Describing Relationships between Two Variables

Up until now, we have dealt, for the most part, with just one variable at a time. This variable, when measured on many different subjects or objects, took the form of a list of numbers. The descriptive techniques we discussed were useful for describing such a list, but more often, science and society are interested in the relationship between two or more variables. To take a mundane example, it is nice to know what the "typical" weight is, and what the typical height is. But more interesting is to know the relationship between weight and height.

For now, and most of this course, we'll stick to relations between only two variables. The sorts of questions we'll examine are:

1. Does y increase with x? Decrease? Does it depend on what values? For example, it seems intuitive that weight increases with height (taller people tend to weight more), but for other things, perhaps y goes up and then goes back down again.

2. Suppose y does increase with x. How fast? If you are 2 inches taller, how much heavier are you likely to weigh?

3. Is the relationship strong? Can you make reliable predictions? That is, if you tell me your height, can I predict your weight?

   To parallel our discussion with just one variable, we'll discuss

   I. Graphical Summaries

   II. Numerical Summaries

<u>Graphical Summary</u>

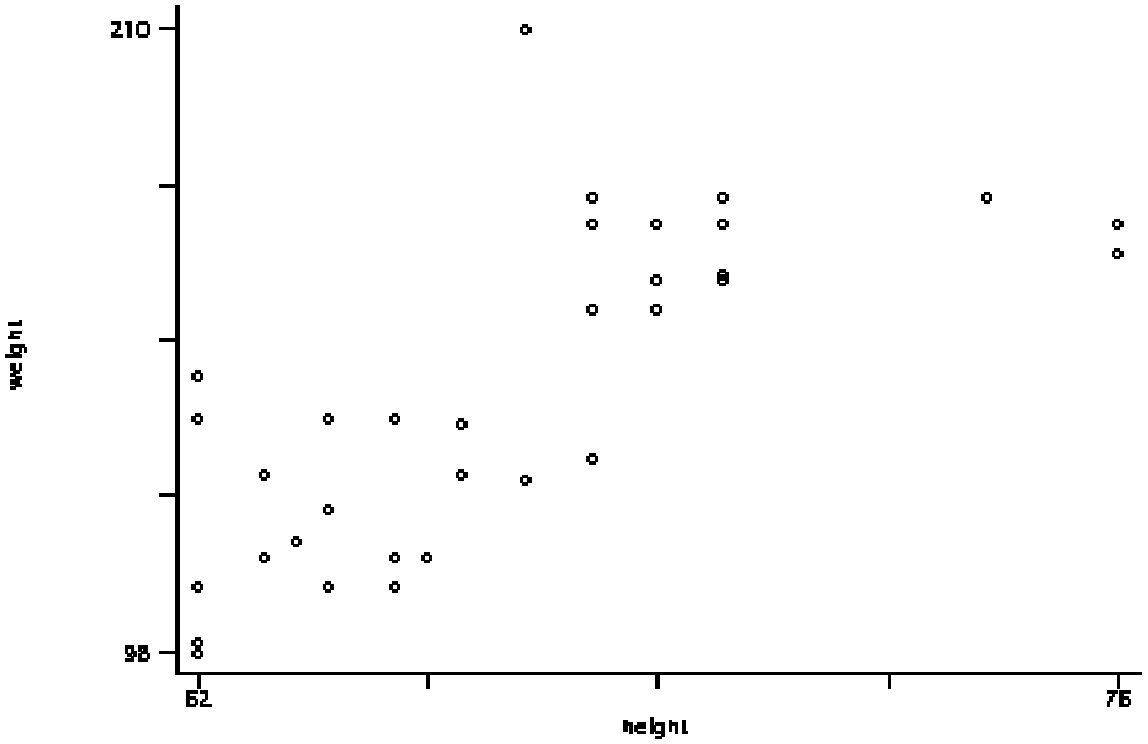
The foremost technique is to use scatterplots. Scatterplots plot points (x,y). They give us a summary of what the relationship looks like. Here are the things to look for:

1. Is the relationship positive (x goes up and y goes up, x goes down and y goes down), negative (x goes up, y goes down), or is there no relationship? (Looks like blob.)

