Exploring the impact of BMI and Schooling on life expectancy by country

by Nick Barra, Lauren Díaz Morgan, Chris Ramirez, & Nive Venkat

COR1-GB.1305.11 – Statistics and Data Analysis Fall 2023

1) Describe your data set (including the source). Why are you interested in it? What do you hope to learn? Before exploring your data set, state some hypotheses (guesses) about how the variables should be related, perhaps based on your knowledge and experience. Be sure to identify the response variable and the predictor variables.

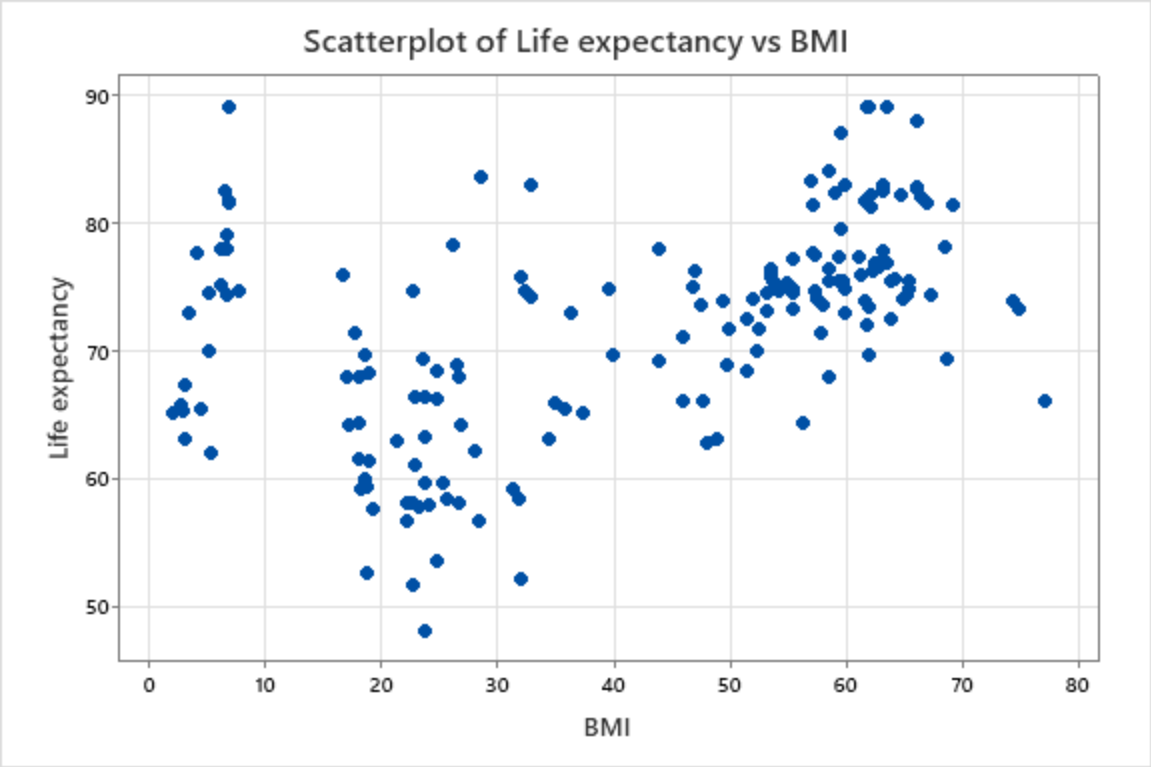
We chose to study Life Expectancy as our response variable, with BMI Index and Years of Schooling as our predictor variables. The data was accessed from Kaggle and originally published by the World Health Organization (WHO). Though the original data set includes observations for each country from 2000 - 2015, our team decided to focus on one specific year. Our data set shows the life expectancy of all countries in 2014 for project usage.

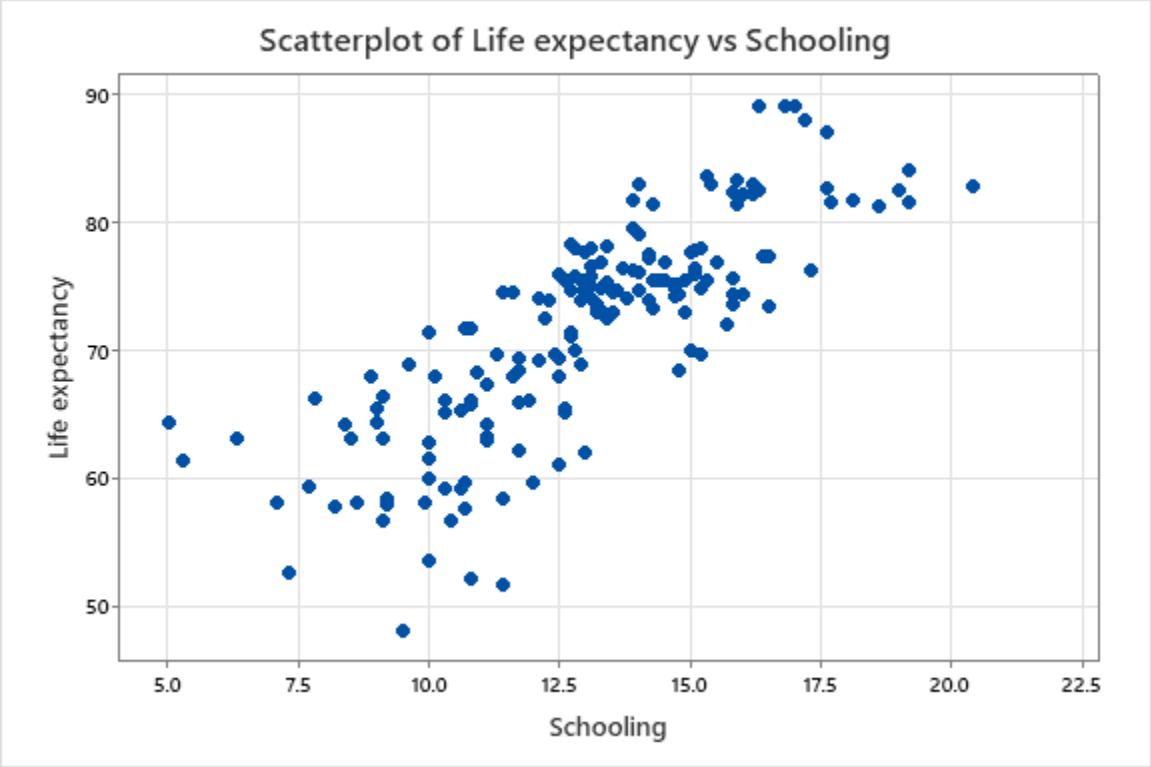
The life expectancy dataset can help us understand how socio-economic factors, education, and lifestyle choice can influence life expectancy outcomes. We hope to learn which areas countries should focus on to efficiently improve the life expectancy of their populations.

