Correlation and Regression

Theme

Part 1: Sign?

The sign of s_{XY}, r_{XY} and $m_{Y,X}$ share all the same properties since the only difference between them is the different scaling via the standard deviations and they are always positive.

Part 2: Check the Slope Equation

```
In[129]:= data = {{2, 20}, {2, 18}, {3, 32}, {5, 40}, {8, 60}};
          data // MatrixForm

Out[130]//MatrixForm=
          ( 2  20 )
          ( 2  18 )
          ( 3  32 )
          ( 5  40 )
          ( 8  60 )

In[131]:= {μx, μy} = Mean[data]

Out[131]= {4, 34}

In[132]:= lm = LinearModelFit[data, x, x]

Out[132]= FittedModel[ 7.53846 + 6.61538 x ]

In[133]:= slope = lm["BestFitParameters"][[2]]

Out[133]= 6.61538

In[134]:= {σx, σy} = StandardDeviation[data] // N

Out[134]= {2.54951, 17.088}

In[135]:= slope * σx
                  ──
                  σy

Out[135]= 0.987007

In[136]:= Correlation[data] // N // MatrixForm

Out[136]//MatrixForm=
          (    1.      0.987007 )
          ( 0.987007      1.    )
```

The result is indeed the same.

Part 3: Slope

A higher value for the correlation coefficient does not necessarily indicate that the slope of the regression line is higher (it is only a measure of how strong the linear relationship is, not how this linear relationship looks like). The thing is that when the slope of the line changes, the standard deviations usually changes as well. E.g. if we increase the slope of the line, s_Y might increase and s_X decrease, so $\frac{s_X}{s_Y}$gets smaller and hence acts as an opposing force.

Part 4: Exact Linear Relationship

- When the points are moved vertically to the line, s_Y decreases since the points mover closer to the mean. When all points are exactly located on the line, we cannot decrease s_Y further. Note: the same argument may also be true for s_X but this is a bit trickier since then also the slope of the regression line changes. This is not true for s_Y since the regression line uses the (quadratic) distance between the points and the line.
- In the case of exact linear relationship, the relation $\frac{s_X}{s_Y}$ is exactly the inverse slope, i.e. m^{-1}.

Part 5: Correlation and Causality

The correlation coefficient says nothing about causality, i.e. whether more police really is the cause of the increase crime (as indicated). Usually, there are two main reasons why the relationship indicated by r_{XY} might not be true.

- Influence of different variables: there might be another variable which influences X and Y and hence results in a high correlation. Here, maybe there is an event where a lot of people attend so that we naturally need more police and we have naturally more crimes (probably not the case here, more likely is the next point).
- Backwards causality: it might be that not X influences Y but the other way around. Here, this would mean that due to higher crime rates we need more police (this is definately a problem here).

Code