Predicting Tuberculosis Fatality Rate

Stat Squad

Introduction

Tuberculosis is a deadly disease that can have extreme consequences if not identified and treated properly. This increasingly deadly disease has motivated us to consider what specific factors affect the fatality rate for TB cases. We investigate several factors such as HIV rate, GDP, and government spending on health care. In terms of government spending on healthcare, according to an article by the National Institutes of Health, "A hospital-based TB diagnosis is a critical opportunity to identify those at high risk of early and overall mortality" (Osman, 2021, pg. 1). In addition, for HIV "recent data estimates show that 3-7 million HIV patients develop TB per year and up to 5 million people develop acute pulmonary TB" (Obeagu, 2023, pg. 128). Lastly, GDP per capita "was highest for the type IV countries (high-income countries), which indicates that a lower TB incidence is accompanied by increasing affluence" (Lei, 2023, pg. 5).

These are our 3 research questions based on the evidence we found: Does a higher HIV rate in a country make the tuberculosis death rate higher? Does a country having a low GDP make the tuberculosis death rate higher? Do the countries with the least amount of government spending on healthcare have the highest tuberculosis death rate?

A majority of our data sets were recorded by countries themselves then were collected by various organizations such as the WHO and IHME. This data was then cleaned and checked for accuracy by these organizations then published for public use. The political regime data set was collected by the organization V-Dem. V-Dem gave surveys to approximately 25 political experts in each country. These experts then ranked their countries based on various metrics then V-Dem averaged the scores together to assign each country a political regime. V-Dem experts also went through the data and checked for accuracy through their own experts.

Methods and Analysis

Prior to building our model and performing any analysis, we decided that it was best to log transform gdp, hc_expenditure, hiv, and multidrug_resistant_tb. We made this transformation because in our EDA we concluded that those explanatory variables had an exponential relationship with the response variable.

In the first stage of our model building process we explored a model containing only the quantitative explanatory variables. Before hypothesizing this model, we analyzed our explanatory variables for the presence of multicollinearity. A correlation matrix of the quantitative explanatory variables showed that gdp and hc_expenditure had a strong pairwise relationship (correlation of 0.96). The VIF of gdp was 12.47 and the VIF of hc_expenditure was 13.37, and the mean VIF of our quantitative variables was 6.53. Because the mean VIF was greater than 3 and the VIF of gdp and hc_expenditure was greater than 10, we concluded that there was a significant concern for multicollinearity. To resolve this concern we split our model building process into two branches: one branch would explore models with gdp and another would explore models with hc_expenditure. We made this decision because we realized that gdp and hc_expenditure would provide overlapping information to the model.

Next, we utilized stepwise regression (with p_ent and p_rem at 0.15) to determine which quantitative variables to include in our model. We executed stepwise regression twice - once for the model containing gdp and once for the model containing hc_expenditure. The stepwise regression process removed multidrug_resistant_tb for both models. Afterwards, we tested the possibility of including a gdp x rate_of_new_tb interaction or a hc_expenditure x rate_of_new_tb interaction for each respective model, but the individual t tests for the corresponding parameters were not significant at a 0.05 significance level.

In the next stage of our model building process we added the qualitative variables: political_regime, hemisphere, and majority_religion. Since hemisphere was described with just one dummy variable, we tested its significance with an individual t test and concluded it was not significant for both models. To test the significance of political_regime and majority_religion we utilized the nested F test since those qualitative variables had more than two levels and thus multiple dummy variables in the models. The nested F tests determined that both variables were not significant for the model built with gdp and the model built with hc_expenditure. In our EDA, we noticed that there was potentially a significant interaction between hemisphere and political regime, but we found this interaction to not be significant across both models through nested F testing. All testing was done at a 0.05 significance level. At this point, no qualitative variables remained in either model.

In the last stage of our model building process, we explored qualitative x quantitative interactions. In our EDA, we noticed that there was potentially a significant interaction between gdp and political regime and hc_expenditure and political regime. We tested these interactions with nested F tests, and it was determined that these interactions were significant at a 0.05 significance level, so we kept those interactions in the models. At this point, our gdp model and hc_expenditure model were essentially identical, with gdp and hc_expenditure being interchangeable across the two models. We decided to select the model containing gdp to be our final model because it had a slightly lower RMSE and slightly higher adjusted R-squared.

