Key Indicators of High or Low Life Expectancy in Developing Countries

Author: Adrian Chavez-Loya

Completed: June, 2024

Objective

Highlight **key factors in increasing life expectancy** within developing countries.

Dataset: `Life Expectancy Data.csv` ***

Background

The Global Health Observatory (GHO) data repository under the World Health Organization (WHO) tracks health status and other related factors for all countries. This dataset is related to life expectancy and associated health factors for 193 countries, aggregated from the WHO, and corresponding economic data was collected from the United Nations. This data spans the years of 2000 to 2015 and is information about developing countries.

Here is a comprehensive list of variables that may have connections to Life Expectancy:

1. Adult Mortality: High adult mortality rates are often correlated with lower life expectancy.

2. Infant Mortality: Similarly, high infant mortality rates can indicate poor health conditions and limited access to healthcare, which may lead to lower life expectancy.

3. Alcohol Consumption: Excessive alcohol consumption can have detrimental effects on health and may reduce life expectancy.

4. Percentage Expenditure on Healthcare: Higher healthcare expenditures may lead to better access to healthcare services and contribute to higher life expectancy.

5. Vaccination Coverage (e.g., Hepatitis B, Polio, Diphtheria): Adequate vaccination coverage can prevent infectious diseases and reduce mortality rates, thereby increasing life expectancy.

6. BMI (Body Mass Index): BMI is often used as an indicator of overall health and can influence life expectancy, with both underweight and obesity associated with higher mortality risks.

7. Income Composition of Resources: Higher income levels and a more equitable distribution of resources are generally associated with better access to healthcare,

education, and other social determinants of health, contributing to higher life expectancy.

8. Education Level (Schooling): Higher levels of education are correlated with better health outcomes and behaviors, which can lead to increased life expectancy.
9. HIV/AIDS Prevalence: HIV/AIDS significantly impacts mortality rates and life expectancy, especially in regions with high prevalence rates.
10. Gross Domestic Product (GDP): GDP can serve as a proxy for overall economic development, which in turn affects access to healthcare, nutrition, sanitation, and other factors influencing life expectancy.

Models to be Presented in Order

1a. **Model 1:** The Effect of GDP on Life Expectancy 1b. **Model 1a:** Quadratic Model (Result of Model Assumptions):

1. **Model 2:** Comprehensive Analysis of Multiple Variables to Explore Their Potential Impact on Life Expectancy.
2. **Model 3:** Comprehensive Analysis Multiple Predictors on Life Expectancy **Excluding GDP and Percentage Expenditure

```
# Import necessary packages
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import statsmodels.api as sm
import seaborn as sns
import matplotlib.pyplot as plt #use for plotting model

# Read in data set
life_expectancy_data =
pd.read_csv('/Users/adrianchavezloya/Desktop/USU Summer 2024/Intro to
Regression:Machine Learning/Module 1 and 2/Unit 1 Test/Life Expectancy
Data.csv')
print(life_expectancy_data.columns) ##Used to check headings

Index(['Country', 'Year', 'Status', 'Life Expectancy ', 'Adult
Mortality',
       'infant deaths', 'Alcohol', 'percentage expenditure',
'Hepatitis B',
       'Measles ', 'BMI ', 'under-five deaths ', 'Polio', 'Total
expenditure',
       'Diphtheria ', ' HIV/AIDS', 'GDP', 'Population',
       ' thinness  1-19 years', ' thinness 5-9 years',
```

```
       'Income composition of resources', 'Schooling'],
      dtype='object')
```

```
life_expectancy_data.head()
```

```
       Country  Year      Status  Life Expectancy   Adult Mortality  \
0  Afghanistan  2015  Developing              65.0            263.0
1  Afghanistan  2014  Developing              59.9            271.0
2  Afghanistan  2013  Developing              59.9            268.0
3  Afghanistan  2012  Developing              59.5            272.0
4  Afghanistan  2011  Developing              59.2            275.0

   infant deaths  Alcohol  percentage expenditure  Hepatitis B
Measles   ...  \
0             62     0.01               71.279624         65.0
1154  ...
1             64     0.01               73.523582         62.0
492  ...
2             66     0.01               73.219243         64.0
430  ...
3             69     0.01               78.184215         67.0
2787  ...
4             71     0.01                7.097109         68.0
3013  ...

   Polio  Total expenditure  Diphtheria    HIV/AIDS         GDP
Population  \
0    6.0               8.16         65.0         0.1  584.259210
33736494.0
1   58.0               8.18         62.0         0.1  612.696514
327582.0
2   62.0               8.13         64.0         0.1  631.744976
31731688.0
3   67.0               8.52         67.0         0.1  669.959000
3696958.0
4   68.0               7.87         68.0         0.1   63.537231
2978599.0

    thinness  1-19 years   thinness 5-9 years  \
0                   17.2                 17.3
1                   17.5                 17.5
2                   17.7                 17.7
3                   17.9                 18.0
4                   18.2                 18.2

   Income composition of resources  Schooling
0                            0.479       10.1
1                            0.476       10.0
2                            0.470        9.9
3                            0.463        9.8
```

