Life Expectancy

DS 6371 Applied Stats

Stephanie Duarte, Caleb Thornsbury, Steven Cox

Table of Contents

Defining Why

The purpose behind this data analysis

Making Objectives

Setting goals and ensuring relevance

Data Analysis

Analyzing data in a manner that helps the audience understand

The Big Picture

Gaining a comprehensive understanding of patterns, trends, and implications from this dataset

Defining Why

Problem Identification

What is the problem were trying to solve? What are some issue in the data that will make this difficult

Continuous Improvement

What can we do to improve on the models? Which predictors are important in this?

Pattern Recognition

What are some patterns we can see in and outside the data? Is the outside influence relevant to our analysis?

Data Analysis

The Life Expectancy dataset consists of 22 Columns and 2928 rows from years 2000-2015 for 183 countries. All predicting variables was then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors.

Making Objectives

01

Identify Key Relationships

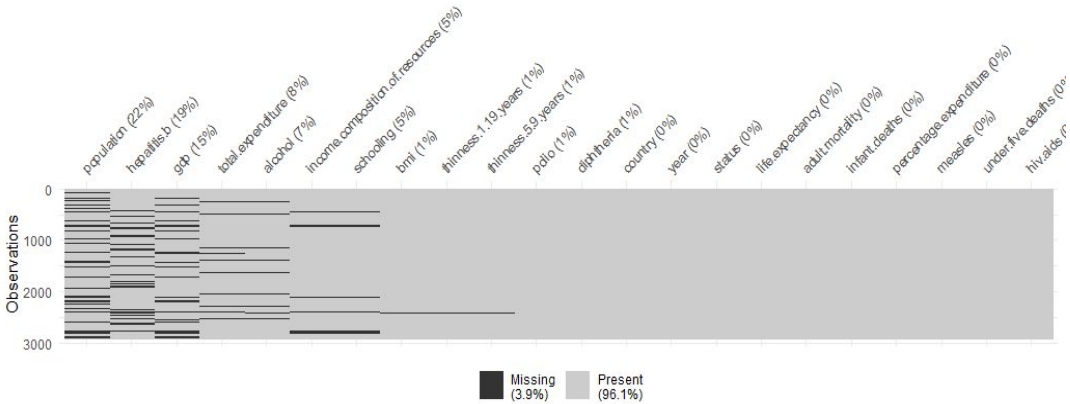
Build a Regression Model with the purpose to Identify and interpret key Relationships in the data.

02

Compare Multiple Models

Build multiple models to compare the predictive nature with the original Regression model

Missing Analysis



name	Correlation	missing_count
schooling	0.7276300	160
income.composition.of.resources	0.7210826	160
bmi	0.5420416	32
thinness.1.19.years	0.4578382	32
thinness.5.9.years	0.4575083	32
gdp	0.4413218	443
alcohol	0.4027183	193
diphtheria	0.3413312	19
polio	0.3272944	19
hepatitis.b	0.1999353	553
total.expenditure	0.1747176	226
population	0.0223050	644

