Linear Relationships: Properties of r

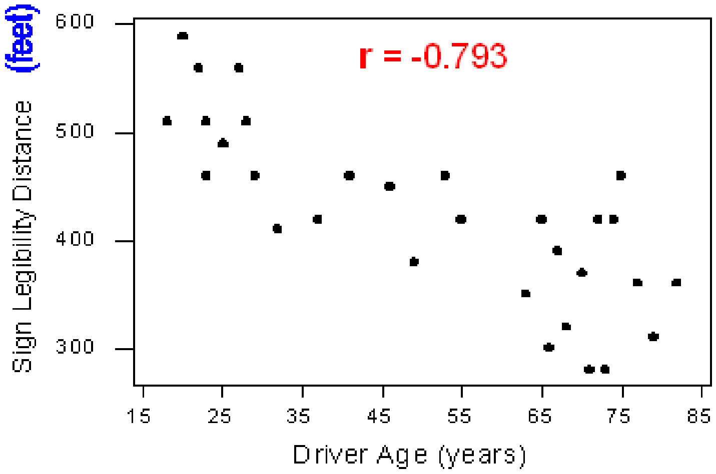
Learning Objective: Interpret the value of the correlation coefficient, and be aware of its limitations as a numerical measure of the association between two quantitative variables.

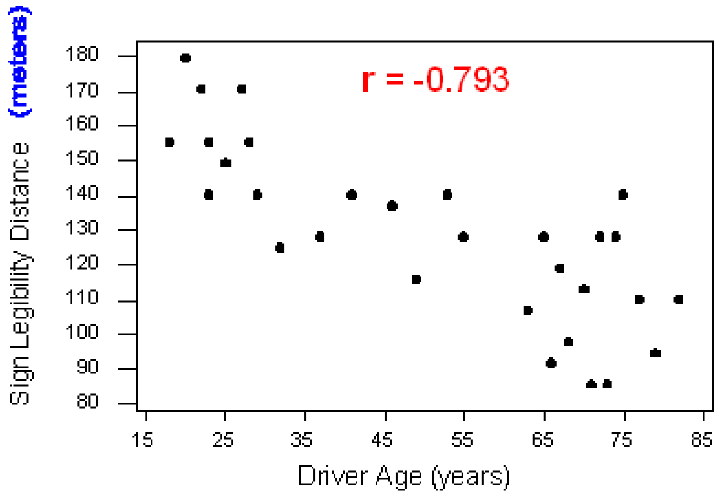
Properties of r

We now discuss and illustrate several important properties of the correlation coefficient as a numerical measure of the strength of a linear relationship.

1. The correlation does not change when the units of measurement of either one of the variables change. In other words, if we change the units of measurement of the explanatory variable and/or the response variable, the change has *no effect on the correlation (r)*.

   To illustrate, following are two versions of the scatterplot of the relationship between sign legibility distance and driver's age:

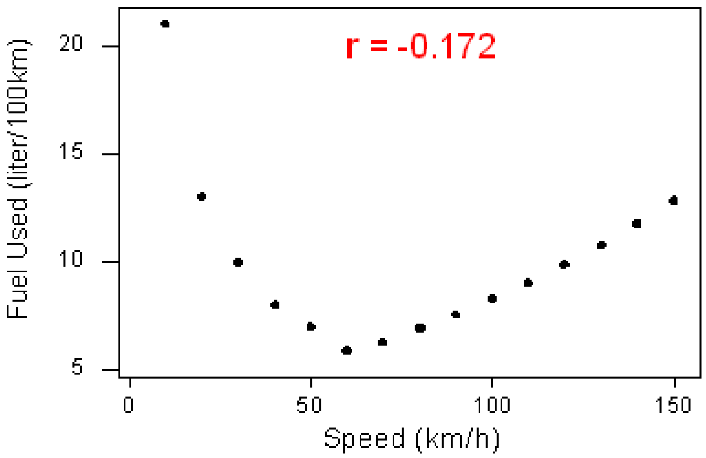




The top scatterplot displays the original data where the maximum distances is measured in *feet*. The bottom scatterplot displays the same relationship but with maximum distances changed to *meters*. Notice that the Y-values have changed, but the correlations are the same. This example illustrates how changing the units of measurement of the response variable has no effect on r, but as we indicated above, the same is true for changing the units of the explanatory variable, or of both variables.

