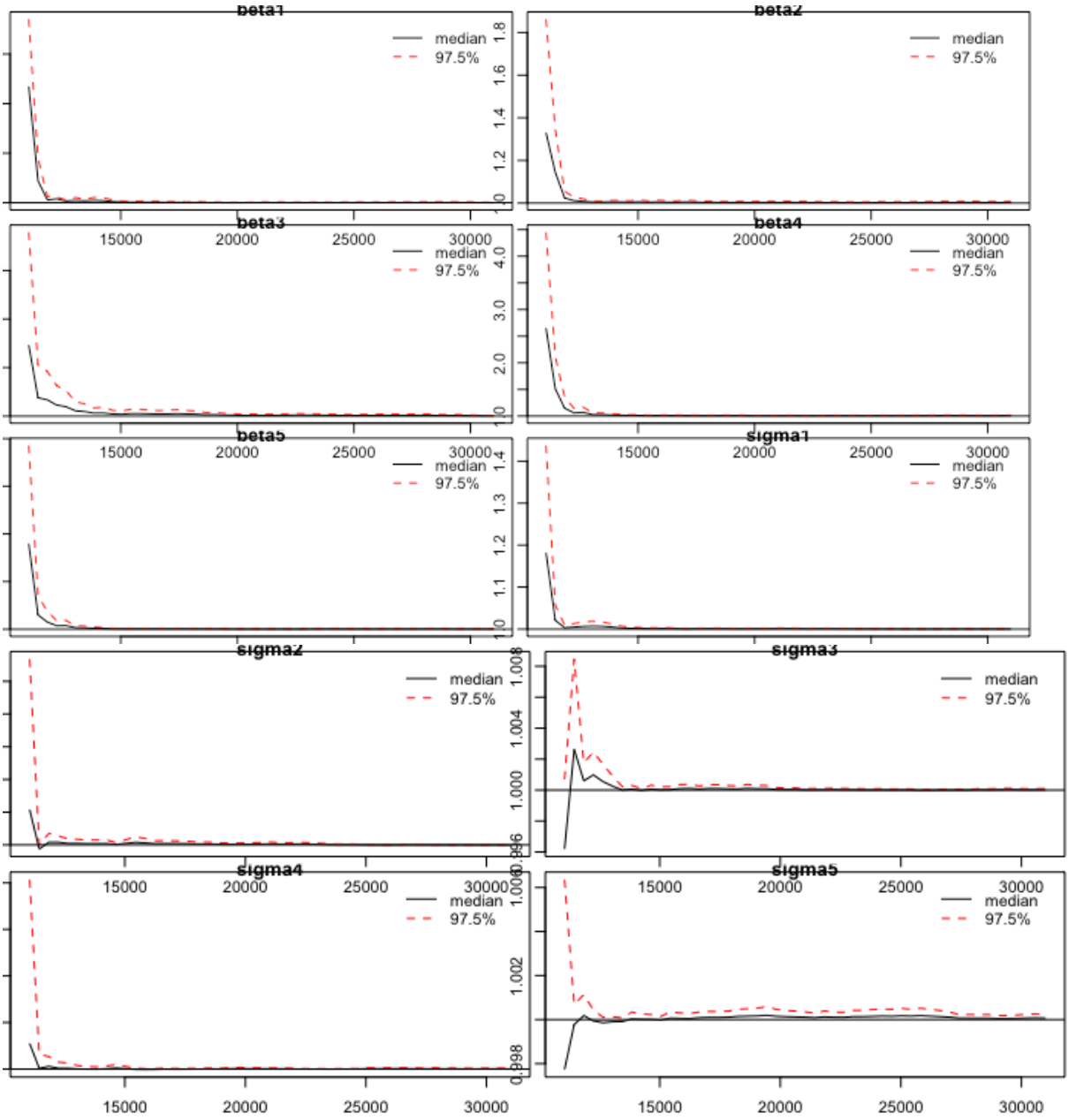


Fig 26. Gelman Beta and Sigma convergence plots for Low Income model

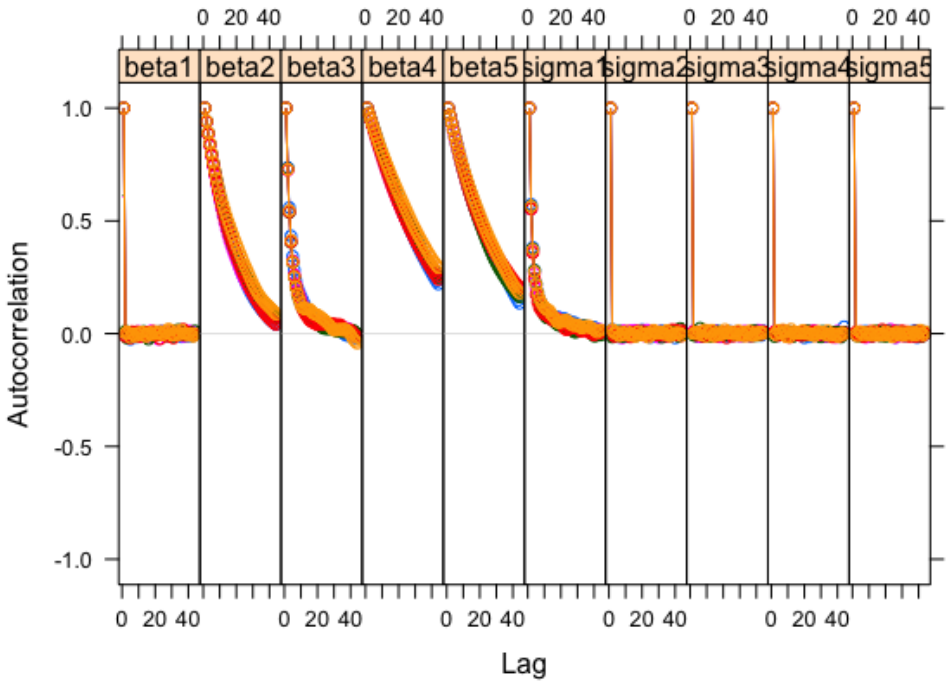


Fig. 27 Autocorrelation Plots for beta and sigma values.

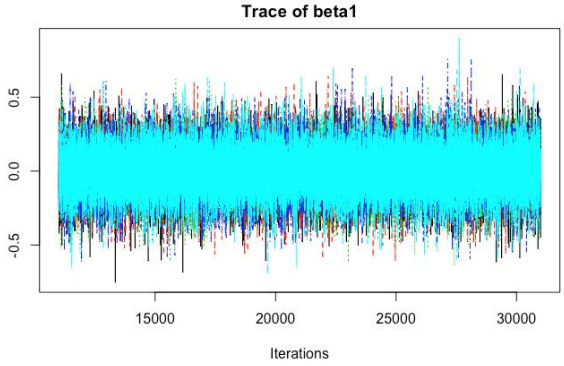
The autocorrelation plot shows that there is no autocorrelation for beta or sigma. The samples were finally inspected as listed in the table below. The sample sizes well exceed any number data observations and easily cross the threshold of 10,000 iterations.

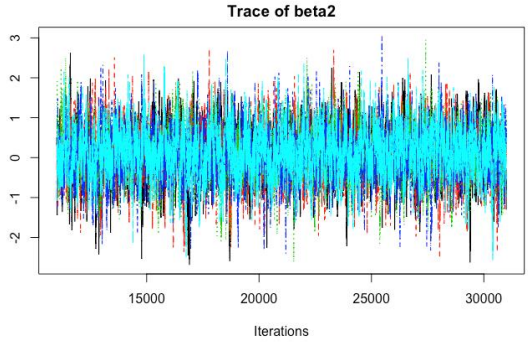
Table2: Effective Sample Sizes [Low Income]

