Chapter 7: Scatterplots, Association and Correlation

The Big Change

So—we've done quite a bit with univariate data (quantitative and qualitative), and we've also looked at bivariate qualitative data (Chi-Square for Independence). Now it's time to look at bivariate *quantitative* data.

As was the case with univariate quantitative data, we begin by looking at the distribution graphically. Since we've now got two variables, our old methods of visualization won't work.

We need a new graph!

Graphing the Data

Enter the scatterplot! Plot one variable horizontally, and the other vertically. Each measurement pair becomes a coordinate pair (point) in the plot. Remember to include a scale and label for each axis! Also notice that the origin is not the point at the lower left…there are some people who insist on that; AP does not.

A title is nice, but not required.

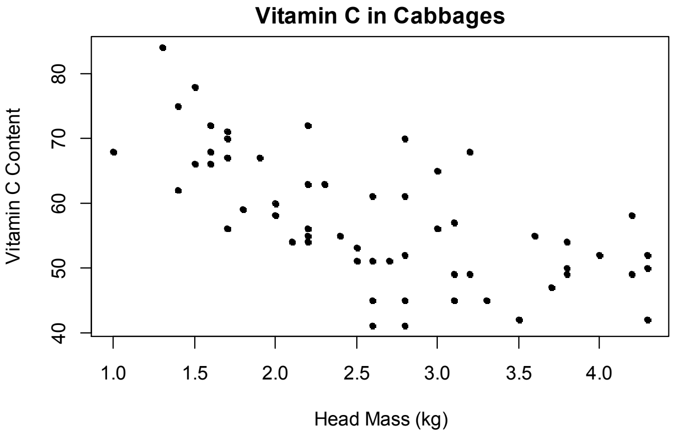

Figure 1 - A Scatterplot

Variables

Deciding which variable to plot horizontally is important. Once again, if we were the ones actually doing the work, we'd know—but we as students will (probably) have to work harder to figure it out.

Every bivariate study includes an **explanatory variable** (independent) and a **response variable** (dependent). We are going to look and see if changes in one of these (the explanatory) causes a change in the other (the response). So, in our mind—and *only* in our mind—do we wonder "does a change in *this* variable cause a change in *that* variable?" Or—and it is OK to say this one aloud—"Can I predict the value of *that* variable from *this* variable?"

The variable that you predict is the **response**. The one that you use as a basis for the prediction is the **explanatory** variable.

Examples

[1.] A realtor wants to construct a model that uses the distance from the center of downtown (in miles) to predict the home price (in dollars). Which variable is explanatory and which is response?

Since the realtor wants to predict the home price, that should be the response variable. Distance from downtown should be the explanatory variable.

[2.] An actuary has collected some data—specifically, for each year, the number of vehicle deaths was found. What are the (likely) explanatory and response variables that the actuary would use?

It should be pretty obvious that the actuary wants to predict the number of deaths—thus, that should be the response variable. Year should be the explanatory variable.

By the way…you'll see lots of examples in textbooks where "year" (or some kind of time variable) is used as an explanatory variable. Technically, this should be handled using a different procedure…but I won't try to explain the difference here.

[3.] A doctor has collected some data concerning the weight of a new mother (in pounds) and the mass of the child (in grams). Which of these is likely to be the response variable that the doctor would use?

I believe that one would want to predict the mass of the child, since measuring the weight of the mother is easy. Thus, the child's birth mass ought to be the response variable.

Main Features

With univariate data, we looked for center, shape and spread. Not so for bivariate data…The primary quality for which we will look is **linearity**—that's the simplest type of function that can relate two variables. Naturally, we don't expect the points to fall exactly on a line—the line is going to be fuzzy. Do you get the feeling of a line when you look at the graph? Can you unfocus your eyes and imagine a line? For us, our responses will either be "linear" or "nonlinear."

Here is an example of a scatterplot that is linear.

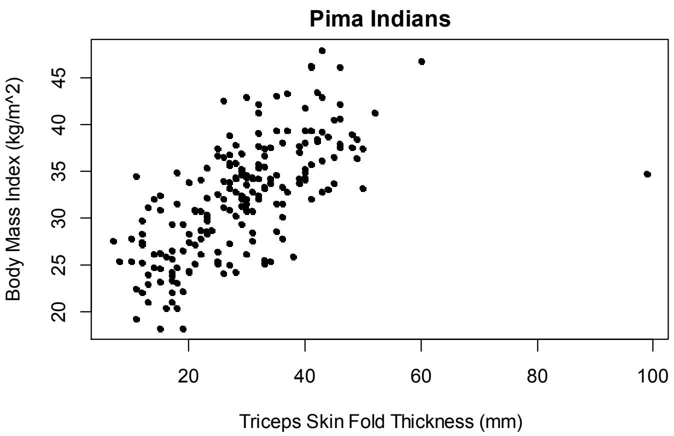

Figure 2 - A Linear Scatterplot

Here is an example of a scatterplot that is clearly non-linear.

