TARC
TUNKU ABDUL RAHMAN
UNIVERSITY COLLEGE

Faculty of Computing and Information Technology

BACHELOR DEGREE IN MANAGEMENT MATHEMATICS WITH COMPUTING (HONOURS)

BAMS3043 Mathematical and Statistical Software

Assignment 4

Tutorial Group : RMM3S1G2
Tutor : Dr. Tan Yan Bin
Submission Due Date : 19 September 2021

Name	Student ID
Ong YiLiang	20WMR09187
Kang Tien Wey	20WMR09182
Chan Teik Chun	20WMR09179

Data Understanding and Preprocessing

```
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                  2928 non-null   float64
 4   Adult Mortality                  2928 non-null   float64
 5   infant deaths                    2938 non-null   int64
 6   Alcohol                          2744 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2385 non-null   float64
 9   Measles                          2938 non-null   int64
 10  BMI                              2904 non-null   float64
 11  under-five deaths                2938 non-null   int64
 12  Polio                            2919 non-null   float64
 13  Total expenditure                2712 non-null   float64
 14  Diphtheria                       2919 non-null   float64
 15  HIV/AIDS                         2938 non-null   float64
 16  GDP                              2490 non-null   float64
 17  Population                       2286 non-null   float64
 18  thinness  1-19 years             2904 non-null   float64
 19  thinness 5-9 years               2904 non-null   float64
 20  Income composition of resources  2771 non-null   float64
 21  Schooling                        2775 non-null   float64
```

Screenshot 1 - Data type

The target variable measures life expectancy from 2000 to 2015 among countries. The dataset provided contains numerous data with different variable names such as country, year, status and other variables. In screenshot 1, the variables are listed for non-null count which start with the variable named "country" with a 2938 non-null count, and end with the variable named "schooling" with a 2775 non-null count. The majority data type for the variables in the dataset is float, only "country" and "status" are strings, and the rest of the variables are integers.

Null values

```
Country                          0
Year                             0
Status                           0
Life Expectancy                 10
Adult Mortality                 10
Infant Deaths                    0
Alcohol Intake(L)              194
Percentage Expenditure           0
HepB Vaccination %             553
Measles                          0
BMI                             34
Under Five Deaths                0
Pol3 Vaccination %              19
Total Expenditure              226
Diphtheria Vaccination %        19
HIV/AIDS                         0
GDP                            448
Population                     652
Thinness 10-19 years            34
Thinness 5-9 years              34
Resources Income Composition   167
Schooling                      163
dtype: int64
```

Screenshot 2 - Number of null values in dataset

```
Country                        0
Year                           0
Status                         0
Life Expectancy                0
Adult Mortality                0
Infant Deaths                  0
Alcohol Intake(L)              0
Percentage Expenditure         0
HepB Vaccination %             0
Measles                        0
BMI                            0
Under Five Deaths              0
Pol3 Vaccination %             0
Total Expenditure              0
Diphtheria Vaccination %       0
HIV/AIDS                       0
GDP                            0
Population                     0
Thinness 10-19 years           0
Thinness 5-9 years             0
Resources Income Composition   0
Schooling                      0
dtype: int64
```

Screenshot 3 - After replace the null values by mean value in dataset

In screenshot 2, you can see that some variables in this dataset contain null values. For example, a variable named “GDP” in the dataset contains 448 numbers of null values. Thus, we replace those null values with their mean value according to the year. In screenshot 3, you can see that every of the variables did not contain null values anymore.

Label Encoding

```
df['Status'] = LabelEncoder().fit_transform(df['Status'])
df['Status'].unique()
#developing : 1, developed : 0
```

Screenshot 4 - replace column Status

In screenshot 4, we have change every row for the column or variable named “Status”, row with ‘developing’ become to 1, and row with ‘developed’ become to 0.