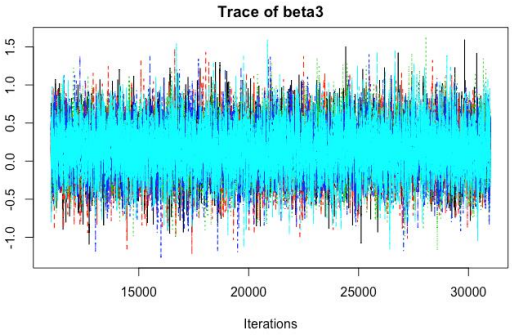
beta1	beta2	beta3	beta4	beta5	sigma1	sigma2	sigma3	sigma4	sigma5
103195.760	2764.902	11433.520	1401.938	1731.647	13924.592	99693.464	99116.237	99158.610	100000.000

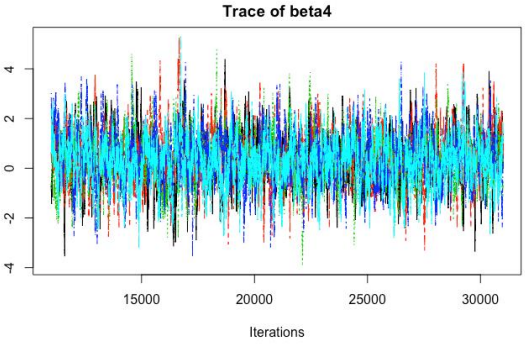
Based on the model diagnostics discussed above, the collider model requires no changes in burn-in period, the use of thinning, or additional runs/chains. The model can improve significantly by incorporating non-linear relationships between variables. Though these types of changes will increase the complexity of standardizing the data, it will provide a better insight into the complex relationships between flfpr and other variables. This model also explores a simply causal relationship among many variables which may not be the case in real life.