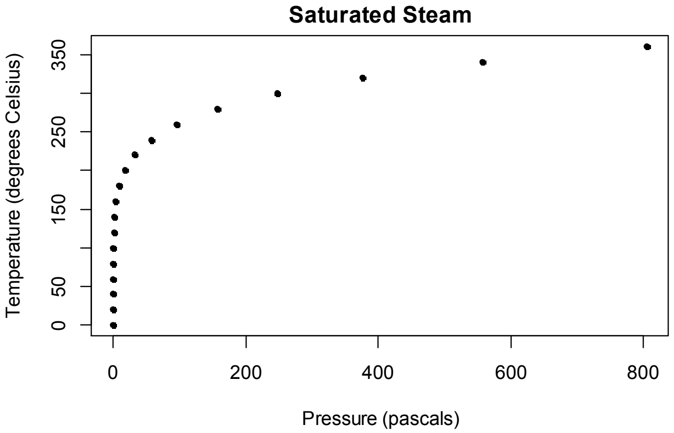

Figure 3 - A Non-Linear Scatterplot

Here is an example of a scatterplot that is neither linear nor curved. Technically, we could call this "non-linear," but it is of a very different quality than the preceding scatterplot!

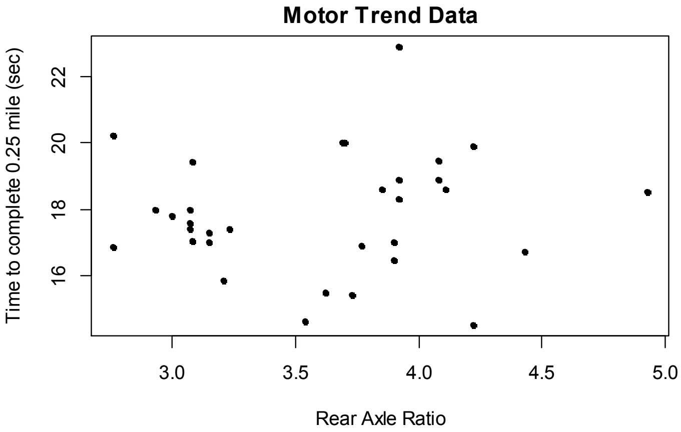

Figure 4 - Neither Linear nor Curved

Since we'd like to predict the value of the response variable from the explanatory variable, we'd very much like for the relationship to have some pattern. The more obvious the pattern in the plot, the stronger the relationship between the variables is. Thus, one feature on which you should comment is the **strength** of the relationship. The cabbage scatterplot above isn't very strong. The graph below shows a *very* strong relationship between carapace width and length.

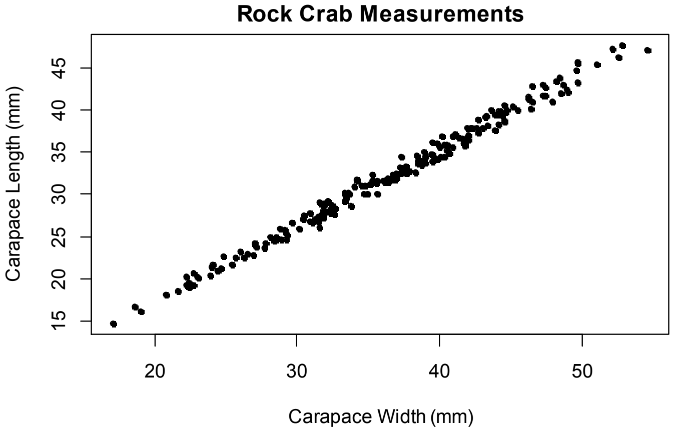

Figure 5 - A Strong Relationship

The fuzzier the pattern, then the weaker the relationship. The graph below shows a weak relationship.

