WHO Life Expectancy

Multilinear Regression Model

Overview

- **Aim:**

To build regression model to predict life expectancy and investigate the effects of multiple factors from demographic variables, income composition and mortality rates to immunization and human development index to give a country which area should be given importance in order to efficiently improve the life expectancy of its population.

- **Business problem:**

WHO wishes to predict life expectancy and determine which factors has significant impact to develop customised action plan to improve life expectancy in countries with low life expectancy.

- **Data:**

The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website.

Our focus

1. Does various predicting factors which has been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting the life expectancy
2. What is the impact of Immunization coverage on life Expectancy?
3. Do densely populated countries tend to have lower life expectancy?
4. What is the impact of schooling on the lifespan of humans?

Approach

- 4 different multilinear regression models were built and evaluated using statistical model library in Python
- Each independent variables/ features relationship with prices were analysed
- 9 significant features affect life expectancy values

OLS Regression Results

Dep. Variable:	Life_expectancy	**R-squared:**	0.820
Model:	OLS	**Adj. R-squared:**	0.819
Method:	Least Squares	**F-statistic:**	698.8
Date:	Tue, 21 Feb 2023	**Prob (F-statistic):**	0.00
Time:	17:29:13	**Log-Likelihood:**	-8267.9
No. Observations:	2938	**AIC:**	1.658e+04
Df Residuals:	2918	**BIC:**	1.670e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	56.7284	0.672	84.444	0.000	55.411	58.046
Adult_Mortality	-0.0199	0.001	-25.151	0.000	-0.021	-0.018
infant_death	0.0997	0.008	11.822	0.000	0.083	0.116
Alcohol	0.0615	0.026	2.376	0.018	0.011	0.112
percentage_expenditure	3.937e-05	9.03e-05	0.436	0.663	-0.000	0.000
Hepatitis_B	-0.0167	0.004	-4.493	0.000	-0.024	-0.009
Measles	-1.934e-05	7.65e-06	-2.527	0.012	-3.43e-05	-4.33e-06
BMI	0.0449	0.005	9.131	0.000	0.035	0.055
under_five_deaths	-0.0747	0.006	-12.083	0.000	-0.087	-0.063
Polio	0.0287	0.004	6.440	0.000	0.020	0.037
Total_expenditure	0.0681	0.034	1.993	0.046	0.001	0.135
Diphtheria	0.0410	0.005	8.834	0.000	0.032	0.050
HIV_AIDS	-0.4698	0.018	-26.766	0.000	-0.504	-0.435
GDP	4.246e-05	1.37e-05	3.089	0.002	1.55e-05	6.94e-05
Population	6.001e-11	1.69e-09	0.036	0.972	-3.25e-09	3.37e-09
thinness_1_19yrs	-0.0833	0.050	-1.655	0.098	-0.182	0.015
thinness_5_9yrs	0.0105	0.050	0.211	0.833	-0.087	0.108
Income_composition_of_resources	5.5131	0.631	8.733	0.000	4.275	6.751
Schooling	0.6583	0.042	15.821	0.000	0.577	0.740
status_Developing	-1.6115	0.270	-5.970	0.000	-2.141	-1.082

Omnibus:	136.306	**Durbin-Watson:**	0.704
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	389.559
Skew:	-0.189	**Prob(JB):**	2.56e-85
Kurtosis:	4.743	**Cond. No.**	5.28e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.28e+08. This might indicate that there are strong multicollinearity or other numerical problems.

Final Model

- Produces predicted value with almost 0 residual errors

  Good predictive ability

- Can explained 80.4% of variance in life expectancy

  High goodness of fit - strong inference ability in explaining variance in property prices

OLS Regression Results

Dep. Variable:	Life_expectancy	**R-squared:**	0.805
Model:	OLS	**Adj. R-squared:**	0.804
Method:	Least Squares	**F-statistic:**	1344.
Date:	Tue, 21 Feb 2023	**Prob (F-statistic):**	0.00
Time:	23:16:53	**Log-Likelihood:**	3916.0
No. Observations:	2938	**AIC:**	-7812.
Df Residuals:	2928	**BIC:**	-7752.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.2419	0.001	3596.968	0.000	4.240	4.244
scaled_adult_mortality	-0.0131	0.001	-10.070	0.000	-0.016	-0.011
scaled_BMI	0.0072	0.001	5.320	0.000	0.005	0.010
scaled_polio	0.0086	0.001	6.230	0.000	0.006	0.011
scaled_diphtheria	0.0102	0.001	7.335	0.000	0.007	0.013
scaled_HIV_AIDS	-0.0782	0.001	-56.204	0.000	-0.081	-0.075
scaled_GDP	0.0215	0.001	15.088	0.000	0.019	0.024
scaled_thinness	-0.0148	0.001	-9.956	0.000	-0.018	-0.012
scaled_schooling	0.0261	0.001	18.773	0.000	0.023	0.029
scaled_status_developing	-0.0125	0.001	-8.824	0.000	-0.015	-0.010

Omnibus:	469.690	**Durbin-Watson:**	0.597
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2213.671
Skew:	-0.690	**Prob(JB):**	0.00
Kurtosis:	7.022	**Cond. No.**	2.70