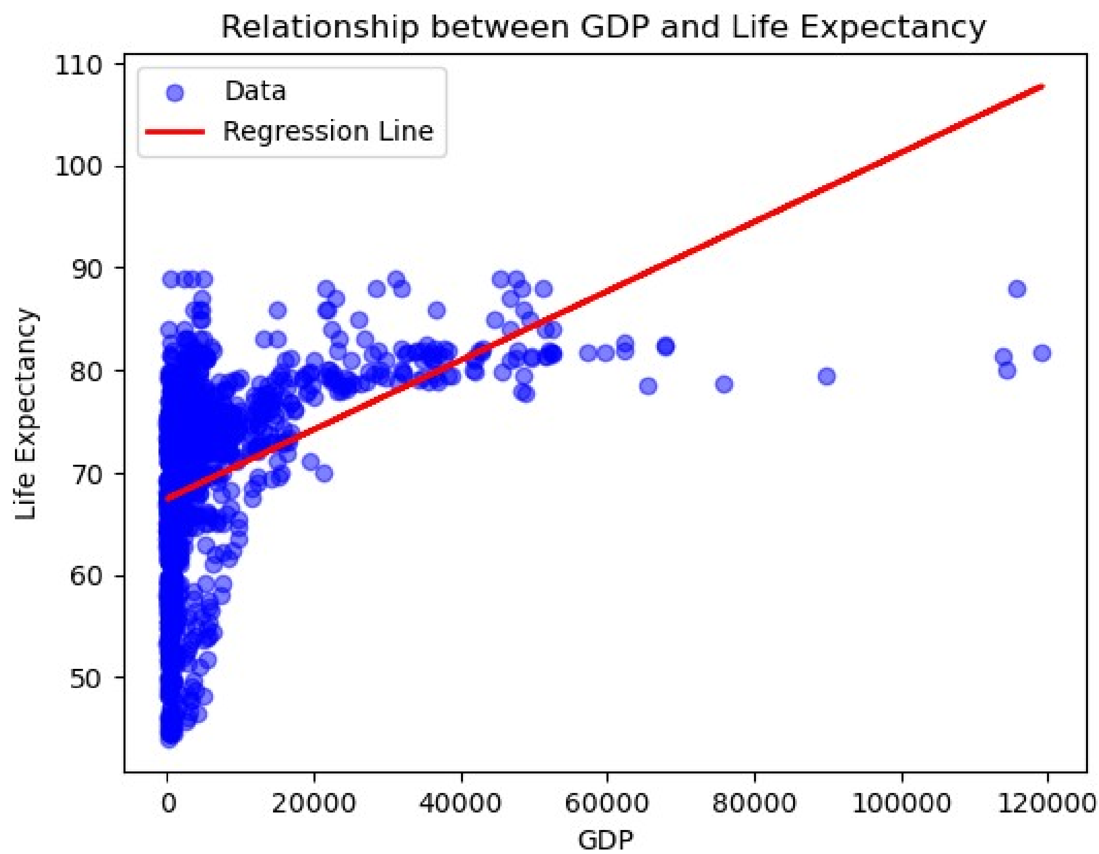
```
4                                   0.454        9.5

[5 rows x 22 columns]

#Drop rows with missing values
life_expectancy_data = life_expectancy_data.dropna()
```

Model 1a: The Effect of GDP on Life Expectancy

Results: To begin, we first examined if there is any correlation betweeen **Body Mass Index (BMI)** and Life Expectancy.

```
# First dataframe
X = life_expectancy_data['GDP']
Y = life_expectancy_data['Life Expectancy ']
X = sm.add_constant(X) # add constant (y-int)
model = sm.OLS(Y, X).fit() # fit model
# Summary of Model 1
print(model.summary())

                            OLS Regression Results

==============================================================================
Dep. Variable:        Life Expectancy    R-squared:
0.195
Model:                            OLS    Adj. R-squared:
0.194
Method:                 Least Squares    F-statistic:
398.4
Date:                Sun, 02 Jun 2024    Prob (F-statistic):
1.50e-79
Time:                        12:49:26    Log-Likelihood:
-5746.3
No. Observations:                1649    AIC:
1.150e+04
Df Residuals:                    1647    BIC:
1.151e+04
Df Model:                           1

Covariance Type:            nonrobust

==============================================================================
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
const         67.4193      0.216    311.942      0.000      66.995
```

```
67.843
GDP             0.0003   1.69e-05     19.959      0.000       0.000
0.000
======================================================================
========
Omnibus:                      150.538   Durbin-Watson:
0.414
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
191.528
Skew:                          -0.804   Prob(JB):
2.57e-42
Kurtosis:                       3.452   Cond. No.
1.42e+04
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 1.42e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.

# Plot of Model 1
plt.scatter(life_expectancy_data['GDP'], life_expectancy_data['Life
Expectancy '], alpha=0.5, color='blue', label='Data')
plt.plot(life_expectancy_data['GDP'], model.predict(X), color='red',
linewidth=2, label='Regression Line') # Plot regression line
plt.xlabel('GDP') # Add labels and title
plt.ylabel('Life Expectancy')
plt.title('Relationship between GDP and Life Expectancy')
plt.legend()
plt.show()
```



Model 1a Results:

- For this model, we have a couple statistical indicators that the correlation between BMI and Life Expectency is statistically significant.

