

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 we're trying to solve?
What are some issues 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.

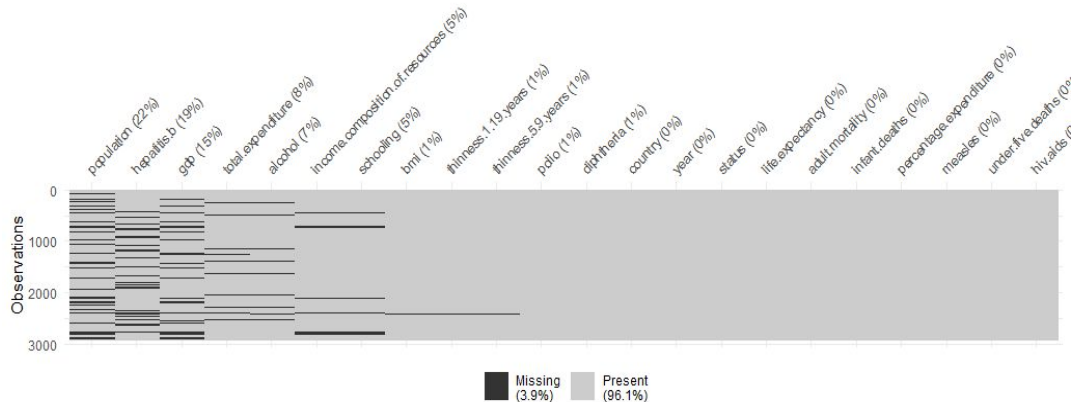
02

Compare **Multiple Models**

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



Missing Analysis



- 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

| 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 |

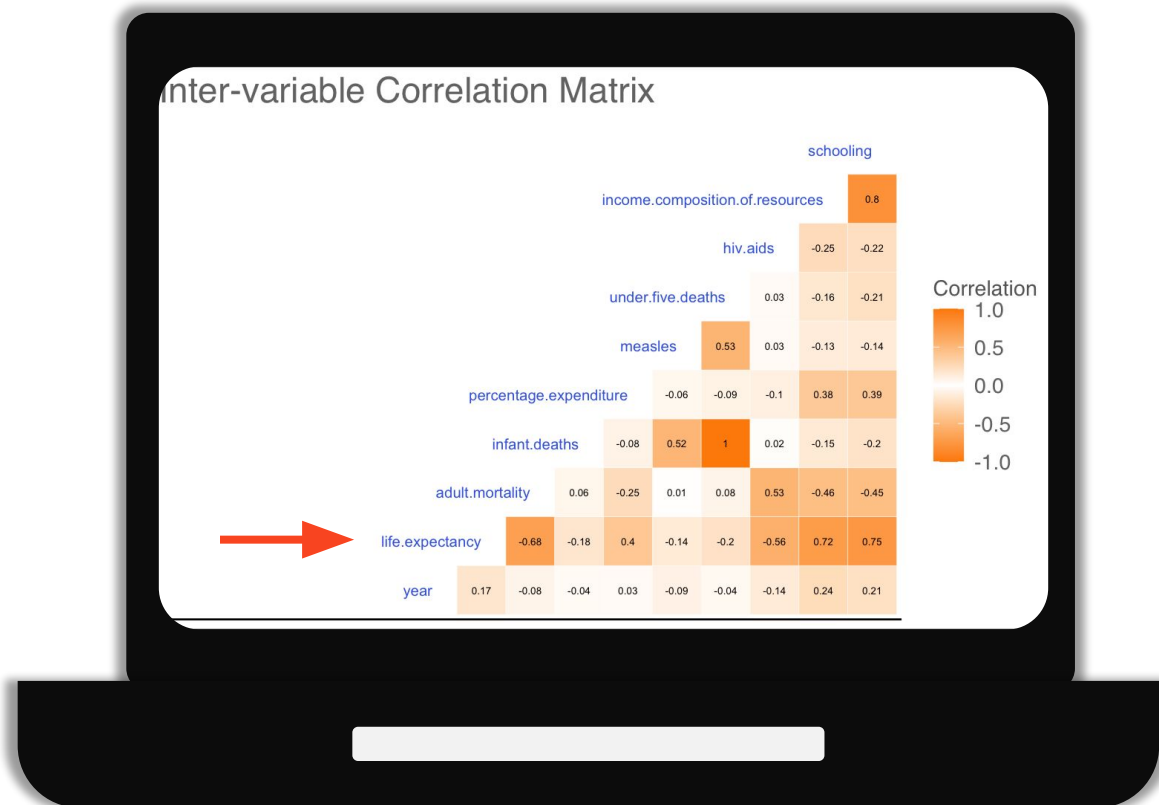
| | |
|---------------------------------|-----------------|
| country | year |
| 0 | 0 |
| status | life.expectancy |
| 0 | 0 |
| adult.mortality | infant.deaths |
| 0 | 0 |
| percentage.expenditure | measles |
| 0 | 0 |
| under.five.deaths | hiv.aids |
| 0 | 0 |
| income.composition.of.resources | schooling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

