Data Analysis

Nentidis Dimitrios - 1st Assignment

Contents

1 Introduction

Within the framework of the first assignment, the dependence of life expectancy is studied on 17 different variables, among which are alcohol consumption, education level, or even the body mass index. The data is given for various countries and years. In the present analysis, data for the years 2001 and 2007 are not taken into account, as per the assignment's instructions.

2 Simple Linear Regression

Initially, a simple linear regression is executed for each of the candidate predictive variables, with the exception of the country of origin of the data and the status of the country, whether it is developed or developing. The null hypothesis for all predictive variables is the absence of a statistical relationship between the variable and the response. This hypothesis is examined at a significance level of 0.01.

For the evaluation of the results, the models are summarized in two tables. The first table presents the equation describing the relationship between the response variable Y and each predictor, while the second shows the coefficients of determination (R^2)—that is, the percentage of variability explained by the model—and the p-values. It can be observed that the models for the population of a country as well as for the year from which the data originates are not statistically significant based on the criterion that was defined for the rejection of the null hypothesis. Additionally, the variables "Percentage Expenditure", "Measles", "GDP" have a very low slope coefficient β. Noteworthy, however, is that the various variables are not considered normalized and therefore their contribution cannot be judged easily. In other words, a coefficient close to zero does not necessarily render the variable insignificant. However, taking into account the large sample size and the coefficients of these values in combination with their p-values and their R^2, it is practically evident that they do not participate in a substantial way in the variation of the values despite the statistically significant correlation.

The predictive variables that are considered to explain the variation in the values, and thus have significant correlation with the response, are first and foremost the completed school years (Schooling), the In-

Predictive Variable	Equation
Year	$Y = -211.026 + 0.140 \cdot X$
Adult Mortality	$Y = 77.788 - 0.050 \cdot X$
Alcohol	$Y = 65.345 + 0.887 \cdot X$
Percentage Expenditure	$Y = 67.862 + 0.002 \cdot X$
Hepatitis B	$Y = 63.499 + 0.073 \cdot X$
Measles	$Y = 69.471 - 0.000072 \cdot X$
BMI	$Y = 60.153 + 0.240 \cdot X$
Under-five Deaths	$Y = 69.774 - 0.010 \cdot X$
Polio	$Y = 58.283 + 0.132 \cdot X$
Total Expenditure	$Y = 65.269 + 0.675 \cdot X$
Diphtheria	$Y = 57.577 + 0.140 \cdot X$
HIV/AIDS	$Y = 71.016 - 0.887 \cdot X$
GDP	$Y = 67.400 + 0.000334 \cdot X$
Population	$Y = 69.351 - 2.775 \times 10^{-9} \cdot X$
Thinness 5-9 Years	$Y = 73.546 - 0.860 \cdot X$
Income Composition of Resources	$Y = 47.318 + 34.804 \cdot X$
Schooling	$Y = 41.410 + 2.298 \cdot X$

Table 1: Equations of simple linear regression.

come Composition of Resources, and adult mortality (Adult Mortality), all with $R^2 > 0.5$. The models for deaths from HIV/AIDS at ages 0–4 years and for body mass index (BMI), with determination coefficients of 0.345 and 0.293 respectively, also appear to explain part of the variation, though not equally satisfactorily.

3 Multiple Linear Regression

While previously the model contained only one variable at a time, in this section the modeling of the response as a function of all predictive variables is examined. The null hypothesis here is that all coefficients of the predictive variables are equal to zero. More specifically, the resulting model has a coefficient of determination equal to 0.83, meaning the model explains 83% of the variability observed in the data.

Adult Mortality, Alcohol consumption, Body Mass Index (BMI), HIV/AIDS cases at ages 0–4, the Income Composition of Resources,

Predictive Variable	R^2	p-value magnitude
Year	0.004	10^{-2}
Adult Mortality	0.505	10^{-225}
Alcohol	0.165	10^{-59}
Percentage Expenditure	0.173	10^{-62}
Hepatitis B	0.045	10^{-16}
Measles	0.007	10^{-3}
BMI	0.293	10^{-112}
Under-five Deaths	0.037	10^{-14}
Polio	0.116	10^{-41}
Total Expenditure	0.032	10^{-12}
Diphtheria	0.120	10^{-42}
HIV/AIDS	0.345	10^{-136}
GDP	0.201	10^{-73}
Population	0.000	10^{-1}
Thinness 5-9 Years	0.208	10^{-76}
Income Composition of Resources	0.527	10^{-239}
Schooling	0.536	10^{-245}

Table 2: Evaluation of simple linear regression.

and the completed years of education are considered the variables for which the null hypothesis is rejected. For all, the p-value is less than 0.001, which indicates that such results would be practically impossible if the null hypothesis were true. For the remaining variables, the null hypothesis cannot be rejected with the available data.

