TARC

TUNKU ABDUL RAHMAN
UNIVERSITY COLLEGE

Faculty of Computing and Information Technology

Department of Mathematical and Data Science

Bachelor of Science (Honours) in Management Mathematics with Computing

Academic Year 2020/2021

BAMS3043 Mathematical and Statistical Software
Assignment 4

Programme of Study: RMM3S1G2

No	Student Name	Student ID
1	Eng Wei Hang	20WMR09180
2	Sim Ka Yee	20WMR09188

Task 1

First, we drop the row with the null value in the column of target variables and then drop three columns with the highest percentage of null value which are hepatitis B, GDP, and population. The percentage of null values of these three columns seems very high if compared to other columns. After that, we check the correlation between all variables to avoid collinearity problems. We found that three pairs of variables are highly correlated. The first pair is infant deaths and under-five deaths. The second pair is thinness 1-19 years and thinness 5-9 years while the third pair is income composition of resources and schooling. Thus, we check the correlation of each variable in each pair with our target variable, life expectancy to decide on dropping which column.

	Life expectancy
infant deaths	-0.196557
under-five deaths	-0.222529
Life expectancy	1.000000

	Life expectancy
thinness 1-19 years	-0.477183
thinness 5-9 years	-0.471584
Life expectancy	1.000000

	Life expectancy
Income composition of resources	0.724776
Schooling	0.751975
Life expectancy	1.000000

Based on the result generated, the columns of infant deaths, thinness 5-9 years, and income composition of resources decided to be dropped because these variables have a lower correlation with life expectancy. After the cleaning process, the dataset has been reduced to 12 features in total. Then, we perform the correlation analysis to each variable with the target variable.

	Life expectancy
Adult Mortality	-0.696359
Alcohol	0.404877
percentage expenditure	0.381864
Measles	-0.157586
BMI	0.567694
under-five deaths	-0.222529
Polio	0.465556
Total expenditure	0.218086
Diphtheria	0.479495
HIV/AIDS	-0.556556
thinness 1-19 years	-0.477183
Schooling	0.751975

According to the result, the 'schooling' variable has the highest percentage of correlation with the target variable among all the other variables. So, we choose '**schooling'** as the **independent variable, x1.**

Before fitting the model, we eliminate the outlier for both x1 and y variables. The elimination of the outlier is important because outliers will increase the variability in the data, which decreases the statistical power (Statistics By Jim, 2019).

OLS Regression Results

Dep. Variable:	Life expectancy	R-squared:	0.621
Model:	OLS	Adj. R-squared:	0.621
Method:	Least Squares	F-statistic:	4415.
Date:	Sat, 18 Sep 2021	Prob (F-statistic):	0.00
Time:	16:02:29	Log-Likelihood:	-8446.6
No. Observations:	2700	AIC:	1.690e+04
Df Residuals:	2698	BIC:	1.691e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	41.7359	0.435	96.034	0.000	40.884	42.588
Schooling	2.2967	0.035	66.448	0.000	2.229	2.365

Omnibus:	283.241	Durbin-Watson:	0.214
Prob(Omnibus):	0.000	Jarque-Bera (JB):	406.788
Skew:	-0.801	Prob(JB):	4.65e-89
Kurtosis:	4.025	Cond. No.	51.7

The simple linear regression model is **y = 41.7359 + 2.2967x1**

Task 2

For the second and third multiple regression models, we decided to fit the model with either 2, 3, or 4 variables based on the adjusted r-squared values. First, we will eliminate the outliers for each combination of x and y variables. After that, we will check the condition number of the combination of x variables and ignore those with condition number greater than 1000 to avoid multicollinearity being high. The high condition number might indicate that there are strong multicollinearity or other numerical problems (Everything is Correlated, 2016). Then, we will choose two combinations of x variables with the highest adjusted r-squared values. This is because adjusted r-squared can indicate how well terms fit a curve or line, but adjusts for the number of terms in a model. If the added variables are useless to a model, adjusted r-squared will decrease. If the added variables are useful to a model, the adjusted r-squared will increase (Statistics How To, n.d.).

