Summarizing Data: Relationships Between Variables

Relationships Between Two Quantitative Variables

In previous sections, we have discussed the relationship between different variables when at least one of them was categorical. Now let us consider the relationship between two quantitative variables using the life expectancy data from previous lectures.

We have already considered life expectancy in different countries, and we have already considered in which regions these different countries and territories are. Now, we are going to consider two more variables:

- GDP per capita = average wealth of the country (per $2,000 USD)
- HIV infection rate = percentage of adults infected with HIV

It is natural to think that the higher a country's GDP per capita, the higher its citizens' life expectancy. But how can we quantitatively examine this relationship?

Life expectancy and GDP per capita are both quantitative variables. One way to graphically display the relationship between two quantitative variables is through a **scatterplot**. In the scatter plot shown in Figure 1, a point along the horizontal axis represents a country's GDP per capita and a point along the vertical axis represents its life expectancy.

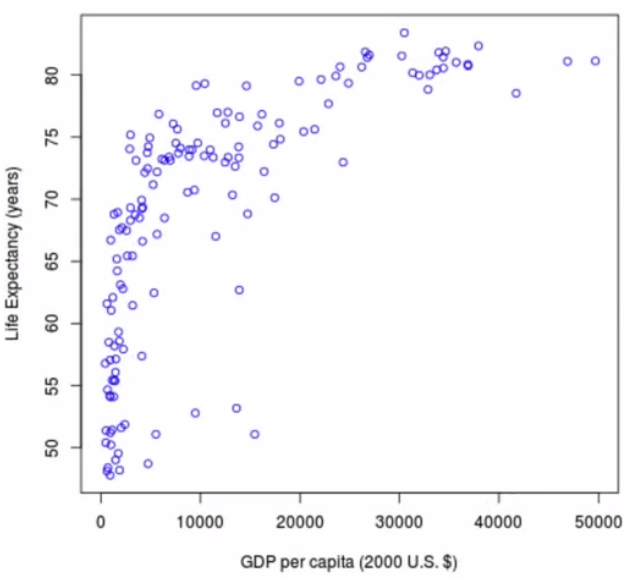

Figure 1: Scatterplot of life expectancy and GDP per capita

We can already see some obvious patterns: it certainly appears that countries with a larger

GDP per capita also tend to have a longer life expectancy.

We will learn later how to fit a line of best fit through a plot like this. If we put such a line of best fit like so in Figure 2, we can see that it is moving upwards to the right, which is an illustration of the *positive relationship* between these two variables. The higher the GDP per capita, the higher the life expectancy.

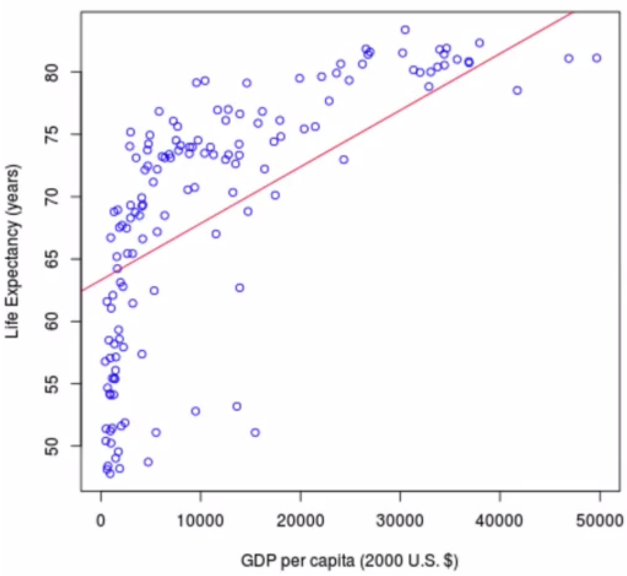

Figure 2: Scatterplot with line of best fit of life expectancy and GDP per capita

We can also compute the **correlation** between two quantitative variables. Denoted by the Greek letter ρ ("rho"), correlation measures the strength of the linear relationship between them. For the ith country, let x_i denote its GDP per capita and let y_i denote its life expectancy. The correlation between two quantitative variables x and y for n countries can be calculated by

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Observe that our choice of denoting GDP per capita by x and life expectancy by y was completely arbitrary and does not change the calculation had we reversed the variable names.

One way to gain some intuition about correlation is if the variables y_i's were actually the same as the x_i's. That is, suppose these two different quantitative variables were actually the same or if they just differed by some sort of a positive linear relationship. Then the correlation would be equal to $+1$, i.e., **perfect positive correlation** between these two variables. On the other hand, if the y_i's were the negative of the x_i's, or more generally had a negative linear relationship, then we could work out that this correlation would actually be equal to -1, i.e., **perfect negative correlation**. All the other correlations would be somewhere between -1 and $+1$. A correlation of $0 < \rho < 1$ illustrates the extent to which the variables increase together, while a correlation of $-1 < \rho < 0$ illustrates how much one variable increases when the other decreases. Of course, a correlation of 0 indicates no linear relationship exists between the x_i's and y_i's.

To summarize,

- $-1 \leq \rho \leq +1$
- if ρ is positive, they tend to both increase together
- if ρ is negative, one tends to increase when the other decreases

We should also remember that correlation, like lines of best fit, only captures the *linear* aspect of the relationship between two variables. In some of the examples we're considering here, e.g., Figure **??**, there are also non-linear effects which are not captured by the correlation or the line of best fit.

Example 1

Let's consider the HIV infection rate of the different countries and territories around the world. In this case we might expect that the higher the HIV rate, the lower the life expectancy. Once again, we have two quantitative variables: life expectancy and the percentage of HIV infections.

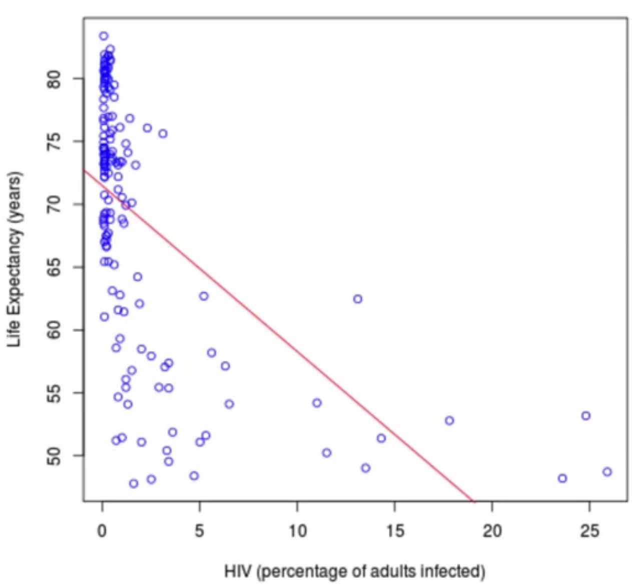

Figure 3: Scatterplot with line of best fit of Life Expectancy and HIV rate

The scatterplot is shown in Figure 3. The horizontal axis corresponds to the percentage of HIV infections of adults in each of the countries and territories, while the vertical axis corresponds to the life expectancy of that same country or territory. Once again, we can see that there is some sort of a relationship. In this case, we can see that countries which have the highest percentage of HIV infections also tend to have the lowest life expectancies. This suggests a negative relationship between HIV infections and life expectancy. We can also compute the correlation coefficient to obtain -0.566. This illustrates that when a country has more HIV infections as a percentage of adults, it probably tends to have a lower life expectancy too.

We will return to issues of relationships between variables throughout the rest of the course, but for now, we can see that using such concepts as scatter plots and correlation, we have our first understanding of the relationship between two quantitative variables.