WESLEYAN UNIVERSITY

CAPSTONE PROJECT

The relation between macro and micro-level health importance and tuberculosis treatment success rate

Yan Anderson Siriano Duarte

A capstone project submitted in fulfillment of the requirements for the Specialization Certificate of Data Analysis and Interpretation

December 9, 2016

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Tuberculosis (TB) remains a major global health problem. In 2012, 1.3 million people were believed to have died because of tuberculosis with an estimated 8.6 million new cases of TB worldwide (McIntosh and Webberley, 2015).The number of TB deaths is unacceptably large given that most are preventable (*Global tuberculosis report* 2013).

The purpose of this project is to enforce and determine what measures of health care are related to the tuberculosis treatment. With this in mind, the idea is to discover how does the relation of the importance give to health of individuals, governments and companies influence the success rate in the treatment of tuberculosis.

As it is a dangerous disease that has a good chance of prevention, it would be interesting to have some measures that countries could take to increase the success rate in the treatment.

Chapter 2

Methods

2.1 Sample

To make this project, the QoG Standard Data set 2016 (Teorell et al., 2016) was used. This data set consists of approximately 2500 variables from more than 100 data sources. The used variables was extracted from four different database:

- Environmental Performance Data (EPI) (*Environmental Performance Data Set* 2016);
- International Monetary Fund (IMF) (*International Monetary Fund Data Set* 2014);
- World bank - World Development Indicators (WDI) (*Worldbank - World Development Indicators Data Set* 2016);
- World Economic Forum (WEF) (*World Economic Forum Data Set* 2015).

In the QoG Standard CS data set, data from 2012 is prioritized, however, if no data are available for a country for 2012, data for 2013 is included. If no data for 2013 exists, data for 2011 is included, and so on up to a maximum of +/- 3 years.

In the code-book you can find a detailed description of all data sources and variables sorted by original data sources.

Every single variable has a different sample number. The variables with the most samples are Incidence of tuberculosis (per 100,000 people), Air Quality and Water and Sanitation with N=191. The variable with the lowest samples are Smoking prevalence, females and Smoking prevalence, males with N = 127.

After dropping the countries with miss information, a total of N = 109 was selected to make the research.

2.2 Measures

The response variable of this project was *Tuberculosis treatment success rate (% of new cases)* and the predictors are:

- Health expenditure per capita, PPP (constant 2011 international dollar)
- Water and Sanitation: Access to Drinking Water and Access to Sanitation
- Air Quality: Household Air Quality, Air Pollution - Average Exposure to PM2.5 and Air Pollution
- Smoking prevalence, females (% of adults)
- Smoking prevalence, males (% of adults)
- Business impact of tuberculosis
- Tuberculosis case detection rate (%, all forms)
- Incidence of tuberculosis (per 100,000 people)
- GDP (PPP) (share of world total) (%)

All variables are quantitative and will be used without any management.

2.3 Analysis

The distributions for the predictors and the tuberculosis treatment success rate response variable were evaluated by examining the mean, standard deviation and minimum and maximum values.

Scatter plots were also examined. For test bi-variate associations between individual predictors and the tuberculosis treatment success rate response variable, Pearson correlation were used.

Lasso regression with the least angle regression selection algorithm was used to identify the subset of variables that best predicted the tuberculosis treatment success rate.

As the data set has few samples, the lasso regression model was estimated on the entire data set (N=109). All predictor variables were standardized to have a mean=0 and standard deviation=1 prior to conducting the lasso regression analysis. Cross validation was performed using k-fold cross validation specifying 10 cross validation folds. The change in the cross validation mean squared error rate at each step was used to identify the best subset of predictor variables.

Chapter 3

Results

3.1 Descriptive Statistics

Table 3.1 shows descriptive statistics for the quantitative data analytic variables. The average of the response variable, tuberculosis treatment success rate, was 78.29%, with a minimum success rate of 0% and a maximum of 100%.

Analysis Variable	N	Mean	Std Dev.	Minimum	Maximum
Air Quality	109	78.97	18.73	14.30	100.00
Water and Sanitation access	109	55.56	33.05	2.88	100.00
GDP PPP share of world total	109	0.81	2.44	0.00	19.57
Health expenditure per capita	109	1424.78	1626.97	34.81	8845.18
Smoking prevalence females	109	11.57	10.22	0.40	39.80
Smoking prevalence males	109	34.44	12.83	8.90	71.80
TB case detection rate	109	75.28	17.89	16.00	120.00
Incidence of TB	109	128.83	195.29	1.60	1042.00
TB treatment success rate	109	78.29	15.64	0.00	100.00
Business impact of TB	109	5.24	1.05	2.27	6.84

TABLE 3.1: Descriptive Statistic for Data Analytic Variables.

