Project Proposal

Problem Statement

In some low- and middle-income countries (LMICs), young women are increasingly expressing a desire to have their first child much later in life compared to older generations. In other LMICs, young women have not significantly changed their intent to delay the timing of their first birth. What factors can best explain these differences?

This analysis aims to generate a model that can identify why views among young women on the timing of their first birth have changed so drastically in some places, but not in others. The analysis will consider variables that draw on existing explanations, such as the increased availability of family planning, new laws on women's right to work, and changes in norms around women's agency and status more broadly.

2. Data Sources

- Demographic and Health Survey – Individual Recode Data
  - Preferred waiting time
  - Ideal number of kids
  - Religion
  - Urban/rural
  - Knowledge of family planning
  - Heard of family planning via TV, newspaper or magazine, and radio
- Demographic and Health Survey – Household Recode Data
  - Educational attainment of mom
  - Educational attainment of older sisters
- World Bank – World Development Indicators
  - Ratio of female to male labor force participation rate (%)
- World Bank – Women, Business and the Law Data
  - Does the law prohibit discrimination in employment based on gender?
  - Is there legislation on sexual harassment in employment?
  - Are there criminal penalties or civil remedies for sexual harassment in employment?
  - Does the law mandate equal remuneration for work of equal value?
  - Can women work the same night hours as men?
  - Can women work in jobs deemed dangerous in the same way as men?
  - Are women able to work in the same industries as men?
- United Nations World Population Policies Database
  - Policy on population size and growth
  - Level of concern about ageing of the population
  - Policy on fertility level

3. Obtaining the Data

Prior to downloading any Demographic and Health Survey data, I will have to scrape information from the DHS Program website to learn which countries and surveys I can use for my analysis. Specifically, I will need to scrape the main list of surveys and drop any surveys with restricted or otherwise unavailable data, as well as omit surveys specific to malaria or AIDS. I will also have to scrape each survey's individual website to collect information on the sample (all women ages 15-49, married women ages 15-49, etc.) and other survey characteristics (e.g., has module on women's status or domestic violence).

Once I have finalized the sample of countries, I can download the Individual Recode data for information on the young women and the Household Recode data for information on their family members. I will need to clean the Household Recode data and calculate the mean years of education for any older sisters in each household. I will then merge the adapted Household Recode data with the Individual Recode data.

The remaining data comes from either the World Bank or the United Nations. I will use the WBSTATS package in R to obtain the World Development Indicator variables, and manually download the other data (Piburn, 2018). The Women, Business and the Law data and the World Population Policies Database have data only for every other year so I will create variables that assume the values of the previous year for years in which not data were collected.

4. Methodology

As described above, this project will use various datasets and I expect that not all of them will be in tidy format. As such, I will use the TIDYVERSE package in R to clean, re-structure, and join the data (Wickham et al., 2019).

The first two data visualizations I would like to produce are plots of how the mean and median preferred waiting time has changed over time, by country. The third visualization I would like to produce is a waffle chart for the three countries with the greatest and smallest changes in mean preferred waiting time. Using the WAFFLE package in R, each square of the waffle chart would represent some number of survey respondents and squares would be different colors depending on the actual preferred waiting time. For example, red squares could be a preferred waiting time of less than one year, orange squares more than one year, but less than two years, etc.

In terms of machine learning techniques I would like to apply, I plan to use K-nearest neighbors, bootstrapping/bagging, and random forest.

5. Defining Success for the Project

I will consider this project a success if the analysis I am able to do is robust in terms of sample size, variables considered, and methods applied.

6. References

- Bob Rudis and Dave Gandy (2017). waffle: Create Waffle Chart Visualizations in R. R package version 0.7.0. https://CRAN.R-project.org/package=waffle
- Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer
- Jesse Piburn (2018). wbstats: Programmatic Access to the World Bank API. Oak Ridge National Laboratory. Oak Ridge, Tennessee. URL https://www.ornl.gov/division/csed/gist
- Hadley Wickham (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
- Hadley Wickham (2019). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.5. https://CRAN.R-project.org/package=rvest
- Hadley Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686