Detecting and dealing with outliers

	Factor	Lower Bound %	Upper Bound %
0	Year	0.00	0.00
1	Life Expectancy	0.58	0.00
2	Adult Mortality	0.00	2.93
3	Infant Deaths	0.00	10.72
4	Alcohol Intake(L)	0.00	0.10
5	Percentage Expenditure	0.00	13.24
6	HepB Vaccination %	7.56	0.00
7	Measles	0.00	18.45
8	BMI	0.00	0.00
9	Under Five Deaths	0.00	13.41
10	Pol3 Vaccination %	9.50	0.00
11	Total Expenditure	0.00	1.74
12	Diphtheria Vaccination %	10.14	0.00
13	HIV/AIDS	0.00	18.45
14	GDP	0.00	10.21
15	Population	0.00	6.91
16	Thinness 10-19 years	0.00	3.40
17	Thinness 5-9 years	0.00	3.37
18	Resources Income Composition	4.42	0.00
19	Schooling	2.18	0.44

Screenshot 5 - Outliers in dataset

In the dataset provided, there are outliers for all of the columns except columns named “BMI”. In screenshot 5, you can notice that most of these variables contain numbers in lower bound, or upper bound or both. Thus, we have normalized it with Winsorization which reduces the effect of possibly spurious outliers.

Correlation

	Factor	Coefficient	P-value	Relation	Result Certainity
0	Life Expectancy	1.000000	0.000000e+00	Strong Positive	Strong
1	HIV/AIDS	-0.817994	0.000000e+00	Strong Negative	Strong
2	Resources Income Composition	0.810609	0.000000e+00	Strong Positive	Strong
3	Schooling	0.757946	0.000000e+00	Strong Positive	Strong
4	Under Five Deaths	-0.629121	9.881313e-324	Strong Negative	Strong
5	Adult Mortality	-0.618449	1.131544e-309	Strong Negative	Strong
6	BMI	0.617655	1.165492e-308	Strong Positive	Strong
7	Infant Deaths	-0.601791	5.158060e-289	Strong Negative	Strong
8	Thinness 5-9 years	-0.597486	7.169131e-284	Strong Negative	Strong
9	Thinness 10-19 years	-0.593594	2.740681e-279	Strong Negative	Strong
10	Pol3 Vaccination %	0.591391	1.011286e-276	Strong Positive	Strong
11	Diphtheria Vaccination %	0.586092	1.249742e-270	Strong Positive	Strong
12	GDP	0.471528	1.494804e-162	Weak Positive	Strong
13	Percentage Expenditure	0.469177	9.702695e-161	Weak Positive	Strong
14	Status	-0.439894	2.595898e-139	Weak Negative	Strong
15	Alcohol Intake(L)	0.375008	9.304893e-99	Weak Positive	Strong
16	Measles	-0.356096	1.469739e-88	Weak Negative	Strong
17	HepB Vaccination %	0.339202	5.100247e-80	Weak Positive	Strong
18	Total Expenditure	0.232347	2.614963e-37	Weak Positive	Strong
19	Year	0.136086	1.285489e-13	Weak Positive	Strong
20	Population	0.017197	3.514344e-01	Weak Positive	Nil

Screenshot 6 - Correlation in dataset

In screenshot 6, the correlation table shows the relationship between independent variable and dependent variable. When the value of the correlation coefficient column is close to -1 or 1, it means the features are strongly correlated to the dependent variable, however if the value of the correlation coefficient is close to 0 means the features are weakly correlated to the dependent variable.

Q1 Modeling Simple Linear Regression

Let xi = value of the index i features

i = index indicate the feature according to the correlation ranking table, i ∈ {1, 2, 3, ..., 20}

Model1:

The HIV/AIDS feature is chosen as the independent variable for Model1 because this feature has the highest correlation(-0.817994) with life expectancy.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.669
Model:	OLS	**Adj. R-squared:**	0.669
Method:	Least Squares	**F-statistic:**	5937.
Date:	Mon, 13 Sep 2021	**Prob (F-statistic):**	0.00
Time:	17:31:11	**Log-Likelihood:**	-8565.4
No. Observations:	2938	**AIC:**	1.713e+04
Df Residuals:	2936	**BIC:**	1.715e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	73.9960	0.105	708.006	0.000	73.791	74.201
x1	-9.3965	0.122	-77.053	0.000	-9.636	-9.157

Screenshot 7 - simple linear regression with OLS method

y = 73.996 - 9.3965x1

Q2 Modeling two Multiple Linear Regression

Model2:

For Model2, we are choosing all the features that have a strong correlation(>0.5 or <-0.5) with life expectancy no matter if it is positive or negative correlation. The features are HIV/AIDS, Resources Income Composition, Schooling, Under Five Deaths, Adult Mortality, BMI, Infant Deaths, Thinness 5-9 years, Thinness 10-19 years, Pol3 Vaccination %, Diphtheria Vaccination %.

