**Introduction**

This case study will focus on the analysis of Life Expectancy data, and on creating predictive models for life expectancy.

For our first objective, we will build a model in order to identify key relationships along with all necessary testing and interpretations.

For our second objective, we will compare multiple models in order to develop the best fitting model, that can highly predict life expectancy for the entire population of Earth.

**Nota Bene**

The Life Expectancy data set is strictly observational data. As is the case with all observational studies, this type of study has no intervention nor treatments done by researchers. This study would provide descriptive and exploratory data, in sharp contrast to randomized studies. We can only observe the overall population as people cannot be forced to be vaccinated, nor have HIV nor Aids, nor assigned to how long one should live or die. It is most important to mention that observational studies cannot be used to make causal statements, and thus with our study we will explore possibilities and draw valid inferences that relate to Life Expectancy.

**Matrix**

**Data Description**

The data were provided by Dr. Turner for his MSDS 6372 class. The data can be found under Files, Project 1, Life Expectancy Data (1).csv within Digital Campus for course DS 6372 Applied Statistics. The data set is made up of 2938 observations and 22 explanatory variables.

\*No independence bc same country over time, repeated measures \* doesn’t take into account since the repeated measures of the same country are correlated , also year to year sampling will contribute to correlation and is outside the scope as it is related to Repeated Measures which will be addressed in future models and courses. Country is a categorical variable with

**Objective 1**

**Restatement of Problem**

As part of our first objective, we will build a model in order to identify key relationships along with all necessary testing and interpretations.

**Preparation**

We start by reading in the data set, getting a sense of the overall data trend as we are interesting in in Life Expectancy as our response variable. We proceed to check all observations for missing points such as NAs. Upon visualizing the data and its missing data, we select variables that will be imputed based on a threshold of about 10% for missing data per variable. All variables, except for 5—Country, Year, Status, Life Expectancy and Hepatitis B—were selected to be imputed as they variables have numeric outputs and fall under the 10% threshold. We proceed to then remove 2 columns—Hepatitis B and Population from the data set and then remove any remainder NAs which were found our response variable—Life Expectancy. Research suggests that it is inadvisable to impute response variables, therefore we opted to only remove the <0.01% remaining NAs. Given this information we felt that the 19% for Hepatitis B and the 22% for Population were too great that we would risk biasing our overall data. Population also is related to the GDP within the data set. The GDP provided appears to be GDP per capita, as research would suggest, and GDP per capita is calculated by using GDP/Population, which population is a variable in our data set. GDP per capita and Population would be too closely related and would possibly attribute to covariance.

(Our 9 variables Variables to consider are HIV, Schooling, Alcho (Alcohol), BMI, Polio, Diphtheria, GDP, Thinness 1-19, Income Composition of Resources.)

Interpretation of coefficients since logged GDP, Practical vs Statistical significance of the predictors

**Intercept Interpretations:**

Intercept

We estimate that the median Life Expectancy for an individual with none of the variables present would theoretically be 43.6420 years old (p-value < 2e-16). We are 95% confident that the Life Expectancy is between 42.43 and 44.85 years.

HIV.AIDS

For every 1,000 live births, there is an increase in HIV.AIDS, where we estimate that the mean of Life Expectancy will decrease by 68.0044 (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true estimate is between 64.10 and 71.90 children who are born with HIV.AIDS.

Schooling

For each year increase in Schooling, the mean of Life Expectancy will increase by 0.8440 (p-value (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true estimate is between 0.7377 and 0.9505.

BMI kg/m^2

For each average kg/m^2 increase in BMI for the population, the mean of Life Expectancy will increase by 0.0561 (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true estimate is between 0.0437 and 0.0685.

Diphtheria

For one percent immunization coverage among 1-year olds vaccinated using DTP3 for Diphtheria, the mean of Life Expectancy will increase by 0.0675 (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true estimate is between 0.0587 and 0.0762.

logGDP

For a doubling of logGDP per capita, the mean of Life Expectancy will change by 0.599579\*ln (2) = 0.4155965 (p-value < 2e-16) holding all other variables constant. We are 95% confident that the true multiplicative increase is between 0.46073069 \*ln(2) = 0.3193542 and 0.73842704 \*ln(2) = 0.5118386.

Thinness

For 1 percent increase in Thinness (10-19 years) among children in adolescence, the mean of Life Expectancy will decrease by 0.1237 (p-value 4.58e-06). We are 95% confident that the true estimate is between 0.1765 and 0.0709.

Income Composition of Resources

For every unit increase in Income Composition of Resources, the mean of Life Expectancy will increase by 7.9522 (p-value < 2e-16). We are 95% confident that the true estimate is between 6.3250 and 9.5793.

Model Conclusion

Within our tailored regression model, it contains independent variables that are statistically significant as each variables p-value is less than our 0.5% level of significance except for Alcohol (p-value 0.814291) and Polio (p-value 0.136951).

(add convo about variables that are not added from dr. turner due to out of scope for now / year is related to time series/ )

≅)

NOTA BENE:

Income Composition of Resources (something to talk about when we fine-tune our model in part 2 as it can be heavily correlated due to it incorporating Life Expectancy, Education, Per Capita Income – variables that are used in the model for part 1 but also since it incorporates L.E which happens to be our Responsse)

The Human Development Index is a statistic composite index of life expectancy, education, and per capita income indicators, which is used to rank countries into four tiers of human development.

**World Map**

Life Expectancy per Country Plot

In this visual example we are interested in seeing the true values for Life Expectancy, which is our response variable, for each country that is within our cleaned data set. This color-coded map only concentrates on the variable Life Expectancy in reference to Country and no other variable.

Country’s Status: Developed v. Developing

This visualization shows us the global story of which countries are considered developed and which countries are considered developing. It should be well noted that Status is a categorical variable, as we only have two options. It would also appear, from visual inference, that our data are Eurocentric in comparison to the rest of the world. Of course, there are a few outliers too.

Income Composition of Resources per Country

In this final visualization we wanted to compare Income Composition of Resources per country as it is a variable that is closely related to Life Expectancy, Education, and Per Capita Income which happen to be other variables within our dataset. It should be well noted that there are many similarities shared between Income Composition of Resources per Country and the results of Life Expectancy per Country.