We hypothesize that there will be a positive linear relationship between Life Expectancy and BMI Index. For example, more developed countries would have a higher BMI and those countries would also have a higher life expectancy.

We also hypothesize a positive linear relationship between Life Expectancy and Schooling. Higher levels of education can correlate with better health, as individuals with more education tend to make healthier lifestyle choices and also have greater access to healthcare.

2) Make a scatterplot of your response variable (on the Y-axis) versus one of the predictor variables (on the X-axis). Describe the pattern you see. Is this pattern consistent with what you expected? Note any apparent outliers in the plot. Repeat the entire procedure for the other predictor variables.


Scatterplot of Life expectancy vs BMI
Life expectancy
BMI
90
80
70
60
50
0
10
20
30
40
50
60
70
80


Scatterplot of Life expectancy vs Schooling
Life expectancy
Schooling
90
80
70
60
50
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5

We see a strong, linear relationship between Life Expectancy and Schooling, which was what we expected. We do not see any major outliers, however, there is one country (Sierra Leone) that has a low Life Expectancy (~48 years) and relatively low Schooling (9.5 years), which may be an outlier.

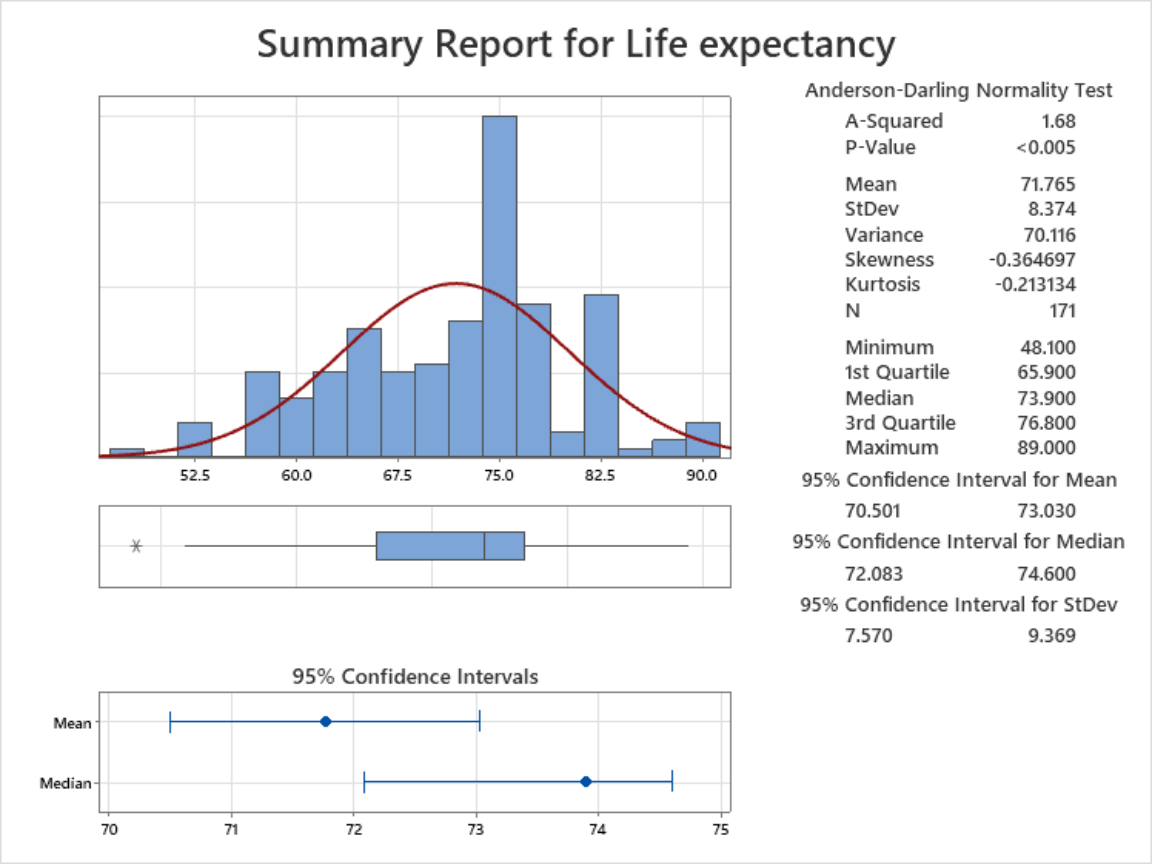
3) Can you think of any other variables (not in your data set) that might be useful in predicting Y? Try to list a few possibilities.

We could have looked at other variables included in the original data set from the WHO, such as: Income Composition of Resources, Total Government Expenditure on Health, GDP, or Alcohol Consumption.

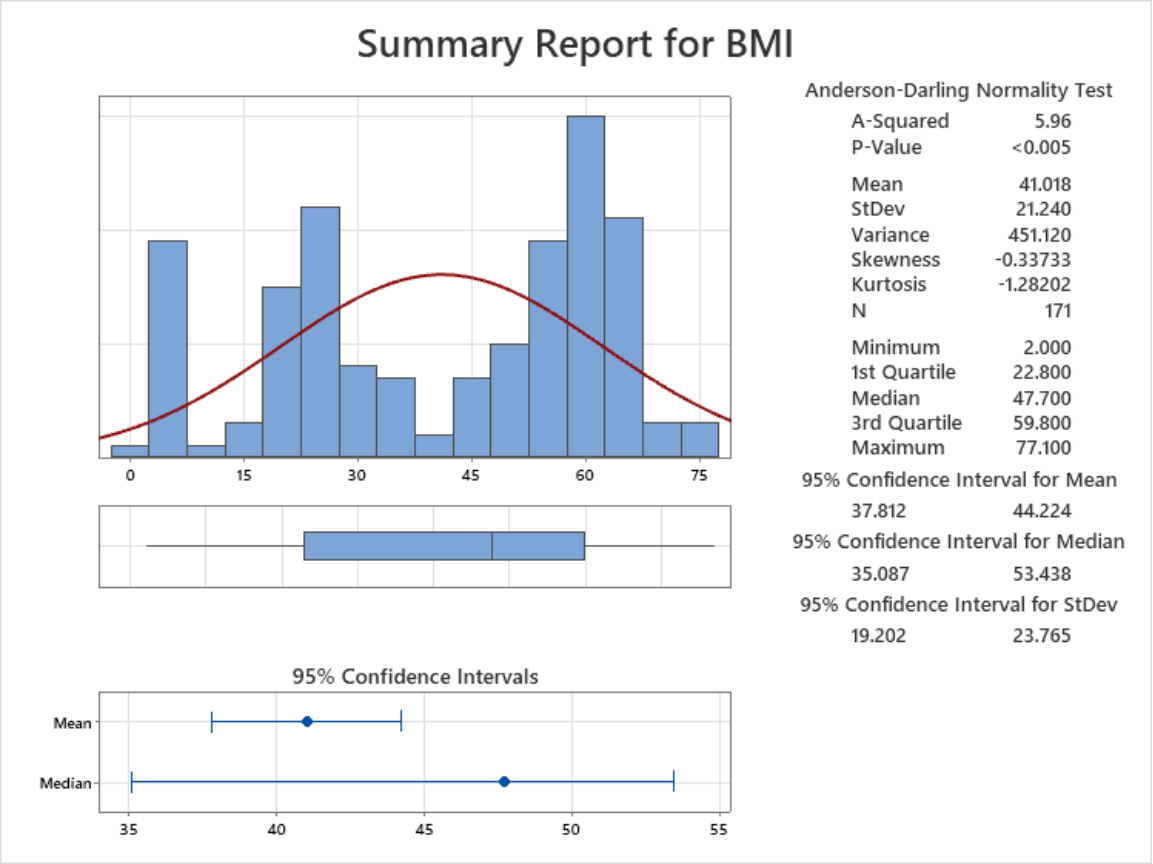
A few other variables that could be useful for predicting Life Expectancy are:

1. *Physical Activity*
2. *Environmental factors (e.g. air quality or water sanitation)*
3. *Tobacco or drug usage*

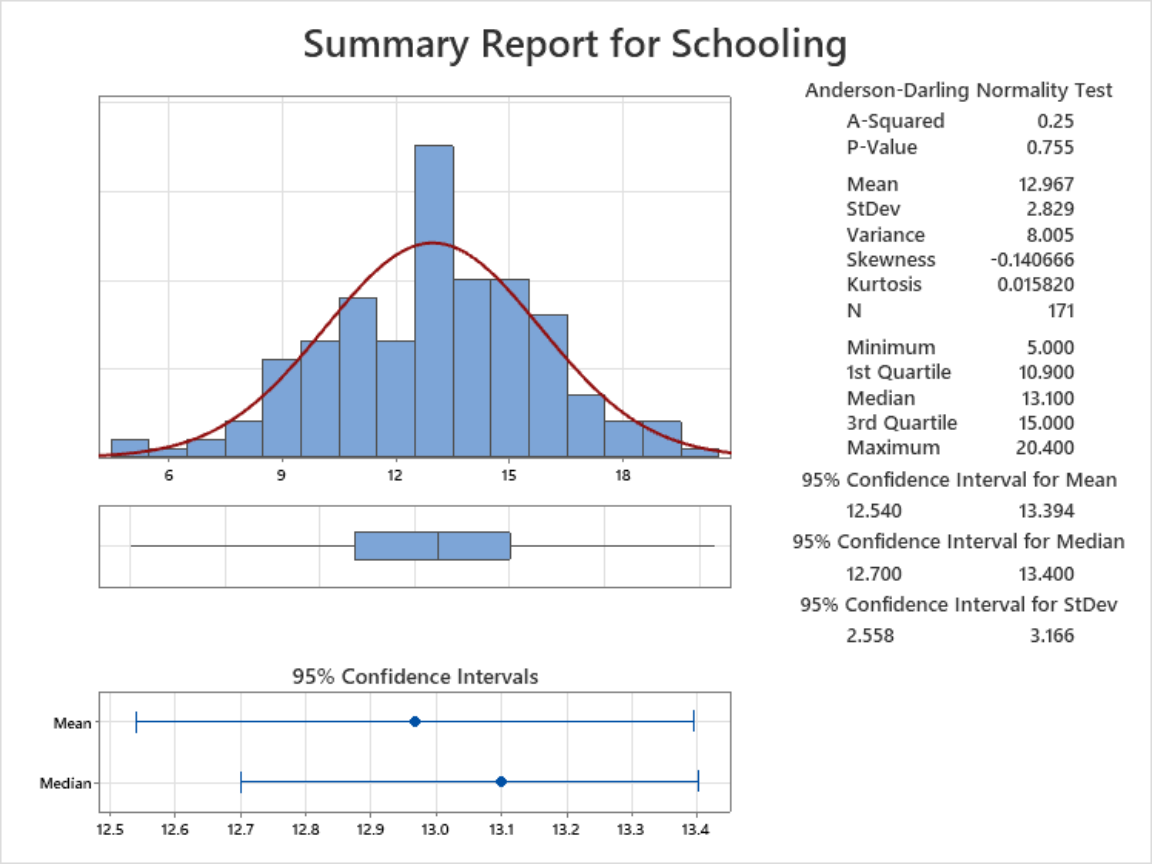
4) For each variable, obtain Minitab's Graphical Summary. For each variable, the graph gives, first, a histogram with a "normal curve" superimposed. The graph also gives a boxplot (on its side, corresponding to the X-axis of the histogram) as well as other numerical and graphical information. Note any observations that are outliers (or at least the two or three most extreme ones), according to the boxplots. Do these correspond to outliers you found in the scatterplots?



*We see one outlier for Life Expectancy (by the 1st quartile - 1.5*IQR definition), which is Sierra Leone. This does correspond to what we found on the scatterplot.*



According to the boxplot above, there are no outliers for BMI. However the sample does not look to be not normally distributed. The histogram suggests a bi-modal distribution.



According to the boxplot above, there are no outliers for Schooling. The histogram suggests that the sample is normally distributed.

5) Often, the variability of a quantity depends on its size. For example, the variation in the incomes of the top 10% of earners is much greater than in the bottom 10% of earners. If one of your variables suffers from this size-dependent variability:
(A) The histogram will show a right-skewed distribution,
(B) The mean will be larger than the median,
(C) The boxplots will show that the median line is towards the low side (in this case, left side) of the box.
(D) The boxplot will show more outliers on the high side than on the low side.
For each variable, based on the descriptive statistics output, decide if your response variable has the problem described above. If so, and if all of the data values for this variable are positive, try taking natural logs of the variable. If, for example, you want to create a variable, LogPrice, from the existing variable Price, type LogPrice in the box marked "Store result in

variable:", and type LN(Price) in the box marked "Expression:". Then create the descriptive statistics graph again for the log of the variable, and decide whether the problem is reduced, according to the criteria (A)-(D) above.