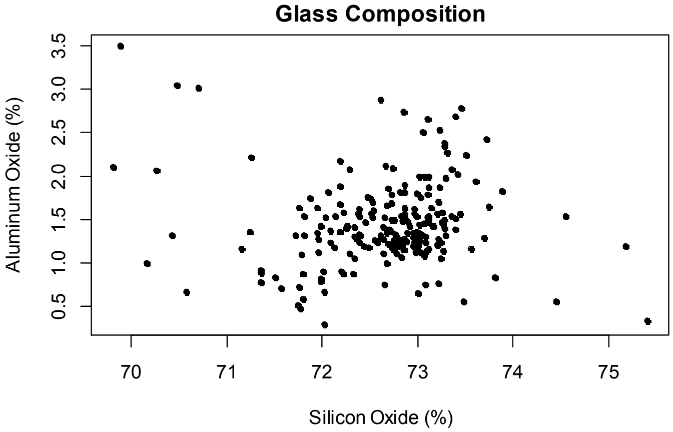

Figure 6 - A Weak Relationship

There is no definitive guide to determine whether a relationship should be called *strong* or *weak* (or *moderate*)—and in truth, the distinction will depend on the context! Don't worry too much about it for this class. Look at lots of examples, and (hopefully) you'll get a feel for it, eventually.

For graphs that are linear (and a few that are not), we should also comment on the **direction**. This is another way of talking about the sign of the slope of the line (if the relationship appears linear). If the line has a positive slope, then the relationship has a **positive** direction. A great way to think about this is that larger values of the x-variable are plotted with larger value of the y-variable, and smaller values of the x-variable are plotted with smaller values of the y-variable (and this will be true regardless of the linearity). The rock crab scatterplot above has a positive direction, and the glass composition scatterplot has a negative direction.

Note that vertical and horizontal lines are said to have *no direction*. This is (in part) because vertical and horizontal lines don't help us make predictions.

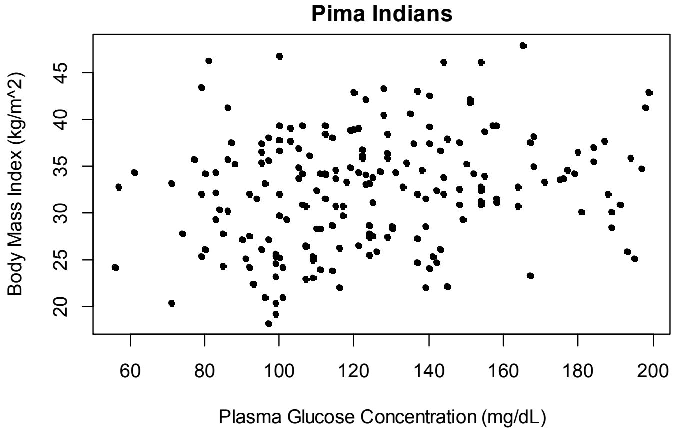

Figure 7 - A Scatterplot with No Direction

In the case of univariate data, we also looked for unusual features. We'll continue that practice with bivariate data. Some unusual features include clusters, gaps and outliers (we'll talk about more kinds later).

Clusters of points within the plot are just what it says—clusters of points! Note that clusters can indicate the presence of another variable—you'll investigate this further if you continue your studies of statistics.

The scatterplot below shows two clear clusters.

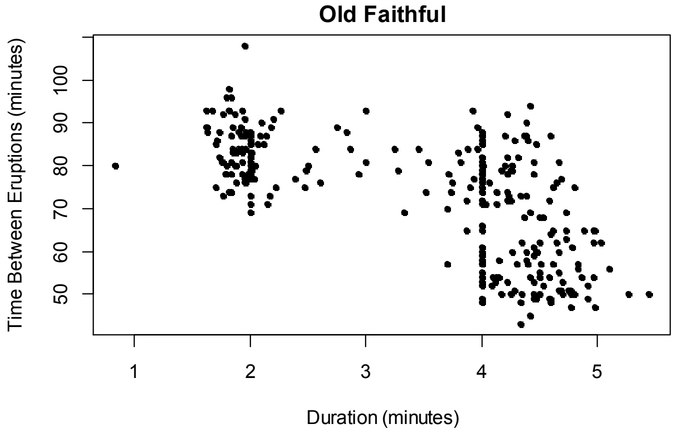

Figure 8 - A Scatterplot with Two Clusters

Gaps are regions (values) of the explanatory variable that have no associated response measurements. It is not the case that a cluster automatically makes a gap! Note the Glass Composition scatterplot above…there is a clear cluster of points in the center of the graph, but no apparent gaps. The Old Faithful data above also show two clusters with no gaps. The scatterplot below shows a clear gap between 4 and 5 (thousand) pounds…but I wouldn't say that there are two clusters!