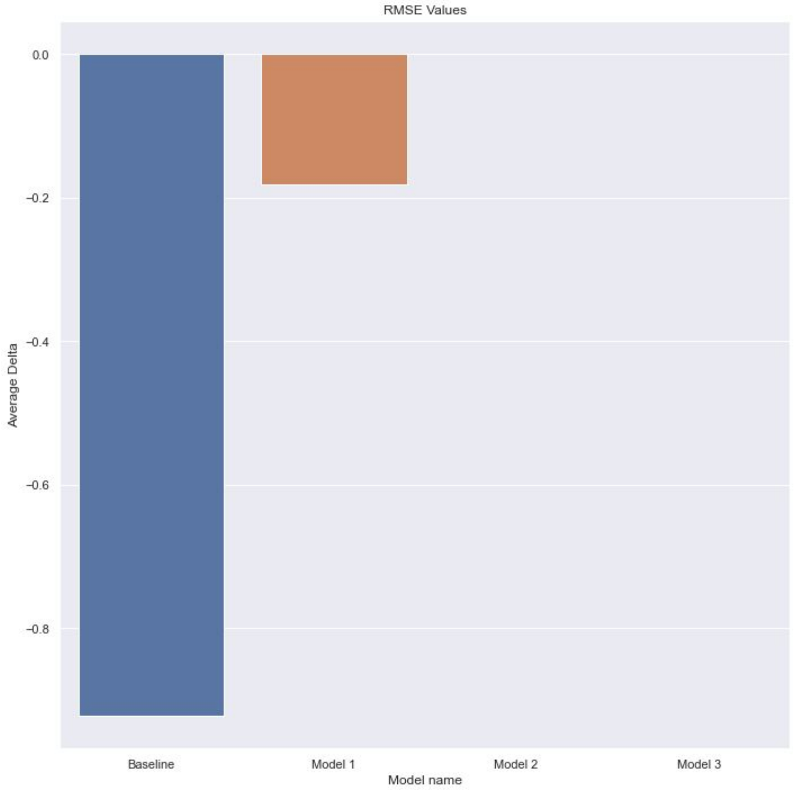
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

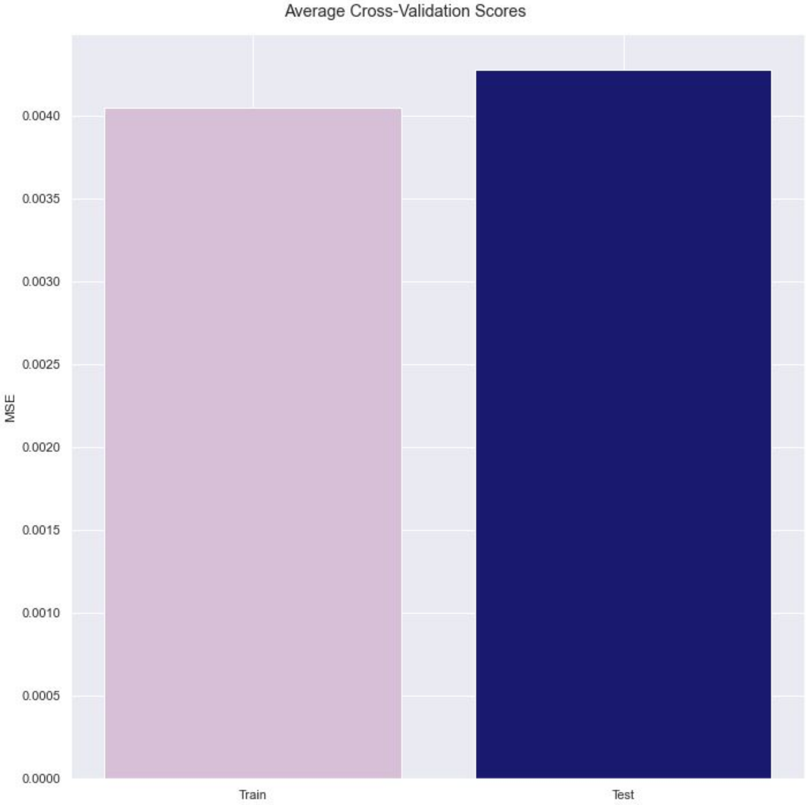
```
X_3 = non_colin_df.drop('Life_expectancy', axis=1)
y_3 = non_colin_df['Life_expectancy']
X3_train, X3_test, y3_train, y3_test = train_test_split(X_3,y_3, random_state=22)
model_3k = LinearRegression()
model_3k.fit(X3_train, y3_train)

y3_hat_train = model_3k.predict(X2_train)
y3_hat_test = model_3k.predict(X2_test)

from sklearn.metrics import mean_squared_error
train_mse_3 = mean_squared_error(y3_train, y3_hat_train)
test_mse_3 = mean_squared_error(y3_test, y3_hat_test)
RSME_3 = test_mse_3 - train_mse_3
print('Train Mean Squared Error:', train_mse_3)
print('Test Mean Squared Error:', test_mse_3)
print('RMSE:', RSME_3)
```

```
Train Mean Squared Error: 0.03742441791173695
Test Mean Squared Error: 0.036934494440102446
RMSE: -0.0004899234716345055
```


RMSE Values
0.0
-0.2
-0.4
-0.6
-0.8
Average Delta
Baseline
Model 1
Model 2
Model 3
Model name


Average Cross-Validation Scores
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
MSE
Train
Test

Conclusions

- HIV/AIDS rate - most significant factor negatively impact life expectancy
- Population size does not play a role in life expectancy
- High Polio & Diphtheria immunisation rate positively affect life expectancy
- Number of years of schooling positively impact life expectancy

Limitations:

- Multilinear regression model perhaps not ideal due to homoskedasticity violation - explore other models for this type of data for more suitable regression model
- Limitation of dataset only includes data 2000-2015 - Further analysis into larger dataset with more up-to-date data.

Thank You!

Email: yen.ho993@gmail.com
GitHub: @NBYH
LinkedIn: linkedin.com/in/yenho93/