	First Feature	Second Feature	Adj. R-Squared
55	HIV/AIDS	Schooling	0.646518
56	thinness 1-19 years	Schooling	0.643987
36	BMI	Schooling	0.635057
46	Polio	Schooling	0.625129
17	Alcohol	Schooling	0.621255

	First Feature	Second Feature	Third Feature	Adj. R-Squared
117	Polio	thinness 1-19 years	Schooling	0.653014
99	BMI	thinness 1-19 years	Schooling	0.652486
63	Alcohol	thinness 1-19 years	Schooling	0.649733
90	BMI	Polio	Schooling	0.643893
125	Diphtheria	HIV/AIDS	Schooling	0.643837

	First Feature	Second Feature	Third Feature	Forth Feature	Adj. R-Squared
17	Adult Mortality	Alcohol	HIV/AIDS	thinness 1-19 years	0.661031
123	BMI	Polio	thinness 1-19 years	Schooling	0.659938
74	Alcohol	BMI	thinness 1-19 years	Schooling	0.656942
90	Alcohol	Polio	thinness 1-19 years	Schooling	0.653788
10	Adult Mortality	Alcohol	BMI	HIV/AIDS	0.653179

From the result generated, adult Mortality, alcohol, HIV/AIDS, and thinness 1-19 years has the highest adjusted r-squared value which is 0.661031 while BMI, polio, thinness 1-19 years, and schooling has the second-highest value which is 0.659938. These two combinations of variables will be used to fit the second and third model. Again, before fitting the models, we will eliminate the outlier for both x and y variables.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life expectancy   R-squared:                       0.662
Model:                            OLS   Adj. R-squared:                  0.661
Method:                 Least Squares   F-statistic:                     1038.
Date:                Sat, 18 Sep 2021   Prob (F-statistic):               0.00
Time:                        16:02:53   Log-Likelihood:                -5892.7
No. Observations:                2129   AIC:                         1.180e+04
Df Residuals:                    2124   BIC:                         1.182e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                  77.8002      0.247    314.751      0.000      77.315      78.285
Adult Mortality        -0.0378      0.001    -31.999      0.000      -0.040      -0.035
Alcohol                 0.4257      0.023     18.699      0.000       0.381       0.470
HIV/AIDS               -5.0515      0.282    -17.885      0.000      -5.605      -4.498
thinness 1-19 years    -0.2809      0.030     -9.413      0.000      -0.339      -0.222
==============================================================================
Omnibus:                       85.761   Durbin-Watson:                   0.727
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              160.784
Skew:                          -0.299   Prob(JB):                     1.22e-35
Kurtosis:                       4.206   Cond. No.                         515.
==============================================================================
```

The second multiple linear regression model is **y = 77.8002 - 0.0378x1 + 0.4257x2 - 5.0515x3 - 0.2809x4**