3.2 Bi-variate Analysis

Scatter plots for the association between the tuberculosis success rate response variable and quantitative predictors (Figure 3.1) revealed that only the variables *GDP PPP share of the world total*, *Smoking prevalence males* and *Incidence of Tuberculosis* increased when the tuberculosis treatment had a greater success rate. However, the other variables decreased when the success treatment rate had a great value.

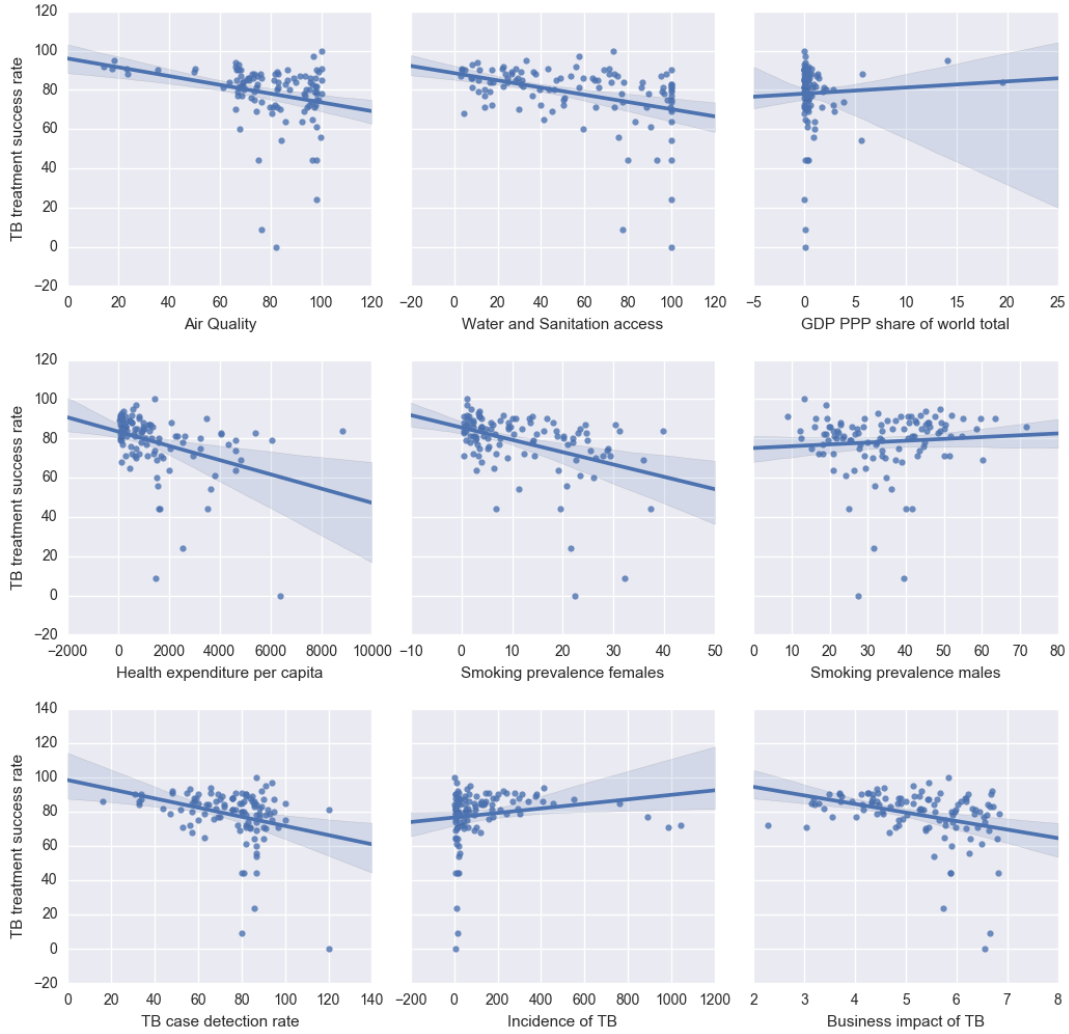

FIGURE 3.1: Association between predictors and tuberculosis success rate.

