Sara Benist MADA Project

Tuberculosis Burden and Health Inequality Measures

Sara Benist

4/7/23

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

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

Warning: package 'ggplot2' was built under R version 4.2.2

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# 1. Questions for Dr. Handel

I would like to be able to predict the main TB outcomes based on the health equity indicators (ex: level of TB incidence predicted from % of people with drug-resistant strains), but I cannot join the rows together correctly. When I tried bind\_rows() on the one of the countries to see if it would work, the outcome rows are still separated and if I remove year to force everything onto the same row, I have 90 variables and they are grouped within the same cell. I added an example below. Do you have any advice?

#path to data  
data\_location <- here::here("data","raw\_data","202206-repository-tb.xlsx")  
  
#load data and assign to rawdata  
rawdata <- readxl::read\_excel(data\_location)  
  
#without year removed: outcomes remain on a separate row  
test <- rawdata %>%   
 select(c(setting, year, indicator\_name,  
 dimension, subgroup,  
 estimate)) %>%   
 pivot\_wider(names\_from = c("indicator\_name", "dimension", "subgroup"), values\_from = "estimate") %>%   
 filter(setting == "Afghanistan") %>% bind\_rows()  
test

# A tibble: 3 × 91  
 setting year BCG i…¹ BCG i…² BCG i…³ BCG i…⁴ BCG i…⁵ BCG i…⁶ BCG i…⁷ BCG i…⁸  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 Afghani… 2010 53.8 61.9 58.1 65.1 77.9 61.1 76.3 85.9  
2 Afghani… 2015 64.7 67.5 72.0 79.5 84.0 70.5 86.1 87.9  
3 Afghani… 2020 NA NA NA NA NA NA NA NA   
# … with 81 more variables:  
# `BCG immunization coverage among one-year-olds (%)\_Place of residence\_Rural` <dbl>,  
# `BCG immunization coverage among one-year-olds (%)\_Place of residence\_Urban` <dbl>,  
# `BCG immunization coverage among one-year-olds (%)\_Sex\_Female` <dbl>,  
# `BCG immunization coverage among one-year-olds (%)\_Sex\_Male` <dbl>,  
# `People who report TB is spread through coughing (%)\_Sex\_Female` <dbl>,  
# `People who report TB is spread through coughing (%)\_Sex\_Male` <dbl>, …

#with year removed: cells contain multiple observations  
test2 <- rawdata %>%   
 select(c(setting, indicator\_name,  
 dimension, subgroup,  
 estimate)) %>%   
 pivot\_wider(names\_from = c("indicator\_name", "dimension", "subgroup"), values\_from = "estimate") %>%   
 filter(setting == "Afghanistan") %>% bind\_rows()

Warning: Values from `estimate` are not uniquely identified; output will contain list-cols.  
\* Use `values\_fn = list` to suppress this warning.  
\* Use `values\_fn = {summary\_fun}` to summarise duplicates.  
\* Use the following dplyr code to identify duplicates.  
 {data} %>%  
 dplyr::group\_by(setting, indicator\_name, dimension, subgroup) %>%  
 dplyr::summarise(n = dplyr::n(), .groups = "drop") %>%  
 dplyr::filter(n > 1L)

test2

# A tibble: 1 × 90  
 setting BCG immu…¹ BCG i…² BCG i…³ BCG i…⁴ BCG i…⁵ BCG i…⁶ BCG i…⁷ BCG i…⁸  
 <chr> <list> <list> <list> <list> <list> <list> <list> <list>   
1 Afghanistan <dbl [2]> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   
# … with 81 more variables:  
# `BCG immunization coverage among one-year-olds (%)\_Place of residence\_Rural` <list>,  
# `BCG immunization coverage among one-year-olds (%)\_Place of residence\_Urban` <list>,  
# `BCG immunization coverage among one-year-olds (%)\_Sex\_Female` <list>,  
# `BCG immunization coverage among one-year-olds (%)\_Sex\_Male` <list>,  
# `People who report TB is spread through coughing (%)\_Sex\_Female` <list>,  
# `People who report TB is spread through coughing (%)\_Sex\_Male` <list>, …

If it is not feasible, I will be analyzing TB outcomes separately from indicators instead.