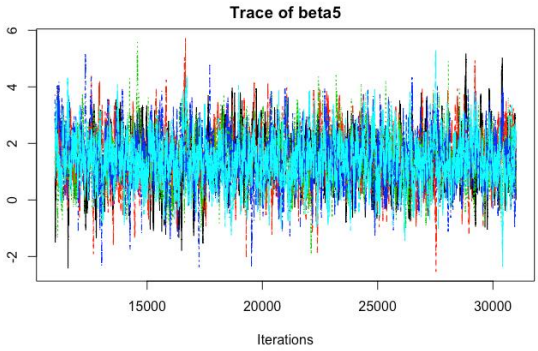
5.4.4 Low-Income Model Output Images

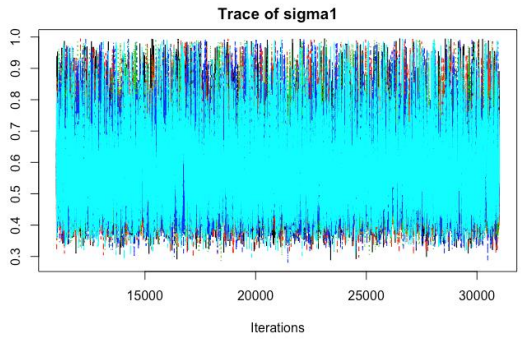


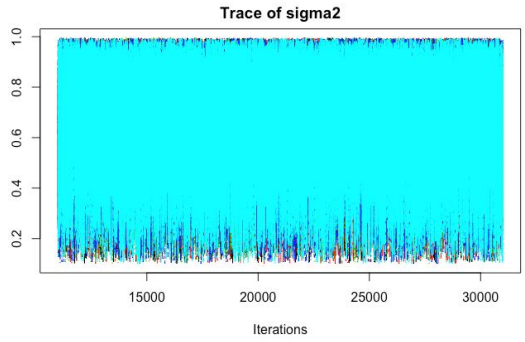


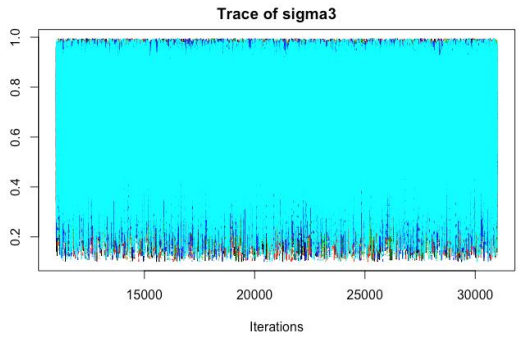


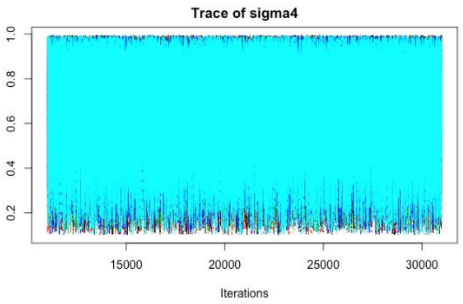


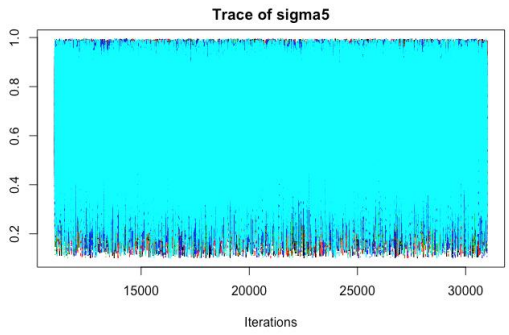












5.5 Assessment

The performed research has the below limitations:

5.5.1 Data Quality

There are limitations with respect to the nature of the data obtained, as is detailed in the Data Quality section of this report. This is especially evident in low-income countries, where data collection is not always standardized, and regular.

5.5.2 Model Selection

For the purposes of the analysis in the report, a relatively simple model was chosen. A more complex and comprehensive model could include maternal mortality, effect of women in managerial positions, etc.

The model chosen in the report selected less nuanced indicators that could influence FLFPR.

5.5.3 Choice of Parameters

The purpose of the research was to quantitatively identify cultural impacts on FLFPR. The motivation behind dividing our research on the basis of income was that western countries, while having higher

GDPs, also have widely differing cultures and eastern countries with respect to gender equality norms.

However, we couldn't achieve the goal entirely, as low income countries are more likely to have missing or un-standardized data points. Several nuances particular to these environments - education for kids, education attitude, gender ratio in low income countries, etc. cannot be quantified due to lack of comprehensive data.

We tried to overcome this by performing separate analysis on the basis of income.

5.5.4 Correlation is not the same as causation

We have identified several correlations in our report, but we cannot definitely identify causation.

5.6 Recommendation

Looking at the alarming decrease in FLFPR rates in the low-income countries as opposed to significant increase in the high-income countries, we advise the OECD to increase allocation of resources to improving women labor force participation in these countries. We also advice the OECD to allocate resources to better data procurement, as many of the low-income countries suffer from poor data quality that could reflect their reality. Improving data collection will definitely provide more insights in the relationships between FLFPR and other variables.

Citations:

International Labour Organization, Labour Force Participation Rate
https://www.ilo.org/ilostat-files/Documents/description_LFPR_EN.pdf

OECD, Fertility Rates
https://data.oecd.org/pop/fertility-rates.htm

UN Women, Facts and Figures: Economic Empowerment
https://www.unwomen.org/en/what-we-do/economic-empowerment/facts-and-figures#notes

UNESCO, Gross Enrolment Ratio
http://uis.unesco.org/en/glossary-term/gross-enrolment-ratio

United Nations, World Economic Situation and Prospects (2014), Country Classification, p.144
https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf