# Research Project

Aaqib Khatibi - Omar Jallouq

January 2025

## Introduction

Gender inequality is a prevalent issue worldwide, influencing various aspects of social and economic life. Despite significant action in recent years towards equality, disparities between genders persist, particularly in areas such as education, health, political representation, and economic participation. This study aims to dive deeper into the underlying factors that contribute to gender inequality, by analyzing several cofactors and their impact on the Gender Inequality Index (GII).

We will conduct a multivariate regression incorporating the seven variables –literacy rate, poverty rate, democracy index, health index, religion prevalence, global peace index, and prevalence of child marriage– to assess their relative influence. By utilizing data from different sources and combining them all together, the study aims to provide meaningful insights into the complex dynamics of gender inequality.

Research Question: How do different factors influence gender inequality worldwide.

## Dataset

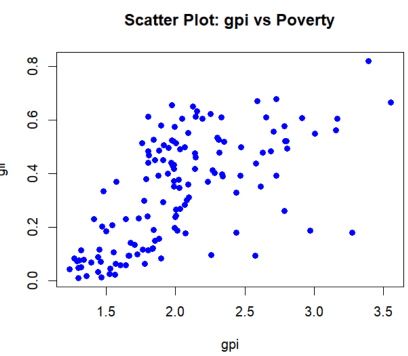
For our analysis, the dataset is compiled from mainly inter-governmental and UN sources, a detailed reference list can be found in the bibliography. We merged the data into a single CSV file, and the column names (in other words, all the covariates of our original model) are specified and explained below. The data was compiled in accordance to a shared year, hence for the purpose of this study, the data obtained for all the covariates are from around the year 2021.

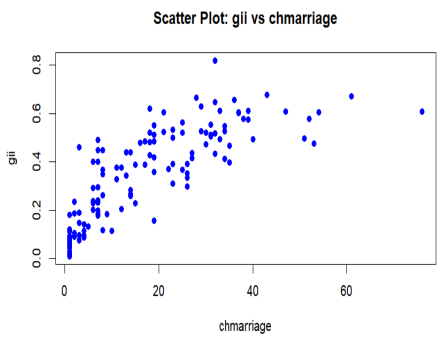
**gii -** UN Gender Inequality Index, from 0 to 1, where a higher value corresponds to a higher gender inequality.
**literacy -** Literacy Rate of the country expressed as a percentage.
**poverty -** Poverty Rate of the country expressed as a percentage.
**dindex -** Democracy Index, in the range of 0 to 10.
**uhssci -** Universal Health Coverage (UHC) Service Coverage Index in the range of 0 to 100, where a higher value denotes higher medical coverage, corresponding to better medical infrastructure.
**religion -** Whether or not the country has a national or a state religion (= 1 if yes; 0 otherwise)
**gpi -** Global Peace Index, where a higher value denotes more instability and conflict in the country.
**chmarriage -** Child Marriage, expressed as a percentage of all marriages.

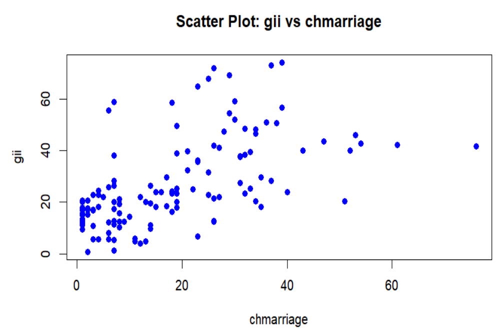
We considered a mix of factors to better explain our model. Hence, the factors include a mix of economic, cultural, and social factors. Cultural factors, especially in the case of our study are not so easy to quantify but we still took them into consideration according to the availability of data. A different factor that we considered is **chmarriage** which is an estimate for child marriage in the country, this will help us to track some prevalent cultural practices. According to different sociological studies, marrying below the age of 18 prevents women from exercising any liberty, further increasing the gender inequality gap. For other estimates (cofactors) we used different sources, which enabled us to create a single CSV file including all the variables. At first, the dataset included 180 countries/territories, which we narrowed down to 138 countries as our data points after checking for missing values and outliers (ndataset_Sheet1.csv).

We constructed scatter plots, to gain better insights on how different independent variables relate to our dependent variable (GII) separately. We wanted to get a prior idea to better understand the hypothesis and the potential results of our research. The plots show a positive correlation with poverty, religion, gpi, chmarriage, while we saw negative correlations in the case of literacy, dindex, and uhssci. This relation reflects our theoretical expectation. To further our understanding, with some statistical data, we came to find out that the mean GII for countries with national/state religion is higher than that of countries that do not follow a religion constitutionally. This statement also relates to why we saw a positive correlation of the religion variable with GII (refer to R script).

