Lab 2: Correlation and Regression
KEY

Javier E. Flores

January 30, 2019

Total Possible Points: 44

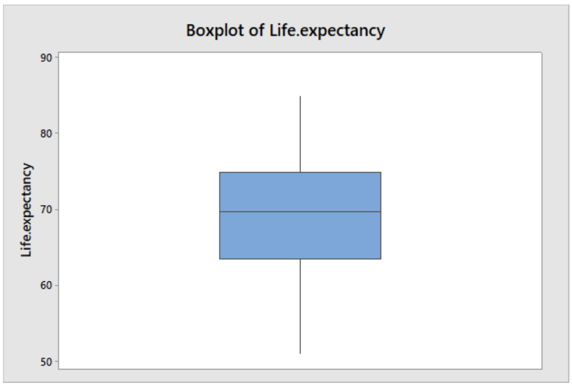
Exploratory Analysis

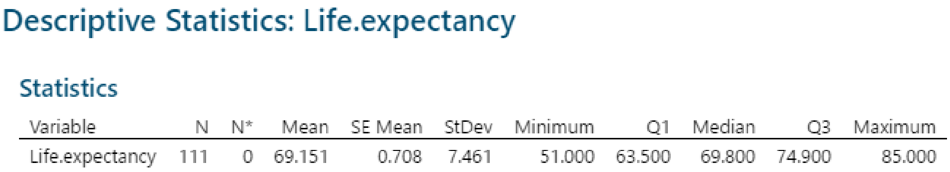
Q1) [1 pts] What population(s) can we generalize these data to?

A1) Several answers possible here. On example would be the that the sample may be generalized to developing countries of the world in 2015. [1 pts]

Q2) [3 pts] Construct a boxplot for the life expectancy variable and compute the IQR. Multiply 1.5∗IQR to assess the presence of outliers. Are there any? If so, do you think those data should be excluded? Include the boxplot and IQR in your write-up.

A2) Look for requested figure and IQR. [1 pts] Since 1.5∗IQR = 17.1, and there are no datapoints less than Q_1 - 1.5∗IQR or greater than Q_3 + 1.5∗IQR, there are no outliers. [2 pts]





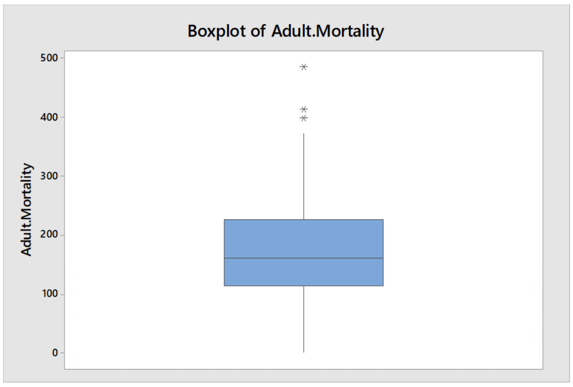
Descriptive Statistics: Life.expectancy

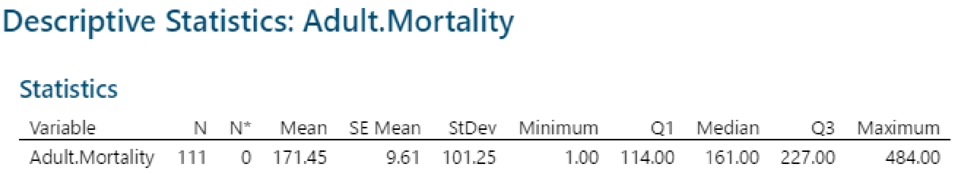
Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Life.expectancy	111	0	69.151	0.708	7.461	51.000	63.500	69.800	74.900	85.000

Q3) [3 pts] Construct a boxplot for adult mortality and compute the IQR. Multiply 1.5∗IQR to assess the presence of outliers. Are there any? If so, do you think those data should be excluded? Include the boxplot and IQR in your write-up