Correlation Analysis

Numerical predictors of interest

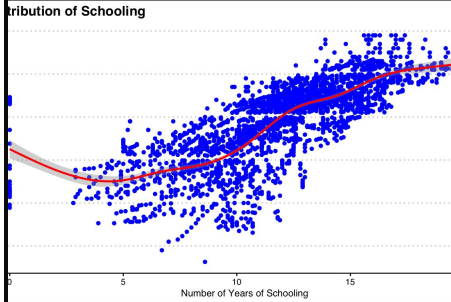
- Schooling
- Income
- Adult Mortality
- HIV/AIDS

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.

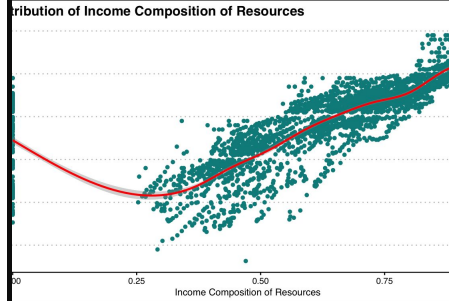


Visualizing Distributions

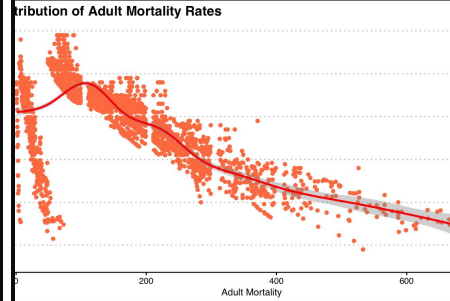
Schooling



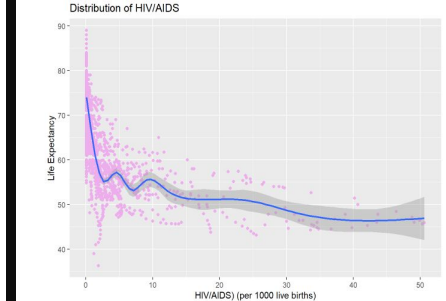
Income



Mortality



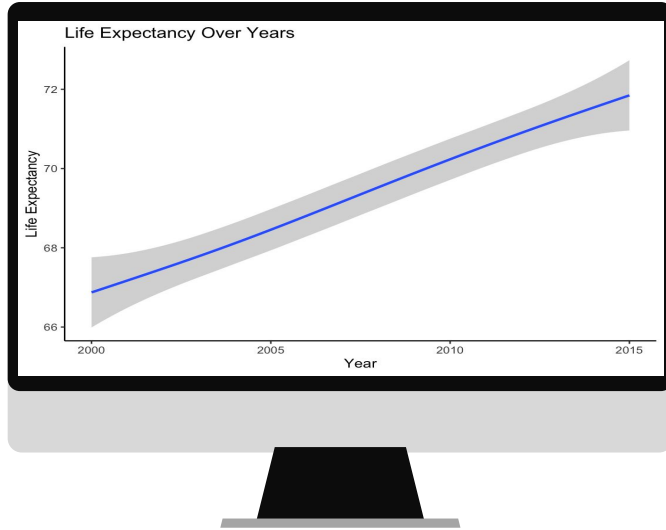
HIV



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.

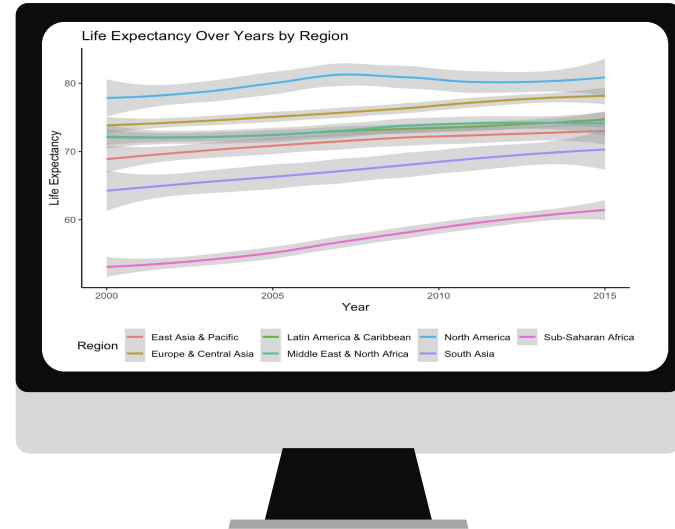
Visualizing Distributions

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.