Looking at the coefficients, it can be concluded that, for example, one additional year of schooling leads to an increase in life expectancy by 0.9 years. For some variables, such as ICOR, reading the results is not easy due to the nature of the variable.

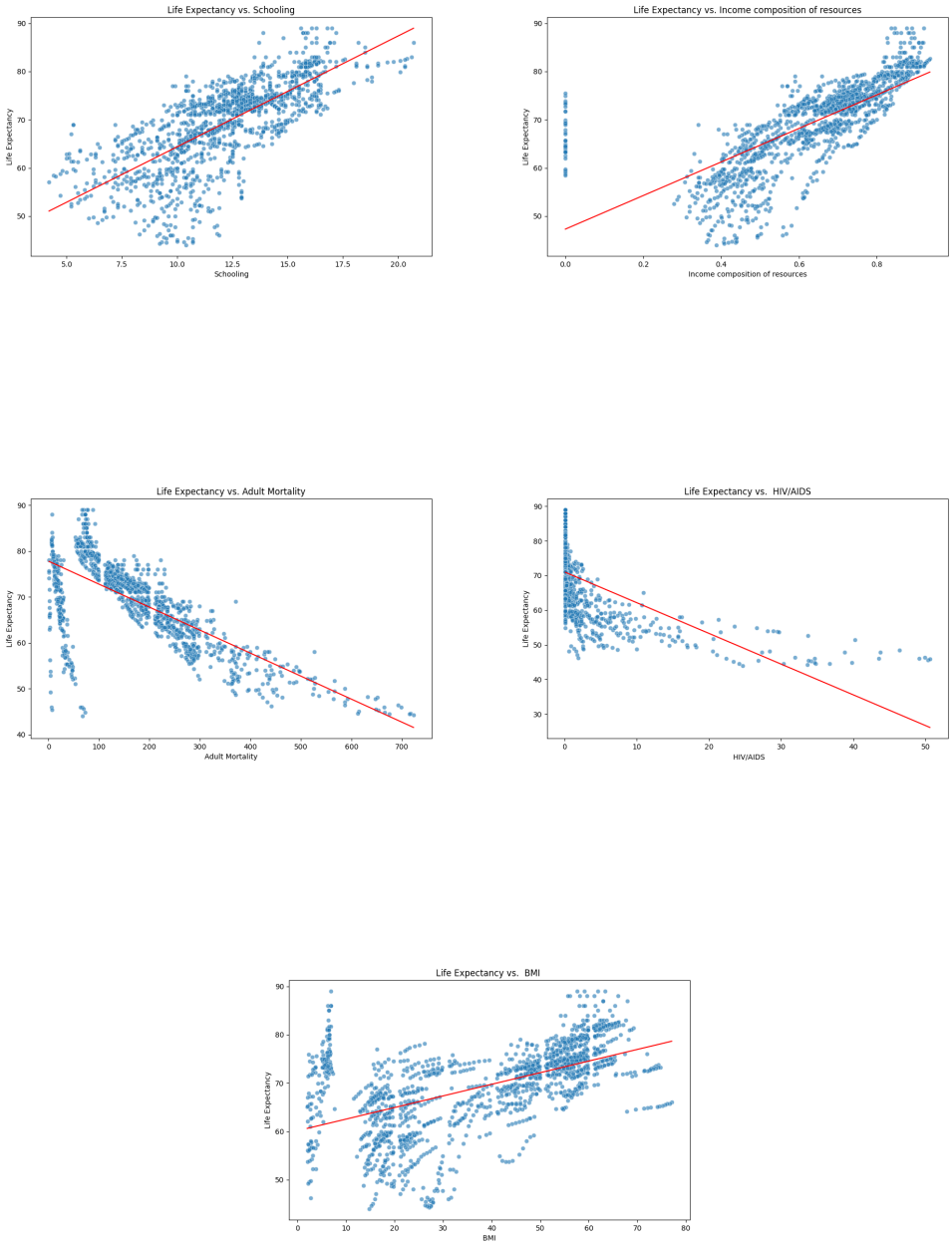
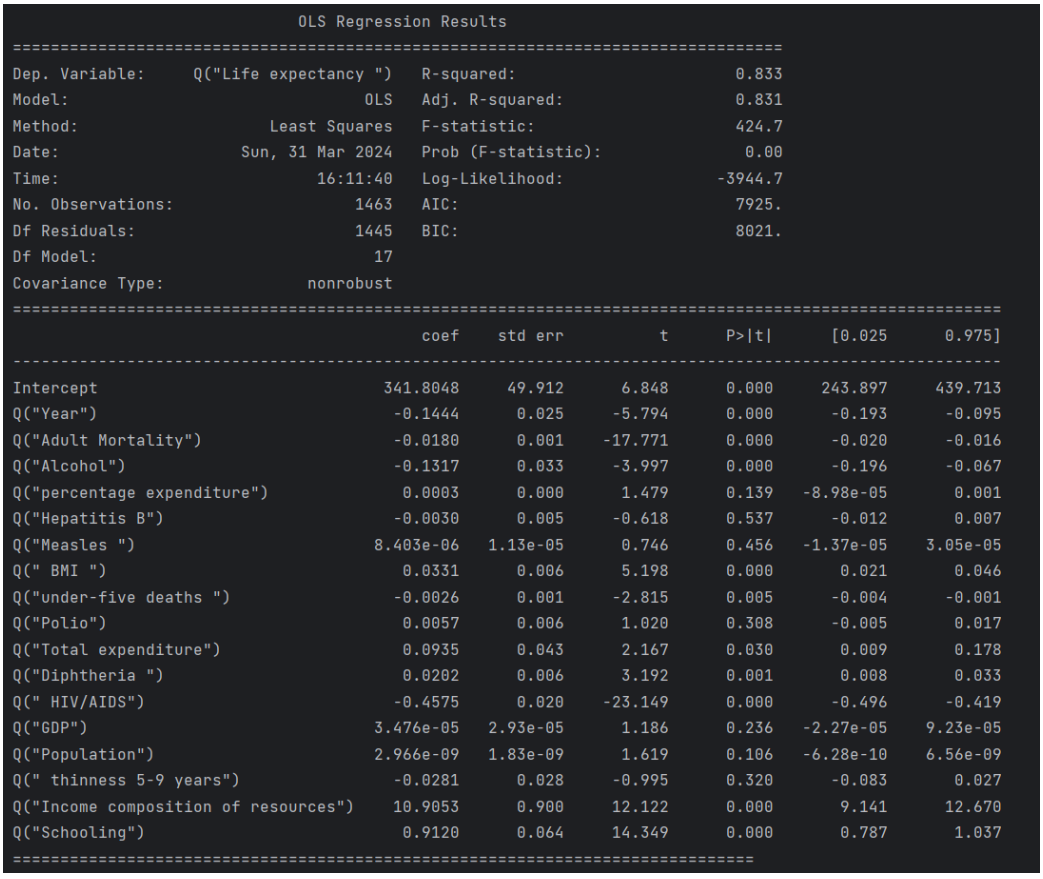


