De_Guzman_Hands_on_Activity_11_1_Linear_Regression_Analysis_Wra

April 24, 2024

1 Hands-on Activity 11.1 Linear Regression Analysis

1.1 Objective(s):

- This activity aims to demonstrate how to apply simple linear regression analysis to solve regression problem

1.2 Intended Learning Outcomes (ILOs):

- Demonstrate how to solve regression problems using simple linear regression
- Use the linear regression model to predict the target value

1.3 Resources:

- Jupyter Notebook

1.4 Files:

- Life Expectancy Data.csv

1.5 Submission Requirements:

- PDF containing initial EDA and Data Wrangling
- PDF showing demonstration of simple linear regression.
- Submit a link to the colab file through the comment section.

1.6 Procedure:

1.6.1 Setup

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

[10]:
```
life_df = pd.read_csv("/content/Life Expectancy Data.csv")
life_df
```

[10]:

	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	
…	…	…	…	…	…	
2933	Zimbabwe	2004	Developing	44.3	723.0	
2934	Zimbabwe	2003	Developing	44.5	715.0	
2935	Zimbabwe	2002	Developing	44.8	73.0	
2936	Zimbabwe	2001	Developing	45.3	686.0	
2937	Zimbabwe	2000	Developing	46.0	665.0	

	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	
…	…	…	…	…	…	
2933	27	4.36	0.000000	68.0	31	
2934	26	4.06	0.000000	7.0	998	
2935	25	4.43	0.000000	73.0	304	
2936	25	1.72	0.000000	76.0	529	
2937	24	1.68	0.000000	79.0	1483	

	…	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	\
0	…	6.0	8.16	65.0	0.1	584.259210	
1	…	58.0	8.18	62.0	0.1	612.696514	
2	…	62.0	8.13	64.0	0.1	631.744976	
3	…	67.0	8.52	67.0	0.1	669.959000	
4	…	68.0	7.87	68.0	0.1	63.537231	
…	…	…	…	…	…	…	
2933	…	67.0	7.13	65.0	33.6	454.366654	
2934	…	7.0	6.52	68.0	36.7	453.351155	
2935	…	73.0	6.53	71.0	39.8	57.348340	
2936	…	76.0	6.16	75.0	42.1	548.587312	
2937	…	78.0	7.10	78.0	43.5	547.358878	

	Population	thinness	1-19 years	thinness 5-9 years	\
0	33736494.0		17.2	17.3	
1	327582.0		17.5	17.5	
2	31731688.0		17.7	17.7	
3	3696958.0		17.9	18.0	
4	2978599.0		18.2	18.2	
…	…	…		…	
2933	12777511.0		9.4	9.4	

```
2934  12633897.0                   9.8                  9.9
2935    125525.0                   1.2                  1.3
2936  12366165.0                   1.6                  1.7
2937  12222251.0                  11.0                 11.2

      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
…                               …          …
2933                            0.407        9.2
2934                            0.418        9.5
2935                            0.427       10.0
2936                            0.427        9.8
2937                            0.434        9.8

[2938 rows x 22 columns]
```

1.6.2 Data Wrangling and Cleaning

[11]:
```
# renaming and rearranging columns
life_df.columns
```

[11]:
```
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')
```

[13]:
```
life_df.columns =␣
 ↪['country','year','status','life_expectancy','adult_mortality','infant_deaths','alcohol','%
                   'bmi','deaths_under_5', 'polio',␣
 ↪'total_expenditure','diphtheria','hiv/
 ↪aids','gdp','population','thinness(minors)',
                  ␣
 ↪'thinness(children)','resource_income_composition','schooling']

life_df.columns
```

[13]:
```
Index(['country', 'year', 'status', 'life_expectancy', 'adult_mortality',
       'infant_deaths', 'alcohol', '%_expenditure', 'hepatitis_b', 'measles',
       'bmi', 'deaths_under_5', 'polio', 'total_expenditure', 'diphtheria',
```

```
       'hiv/aids', 'gdp', 'population', 'thinness(minors)',
       'thinness(children)', 'resource_income_composition', 'schooling'],
      dtype='object')
```

