

## MSSC 5780: Project Proposal

Due 12/01/2023 Fall 2023

Edward Liu and Anesti Sotirovski

We will be using the life expectancy data set from the WHO to do complete option DA, data analysis, of the online data set to compute multiple types of regression analysis. Here is an example of what kind of parameters exist in the data set provided in table 1 below:

Country	Year	Status	Life expectancy	Alcohol	Hepatitis B	BMI	Diphtheria	HIV/AIDS	GDP	Population	Income composition of resources	Schooling
Afghanistan	2015	Developing	65	0.01	65	19.1	65	0.1	584.25921	33736494	0.479	10.1
Afghanistan	2014	Developing	59.9	0.01	62	18.6	62	0.1	612.696514	327582	0.476	10
Afghanistan	2013	Developing	59.9	0.01	64	18.1	64	0.1	631.744976	31731688	0.47	9.9
Afghanistan	2012	Developing	59.5	0.01	67	17.6	67	0.1	669.959	3696958	0.463	9.8
Afghanistan	2011	Developing	59.2	0.01	68	17.2	68	0.1	63.537231	2978599	0.454	9.5
Afghanistan	2010	Developing	58.8	0.01	66	16.7	66	0.1	553.32894	2883167	0.448	9.2
Afghanistan	2009	Developing	58.6	0.01	63	16.2	63	0.1	445.8932979	284331	0.434	8.9
Afghanistan	2008	Developing	58.1	0.03	64	15.7	64	0.1	373.3611163	2729431	0.433	8.7
Afghanistan	2007	Developing	57.5	0.02	63	15.2	63	0.1	369.835796	26616792	0.415	8.4
Afghanistan	2006	Developing	57.3	0.03	64	14.7	58	0.1	272.56377	2589345	0.405	8.1
Afghanistan	2005	Developing	57.3	0.02	66	14.2	58	0.1	25.2941299	257798	0.396	7.9
Afghanistan	2004	Developing	57	0.02	67	13.8	5	0.1	219.1413528	24118979	0.381	6.8
Afghanistan	2003	Developing	56.7	0.01	65	13.4	41	0.1	198.7285436	2364851	0.373	6.5
Afghanistan	2002	Developing	56.2	0.01	64	13	36	0.1	187.84595	21979923	0.341	6.2
Afghanistan	2001	Developing	55.3	0.01	63	12.6	33	0.1	117.49698	2966463	0.34	5.9
Afghanistan	2000	Developing	54.8	0.01	62	12.2	24	0.1	114.56	293756	0.338	5.5
Albania	2015	Developing	77.8	4.6	99	58	99	0.1	3954.22783	28873	0.762	14.2

Table 1 above

We will first make scatter plots of each predictor against the Y variable being life expectancy (years) and see if there is any correlation between things like, Country, Status, Alcohol, Hepatitis B, BMI, Diphtheria, HIV/AIDS, GDP, Population, Income composition of resources, and Schooling. We will then create multiple linear regression on each  $X_n$  predictor against our life expectancy. In addition we will see if there is any increase in predictive power due to interaction effects such as alcohol consumption and BMI have on our life expectancy. We will also run hypothesis testing with critical value less than 0.05 to test for significance in our results based on the predictor and life expectancy. We can then create a mathematical model with predictive power to output an estimate of an individual's life expectancy given the parameters mentioned above ; Country, Year, Status, Alcohol, Hepatitis B, BMI, Diphtheria, HIV/AIDS, GDP, Population Income composition of resources, and Schooling.

Sources link: <https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>