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	Fong Wei Chen	20WMR09181
2	Wong Ke Ying	20WMR09191
3	Wong Yet Xiang	20WMR09193

Question 1

In this project, we are interested in investigating the relationship between the dependent variable 'Life expectancy' and the independent variables in the developed country only. Therefore, we will subset the rows with status equals to 'Developed' in the data and use it for our regression model.

Variable Choosing

Correlations

		Lifeexpectancy	LINT (Incomecompositionofresources)
Pearson Correlation	Lifeexpectancy	1.000	.712
	LINT (Incomecompositionofresources)	.712	1.000
Sig. (1-tailed)	Lifeexpectancy	.	.000
	LINT (Incomecompositionofresources)	.000	.
N	Lifeexpectancy	512	512
	LINT (Incomecompositionofresources)	512	512

Figure 1.1

After investigation done on all the independent variables, we decide to choose 'Income composition of resources' as the independent variable. This variable has a strong linear relationship with the dependent variable 'Life expectancy'. A great linear relationship defines that the linear function can be explained clearly with the independent variable. The closer the correlation coefficient to 1 or -1, the stronger the linear relationship between the two variables. 'Income composition of resources' has a Pearson correlation coefficient of 0.712, which is close to 1. (Utexas.edu, 2016)

Missing Value

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Incomecompositionofresources	463	90.4%	49	9.6%	512	100.0%

Figure 1.2

From Figure 1.2, there are some missing values found in the column of our chosen variable 'Income composition of resources'.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
LINT (Incomecompositionofresources)	512	100.0%	0	0.0%	512	100.0%

Figure 1.3

To solve this problem, we replace the missing values with the mean value of the column.

Outliers

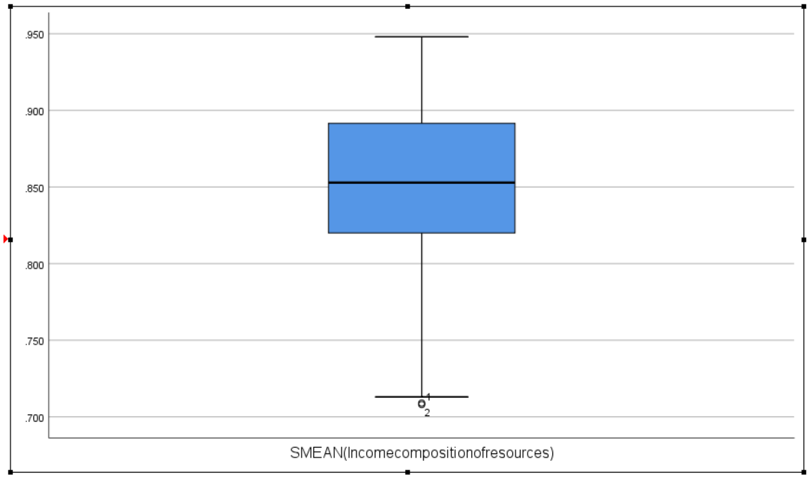

Figure 1.4

We run the descriptive statistics to figure out whether there are outliers in the data. As we can see from Figure 1.4, there are some outliers in the boxplot chart. Outliers will dramatically change the magnitude of regression coefficients and the direction of coefficient signs. Therefore, it is important to fix the problems (Choi, 2009).

