Chapter 14: Describing Relationships between Variables

Overview: In Chapters 10, 11, 12 and 13, we saw how to visually display information about a single variable (using pie charts, bar graphs, histograms etc...), how to describe the distribution of a single variable numerically (using median, mean, quartiles and standard deviation) and we saw a special distribution called the Normal Distribution and looked at the properties of variables with this distribution.

In this chapter, we will now see how to investigate relationships between 2 variables. We will learn how to visualize this relationship, how to describe the relationship in words, and how to quantify the relationship numerically.

Motivating Example: Suppose we had a data set consisting of the following information: Country, Year, Continent, Life Expectancy, Population, and GDP per capita. Perhaps we expect there to be a relationship between some of these variables, for example, the life expectancy and the GDP could be related, or the year and the population could be related. In this Chapter, we will see how to explore these possible relationships.

Data Set: In this section, we will be working with the *gapminder* data set which is not built into R but can be installed in R. To access the *gapminder* data set, do one of the following:

- Type in the command install.packages('gapminder') and run this line of code. Then type in library(gapminder) to load the package into your R session. You should now be able to access the gapminder dataset.

- Or, if you do not have a version of R that supports this, download the *gapminder.csv* dataset that I have posted in Brightspace and read it into R, saving it as *gapminder*.

Visualizing the Relationship between 2 Variables: The first step to determining whether or not 2 variables are related is to plot them against one another (that is, plot one variable on the x-axis and one variable on the y-axis). The best type of plot to use is called a **scatterplot**.

Question: How do you decide which variable should be plotted on the x-axis and which variable should be plotted on the y-axis?

Answer: If you suspect on variable helps to explain the other variable, then you should put the first variable on the x-axis.

We often refer the variable on the x-axis independent variable/ explanatory variable.
The variable on the y-axis, we call it the dependent variable/ response.

y= f(x)

Caution: Just because you suspect that one variable may depend on the other (or one variable may be an explanatory variable for the other) does not mean that this relationship actually exists. You set up your scatterplot based on the relationship you hypothesize might exist but the scatterplot is just one step in trying to explore the true relationship between the variables.

How to Create a Scatterplot in R: We use the plot() function.

```
plot(x=xdata, y=ydata, main= "title", xlab= "x-axis title", xlim= c(0.50), ...)
```

Example: Create a scatterplot comparing the GDP per capita and the Life Expectancy.

We supose that GDP helps explain life expectancy so we plot GDP on the x-axis and life expectancy on the y-axis.

```
plot(x=gapminder$gdpPercap, y=gapminder$lifeExp, main="GDP vs Life Exp",
xlab="GDP", ylab="Life Exp")
```

Note 1: The scale of the x-axis is in scientific notation, if you would like to turn off scientific notation for the session the you can start your session with the code:

```
options(scipen=999)
```

Note 2: The other thing we notice is that there seem to be a lot of points crowded together, which makes the scatterplot appear messy and difficult to interpret. We should consider what data is being plotted and see if there is anything that should be changed.

Question: What are some possible issues with the scatterplot that we just made?

Answer: The plot is very crowded

-We are plotting gap and life expectancy for each country and each year. So there are several points plotted for each country.

- Perhaps, we should focus on a specific year or a specify country.
- Everything is the same colour which can make ir dificult to distinguish points.

Example: Create a scatterplot comparing the GDP per capita and the Life Expectancy in the year 1977.

Step 1: We first need to filter the data so that we have only the gdp per capita and life expectancy data for the year 1977. We can do this in the following way:

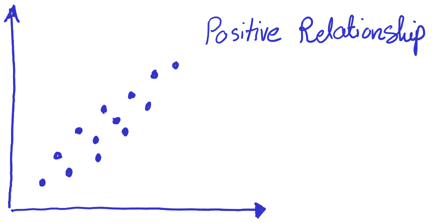
```
rows_1977 = which(gapminder$year==1977)

gdp_1977 = gapminder$gdpPercap[rows_1977]

life_exp_1977= gapminder$lifeExp[rows_1977]
```

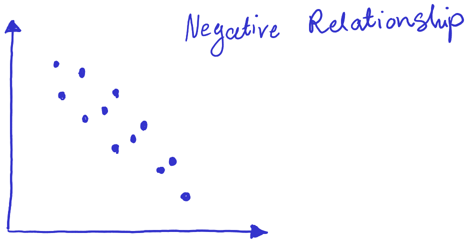
Step 2: Now plot the filtered data in a scatterplot with updated titles to reflect the change.

```
plot (x=gdp_1977, y=life_exp_1977, main="GDP vs Life Exp. in 1977", xlab="GDP",
 ylab="Life Exp. (in yrs)")
```

Words to Describe the Relationship seen in a Scatterplot: There are 3 traits that we look for when visualizing the relationship between two variables; direction, form and strength of the relationship.