[26]:
```
# Selecting rows and columns with NaN values
nan_values = life_df[life_df.columns[life_df.isna().any()]][life_df.isna().
 ↪any(axis=1)]
nan_values
```

[26]:

	life_expectancy	adult_mortality	alcohol	hepatitis_b	bmi	polio	\
32	75.6	19.0	NaN	95.0	59.5	95.0	
44	71.7	146.0	0.34	NaN	47.0	87.0	
45	71.6	145.0	0.36	NaN	46.1	86.0	
46	71.4	145.0	0.23	NaN	45.3	89.0	
47	71.3	145.0	0.25	NaN	44.4	86.0	
…	…	…	…	…	…	…	
2918	46.4	64.0	2.33	NaN	17.6	85.0	
2919	45.5	69.0	2.44	NaN	17.3	85.0	
2920	44.6	611.0	2.61	NaN	17.1	86.0	
2921	43.8	614.0	2.62	NaN	16.8	85.0	
2922	67.0	336.0	NaN	87.0	31.8	88.0	

	total_expenditure	diphtheria	gdp	population	\
32	NaN	95.0	4132.762920	39871528.0	
44	3.60	87.0	294.335560	3243514.0	
45	3.73	86.0	1774.336730	3199546.0	
46	3.84	89.0	1732.857979	31592153.0	
47	3.49	86.0	1757.177970	3118366.0	
…	…	…	…	…	
2918	8.18	83.0	429.158343	11421984.0	
2919	6.93	84.0	377.135244	111249.0	
2920	6.56	85.0	378.273624	1824125.0	
2921	7.16	85.0	341.955625	1531221.0	
2922	NaN	87.0	118.693830	15777451.0	

	thinness(minors)	thinness(children)	resource_income_composition	\
32	6.0	5.8	0.743	
44	6.3	6.1	0.663	
45	6.3	6.2	0.653	
46	6.4	6.3	0.644	
47	6.5	6.4	0.636	
…	…	…	…	
2918	7.3	7.2	0.443	
2919	7.4	7.3	0.433	
2920	7.4	7.4	0.424	
2921	7.5	7.5	0.418	
2922	5.6	5.5	0.507	

```
      schooling
32         14.4
44         11.5
45         11.1
46         10.9
47         10.7
…           …
2918       10.2
2919       10.0
2920        9.8
2921        9.6
2922       10.3

[1289 rows x 14 columns]
```

