Gov 51: Summarizing Bivariate Relationships: Cross-tabs, Scatterplots, and Correlation

Matthew Blackwell

Harvard University

Effect of assassination attempts

```r
leaders <- read.csv("data/leaders.csv")
head(leaders[, 1:7])
```

```
##   year     country       leadername age politybefore
## 1 1929 Afghanistan Habibullah Ghazi  39           -6
## 2 1933 Afghanistan       Nadir Shah  53           -6
## 3 1934 Afghanistan      Hashim Khan  50           -6
## 4 1924     Albania             Zogu  29            0
## 5 1931     Albania             Zogu  36           -9
## 6 1968     Algeria      Boumedienne  41           -9
##   polityafter interwarbefore
## 1       -6.00              0
## 2       -7.33              0
## 3       -8.00              0
## 4       -9.00              0
## 5       -9.00              0
## 6       -9.00              0
```

Contingency tables

- With two categorical variables, we can create **contingency tables**.
  - Also known as **cross-tabs**
  - Rows are the values of one variable, columns the other.

```r
table(Before = leaders$civilwarbefore,
      After = leaders$civilwarafter)
```

```
##       After
## Before   0   1
##      0 177  19
##      1  27  27
```

- Quick summary how the two variables "go together."

Cross-tabs with proportions

- Use the `prop.table()` for proportions:

```
prop.table(table(Before = leaders$civilwarbefore,
                 After = leaders$civilwarafter))
```

```
##       After
## Before     0     1
##      0 0.708 0.076
##      1 0.108 0.108
```

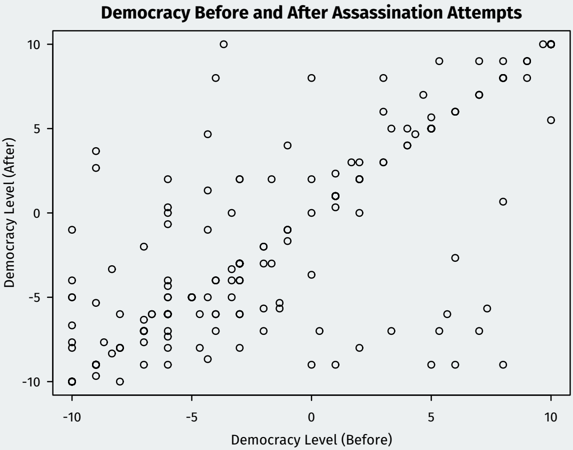
- We can also ask R to calculate proportions within each row:

```
prop.table(table(Before = leaders$civilwarbefore,
                 After = leaders$civilwarafter),
           margin = 1)
```

```
##       After
## Before      0      1
##      0 0.9031 0.0969
##      1 0.5000 0.5000
```

Scatterplot

- Direct graphical comparison of two continuous variables.



How to create a scatterplot

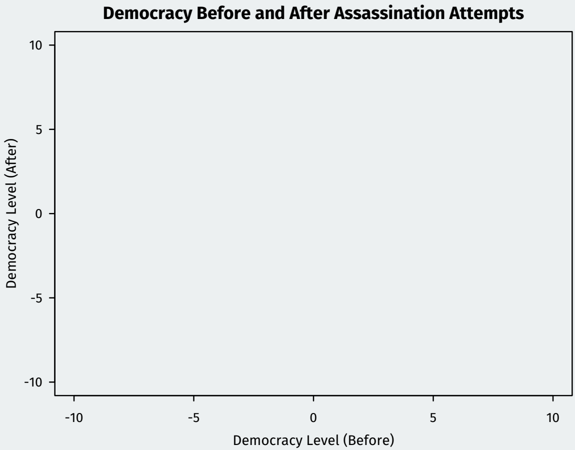
- Each point on the scatterplot (x_i, y_i)
- Use the `plot()` function

```
plot(x = leaders$politybefore, y = leaders$polityafter,
     xlab = "Democracy Level (Before)",
     ylab = "Democracy Level (After)",
     main = "Democracy Before and After Assassination Attempts")
```

