Correlation Coefficients

Computing covariance & correlation

Jenine Harris
Brown School

Exploring the data

- Importing the data using the `here()` function

```r
# import the water data
water.educ <- read.csv("/Users/harrisj/Box/teaching/Teaching/Fall2020/da

# examine the data
summary(object = water.educ)
```

```
##    country              med.age        perc.1dollar   perc.basic2015sani
##  Length:97          Min.   :15.00   Min.   : 1.00   Min.   :  7.00
##  Class :character   1st Qu.:22.50   1st Qu.: 1.00   1st Qu.: 73.00
##  Mode  :character   Median :29.70   Median : 1.65   Median : 93.00
##                     Mean   :30.33   Mean   :13.63   Mean   : 79.73
##                     3rd Qu.:39.00   3rd Qu.:17.12   3rd Qu.: 99.00
##                     Max.   :45.90   Max.   :83.80   Max.   :100.00
##                                     NA's   :33
##  perc.safe2015sani perc.basic2015water perc.safe2015water perc.in.school
##  Min.   :  9.00    Min.   : 19.00      Min.   : 11.00     Min.   :33.32
##  1st Qu.: 61.25    1st Qu.: 88.75      1st Qu.: 73.75     1st Qu.:83.24
##  Median : 76.50    Median : 97.00      Median : 94.00     Median :92.02
##  Mean   : 71.50    Mean   : 90.16      Mean   : 83.38     Mean   :87.02
##  3rd Qu.: 93.00    3rd Qu.:100.00      3rd Qu.: 98.00     3rd Qu.:95.81
##  Max.   :100.00    Max.   :100.00      Max.   :100.00     Max.   :99.44
##  NA's   :47        NA's   :1           NA's   :45
##  female.in.school male.in.school
##  Min.   :27.86    Min.   :38.66
##  1st Qu.:83.70    1st Qu.:82.68
```

Codebook

Definitions of the variables:

- country: the name of the country
- med.age: the median age of the citizens in the country
- perc.1dollar: percentage of citizens living on $1 per day or less
- perc.basic2015sani: percentage of citizens with basic sanitation access
- perc.safe2015sani: percentage of citizens with safe sanitation access
- perc.basic2015water: percentage of citizens with basic water access
- perc.safe2015water: percentage of citizens with safe water access
- perc.in.school: percentage of school-age people in primary and secondary school
- female.in.school: percentage of female school-age people in primary and secondary school
- male.in.school: percentage of male school-age people in primary and secondary school

The data were all from 2015.

Computing and interpreting the covariance between two variables

- The relationship between two variables can be checked in a few different ways.
- One method for measuring this relationship is **covariance**, which quantifies whether two variables vary together (co-vary).
- $cov_{xy} = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{n-1}$
- The equations shows the summation from the first observation in the data, i = 1, to the last observation in the data set, n.
- The sum is of the product of (1) the difference between each individual observation value for the first variable x_i and the mean of that variable m_x and the same thing for the second variable, y.
- The numerator adds up how far each observation is away from the mean values of the two variables being examined, so this ends up being a very large number quantifying how far away all the observations are from the mean values.
- The denominator divides this by the Bessel correction of n - 1, which is close to the sample size and essentially finds the average deviation from the means for each observation.

Interpreting the covariance

- If the numerator is positive, the covariance will be positive, representing a positive relationship between two variables.

- This happens when many of the observations have x and y values that are either:

  - both higher values than the mean, or
  - both lower than the mean

- When x_i and y_i are **both** greater than m_x and m_y, respectively, the contribution of that observation to the numerator is a positive amount.

- Likewise, when x_i and y_i are **both** less than m_x and m_y, respectively, the contribution of that observation to the numerator is also a positive amount because multiplying two negatives results in a positive.

Visualizing the covariance

- A graph showing the means of x and y and highlighting the points that were either above or below m_x and m_y can help.
- There are a lot more points above m_x and m_y than below, which was consistent with the positive value of the covariance.
- The observations with x and y values both above or below the means contribute positive amounts to the sum in the numerator, while the other observations contributed negative amounts to the sum in the numerator.
- Since there were so many more positive contributing data points in the figure, the sum was positive and the covariance was positive.

Negative values in covariance

- Likewise, if there were more negative values contributed to the numerator, the covariance is likely to be negative.

Computing covariance in R

- Females in school and basic water access appeared to have a positive relationship while poverty and basic water access had a negative relationship; the covariance can help quantify it.

- Rather than `drop_na()`, use `use = "complete"` to compute the covariance on the complete cases only.

```
# covariance of females in school, poverty, and
# percentage with basic access to drinking water
water.educ %>%
  summarize(cov.females.water = cov(x = perc.basic2015water,
                                    y = female.in.school,
                                    use = "complete"),
            cov.poverty.water = cov(x = perc.basic2015water,
                                    y = perc.1dollar,
                                    use = "complete"))
```

```
##   cov.females.water cov.poverty.water
## 1           194.027         -261.2131
```

Why not use drop_na?