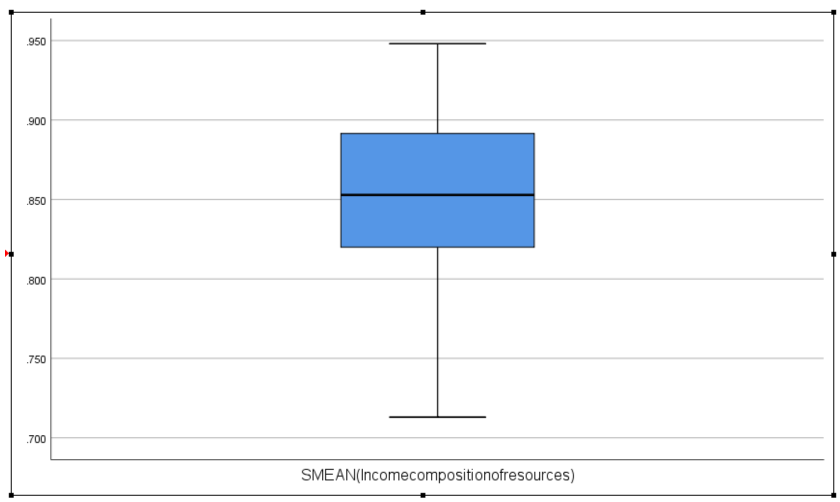

Figure 1.5

To fix the outliers, we replace the two outliers found with the minimum value in the variable. After dealing with the outliers, the data is ready to fit into linear regression.

Simple Linear Regression Fitting (Model A)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.712[a]	.507	.506	2.7636	1.858

a. Predictors: (Constant), SMEAN(Incomecompositionofresources)

b. Dependent Variable: Lifeexpectancy

Figure 1.6

R represents the correlation between the dependent variable and independent variable. 0.712 shows that these 2 variables are having a strong correlation. Besides that, R square value determines how much of the total variation in the dependent variable, Life expectancy, can be explained by the independent variable, Income composition of resources. In this case, 50.7% can be explained, which is quite large. (Laerd Statistics, 2018)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4001.109	1	4001.109	523.891	.000[b]
	Residual	3895.018	510	7.637		
	Total	7896.128	511			

a. Dependent Variable: Lifeexpectancy

b. Predictors: (Constant), SMEAN(Incomecompositionofresources)

Figure 1.7

The regression model predicts the dependent variable significantly. This is because the significant value in Figure 1.7 is less than 0.05.

Coefficients[a]										
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	31.290	2.097		14.924	.000	27.171	35.409		
	SMEAN (Incomecompositionofresources)	56.175	2.454	.712	22.889	.000	51.353	60.997	1.000	1.000

a. Dependent Variable: Lifeexpectancy

Figure 1.8

The regression equation can be written as:

$$\hat{Y} = 31.290 + 56.175x$$

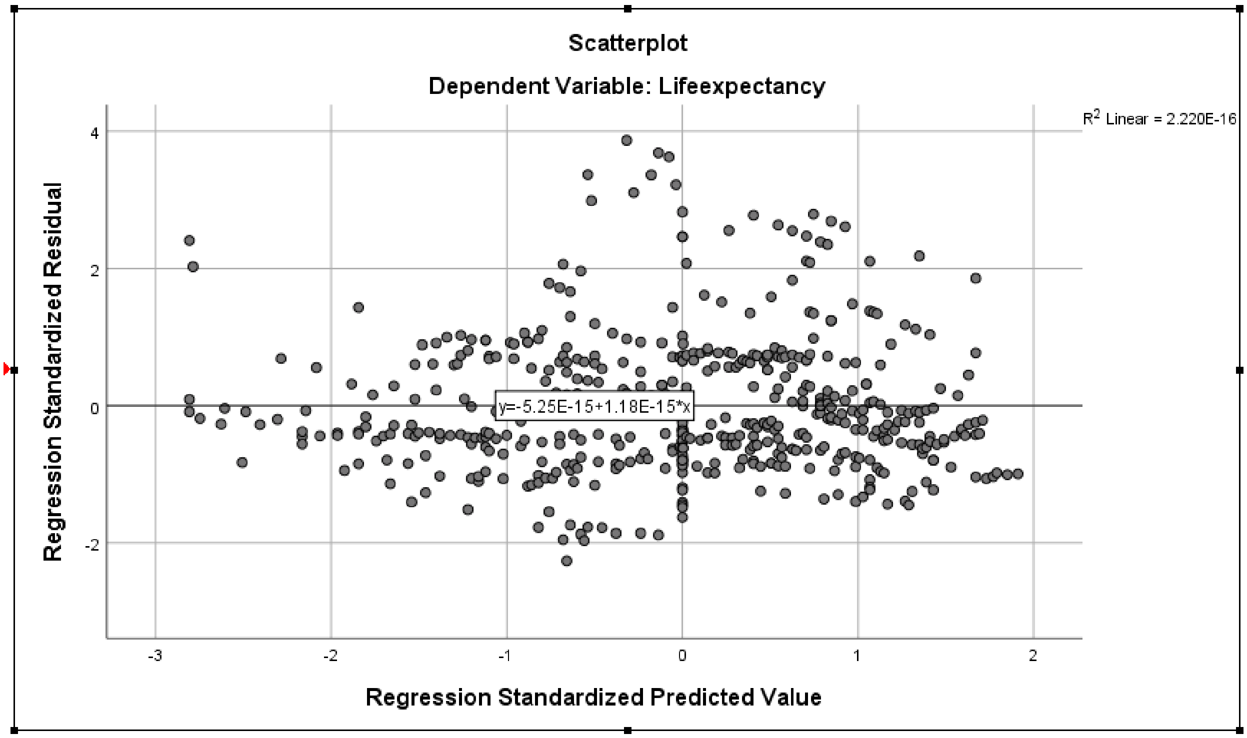

Figure 1.9

Question 2

Independent Variables Choosing

First Assumption

Our first assumption is based on the correlation between the independent variables and the dependent variable. The closer the coefficient of correlation to 1 or -1, the stronger the linear relationship between the variables. We have decided to choose those independent variables that have coefficients of correlation greater than 0.1 or smaller than 0.1.

Correlations

		Lifeexpectancy	LINT (Incomecompositionofresources)	AdultMortality	percent pendi
Pearson Correlation	Lifeexpectancy	1.000	.727	-.432	
	LINT (Incomecompositionofresources)	.727	1.000	-.403	
	AdultMortality	-.432	-.403	1.000	
	percentageexpenditure	.421	.535	-.228	
	thinness119years	-.742	-.634	.520	
	thinness59years	-.725	-.620	.521	
	Schooling	.357	.643	-.212	
	GDP	.387	.506	-.228	
	Alcohol	-.073	.020	-.013	
	infantdeaths	-.079	-.088	.220	
	HepatitisB	-.078	-.213	.163	
	Measles	-.051	-.041	.091	
	BMI	.011	-.008	-.013	
	underfivedeaths	-.032	-.031	.209	
	Polio	.060	.032	.071	
	Totalexpenditure	.179	.131	-.174	
	Diphtheria	-.015	-.018	-.014	
	HIVAIDS	.	.	.	
	Population	.123	.117	-.032	

Figure 2.1

After the first assumption, we have 9 independent variables left. Then we will fix the missing values problem in these 9 columns. We replace the missing value with the mean of each column.

Correlations

		Lifeexpectancy	LINT (Incomecompositionofresources)	AdultMortality	percentageexpenditure	thinness119years	thinness59years	Schooling	GDP	Totalexpenditure	Population
Pearson Correlation	Lifeexpectancy	1.000	.719	-.465	.402	-.664	-.662	.384	.377	.106	.081
	LINT (Incomecompositionofresources)	.719	1.000	-.430	.539	-.632	-.641	.670	.523	.151	.051
	AdultMortality	-.465	-.430	1.000	-.240	.487	.496	-.253	-.244	-.146	-.024
	percentageexpenditure	.402	.539	-.240	1.000	-.381	-.388	.272	.926	.107	.054
	thinness119years	-.664	-.632	.487	-.381	1.000	.991	-.407	-.379	-.222	-.096
	thinness59years	-.662	-.641	.496	-.388	.991	1.000	-.417	-.380	-.231	-.077
	Schooling	.384	.670	-.253	.272	-.407	-.417	1.000	.207	.190	.043
	GDP	.377	.523	-.244	.926	-.379	-.380	.207	1.000	.120	.017
	Totalexpenditure	.106	.151	-.146	.107	-.222	-.231	.190	.120	1.000	-.156
	Population	.081	.051	-.024	.054	-.096	-.077	.043	.017	-.156	1.000

Figure 2.2

Notice that after missing values are replaced, the correlation coefficient may vary compared to Figure 2.1. For example, the correlation coefficient of the independent variable ‘Population’ has decreased to a value lower than 0.1. However, we will still remain this independent variable since it already passed the first assumption.