Here are some of the plots which show the relation of different factors with the dependent variable GII:




# Multivariate Linear Regression

The purpose of our study is to find and explain how different variables relate to gender inequality worldwide. And also how on the basis of different characteristics of the country it can theoretically direct the gender inequality index. Hence, we decided to perform a multivariate regression. Our original model consists of seven covariates. From the R output, we saw some of them seem to be quite significant if we consider the p-value significance level below 0.05.

The value of $R^2$ and the Adjusted $R^2$ in our original full model (full_model in the R script) come out to be 0.8706 and 0.8636 respectively, which are very satisfying values. Considering the Adjusted $R^2$ in our case as the value of $R^2$ can just increase while adding more cofactors even if they contribute little to the model's overall performance, and we want to penalize bigger models. As per the Adjusted $R^2$, our cofactors infact explain 86% variance.

For our model, we decided to find a better explanatory model while running our first regression and some selection methods.

## Model Passage

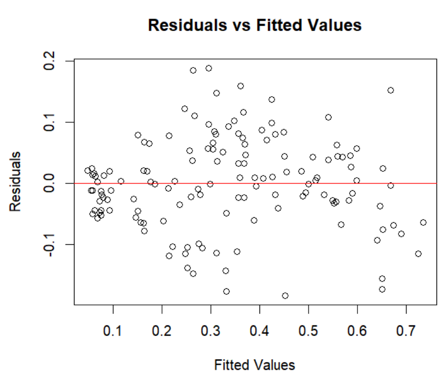
Throughout our model selection, we kept in mind that we needed a model that will be better explanatory and also simpler. To achieve that, we did apply both step-up and step-down methods, which could help us find a smaller and better model. We also used the AIC penalty method to compare between our models. After applying both the step-up and step-down methods, we reduced our model to a simpler model with three covariates: unssci, gpi, chmarriage. Both the step-up and step-down methods agreed with this model. It is evident from the summary in the R script that all three models are quite significant, at a significance of 0.05. While we got a negative correlation with uhssci (health), a positive correlation was seen in the case for gpi and chmarriage.

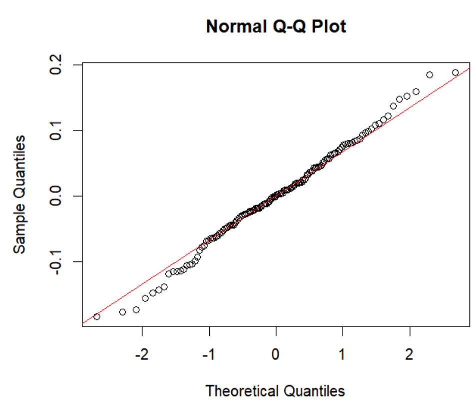
```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.18399 -0.04447  0.00064  0.04598  0.18867 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.7059816  0.0670509  10.529  < 2e-16 ***
uhssci      -0.0081846  0.0006616 -12.371  < 2e-16 ***
gpi          0.0610161  0.0165189   3.694 0.000321 ***
chmarriage   0.0023083  0.0007224   3.195 0.001743 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07523 on 134 degrees of freedom
Multiple R-squared:  0.8631,	Adjusted R-squared:  0.8601 
F-statistic: 281.7 on 3 and 134 DF,  p-value: < 2.2e-16
```

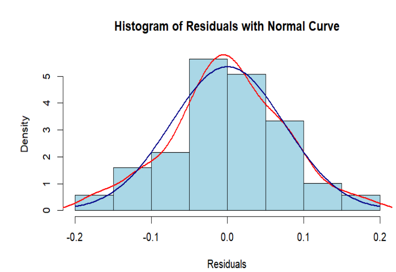
With the Akaike Information Criteria (AIC), we on the other hand found a different model as compared to the step-up and step-down methods. This model is actually a bigger model as compared to the one we obtained in the previous methods, which includes five covariates instead of three, poverty and religion be the two additional ones. If we consider the significance level of p-values to be 0.05, we see that they don’t seem to be very significant compared to the other three covariates. In the bigger model (stepwise_model), the AIC value turns out to be the lowest. While the thing to be noted down is that the difference between the AIC value in both models is not much. The lower value in the case of the bigger model can be explained in such a way that in the case of smaller models/smaller datasets, which we are currently dealing with, increasing the number of covariates often increases the explained variance. Hence, AIC prefers bigger models in such scenarios.

Indeed, to verify the assumptions of a linear model, we need to analyse the residuals. We made a scatterplot of the residuals, where no symmetric change on the spread is seen. It is asymmetrically scattered, further validating the assumption of homoscedasticity in our model. We further also checked the multicollinearity between our variables through Variance Inflation Factor (VIF).