R-squared: Approximately 19.5% of the variance in life expectancy is explained by GDP.

F-statistic: The overall model is statistically significant with a very small p-value (1.50e-79).

Coefficients:

- Intercept (const): Approximately 67.42.
- GDP: For each unit increase in GDP, life expectancy increases by approximately 0.0003 years.

P-values: Both the intercept and GDP coefficients have very small p-values, indicating they are statistically significant predictors of life expectancy.

Model 2b: Change to Quadratic Model (Result of Model Assumptions):

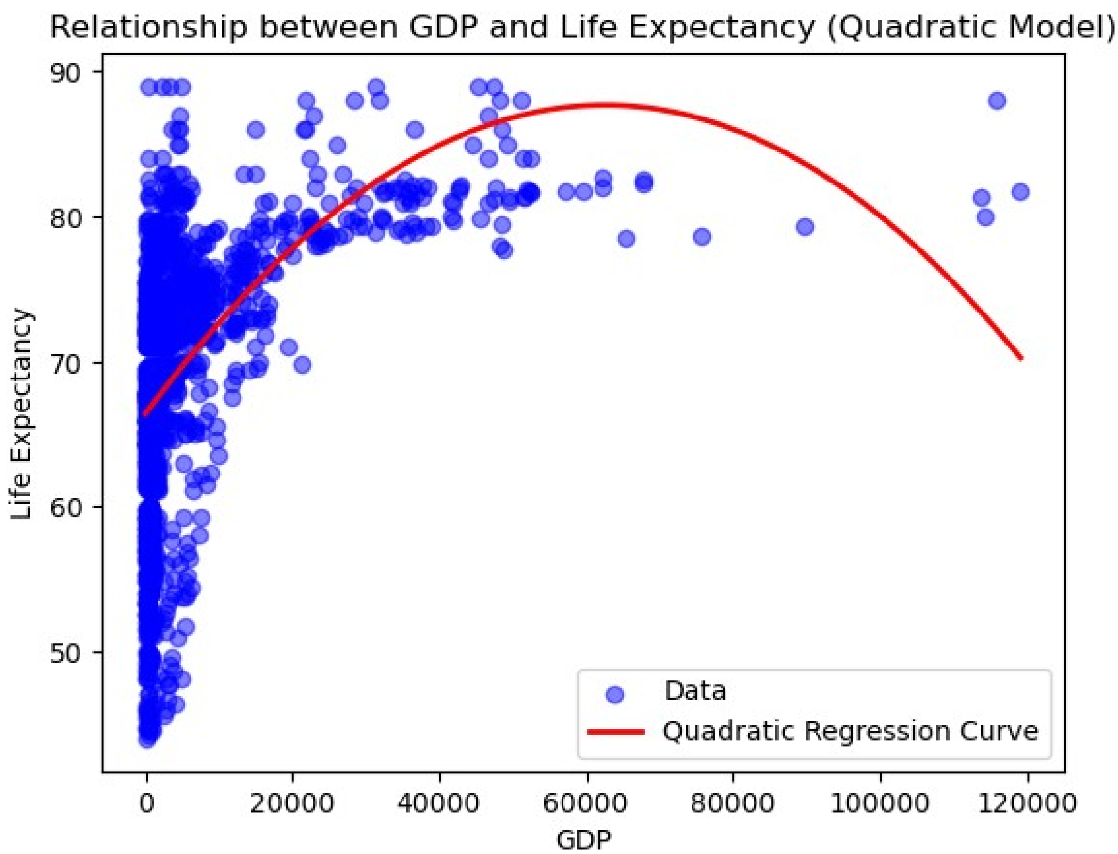
- Our model lacks linearity although there is definitely correlation between the two variables
- Our model seems to display more of a quadratic relationship

Therefore, we will adjust our model and make it a quadratic one.

```
# Create quadratic term
life_expectancy_data['GDP_squared'] = life_expectancy_data['GDP'] ** 2

# Fit a quadratic regression model
X_quad = life_expectancy_data[['GDP', 'GDP_squared']]
X_quad = sm.add_constant(X_quad)
model_quad = sm.OLS(Y, X_quad).fit()

# Scatter plot of the data
plt.scatter(life_expectancy_data['GDP'], life_expectancy_data['Life
Expectancy '], alpha=0.5, color='blue', label='Data')

# Plot the quadratic regression curve
X_plot = np.linspace(life_expectancy_data['GDP'].min(),
life_expectancy_data['GDP'].max(), 100)
X_plot_quad = sm.add_constant(np.column_stack((X_plot, X_plot**2)))
plt.plot(X_plot, model_quad.predict(X_plot_quad), color='red',
linewidth=2, label='Quadratic Regression Curve')

plt.xlabel('GDP')# Add labels and title
plt.ylabel('Life Expectancy')
plt.title('Relationship between GDP and Life Expectancy (Quadratic
Model)')
plt.legend()

# Print summary & plot of the quadratic model
print(model_quad.summary())
plt.show()
```

```
                            OLS Regression Results

======================================================================
========
Dep. Variable:       Life Expectancy   R-squared:
0.258
Model:                           OLS   Adj. R-squared:
0.257
Method:                Least Squares   F-statistic:
286.2
Date:               Sun, 02 Jun 2024   Prob (F-statistic):
2.10e-107
Time:                       12:49:34   Log-Likelihood:
-5678.8
No. Observations:               1649   AIC:
1.136e+04
Df Residuals:                   1646   BIC:
1.138e+04
Df Model:                          2
```

```
Covariance Type:            nonrobust

======================================================================
=========
                 coef    std err          t      P>|t|      [0.025
0.975]
----------------------------------------------------------------------
---------
const         66.3997      0.225    295.553      0.000      65.959
66.840
GDP            0.0007   3.31e-05     20.529      0.000       0.001
0.001
GDP_squared -5.437e-09   4.59e-10    -11.848      0.000   -6.34e-09
-4.54e-09
======================================================================
========
Omnibus:                    111.328   Durbin-Watson:
0.467
Prob(Omnibus):                0.000   Jarque-Bera (JB):
133.280
Skew:                        -0.665   Prob(JB):
1.14e-29
Kurtosis:                     3.414   Cond. No.
1.02e+09
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 1.02e+09. This might indicate that
there are
strong multicollinearity or other numerical problems.
```



