In the last class we learned how to find the slope and intercept of the line that represents the best fit to a set of \((x,y)\) data points. What we didn’t say was how good that best-fit line actually is. In fact, we can find a best-fit line for basically any set of points, even if those points clearly aren’t related linearly. Consider the following points:

- \((x_1, y_1) = (1, 2)\)
- \((x_2, y_2) = (3, 3)\)
- \((x_3, y_3) = (2, 3)\)
- \((x_4, y_4) = (0, 0)\)
- \((x_5, y_5) = (4, 2)\)

By applying the linear regression procedure, we find that the best-fit line, \(y = A + Bx\), has \(A = 1.0\) and \(B = 0.5\). How does this look when it’s plotted up? In Fig. 1, the line \(y = 1 + 0.5x\) seems like a reasonable fit if the goal is to fit the data to a line, but from inspection of the points it hardly looks like \(x\) and \(y\) are linearly related. It looks more like the points form something like an inverted parabola. The linear regression procedure doesn’t actually care about that; it’s going to give you values for \(A\) and \(B\) regardless of what the \((x, y)\) points look like.

There is a tool for determining how well a set of points observes a linear relation. This is known as the correlation coefficient, or more precisely the linear correlation coefficient, and it is typically denoted as \(r\). To see where this idea comes from, consider a line, \(y = A + Bx\). We pick a point on that line and call it \((x_0, y_0)\). Now suppose we move a distance \(\Delta x\) in the \(x\)-direction. To remain on the line, we have to move a distance \(\Delta y\) in the \(y\)-direction, where \(\Delta x\) and \(\Delta y\) are related by

\[
\frac{\Delta y}{\Delta x} = B.
\]

This is, of course, no revolutionary result. We do want to make one useful, if trivial, observation, namely that the sign of \(\Delta y/\Delta x\) is fixed for a given line. If the slope of the line, \(B\), is positive, then \(\Delta x\) and \(\Delta y\) must always have the same sign. If \(B\) is negative, then \(\Delta x\) and \(\Delta y\) must have opposite sign.
We can use this observation in analyzing the linearity of a set of points \((x_i, y_i)\). First, let \(\overline{x}\) be the mean of the \(x_i\) and let \(\overline{y}\) be the mean of the \(y_i\). For a given point \((x_i, y_i)\), the distance from the mean on the \(x\)-axis is \(\Delta x = x_i - \overline{x}\), and the distance from the mean on the \(y\)-axis is \(\Delta y = y_i - \overline{y}\). Now look at the following summation:

\[
\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{N} \Delta x \Delta y = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})
\]  

(this quantity \(\sigma_{xy}\) is known as the covariance). If \(\Delta x\) and \(\Delta y\) always have the same sign, then \(\sigma_{xy}\) will be strongly positive, while if \(\Delta x\) and \(\Delta y\) always have opposite sign, \(\sigma_{xy}\) will be strongly negative. If \(\Delta x\) and \(\Delta y\) don’t have any particular linear relationship, then the terms \(\Delta x \Delta y\) in the sum are as likely to be positive as negative, and \(\sigma_{xy}\) will be close to zero. Based on these observations, and on what we know about lines, we can say:

- If \(\sigma_{xy}\) for a set of points \((x_i, y_i)\) is strongly positive, then the points \((x_i, y_i)\) probably look something like a line with positive slope.
- If \(\sigma_{xy}\) for a set of points is strongly negative, then the points \((x_i, y_i)\) probably look something like a line with positive slope.
- If \(\sigma_{xy}\) for a set of points is close to zero, then the points \((x_i, y_i)\) probably don’t look anything like a line.

The quantity \(\sigma_{xy}\) is thus a pretty useful measure of linearity, but it would be nice if we had a more concrete measure of what ‘strongly negative’ or ‘strongly positive’ or ‘close to zero’ means. This amounts to getting a handle on the magnitudes of the \(\Delta x\) and \(\Delta y\) terms. We already have a measure for these, namely the standard deviations:

\[
\sigma_x = \left[ \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2 \right]^{1/2}
\]
\[
\sigma_y = \left[ \frac{1}{N} \sum_{i=1}^{N} (y_i - \overline{y})^2 \right]^{1/2}
\]

(2)

Instead of using the raw \(\Delta x\) and \(\Delta y\) to analyze the linearity of a set of points, then, we can use \(\Delta x/\sigma_x\) and \(\Delta y/\sigma_y\). The equivalent form of the covariance is then

\[
\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}}
\]  

(3)

which is the correlation coefficient. This quantity gives us some of the same insight as the covariance, \(\sigma_{xy}\) (in fact, you can see that it’s just the covariance scaled by the standard deviations of \(x\) and \(y\)). For example, if two variables don’t have any discernible linear relation, \(r\) will still be near zero. When \(x\) and \(y\) do have a linear relation, however, then the value of \(r\) can be interpreted more objectively than the value of \(\sigma_{xy}\).

The first thing to notice is that \(|r| \leq 1\) (we won’t prove this; take my word for it). Now suppose that the points \((x_i, y_i)\) all lie exactly on a line \(y = A + B x\). Then \(y_i = A + B x_i\) for all \(i\), and \(\overline{y} = A + B \overline{x}\), which lets us write

\[
y_i - \overline{y} = B(x_i - \overline{x}),
\]
and plugging this into Eq. 3, we get

\[
r = \frac{B \sum (x_i - \bar{x})^2}{\sqrt{\sum (x_i - \bar{x})^2 B^2 \sum (x_i - \bar{x})^2}} = \frac{B}{|B|}.
\]

We can therefore say the following about the correlation coefficient.

- If \( r = +1 \), then the points lie perfectly on a line with positive slope. Any positive \( r \) means that \( x \) and \( y \) are positively correlated.
- If \( r = -1 \), then the points lie perfectly on a line with negative slope. Any negative \( r \) means that \( x \) and \( y \) are negatively correlated.
- If \( r = 0 \), then \( x \) and \( y \) are referred to as uncorrelated.

Whether, and in what sense, two variables \( x \) and \( y \) are correlated thus depends on whether \( r \) is close to \( \pm 1 \) or close to zero.

Let’s look at the four \((x_i, y_i)\) data points from the last lecture. Using Eq. 3, we find that \( r = 0.98 \). This indicates that the data are close to lying perfectly on a line with positive slope, which is what we saw visually. On the other hand, the five \((x_i, y_i)\) points from the start of this lecture have a correlation value \( r = 0.65 \). This indicates that the points are positively correlated, meaning that increasing \( x_i \) tends to go with increasing \( y_i \) (ditto for decreasing \( x_i \) and \( y_i \)), but that the points aren’t that close to a single line, which is again what Fig. 1 shows. Sometimes, people consider \( r^2 \) instead of just \( r \), which no longer tells you whether points are positively or negatively correlated, but accentuates differences in the numerical correlation values. For example, \( r \) values of 0.98 and 0.65 become \( r^2 \) values of 0.96 and 0.42, which makes it easier to see that the five points in Fig. 1 are far from being linear, particularly in comparison with the four data points from the last lecture.

(The term ‘correlation’ is used because the common-sensical explanation of correlated variables is that they ‘move together’, in this case linearly. This explanation is easily understood by inspection of the equations for \( \sigma_{xy} \) and \( r \), or by looking at plots of data that are highly correlated.)