Please note that if a variable has any zero or negative values, then taking logs is NOT appropriate, so there is no point in trying it in this case. (Minitab will simply generate an error message).
The reason we worry so much about taking logs is that it often helps the subsequent statistical analysis. In particular, taking logs tends to bring the high outliers more in line with the rest of the data, while at the same time "blowing up" the picture at the low end, so that these points can now be seen more clearly.

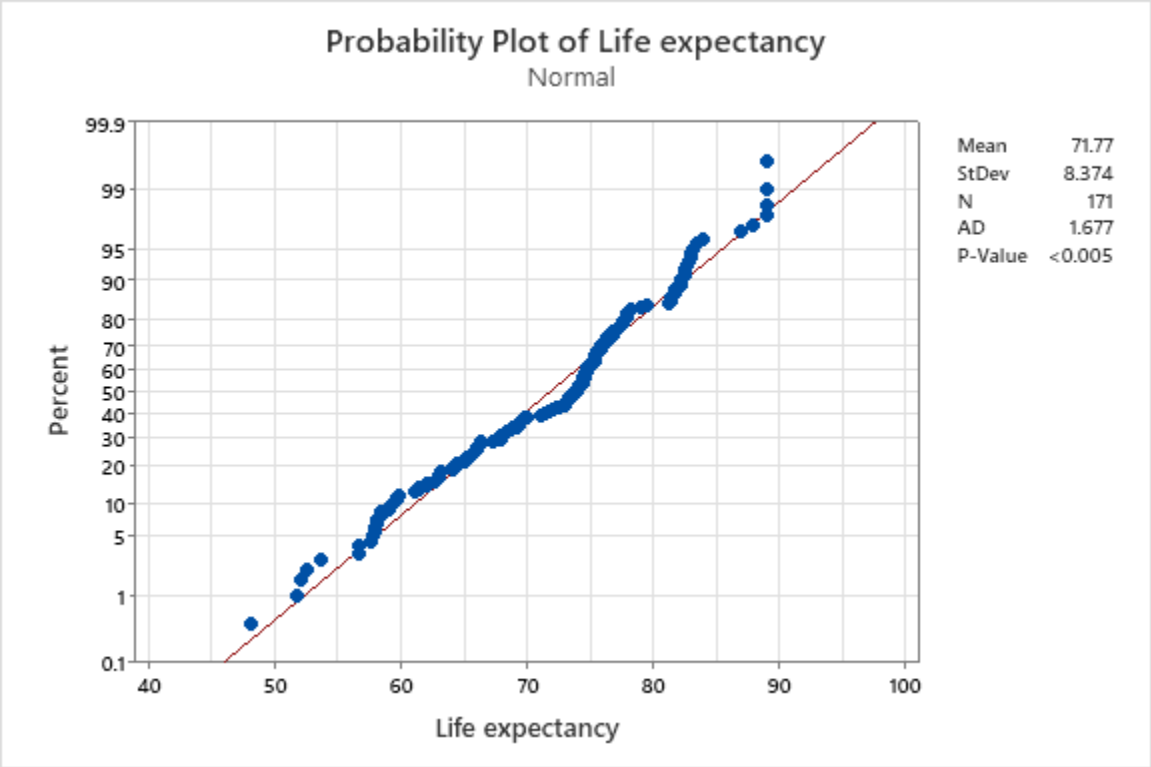
Our response variable (Life Expectancy) skews to the left, not to the right. There are a few countries that have very low Life Expectancy, which is skewing our distribution. Neither of our predictor variables – Schooling nor BMI – looks to be affected by size-dependent variability.

6) Rerun the scatterplots (and answer the rest of question 2) using the logged variables wherever this was found appropriate in question 5). Here are some examples of what I mean: If you decided to take logs of predictor variable X2 only, then you should run a scatterplot of your response variable (let's call it Y) against log(X2). If you decided to take logs of X2 and X3, then you should run scatterplots of Y versus log(X2) and Y versus log(X3). If you decided to take logs of Y only, then you should run scatterplots of log(Y) versus all of the (non-logged) predictor variables. If you decided not to take the log of any of the variables, you do not need to do anything. For each scatterplot you create here), compare it with the corresponding one from question 2). Did taking logs help you to uncover a relationship between the variables?

Based on our answer to Question 5, we do not find it appropriate to return the scatterplots using logged variables

7) For your response variable, use Minitab to create a Normal Probability Plot. In the “Variable” box, put your response variable. Make sure the box for “Anderson-Darling” is checked. You can type in a title for your plot. This plot gives the percentiles of a normal distribution vs. the percentiles of your data set. If the data set came from a normal distribution, the plot should produce a straight line. For guidance, Minitab draws in this ideal straight

line. Non-normal distributions will produce curvature in the plot. Describe the pattern you see in your plot. Does this pattern seem to indicate non-normality? If so, in what way? (For example, is the upper tail longer than that of a normal distribution? To diagnose this, see if the points seem to be bending downwards from the line on the right hand side of the plot.)



The normal probability plot indicates that our Life Expectancy data is not quite normally distributed. In particular, we see that the upper tail is shorter than that of a normal distribution – the points bend sharply upwards from the line on the right side of the plot. We also see a dip downwards from the line toward the middle of the data, whereas the lower tail curves upwards from the line.

8) On the upper right hand side of the normal probability plot, you will see a box containing five numbers. The fourth number ("AD") is the Anderson-Darling Statistic. The larger this number is, the stronger the evidence of non-normality. The fifth number ("*p*-value") gives the probability that we would get such a large value for AD if the distribution were actually normal. The closer this *p*-value is to zero, the stronger the

evidence of non-normality. Does this *p*-value seem to indicate non-normality in your data?