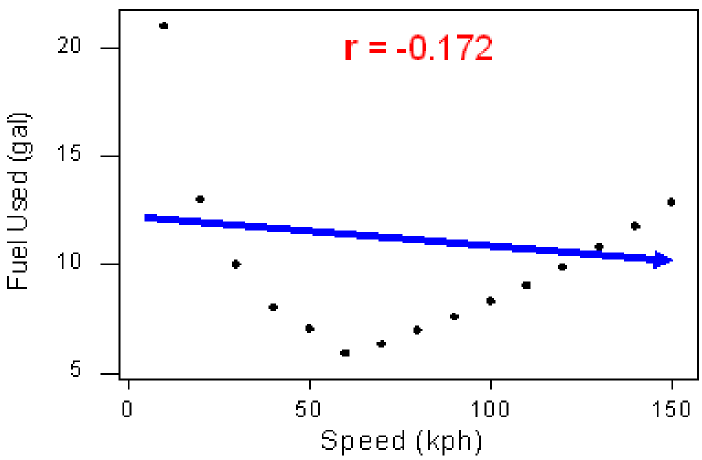
This might be a good place to comment that the correlation (r) is *unitless*. It is just a number.

2. The correlation measures only the *strength* of a linear relationship between two variables. It **ignores** any other type of relationship, no matter how strong it is. For example, consider the relationship between the average fuel usage of driving a fixed distance in a car, and the speed at which the car drives:



Our data describe a fairly simple curvilinear relationship: the amount of fuel consumed decreases rapidly to a minimum for a car driving 60 kilometers per hour, and then increases gradually for speeds exceeding 60 kilometers per hour. The relationship is very strong, as the observations seem to perfectly fit the curve.

Although the relationship is strong, the correlation r = -0.172 indicates a weak *linear* relationship. This makes sense considering that the data fails to adhere closely to a linear form:

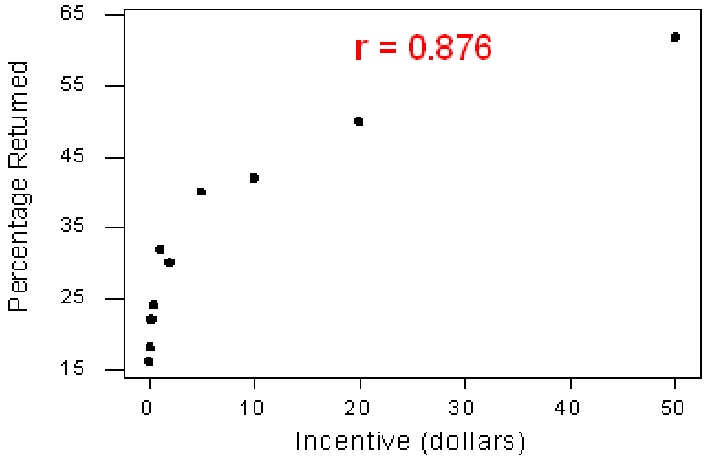


The correlation is useless for assessing the strength of any type of relationship that is not linear (including relationships that are curvilinear, such as the one in our example). Beware, then, of interpreting the fact that r is close to 0 as an indicator of a weak relationship rather than a weak *linear* relationship. This example also illustrates how important it is to *always look at the data in the scatterplot* because, as in our example, there might be a strong nonlinear relationship that r does not indicate.

Since the correlation was nearly zero when the form of the relationship was not linear, we might ask if the correlation can be used to determine whether or not a relationship is linear.