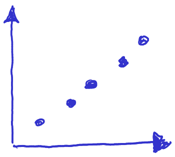
- The **direction** of a relationship describes whether or not the overall trend of the relationship is positive (increasing in one variable leads to increasing in the other variable) or negative (increasing in one variable leads to decreasing in the other variable).



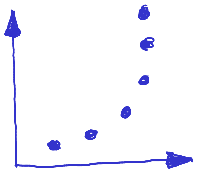


- The **form** of a relationship describes the type of function that would have a similar shape, common forms are:

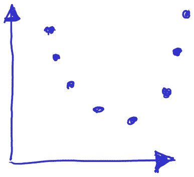
linear



Exponential



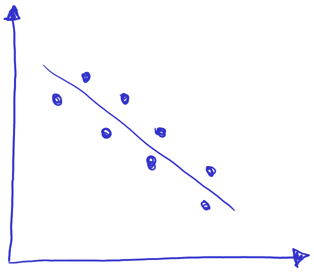
Quadratic



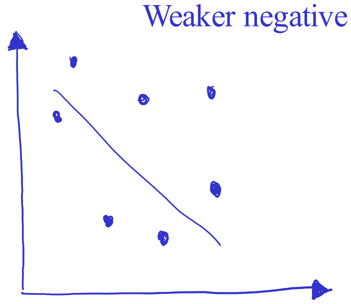
- The **strength** of the relationship describes

How clear the direction and form of the relationship are displayed with data.

Stronger negative linear relationship



Weaker negative linear relationship

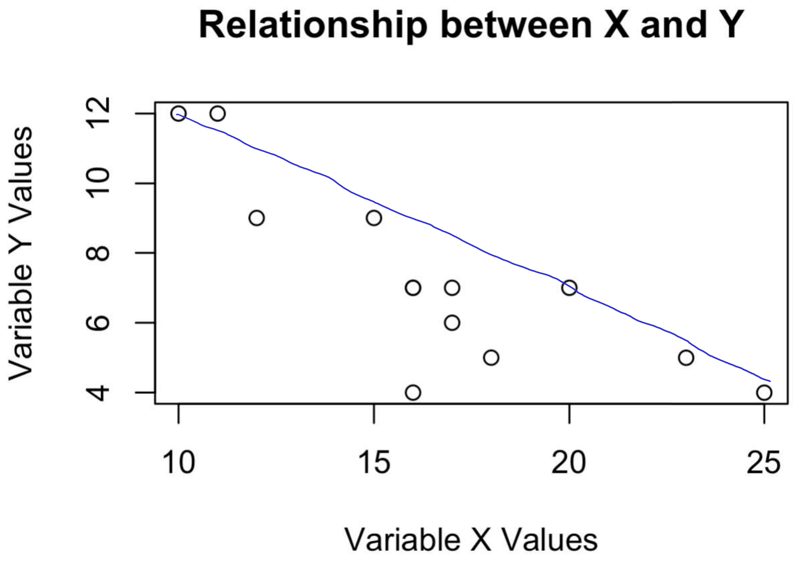


Note: It is really difficult to measure strength simply by looking at a scatterplot. Later in this chapter we will see a numerical way to measure the strength of a linear relationship.

Question: Describe the direction, form and your best judge of strength of the relationship between the average GDP of a country and its life expectancy in 1977.

Direction:Positive
Form: Exponential or logarithmic
Strength: Strong

Practice Question: Consider the scatterplot of variables X and Y below.



(a) Determine the direction of the relationship.

(A) Positive [✓ (B) Negative]

(b) Determine the form of the relationship.

(A) Exponential (B) Linear (C) ~~Quadratic~~ (D) Logarithmic

(C) To the best of your ability, describe the strength of the relationship.