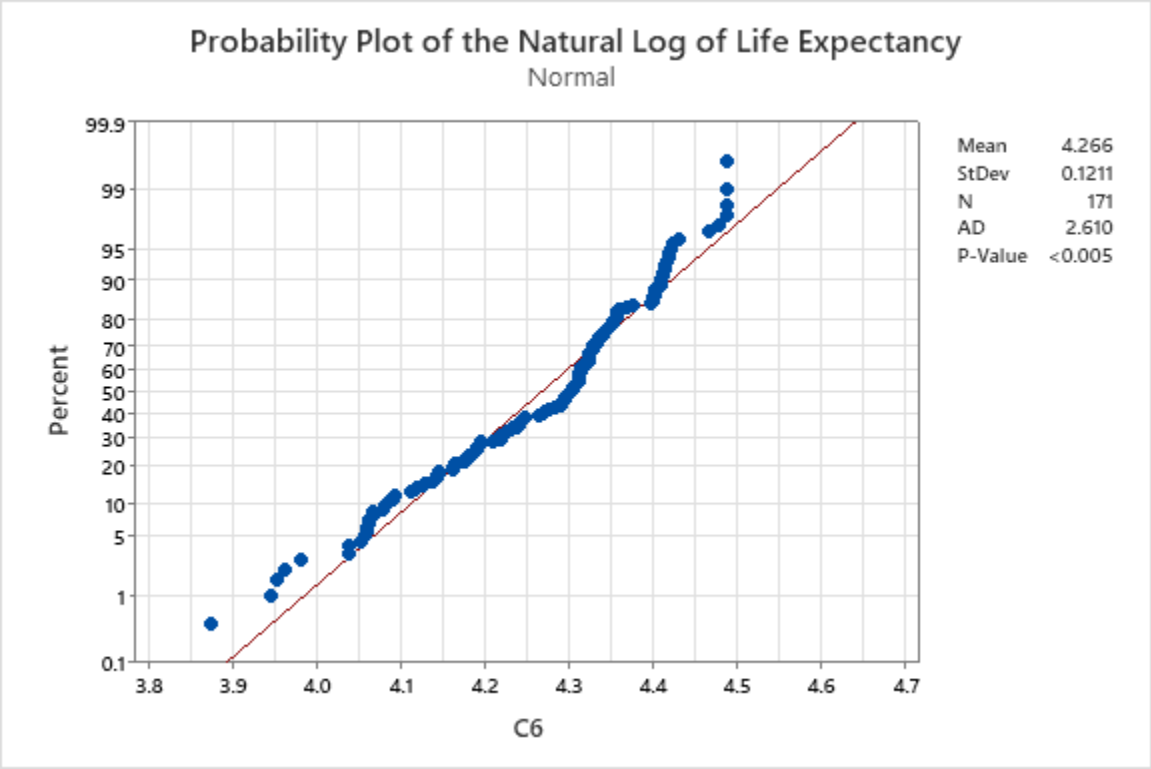
AD = 1.677; P-Value = <.005

Due to a small p-value less than .05, there is evidence of non-normality

9) Do your findings based on this plot agree with what you found based on the Descriptive Statistics plot in problem 4).

Yes, the histogram from the Descriptive Statistics plot does not show a normal distribution (instead it looks to be potentially bi-modal or left-skewed), so our findings on the Normal Probability Plot do agree with our Descriptive Statistics.

10) Repeat questions (7-9) for the logarithm of the response variable. If it's not possible to take logs of the response variable values, skip this question.



As we saw with our initial normal probability plot for Life Expectancy, the normal probability plot for log(Life Expectancy) indicates that the data is not quite normally distributed. The upper tail is even shorter than we saw in

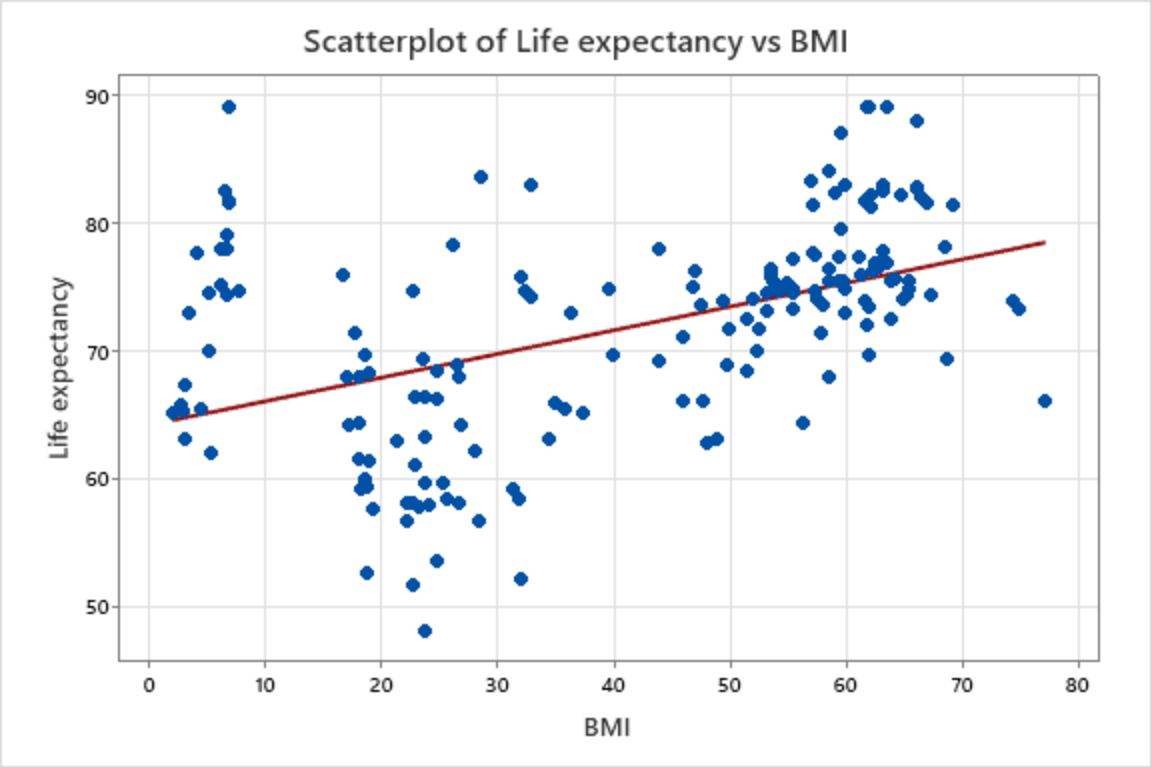
the previous plot – the points start curving upward from the line earlier. There is still a dip downward from the line toward the middle of the data, and the upward curve of the lower tail is even more pronounced than we previously saw.

AD = 2.610; P-Value = <.005

Due to a small p-value close to zero, there is evidence of non-normality.

Similar to Question 9, the histogram from the Descriptive Statistics plot does not indicate a normal distribution (it looks to be potentially bi-modal or left-skewed), so our findings on the Normal Probability Plot for the natural log of Life Expectancy also agree with our Descriptive Statistics.

11) Run simple regressions of your response variable (use either the actual response or the logged response depending upon your answers to previous parts of the project) against each of the individual explanatory variables (remember to use logged predictors or actual predictors depending upon your previous work). Interpret the slope coefficients. Determine the p-values for the slopes. (Remember to take account of the direction of the alternative (research) hypothesis you had thought of in Question 1). Are the slopes statistically significant?