[33]:
```
nan_cols = list(nan_values.columns)
countries = list(life_df['country'].unique())

for i in nan_cols:
  for j in countries:
    mean = life_df[i].loc[life_df['country'] == j].mean()
    life_df[i].fillna(mean,inplace=True)

life_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   country                      2938 non-null   object
 1   year                         2938 non-null   int64
 2   status                       2938 non-null   object
 3   life_expectancy              2938 non-null   float64
 4   adult_mortality              2938 non-null   float64
 5   infant_deaths                2938 non-null   int64
 6   alcohol                      2938 non-null   float64
 7   %_expenditure                2938 non-null   float64
 8   hepatitis_b                  2938 non-null   float64
 9   measles                      2938 non-null   int64
 10  bmi                          2938 non-null   float64
 11  deaths_under_5               2938 non-null   int64
 12  polio                        2938 non-null   float64
 13  total_expenditure            2938 non-null   float64
 14  diphtheria                   2938 non-null   float64
 15  hiv/aids                     2938 non-null   float64
```

```
 16  gdp                              2938 non-null   float64
 17  population                       2938 non-null   float64
 18  thinness(minors)                 2938 non-null   float64
 19  thinness(children)               2938 non-null   float64
 20  resource_income_composition      2938 non-null   float64
 21  schooling                        2938 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

[39]:
```
# Selecting only numerical columns for correlation
life_df_nums = life_df.select_dtypes(exclude='object')
life_df_nums
```

[39]:

	year	life_expectancy	adult_mortality	infant_deaths	alcohol	\
0	2015	65.0	263.0	62	0.01	
1	2014	59.9	271.0	64	0.01	
2	2013	59.9	268.0	66	0.01	
3	2012	59.5	272.0	69	0.01	
4	2011	59.2	275.0	71	0.01	
…	…	…	…	…	…	
2933	2004	44.3	723.0	27	4.36	
2934	2003	44.5	715.0	26	4.06	
2935	2002	44.8	73.0	25	4.43	
2936	2001	45.3	686.0	25	1.72	
2937	2000	46.0	665.0	24	1.68	

	%_expenditure	hepatitis_b	measles	bmi	deaths_under_5	polio	\
0	71.279624	65.0	1154	19.1	83	6.0	
1	73.523582	62.0	492	18.6	86	58.0	
2	73.219243	64.0	430	18.1	89	62.0	
3	78.184215	67.0	2787	17.6	93	67.0	
4	7.097109	68.0	3013	17.2	97	68.0	
…	…	…	…	…	…	…	
2933	0.000000	68.0	31	27.1	42	67.0	
2934	0.000000	7.0	998	26.7	41	7.0	
2935	0.000000	73.0	304	26.3	40	73.0	
2936	0.000000	76.0	529	25.9	39	76.0	
2937	0.000000	79.0	1483	25.5	39	78.0	

	total_expenditure	diphtheria	hiv/aids	gdp	population	\
0	8.16	65.0	0.1	584.259210	33736494.0	
1	8.18	62.0	0.1	612.696514	327582.0	
2	8.13	64.0	0.1	631.744976	31731688.0	
3	8.52	67.0	0.1	669.959000	3696958.0	
4	7.87	68.0	0.1	63.537231	2978599.0	
…	…	…	…	…	…	
2933	7.13	65.0	33.6	454.366654	12777511.0	

```
2934                6.52        68.0      36.7  453.351155  12633897.0
2935                6.53        71.0      39.8   57.348340    125525.0
2936                6.16        75.0      42.1  548.587312  12366165.0
2937                7.10        78.0      43.5  547.358878  12222251.0

      thinness(minors)  thinness(children)  resource_income_composition  \
0                 17.2                17.3                        0.479
1                 17.5                17.5                        0.476
2                 17.7                17.7                        0.470
3                 17.9                18.0                        0.463
4                 18.2                18.2                        0.454
…                    …                   …                            …
2933               9.4                 9.4                        0.407
2934               9.8                 9.9                        0.418
2935               1.2                 1.3                        0.427
2936               1.6                 1.7                        0.427
2937              11.0                11.2                        0.434

      schooling
0          10.1
1          10.0
2           9.9
3           9.8
4           9.5
…           …
2933        9.2
2934        9.5
2935       10.0
2936        9.8
2937        9.8