Scatterplot

```r
leaders[1, c("politybefore", "polityafter")]
```

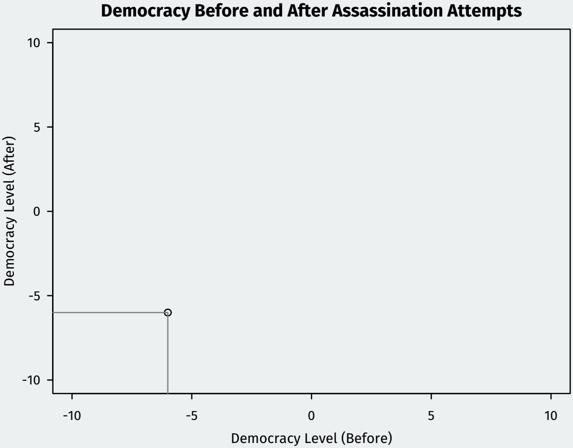
```
##   politybefore polityafter
## 1           -6          -6
```



Scatterplot

```r
leaders[1, c("politybefore", "polityafter")]
```

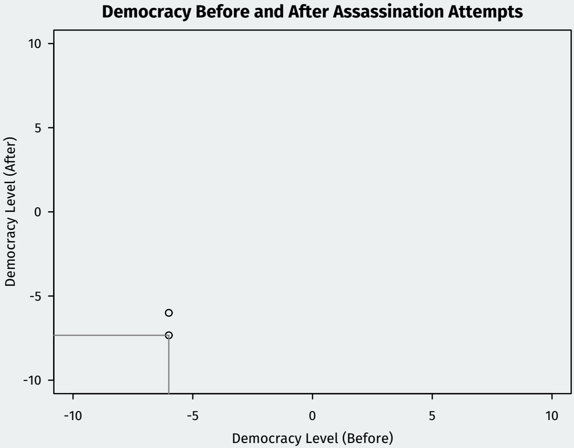
```
##   politybefore polityafter
## 1           -6          -6
```



Scatterplot

```r
leaders[2, c("politybefore", "polityafter")]
```

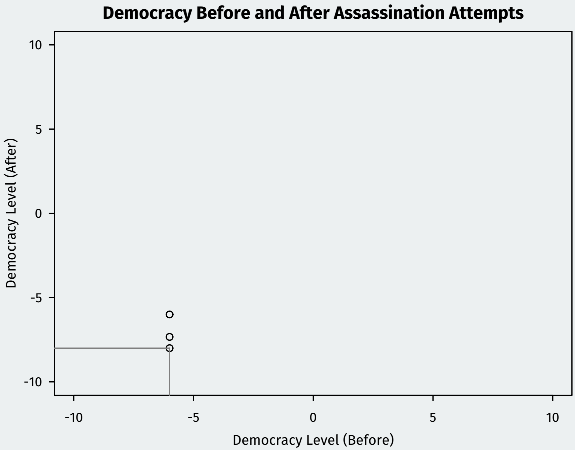
```
##   politybefore polityafter
## 2           -6       -7.33
```



Scatterplot

```r
leaders[3, c("politybefore", "polityafter")]
```

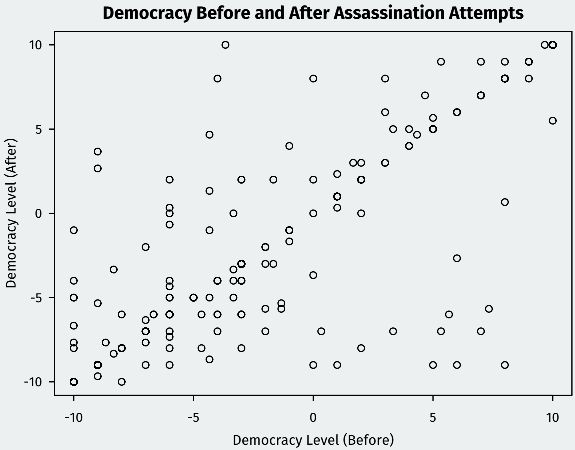
```
##   politybefore polityafter
## 3           -6          -8
```



Scatterplot

```
leaders[3, c("politybefore", "polityafter")]
```

```
##   politybefore polityafter
## 3           -6          -8
```



How big is big?