Figure 1: Scatter plots



OLS Regression Results

Dep. Variable:	Q("Life expectancy ")	R-squared:	0.833
Model:	OLS	Adj. R-squared:	0.831
Method:	Least Squares	F-statistic:	424.7
Date:	Sun, 31 Mar 2024	Prob (F-statistic):	0.00
Time:	16:11:40	Log-Likelihood:	-3944.7
No. Observations:	1463	AIC:	7925.
Df Residuals:	1445	BIC:	8021.
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	341.8048	49.912	6.848	0.000	243.897	439.713
Q("Year")	-0.1444	0.025	-5.794	0.000	-0.193	-0.095
Q("Adult Mortality")	-0.0180	0.001	-17.771	0.000	-0.020	-0.016
Q("Alcohol")	-0.1317	0.033	-3.997	0.000	-0.196	-0.067
Q("percentage expenditure")	0.0003	0.000	1.479	0.139	-8.98e-05	0.001
Q("Hepatitis B")	-0.0030	0.005	-0.618	0.537	-0.012	0.007
Q("Measles ")	8.403e-06	1.13e-05	0.746	0.456	-1.37e-05	3.05e-05
Q(" BMI ")	0.0331	0.006	5.198	0.000	0.021	0.046
Q("under-five deaths ")	-0.0026	0.001	-2.815	0.005	-0.004	-0.001
Q("Polio")	0.0057	0.006	1.020	0.308	-0.005	0.017
Q("Total expenditure")	0.0935	0.043	2.167	0.030	0.009	0.178
Q("Diphtheria ")	0.0202	0.006	3.192	0.001	0.008	0.033
Q(" HIV/AIDS")	-0.4575	0.020	-23.149	0.000	-0.496	-0.419
Q("GDP")	3.476e-05	2.93e-05	1.186	0.236	-2.27e-05	9.23e-05
Q("Population")	2.966e-09	1.83e-09	1.619	0.106	-6.28e-10	6.56e-09
Q(" thinness 5-9 years")	-0.0281	0.028	-0.995	0.320	-0.083	0.027
Q("Income composition of resources")	10.9053	0.900	12.122	0.000	9.141	12.670
Q("Schooling")	0.9120	0.064	14.349	0.000	0.787	1.037

Figure 2: Summary of multiple linear regression analysis.

4 Forward Selection

At this point, it is examined whether the selection of only specific predictive variables for the response could maintain the ability of the model to explain the variability of the values, keeping R^2 at high levels. For the selection of the most suitable variables, the method of forward selection is used. Initially, there is an empty model in which the addition of variables that pass a criterion is examined, in this case p-value < 0.001. The process starts from the variable with the highest correlation with the response. Once a variable has been inserted into the model, it cannot be removed afterward. The criterion selected here for adding a variable to the model is R^2. If the addition of a variable significantly improves it, then the variable is deemed appropriate to be added. The results are summarized in the table below, where any variable not shown has a coefficient of zero. In total, the R^2 is equal to 0.835 for this model, with p-value < 0.001 making it statistically significant. As seen, the "predictive power" of the model was maintained with a reduction of the variables from 17 to 9.

Predictive Variable	**Coefficient**
Intercept	321.7432
Schooling	1.0530
Adult Mortality	-0.0192
HIV/AIDS	-0.4258
Income Composition of Resources	9.9186
Percentage Expenditure	0.0005
BMI	0.0341
Year	-0.1347
Diphtheria	0.0225
Alcohol	-0.1269

Table 3: Multiple linear regression coefficients with forward selection.

5 Non-Linear Regression

By observing the residual plots for the requested variables using the linear models developed for the first question, it is seen that for Schooling and BMI there is clearly non-linear behavior, while for Year and Alcohol less so. As mentioned earlier, there is no indication of dependence of Life Expectancy on the year from which the data originates.

Predictive Variable	Equation
Year	$Y = 0.109 + 72.739X - 0.0725X^2 + 1.806 \times 10^{-5}X^3$
Alcohol	$Y = 65.638 + 0.423X + 0.0901X^2 - 0.0044X^3$
BMI	$Y = 73.809 - 1.353X + 0.0437X^2 - 0.0003X^3$
Schooling	$Y = 81.052 - 8.596X + 0.936X^2 - 0.0254X^3$

Table 4: Non-linear regression equations.

Predictive Variable	R^2	p-value magnitude
Year	0.007	0.00347
Alcohol	0.163	10^{-63}
BMI	0.430	10^{-200}
Schooling	0.551	10^{-285}

Table 5: Evaluation of polynomial regression.

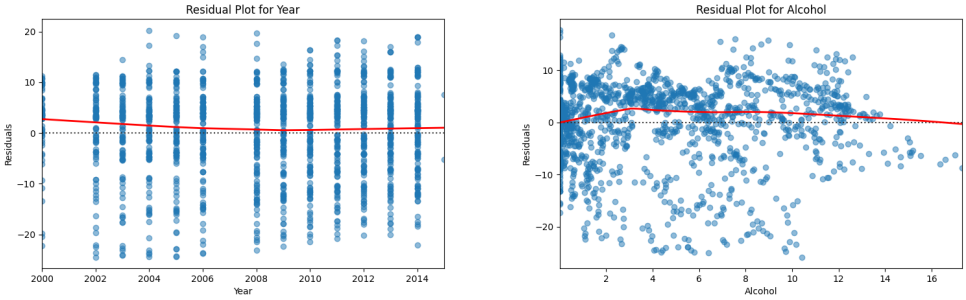


Figure 3: Residual plots.