Regression Equation

Life expectancy = 64.17 + 0.1851 BMI

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	64.17	1.24	51.92	0.000	
BMI	0.1851	0.0268	6.91	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7.41539	22.04%	21.58%	19.93%

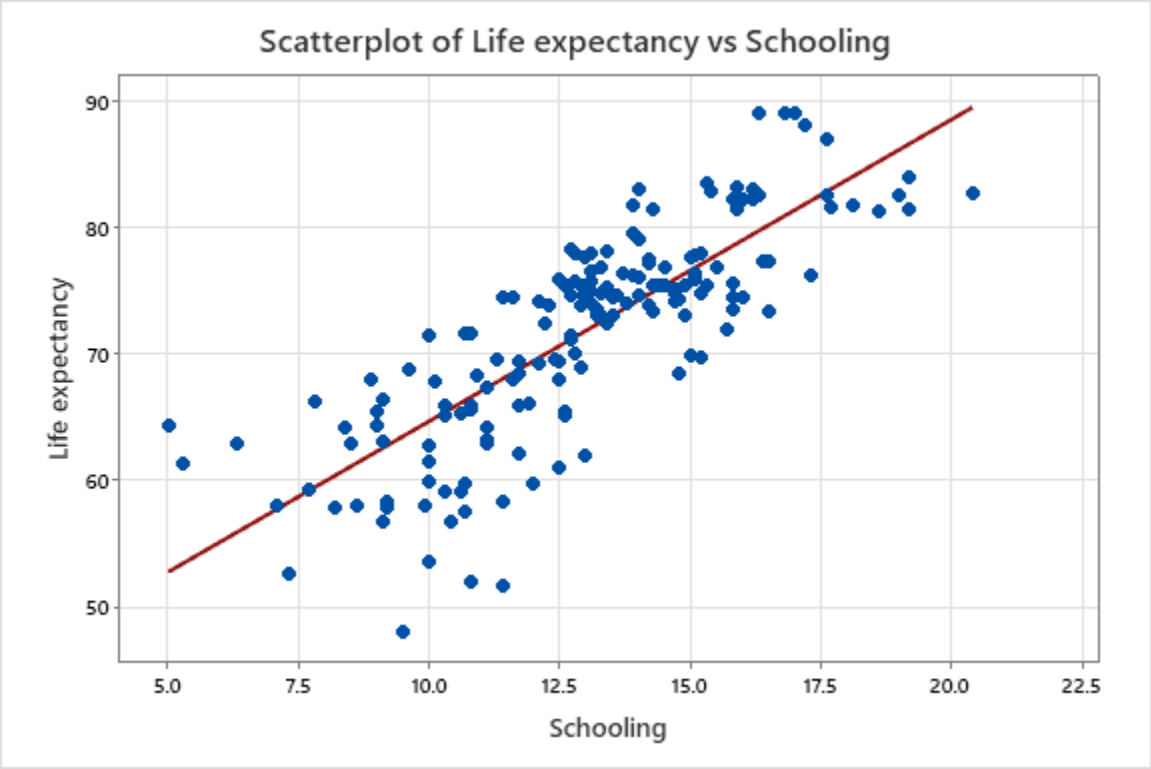
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	2627	2626.68	47.77	0.000
BMI	1	2627	2626.68	47.77	0.000
Error	169	9293	54.99		
Lack-of-Fit	134	7981	59.56	1.59	0.056
Pure Error	35	1312	37.48		
Total	170	11920			

The slope coefficient of BMI is 0.1851, which means as BMI increases by 1 unit, life expectancy increases by 0.1851 years.

P-value for BMI is ~0.000, which gives evidence the model is useful and BMI is a useful predictor of life expectancy. The slope is statistically significant since the p-value is ~0 and less than any alpha one could choose.

The R-squared value is relatively low at 22.04%, therefore the regression line doesn't show a strong "goodness of fit," implying that only 22.04% of the variability in life expectancy is explained by BMI.



Regression Equation

Life expectancy = 40.88 + 2.382 Schooling

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	40.88	1.79	22.80	0.000	
Schooling	2.382	0.135	17.63	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.98482	64.77%	64.56%	63.78%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	7720	7720.25	310.69	0.000
Schooling	1	7720	7720.25	310.69	0.000
Error	169	4199	24.85		
Lack-of-Fit	83	2459	29.63	1.46	0.040
Pure Error	86	1740	20.24		
Total	170	11920			

The slope coefficient for schooling is 2.382, which means as schooling increases by 1 year, life expectancy increases by 2.382 years.

The P-value for Schooling is ~0.000, which gives evidence the model is useful and Schooling is a useful predictor of life expectancy.

The R-squared value is higher at 64.77%, therefore the regression line shows a strong "goodness of fit," implying that 64.77% of the variability in life expectancy is explained by schooling.

12) Run a multiple regression, using all of your predictor variables (remember to select the logged response or not and remember to select the logged predictors are not). Are all of the coefficients significant? Which variables (if any) appear to be useless for predicting the response variable? Check the F-statistic. (Interpret it briefly). What is the value of the R^2? Is it appreciably higher than what you got in the simple regressions?

Regression Equation

Life expectancy = 41.12 + 2.226 Schooling + 0.0435 BMI

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	41.12	1.78	23.13	0.000	
Schooling	2.226	0.152	14.63	0.000	1.30
BMI	0.0435	0.0203	2.15	0.033	1.30

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.93245	65.71%	65.30%	64.39%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	7832.4	3916.18	160.97	0.000
Schooling	1	5205.7	5205.69	213.97	0.000
BMI	1	112.1	112.11	4.61	0.033
Error	168	4087.3	24.33		
Total	170	11919.6			

Fits and Diagnostics for Unusual Observations

Obs	Life expectancy	Fit	Resid	Std Resid	
4	51.700	67.480	-15.780	-3.22	R
50	64.400	53.030	11.370	2.37	R X
70	82.500	83.696	-1.196	-0.25	X
89	52.100	66.549	-14.449	-2.94	R
115	53.600	64.451	-10.851	-2.21	R
116	81.600	80.811	0.789	0.17	X
125	89.000	78.812	10.188	2.13	R X
138	48.100	63.299	-15.199	-3.10	R

R Large residual
X Unusual X

Checking if predictors are useful:

1. *Schooling: In the context of this model, p-value for schooling is close to 0. This implies that the p-value is less than any value of alpha we*

would normally use. This indicates this coefficient has a linear relationship and this variable is useful.

2. *BMI: In the context of this model, p-value is 0.033. Since we normally use an alpha of .05, the p-value indicates this coefficient has a linear relationship and this variable is useful.*

Because the p-value for BMI is 0.033, whether or not this variable would be considered useful would depend on what alpha one would choose. If one were to choose an alpha of .05, then BMI would be interpreted as a useful variable in the context of this regression. However, if one were to pick a lower alpha (e.g., .01), BMI would not be considered a useful variable in the context of this regression. That said, the coefficient for BMI isn't very large, so while it may be considered statistically significant, we're not sure how much substantive meaning it holds since a one unit increase in BMI is associated with only a 14 day increase in lifespan.