- Would be nice to have a standard summary of how similar variables are.
  - Problem: variables on different scales!
  - Need a way to put any variable on common units.
- **z-score** to the rescue!

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

- Crucial property: z-scores don't depend on units

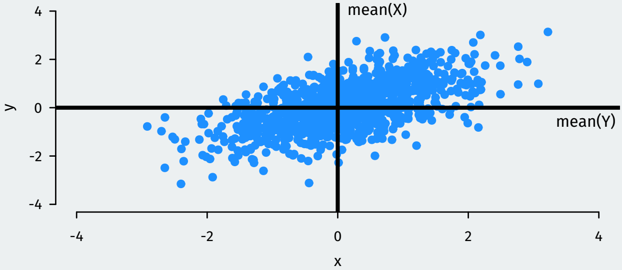
$$\text{z-score of } (ax_i + b) = \text{z-score of } x_i$$

Correlation

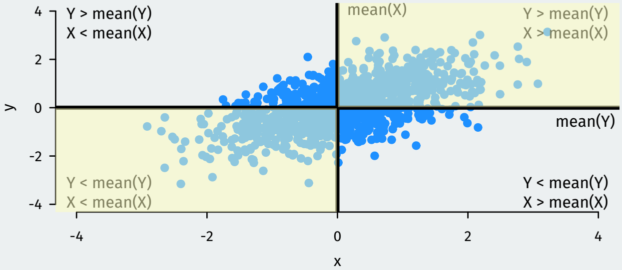
- How do variables move together on average?
- When x_i is big, what is y_i likely to be?
  - Positive correlation: when x_i is big, y_i is also big
  - Negative correlation: when x_i is big, y_i is small
  - High magnitude of correlation: data cluster tightly around a line.
- The technical definition of the **correlation coefficient**:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left[(\text{z-score for } x_i) \times (\text{z-score for } y_i)\right]$$

Correlation intuition

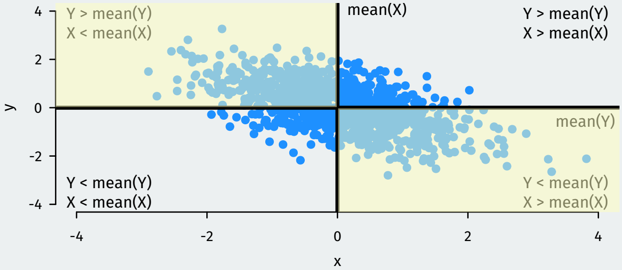


Correlation intuition



- Large values of X tend to occur with large values of Y:
  - $(\text{z-score for } x_i) \times (\text{z-score for } y_i) = (\text{pos. num.}) \times (\text{pos. num}) = +$
- Small values of X tend to occur with small values of Y:
  - $(\text{z-score for } x_i) \times (\text{z-score for } y_i) = (\text{neg. num.}) \times (\text{neg. num}) = +$
- If these dominate $\rightsquigarrow$ positive correlation.

Correlation intuition



- Large values of X tend to occur with small values of Y:
  - $(\text{z-score for } x_i) \times (\text{z-score for } y_i) = (\text{pos. num.}) \times (\text{neg. num}) = -$
- Small values of X tend to occur with large values of Y:
  - $(\text{z-score for } x_i) \times (\text{z-score for } y_i) = (\text{neg. num.}) \times (\text{pos. num}) = -$
- If these dominate $\rightsquigarrow$ negative correlation.

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
  - Correlation is between -1 and 1
  - Correlation of 0 means no linear association.
  - Positive correlations $\rightsquigarrow$ positive associations.
  - Negative correlations $\rightsquigarrow$ negative associations.
  - Closer to -1 or 1 means stronger association.
- Order doesn't matter: `cor(x,y) = cor(y,x)`
- Not affected by changes of scale:
  - `cor(x,y) = cor(ax+b, cy+d)`
  - Celsius vs. Fahreneheit; dollars vs. pesos; cm vs. in.

Correlation in R

- Use the `cor()` function
- Missing values: set the `use = "pairwise"` $\rightsquigarrow$ available case analysis

```
cor(leaders$politybefore, leaders$polityafter,
    use = "pairwise")
```

```
## [1] 0.828
```

- Very highly correlation!