[2938 rows x 20 columns]
```

1.6.3 Exploratory Data Analysis

[42]:
```
# Correlation
life_df_nums.corr()
```

[42]:
```
                              year  life_expectancy  adult_mortality  \
year                      1.000000         0.164554        -0.075371
life_expectancy           0.164554         1.000000        -0.697240
adult_mortality          -0.075371        -0.697240         1.000000
infant_deaths            -0.037415        -0.195075         0.077919
alcohol                  -0.155781         0.372304        -0.187863
%_expenditure             0.031400         0.382124        -0.243394
hepatitis_b               0.182478         0.238710        -0.151727
measles                  -0.082493        -0.156384         0.030534
```

bmi	0.106656	0.564583	-0.388045
deaths_under_5	-0.042937	-0.220964	0.093274
polio	0.100888	0.466427	-0.278584
total_expenditure	0.168512	0.197784	-0.099697
diphtheria	0.139863	0.479706	-0.278265
hiv/aids	-0.139741	-0.553916	0.522176
gdp	0.090686	0.429977	-0.282525
population	0.014734	-0.021558	-0.010972
thinness(minors)	-0.043795	-0.485543	0.316336
thinness(children)	-0.047279	-0.480154	0.320855
resource_income_composition	0.227183	0.686130	-0.442175
schooling	0.195565	0.706699	-0.437362

	infant_deaths	alcohol	%_expenditure	\
year	-0.037415	-0.155781	0.031400	
life_expectancy	-0.195075	0.372304	0.382124	
adult_mortality	0.077919	-0.187863	-0.243394	
infant_deaths	1.000000	-0.105479	-0.085612	
alcohol	-0.105479	1.000000	0.353509	
%_expenditure	-0.085612	0.353509	1.000000	
hepatitis_b	-0.183863	0.038084	-0.029464	
measles	0.501128	-0.042626	-0.056596	
bmi	-0.227005	0.299847	0.231397	
deaths_under_5	0.996629	-0.102058	-0.087852	
polio	-0.168433	0.213648	0.149324	
