6.1: Bivariate Data

Textbook: 11.8

Objectives

- Be familiar with the idea of bivariate data
- Know how to produce and read a scatterplot
- Use R to compute the correlation coefficient and know how to interpret it

Motivation

Up until this point, we've been examining variables one at a time, attempting to gain an understanding of how they behave.

Examples

- What is a typical waiting time in a Tim Horton's drive-thru?
- What proportion of freshman (first year students) took a "gap year" before starting university?
- How much do the heights of MLB players vary?

While it is important to be able to describe individual variables in a given sample (and use that information to make an inference about the corresponding population), there are many situations in which the relationship between two variables might be of more interest.

Examples

- How can we describe the relationship between height and weight in high school students?
- To what degree are temperature and rainfall levels related?
- Is there an association between the duration of an eruption of the Old Faithful geyser and the amount of time between eruptions?

In this set of notes, we will introduce the idea of bivariate data and talk briefly about some of the things we look at when examining two variables at the same time.

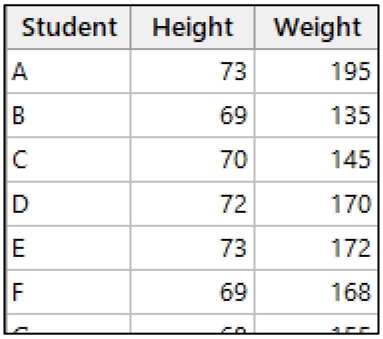
Bivariate Data

When we refer to bivariate data, we're referring to information collected on two variables from the same set or sample of observations.

Example

Suppose we were interested in the relationship between height and weight in high school students. We take a sample of $n = 43$ students and measure each students' height and weight.

In this case, we have two variables of interest: height and weight. Each individual in our sample is measured on both of these variables. We could keep track of this information in a spreadsheet as follows:



Student	Height	Weight
A	73	195
B	69	135
C	70	145
D	72	170
E	73	172
F	69	168

Notice that these measurements of height and weight are "paired" in the sense that there is one height and one weight for each of the n students. If we let X represent the "Height" variable and Y represent the "Weight" variable, then we have n paired observations: $(X_1, Y_1,), (X_2, Y_2,), (X_3, Y_3,)$ ... etc.

What are we interested in when we look at bivariate data? We're interested in how the two variables relate to one another. Specifically…

1. Does there appear to be a relationship between the two variables? If so, how can we describe the "direction" of the relationship?
    - Does weight increase as height increases?
    - Is there a linear relationship between height and weight?

2. If there is a relationship between the two variables, how strong is this relationship?
    - Is there only a slight relationship between height and weight, or is it a fairly apparent relationship?
    - Might there be other variables that explain the relationship between height and weight?

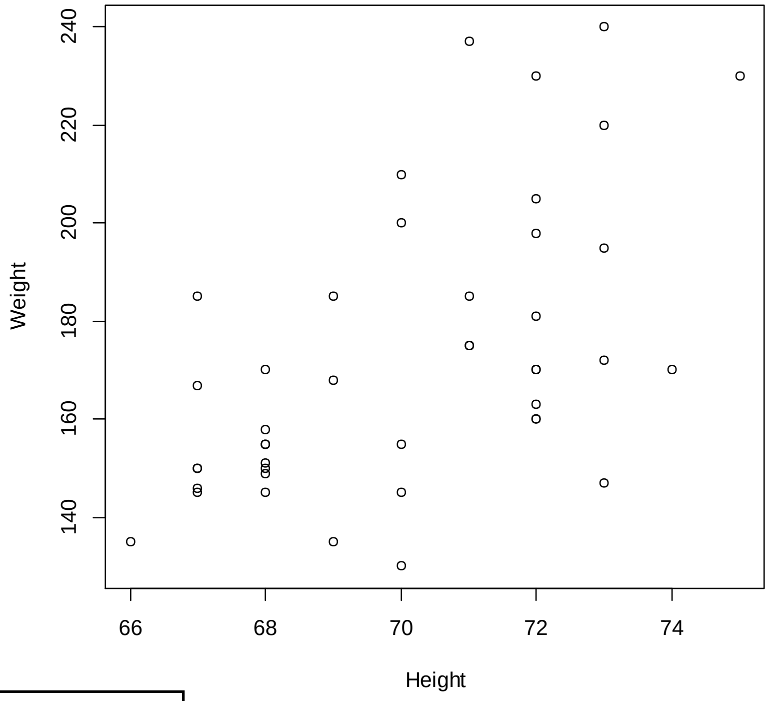
3. If the relationship between the two variables is strong, can said relationship be used to predict what will happen in the future?
   - If we know the height of person X, can we predict their weight with a reasonable degree of accuracy?

