Stat 531 Project

Owen Bachhuber, Melanie Mitton, Magdalene Lo, Corban Lethcoe

March 14, 2024

Contents

Chapter 1

Introduction

1.1 Motivation and Data Description

As reported by the New York Times, tuberculosis (TB) has reclaimed the title of the world's leading infectious disease killer as of November 2023, after being briefly removed from its long reign by Covid-19. The World Health Organization reports that nearly 40 percent of people who are living with TB are untreated and undiagnosed. Deaths from tuberculosis remain pronounced, with 1.36 million people dying from tuberculosis in 2022.

This project aims to better contextualize the relationship between government spending and tuberculosis on a country-by-country basis. We hypothesize that countries that spend more on healthcare per person also see fewer deaths from tuberculosis.

Two data sets were chosen for this analysis:

1. The estimated numbers of deaths caused by TB (all forms) among 100,000 residents during the given year from 2000 to 2015 sorted by country.

2. The average health expenditures per person paid by government entities from 1995 to 2010 expressed in US dollars using the average exchange rate.

The data sets were joined by country, and only complete observations from 2000 to 2010 were included.

1.2 Sources

- Nolen, S. (2023, November 6) Ending TB Is Within Reach — So Why Are Millions Still Dying? *The New York Times*. `https://www.nytimes.com/2023/11/06/health/tuberculosis-tb-treatment-vaccine-diagnosis.html`
- TB deaths per 100,000, estimated, all forms of TB: World Health Organization through gapminder.org
- Govt. health spending per person (US$): World Health Organization through gapminder.org

1.3 Data Cleaning

The following code was used to clean the data set:

```
library(tidyverse)
library(here)
library(ggplot2)
library(gridExtra)
library(gganimate)
library(gifski)

# Read data
tb_deaths_long <- read_csv(here("all_forms_of_tb_deaths_per_100000_estimated.csv"))
gov_health_spending_long <- read_csv(here("government_health_spending_per_person_us.csv"))

# Pivot the Tuberculosis Deaths Data
tb_deaths_long <- tb_deaths_long |>
  pivot_longer(cols = -country, names_to = "year", values_to = "deaths_per_100k", names_prefix = "X") |>
  mutate(year = as.integer(year)) |>
  drop_na()

# Pivot the Government Health Spending Data
gov_health_spending_long <- gov_health_spending_long |>
  pivot_longer(cols = -country, names_to = "year", values_to = "gov_health_spending_usd", names_prefix = "X") |>
  mutate(year = as.integer(year)) |>
  drop_na()
```

```r

# Join the cleaned, long-format datasets
combined_dataset_long_cleaned <- left_join(tb_deaths_long, gov_health_spending_long, by = c("country", "year"))

# Remove any remaining rows with NAs
combined_dataset_long_cleaned <- drop_na(combined_dataset_long_cleaned)

# Write the cleaned, combined dataset to a CSV file
write_csv(combined_dataset_long_cleaned, here("combined_dataset_long_cleaned.csv"))
```

Chapter 2

Data Visualization

2.1 Scatter Plot Over Time

To examine how the relationship between our two variables of interest changed over time, we created an animated plot that shows deaths per 100k vs government spending per person in USD for every country in every year of the study (2000 - 2010).

The following code was used to generate the plot:

```
plot <- combined_dataset_long_cleaned |>
  ggplot(aes(
    x = gov_health_spending_usd,
    y = deaths_per_100k
  )) +
  geom_point() +
  labs(title = "Deaths Per 100k vs Gov Spending - Year: {frame_time}",
       y = "",
       x = "Government spending per person, USD") +
  transition_time(year) +
  ease_aes('linear') +
  theme(aspect.ratio = 1)+
  theme_bw()

# Create the animation object
anim <- animate(plot,nframes = 10 , duration = 20, renderer=gifski_renderer())

# Save the animation
anim_save("Animation_trials/Generated_animations/animation.gif",
```

```
    animation = anim)
anim
```

Over time, the amount of government spending per person increased on average. The relationship between deaths per 100k and government spending displays an exponential decay overtime, suggesting that as government spending increases, the number of deaths per 100k decreases. To confirm this inference, we created density plots of deaths per 100k and government spending per person to show how the proportional distribution changes overtime.

2.2 Density Plot Over Time

The following code was used to generate the density plots:

```
## Density plot of deaths per 100k
plot_2 <- combined_dataset_long_cleaned |>
  ggplot(aes(x = deaths_per_100k)) +
  geom_density(fill = "skyblue", color = "black") +
  labs(title = "Density Plot of Deaths Per 100k - Year: {frame_
    time}") +
  transition_time(year) +
  ease_aes('linear') +
  theme(aspect.ratio = 1)

anim_2 <- animate(plot_2,nframes = 10 , duration = 20, renderer=
    gifski_renderer())

# Density plot of gov spending.
plot_3 <- combined_dataset_long_cleaned |>
  ggplot(aes(x = gov_health_spending_usd)) +
  geom_density(fill = "skyblue", color = "black") +
  labs(title = "Density Plot of Gov Spending Per Person in USD -
     Year: {frame_time}") +
  transition_time(year) +
  ease_aes('linear') +
  theme(aspect.ratio = 1)

anim_3 <- animate(plot_3,nframes = 10 , duration = 20, renderer=
    gifski_renderer())
anim_2
anim_3
```

From these plots, it becomes clearer to see that as government spending increased overtime, the number of deaths per 100k decreased.

Chapter 3

Linear Regression

3.1 Visualization

Before making the linear model, we wanted a general idea of what it would look like, so we fit a line to the scatter plot of average deaths per 100k vs. average government spending per person.

The following code was used to generate this plot:

```
combined_dataset_long_cleaned |>
  group_by(country) |>
  summarise(mean_deaths = mean(deaths_per_100k),
            mean_spending = mean(gov_health_spending_usd)) |>
  ggplot(aes(x = mean_deaths, y = mean_spending)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("Mean Government Spending in USD") +
  ylab("") +
  ggtitle("Mean Deaths Per 100 k")
```

3.2 Making the Model

Now that we had a better idea of what the linear regression would look like, we made the model.

The following code was used to create the linear regression model:

```
combined_dataset_long_av <- combined_dataset_long_cleaned |>
```

```
  group_by(country) |>
  summarise(av_deaths_per_100k = mean(deaths_per_100k),
            av_gov_health_spending_usd = mean(gov_health_
    spending_usd)) |>
  ungroup()

combined_dataset_lm <- lm(av_deaths_per_100k ~ av_gov_health_
    spending_usd ,
                          data = combined_dataset_long_av)

broom::tidy(combined_dataset_lm)
```

3.3 Regression Results and Equation

average deaths per 100k = 44.52 - 0.017(average govhealth spending usd)

3.4 Interpretation

There is a negative correlation between Deaths per 100K and Government Spending. Our model suggests that for every $1000(USD) spent by a government on healthcare, 17 people per 100,000 can be saved from dying of TB.

With an R-squared value of about 0.051, our suspicions from the regression visualization are confirmed and our linear model is not a great fit for the data. Though our model is a poor predictor of the data trends, we can tell from our P-values (both ¡ 2e-16) that there is a very strong correlation between these two variables.