# 2. Summary/Abstract

Full summary to be added (TBA): background of topic, objective of project, quick overview of methods, main findings

# 3. Introduction

## 3.1 General Background Information

Tuberculosis is an infectious disease caused by *Mycobacterium tuberculosis* and is the leading infectious cause of death in the world even though the disease is both treatable and curable(Organization, 2022c). Approximately 10 million people are infected each year, and 1.5 million die from the disease (Organization, 2022c). Humanity has been impacted by TB for centuries, but the burden of disease is not equal across countries and populations (Organization, 2022a). The World Health Organization marked 30 countries as having highest burden of TB due to the number of cases, presence of multi-drug resistance strains, and the high mortality within certain populations such as HIV (Organization, 2021b). To understand more about why some countries or groups of people experience the higher burden, the World Health Organization gathered data on several health equity indicators subgroups. The Global Tuberculosis Programme houses the End TB Strategy with a goal to eradicating TB. Their current objective is to reduce TB incidence by 80% and TB deaths by 90% by 2030 (Organization, 2022b). Previous research has focused on trends in TB outcomes,; this project will explore and analyze the disparities within the indicators based on subgroup and burden of disease as well as analyizing trends in tuberculosis outcomes.

## 3.2 Description of data and data source

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 (World Health Organization, 2021). 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” (Organization, 2021a). The dataset can be found [here](https://www.who.int/data/inequality-monitor/data#PageContent_C158_Col00) under “Tuberculosis Indicators” (Organization, 2021a).

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 (World Health Organization, 2021). 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.

## 3.3 Questions/Hypotheses to be addressed

*update depending on ability to bind indicators with outcomes*

Research question: Does high disparities between inequality measures show high correlation to TB incidence, prevalence, and mortality of a country? How do these relationships differ for high burden countries compared to all countries?

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. This can be completed by subsetting the high burden countries into a smaller data set.

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.

# 4. Methods

The data cleaning will consist of widening the data set so each TB indicator is placed in a column in order to explore the indicators separately. The data set will also be split into a full data set containing all countries (wide\_data) and high burden countries (highburden). The data will then be summarized and described using exploratory analysis approaches such as graphs, plots, and tables. The statistical analysis will use cross-validation and decision tree modeling to model TB outcomes based on the available predictor variables. A similar approach will be used for the health equity indicators as outcomes.

## 4.1 Data aquisition, import, and cleaning

The WHO dataset can be found [here](o%20https://www.who.int/data/inequality-monitor/data#PageContent_C158_Col00) under “Tuberculosis Indicators”. This data was downloaded and stored into the raw\_data folder of the repository as rawdata. Information on the indicators, social determinants dimensions, and subgroups can be found in the raw\_data folder under 202206-metadata-tb.pdf

#path to data  
data\_location <- here::here("data","raw\_data","202206-repository-tb.xlsx")  
  
#load data and assign to rawdata  
rawdata <- readxl::read\_excel(data\_location)  
  
str(rawdata)

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" ...

Please see processingfile.qmd for full code describing data importing and cleaning.

The rawdata file contains information on the country, year of data collection, the indicator they are studying, the social determinants of health they are considering, and the subgroup of each dimension. The dataset is mostly complete with only a few main variables missing data points. The indicators being studied are shown below.

#look at what indicators are in the dataset  
unique(rawdata$indicator\_name)

[1] "BCG immunization coverage among one-year-olds (%)"   
 [2] "People who report TB is spread through coughing (%)"   
 [3] "People who report TB is spread through coughing - Female (%)"   
 [4] "People who report TB is spread through coughing - Male (%)"   
 [5] "Case detection rate (%)"   
 [6] "TB incidence (new infections per 100 000 population)"   
 [7] "TB mortality (deaths per 100 000 population)"   
 [8] "People with MDR/RR-TB (%)"   
 [9] "People who would want a family member's TB kept secret (%)"   
[10] "People who would want a family member's TB kept secret - Male (%)"   
[11] "TB prevalence (cases per 100 000 population)"   
[12] "Prevalence to notification ratio (years)"   
[13] "Families affected by TB facing catastrophic costs due to TB (%)"   
[14] "People who would want a family member's TB kept secret - Female (%)"