Model 1b Results:

- R-squared: The model explains 25.8% of the variance in life expectancy.
- F-statistic: The overall model is statistically significant with a p-value of 2.10e-107.
- Coefficients:
  - Intercept (const): 66.3997 (p-value: <0.0001)
    - The intercept is statistically significant.
- GDP: 0.0007 (p-value: <0.0001)
  - There is a positive relationship between GDP and life expectancy, which is statistically significant.
- GDP_squared: -5.437e-09 (p-value: <0.0001)
  - The negative coefficient indicates a concave relationship, meaning the positive effect of GDP on life expectancy diminishes as GDP increases. This term is also statistically significant.

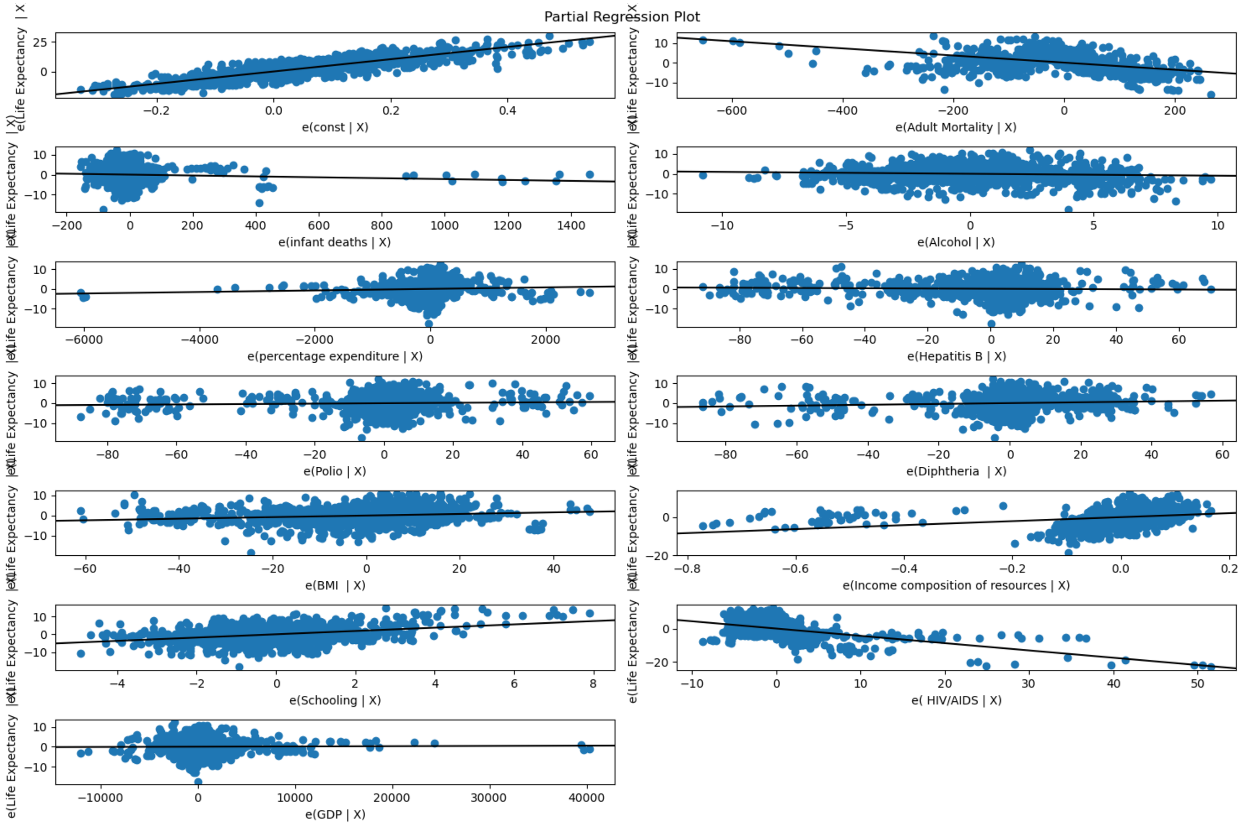
Model 2: Comprehensive Analysis of Multiple Variables to Explore Their Potential Impact on Life Expectancy.

```
X = life_expectancy_data[['Adult Mortality', 'infant deaths',
'Alcohol', 'percentage expenditure',
                          'Hepatitis B', 'Polio', 'Diphtheria ', 'BMI
',
                          'Income composition of resources',
'Schooling', ' HIV/AIDS', 'GDP']]
Y = life_expectancy_data['Life Expectancy ']

X = sm.add_constant(X) #Add constant

# Fit and print the multiple linear regression model and summary
model_multiple = sm.OLS(Y, X).fit()
print(model_multiple.summary())

# Partial regression plots for each variable
fig = plt.figure(figsize=(15, 10))
sm.graphics.plot_partregress_grid(model_multiple, fig=fig)
plt.tight_layout()
plt.show()
```

```
                            OLS Regression Results

======================================================================
========
Dep. Variable:       Life Expectancy    R-squared:
0.825
Model:                           OLS    Adj. R-squared:
0.824
Method:                Least Squares    F-statistic:
644.3
Date:               Sun, 02 Jun 2024    Prob (F-statistic):
0.00
Time:                       12:56:23    Log-Likelihood:
-4486.2
No. Observations:               1649    AIC:
8998.
Df Residuals:                   1636    BIC:
9069.
Df Model:                         12

Covariance Type:           nonrobust

======================================================================
=============================
                                      coef    std err          t
P>|t|      [0.025      0.975]
----------------------------------------------------------------------
```

```
-----------------------------
const                              51.9753      0.667     77.882
0.000      50.666      53.284
Adult Mortality                    -0.0180      0.001    -18.755
0.000      -0.020      -0.016
infant deaths                      -0.0022      0.001     -2.806
0.005      -0.004      -0.001
Alcohol                            -0.0964      0.030     -3.181
0.001      -0.156      -0.037
percentage expenditure              0.0004      0.000      2.089
0.037    2.35e-05       0.001
Hepatitis B                        -0.0063      0.005     -1.381
0.167      -0.015       0.003
Polio                               0.0104      0.005      1.970
0.049    4.73e-05       0.021
Diphtheria                          0.0209      0.006      3.454
0.001       0.009       0.033
BMI                                 0.0395      0.006      6.890
0.000       0.028       0.051
Income composition of resources    10.4204      0.851     12.249
0.000       8.752      12.089
Schooling                           0.9229      0.060     15.286
0.000       0.804       1.041
 HIV/AIDS                          -0.4383      0.018    -24.088
0.000      -0.474      -0.403
GDP                              1.265e-05   2.91e-05      0.435
0.664   -4.44e-05    6.97e-05
======================================================================
========
Omnibus:                       41.284   Durbin-Watson:
0.728
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
69.270
Skew:                          -0.207   Prob(JB):
9.08e-16
Kurtosis:                       3.915   Cond. No.
1.22e+05
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 1.22e+05. This might indicate that
there are
strong multicollinearity or other numerical problems.

/var/folders/vq/l_8lvyx12cxb7kcq563rstwh0000gn/T/
ipykernel_1351/3713450831.py:15: UserWarning: The figure layout has
```