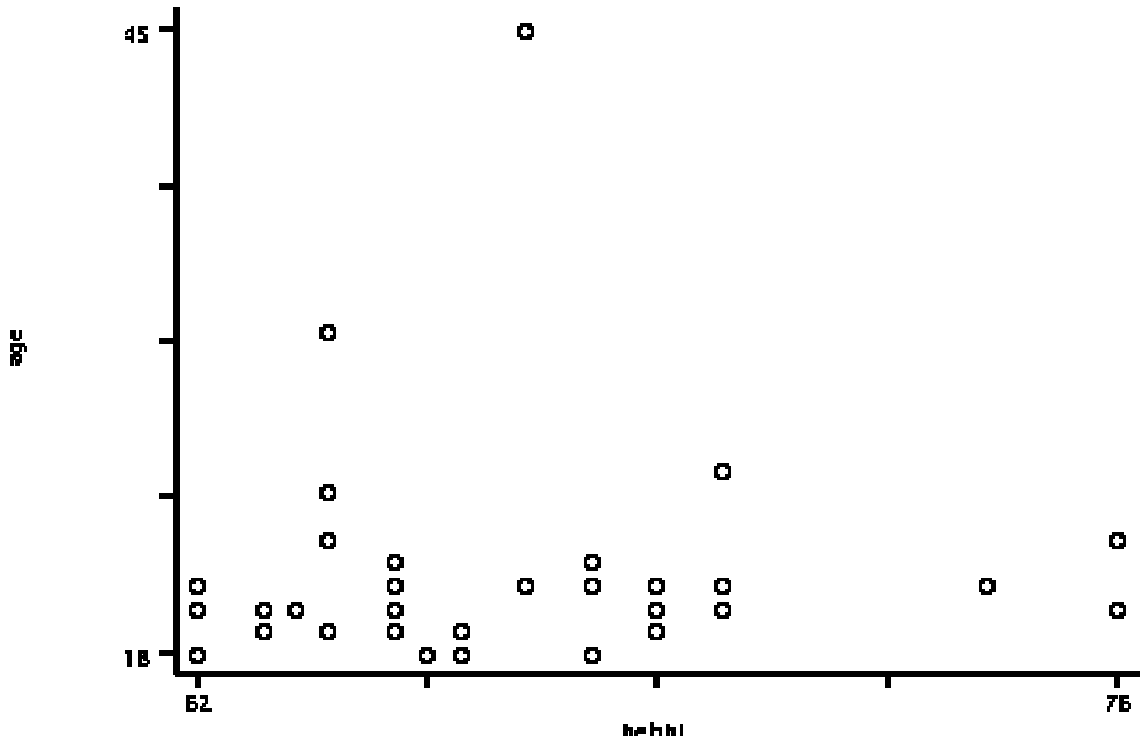
2. Is the relationship linear, quadratic, something else?

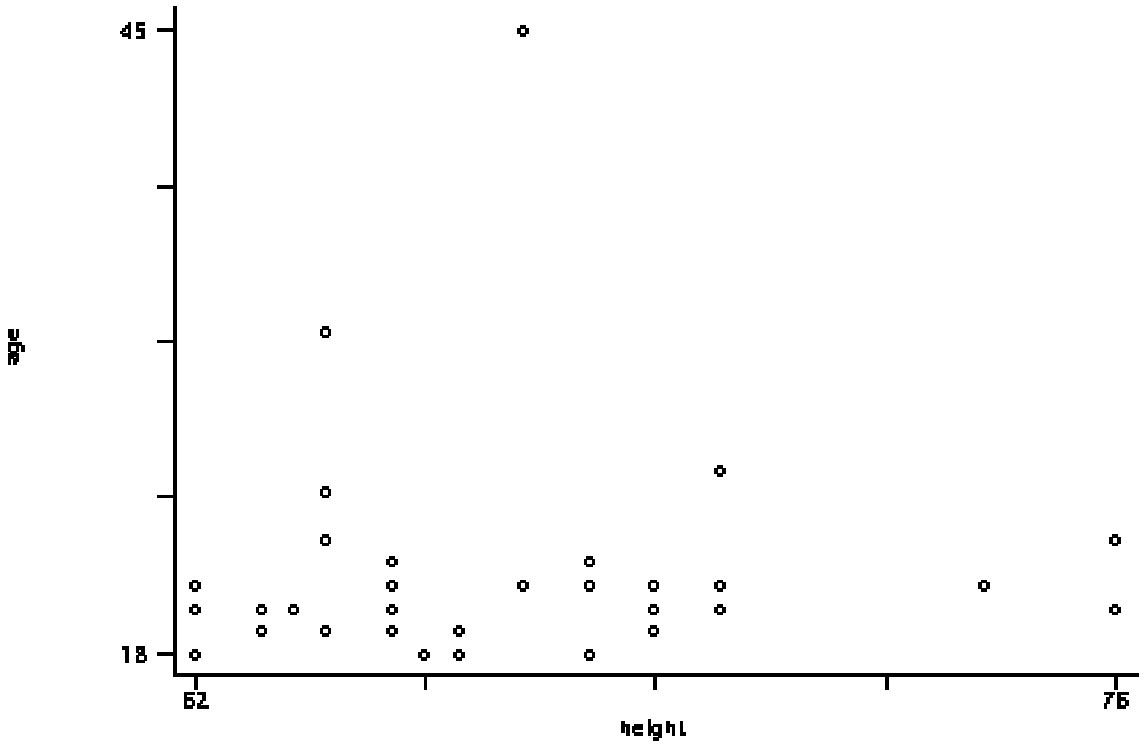
3. Is the relationship strong (clear patterns) or weak (fuzzy patterns)?

