Lab 2: Correlation and Regression

Javier E. Flores

February 8, 2019

Introduction

In today's lab, we will apply the correlation and regression methods discussed in class to data collected from the World Health Organization (WHO). A cleaned version of these data may be accessed on the course website.

As you work through the lab with your group, you will be asked to answer several questions. Please submit your responses (as a group) in a single, separate document. Include the original questions as well as your group's response in the final submission.

Data

These data are pulled from the World Health Organization's (WHO) Global Health Observatory (GHO) data repository. While this repository offers a vast amount of data, we will be exploring a dataset containing a few health indicators for over a hundred different developing countries in the year 2015. The data contain the following variables:

- **Country**: Country
- **Life Expectancy**: Life Expectancy in age
- **Adult Mortality**: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- **Infant Deaths**: Number of Infant Deaths per 1000 population
- **Hepatitis B**: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- **Measles**: Measles - number of reported cases per 1000 population
- **BMI**: Average Body Mass Index of entire population
- **Under-five Deaths**: Number of under-five deaths per 1000 population
- **Polio**: Polio (Pol3) immunization coverage among 1-year-olds (%)
- **Diptheria**: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- **HIV/AIDS**: Deaths per 1 000 live births HIV/AIDS (0-4 years)
- **GDP**: Gross Domestic Product per capita (in USD)
- **Population**: Population of the country
- **Thinness 5-9 years**: Prevalence of thinness among children and adolescents for Age 5 to 9 (%)
- **Thinness 10-19 years**: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)

- **Income Composition of Resources**: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- **Schooling**: Number of years of Schooling (years)

Today, our investigation will center around life expectancy. Specifically, we'll consider the following questions:

1) How are different health indicators related to life expectancy?
2) Can we use any health indicators to predict a country's life expectancy?
3) Suppose we had the ability to implement some form of intervention with promise to improve life expectancy. Which country would be in greatest need of this intervention?

Exploratory Analysis

By the end of our first lab, we were equipped with several techniques for data visualization and exploration. These tools are often the first deployed in any statistical analysis since they help orient statisticians to the data and variables it contains. As these methods are applied, the following questions are often considered:

- How were these data collected? What population(s) might they describe?
- What variable types do these data contain?
- Are there any **outliers**, or anomalous observations, in these data?
- Are there patterns of association between variables?

Use the data to address the following questions:

Q1) What population(s) can we generalize these data to?

Q2) Construct a boxplot for the life expectancy variable and compute the IQR. Multiply 1.5*IQR and add (subtract) to (from) Q_3 (Q_1) 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.

Q3) Construct a boxplot for adult mortality and compute the IQR. Multiply 1.5*IQR and add (subtract) to (from) Q_3 (Q_1) 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

Q4) 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.

Q5) 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.

Oftentimes, we aren't able to perform an exhaustive search of all potential variable relationships within our data. To emphasize this point, if we were interested in characterizing all two way associations in our current data, we would have to consider 120 different variable combinations! Even if we had the time to perform all these analyses, it is generally ill-advised. With an increasing amount of associations assessed, you are more likely to observe a relationship by chance alone. This process of dredging through data (exhaustively) in order to find an association is often referred to as "p-hacking".

To avoid this bad practice, and to save hours of time, statisticians investigate associations that they believe *a priori* to be true. These prior beliefs are usually informed by personal knowledge or through consultation with experts in the appropriate field.

Q6) 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.

Q7) 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.

Data Analysis

So far, we've investigated associations between a handful of variables that we've hypothesized as being related to life expectancy. We've also assessed the presence of outliers in a few variables and considered populations our sample may generalize to. Having done these things, we should now have a solid understanding of the contents of our data. Next, we'll work to address each of our research questions.

Q8) 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.

Our second research question is one of prediction. In class, we learned that prediction could be accomplished using correlation or regression. For subsequent analyses, consider only the response variable and most strongly associated explanatory variable.

Q9) 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.

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

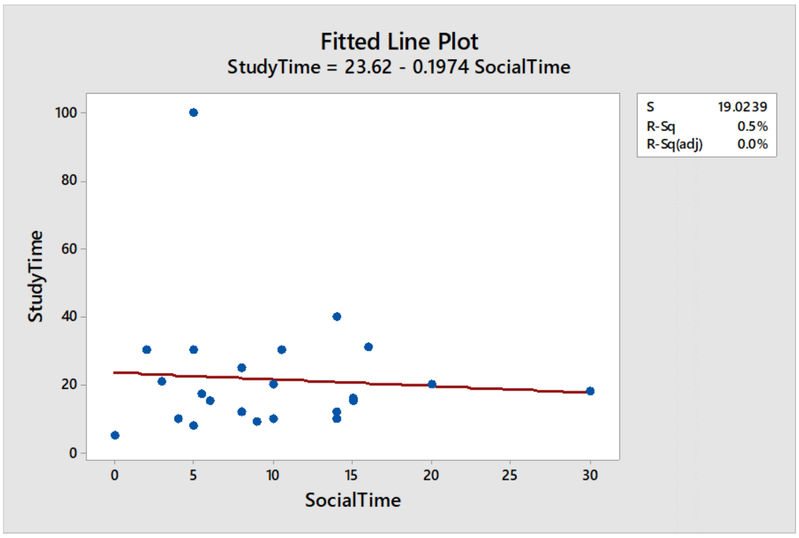
In Minitab, there are several ways that you can perform a regression analysis. For this lab, fit a regression line by following these steps:

1) Navigate to "Stat" -> "Regression" -> "Fitted Line Plot"

2) Select your response (Y) and explanatory (X) variables.

3) Select "Linear" for the type of regression model and click "Ok".

An example of the resulting output is shown below:



Notice that the equation for the regression line is displayed at the top of the plot. From this equation we can obtain the slope and intercept. In addition to this information, the plot also provides S, R^2, and the adjusted R^2.

- S measures the average prediction error of the fitted regression line.
- R^2, the **coefficient of determination**, describes the proportion of variability in the outcome variable that is explained by the explanatory variable. In other words, this describes how much of the changes observed in our response are explained by changes in our predictor.

$$R^2 = \frac{\text{Total squared deviation of the predictions from the mean}}{\text{Total squared deviation of the observed data from the mean}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Adjusted R^2 will be discussed later in the course.

Q11) 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?

Q12) 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.

For the final research question, consider the following:

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

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

Q15) 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.

Q16) 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.

Data Analysis, Part 2

Q17) 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.

Challenge (Optional)

Using R, repeat the analyses done for Q17. Include all relevant figures and statistics. It is not necessary for you to repeat the explanations/descriptions provided for Q17. In your attempt to complete this challenge, try and replicate the Minitab output as much as possible in R. For example, should you fit a regression line, be sure to include a plot similar to what is output by Minitab (shown on p.3) for the equivalent analysis. More points will be awarded the closer your figures are in appearance to the Minitab output. You may work on this challenge as a group, but I expect individuals to turn in separate copies of their own code and generated statistics/figures.

Turning in this challenge may prompt a brief assessment in which you will be asked to state the code used to generate any one of the results produced.

This challenge may be turned in and redeemed for credit at any point prior to the first exam, which will tentatively be scheduled one week after finishing chapter 3.