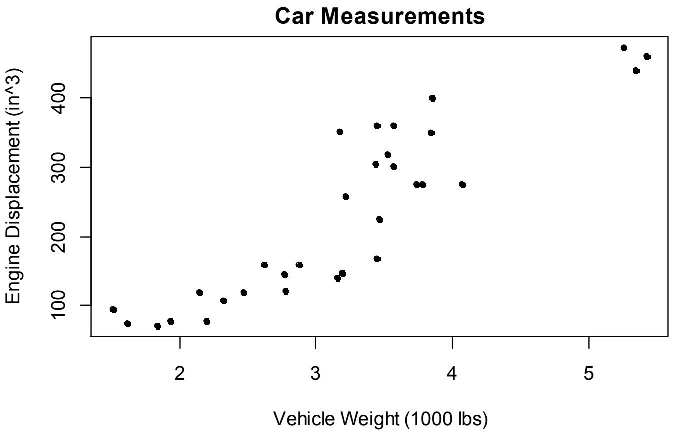

Figure 9 - A Scatterplot with a Gap

Any point which does not seem to be in accord with the others is an **outlier**. Unlike before, there is no numerical rule (*e.g., Tukey's Rule*) for determining if a point is an outlier—you just have to look at the graph. The graph below shows two clusters and one outlier (at the bottom left).

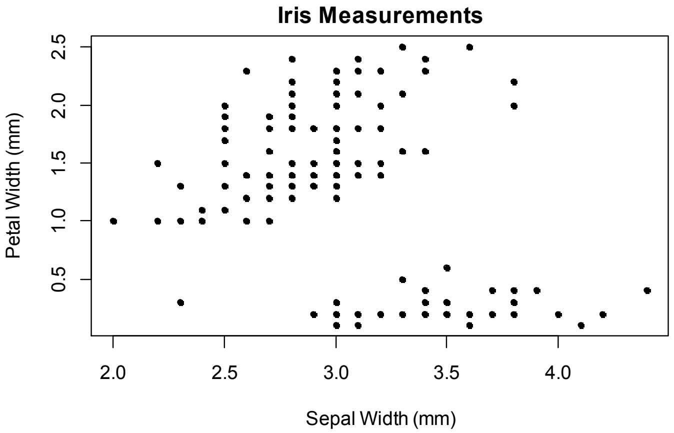

Figure 10 - Two Clusters, A Gap, and an Outlier

There was another graph earlier in this document with a clear outlier…see if you can find it!

Examples

[1.] Data were collected from a sample of Loblolly pine trees—specifically, age (in years) and height (in feet). Construct and comment on a scatterplot suitable for predicting age from height.

Height	4.51	10.89	28.72	41.74	52.7	60.92
Age	3	5	10	15	20	25

First, the graph:

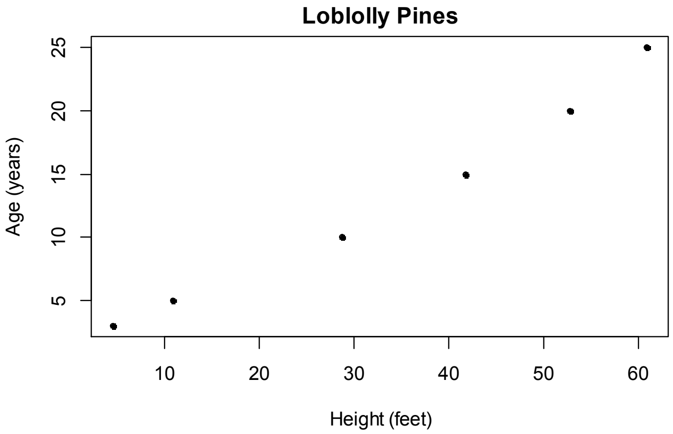

Figure 11 - Scatterplot for Example 4

There appears to be a strong, positive linear relationship between height and age of Loblolly pine trees.

If you thought that the graph was curved, I won't say you're wrong…but the curve is so slight as to be hardly recognizable. This plot is not clearly curved; it is basically linear. Also…I'm not willing to say that there is a gap in the data. I wouldn't blame you if you thought that there was one.

[2.] Data from vintage 1974 cars were collected. Construct and comment on a scatterplot suitable for predicting mileage (miles per gallon) from horsepower.