Analysis Variable	Pearson	p-value
Air Quality	-0.26776	0.0049
Water and Sanitation access	-0.38838	3.0049e-05
GDP PPP share of world total	0.04870	0.6150
Health expenditure per capita	-0.37709	5.3036e-05
Smoking prevalence females	-0.41092	9.0657e-06
Smoking prevalence males	0.07624	0.43071
TB case detection rate	-0.30539	0.00124
Incidence of TB	0.16489	0.08664
Business impact of TB	-0.33497	0.00037

TABLE 3.2: Pearson values of the association between predictors and tuberculosis success rate.

Table 3.2 shows all the Pearson values of the variables. The variables *GDP PPP*

share of world total, *Smoking prevalence males* and *Incidence of tuberculosis* were not significantly associated with the response variable *tuberculosis treatment success rate*.

3.3 Multivariable Analysis

Figure 3.2 shows that only four variables were retained in the model selected by the lasso regression analysis. The other five predictors were excluded. The *Smoking prevalence females (%)* and the *Health expenditure per capita* were most strongly associated with *tuberculosis success treatment rate*, followed by *Air quality* and *Water and sanitation access* (Table 3.3).

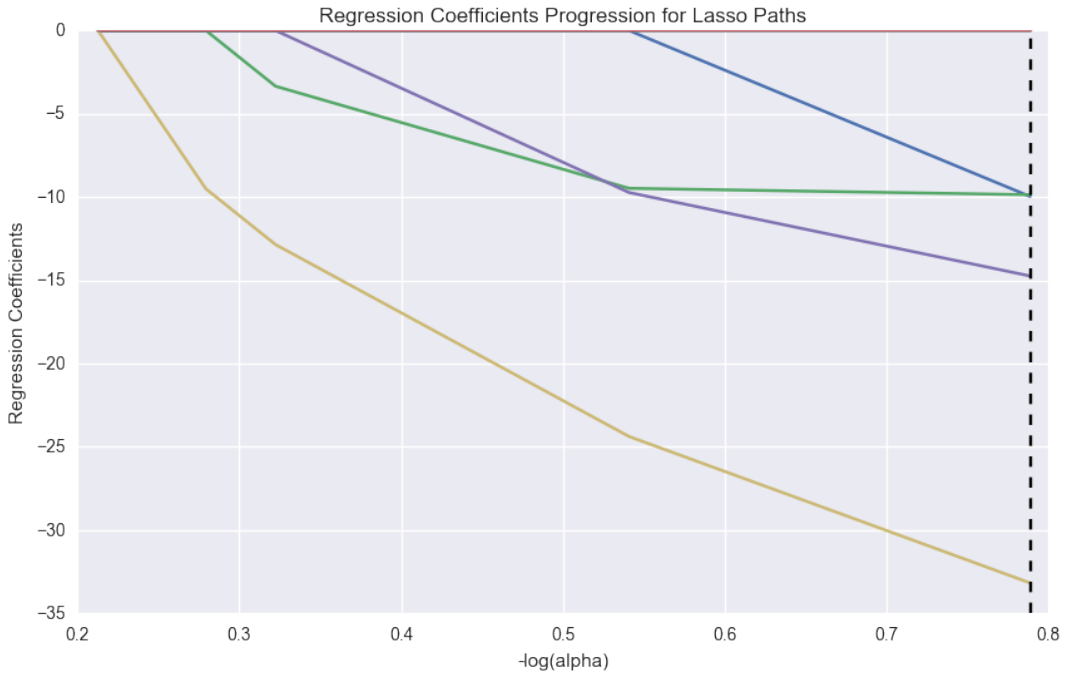

FIGURE 3.2: Regression Coefficients Progression for Lasso Paths.

Analysis Variable	Coef
Air Quality	-0.95440
Water and Sanitation access	-0.94431
Health expenditure per capita	-1.41151
Smoking prevalence females	-3.17685

TABLE 3.3: Lasso Regression Coefficients.

Figure 3.3 shows that there is variability across the individual cross-validation folds in the training data set, but the change in the mean square error as variables are added to the model follows the same pattern for each fold.