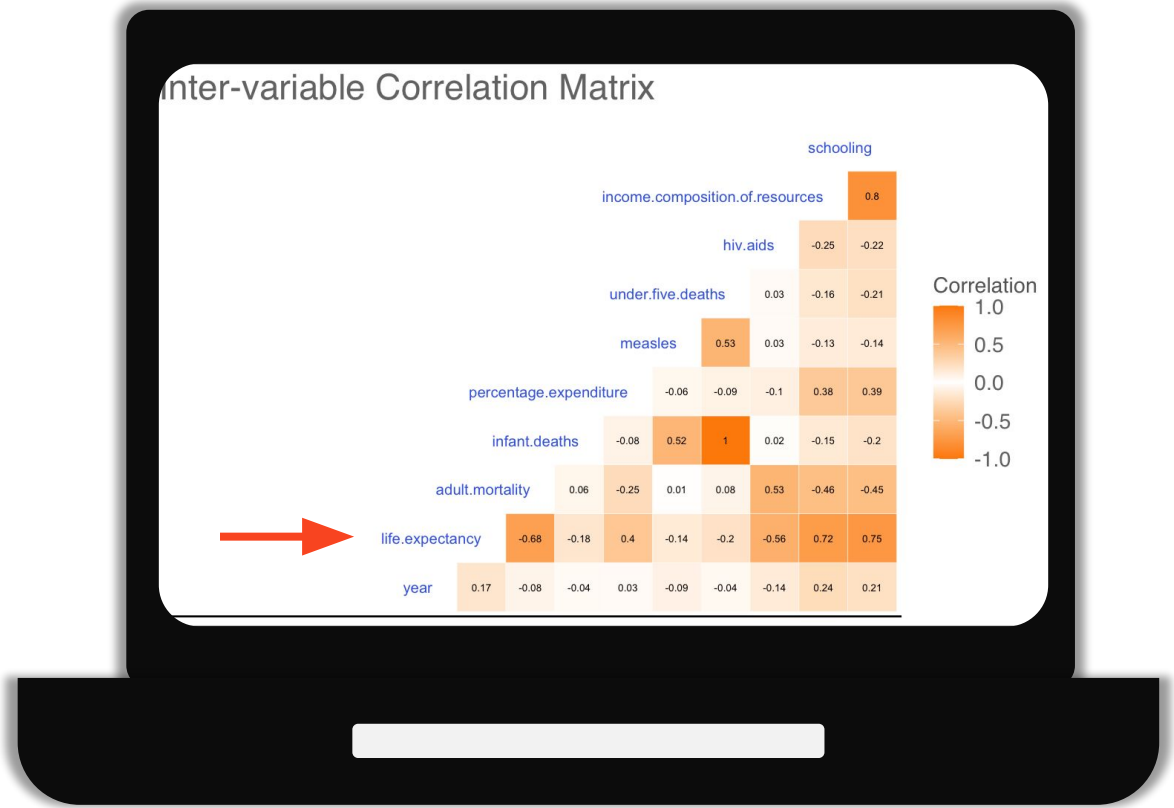
- Of the remaining 160 missing data points within income and schooling, we found those to be the same observations. After removing those, our dataset now has a size of 2768 observations with 10 predictors

country	year
0	0
status	life.expectancy
0	0
adult.mortality	infant.deaths
0	0
percentage.expenditure	measles
0	0
under.five.deaths	hiv.aids
0	0
income.composition.of.resources	schooling
160	160

Correlation Analysis

Numerical predictors of interest

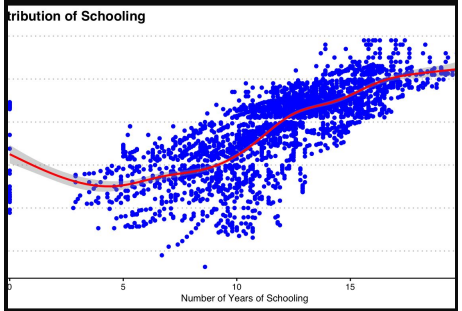
- Schooling
- Income
- Adult Mortality
- HIV/AIDS



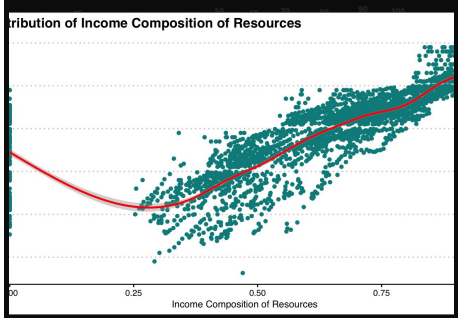
Predictors displaying very high correlation is infant deaths and deaths under five. We removed infant deaths after further inspection showing they represented the same information.