All of our assumptions were met except for a slight violation of the constant variance assumption. The normality assumption was met because the data points did not stray from the middle line of the QQplot and the histogram was unimodal and roughly symmetric. Based on the residual plots, there did not appear to be any concerning trends, thus the lack of fit assumption was met. For the independence assumption, we did not have time series data. For the constant variance assumption, we found a minor violation in the rate_of_new_TB residual plot. There was a small case of fanning out, and we log transformed rate_of_new_tb to fix the issue. However, after correcting the violation and transforming an explanatory variable, we realized that it made our model less statistically significant. The Adjusted R^2 value decreased and the RSME increased. We made the decision to keep the minor violation to have our model stay as statistically significant as possible. After analyzing the assumptions, we used the cooks distance models and the influence plots to find our outliers. We had 5 outliers in our model: Equatorila Guinea, Ghana, Qatar, Sudan, and United Arab Emirates. We removed these outliers and we noticed an improvement in our models Adjusted R^2 and a decrease in the RMSE.

For our additional techniques, we used weighted least squares regression. For our weight we used the inverse of the residuals squared. This significantly overfitted our model (Adj-R^2: 0.99) so we weren't able to proceed with this additional technique. We then tried weighting with different explanatory variables, but this didn't help either.

Results

See Appendix C for the final model. The analysis reveals a compelling association between higher HIV prevalence, an increased number of new tuberculosis cases, and elevated fatality rates, if all other variables are held constant. Notably, the impact of GDP on fatality rates is nuanced within specific political regimes.

The electoral autocracy, electoral democracy, and liberal democracy regimes exhibit higher fatality rates compared to the baseline, however an increase in gdp for those regimes causes the fatality rate to decrease, as evinced by the negative coefficients for these interaction terms. The baseline political regime, closed autocracy, has a contrasting relationship with fatality rate - the fatality rate increases as gdp increases in countries with a closed autocracy. The complex interplay between a country's GDP and political regime underscores the intricate relationship between economic development and health outcomes across different political landscapes. Despite the model's statistical significance, indicated by a remarkably low p-value (<0.001), the adjusted R^2 value suggests that 36.25% of the variability in fatality rates is accounted for in the model. This discrepancy may signify the necessity for additional variables, more sophisticated modeling techniques, or the inherent high variability within the dataset.

Conclusions

$$
\begin{aligned}
\widehat{fatality\_rate} = &- 1.11 + 0.76 \log(\text{gdp}) + 0.42 \log(\text{hiv}) + 0.01(\text{rate\_of\_new\_tb}) && (1)\\
&+ 35.71 \text{electoral\_autocracy} + 46.38 \text{electoral\_democracy} + 67.25 \text{liberal\_democracy} && (2)\\
&- 3.86 \log(\text{gdp}) \times \text{electoral\_autocracy} - 4.79 \log(\text{gdp}) \times \text{electoral\_democracy} && (3)\\
&- 6.46 \log(\text{gdp}) \times \text{liberal\_democracy} && (4)
\end{aligned}
$$

Higher HIV prevalence, and a greater number of new tuberculosis cases are associated with higher fatality rates, if all other variables are held constant. However, the impact of GDP on fatality rates depends on the political regime of a country. As GDP increases for a country with an electoral autocracy, electoral democracy, or liberal democracy, the fatality rate decreases. The opposite trend occurs for the baseline political regime, closed autocracy. This suggests a complex interplay between economic development and health outcomes under different political regimes. Overall, all of our research hypotheses were correct. Higher HIV rates correspond with higher death rates, and in most cases (except for the closed autocracy case), higher gdps correspond with lower fatality rates.

The model is statistically significant as indicated by the very low P-value, suggesting that the predictors have a meaningful contribution to the model. However, the Adjusted R^2 value shows that 36.25% of variability in the fatality rate has been captured by the model. This might suggest the need for additional variables, more complex modeling, or that inherent variability in the dataset is high. The prediction equation was tested to discover how accurately it could predict the fatality rate of TB in Austria based on its GDP (55806.43), HIV rate (17405.05), rate of new TB cases (6.0), and political regime (liberal democracy). Our model predicts a fatality rate of 8.05%, and the actual fatality rate is 8.00%, resulting in a residual of 0.05.