Dep. Variable:	y	**R-squared:**	0.868
Model:	OLS	**Adj. R-squared:**	0.867
Method:	Least Squares	**F-statistic:**	1743.
Date:	Mon, 13 Sep 2021	**Prob (F-statistic):**	0.00
Time:	17:34:33	**Log-Likelihood:**	-7220.2
No. Observations:	2938	**AIC:**	1.446e+04
Df Residuals:	2926	**BIC:**	1.454e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	59.2565	0.567	104.561	0.000	58.145	60.368
HIV/AIDS	-4.3640	0.111	-39.151	0.000	-4.583	-4.145
Resources Income Composition	14.7097	0.878	16.762	0.000	12.989	16.430
Schooling	0.0973	0.052	1.884	0.060	-0.004	0.199
Under Five Deaths	-0.2870	0.034	-8.546	0.000	-0.353	-0.221
Adult Mortality	-0.0138	0.001	-18.619	0.000	-0.015	-0.012
BMI	0.0067	0.004	1.582	0.114	-0.002	0.015
Infant Deaths	0.3375	0.047	7.124	0.000	0.245	0.430
Thinness 5-9 years	-0.2714	0.061	-4.480	0.000	-0.390	-0.153
Thinness 10-19 years	0.0225	0.061	0.372	0.710	-0.096	0.141
Pol3 Vaccination %	0.0242	0.007	3.538	0.000	0.011	0.038
Diphtheria Vaccination %	0.0342	0.006	5.357	0.000	0.022	0.047

Screenshot 8 - multiple linear regression for model2 with OLS method

y = 59.2565 - 4.364x1 + 14.7097x2 + 0.0973x3 - 0.287x4 - 0.0138x5 + 0.0067x6 + 0.3375x7 - 0.2714x8 + 0.0225x9 + 0.0242x10 + 0.0342x11

Model3:

Life expectancy is affected by many factors such as: socioeconomic status, including employment, income, education and economic wellbeing. (The Department of Health, 2012).

Thus, we improve the model as Model3 by adding the Status feature in Model2.

Dep. Variable:	y	**R-squared:**	0.868
Model:	OLS	**Adj. R-squared:**	0.868
Method:	Least Squares	**F-statistic:**	1609.
Date:	Mon, 13 Sep 2021	**Prob (F-statistic):**	0.00
Time:	17:42:52	**Log-Likelihood:**	-7210.5
No. Observations:	2938	**AIC:**	1.445e+04
Df Residuals:	2925	**BIC:**	1.452e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	60.1671	0.602	100.018	0.000	58.988	61.347
HIV/AIDS	-4.3958	0.111	-39.477	0.000	-4.614	-4.177
Resources Income Composition	14.3827	0.878	16.382	0.000	12.661	16.104
Schooling	0.0757	0.052	1.464	0.143	-0.026	0.177
Under Five Deaths	-0.3063	0.034	-9.071	0.000	-0.372	-0.240
Adult Mortality	-0.0133	0.001	-17.917	0.000	-0.015	-0.012
BMI	0.0075	0.004	1.795	0.073	-0.001	0.016
Infant Deaths	0.3647	0.048	7.658	0.000	0.271	0.458
Thinness 5-9 years	-0.2575	0.060	-4.259	0.000	-0.376	-0.139
Thinness 10-19 years	0.0399	0.061	0.658	0.511	-0.079	0.159
Pol3 Vaccination %	0.0238	0.007	3.485	0.000	0.010	0.037
Diphtheria Vaccination %	0.0340	0.006	5.344	0.000	0.022	0.046
Status	-0.7320	0.166	-4.406	0.000	-1.058	-0.406