total_expenditure	-0.125881	0.204966	0.141271	
diphtheria	-0.173320	0.209819	0.145546	
hiv/aids	0.025231	-0.034509	-0.097857	
gdp	-0.101404	0.302695	0.901868	
population	0.549684	-0.026012	-0.022760	
thinness(minors)	0.450986	-0.406835	-0.251818	
thinness(children)	0.459543	-0.397093	-0.254134	
resource_income_composition	-0.147503	0.364723	0.390984	
schooling	-0.194417	0.440712	0.398214	

	hepatitis_b	measles	bmi	deaths_under_5	\
year	0.182478	-0.082493	0.106656	-0.042937	
life_expectancy	0.238710	-0.156384	0.564583	-0.220964	
adult_mortality	-0.151727	0.030534	-0.388045	0.093274	
infant_deaths	-0.183863	0.501128	-0.227005	0.996629	
alcohol	0.038084	-0.042626	0.299847	-0.102058	
%_expenditure	-0.029464	-0.056596	0.231397	-0.087852	
hepatitis_b	1.000000	-0.109646	0.163716	-0.192762	
measles	-0.109646	1.000000	-0.173695	0.507809	
bmi	0.163716	-0.173695	1.000000	-0.237775	
deaths_under_5	-0.192762	0.507809	-0.237775	1.000000	
polio	0.449330	-0.133235	0.289049	-0.186499	

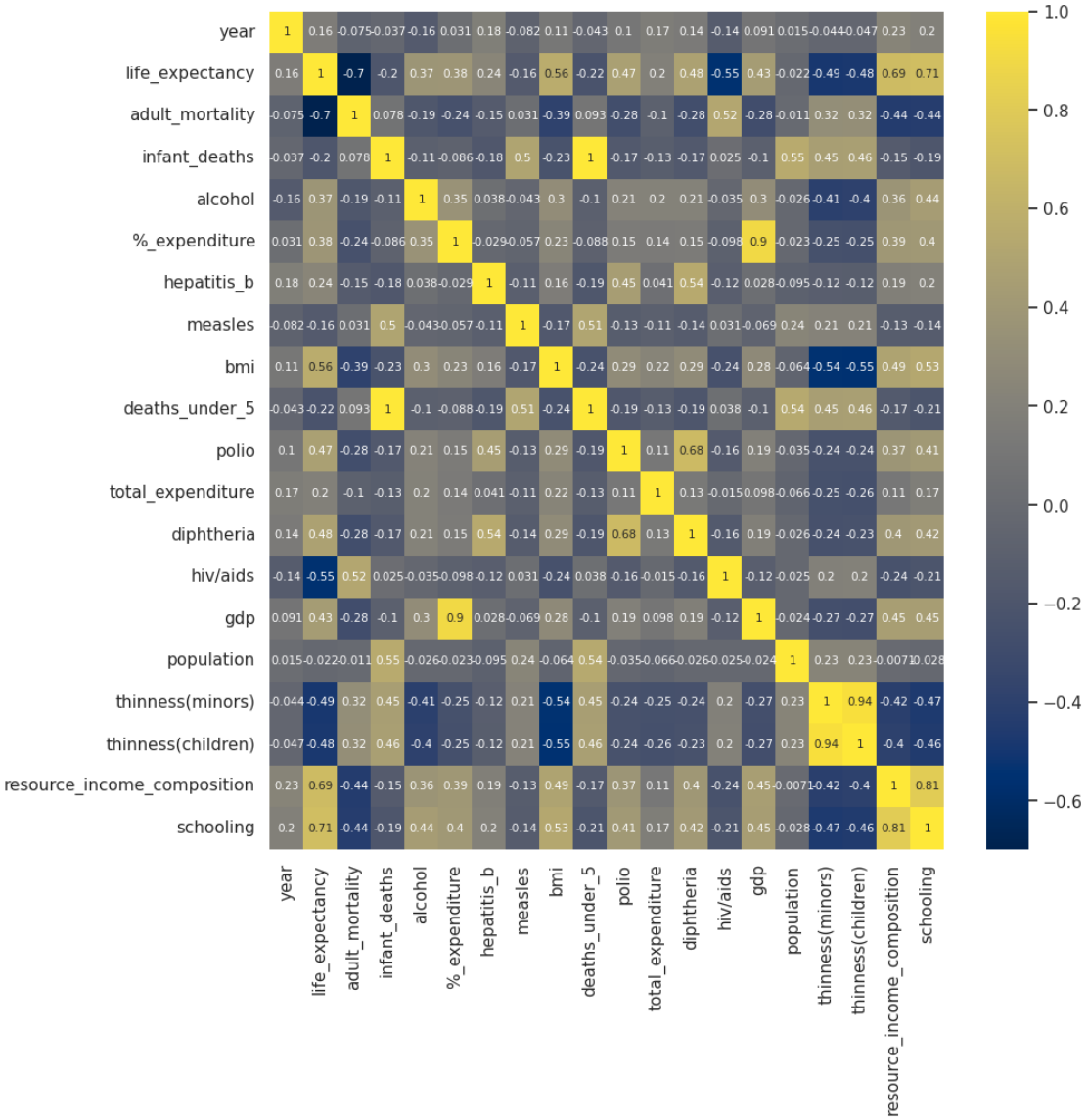
```
total_expenditure               0.041075 -0.106350  0.224941       -0.127693
diphtheria                      0.544834 -0.139408  0.287348       -0.193847
hiv/aids                       -0.120402  0.030899 -0.242053        0.038062
gdp                             0.028037 -0.069292  0.277227       -0.104590
population                     -0.095443  0.236648 -0.064073        0.537078
thinness(minors)               -0.118642  0.213993 -0.541140        0.454032
thinness(children)             -0.121277  0.212234 -0.548344        0.461368
resource_income_composition     0.188415 -0.131353  0.492557       -0.165510
schooling                       0.200507 -0.139708  0.529127       -0.209965