Second Assumption

The second assumption is the multicollinearity. Multicollinearity exists when the independent variables in a regression model are highly correlated. If the correlation coefficient is between 0 and 0.3 (0 and -0.3), it means that the relationship between two variables is weak. On the other hand, if the coefficient is between 0.7 and 1.0 (-0.7 and -1.0), it means that the two variables are strongly correlated (Ratner, 2009). Multicollinearity brings negative impacts such as reducing the precision of the estimated coefficients that weakens the statistical power of the regression model. We might not be able to rely on the p-values to determine the independent variables that are statistically significant (Frost, n.d).

From Figure 2.2, we can notice that the independent variables ‘thinness 1 - 19 years’ and ‘thinness 5 - 9 years’ are strongly correlated. The correlation coefficient between these two independent variables is 0.991. Therefore, we will remove one of these variables from our model. ‘thinness 1 - 19 years’ is removed since ‘thinness 5 - 9 years’ has stronger linear relationship with the dependent variable compared to it.

		Lifeexpectancy	AdultMortality	percentageexpenditure	thinness59years	Totalexpenditure	SMEAN (Incomecompositionofresources)	SMEAN (Schooling)	SMEAN(GDP)	SMEAN (Population)
Pearson Correlation	Lifeexpectancy	1.000	-.455	.405	-.604	.072	.709	.383	.358	.075
	AdultMortality	-.455	1.000	-.213	.431	-.155	-.430	-.245	-.225	-.022
	percentageexpenditure	.405	-.213	1.000	-.315	.005	.514	.269	.871	.048
	thinness59years	-.604	.431	-.315	1.000	-.318	-.592	-.451	-.317	-.072
	Totalexpenditure	.072	-.155	.005	-.318	1.000	.138	.227	.075	-.125
	SMEAN (Incomecompositionofresources)	.709	-.430	.514	-.592	.138	1.000	.666	.499	.049
	SMEAN(Schooling)	.383	-.245	.269	-.451	.227	.666	1.000	.162	.038
	SMEAN(GDP)	.358	-.225	.871	-.317	.075	.499	.162	1.000	.017
	SMEAN(Population)	.075	-.022	.048	-.072	-.125	.049	.038	.017	1.000

Figure 2.3

After the two assumptions above, the eligible 8 independent variables are shown in Figure 2.3.

