Sara Benist MADA Project

Tuberculosis Burden and Health Inequality Measures

Sara Benist

2/3/23

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

Warning: package 'here' was built under R version 4.2.2

Warning: package 'knitr' was built under R version 4.2.2

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

## 2.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis*

## 2.2 Description of data and data source

*(Part 1 of project: data source, description, loading into raw folder, general aspects)*

The World Health Organization collected data on the inequity surrounding tuberculosis, HIV, and malaria for the [State of inequality report](https://www.who.int/data/inequality-monitor/publications/report_2021_hiv_tb_malaria), and I will be exploring the dataset for TB. More information about the data can be found [here](https://heatrepository.blob.core.windows.net/documents/data-repository-indicator-list.pdf?sp=r&st=2022-06-07T14:16:36Z&se=2023-12-30T23:16:36Z&spr=https&sv=2020-08-04&sr=b&sig=4kzThU1QDo55UOQyhWcUd8rPWJ9LxUZdRRI3zl6wKNs%3D) under “Tuberculosis Indicators”. The dataset can be found [here](o%20https://www.who.int/data/inequality-monitor/data#PageContent_C158_Col00) under “Tuberculosis Indicators”.

For the tuberculosis dataset, the data was collected from the WHO Global TB programme, TB prevalence surveys, country-specific TB programmes, the WHO Health Equity Monitor database, TB patient cost surveys, and other sources. The WHO organized the dataset to be used with the Health Equity Assessment Toolkit which is the built in data analysis and exploration tool. The database contains 10 variables regarding the burden, detection, prevention, knowledge, and social protection and observations for 194 countries over various years. The observations for each variable is further separated by up 7 inequality domains. Not all countries have data available for each year or for each inequality domain. The dataset contains a total of 7473 observations.

raw\_data <- readxl::read\_xlsx(path = "../../data/raw\_data/202206-repository-tb.xlsx")  
str(raw\_data)

tibble [7,473 × 21] (S3: tbl\_df/tbl/data.frame)  
 $ setting : chr [1:7473] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...  
 $ year : num [1:7473] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...  
 $ source : chr [1:7473] "MICS" "MICS" "MICS" "MICS" ...  
 $ indicator\_abbr : chr [1:7473] "bcg" "bcg" "bcg" "bcg" ...  
 $ indicator\_name : chr [1:7473] "BCG immunization coverage among one-year-olds (%)" "BCG immunization coverage among one-year-olds (%)" "BCG immunization coverage among one-year-olds (%)" "BCG immunization coverage among one-year-olds (%)" ...  
 $ dimension : chr [1:7473] "Economic status (wealth quintile)" "Economic status (wealth quintile)" "Economic status (wealth quintile)" "Economic status (wealth quintile)" ...  
 $ subgroup : chr [1:7473] "Quintile 1 (poorest)" "Quintile 2" "Quintile 3" "Quintile 4" ...  
 $ estimate : num [1:7473] 53.8 61.9 58.1 65.1 77.9 ...  
 $ se : num [1:7473] 4.2 3.12 3.37 3.77 2.24 ...  
 $ ci\_lb : num [1:7473] 45.5 55.6 51.4 57.4 73.1 ...  
 $ ci\_ub : num [1:7473] 61.8 67.8 64.6 72.2 81.9 ...  
 $ population : num [1:7473] 532 549 495 473 447 ...  
 $ flag : chr [1:7473] NA NA NA NA ...  
 $ setting\_average : num [1:7473] 62.9 62.9 62.9 62.9 62.9 ...  
 $ iso3 : chr [1:7473] "AFG" "AFG" "AFG" "AFG" ...  
 $ favourable\_indicator: num [1:7473] 1 1 1 1 1 1 1 1 1 1 ...  
 $ indicator\_scale : num [1:7473] 100 100 100 100 100 100 100 100 100 100 ...  
 $ ordered\_dimension : num [1:7473] 1 1 1 1 1 1 1 1 0 0 ...  
 $ subgroup\_order : num [1:7473] 1 2 3 4 5 1 2 3 0 0 ...  
 $ reference\_subgroup : num [1:7473] 0 0 0 0 0 0 0 0 0 1 ...  
 $ topic : chr [1:7473] "TB" "TB" "TB" "TB" ...

## 2.3 Questions/Hypotheses to be addressed

*(Part 1 of project: RQ, outcomes, predictors, patterns)*

Research question: How do the TB indicators relate to the TB incidence, prevalence, and mortality of a country, and what inequality measure shows the greatest disparity in TB outcomes?

The overall outcome I would study is TB mortality since effective health programs ideally reduce disease-specific mortality. By the end of my analysis, I would like to be able to identify populations that could be a focus of TB health improvement programs. Other outcomes I would like to explore include regional differences in drug-resistant TB and the attitudes and perceptions for high burden areas compared to medium or low burden regions.

In addition to the inequality and indicator predictors provided with the dataset, I would like to examine differences in TB outcomes based on level of TB burden. I would need to add another classification based on the literature to indicate what levels of TB incidence and prevalence would fall into each level of disease burden.

In the data, the pattern I would expect to see is a higher burden of disease on populations with greater inequality. However, I am not confident in predicting how the indicator categories, specifically the TB attitudes and perceptions, would relate to the TB outcomes.

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above (here dataanalysis\_template\_references.bib) and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020)

# 3. Methods

*(Part 1 of project: initial analysis thoughts)*

To analyze, I would need to determine how to handle the missing data. The years between countries are not specific, and not all indicators were collected for each country. At the moment, I would subset the data based on indicator and remove incomplete observations to analyze each indicator individually. For the final analysis, I will most likely focus on TB indicators that are significantly different between subgroups or complete enough for further analysis. As the class progresses, I look forward to learning other analysis techniques. Eventually, I would like to create a dataset that can produce the statistical information, plots, and models that explore TB outcomes based on inequality measures, TB indicators, and/or level of burden.

## 3.1 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 3.2 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

## 3.3 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 1](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below.

Table 1: Data summary table.

| skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA |
| numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ |
| numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 1](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 1: Height and weight stratified by gender. |

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

Table 2: Linear model fit table.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 |
| Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 |
| GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 |
| GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, et al., 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 6. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>