Correlation

Max Turgeon

DATA 2010–Tools and Techniques in Data Science

Lecture Objectives

- Explain the purpose of correlation analysis in data science
- Describe the difference between different measures of correlation
- Discuss the main differences between correlation and causation

Motivation

- In data science, we rarely look at one variable at a time.
  - Not just the distribution of GPAs, but the distribution for each major separately.
- In fact, we are mostly interested in the *relationship* between different variables.
  - GPA vs high-school grades
- **Correlation** is the main language we use to describe these relationships in statistics

Pearson correlation i

- This is the main measure of correlation.
- Given two samples $X_1, \ldots, X_n, Y_1, \ldots, Y_n$.
  - Same sample size
  - X_i, Y_i are measured on the **same** experimental/observational unit.
  - $\bar{X}$ is the sample mean of X_i's, $\bar{Y}$ is the sample mean of Y_i's

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}.$$

Pearson correlation ii

- Similarly, it can be defined in terms of the sample variances s_X^2, s_Y^2 and the sample covariance s_{XY}:

$$r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}.$$

- We have $r \in [-1, 1]$.
- Pearson correlation measures **linear relationships**, and it especially suited for normally distributed measurements.
  - We will have $r = 1$ if $Y_i = aX_i$ for $a > 0$, and similarly for $r = -1$.

Pearson correlation iii

    - The maximum value of r can be strictly less than 1 for some distributions.
- In particular, the Pearson correlation can change dramatically after transformation of the data.
    - It is also sensitive to outliers

Spearman correlation i

- Spearman correlation uses rank information.
    - Do X_i and Y_i have the same rank?
- Let $\mathrm{rank}(X_i)$ be the rank of X_i after having sorted the X_i's, and similarly for $\mathrm{rank}(Y_i)$.
    - So $\mathrm{rank}(X_i) \in 1, \ldots, n$
- If we let $d_i = \mathrm{rank}(X_i) - \mathrm{rank}(Y_i)$, we define the Spearman correlation as

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

Spearman correlation ii

- Just like Pearson correlation, we have $\rho \in [-1, 1]$.
    - **Exercise**: what ranks should we have to get $\rho = 1$? What about $\rho = -1$?
- Spearman measures **monotone relationships** and less sensitive to outliers.

Example i

- `divorce_margarine` contains data on divorce rates per year in Maine and margarine consumption per capita from 2000 to 2009

```
library(tidyverse)
library(dslabs)

# Rename variables
dataset <- divorce_margarine |>
  rename(div = divorce_rate_maine,
         marg = margarine_consumption_per_capita)
```

Example ii

```
# 1. Pearson correlation
dataset |>
  summarise(sxy = cov(div, marg),
            s2x = var(div),
            s2y = var(marg)) |>
  mutate(corr = sxy/sqrt(s2x * s2y))
```

```
##         sxy        s2x      s2y      corr
## 1 0.4326667 0.08844444 2.148444 0.9925585
```

Example iii

```r
# Or even simpler
dataset |>
  summarise(corr = cor(div, marg))
```

```
##        corr
## 1 0.9925585
```

```r
# 2. Spearman correlation
dataset |>
  summarise(corr = cor(div, marg, method = "spearman"))
```

Example iv

```
##          corr
## 1 0.9847319
```

Exercise

Calculate the Spearman correlation between divorce rate and margarine consumption using the definition. Use the `rank` function from the tidyverse.

Solution i

```
nobs <- nrow(dataset)
dataset |>
  mutate(div_rk = rank(div),
         marg_rk = rank(marg),
         diff = div_rk - marg_rk) |>
  summarise(corr = 1 - 6*sum(diff^2)/(nobs*(nobs^2 - 1))
```

```
##        corr
## 1 0.9848485
```

Solution ii

```r
# Also equal to Pearson corr. of ranks
dataset |>
  mutate(div_rk = rank(div),
         marg_rk = rank(marg)) |>
  summarise(corr = cor(div_rk, marg_rk))
```

```
##        corr
## 1 0.9847319
```

Auto-correlation i

- **Auto-correlation** is a main feature of time series data, i.e. measurements taken over time.
    - E.g. stock markets, temperatures, sales numbers
- By their nature, successive measurements tend to be correlated.
    - E.g. yesterday's temperature is correlated with today's
- **Periodicity** is also a common feature.
    - E.g. Sales tend to be higher on Saturdays, and around Christmas, etc.
- Auto-correlation is measured by taking the correlation between the time series and its lagged counterpart.