For the main part of the data cleaning, I used pivot\_wider() to allow each indicator to have its own column, and subset the high burden countries (as indicated by the WHO) into a new object called highburden. I also checked each indicator using filter(), select(), and summary() functions. These methods will allow for deeper exploration of each indicator in order to explore trends and patterns for each subgroup. An example of the data cleaning is shown below.

wide\_data <- rawdata %>%   
 select(c(setting, year, indicator\_name,  
 indicator\_abbr, dimension, subgroup,  
 estimate, population)) %>%   
 pivot\_wider(names\_from = "indicator\_name", values\_from = "estimate")  
head(wide\_data)

# A tibble: 6 × 20  
 setting year indic…¹ dimen…² subgr…³ popul…⁴ BCG i…⁵ Peopl…⁶ Peopl…⁷ Peopl…⁸  
 <chr> <dbl> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 Afghani… 2010 bcg Econom… Quinti… 532. 53.8 NA NA NA  
2 Afghani… 2010 bcg Econom… Quinti… 549. 61.9 NA NA NA  
3 Afghani… 2010 bcg Econom… Quinti… 495. 58.1 NA NA NA  
4 Afghani… 2010 bcg Econom… Quinti… 473. 65.1 NA NA NA  
5 Afghani… 2010 bcg Econom… Quinti… 447. 77.9 NA NA NA  
6 Afghani… 2010 bcg Educat… No edu… 2267. 61.1 NA NA NA  
# … with 10 more variables: `Case detection rate (%)` <dbl>,  
# `TB incidence (new infections per 100 000 population)` <dbl>,  
# `TB mortality (deaths per 100 000 population)` <dbl>,  
# `People with MDR/RR-TB (%)` <dbl>,  
# `People who would want a family member's TB kept secret (%)` <dbl>,  
# `People who would want a family member's TB kept secret - Male (%)` <dbl>,  
# `TB prevalence (cases per 100 000 population)` <dbl>, …

# BCG coverage indicator  
wide\_data %>%   
 filter(indicator\_abbr == "bcg") %>%   
 select(c(1,2,3,4,5,6,7)) %>%   
 summary()

setting year indicator\_abbr dimension   
 Length:4352 Min. :1991 Length:4352 Length:4352   
 Class :character 1st Qu.:2003 Class :character Class :character   
 Mode :character Median :2009 Mode :character Mode :character   
 Mean :2008   
 3rd Qu.:2014   
 Max. :2019   
   
 subgroup population   
 Length:4352 Min. : 18.9   
 Class :character 1st Qu.: 210.3   
 Mode :character Median : 377.8   
 Mean : 1666.4   
 3rd Qu.: 719.8   
 Max. :696209.9   
 NA's :96   
 BCG immunization coverage among one-year-olds (%)  
 Min. : 5.57   
 1st Qu.: 85.97   
 Median : 93.65   
 Mean : 88.73   
 3rd Qu.: 97.25   
 Max. :100.00   
 NA's :96

Due to the pivot\_wider() function, there was several missing observations for countries and years that had data for only some subgroups (ex: data on males but no data on females). Since I will be maintaining all the indicators in their own data set, drop\_na() would remove all data points. Instead, I will handle missing data points after exploring the data further.

The processed and cleaned data was then saved into the processed\_data folder under processeddata.rda to be used during the exploratory analysis phase. I have loaded the processed data below.

#path to data  
data\_location <- here::here("data","processed\_data","processeddata.rda")  
#load data.   
load(data\_location)

## 4.2 Statistical analysis

For the statistical analysis of the data, I created decision tree models for the main outcomes (TB incidence, TB mortality, and TB prevalence) based on the subgroups of the indicator and tuned the models using cross-validation folds. The data was initially split into test/train data sets to better evaluate model performance. An example is shown below. This method was repeated for the wide\_data and highburden countries.

# fit linear model using TB incidence as outcome, subgroup as predictor  
lm\_mod <- linear\_reg() #set linear regression  
  
#fit model with high burden countries  
inclm\_fit <- lm\_mod %>%   
 fit(`TB incidence (new infections per 100 000 population)` ~ subgroup,  
 data = highburden) #fit linear model  
incfittable <- tidy(inclm\_fit)   
print(incfittable) #produce tidy table of fitted model

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 194. 24.2 8.03 2.49e-12  
2 subgroupMale 126. 34.2 3.69 3.70e- 4

Other indicators were also analyzed and fit to a decision tree model using the same methods.

