Author: Andy Duong
Email: aduong.cs@gmail.com

Problem Statement

One of the questions during the Covid-19 pandemic was why different countries have such different proportions of deaths relative to the number of disease cases. There were several hypotheses about this, including differences in medical facilities (countries with more hospitals should see fewer deaths), age demographics (countries with a larger proportion of older people will see more deaths), and existing mortality rates between countries (indicating pre-existing differences in overall health between countries).

The goal of this project was to use linear modeling to quantify some of the variation in mortality from Covid-19 in different counties due to differences in hospital infrastructure and demographics.

Data Wrangling

My process of data wrangling started with the data for global confirmed cases. I read the data using the function `read_csv()` and then used `pivot_longer()` because it included column headers as values, in this case, the dates where data was collected. I then moved on to read in the data for deaths. For the same reason as the data for confirmed cases, I pivoted the data for deaths and then did a full join with the confirmed and deaths data into a new table that I named `covid`. This table will eventually be the final product after data wrangling. I then cleaned up the table, discarding unnecessary data and renaming columns, to later make joins.

The next step was to read in the hospital beds data. I aggregated the data to only include data from the most recent year for each country. I then renamed columns to make them easier to read and fixed the names of countries to be consistent with my other tables. I then did an inner join of this table with my `covid` table, discarding countries that didn't have sufficient data and making the names of countries consistent.

Linear Modeling

As the number of confirmed cases is the most relevant predictor, I first created a model using it as the single independent variable and the number of deaths as the dependent variable to get a control before seeing how the other predictor variables affect the deaths. This gave me a multiple r-squared value of 0.8921, which I will use to compare with models created with different combinations of predictors. All the combinations of predictors include the number of confirmed cases since other predictors have little effect on the number of deaths without the number of cases. Because of this, we will always have more independent variables than our control case and the r-squared value should always be higher. We will see what combinations of predictors with the number of confirmed cases has the highest effect on the number of deaths.

Author: Andy Duong
Email: aduong.cs@gmail.com

I chose to create models based on the following combinations of predictors:

combination of predictors	resulting r-squared value	notes
cases	0.808	control
cases + life expectancy	0.808	
cases + mortality	0.808	
cases + beds	0.8115	
cases + POP.65UP (normalized)	0.809 (0.8085)	
cases + POP.1564 (normalized)	0.8093 (0.8084)	
cases + PROP.65UP + beds (normalized)	0.8131 (0.812)	
cases + PROP.1564 + beds (normalized)	0.8127 (0.8139)	

I first created models using cases and life expectancy as predictors. The resulting r-squared value did not change from by control which suggests that existing life expectancy has little to no effect on the death rate of Covid-19 within a country. The same occurred when using existing mortality as a predictor.

I then created a model using cases and hospital bed density as predictors. For this analysis, hospital bed density is essentially a measure of hospital infrastructure in a country. This single predictor ended up creating the largest change in the r-squared value suggesting that the varying death rate of Covid-19 is greatly influenced by hospital infrastructure.

I chose to focus on two different age groups. Those between ages 15 and 64, and those of age 65 and up. I chose these age groups because it is commonly known that people over the age of 65 should be a high-risk group for covid and people between ages 15 and 64 would usually be at a lower risk.

I added the age groups to be an independent variable of the model, but the resulting r-squared values were not as I expected. When using the age group of 15 to 64 as a predictor, the resulting r-squared value was higher than that of when using the age group of those over 65. Adding in hospital bed density flipped the comparison, resulting in a higher r-squared value for the age group of those over 65. I tried transforming the age groups data to be proportions of the countries' total populations. When using this predictor as an independent variable, the resulting r-squared values became closer to what was expected. However, when using these transformed proportions as predictors with hospital bed density, the comparison flipped again, with proportion of those between 15 to 64 resulting in a higher r-squared value. My guess for why this happened is that different age groups may demand resources differently from hospitals.

With this analysis, I figured that of the different hypotheses previously covered, medical facilities are the reason that different countries have such varying death rates from Covid-19.