(A) ~~Weak~~ (B) Moderate (C) Strong (D) ~~Very Strong~~

Measuring the Strength of a Linear Relationship: If we have two variables which we suspect have a linear relationship (either positive linear or negative linear) then we can measure the strength of the linear relationship by calculating something called **correlation**.

Definition: The **correlation** describes the direction and strength of a linear relationship. It canot be used for non-linear relationships, except to show the relationship is not linear.
Nototation is r

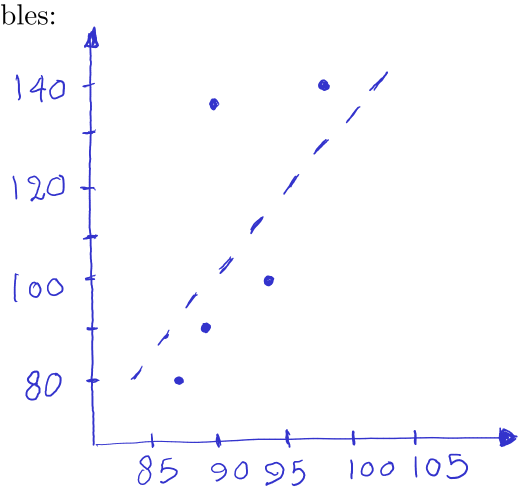
Calculating r by hand: The formula for the correlation r between variables X and Y is

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_x} \frac{(y_i - \bar{y})}{S_y}$$

Where $\bar{x}$ and $\bar{y}$ are sample mean for x and y.
S_x and S_y are the standard deviation.

Example: Suppose we had the following observations of variables A and B, $A = 100, 90, 95, 92, 88$ and $B = 140, 90, 100, 135, 80$ and we assumed that B was the dependent variable.

Plotting the values in a scatterplot, we see that there appears to be a positive linear relationship between the variables:



Suppose we were given (or used R to compute) the sample means and sample standard deviations for A and B. We could use these values to calculate the correlation between A and B:

	Sample Mean	Sample Std. Dev.
A	93	4.69
B	109	27.02

$$r = \frac{1}{5-1}\sum_{i=1}^{5} \frac{(A_i - 93)}{4.69} \frac{(B_i - 109)}{27.02} = \frac{1}{4}\left(\frac{1}{4.69}\right)\left(\frac{1}{27.02}\right)\Big[(100-93)(140-109) +$$

$$(90-93)(90-109) + (95-93)(135-109) + (88-93)(80-109)\Big]$$

$$= 0.7398$$

Interpreting the Correlation r:

- r will always be a value between -1 and 1.

- If r is positive then, that means the direction of the relationship is positive

- If r is negative then, the direction of the relationship is negative

- The closer r is to -1 or 1 the stronger the linear relationship.

  If r=1, that means you can draw a straight line that goes through all data points

- If r is close to 0, then we say that there is a very weak linear relationship

  There may be other relationship (non-linear relationship)

How to Compute Correlation (r) in R: We can compute correlation much faster using the cor() funciton in R:

Example: Create the vectors with the observations of variables A and B from the previous example then use the cor() function to compute the correlation. What does this correlation value imply about the strength of the linear relationship?

```
A= c (100, 90, 95, 92, 88)
B= c (140, 90, 100, 135, 80)
cor (A, B)  or cor (B, A)  gives the exact same number
```

Answer: 0.7398

Caution: Remember that r can only measure the strength of a linear relationship. If you compute r for a strong exponential relationship, r will not reflect that strength (in fact, r would be close to 0 since a strong exponential relationship would be a weak linear relationship).

Practice Problem: Going back to the gapminder data set:

(a) Create a vector which contains the years in the data set that are present for the country Canada.

```
rows_canada= which(gapminder$country== "Canada")
```

(b) Create a vector which contains the population for Canada throughout those years.

```
population_canada= gapminder$pop[rows_canada]
years_canada= gapminder$year[rows_canada]
```

(c) Which variable do you think is the explanatory variable (i.e. which variable should be plotted on the x-axis)?

(A) Population [✓ (B) Year]

(d) Plot a scatterplot of the variables. What is the direction of the relationship?

[✓ (A) Positive] (B) Negative

(e) What is the form of the relationship?

(A) Exponential (B) Quadratic (C) Logarithmic [✓ (D) Linear]

(f) What is the correlation r between the two variables? Interpret this value.

[✓ (A) 0.999] (B) 0.5 (C) 0 (D) −0.999