# fit linear model with BCG coverage as outcome, subgroups as predictors  
  
#fit model with high burden countries  
BCGlm\_fit <- lm\_mod %>%   
 fit(`BCG immunization coverage among one-year-olds (%)` ~ subgroup,  
 data = highburden) #fit linear model  
BCGfittable <- tidy(BCGlm\_fit)   
print(BCGfittable) #produce tidy table of fitted model

# A tibble: 12 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
 1 (Intercept) 88.4 0.884 100. 0   
 2 subgroupMale 0.363 1.25 0.291 7.71e- 1  
 3 subgroupNo education -10.6 1.30 -8.18 4.67e-16  
 4 subgroupPrimary education 0.169 1.28 0.133 8.95e- 1  
 5 subgroupQuintile 1 (poorest) -6.46 1.25 -5.15 2.83e- 7  
 6 subgroupQuintile 2 -1.60 1.25 -1.27 2.03e- 1  
 7 subgroupQuintile 3 1.06 1.25 0.843 3.99e- 1  
 8 subgroupQuintile 4 4.50 1.25 3.59 3.40e- 4  
 9 subgroupQuintile 5 (richest) 7.01 1.25 5.59 2.52e- 8  
10 subgroupRural -2.12 1.25 -1.69 9.06e- 2  
11 subgroupSecondary or higher education 6.66 1.25 5.32 1.14e- 7  
12 subgroupUrban 5.10 1.25 4.08 4.72e- 5

I would still like to try to manipulate the data set so I can use the social determinant indicators as predictors for the main outcomes. This will most likely be limited to the high burden countries.

# 5. Results

## 5.1 Exploratory/Descriptive analysis

Please see exploratory\_analysis.qmd for full code describing data exploratory analysis.

The processedcode was loaded into the exploratory\_analysis.qmd file which pulls coding from the exploratoryanalysis.r script.

#Path to data.  
data\_location <- here::here("data","processed\_data","processeddata.rda")  
#load data  
load(data\_location)

For each indicator, I produced a summary table using skim() to determine the number of dimensions and subgroups as well as the summary statistics to see any overall trends in the data. The summary tables are stored in the results folder. An example of the code is shown below.

#incidence summary and save to file location  
summary\_inc <- skimr::skim(wide\_data$`TB incidence (new infections per 100 000 population)`)  
print(summary\_inc)

── Data Summary ────────────────────────  
 Values   
Name wide\_data$TB incidence (n...  
Number of rows 7473   
Number of columns 1   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Column type frequency:   
 numeric 1   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Group variables None   
  
── Variable type: numeric ──────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate mean sd p0 p25 p50 p75 p100 hist   
1 data 7085 0.0519 107. 147. 0 9.28 40.9 146. 908. ▇▁▁▁▁

#mortality summary and save to file location  
summary\_mort <- skimr::skim(wide\_data$`TB mortality (deaths per 100 000 population)`)  
print(summary\_mort)

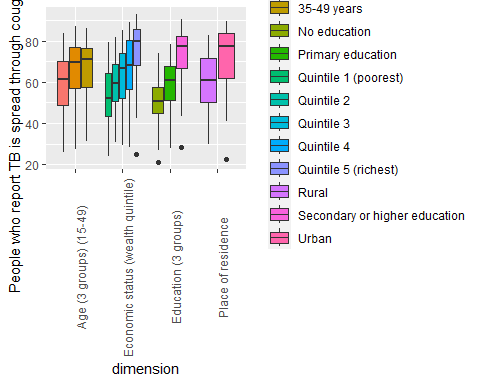
── Data Summary ────────────────────────  
 Values   
Name wide\_data$TB mortality (d...  
Number of rows 7473   
Number of columns 1   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Column type frequency:   
 numeric 1   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Group variables None   
  
── Variable type: numeric ──────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate mean sd p0 p25 p50 p75 p100 hist   
1 data 7085 0.0519 13.2 21.1 0 0.630 3.39 16.9 140. ▇▁▁▁▁

The population for most variables were relatively low and skewed to the right. The TB incidence and mortality indicators also showed a skewed distribution. The highest incidence rate was 908 cases/100,000 and highest mortality rate was 140 deaths/100,000.