Here are some examples:

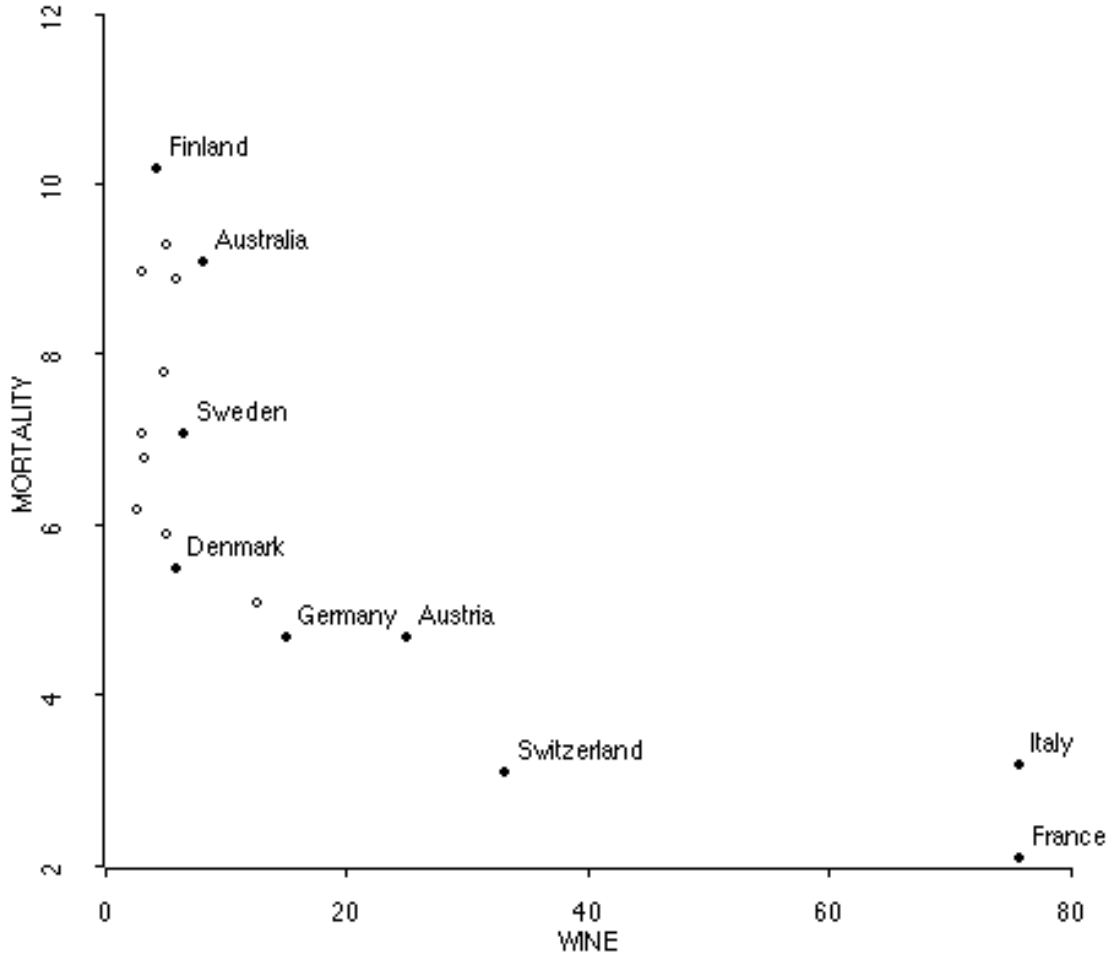


This data comes from class. The picture comes from plotting each person's height and weight. For example (74, 180), (69, 175), (76, 170), etc. This is a positive relation, fairly weak, and possibly linear.