MPG	21.0	21.0	22.8	32.4	30.4	33.9	27.3	26.0	30.4	15.8	19.7	15.0	21.4
HP	110	110	93	66	52	65	66	91	113	264	175	335	109

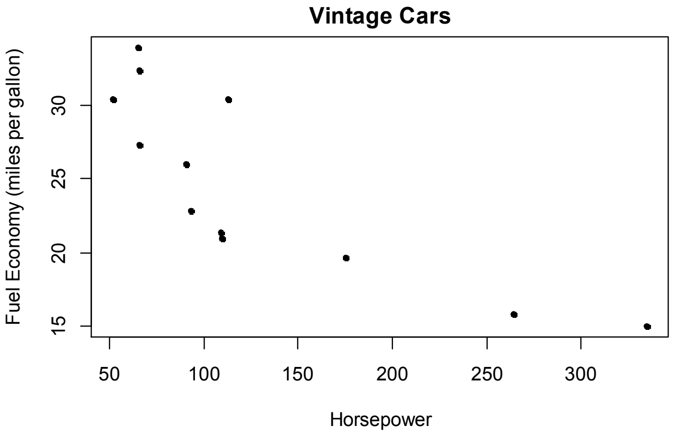

Figure 12 - Scatterplot for Example 5

There appears to be a fairly strong, negative non-linear relationship between horsepower and mileage.

Correlation

Not all linear relationships are equal. Some are very nearly perfect lines while others are hardly linear at all. This is quite subjective, though—it would be nice to firm up these observations with some numbers…

Enter the **Correlation Coefficient**! Officially named Pearson's Product-Moment Correlation Coefficient (*say that three times, fast*), most people simply refer to it by the symbol that we use for it: *r*.

This measures (quantitatively) the *strength* and *direction* of a *linear* relationship. Note the last part of that—if the variables do not have a (vague, at least) linear relationship, then the correlation coefficient should not be used. Specifically, one should NOT use the value of *r* to determine whether or not the relationship between two variables is linear!

If you are asked to interpret the correlation coefficient, be sure to specifically address (*in context!*) what it tells you about the strength, direction and linearity of the relationship between the variables.

Notes about the correlation coefficient:

[a] *r* has a value between -1 and 1 (inclusive). The farther the value is from zero, the closer the points are to falling on a line. The nearer to zero, the less the points look like a line (more like a blob). This is the *strength*. Remember that this only makes sense if the data appear to have some kind of linear relationship! It is easy to find examples of relationships that are very strong, but have correlation coefficients close to zero (because they are curved). Conversely, it is easy to find relationships with correlation coefficients close to 1 that are not linear.

[b] the sign of *r* is also the type of association the data exhibit (positive or negative). This is the *direction*. Again, this direction only makes sense if the relationship is (vaguely) linear.

[c] *r* has no units, and is unaffected by changing units in the original data. The formula might make this clearer…

Equation 1 - The Correlation Coefficient

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n-1}$$

Notice in the formula that the numerator looks a lot like the standardizing formula—in fact, that's exactly what it is. Standardized scores are the same no matter what transformations are applied to the data—so *r* stays the same, also!

[d] *r* is the same regardless of which variable is assigned to be explanatory and response.

Examples

[6.] Calculate and interpret the correlation coefficient for the Lobolly Pine data above.

I get *r* = 0.9911. Since this value is very close to 1, it confirms my observation of a positive linear relationship between height and age of Loblolly pine trees.

Note that you do not need to show the formula for r! No one does that by hand anymore…just use your calculator.

[7.] Calculate and interpret the correlation coefficient for the 1974 car data above.

I get *r* = −0.8001. As this value is closer to -1 than 0, it indicates a moderate to strong negative association between mileage and horsepower.

[8.] A study found the correlation between height (cm) and weight (kg—though that's really mass) for a group of athletes to be 0.781. A reporter wants to convert the measurements to pounds and inches for U.S. readers. How will this conversion affect the correlation between the variables?

The correlation will remain the same! Changing the units of measure will not change the correlation coefficient.

Well…that's not entirely true. If one of the variables is made negative, then the sign of the correlation coefficient would change. Perhaps it would be better to say that the absolute value of the correlation coefficient does not change…and even then, that only applies to linear transformations.

Warnings!

Just because two variables are strongly correlated doesn't mean that that there is a causal relationship. It is entirely possible that there is a third variable that is really causing changes in the two variables that you've plotted. Thus, statisticians are fond of saying: *correlation does not imply causation!*

Another warning concerns the use of averaged data. Every point on a scatterplot should represent a single measurement—not an *average* measurement. When you plot averages, the amount of variation in the plot decreases and makes relationships look stronger than they really are.

Straightening Scatterplots

The textbook now begins describing methods for making a non-linear scatterplots more linear. I'm going to save that for Chapter 10.