                                   polio  total_expenditure  diphtheria  \
year                            0.100888           0.168512    0.139863
life_expectancy                 0.466427           0.197784    0.479706
adult_mortality                -0.278584          -0.099697   -0.278265
infant_deaths                  -0.168433          -0.125881   -0.173320
alcohol                         0.213648           0.204966    0.209819
%_expenditure                   0.149324           0.141271    0.145546
hepatitis_b                     0.449330           0.041075    0.544834
measles                        -0.133235          -0.106350   -0.139408
bmi                             0.289049           0.224941    0.287348
deaths_under_5                 -0.186499          -0.127693   -0.193847
polio                           1.000000           0.112300    0.677326
total_expenditure               0.112300           1.000000    0.128324
diphtheria                      0.677326           0.128324    1.000000
hiv/aids                       -0.159240          -0.014722   -0.164668
gdp                             0.192364           0.097875    0.185063
population                     -0.035294          -0.065720   -0.025642
thinness(minors)               -0.236948          -0.247084   -0.242087
thinness(children)             -0.236655          -0.255694   -0.234909
resource_income_composition     0.374446           0.112728    0.396228
schooling                       0.405910           0.171326    0.416232

                                hiv/aids       gdp  population  thinness(minors)  \
year                           -0.139741  0.090686    0.014734         -0.043795
life_expectancy                -0.553916  0.429977   -0.021558         -0.485543
adult_mortality                 0.522176 -0.282525   -0.010972          0.316336
infant_deaths                   0.025231 -0.101404    0.549684          0.450986
alcohol                        -0.034509  0.302695   -0.026012         -0.406835
%_expenditure                  -0.097857  0.901868   -0.022760         -0.251818
hepatitis_b                    -0.120402  0.028037   -0.095443         -0.118642
measles                         0.030899 -0.069292    0.236648          0.213993
bmi                            -0.242053  0.277227   -0.064073         -0.541140
deaths_under_5                  0.038062 -0.104590    0.537078          0.454032
polio                          -0.159240  0.192364   -0.035294         -0.236948
total_expenditure              -0.014722  0.097875   -0.065720         -0.247084
diphtheria                     -0.164668  0.185063   -0.025642         -0.242087
hiv/aids                        1.000000 -0.119220   -0.024827          0.196772
```

```
gdp                           -0.119220  1.000000   -0.024231         -0.265523
population                    -0.024827 -0.024231    1.000000          0.226585
thinness(minors)               0.196772 -0.265523    0.226585          1.000000
thinness(children)             0.201320 -0.269499    0.226532          0.942874
resource_income_composition   -0.237470  0.453966   -0.007058         -0.417428
schooling                     -0.208304  0.447429   -0.027547         -0.471674

                              thinness(children)  resource_income_composition  \
year                                   -0.047279                    0.227183
life_expectancy                        -0.480154                    0.686130
adult_mortality                         0.320855                   -0.442175
infant_deaths                           0.459543                   -0.147503
alcohol                                -0.397093                    0.364723
%_expenditure                          -0.254134                    0.390984
hepatitis_b                            -0.121277                    0.188415
measles                                 0.212234                   -0.131353
bmi                                    -0.548344                    0.492557
deaths_under_5                          0.461368                   -0.165510
polio                                  -0.236655                    0.374446
total_expenditure                      -0.255694                    0.112728
diphtheria                             -0.234909                    0.396228
hiv/aids                                0.201320                   -0.237470
gdp                                    -0.269499                    0.453966
population                              0.226532                   -0.007058
thinness(minors)                        0.942874                   -0.417428
thinness(children)                      1.000000                   -0.403985
resource_income_composition            -0.403985                    1.000000
schooling                              -0.456930                    0.809767

                              schooling
year                           0.195565
life_expectancy                0.706699
adult_mortality               -0.437362
infant_deaths                 -0.194417
alcohol                        0.440712
%_expenditure                  0.398214
hepatitis_b                    0.200507
measles                       -0.139708
bmi                            0.529127
deaths_under_5                -0.209965
polio                          0.405910
total_expenditure              0.171326
diphtheria                     0.416232
hiv/aids                      -0.208304
gdp                            0.447429
population                    -0.027547
thinness(minors)              -0.471674
```