Outliers

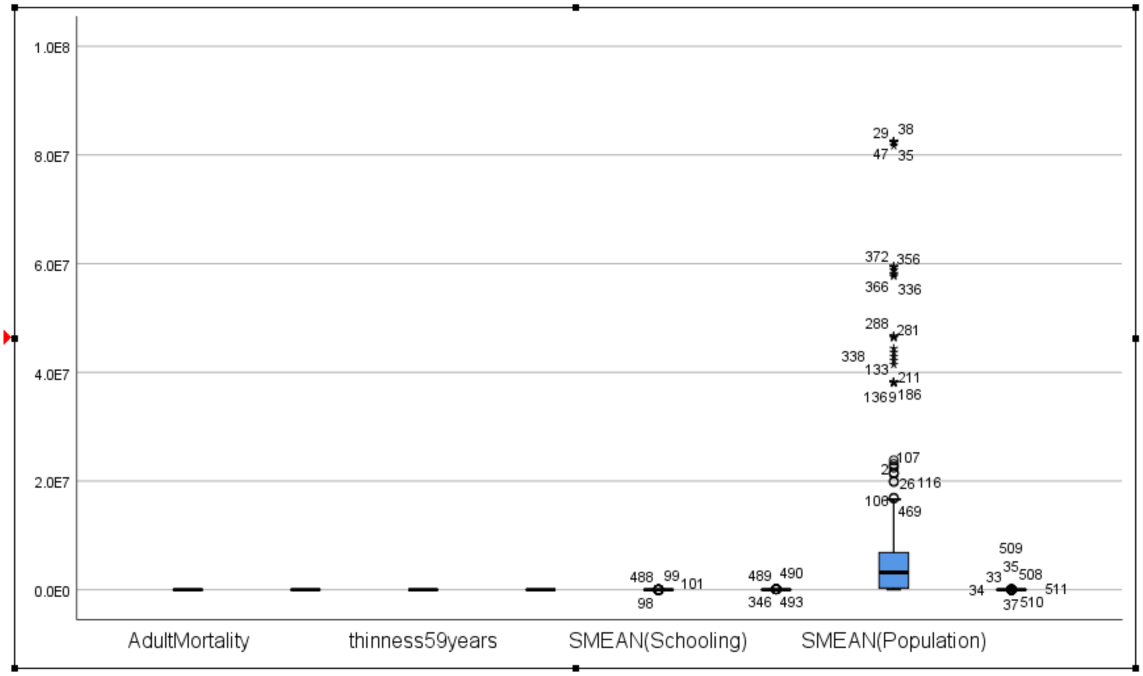

Figure 2.4

We can notice that 4 of the 8 chosen independent variables have outliers. We will replace the outliers of each variable with the maximum or minimum value of that particular variable range.

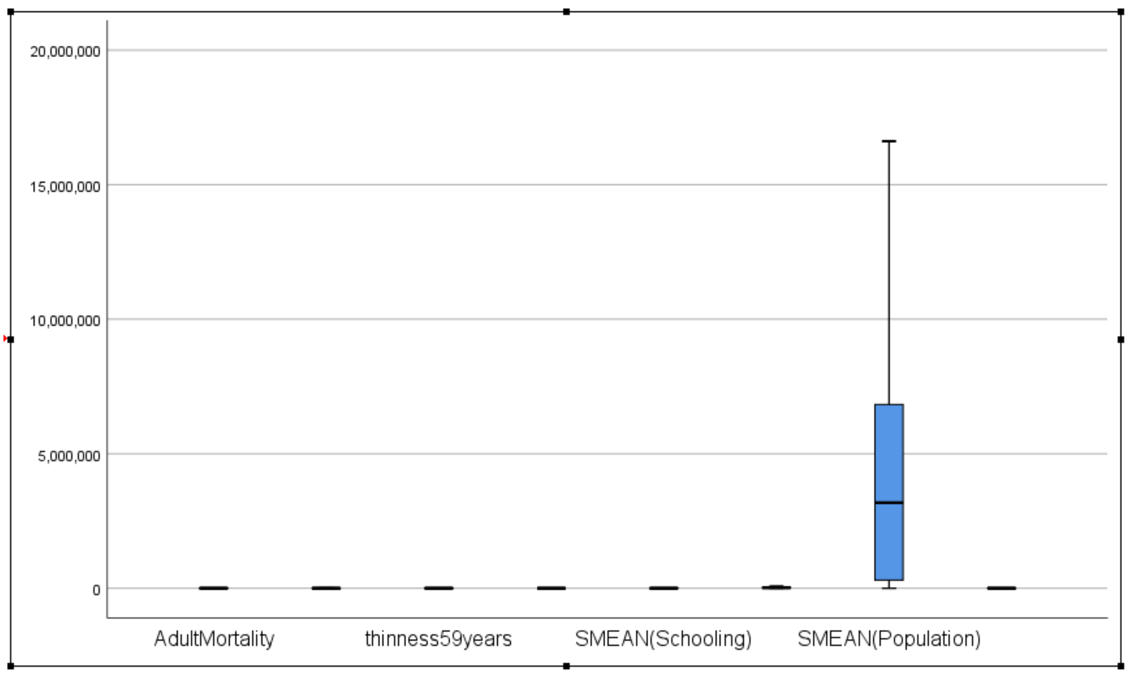

Figure 2.5

Multi Linear Regression Fitting

First Model (Model B)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.778[a]	.605	.599	2.4901	1.695