A3) Look for requested figure and IQR. [1 pts] Since 1.5∗IQR = 169.5, and there are three datapoints greater than Q_3 + 1.5∗IQR = 396.5, there are three outliers. These outliers should not be excluded unless there is reason to believe that they are typos. Later on we will discuss other instances when we may want to remove outliers.[2 pts]





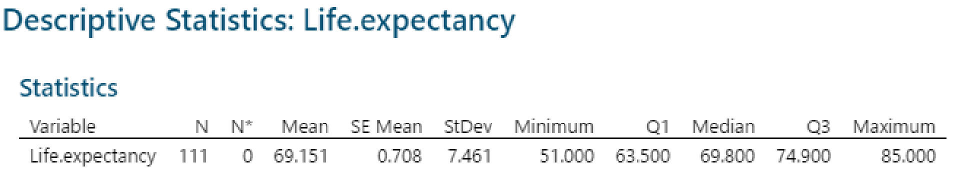
Descriptive Statistics: Adult.Mortality

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Adult.Mortality	111	0	171.45	9.61	101.25	1.00	114.00	161.00	227.00	484.00

Q4) [3 pts] Report the mean, standard deviation, and five number summary of the variable ”Life Expectancy”. Construct a histogram for this variable as well. How would you describe the shape of the distribution? Include the histogram in your write-up.

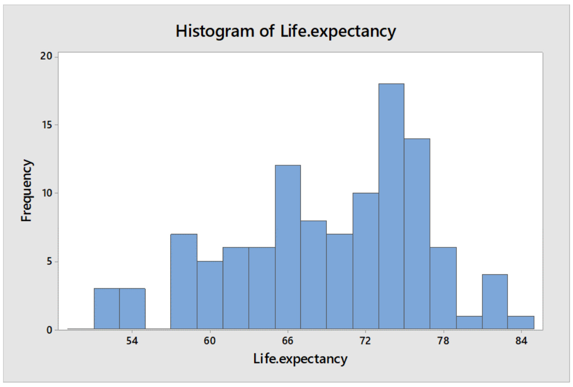
> **A4)** Check for the five number summary. [1 pts] Check for the histogram and description of shape (slightly left-skewed). [2 pts]



Descriptive Statistics: Life.expectancy

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Life.expectancy	111	0	69.151	0.708	7.461	51.000	63.500	69.800	74.900	85.000



Q5) [4 pts] Explore three additional variables using appropriate plots and/or summary statistics. In your write-up, include the most interesting of these explorations along with the results and a brief explanation of what you learned.

> **A5)** Check for plots/statistics for one additional variable. [2 pts] Check for reasonable conclusion drawn from results. [2 pts]

Q6) [1 pts] Based on your personal knowledge, identify five explanatory variables that you believe are most associated with life expectancy. List your choices in your write-up.

A6) Check for list of five variables. [1 pts]

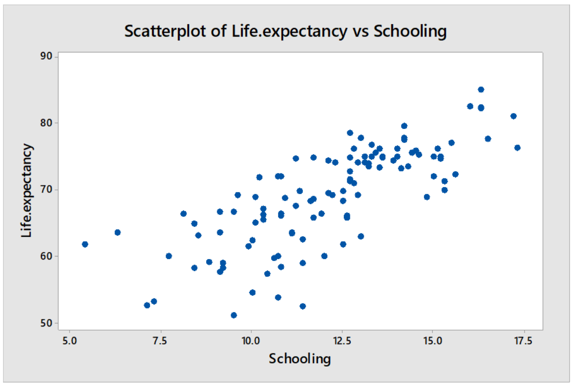
Q7) [3 pts] Explore how your choice of variables relate to one another and to the response variable, life expectancy. Include the three most interesting statistics (either visual descriptions or numerical summaries) along with a few sentences for each describing what was learned.