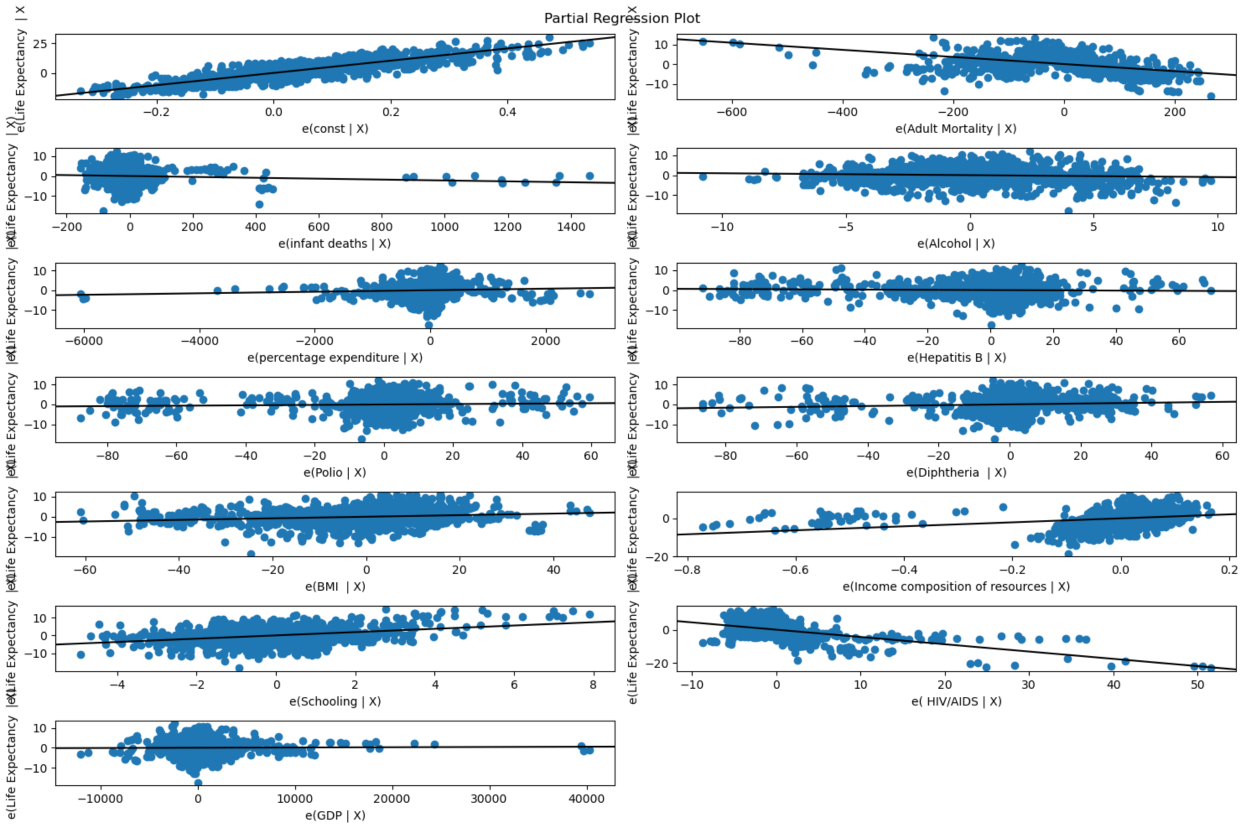
```
changed to tight
  plt.tight_layout()
```




```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Life Expectancy   R-squared:
0.825
Model:                           OLS   Adj. R-squared:
0.824
Method:                Least Squares   F-statistic:
644.3
Date:               Sun, 02 Jun 2024   Prob (F-statistic):
0.00
Time:                       12:55:17   Log-Likelihood:
-4486.2
No. Observations:               1649   AIC:
8998.
Df Residuals:                   1636   BIC:
9069.
Df Model:                         12
```

```
Covariance Type:            nonrobust

======================================================================
==============================
                                      coef    std err          t
P>|t|      [0.025      0.975]
----------------------------------------------------------------------
------------------------------
const                              51.9753      0.667     77.882
0.000      50.666      53.284
Adult Mortality                    -0.0180      0.001    -18.755
0.000      -0.020      -0.016
infant deaths                      -0.0022      0.001     -2.806
0.005      -0.004      -0.001
Alcohol                            -0.0964      0.030     -3.181
0.001      -0.156      -0.037
percentage expenditure              0.0004      0.000      2.089
0.037    2.35e-05       0.001
Hepatitis B                        -0.0063      0.005     -1.381
0.167      -0.015       0.003
Polio                               0.0104      0.005      1.970
0.049    4.73e-05       0.021
Diphtheria                          0.0209      0.006      3.454
0.001       0.009       0.033
BMI                                 0.0395      0.006      6.890
0.000       0.028       0.051
Income composition of resources    10.4204      0.851     12.249
0.000       8.752      12.089
Schooling                           0.9229      0.060     15.286
0.000       0.804       1.041
 HIV/AIDS                          -0.4383      0.018    -24.088
0.000      -0.474      -0.403
GDP                              1.265e-05   2.91e-05      0.435
0.664   -4.44e-05    6.97e-05
======================================================================
========
Omnibus:                       41.284   Durbin-Watson:
0.728
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
69.270
Skew:                          -0.207   Prob(JB):
9.08e-16
Kurtosis:                       3.915   Cond. No.
1.22e+05
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
```

```
correctly specified.
[2] The condition number is large, 1.22e+05. This might indicate that
there are
strong multicollinearity or other numerical problems.

/var/folders/vq/l_8lvyx12cxb7kcq563rstwh0000gn/T/
ipykernel_1351/1776638306.py:18: UserWarning: The figure layout has
changed to tight
  plt.tight_layout()
```