After getting an overview of the indicators, I wanted to explore the differences between subgroups and start identifying the largest disparities. Using ggplot(), I created plots of the indicators separated by dimension and subgroup. For most variables, these disparities were described using the boxplot() function as shown below.

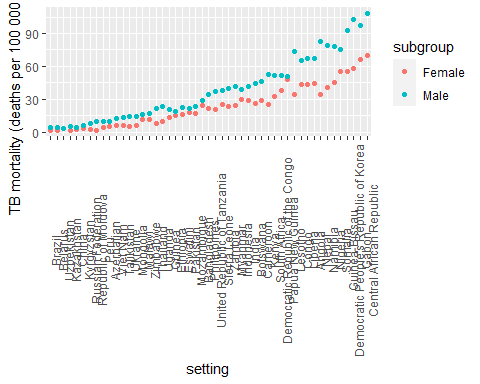
p13 <- highburden %>%  
 filter(!dimension %in% c("Age (2 groups) (0-15+)", "TB drug resistance", "Sex")) %>%  
 ggplot(aes(x=dimension,  
 y = `People who report TB is spread through coughing - Female (%)`,  
 fill = subgroup)) +  
 geom\_boxplot()+  
 theme(axis.text.x = element\_text(angle = 90))   
plot(p13) # plots knowledge about TB by dimension and subgroup for females

Warning: Removed 2504 rows containing non-finite values (`stat\_boxplot()`).



The main outcomes of interest for TB incidence, mortality, and prevalence were plotted using the geom\_point() function as shown below.

p6 <- highburden %>%   
 filter(dimension %in% "Sex") %>%  
 filter(indicator\_abbr == "mortality") %>%   
 select(c(1,2,3,4,5,6,13)) %>%   
 ggplot(aes(x=fct\_reorder(  
 setting, `TB mortality (deaths per 100 000 population)`),  
 y = `TB mortality (deaths per 100 000 population)`,  
 color = subgroup)) +  
 geom\_point()+  
 theme(axis.text.x = element\_text(angle = 90))+  
 scale\_fill\_brewer(palette = "Spectral")+  
 xlab("setting")  
plot(p6) #plots mortality by country from lowest mortality to highest, colored by subgroup



Most indicators showed the greatest disparities due to economic status and education level.

## 5.2 Basic statistical analysis

*update depending on binding outcomes and predictors: will probably do linear model predictions from TB outcomes from one indicator if possible. otherwise will do line*

The linear models summary for TB incidence are listed below.The high burden model has the linear equation of TB incidence = 194 + 126(Male) where male would equal 1. The full data model has the linear equation of TB incidence = 82.6 + 48(Male). Based on the p-value for the subgroup variable, sex has a significant association with TB incidence in both models. Males having higher TB incidence is congruent with the exploratory analysis findings. The full model has smaller coefficients compared to highburden model.

readRDS(file = here("results", "model fit tables", "incfittable.rds")) #TB incidence model for high burden countries

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 194. 24.2 8.03 2.49e-12  
2 subgroupMale 126. 34.2 3.69 3.70e- 4

readRDS(file = here("results", "model fit tables", "incfittableFD.rds")) #TB incidence model for full data

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 82.6 10.4 7.91 2.78e-14  
2 subgroupMale 48.0 14.8 3.25 1.25e- 3

The fitted model summaries below are the TB mortality for high burden countries and full data. For high burden countries, the linear equation is TB mortality = 23.6 + 14.9(Male); for the full data model, the linear equation is TB mortality = 9.98 + 6.43(Male). Sex is a statistically significant predictor of TB mortality (p-values <0.05), but this does not consider any other factors. The next step in my analysis would be to rearrange the data set so I can predict the main outcomes using the sociodemographic indicators.

readRDS(file = here("results", "model fit tables", "mortfittable.rds")) #TB mortality model for high burden countries

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 23.6 3.55 6.65 0.00000000179  
2 subgroupMale 14.9 5.03 2.97 0.00380

readRDS(file = here("results", "model fit tables", "mortfittableFD.rds")) #TB mortality model for full data