A7) Check for requested graphs/statistics/explanations. [3 pts]

Data Analysis

Q8) [2 pts] Investigate the first question, ”How are select health indicators related to life expectancy?”, by numerically and visually summarizing the strongest relationship between your choices of variables and the outcome. Briefly describe how you determined which association was strongest. Include only the numerical and visual summaries corresponding to the strongest association identified.

A8) For the visual summary, look for a scatterplot. For numerical summary, look for the correlation coefficient. The determination of strongest association should be based on the appropriate use of the correlation coefficient (i.e. relationship must be linear). [2 pts] An example is provided below ($r = 0.768$).



Q9) [2 pts] Using the correlation coefficient, predict the life expectancy of a country that is 1.5 standard deviations below average for the explanatory variable. Show all of your work.

A9) Check that the proper formula was used. [2 pts] Using the variables from the example above, we have

$$0.768 * (-1.5 * s_y) + \bar{y} = 0.768 * (-1.5 * 7.461) + 69.151 = 60.556$$

Q10) [2 pts] Do you expect the correlation coefficient and regression slope to be the same? Why or why not?

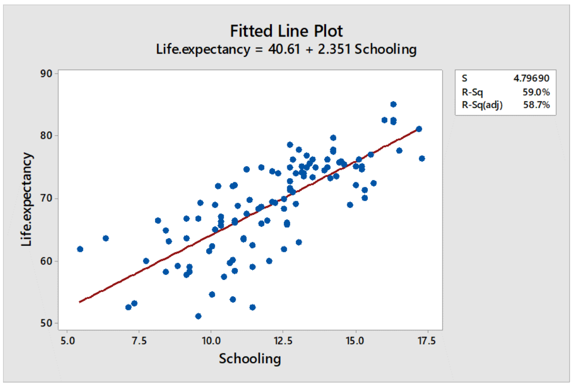
A10) If the explanatory and response variables have the same standard deviation, then yes. Otherwise, no. [2 pts]

Q11) [2 pts] Perform a regression in Minitab using your explanatory and response variables. Take the square root of the obtained R^2. What does this quantity tell you? Have you seen this number at any point thus far?

A11) The quantity is the same as the correlation computed earlier (and should be interpreted as such). [2 pts]

Q12) [2 pts] Using the fitted regression line from the previous question, predict the life expectancy of a country that is 1.5 standard deviations below average for the explanatory variable. Show all of your work and include the regression line plot in your write-up.

A12) The value 1.5 should be unstandardized and plugged in to the regression equation. See example below. [2 pts]



Unstandardize: $-1.5 * 2.438 + 12.137 = 8.48$

Plug in: $40.61 + 2.351 * 8.48 = 60.546$

Q13) [3 pts] Non-statistically speaking, and considering the explanatory variable you've identified, how would you determine which country is most in need?

A13) Lots of answers possible, give credit to any reasonable explanation. [3 pts]

Q14) [3 pts] Using **residuals** and your constructed regression line, explain how would you determine which country is most in need.

A14) The residuals may be used to identify those countries who are performing much below what is predicted given their level of explanatory variable. This would indicate a greater need for intervention (in the example case an intervention directed towards schooling). [3 pts]

Q15) [2 pts] Using a Minitab formula, create a new variable called ”le.residuals” that contains the residual of each predicted life expectancy. Include a screenshot of your formula in your write-up.

A15) Check for requested screenshot. [2 pts]

Q16) [3 pts] Using this new variable, find three countries most in need of intervention. Include the predicted life expectancies, actual life expectancies, and explanations of why you identified these countries as most in need in your write-up.

A16) Check for requested information. The provided answers should correspond to those with the largest negative residuals. [3 pts]

Data Analysis, Part 2

Q17) [5 pts] Formulate a separate research question (using your choice of response variable) that may be answered using these data. Be sure to clearly state your research question, include a brief description (1-2 sentences) of the statistical approach used, and a brief description (1-2 sentences) of your obtained results. Include any relevant summary statistics or graphs.

A17) Check for all of question criteria. [5 pts]