We will spend the rest of the 6.1 notes briefly discussing two “quick and easy” methods of assessing the relationship between two variables.

Scatterplots

Probably the quickest way to get an idea of the relationship between two variables – let’s call them X and Y – is to graph them. A scatterplot is a two-dimensional plot with one variables’ values plotted along the horizontal axis (x-axis) and the other variables’ values plotted along the vertical axis (y-axis).

Example 6.1.1
The “6.1 R” file contains the heights (in inches) and weights (in pounds) of a sample of $n = 43$ high school students. The following is a scatterplot of these data.



R function: `plot(x,y)`

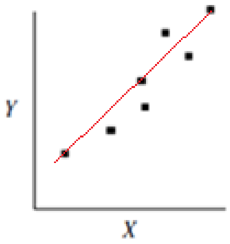
`plot(                )`

Notice that there is a point on the scatterplot for each of the $n = 43$ students. The points represent each pair of measurements (height, weight) for each student.

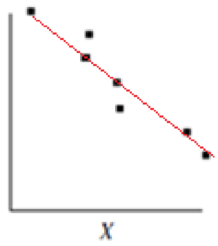
As you can see, the scatterplot can give a general idea of the relationship between the two variables. In general, we like to describe this relationship in terms of both its *direction* and its *strength*.

The *direction* of a relationship can generally be described as positive, negative, curvilinear, or non-existent.

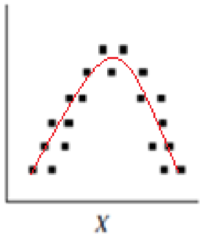
- A *positive* linear relationship exists when, as the values of one of the variables increase, the values of the other generally increase as well. (Plot A)
- A *negative* linear relationship exists when, as the values of one of the variables increase, the values of the other generally decrease. (Plot B)
- A *curvilinear* relationship exists when the relationship between the variables' values can best be descried by a curve (rather than a line). (Plot C)
- A *non-existent* relationship exists when there is no consistent relationship between the values of the two variables. (Plot D)



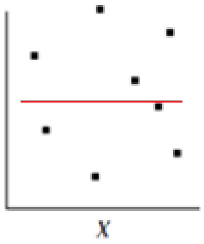
Plot A



Plot B

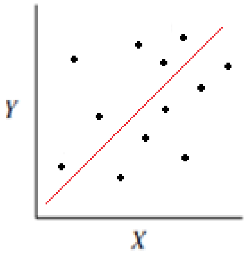


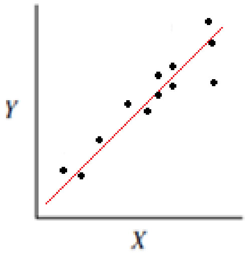
Plot C



Plot D

The *strength* (or *magnitude*) of a relationship is based on how tightly the "cloud" of points is clustered about the trend line (the line/curve that best describes the direction of the relationship. The more tightly clustered the cloud the stronger the relationship is between the two variables.





Example 6.1.1 (continued)
The relationship between height and weight in the example with $n = 43$ high school students appears to be a moderately strong positive linear relationship.

Correlation

Using a scatterplot is a good way to get a quick general idea of the relationship between two variables, but what if we wanted some way to quantify this relationship beyond the interpretation of a scatterplot?

Pearson's correlation coefficient (often just referred to as "correlation") is a measure of the direction and strength of a linear relationship between two quantitative variables $X$and Y. The sample correlation coefficient is usually denoted r and is computed as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_X s_Y}$$

where X_i is the i$^{\text{th}}$ observation of the X variable, Y_i is the i$^{\text{th}}$ observation of the Y variable, $\bar{\mathrm{x}}$ and $\bar{\mathrm{y}}$ are the means of the X and Y variables, respectively, $\mathrm{s_x}$ and $\mathrm{s_y}$ are the standard deviations of the X and Y variables, respectively, and n is the sample size.

The correlation coefficient is a scaled version of the covariance of two variables, which makes it a bit easier to interpret. The following are important features of the correlation coefficient:

- -1 ≤ r ≤ 1
- The sign of r indicates the direction of the linear relationship
  - +r indicates a positive relationship
  - -r indicates a negative relationship
- Values of r closer to -1 or 1 suggest a strong linear relationship (with $r = 1$ and $r = -1$ representing perfect positive and negative correlations, respectively); values of r closer to 0 suggest a weak linear relationship

Example 6.1.1 (continued)
The "6.1 R" file contains the heights (in inches) and weights (in pounds) of a sample of $n = 43$ high school students. Use R to compute the correlation between height and weight and interpret its meaning.

R function: `cor(x,y)`

`cor(          ,          )`