Visualizing Distributions

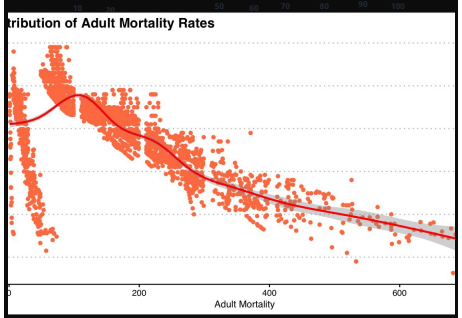
Schooling



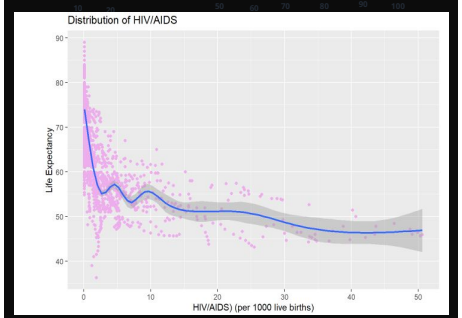
Income



Mortality



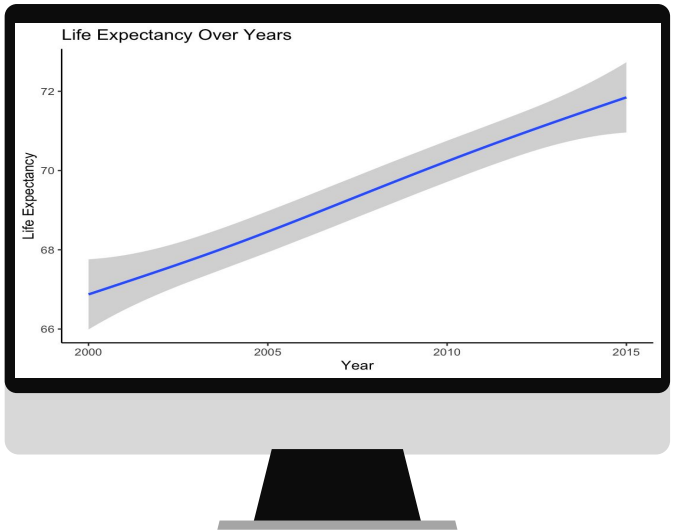
HIV



There is a big concern regarding the outliers and how they are affecting the correlation. However it is unclear on the best way to approach these, therefore we are interested in how the modeling will handle the extreme cases.

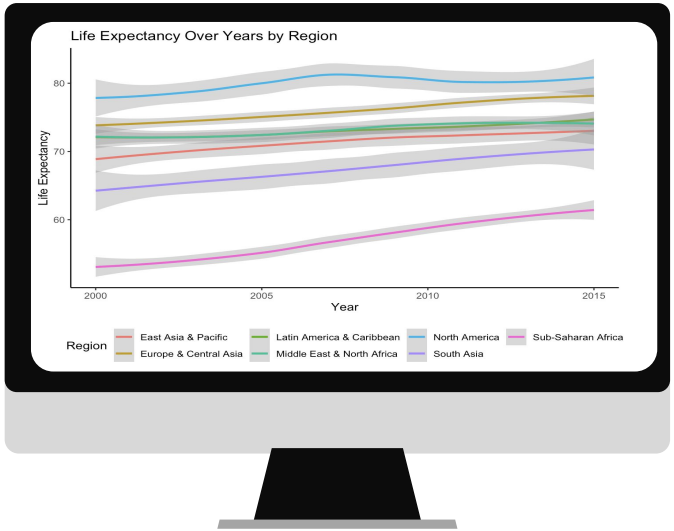
Visualizing Distributions

Life Expectancy by Year



The correlation matrix earlier gave us a value of only 0.17, however here we can see that there is definitely a strong positive relationship between a person's life expectancy and the year.