To check normality, the QQ-plot validates it along with Shapiro-Wilk Test and Kolmogorov-Smirnov test that we performed (more details in R script).

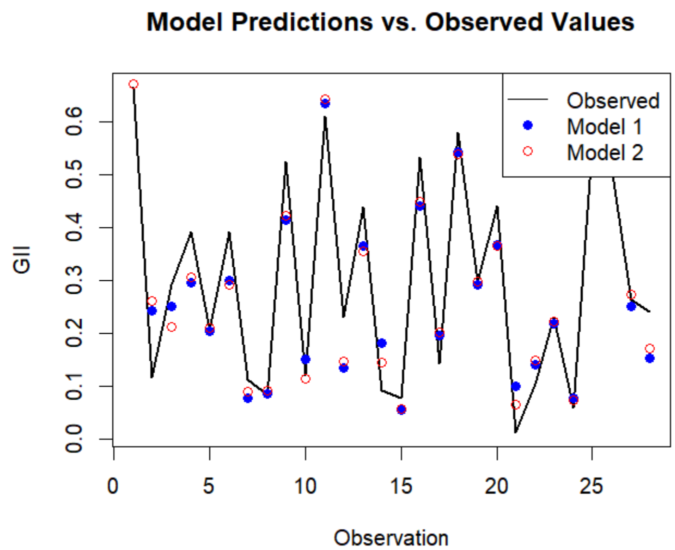


# Prediction

For our understanding and to gain insights, we wanted to run a prediction test for GII. Hence, we split our dataset (ndataset_Sheet1) into two parts, 80% as the training set, and the remaining 20% as the testing set. The results show that the predictions in both models do not differ much. It is quite clear as seen in the figure below.




As per the Root Mean Square Error metric, we obtain the value of 0.063 for the step-up/step-down model (final_model) compared to the value of 0.067 for the AIC model (stepwise_model). Hence, this method favors the simpler model slightly. Considering the fact that simpler models are often preferred, and also an understanding of the AIC model that potentially the lower value in the bigger model does not signify that the two extra covariates in the bigger model are really significant in large scale, we decided to prefer the model we got from the step-up/step-down p-value method over the other bigger one.

**Test for distribution of Gender Inequality**

We examined how gender inequality varies with a country's economic condition, dividing countries into three income groups: High Income (poverty < 20%), Middle Income (20-40%), and Low Income (poverty > 40%). Using the Wilcoxon rank-sum test with Bonferroni correction, we compared the distributions of GII across these groups to control for Type I error from multiple tests (see appendix for details related to Bonferroni correction). The null hypothesis assumed no difference in GII distributions between income groups, while the alternative suggested differences. Results align with the scatterplot showing a positive correlation between poverty and GII, with p-values significantly below 0.05 for all pairs, indicating substantial differences in GII across income groups.

Null Hypothesis ($H_0$ ): The distributions of GII are the same between the two income groups.

Alternative Hypothesis ($H_1$): The distributions of GII are different between the two income groups.

The results in some way agree with the scatterplot which shows a positive correlation of **poverty** with **gii.** As seen from the R output below, the p value is much lower than the generally considered significant level of 0.05 for each pair.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  ndataset_Sheet1$gii and ndataset_Sheet1$income_group

              High Income Middle Income
Middle Income 2.2e-06     -
Low Income    3.1e-12     2.5e-05