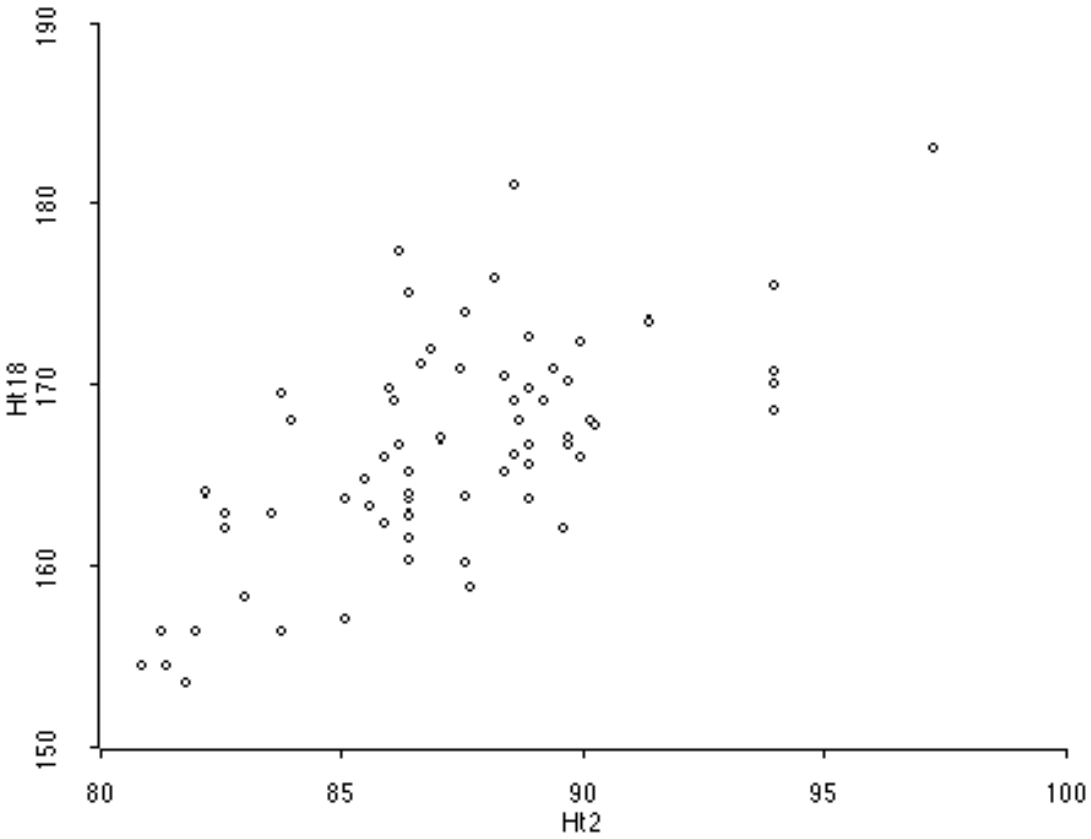




There does not seem to be any relationship here. People of a given height can be any age.