a. Predictors: (Constant), SMEAN(Totalexpenditure), SMEAN(Population), percentageexpenditure, AdultMortality, SMEAN(Schooling), thinness59years, SMEAN(Incomecompositionofresources), SMEAN(GDP)

b. Dependent Variable: Lifeexpectancy

Figure 2.6

As we can see in Figure 2.6, the R value is the multiple correlation coefficient. It can be used to determine the quality of the prediction of the dependent variable (Laerd Statistics, 2018). Our first multiple linear regression model has a R value of 0.778 that indicates a good level of prediction. The R Square value is 0.605 which means that 60.5% of the variability of the dependent variable is explained by the chosen 8 independent variables.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4777.234	8	597.154	96.306	.000[b]
	Residual	3118.893	503	6.201		
	Total	7896.128	511			

a. Dependent Variable: Lifeexpectancy

b. Predictors: (Constant), SMEAN(Totalexpenditure), SMEAN(Population), percentageexpenditure, AdultMortality, SMEAN(Schooling), thinness59years, SMEAN (Incomecompositionofresources), SMEAN(GDP)

Figure 2.7

The first multiple linear regression model predicts the dependent variable significantly. This is because the significant value in Figure 2.7 is less than 0.05.

Coefficients[a]										
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	46.388	2.814		16.483	.000	40.859	51.917		
	AdultMortality	-.012	.003	-.126	-3.880	.000	-.018	-.006	.748	1.338
	percentageexpenditure	.000	.000	.196	3.448	.001	.000	.000	.243	4.114
	thinness59years	-1.337	.180	-.279	-7.439	.000	-1.690	-.984	.559	1.788
	SMEAN (Incomecompositionofresources)	51.851	3.863	.657	13.422	.000	44.261	59.441	.328	3.052
	SMEAN(Schooling)	-.498	.099	-.203	-5.009	.000	-.693	-.302	.476	2.101
	SMEAN(GDP)	-4.360E-5	.000	-.220	-3.710	.000	.000	.000	.223	4.484
	SMEAN(Population)	1.514E-8	.000	.019	.655	.513	.000	.000	.945	1.058
	SMEAN(Totalexpenditure)	-.084	.052	-.050	-1.624	.105	-.185	.018	.825	1.212

a. Dependent Variable: Lifeexpectancy

Figure 2.8

The regression equation can be written as:

$$\hat{Y} = 46.388 - 0.012\, x1 - 1.337\, x3 + 51.851\, x4 - 0.498\, x5 - 4.360E\text{-}5\, x6 + 1.514E\text{-}8\, x7 - 0.084\, x8$$

From the ‘Sig’ column in Figure 2.8, we can see that 2 out of 8 independent variables are not significant in predicting the dependent variable. The not significant independent variables are ‘Population’ and ‘Total expenditure’.

Second Model (Model C)

The second multiple linear regression model is the reduced model of our first model. Stepwise regression is a technique that uses an algorithm to perform a number of times of multiple regression by removing the weakest correlated variable each time. (ScaleStatistics.com, n.d.) We used it to choose the best combination of independent variables to predict the outcomes.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.712[a]	.507	.506	2.7636
2	.747[b]	.558	.557	2.6178
3	.761[c]	.579	.577	2.5567
4	.768[d]	.590	.587	2.5258
5	.771[e]	.594	.590	2.5175

a. Predictors: (Constant), SMEAN (Incomecompositionofresources)

b. Predictors: (Constant), SMEAN (Incomecompositionofresources), thinness59years

c. Predictors: (Constant), SMEAN (Incomecompositionofresources), thinness59years, SMEAN(Schooling)

d. Predictors: (Constant), SMEAN (Incomecompositionofresources), thinness59years, SMEAN(Schooling), AdultMortality

e. Predictors: (Constant), SMEAN (Incomecompositionofresources), thinness59years, SMEAN(Schooling), AdultMortality, SMEAN (Totalexpenditure)