OLS Regression Results

Dep. Variable:	Life expectancy	R-squared:	0.661
Model:	OLS	Adj. R-squared:	0.660
Method:	Least Squares	F-statistic:	1140.
Date:	Sat, 18 Sep 2021	Prob (F-statistic):	0.00
Time:	16:02:53	Log-Likelihood:	-7064.3
No. Observations:	2349	AIC:	1.414e+04
Df Residuals:	2344	BIC:	1.417e+04
Df Model:	4		
Covariance Type:	nonrobust		

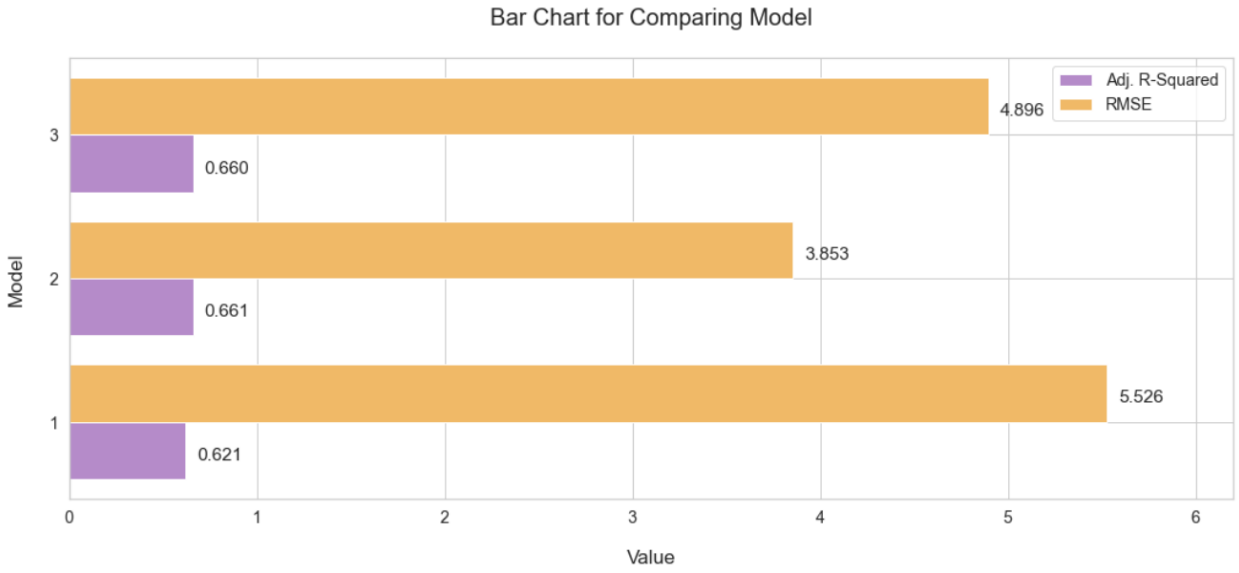
	coef	std err	t	P>|t|	[0.025	0.975]
const	36.7246	0.925	39.682	0.000	34.910	38.539
BMI	0.0460	0.007	6.982	0.000	0.033	0.059
Polio	0.1570	0.011	14.575	0.000	0.136	0.178
thinness 1-19 years	-0.3448	0.037	-9.333	0.000	-0.417	-0.272
Schooling	1.5610	0.049	31.733	0.000	1.465	1.657

Omnibus:	239.204	Durbin-Watson:	0.290
Prob(Omnibus):	0.000	Jarque-Bera (JB):	392.805
Skew:	-0.720	Prob(JB):	5.05e-86
Kurtosis:	4.394	Cond. No.	919.

The third multiple linear regression model is **y = 36.7246 + 0.0460x1 + 0.1570x2 - 0.3448x3 + 1.5610x4**

Task 3

After doing some research, we decided to use **adjusted R-squared** and **Root Mean Square Error(RMSE)** to compare the three models. Same as what has been said above, Adjusted R2 can indicate how well terms fit a curve or line, but adjusts for the number of terms in a model. On the other hand, RMSE can indicate the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction (The Analysis Factor, 2008).



Based on the result generated, model 2 has the highest value of adjusted r-squared and the lowest value of RMSE which are 0.6610 and 3.8532. Therefore, we conclude that **model 2** is the best model.

Task 4

Based on the answer in task 3, we know that the best model is model 2 and this model will be used to find the prediction interval of life expectancy. At first, we choose the X values using a measure of central tendency which is mean. We get the values of x as below:

x1 - Adult Mortality = 129.4814
x2 - Alcohol = 5.0305
x3 - HIV/AIDS = 0.2409
x4 - thinness 1-19 years = 3.6882

Based on the prediction interval calculated, we are 95% confident that the life expectancy is between 65.2323 years old and 80.3664 years old when the values of x variables are as above.

Reference

1. Statistics By Jim. 2019. *Guidelines for Removing and Handling Outliers in Data*. Available at: <https://statisticsbyjim.com/basics/remove-outliers> [Accessed 16 Sep. 2021].
2. Everything is Correlated. 2016. *Multiple Linear Regression - Python*. Available at: <https://lilithelina.tumblr.com/post/147984528439/multiple-linear-regression-python> [Accessed 16 Sep. 2021].
3. Statistics How To. (n.d.). *Adjusted R2 / Adjusted R-Squared: What is it used for?* Available at: <https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/adjusted-r2/> [Accessed 16 Sep. 2021].
4. The Analysis Factor. 2008. *Assessing the Fit of Regression Models - The Analysis Factor*. Available at: <https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/> [Accessed 16 Sep. 2021].