The F-statistic is 160.97. If we were to use an alpha level of .05, the F(alpha) would be between 3.04 and 3.09. Therefore, since 160.97 is greater than (3.04-3.09), we would reject the null in favor of the alternative. We have evidence that regression is useful.

The value of R^2 is 65.71%. This is much higher than the R^2 of the simple regression using BMI as the predictor variable and only slightly higher than the R^2 of the simple regression using Schooling as the predictor variable.

13) Do you find any apparent inconsistencies in the coefficients you get in the full multiple regression model, compared with the coefficients for the corresponding variable in the simple regression? Did the coefficient values change appreciably from the simple model to the full model? Please explain.

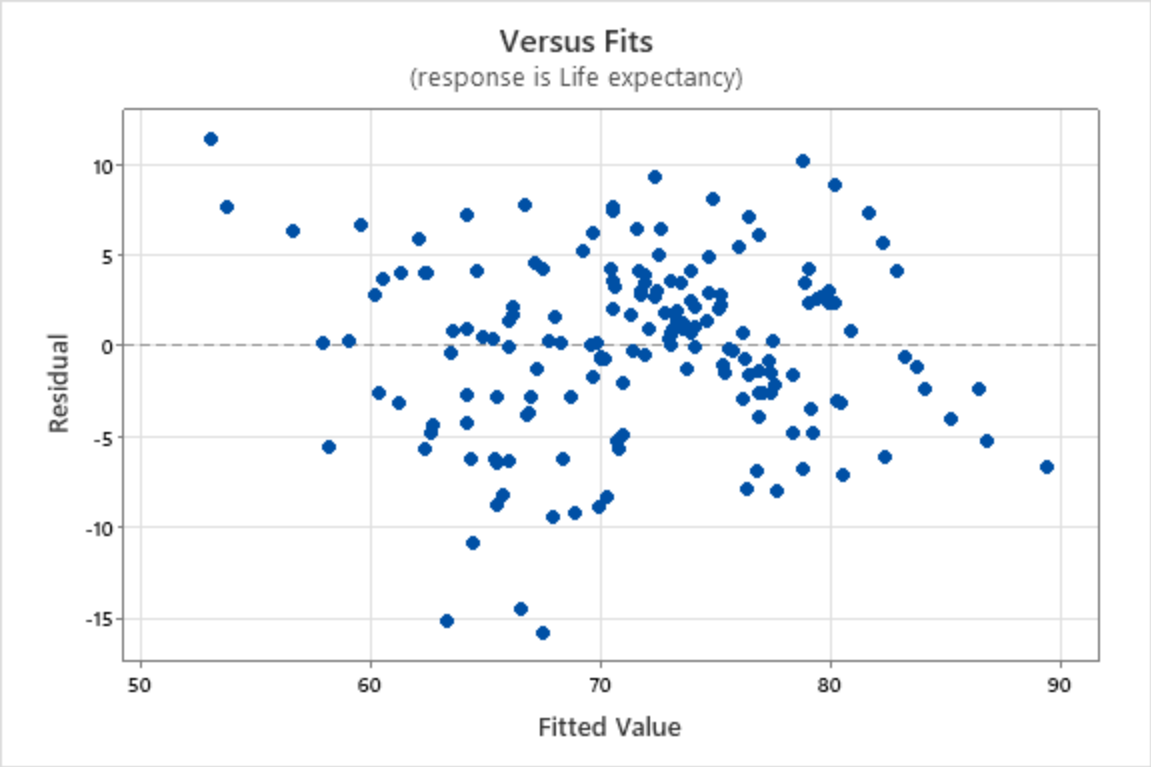
Label	***Simple Regression (Schooling)***	***Simple Regression (BMI)***	***Multiple Regression***
b0	*40.88*	*64.17*	*41.12*
Sb0	*1.79*	*1.24*	*1.78*
b1 (Schooling)	*2.382*	*N/A*	*2.226*
Sb1(Schooling)	*0.135*	*N/A*	*0.152*

b2 (BMI)	*N/A*	*0.1851*	*0.0435*
Sb2(BMI)	*N/A*	*0.0268*	*0.0203*

A slight increase in the coefficient of constant and predictor variables is observed between the simple and multiple regression. The main difference between the simple and multiple regression is that simple regression has only one independent variable whereas multiple regression has multiple independent variables. Each independent variable in multiple regression has its own coefficient to ensure that each variable is weighted appropriately.

It is also worthwhile to note that the coefficient for BMI changes almost 4-fold. Controlling for Schooling, the effect of BMI on Life Expectancy is lessened in the multiple regression model.

14) For the full multiple regression model, get Cook's D and leverage, as well as a standardized residual vs. fits plot. Briefly discuss the results. (In multiple regression, the leverage is large if it exceeds 3(k+1)/n, where k is the number of explanatory variables, and Cook's D is large if it exceeds 1). Identify any outliers, and discuss the meaning of the outliers, if possible. Do all of these outliers correspond to the ones found in the scatterplots and descriptive statistics graphs from questions 2-6? If not, discuss briefly. Overall, considering the R^2, the significance of the individual coefficients, and the Cook's D values, does the full model seem to fit well?



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	41.12	1.78	23.13	0.000	
BMI	0.0435	0.0203	2.15	0.033	1.30
Schooling	2.226	0.152	14.63	0.000	1.30