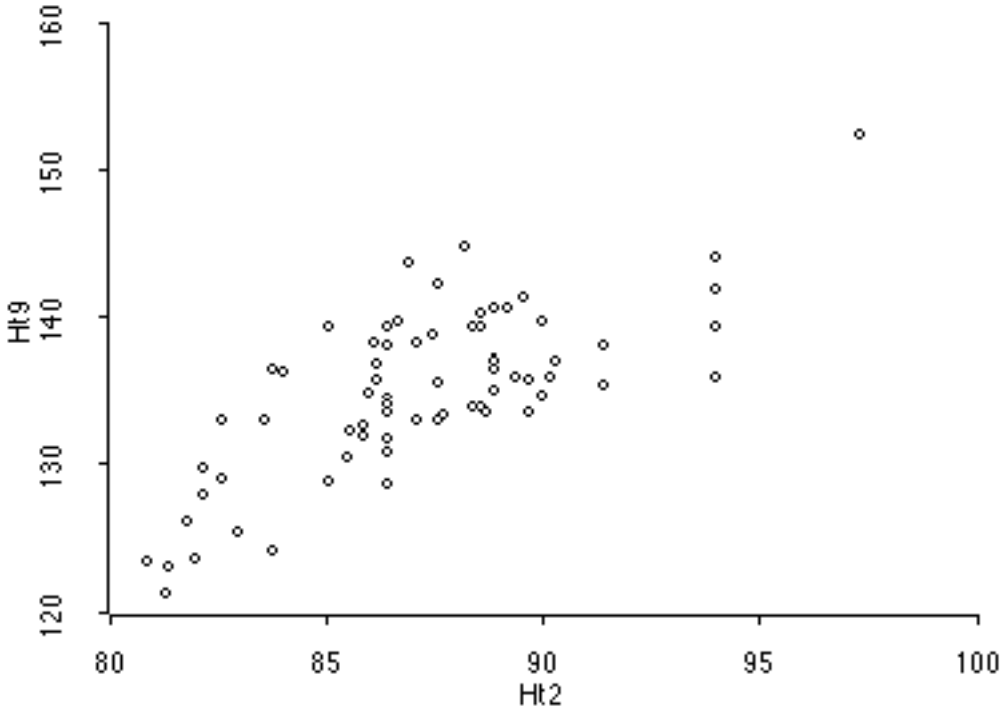


A negative relation. Mortality decreases as per capita wine consumption increases. The relation is not linear. (A straight-line would be a poor description of the trend.) The trend is fairly strong.



This shows the heights of girls at age 2 and later at age 18 (in cm). Below is height at age 2 and then at 9.

A fairly strong positive linear trend. However, it is not as strong as the positive linear trend below. Does this make sense?



Some guidelines.

--When dealing with a relation, the x and y variables have particular roles to play. The x (horizontal axis) variable represents the variable that explains what we see in y (the vertical axis.) Often, x is a quantity which we can change or have control over. Or at least, we want to see how y reacts to different values of x. So x is usually called the independent or explanatory variable, and y the dependent or response variable. Another way to think about it, sometimes, is that x is used to predict y. So x is something that comes before y in time.

That is why, for example, height at age 2 is on the x axis and height at age 18 on the y.

We want to know how what we learn about height at age 2 predicts height at a later age.

Does it make sense, then, to have mortality on the x axis, and wine consumption on the y? Perhaps, but probably this would be an unusual arrangement. It's hard to imagine a situation in which someone says, "I'll tell you the death rate of a country, and you predict their wine consumption." Probably for most medical purposes, since wine consumption is something that is changeable, the arrangement shown above is best.

Numerical summary

It is a bit much to expect a single number to summarize a relation between two variables. And it is. However, it turns out that linear relations are quite prevalent in many studies, and even if a relation is not linear, there is often a way of making it so. For this reason, the numerical summary we use applies only to linear relations.

So assume we are talking about a linear relation. The correlation coefficient is a number that measures the strength of a linear relation. It is represented by the letter r, and it has these qualities:

1) It is a number between -1 and 1 (inclusive).

2) If it is negative, then the relation is a negative one (x goes up y goes down) and if positive, then the relation is positive.

3) If it is 0, then there is no linear relation.

4) The closer to +1 or -1, the stronger the relation.

There are particular shapes associated with particular values of r. If r is 0, the scatterplot is a blob. If r = 1, then it is a straight line with positive slope. If $0 < r < 1$, then the scatterplot falls somewhere in between a blob and a straight line.

Reminder: this applies only to linear relations. If the relation is non-linear, r might be close to 0 even if the relation is quite strong. This happens with quadratics, as you'll see later.

r measures things with respect to averages. If someone says "x and y are highly positively correlated", this means that if x is above average, then y tends to be above average. If x is below average, then y tends to be below average. So if height and weight are positively correlated, this means that people who are taller than average tend to be heavier than average. People who are taller than average at age 2 tend to be taller than average at age 9.

The correlations for the above graphs are:

Height, Weight: r = 0.72

Age, Height r = 0.06

Wine, Mortality: non-linear; don't calculate r

Ht2, Ht 9 r = 0.738

Ht2, Ht 18, r = 0.663

<u>Golden Rule of Correlations</u>

Correlation does not imply cause-and-effect

blanket sales in canada and brush fires in Australia are positively correlated. Does buying a blanket cause a fire in Australia?

The amount of damage at a fire is positively correlated with the number of fire engines that show up to put out the fire. Can damages be reduced by not sending fire trucks?

Your age is negatively correlated with the distance of Halley's comet from the earth. When the comet returns, will you grow younger?

The bottom line is that confounding variables prevent us from making a cause-and-effect

conclusion. The best we can do is accept that we are describing a relationship without knowing what makes the relationship happen, or if the relation is "real."

Calculating r

There is a formula for calculating r. I don't expect you to memorize it, but you should know how to use it to find r for a list of pairs of numbers, and you must know what it means.

First we have to define a new obj