```
thinness(children)            -0.456930
resource_income_composition    0.809767
schooling                      1.000000
```

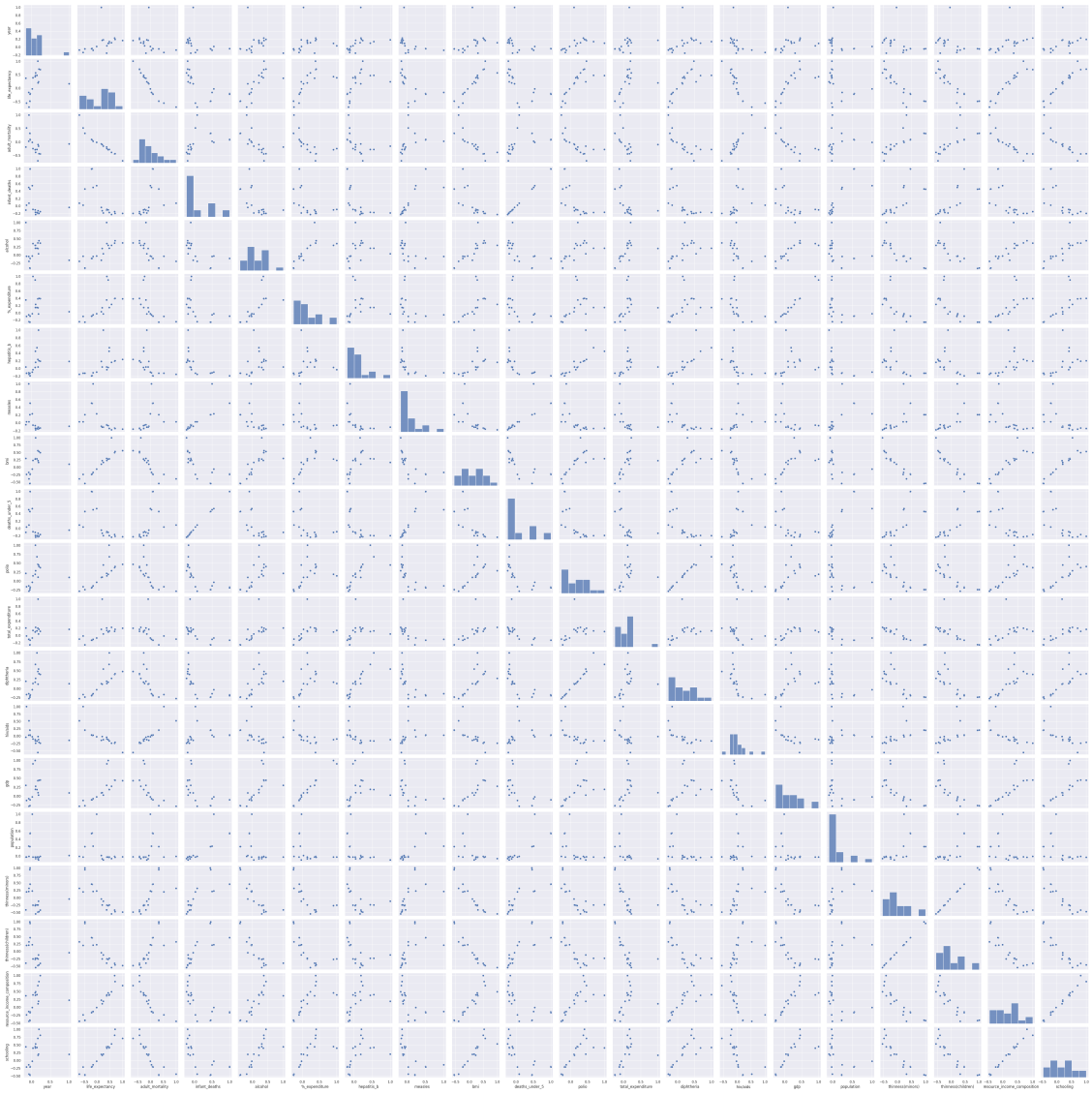
[55]:
```
# Heatmap of correlation
sns.set(rc = {'figure.figsize':(10, 10)})
sns.heatmap(life_df_nums.corr(), annot=True, cmap='cividis',annot_kws={'size':␣
↪7.5})
```

[55]: <Axes: >