- Use the `drop_na()` for all three variables first and then used `cov()` without the `use = "complete"` option.

```
# covariance of females in school, poverty, and
# percentage with basic access to drinking water
water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  drop_na(perc.1dollar) %>%
  summarize(cov.females.water = cov(x = perc.basic2015water,
                                    y = female.in.school),
            cov.poverty.water = cov(x = perc.basic2015water,
                                    y = perc.1dollar))
```

```
##   cov.females.water cov.poverty.water
## 1          162.2263         -261.2131
```

The perils of drop_na

- The `drop_na()` function dropped the `NA` *for all three variables* before computing the two covariances for the second coding option.
- The calculations using `use = "complete"` only dropped the `NA` from the two variables *in that specific calculation*.
- The version with the `drop_na()` is dropping some observations that could be used in each of the `cov()` calculations.
- If you prefer `drop_na()`, use it in two separate code chunks with each `cov()` function having `drop_na()` only for the relevant variables.

Using drop_na effectively for correlation

```
# covariance of females in school and
# percentage with basic access to drinking water
water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  summarize(cov.females.water = cov(x = perc.basic2015water,
                                    y = female.in.school))
```

```
##   cov.females.water
## 1           194.027
```

```
# covariance of poverty and
# percentage with basic access to drinking water
water.educ %>%
  drop_na(perc.basic2015water) %>%
  drop_na(perc.1dollar) %>%
  summarize(cov.poverty.water = cov(x = perc.basic2015water,
                                    y = perc.1dollar))
```

```
##   cov.poverty.water
## 1         -261.2131
```

Covariance is less useful than correlation

- The covariance does not have a useful inherent meaning; it is not a percentage or a sum or a difference.
- The size of the covariance depends largely on the size of what is measured.
    - For example, something measured in millions might have a covariance in the millions or hundreds of thousands.
- The value of the covariance indicates whether there is a relationship at all and the direction of the relationship---that is, whether the relationship is positive or negative.
- In this case, a non-zero value indicates that there is some relationship and the positive value indicates the relationship is positive.

Computing the Pearson's r correlation between two variables

- The covariance is not reported very often to quantify the relationship between two continuous variables.
- Instead the covariance is **standardized** by dividing it by the standard deviations of the two variables involved.
- The result is called the correlation coefficient and is referred to as r
  - $r_{xy} = \frac{cov_{xy}}{s_x s_y}$

Interpreting the direction of Pearson's r

- *Negative correlations* are when one variable goes up, the other goes down
- *No correlation* is when there is no discernable pattern in how two variables vary
- *Positive correlations* are when one variable goes up, the other also goes up (or when one goes down the other does too); both variables move together in the same direction

Graphing the correlation with a line

- To add a line to a scatterplot, add a `geom_smooth()` layer.
- The first argument is `method =` which is the method used for drawing the line.
    - In this case, use the `lm` method, with `lm` standing for **linear model**.
- The legend is getting more complicated with two different types of symbols, points and lines.
    - The legend is generated from attributes included in the `aes()` argument and that different symbols can be generated by using different attributes.
    - In this case, use the `color =` attribute for the points and the `linetype =` attribute for the lines.

```
# explore plot of female education and water
water.educ %>%
  ggplot(aes(y = female.in.school/100, x = perc.basic2015water/100)) +
  geom_smooth(method = "lm", se = FALSE, aes(linetype = "Fit line"), col
  geom_point(size = 2, aes(color = "Country"), alpha = .6) +
  theme_minimal() +
  labs(y = "Percent of school-aged females in school",
       x = "Percent with basic water access") +
  scale_x_continuous(labels = scales::percent) +
```

Graphing the correlation with a line

- To add a line to a scatterplot, add a `geom_smooth()` layer.
- The first argument is `method =` which is the method used for drawing the line.
  - In this case, use the `lm` method, with `lm` standing for **linear model**.
- The legend is getting more complicated with two different types of symbols, points and lines.
  - The legend is generated from attributes included in the `aes()` argument and that different symbols can be generated by using different attributes.
  - In this case, use the `color =` attribute for the points and the `linetype =` attribute for the lines.

Correlation with tidyverse

```
# correlation between water access and female education
water.educ %>%
  summarize(cor.females.water = cor(x = perc.basic2015water,
                                    y = female.in.school,
                                    use = "complete"))
```

```
##   cor.females.water
## 1         0.8086651
```

- Interpretation: The Pearson's correlation coefficient demonstrated that the percentage of females in school is positively correlated with the percentage of citizens with basic access to drinking water ($r = 0.81$). Essentially, as access to water goes up, the percentage of females in school also increases in countries.

Interpreting the strength of the Pearson's product-moment correlation coefficient

- r is not only positive, but it also shows a very strong relationship.
- Most values describing the strength of r are similar to these:
    - r = -1.0 is perfectly negative
    - r = -.8 is strongly negative
    - r = -.5 is moderately negative
    - r = -.2 is weakly negative
    - r = 0 is no relationship
    - r = .2 is weakly positive
    - r = .5 is moderately positive
    - r = .8 is strongly positive
    - r = 1.0 is perfectly positive