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 9.98 1.50 6.64 1.05e-10  
2 subgroupMale 6.43 2.12 3.03 2.65e- 3

The last outcome model I fitted was TB prevalence for high burden countries and the full data. For high burden countries, the linear equation is TB Prevalence = 399 + 151(Urban) with urban residences receiving a 1 and rural residences receiving a 0. The full data model has an equation of TB Prevalence = 410 + 95.5(Urban).

readRDS(file = here("results", "model fit tables", "prevfittable.rds")) #TB prevalence model for high burden countries

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 399. 53.2 7.50 0.0000000361  
2 subgroupUrban 151. 75.3 2.01 0.0544

readRDS(file = here("results", "model fit tables", "prevfittableFD.rds")) #TB prevalence model for full data

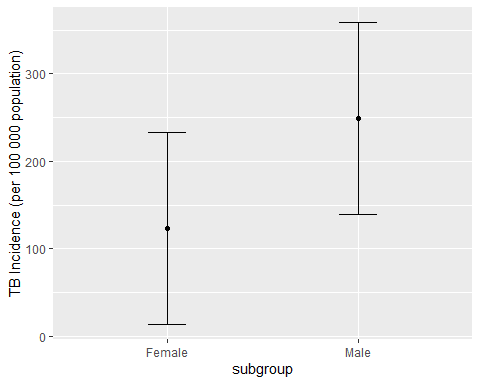
# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 410. 49.2 8.34 4.15e-10  
2 subgroupUrban 95.5 69.6 1.37 1.78e- 1

The models can be further explored by including setting as a predictor, shown in the prediction model below. The prediction model is basic prediction of TB incidence from subgroup and setting. As an example, I predicted the incidence for females and males in India (shown in the plot). The estimates for males (249.3660; CI = 139.68336, 359.0486) and females (123.1038; CI = 13.42112, 232.7864) are shown using geom\_point() and geom\_errorbar(). I need to explore the prediction modelling further before attempting to predict my main outcomes with more variables.

#prediction model for TB incidence with subgroup and setting as predictors  
predinclm\_fit <- lm\_mod %>%   
 fit(`TB incidence (new infections per 100 000 population)` ~ subgroup + setting,  
 data = highburden) #fit linear model  
predincfittable <- tidy(predinclm\_fit)   
print(predincfittable)

# A tibble: 50 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
 1 (Intercept) 287. 54.6 5.27 3.20e- 6  
 2 subgroupMale 126. 15.4 8.18 1.17e-10  
 3 settingAzerbaijan -292. 76.4 -3.83 3.73e- 4  
 4 settingBangladesh -132. 76.4 -1.73 8.97e- 2  
 5 settingBelarus -324. 76.4 -4.24 1.01e- 4  
 6 settingBotswana -111. 76.4 -1.45 1.54e- 1  
 7 settingBrazil -305. 76.4 -4.00 2.21e- 4  
 8 settingCameroon -177. 76.4 -2.32 2.45e- 2  
 9 settingCentral African Republic 189. 76.4 2.47 1.70e- 2  
10 settingChina -293. 76.4 -3.83 3.71e- 4  
# … with 40 more rows

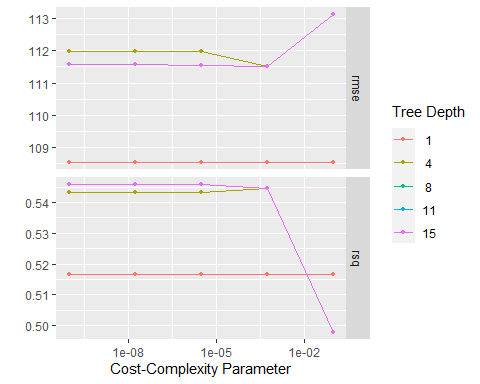
#prediction plot using the above model estimating TB incidence for females and males in India  
readRDS(file = here("results", "model fit tables", "prediction1.rds"))