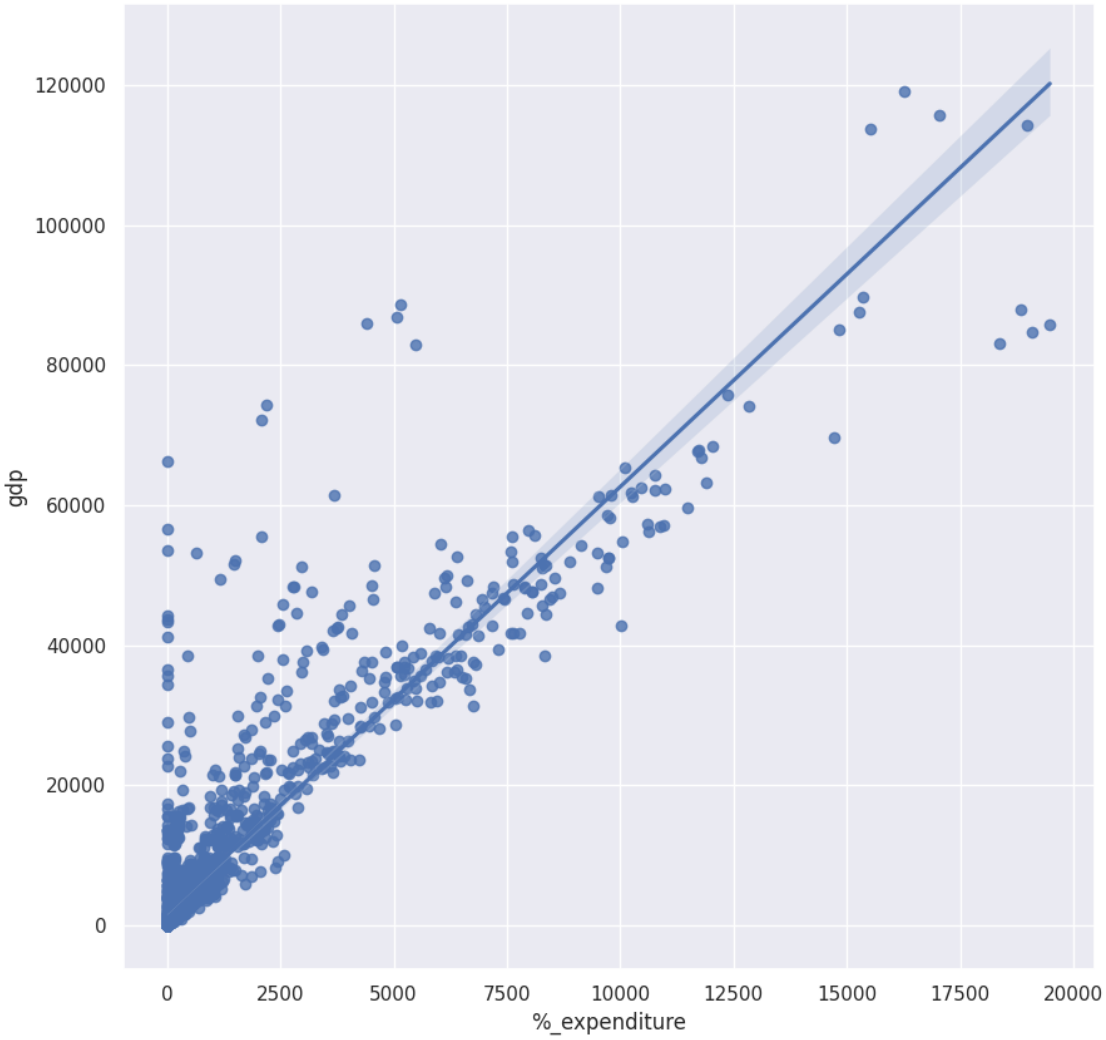
[56]:
```
sns.pairplot(life_df_nums.corr())
```

[56]: <seaborn.axisgrid.PairGrid at 0x788754e2b310>



[57]:
```
sns.regplot(data=life_df_nums, x='%_expenditure', y='gdp')
```

[57]: <Axes: xlabel='%_expenditure', ylabel='gdp'>


120000
100000
80000
60000
40000
20000
0
gdp
0
2500
5000
7500
10000
12500
15000
17500
20000
%_expenditure