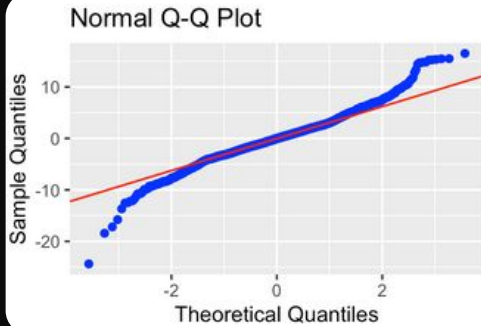
Life Expectancy by Region



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

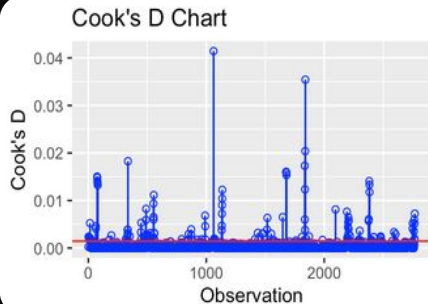


01

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

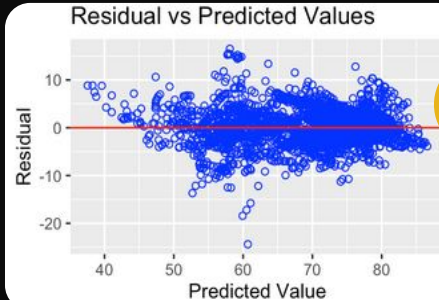
02

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.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.680e+01  3.297e+01  -2.330  0.01990 *
regionEurope & Central Asia  7.187e-01  2.507e-01  2.867  0.00418 **
regionLatin America & Caribbean  2.059e+00  2.557e-01  8.053  1.19e-15 ***
regionMiddle East & North Africa  1.572e+00  2.840e-01  5.535  3.41e-08 ***
regionNorth America  4.320e+00  9.651e-01  4.476  7.91e-06 ***
regionSouth Asia  8.318e-01  4.061e-01  2.048  0.04062 *
regionSub-Saharan Africa -5.063e+00  2.804e-01 -18.052 < 2e-16 ***
year  6.902e-02  1.651e-02  4.182  2.98e-05 ***
statusDeveloping -2.318e+00  2.713e-01  -8.545 < 2e-16 ***
adult.mortality -1.375e-02  7.690e-04 -17.861 < 2e-16 ***
percentage.expenditure  3.348e-04  4.081e-05  8.202  3.57e-16 ***
under.five.deaths -3.310e-03  4.744e-04  -6.976  3.79e-12 ***
hiv.aids -3.672e-01  1.725e-02 -21.292 < 2e-16 ***
schooling  7.311e-01  3.881e-02  18.840 < 2e-16 ***
income.composition.of.resources  6.368e+00  5.886e-01  10.819 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.687 on 2753 degrees of freedom
Multiple R-squared:  0.8457, Adjusted R-squared:  0.8449
F-statistic: 1077 on 14 and 2753 DF, p-value: < 2.2e-16
```

Objective 02

Glmnet

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

$\alpha = .1$
 $\lambda = .014$
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.

Glmnet

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

Bootstrapping

MLR

KNN

```

# Neighbor
n = 2760 samples
p = 9 predictor

Pre-processing: centered (12), scaled (12), Van-Der Waerden Transformation
Principal component analysis extraction (12), remove (12)
Resampling: Cross-validated (18 fold)
Number of sample sizes: 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, ...
Number of resampling iterations: 100
Number of resampling iterations per sample size: 100
Number of resampling iterations per sample size and iteration: 100

k   RMSE      Required RMSE
1   0.5647000   0.5644399   4.01512
2   0.272508   0.0499356   1.95967
3   0.151238   0.0306486   1.07667
4   0.103735   0.0189409   0.578135
5   0.048475   0.0120284   0.28464
6   0.249322   0.0230761   0.64251
7   0.089897   0.0129747   0.717139
8   0.242766   0.0230193   0.781339
9   0.508868   0.0408612   1.04521
10  0.508868   0.0408612   1.04521
11  0.508868   0.0408612   1.04521
12  0.508868   0.0408612   1.04521
13  0.508868   0.0408612   1.04521
14  0.508868   0.0408612   1.04521
15  0.508868   0.0408612   1.04521
16  0.508868   0.0408612   1.04521
17  0.508868   0.0408612   1.04521
18  0.508868   0.0408612   1.04521
19  0.508868   0.0408612   1.04521
20  0.508868   0.0408612   1.04521

RMSE used to select the optimal model using the smallest value.
The final value used for the model was k = 2.

```

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