While our model is significant, there are significant improvements that could be made to improve the amount of variation that is accounted for by our model. To help account for variations in the data we should include more explanatory variables to our model. These variables could include metrics measuring average air quality in countries and alcohol consumption in countries as these two facts can impact the fatality rate of Tuberculosis. Furthermore, our model would be more accurate if we were able to have access to more recent data as opposed to data from 2019. Lastly, we could do further research to implement a more complex modeling procedure to hopefully make our model more accurate.

Appendix A: Data Dictionary

Variable Name	Abbreviated Name	Description	Units	Levels (if Qualitative)
Fatility Rate	Fatality Rate	Percentage of deaths among diagnosed tuberculosis cases	Fatality percentage	
GDP per capita	GDP	The gross domestic product that measures a country's economic well-being	Currency in international-$	
Healthcare expenditure per capita	HC expenditure	The amount spent on healthcare services divided by a country's population	Currency in international-$	
Number of people living in a country with HIV	HIV	The number of people living with diagnosed HIV (human immunodeficiency virus) in each country	Number of people living with diagnosed HIV	
Multidrug resistant Tuberculosis	Multidrug Resistant TB	A type of TB that is resistant to at least 2 different types of anti-TB drugs	Number of people with diagnosed multidrug resistant TB	
Rate of New Cases	Rate of New TB	The proportion of people recently diagnosed with TB per 100,000 people in each country	Proportion of people with recently diagnosed TB	
Political Regime	Political Regime	A set of rules, protocols, and cultural norms that regulate how a government functions	N/A	Closed autocracy, Electoral autocracy, Electoral democracy,Liberal democracy
Hemisphere	Hemisphere	Hemisphere that a country belongs to	N/A	Northern, Southern

Variable Name	Abbreviated Name	Description	Units	Levels (if Qualitative)
Dominant Religious Affiliation	Majority Religion	Dominant religion of a country	N/A	Christianity, Islam, Buddhism, Hinduism, Judaism

Appendix B: Data Rows

	country	year	gdp	fatality_rate	hc_expenditure	hiv	multidrug_resistant_tb	rate_of_new_tb	political_regime	hemisphere	majority_religion
1	Afghanistan	2019	2079.922	14	285.5581	5125.301	1762.45100	189.0	electoral autocracy	North	Islam
2	Algeria	2019	11627.280	11	750.4487	9485.271	221.38165	61.0	electoral autocracy	North	Islam
3	Angola	2019	6602.424	18	178.0261	383909.750	3405.68300	351.0	electoral autocracy	South	Christianity
4	Argentina	2019	22071.748	6	2198.8804	188657.250	120.33472	29.0	electoral democracy	South	Christianity
5	Armenia	2019	14317.553	6	1616.1779	1389.422	214.82237	26.0	electoral democracy	North	Christianity
6	Australia	2019	49379.094	4	5294.4630	17305.775	37.55193	6.8	liberal democracy	South	Christianity