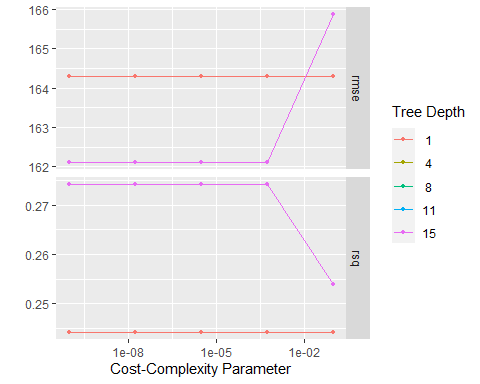
## 5.3 Full analysis

For the majority of the statistical analysis, the tuberculosis outcomes and the health equity indicators with high disparity between subgroups (as identified during the exploratory analysis) were modeled and tuned using the decision tree model and cross-validation methods from the tidymodels guide. The models were all used the training data to tune the cost\_complexity and tree\_depth parameters to find the model with the lowest value of RMSE. An example output is shown below.

#tree plot models for full data TB incidence  
readRDS(file = here("results", "tune plots", "wdinctreeplot.rds"))



#tree plot models for high burden TB incidence  
readRDS(file = here("results", "tune plots", "hbinctreeplot.rds"))



Please refer to the statistical\_analysis.qmd for the full code and results.

A summary of the model performances are shown below. The full data set models are marked with ‘wd’ and the high burden models are marked as ‘hb’. The outcomes being modeled are abbreviated in the model column. The best performing model predicted prevalence to notification ratio for the full data set (RMSE = 0.971). The TB outcome models all performed poorly in predicting TB incidence, mortality, and prevalence. This is most likely due to the only significant predictor being country.

readRDS(file = here("results", "summaryrmse.rds")) #summary table of RMSE from decision tree models

# A tibble: 18 × 4  
 .metric .estimator .estimate model   
 <chr> <chr> <dbl> <chr>   
 1 rmse standard 116. wdinc   
 2 rmse standard 163. hbinc   
 3 rmse standard 20.7 wdmort  
 4 rmse standard 21.6 hbmort  
 5 rmse standard 241. wdprev  
 6 rmse standard 184. hbprev  
 7 rmse standard 5.94 wdbcg   
 8 rmse standard 5.83 hbbcg   
 9 rmse standard 16.4 wdcata  
10 rmse standard 17.8 hbcata  
11 rmse standard 13.3 wdcase  
12 rmse standard 16.2 hbcase  
13 rmse standard 0.972 wdptn   
14 rmse standard 1.03 hbptn   
15 rmse standard 5.56 wdatt   
16 rmse standard 5.82 hbatt   
17 rmse standard 10.0 wdknow  
18 rmse standard 11.0 hbknow

The second research question was to explore the differences between the full data set and high burden countries. I hypothesized that the high burden models would have a higher performance because these countries have the highest concentration of TB cases. However, the high burden models consistently preformed worse than the full data models. This is most likely due to the lower amount of available data.

# 6. Discussion

## 6.1 Summary and Interpretation

TBA: summary of findings and why it matters

## 6.2 Strengths and Limitations

TBA: strengths and limitations of analysis

## 6.3 Conclusions

TBA: take away messages, citations from bibtex

# 7. References

Organization, W. H. (2021a). *Data*. Retrieved from <https://www.who.int/data/inequality-monitor/data#PageContent_C158_Col00>

Organization, W. H. (2021b). *WHO releases new global lists of high-burden countries for TB, HIV-associated TB and drug-resistant TB*. Retrieved from <https://www.who.int/news/item/17-06-2021-who-releases-new-global-lists-of-high-burden-countries-for-tb-hiv-associated-tb-and-drug-resistant-tb>

Organization, W. H. (2022a). *Global tuberculosis report 2022*. Retrieved from <https://www.who.int/publications-detail-redirect/9789240061729>

Organization, W. H. (2022b). *The End TB Strategy*. Retrieved from <https://www.who.int/teams/global-tuberculosis-programme/the-end-tb-strategy>

Organization, W. H. (2022c). *Tuberculosis (TB)*. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/tuberculosis>

World Health Organization. (2021). STATE OF INEQUALITY: HIV, TUBERCULOSIS AND MALARIA. *States News Service*. Retrieved from <https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=edsgin&AN=edsgcl.685964785&site=eds-live&custid=uga1>