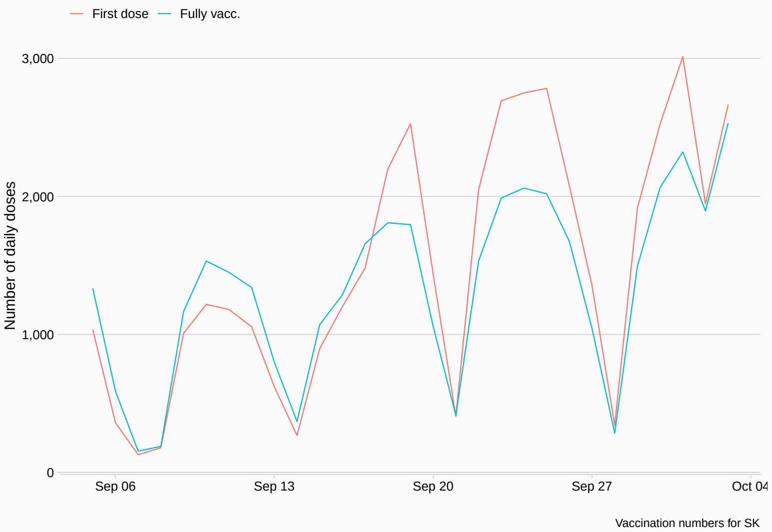
Auto-correlation ii

- Let $Y_t, t = 0, 1, 2, \ldots, T$ be a time series. The k-th auto-correlation is given by

$$r_k = \frac{\sum_{t=k+1}^{T}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}.$$

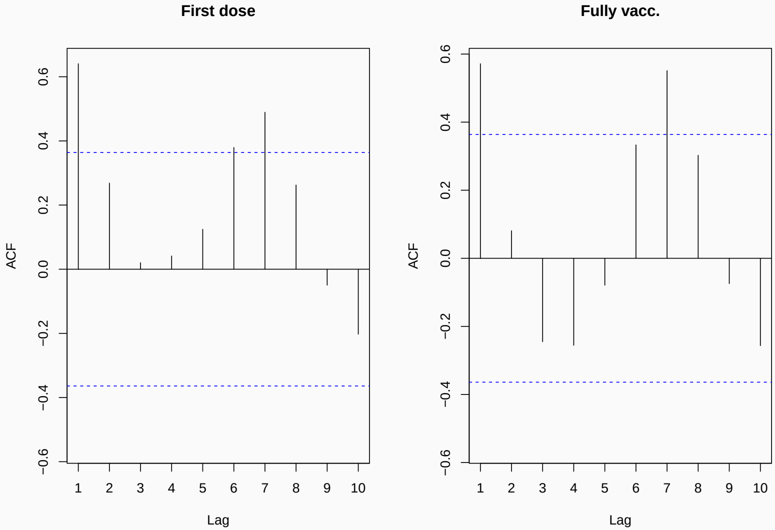
- Unexpectedly large autocorrelation can help identify periodicity in the data.

Example i

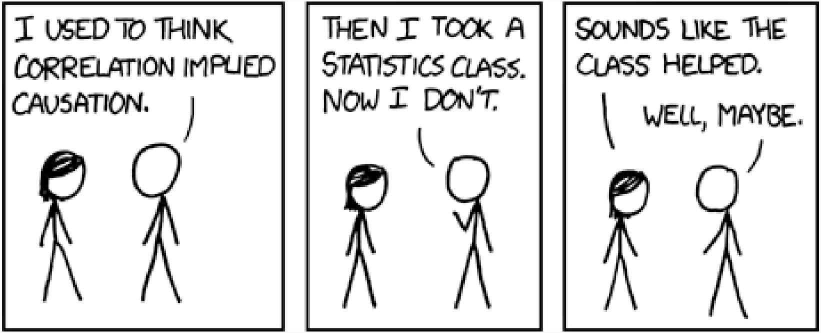


First dose
Fully vacc.
Number of daily doses
3,000
2,000
1,000
0
Sep 06
Sep 13
Sep 20
Sep 27
Oct 04
Vaccination numbers for SK

Example ii



Correlation vs Causation i



- Sometimes correlation happens without causation.
    - E.g. Ice cream sales and drownings
- Sometimes causation happens without correlation.

Correlation vs Causation ii

- E.g Pressing the gas pedal while going up a hill at constant speed.
- **Be careful about conclusions and language used**
- Causation can be inferred from correlation in some situations.
    - Randomized experimental designs.
    - Causal inference.