Model 2 Results:

Our multiple regression model which tests predictor variables that were deemed to be the most likely to have an effect on Life Expectancy shows statistical significance.

Correlation Values and P Values:

There are strong correlations between the predictor variables and the dependent variable (Life Expectancy) for every variable.

- Adult Mortality: Statistically significant (p-value < 0.0001) with a negative coefficient, indicating that higher adult mortality rates are associated with lower life expectancy.

- Infant Deaths: Statistically significant (p-value = 0.005) with a negative coefficient, suggesting that higher infant mortality rates are associated with lower life expectancy.
- Alcohol: Statistically significant (p-value = 0.001) with a negative coefficient, indicating that higher alcohol consumption is associated with lower life expectancy.
- Percentage Expenditure: Statistically significant (p-value = 0.037) with a positive coefficient, suggesting that higher healthcare expenditure percentage is associated with higher life expectancy.
- Polio: Statistically significant (p-value = 0.049) with a positive coefficient, implying that higher polio vaccination coverage is associated with higher life expectancy.
- Diphtheria: Statistically significant (p-value = 0.001) with a positive coefficient, indicating that higher diphtheria vaccination coverage is associated with higher life expectancy.
- BMI (Body Mass Index): Statistically significant (p-value < 0.0001) with a positive coefficient, suggesting that a higher BMI is associated with higher life expectancy.
- Income Composition of Resources: Statistically significant (p-value < 0.0001) with a positive coefficient, indicating that higher income composition of resources is associated with higher life expectancy.
- Schooling: Statistically significant (p-value < 0.0001) with a positive coefficient, implying that higher levels of schooling are associated with higher life expectancy.
- HIV/AIDS: Statistically significant (p-value < 0.0001) with a negative coefficient, suggesting that higher HIV/AIDS prevalence is associated with lower life expectancy.