3. The correlation by itself is *not* sufficient to determine whether a relationship is linear. To see this, let's consider the study that examined the effect of monetary incentives on the return rate of questionnaires. Below is the scatterplot relating the percentage of participants who completed a

survey to the monetary incentive that researchers promised to participants, in which we find a *strong curvilinear relationship*:



The relationship is curvilinear, yet the correlation r = 0.876 is quite close to 1.

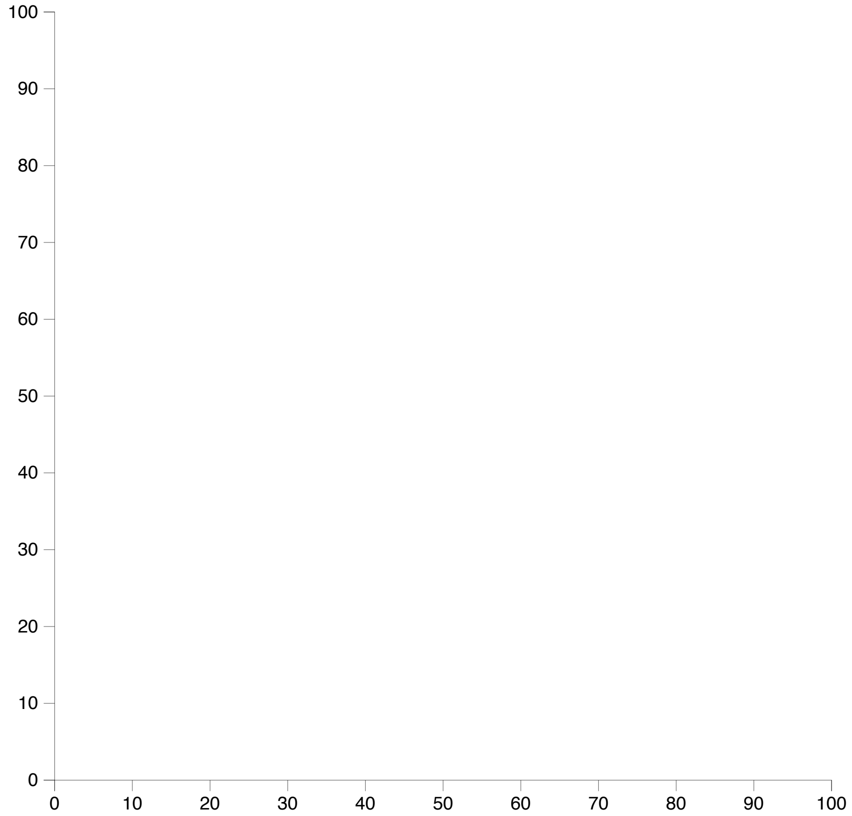
In the last two examples, we have seen two very strong curvilinear relationships, one with a correlation close to 0 and one with a correlation close to 1. Therefore, the correlation alone does not indicate whether a relationship is linear. The important principle here is:

Always look at the data!

4. The correlation is heavily influenced by outliers. As you will learn in the next two activities, the way in which the outlier influences the correlation depends upon whether or not the outlier is consistent with the pattern of the linear relationship.

   Using the simulation below, we will explore how an outlier affects the correlation.

Correlation r = 0 # points = 0



Draw line Add point Draw Least Squares Show mean X & Y Clear

To see how an outlier affects the correlation, do the following:

Fill the scatterplot with a hypothetical positive linear relationship between X and Y (by clicking on the graph about a dozen times starting at lower left and going up diagonally to the top right). Pay attention to the correlation coefficient calculated at the top left. (Clicking on the garbage can ("Clear" button) will let you start over.)

Once you are satisfied with your hypothetical data, create an outlier by clicking on one of the data points in the upper right of the graph, and dragging it down along the right side of the graph. Again, pay attention to what happens to the value of the correlation.

Hopefully, you've noticed the correlation decreasing when you created this kind of outlier, which is *not consistent* with the pattern of the relationship.

The next activity will show you how an outlier that *is consistent* with the direction of the linear relationship actually strengthens it.