Figure 2.9

From the figure above, we can see that the R Square and Adjusted R Square values are increasing when the independent variables increase. Therefore, our second model will be the final model in row five which consists of 5 independent variables. It has a R value of 0.771 that indicates a good level of prediction and a R Square value of 0.594 which means 59.4% of the variability of the dependent variable is explained by the 5 independent variables.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
5	Regression	4689.076	5	937.815	147.966	.000[f]
	Residual	3207.052	506	6.338		
	Total	7896.128	511			

f. Predictors: (Constant), SMEAN(Incomecompositionofresources), thinness59years, SMEAN(Schooling), AdultMortality, SMEAN(Totalexpenditure)

Figure 2.10

The second multiple linear regression model predicts the dependent variable significantly. This is because the significant value in Figure 2.10 is less than 0.05.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
5	(Constant)	47.540	2.668		17.821	.000
	SMEAN (Incomecompositionofresources)	48.794	3.492	.618	13.975	.000
	thinness59years	-1.358	.181	-.283	-7.489	.000
	SMEAN(Schooling)	-.411	.096	-.168	-4.262	.000
	AdultMortality	-.011	.003	-.123	-3.763	.000
	SMEAN(Totalexpenditure)	-.108	.052	-.064	-2.084	.038

a. Dependent Variable: Lifeexpectancy

Figure 2.11

The regression equation can be written as:

$$\hat{Y} = 47.540 + 48.794\, x1 - 1.358\, x2 - 0.411\, x3 - 0.011\, x4 - 0.108\, x5$$

From the 'Sig' column in Figure 2.11, we can see that all the independent variables are significant in predicting the dependent variable.

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
5	percentageexpenditure	.021[f]	.661	.509	.029	.776
	SMEAN(GDP)	-.051[f]	-1.511	.131	-.067	.716
	SMEAN(Population)	-.002[f]	-.071	.943	-.003	.981

f. Predictors in the Model: (Constant), SMEAN(Incomecompositionofresources), thinness59years, SMEAN(Schooling), AdultMortality, SMEAN(Totalexpenditure)

Figure 2.12

These are the variables excluded from the final model when the system starts to test with stepwise regression. It is because the p - value of the independent variables in that model are greater than 0.05.

Question 3

In conclusion, we think that model B is the best model because the adjusted R - square value among these three models is the highest which is 0.599. Adjusted R square is mainly used to compare the goodness of fit for regression models that contain different numbers of independent variables. Besides, the R square value in model B also indicates that 60.5% of the variability of the dependent variable is explained by the chosen 8 independent variables. Thus, it indicates that the independent variables have a better fit in model B compared to the other models.

Question 4

By using mean value of x1 = 77, x2 = 2502.891616, x3 = 1.3, x4 = 0.853, x5 = 15.82, x6 = 21595.8638400, x7 = 4521699.17, x8 = 7.56, the 95% Prediction Interval of life expectancy is $74.30050 < Y^* < 84.09459$.

References

Choi S.W. (2009). *"The Effect of Outliers on Regression Analysis: Regime Type and Foreign Direct Investment | Request PDF.* [online] Available at: https://www.researchgate.net/publication/267025663_The_Effect_of_Outliers_on_Regression_Analysis_Regime_Type_and_Foreign_Direct_Investment

Frost, J. (n.d.). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions.* Available at: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Laerd Statistics. (2018). *Linear Regression Analysis in SPSS Statistics - Procedure, Assumptions and Reporting the output.* [online] Laerd.com. Available at: https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php

Laerd Statistics. (2018). *Multiple Linear Regression Analysis using SPSS Statistics.* Available at: https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php

Ratner, B. (2009). *The correlation coefficient: Its values range between +1/−1, or do they?.* [online] Available at: https://link.springer.com/article/10.1057/jt.2009.5

ScaleStatistics.com. (n.d.). *Use and Interpret Stepwise Regression in SPSS.* [online] Available at: https://www.scalestatistics.com/stepwise-regression.html

Utexas.edu. (2016). *Pearson Correlation and Linear Regression.* [online] Available at: http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/