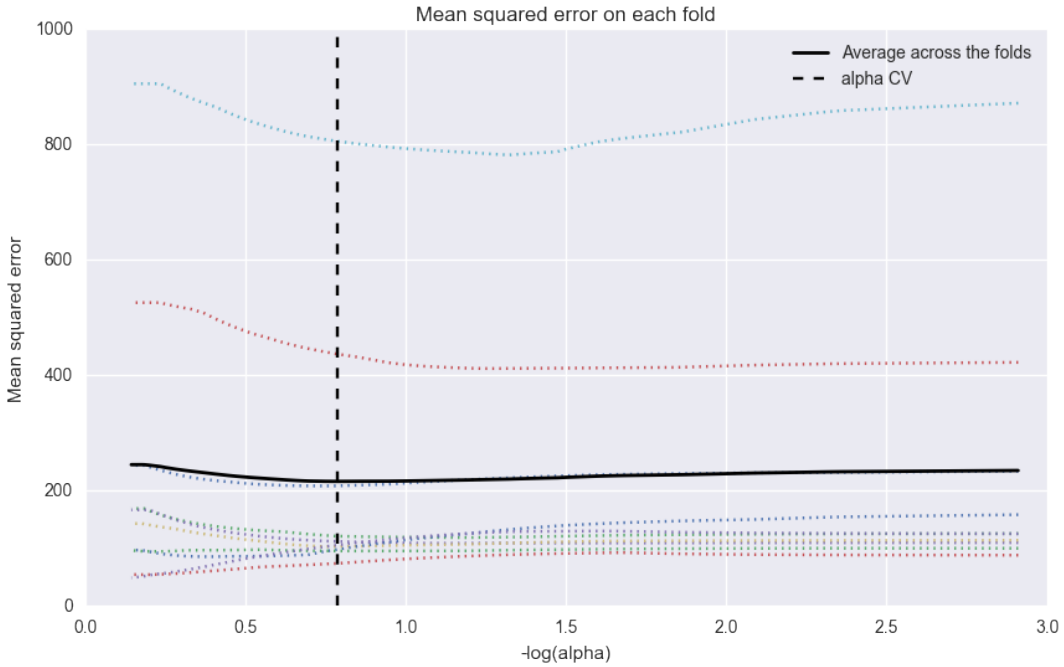

FIGURE 3.3: Mean squared error on each fold.

The mean squared error for the data was MSE = 193.10 and the R-square value was 0.2034, indicating that the selected model explained 20.34% of the variance in tuberculosis success treatment rate for the data set.

Chapter 4

Conclusion

This project used lasso regression analysis to identify a subset of health importance variables that best predicted the success rate of tuberculosis treatment. There were N=109 samples were data was collected between 2009 and 2015.

The lasso regression analysis indicated that 5 of the 9 health importance predictor variable were removed in the final model. The strongest predictors of tuberculosis treatment success rate were Smoking prevalence females (%) and Health expenditure per capita. Both predictors have a negative correlation meaning that if they have a low value, the tuberculosis treatment success rate will be increased.

The results of this project indicate that countries with low air quality, difficult access to water and sanitation, low health expenditure per capita and low % of smoking prevalence female have a better tuberculosis treatment success rate. This is an unexpected and unimpressive result since R-square value was too low indicating that the selected model explained 20.34% of the variance in tuberculosis success treatment rate for the data set.

This project developed a predictive algorithm for tuberculosis success treatment rate that appears to have high bias and variance in a different sample. In addition, it provides more information on which health importance factors are most likely to have a significant impact on tuberculosis treatment success rate. However, there are some limitations that should be taken into account when considering the results of this project. First, some of the data collected had a date divergence between then and may contain data from 2009 to 2015. To maintain consistency, it would be interesting to get all the data for the same year. Second, the number of samples was too low. This can drastically influence the results. Thought that, we can not assume that the predictive algorithm developed in this study will be valid or useful for predicting the tuberculosis treatment success rate. Finally, there is a large number of health

importance related factors that could impact the results of this work, but the current project examined only a few of these factors. Therefore, future efforts to develop a solid predictive algorithm for tuberculosis treatment success rate should expand the algorithm by adding more health importance related predictors to the statistical model.

Bibliography

Environmental Performance Data Set (2016). URL: http://epi.yale.edu/downloads.

Global tuberculosis report (2013). URL: http://apps.who.int/iris/bitstream/10665/91355/1/9789241564656_eng.pdf.

International Monetary Fund Data Set (2014). URL: http://www.imf.org/external/pubs/ft/weo/2014/01/weodata/weoselgr.aspx.

McIntosh, J. and H. Webberley (2015). *Tuberculosis: Causes, Symptoms and Treatments*. URL: http://www.medicalnewstoday.com/articles/8856.php.

Teorell et al. (2016). *The Quality of Government Standard Dataset*. URL: http://qog.pol.gu.se/data/datadownloads/qogstandarddata.

World Economic Forum Data Set (2015). URL: http://reports.weforum.org/global-competitiveness-report-2014-2015/.

Worldbank - World Development Indicators Data Set (2016). URL: http://data.worldbank.org/data-catalog/world-development-indicators.