Unsignificant Variables (with a higher p-value):

- Hepatitis B: Not statistically significant (p-value = 0.167). The coefficient may not accurately estimate the effect of Hepatitis B vaccination coverage on life expectancy due to insufficient evidence.
- GDP: Not statistically significant (p-value = 0.664). The coefficient suggests that there is no significant linear relationship between GDP and life expectancy in this model.

Model Performance:

- R-squared: The model explains 82.5% of the variance in life expectancy, indicating a strong overall fit.
- Adjusted R-squared: After adjusting for the number of predictors, the model still explains 82.4% of the variance, suggesting that the model's explanatory power remains high.
- F-statistic: The F-statistic is 644.3 with a p-value of 0.00, indicating that the overall model is statistically significant.

Drawback:

- Condition Number: The **high condition number (1.22e+05) indicates potential multicollinearity issues, suggesting that the model's predictive power might be compromised**. To address this, we will use the Variance Inflation Factor (VIF) to identify which independent variables are contributing to multicollinearity.

**Using VIF (Variance Inflation Factor):

** Used to see which of our predictor variables may be causing multicollinearity (having an influence on other predictor variables)

```
## Using VIF, we will try to rule out variables that may be causing
multi collinearity.
from statsmodels.stats.outliers_influence import
variance_inflation_factor
independent_vars = [
    'Adult Mortality', 'infant deaths', 'Alcohol', 'percentage
expenditure',
    'Hepatitis B', 'Polio', 'Diphtheria ', 'BMI ', 'Income composition
of resources',
    'Schooling', ' HIV/AIDS', 'GDP' ]
# Calculate VIF
X = life_expectancy_data[independent_vars]
X = sm.add_constant(X)  # Add a constant term for the intercept
vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in
range(X.shape[1])]

print(vif_data)
```

	Variable	VIF
0	const	53.945068
1	Adult Mortality	1.759946
2	infant deaths	1.128156
3	Alcohol	1.803813
4	percentage expenditure	12.794993
5	Hepatitis B	1.641961
6	Polio	1.699198
7	Diphtheria	2.060919
8	BMI	1.553816
9	Income composition of resources	2.936492
10	Schooling	3.447844
11	HIV/AIDS	1.458779
12	GDP	13.490625

The predictor variables GDP and percentage expenditure with high VIF may be responsible for inflating our condition number. We will attempt to remove these variables to create a even more reliable model that will show which variables most affect Life Expectancy.