Screenshot 9 - multiple linear regression for model 3 with OLS method

y = 60.1671 - 4.3958x1 + 14.3827x2 + 0.0757x3 - 0.03063x4 - 0.0133x5 + 0.0075x6 + 0.3647x7 - 0.2575x8 + 0.0399x9 + 0.0238x10 + 0.034x11 - 0.732x12

Q3 Compare the three models

R-squared (R^2) is a statistical measure that shows the proportion of the variance for a dependent variable that's explained by independent variables in a regression model. The correlation represents the strength of relationship between the dependent variable and independent variables. (Free Code Camp, 2017). R-Squared only works as intended in a simple linear regression model with one explanatory variable. With a multiple regression made up of several independent variables, the R-Squared must be adjusted. The adjusted R^2compares the descriptive power of regression models that include diverse numbers of predictors. (Jason Fernando, 2021).

Ranking of Adj R^2

	Model	Adjusted R^2
1	Model 3	0.867910
2	Model 2	0.867079
3	Model 1	0.669002

Screenshot 10 - compare models with adjusted R^2

Model 3 is the best model due to the highest value of R^2 that indicates the goodness of fitness to the model is 0.86791, which means that 86.791% of data fit into this model. However, model 2 and model 1 only got 86.7079% and 66.9002% respectively.

To improve the Life Expectancy:

- Polio, Diphtheria vaccination coverage should be increased
- Measures should be taken to ensure food security
- Measures should be taken to provide education and reduce the risks of infant mortality
- Resources should be utilized productively
- AIDS awareness campaigns should be organized.

Q4 95% confidence interval of life expectancy

AIM: find the optimal interval of life expectancy when people live in perfect and optimal condition.

Using Model 3, we are 95% confident to say that people are able to live for 86.54years to 92.51years when people likely to go schooling for 14.9years and control their BMI at 38.16, at the same time the country is developed, utilizes its resources productively at the index of 0.803, improve the vaccination coverage of Pol3 and Diphtheria to 98%, take action to lower the death rate of HIV/AIDS (0-4years), under five deaths and infant deaths to 0.1, 0 and 0.031 respectively, lastly lower the adult mortality rate to 0.019. Moreover, the prevalence of thinness in the society among 5-9years and among 10-19 years should be at 0.8% and 8% respectively.

Using central tendency - mean

Model 3 is the best model and we are 95% confident to say that people are able to live for 69.36years to 69.93years when people likely to go schooling for 11.97years and control their BMI at 38.16, at the same time the country is developed, utilizes its resources productively at the index of 0.636, improve the vaccination coverage of Pol3 and Diphtheria to around 85%, take action to lower the death rate of HIV/AIDS (0-4years), under five deaths and infant deaths to 0.52, 0.014 and 0.01 respectively, lastly lower the adult mortality rate to 0.1467. Moreover, the prevalence of thinness in the society among 5-9years and among 10-19 years should be at 4.15% and 4.12% respectively.

References

The Department of Health. 2012. *Life expectancy and wellbeing: Life expectancy at birth.* Viewed on 13 Sep 2021. Available from: <https://www1.health.gov.au/internet/publications/publishing.nsf/Content/oatsih-hpf-2012-toc~tier1~life-exp-wellb~119>

Free Code Camp. 2017. *Learn how to select the best performing linear regression for univariate models.* Viewed on 13 Sep 2021. Available from: <https://www.freecodecamp.org/news/learn-how-to-select-the-best-performing-linear-regression-for-univariate-models-e9d429c40581/>

Jason Fernando. 2021. *R-Squared.* Viewed on 13 Sep 2021. Available from: <https://www.investopedia.com/terms/r/r-squared.asp>

Statsmodels. 2021. *Statsmodels*. Viewed on 13 Sep 2021. Available at: <https://www.statsmodels.org/stable/index.html.>

Kaggle. 2021. *DS Project-Life Expectancy (WHO)*. Viewed on 13 Sep 2021. Available at: <https://www.kaggle.com/shabanamir/ds-project-life-expectancy-who.>