P value adjustment method: bonferroni
```

This provides us with strong statistical evidence to reject the null hypothesis that the distributions of **gii** are the same between groups. For better visualization, we have also decided to show the distribution of **gii** for each income group in the Appendix.

## Final Discussion and Conclusion

Our study identified health (uhssci), political instability (gpi), and child marriage (chmarriage) as significant contributors to gender inequality (gii), with a robust model selection process yielding a simpler model supported by an adjusted R square of 0.8636. Indeed, there are further limitations to our study, especially when we decided to remove countries because of the unavailability of data, we removed a lot of countries where theoretically different patterns of gender inequality exist but our data set and study failed to capture that. Furthermore, sociocultural factors we considered such as religion and child marriage might not be able to capture the dynamic way these factors work in real life. In the regression analysis, different models such as ANOVA could have been applied to find out the best subset of the model which may be able to track some interesting relationships among the different covariates.

Furthermore, in the final prediction we applied because of the explanation of the AIC model and the preference of RMSE value in the prediction along with the common knowledge that simpler models are preferred we reached our conclusion and selected the simpler model. More statistical models and estimates could have been applied to further prove why that model is being preferred. Lastly, to be consistent with our syllabus in the class and our coursebook we decided to use the Wilcoxon test in pairs, in such a situation a different testing method like Kruskal-Wallis test paired up with Dunns test would be a more general alternative.

This study underscores the role of health, political instability, and cultural practices in shaping gender inequality while acknowledging the need for further refinement in data and methodologies.

# Bibliography:


- https://www.who.int/news/item/18-09-2023-billions-left-behind-on-the-path-to-universal-health-coverage
- https://freedomhouse.org/countries/freedom-world/scores
- https://worldpopulationreview.com/country-rankings/poverty-rate-by-country
- https://worldpopulationreview.com/country-rankings/poverty-rate-by-countryhttps://worldpopulationreview.com/country-rankings/poverty-rate-by-countryhttps://worldpopulationreview.com/country-rankings/poverty-rate-by-countryhttps://worldpopulationreview.com/country-rankings/poverty-rate-by-country
- https://data.unwomen.org/global-database-on-violence-against-women/data-form

# Appendix

Here in the appendix, we decided to add the information that will make the content of our research more robust.

**What is the Bonferroni Correction?**

The Bonferroni correction is a statistical method used when performing multiple comparisons. It adjusts the significance level to control the overall error rate, reducing the likelihood of mistakenly identifying a result as significant (Type I error).

Why is the Bonferroni Correction Needed?

1. Increased Error Rates with Multiple Comparisons:
   - When performing multiple tests, the chance of a false positive increases.
   - For example, if we perform 10 independent tests with a significance level of 0.05, there is about a 40% chance of getting at least one false positive.
2. The Bonferroni Solution:
   - The correction divides the significance level (such as 0.05) by the number of tests. This reduces the threshold for determining significance, making the criteria stricter.

How Does the Bonferroni Correction Work?

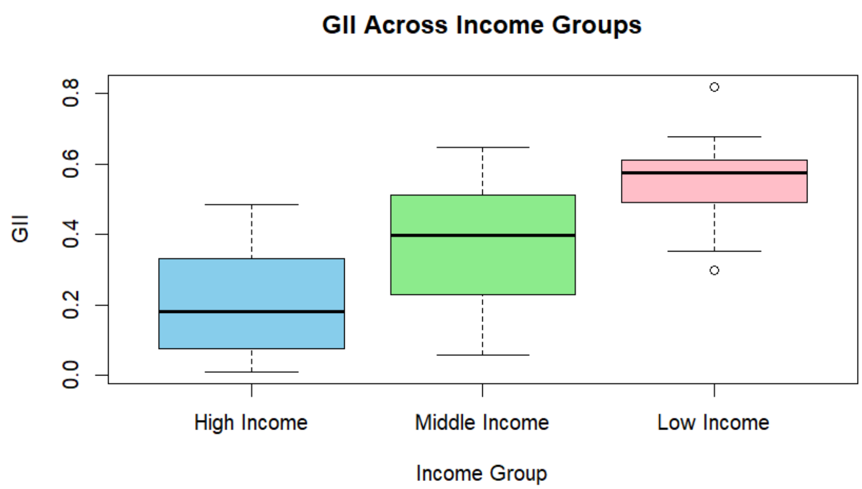
1. Adjusting the Significance Level:
   - Instead of using the original significance level (e.g., 0.05), you divide it by the number of comparisons being made. For example, if there are 3 tests, the adjusted level is 0.05 divided by 3, or approximately 0.0167.
2. Adjusting the p-values:
   - Alternatively, we can multiply each p-value by the number of tests instead of adjusting the significance level. For instance, if the original p-value is 0.01 and there are 3 comparisons, the adjusted p-value would be 0.01 times 3, which equals 0.03.
3. Comparison:
   - Compare the adjusted p-values to the original significance level (0.05). If the adjusted p-value is smaller than 0.05, the result is significant.

Example in our Analysis

In your case:

- We performed 3 pairwise comparisons (High Income vs. Middle Income, High Income vs. Low Income, Middle Income vs. Low Income).

- The Bonferroni correction adjusted the p-values by multiplying them by 3.
- For example:
  - If the original p-value was 0.01, the adjusted p-value is 0.03.
  - Since 0.03 is less than 0.05, the result remains significant.




```
Residuals:
      Min        1Q    Median        3Q       Max
-0.199243 -0.043024  0.000034  0.046084  0.171549

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.6518928  0.0711801   9.158 8.74e-16 ***
ndataset_Sheet1$religion    0.0311499  0.0154882   2.011  0.04634 *
ndataset_Sheet1$poverty     0.0009252  0.0005178   1.787  0.07629 .
ndataset_Sheet1$uhssci     -0.0076416  0.0007001 -10.916  < 2e-16 ***
ndataset_Sheet1$gpi         0.0540082  0.0165725   3.259  0.00142 **
ndataset_Sheet1$chmarriage  0.0023757  0.0007321   3.245  0.00149 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```