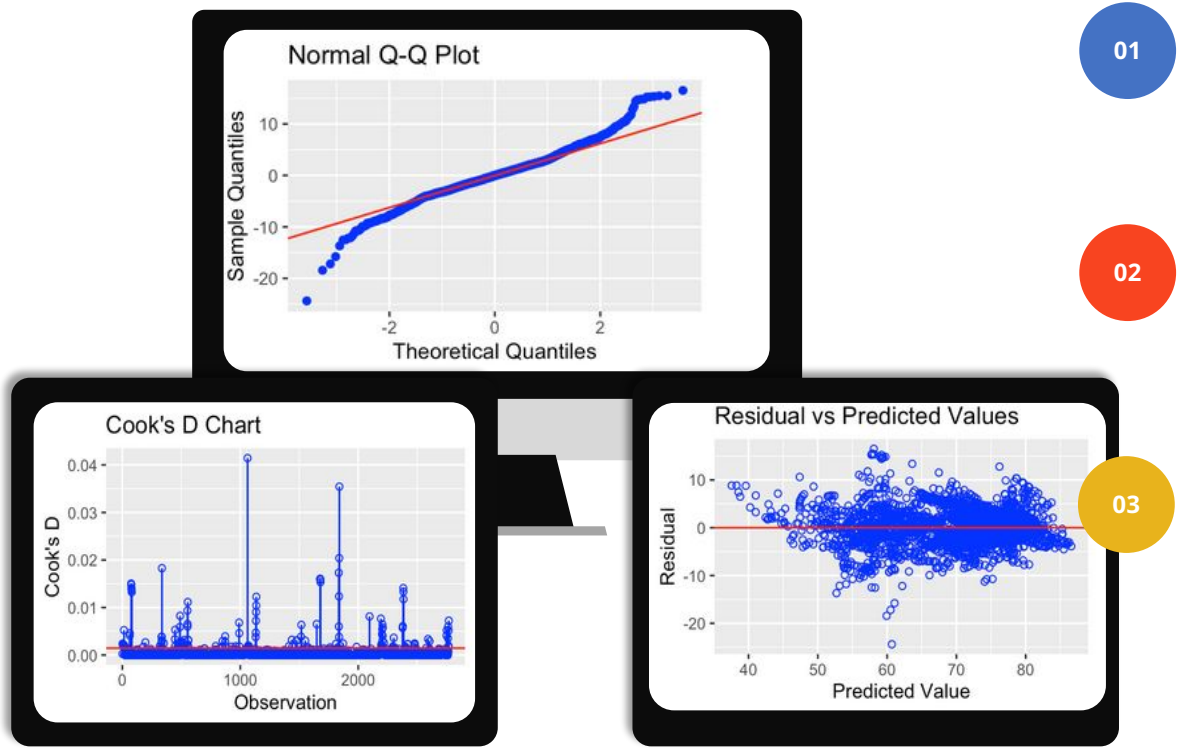
Life Expectancy by Region



Taking this a step further, we broke down Countries by Region. This gives us a good visual on the relationship between Regions and life expectancy over the years.

*Note: Visualizing the 183 unique countries didn't seem practical

Objective 01



01

Residual vs Fitted plot shows a random cloud spread, meaning this model is a good fit.

02

The QQ Plot shows little evidence of non-linearity particularly at the ends of the distribution.

03

The Cook's Distance plot shows data points that have high influence over the dataset, these will need to be removed

Objective 01

Identify and Interpret Key Relationships

We can see from the R code that the relationships with the most significance include:

- Status Developing
- Adult.Mortality
- Infant.deaths
- Alcohol

The calculated RMSE for this model is 3.81.

Objective 02

Glmnet

The glmnet model we used shows final values used for the model were alpha = 0.1 and lambda = 0.01407562.