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.93245	65.71%	65.30%	64.39%

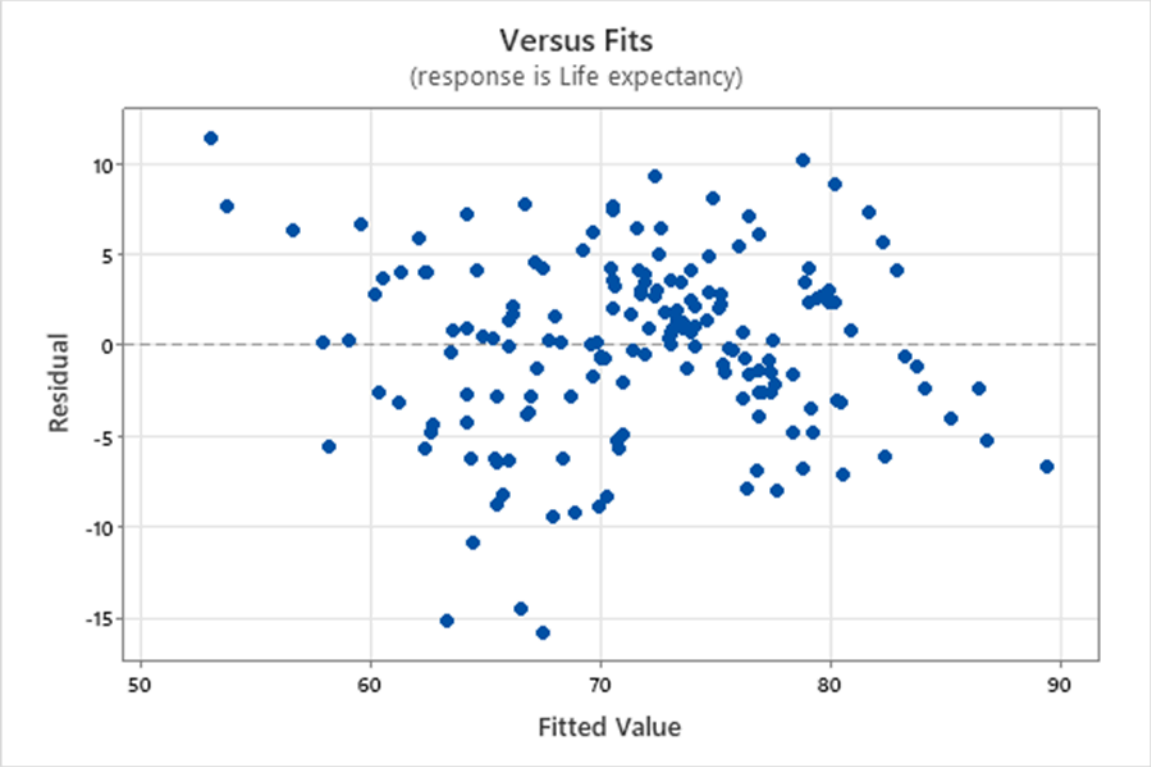
We have four leverage points using the formula 3(k+1)/n = .0526. However, we have no data points exceeding 1 for Cook's Distance, indicating that while we have four leverage points, none of them are bad leverage points.

We identified six outliers in our regression analysis. One of these outliers was the outlier we found earlier in Question 4 (Life Expectancy in Sierra Leone). The other five (Angola, Eritrea, Lesotho, Nigeria, Portugal) are not necessarily surprising as outliers in the regression. Angola, Lesotho and Nigeria are all toward the bottom of life expectancy, while Portugal is at the

top, indicating these are all data points toward the outside of the distribution of life expectancy. Eritrea has the lowest schooling index, which also likely contributed to it being an outlier in the regression.

With the value of R^2 65.71%, the p-values of the predictor variables, and the Cook's Distance values, the full model does seem to fit well. However, we did see evidence of a non-normal distribution earlier in the analysis.

15) Based on the standardized residuals vs. fits plot, is there evidence of non-constant variance? Based on your results on normality of the response variable from Questions 7)-9) and possibly 10), together with the evidence of the residual plots here, do you think that the output can be trusted (in the sense that the inferential statistics are accurate)?



Our standardized residuals vs. fits plot shows a violation of assumption two for constant variance. There is a gap in negative residuals around the ~75 life expectancy value, which indicates the output should not be fully trusted. There may be other variables that are impacting our output.

In combination with our findings in Questions 7-9 that the data we're working with is not normally distributed (a violation of assumption three),

we must surmise that the output cannot be trusted and the inferential statistics are not entirely accurate.

16) Finally, we are going to use two "automatic" methods for selecting the "best" predictor variables. For each of the models you have fitted in parts 1 and 2, you will use the residual sum of squares SSE to compute a number called AICC. The model with the smallest AICC is the "best". AICC is computed as AICC=LN(SSE)+2(k+2)/(n− k−3), where "LN" is the natural log (that is, "ln" on most calculators) and k is the number of predictor variables in the model. If any of the AICC values are negative, then the most negative value is the "best". Determine which of your possible models is "best" according to AICC. Are all of the coefficients in this model statistically significant? Interpret the coefficients of this "best" model, and say what it means in terms of the things you said you wanted to learn in questions 1 and 2. Please repeat this question regarding selecting the best model by utilizing the adjusted R^2 measure (the model with largest adjusted R^2 is best). Do your answers differ?

AICC=LN(SSE)+2(k+2)/(n− k−3)

Multiple Regression:	*Multiple Regression*	*Simple Regression Schooling*	*Simple Regression BMI*
SSE = sq(MSE)	*4087.3*	*4199*	*9293*
k	*2*	*1*	*1*
n	*171*	*171*	*171*
AICC	*8.364*	*8.379*	*9.173*
R-Sq (Adj)	*65.3%*	*64.6%*	*21.6%*

Based on AICC, our multiple regression model is the best. The best model to choose is the one with lowest value of AICC and highest adjusted R-Sq.

In the context of our regression model, the coefficient for Schooling is certainly statistically significant (it has a p-value of 0.00). The coefficient for BMI may or may not be interpreted as statistically significant, depending on what is chosen as alpha. If we choose an alpha of .05, we would interpret the

BMI coefficient as statistically significant, as its p-value is 0.033. However, if we were to choose an alpha of .01, we would determine the coefficient for BMI is not statistically significant.

In terms of what we wanted to learn in Questions 1 and 2, this tells us that as we add more predictor variables to predict life expectancy, the regression model becomes better. This also tells us that Schooling has a more positive impact and is a better predictor of life expectancy than BMI.

Since the adjusted R-Sq (Adj) value for simple regression for Schooling is 64.6 % compared to 21.6% using BMI. The simple linear regression using Schooling is better compared to model using BMI.

17) Provide an executive summary describing what you have learned and what importance (if any) your analysis provides.

- *In the context of our regressions, we found the two variables to be significant linear predictors of Life Expectancy, but there are likely more variables we would need to consider to more accurately predict Life Expectancy.*

- *We also looked at Life Expectancy for a specific year. There could have been specific events within that year that caused our data to be impacted. For example, the Ebola outbreak in Western Africa may have impacted our results, particularly for countries like Sierra Leone.*

- *We did find that the multiple regression was a better predictor of Life Expectancy than the single regressions, which makes sense as Life Expectancy is a complicated variable with (likely) many inputs that contribute to it (and likely many extraneous variables that impact it).*

- *This tells us that as countries strive to increase the Life Expectancy of their population, they need to think more broadly than Schooling and BMI. For example, Nutrition Level, Immunization Percentage, and Income may also be important predictor variables when it comes to Life Expectancy. Countries should look to not only encourage education and healthy living, but also provide resources to combat food insecurity, increase access to healthcare, and establish economic support systems for those living in poverty.*