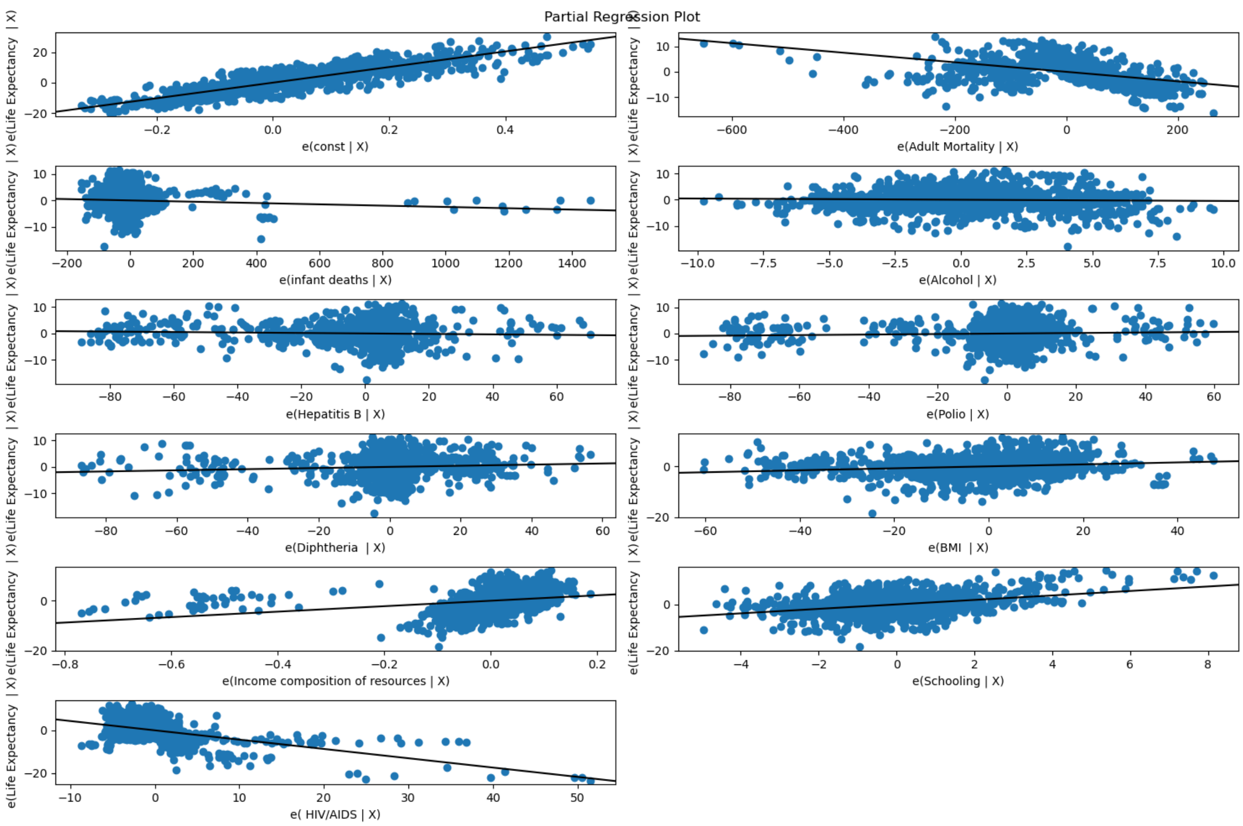
Model 3: Comprehensive Analysis Multiple Predictors on Life Expectancy **Excluding GDP and Percentage Expenditure

```
# Define the predictor and response variables, excluding GDP and
percentage expenditure
X = life_expectancy_data[['Adult Mortality', 'infant deaths',
'Alcohol',
                          'Hepatitis B', 'Polio', 'Diphtheria ', 'BMI
',
                          'Income composition of resources',
'Schooling',
                          ' HIV/AIDS']]
Y = life_expectancy_data['Life Expectancy ']
X = sm.add_constant(X) #Add constant

# Fit and print the multiple linear regression model and summary
model_multiple = sm.OLS(Y, X).fit()
print(model_multiple.summary())

# Partial regression plots for each variable
fig = plt.figure(figsize=(15, 10))
sm.graphics.plot_partregress_grid(model_multiple, fig=fig)
plt.tight_layout()
plt.show()
```

```
                            OLS Regression Results

======================================================================
========
Dep. Variable:       Life Expectancy   R-squared:
0.819
Model:                           OLS   Adj. R-squared:
0.818
Method:                Least Squares   F-statistic:
740.3
Date:               Sun, 02 Jun 2024   Prob (F-statistic):
0.00
Time:                       13:34:00   Log-Likelihood:
-4516.4
No. Observations:               1649   AIC:
9055.
Df Residuals:                   1638   BIC:
9114.
Df Model:                         10

Covariance Type:           nonrobust

======================================================================
```

```
=============================
                                           coef    std err          t
P>|t|      [0.025      0.975]
----------------------------------------------------------------------
-----------------------------
const                                   51.4833      0.672     76.626
0.000      50.165      52.801
Adult Mortality                         -0.0187      0.001    -19.144
0.000      -0.021      -0.017
infant deaths                           -0.0025      0.001     -3.037
0.002      -0.004      -0.001
Alcohol                                 -0.0461      0.030     -1.529
0.127      -0.105       0.013
Hepatitis B                             -0.0089      0.005     -1.926
0.054      -0.018       0.000
Polio                                    0.0099      0.005      1.843
0.066      -0.001       0.020
Diphtheria                               0.0215      0.006      3.498
0.000       0.009       0.034
BMI                                      0.0386      0.006      6.612
0.000       0.027       0.050
Income composition of resources    10.9784      0.861     12.755
0.000       9.290      12.667
Schooling                                0.9717      0.061     15.950
0.000       0.852       1.091
 HIV/AIDS                               -0.4361      0.019    -23.544
0.000      -0.472      -0.400
======================================================================
========
Omnibus:                       39.834   Durbin-Watson:
0.696
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
70.007
Skew:                          -0.181   Prob(JB):
6.28e-16
Kurtosis:                       3.942   Cond. No.
2.32e+03
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 2.32e+03. This might indicate that
there are
strong multicollinearity or other numerical problems.

/var/folders/vq/l_8lvyx12cxb7kcq563rstwh0000gn/T/
ipykernel_1351/360376070.py:16: UserWarning: The figure layout has
```

```
changed to tight
  plt.tight_layout()
```



Model 3: Key Observations

- The R-squared and adjusted R-squared values are slightly lower in the new model, but still very high, indicating a strong fit.
- The condition number has significantly decreased from 1.22e+05 to 2.32e+03, indicating a substantial reduction in multicollinearity issues.

Conclusion:

After analyzing these three models, Model 3 is the best model which contains predictor variables that are statistically significant in correlation with the outcome (Life Expectancy). Model 3 also shows a significantly lower test condition number:

Before Excluding GDP and Percentage Expenditure Condition Number: 1.22e+05 (122,000) After Excluding GDP and Percentage Expenditure

Condition Number: 2.32e+03 (2,320) R-squared: 0.819 Adj. R-squared: 0.818

The R-squared is still extremely high which explains the variance in the these predictor variables explain the variance in Life Expectancy at about 82%.

Adj. R-Squared is almost the same value, which indicates that we are not overfitting (adding too many predictor variables which may increase R-squared although some predictors are not significant).

GDP has a correlation with life expectancy, although it seems to also be affect other variables and/or be affected itself by other predictor variables (multi-collinearity).

As a result, here is the list of predictor variables that are most likely to effect life expectancy in developing countries:

- Adult Mortality
- Infant Deaths
- Alcohol
- Hepatitis B
- Polio
- Diphtheria
- BMI
- Income Composition of Resources
- Schooling
- HIV/AIDS