Appendix C: Tables and Figures

$$\begin{aligned}
\widehat{fatality_rate} = & -1.11 + 0.76\log(\text{gdp}) + 0.42\log(\text{hiv}) + 0.01(\text{rate_of_new_tb}) && (5)\\
& + 35.71\text{electoral_autocracy} + 46.38\text{electoral_democracy} + 67.25\text{liberal_democracy} && (6)\\
& - 3.86\log(\text{gdp}) \times \text{electoral_autocracy} - 4.79\log(\text{gdp}) \times \text{electoral_democracy} && (7)\\
& - 6.46\log(\text{gdp}) \times \text{liberal_democracy} && (8)
\end{aligned}$$

```
Call:
lm(formula = fatality_rate ~ gdp + hiv + rate_of_new_tb + political_regime +
    gdp * political_regime, data = log_final)

Residuals:
     Min       1Q   Median       3Q      Max
-13.6915  -4.0349  -0.8722   4.2954  19.1930

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -1.110501  27.323862  -0.041  0.96764
gdp                                      0.763102   2.802289   0.272  0.78580
hiv                                      0.420769   0.245460   1.714  0.08880
rate_of_new_tb                           0.012539   0.004763   2.632  0.00947
political_regimeelectoral autocracy     35.714044  27.796633   1.285  0.20107
political_regimeelectoral democracy     46.376506  28.217396   1.644  0.10261
political_regimeliberal democracy       67.245036  36.710152   1.832  0.06920
gdp:political_regimeelectoral autocracy -3.864839   2.903092  -1.331  0.18536
gdp:political_regimeelectoral democracy -4.793861   2.935421  -1.633  0.10479
gdp:political_regimeliberal democracy   -6.459627   3.641875  -1.774  0.07838

(Intercept)
gdp
hiv                                     .
rate_of_new_tb                          **
political_regimeelectoral autocracy
political_regimeelectoral democracy
political_regimeliberal democracy       .
gdp:political_regimeelectoral autocracy
gdp:political_regimeelectoral democracy
gdp:political_regimeliberal democracy   .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.264 on 134 degrees of freedom
Multiple R-squared:  0.4026,    Adjusted R-squared:  0.3625
F-statistic: 10.04 on 9 and 134 DF,  p-value: 0.00000000001082
```

Appendix D: References

Background

1. Lei, Y., Wang, J., Wang, Y., & Xu, C. (2023). Geographical evolutionary pathway of global tuberculosis incidence trends. BMC Public Health, 23(1), 755.

2. Obeagu, E. I., & Onuoha, E. C. (2023). Tuberculosis among HIV Patients: A review of Prevalence and Associated Factors. Int. J. Adv. Res. Biol. Sci, 10(9), 128-134.

3. Osman, M., Meehan, S. A., von Delft, A., Du Preez, K., Dunbar, R., Marx, F. M., Boulle, A., Welte, A., Naidoo, P., & Hesseling, A. C. (2021). Early mortality in tuberculosis patients initially lost to follow up following diagnosis in provincial hospitals and primary health care facilities in Western Cape, South Africa. PloS one, 16(6), e0252084. https://doi.org/10.1371/journal.pone.0252084

4. Pai, M., Dewan, P. K., & Swaminathan, S. (2023). Transforming tuberculosis diagnosis. Nature Microbiology, 8(5), 756-759.

Data

1. Healthcare expenditure vs. GDP per capita. Our World in Data. (n.d.). https://ourworldindata.org/grapher/healthcare-expenditure-vs-gdp?tab=table

2. Northern Hemisphere countries 2024. (n.d.). https://worldpopulationreview.com/country-rankings/northern-hemisphere-countries

3. Number of people living with HIV. Our World in Data. (n.d.-b). https://ourworldindata.org/grapher/number-of-people-living-with-hiv?tab=table Pew Research Center. (2022, December 21). Religious composition by country, 2010-2050. Pew Research Center's 4. Religion & Public Life Project. https://www.pewresearch.org/religion/interactives/religious-composition-by-country-2010-2050/

4. Saloni Dattani, Fiona Spooner, Hannah Ritchie and Max Roser (2023) - "Tuberculosis" Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/tuberculosis' [Online Resource]

5. Southern Hemisphere countries 2024. (n.d.). https://worldpopulationreview.com/country-rankings/southern-hemisphere-countries V-Dem (2023) – with major processing by Our World in Data. "Political regime – Regimes of the World" [dataset]. V-Dem, "Democracy and Human rights, OWID based on Varieties of Democracy (v13) and Regimes of the World v13" [original data]. Retrieved March 23, 2024 from https://ourworldindata.org/grapher/political-regime

6. WHO (2023); Population based on various sources (2023) – with minor processing by Our World in Data. "Rate of new tuberculosis cases" [dataset]. WHO, "Global Tuberculosis Report"; Various sources, "Population" [original data]. Retrieved March 23, 2024 from https://ourworldindata.org/grapher/incidence-of-tuberculosis-sdgs

7. World Bank (2023) – with minor processing by Our World in Data. "GDP per capita – World Bank" [dataset]. World Bank, "World Bank World Development Indicators" [original data]. Retrieved March 23, 2024 from https://ourworldindata.org/grapher/gdp-per-capita-worldbank

Supplemental Code and Analysis Help

1. ChatGPT