α =.1
λ =.014
RMSE =3.7

KNN

The KNN model we used did a 10 fold cross validation from 1-30, picking K=2 as the best representation of the data

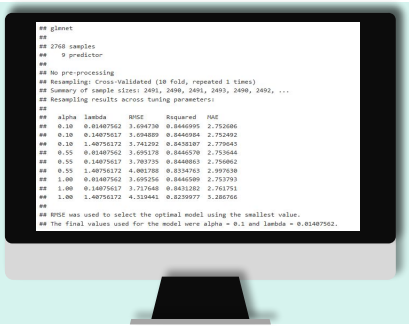
k =2
RMSE=2.3

	2.5%	97.5%
HIV/AIDS	-40	-33
Schooling	65	80
Income	5.2	7.5
Adult Mortality	-152	-123

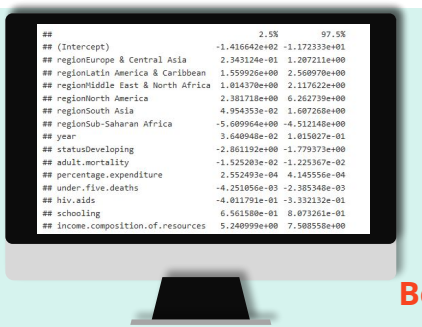
Bootstrapping

These are just a few of the confidence intervals that we found.

The Big Picture



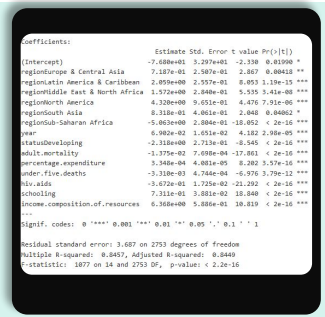
Glmnet



Bootstrapping

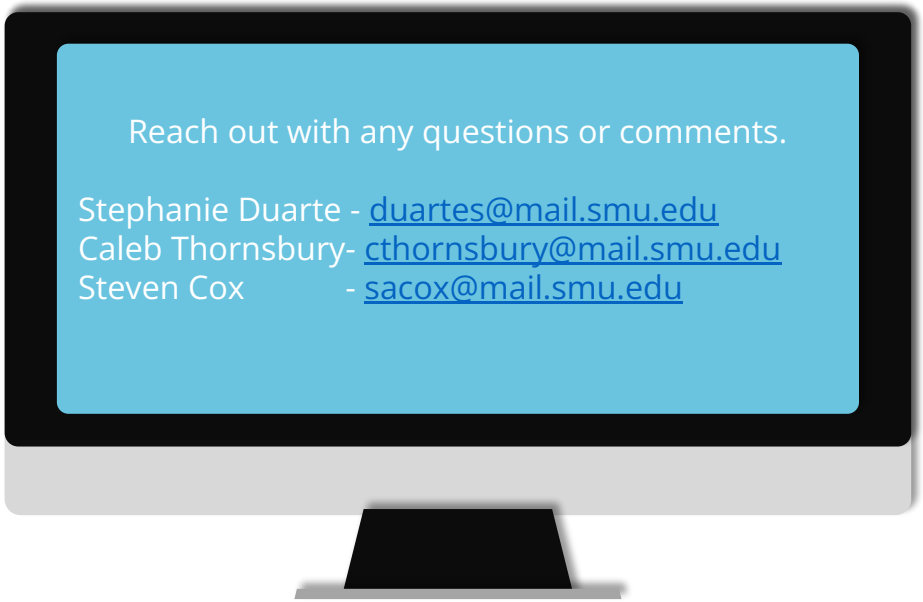
Looking at the parametric and non-parametric models we used, we can see that the models are a good fit for this data

MLR



KNN

Thank you!


Reach out with any questions or comments.
Stephanie Duarte - duartes@mail.smu.edu
Caleb Thornsbury- cthornsbury@mail.smu.edu
Steven Cox - sacox@mail.smu.edu