All models are statistically significant, with p-values clearly smaller than 0.001, with the exception of the predictive variable Year, whose

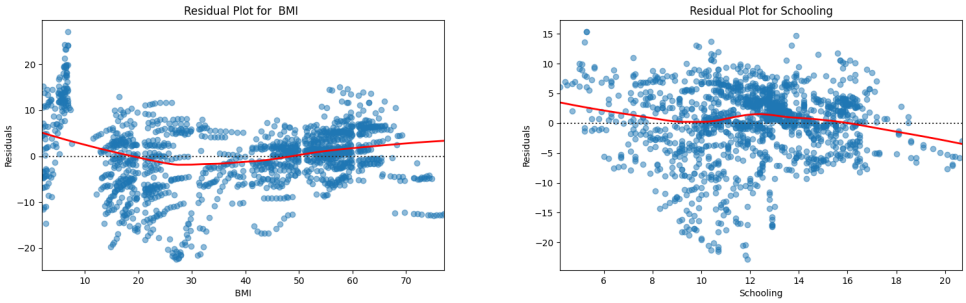


Figure 4: Residual plots continued.

model is only marginally significant statistically. By forming the polynomial models, it can be seen that there is an improvement in the fit. Regarding the predictive variable Year, as with the linear model, there does not appear to be any substantial correlation with the response. Alcohol consumption explains 16.3%, slightly less than in the linear model, while the model of completed school years shows $R^2 = 55.1\%$, slightly increased compared to the linear model. A substantial difference appears in BMI, where the polynomial model is improved by about 50%, from $R^2 = 0.29$ in the linear to 43% here. Judging from the fact that increasing the degree of the model significantly increases its complexity, it is considered appropriate to adopt only the model for BMI. Having said this, however, the polynomial model for Schooling is also better. It is noteworthy that for all polynomial models except for Schooling, the third-degree term is clearly smaller than the second-degree one.

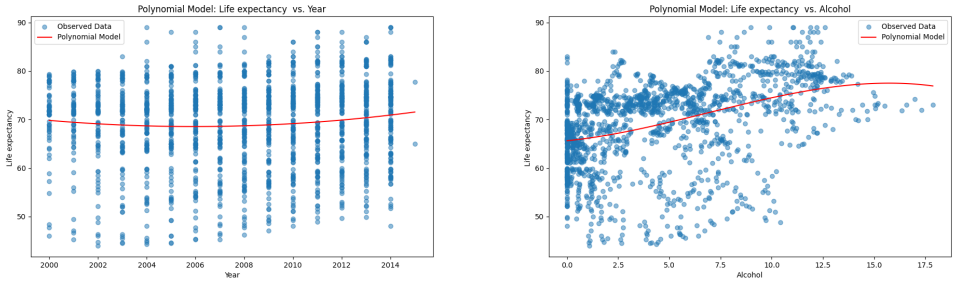


Figure 5: Non-linear models.

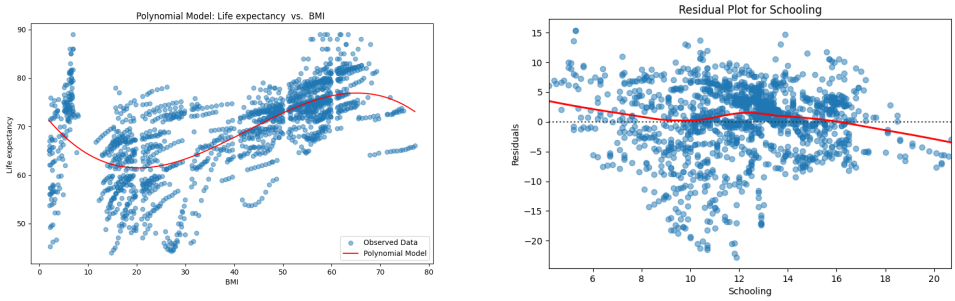


Figure 6: Non-linear models continued.