QMBU450 - Homework 2 - Report

By Selin Öztürk, 60160

To start working on this project, I looked for open-source numeric datasets and their API's. Then, I found a Python package called `quandl` that allowed querying World Bank Data. I needed to gather my own data table from time-series columns that World Bank provided, thus there had to be a consistency between the year scopes of the datasets. While checking which datasets are suitable to come up with a hypothesis, I could find 5 different datasets including life expectancy (in ages), urban population (%), fertility rate (births per woman), child population (0-14 y/o's, %) and GDP per capita (in $) for every country and the overall world that covered the years 1960-2018.

Using my prior knowledge about development of countries, I hypothesized that **life expectancy in Turkey should have been positively correlated with the percentage of urban population and the GDP of that country, whereas it should have been negatively correlated with fertility rate and the percentage of child population.**

I received the following coefficients for the covariates for Turkey:

	`urban_population_percent`	`fertility_rate`	`child_population`	`GDP_in_dollars`
$\beta_{estimated}$	0.778	-0.206	0.561	0.00047
95% c. i.	[0.776, 0.780]	[-0.760, 0.347]	[0.541, 0.580]	[0.00047004, 0.00047009]

According to these results, life expectancy had a positive correlation with urban population percentage and GDP per capita, and a negative correlation with fertility rate, just like I expected. However, the analysis resulted in a positive correlation between an increase in child population and life expectancy, which actually contradicted with what came out with fertility rate. I wanted to check if my dataset was problematic, so I imported `sklearn`'s `LinearRegression` function to perform linear regression on the same dataset and received the following coefficients:

`urban_population_percent`	`fertility_rate`	`child_population`	`GDP_in_dollars`
0.0438	-4.33	-0.219	-0.00038

These coefficients are in line with my hypothesis. The difference between two methods could be a result of the error minimization mechanism that my implementation lacks. The function that I implemented doesn't update the regression line at all: It finds the coefficients only in one step by multiplying the variables. Although we don't really see what is going on inside `sklearn`'s `LinearRegression` function, I believe that it returns the best fitted line that yields the least errors by making use of a mechanism that updates the regression line iteratively. As a result, `sklearn`'s `LinearRegression` function gave the result that I hypotesized.

You can find the estimated life expectancies along with errors in the following page.

year	y	y_estimated	error
2018	77.437	76.32	1.10
2017	77.161	76.67	0.48
2016	76.86	76.59	0.264
2015	76.532	76.39	0.139
2014	76.172	76.66	-0.49
2013	75.784	76.59	-0.81
2012	75.373	75.92	-0.54
2011	74.944	75.44	-0.49
2010	74.507	74.82	-0.313
2009	74.074	73.77	0.30
2008	73.649	74.35	-0.7
2007	73.235	73.52	-0.28
2006	72.83	72.44	0.382
2005	72.424	71.87	0.54
2004	72.004	70.96	1.03
2003	71.559	70.07	1.48
2002	71.078	69.30	1.77
2001	70.56	68.78	1.77
2000	70.005	69.07	0.92
1999	69.417	68.82	0.59
1998	68.807	68.85	-0.0441
1997	68.189	68.06	0.125
1996	67.57	67.87	-0.308
1995	66.963	67.67	-0.70
1994	66.377	67.26	-0.88
1993	65.815	67.58	-1.76
1992	65.275	67.30	-2.0
1991	64.757	67.11	-2.36
1990	64.256	66.82	-2.57
1989	63.763	65.72	-1.96
1988	63.266	64.83	-1.56
1987	62.758	64.01	-1
1986	62.231	63.09	-0.86
1985	61.681	62.03	-0.354
1984	61.111	60.76	0.344
1983	60.52	59.54	0.97
1982	59.915	58.32	1.5
1981	59.297	57.12	2.17
1980	58.667	56.13	2.52
1979	58.023	56.15	1.8
1978	57.37	55.67	1.6
1977	56.709	55.36	1.34
1976	56.046	55.03	1.01
1975	55.387	54.61	0.77
1974	54.741	54.09	0.64
1973	54.109	53.54	0.55
1972	53.492	53.05	0.43
1971	52.887	52.57	0.313
1970	52.286	52.11	0.169
1969	51.678	51.64	0.0306
1968	51.053	51.13	-0.080
1967	50.406	50.60	-0.20
1966	49.733	50.02	-0.295
1965	49.035	49.41	-0.384
1964	48.312	49.01	-0.70
1963	47.573	48.53	-0.96
1962	46.83	47.97	-1.14
1961	46.093	47.40	-1.31
1960	45.369	46.95	-1.5