6.3: Correlation

Correlation is closely related to covariance. In essence, correlation standardizes covariance so it can be compared across variables. Correlation is represented by a correlation coefficient, ρρ, and is calculated by dividing the covariance of the two variables by the product of their standard deviations. For populations it is expressed as:

ρ=cov(X,Y)σxσy(6.4)(6.4)ρ=cov(X,Y)σxσy

For samples it is expressed as:

r=∑(X−¯X)(Y−¯Y)/(n−1)sxsy(6.5)(6.5)r=∑(X−X¯)(Y−Y¯)/(n−1)sxsy

Like covariance, correlations can be positive, negative, and zero. The possible values of the correlation coefficient rr, range from -1, perfect negative relationship to 1, perfect positive relationship. If r=0r=0, that indicates no correlation. Correlations can be calculated in `R`, using the `cor` function.

```
ds %>% dplyr::select(education, ideol, age, glbcc_risk) %>% na.omit() %>%
  cor()
```

```
##              education       ideol         age  glbcc_risk
## education   1.00000000 -0.13246843 -0.06149090  0.09115774
## ideol      -0.13246843  1.00000000  0.08991177 -0.59009431
## age        -0.06149090  0.08991177  1.00000000 -0.07514098
## glbcc_risk  0.09115774 -0.59009431 -0.07514098  1.00000000
```

Note that each variable is perfectly (and positively) correlated with itself - naturally! Age is slightly and surprisingly negatively correlated with education (-0.06) and unsurprisingly positively correlated with political ideology (+0.09). What this means is that, in this dataset and on average, older people are slightly less educated and more conservative than younger people. Now notice the correlation coefficient for the relationship between ideology and perceived risk of climate change (glbcc_risk). This correlation (-0.59) indicates that on average, the more conservative the individual is, the less risky climate change is perceived to be.