COVARIANCE, CORRELATION AND OUTLIERS

Applied math for data science (DSC 482/MTH 445) Fall 2022

Table of Contents

Figure 1: Photo from one of Milgram's experiments 1961-63, Yale U.

1. Covariance
2. Correlation
3. Outliers

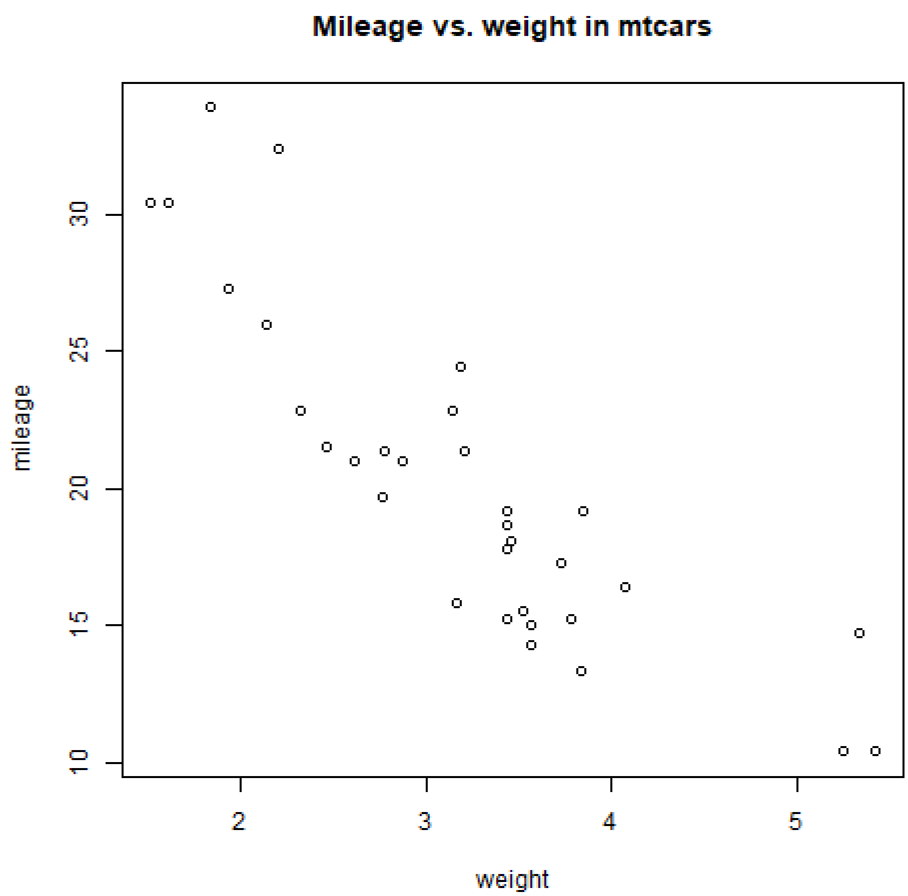
1 Measures of joint variability

- To identify trends, you can assess the *relationship* between two numeric variables: do they increase or decrease together, and how?
- Two linked measures are used to express this quality: covariance and correlation, quantifying degree and direction of joint change
- Their link is comparable to variance vs. standard deviation in the sense that the correlation coefficient is the more commonly used measure to (quickly) assess the relationship

2 Example: mileage and weight in `mtcars`

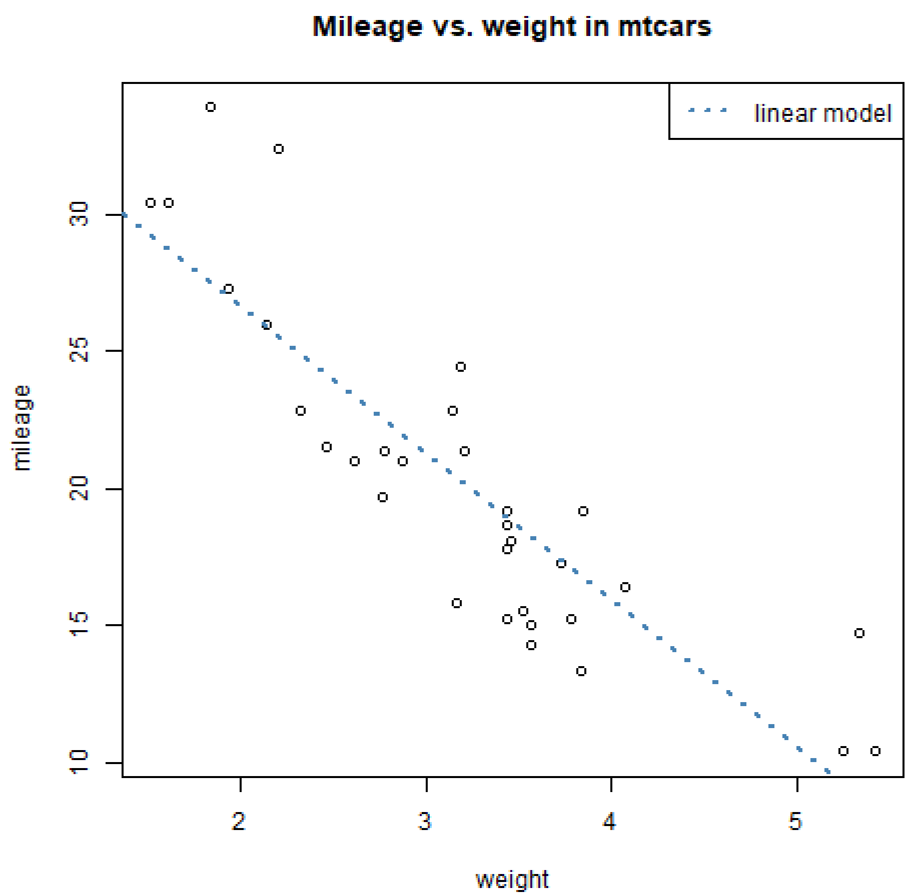
- Example: mileage (`mpg`) and weight (`wt`) in `mtcars`

```
plot(mtcars$mpg ~ mtcars$wt,
     xlab="weight",
     ylab="mileage",
     main="Mileage vs. weight in mtcars")
```



- We can easily fit a linear model `lm` through the sample:

```
plot(mtcars$mpg ~ mtcars$wt,
     xlab="weight",
     ylab="mileage",
     main="Mileage vs. weight in mtcars")
abline(lm(mpg~wt,data=mtcars),      # the model
       lty=3,lwd=2,col="steelblue") # layout details
legend(x="topright",                # plot legend
       legend="linear model",
       lty=3,lwd=2,col="steelblue")
```



- We can read some qualities of the joint change straight from the plot or the associated numbers:

```
lm(mpg~wt,data=mtcars)
```

```

Call:
lm(formula = mpg ~ wt, data = mtcars)

Coefficients:
(Intercept)           wt
      37.29        -5.34
```

3 Covariance

- The covariance expresses how much two variables change together and the nature of this change (positive or negative)
- Positive change means that both variables increase together
- Negative change means that both variables decrease together

- Covariance for a sample of `n` observations for two variables `x` and `y` in relation to the respective sample mean values[1]:

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Positive r_{xy} indicates a positive relationship
- Negative r_{xy} indicates a negative relationship
- $r_{xy}= 0$ indicates that there is no linear relationship
- Also, not that $r_{xy} \equiv r_{yx}$ i.e. the order of variables is irrelevant
- The variance is a special case of the covariance, in which the two variables are identical - i.e. it measures "covariance with itself"

- What is the unit of measurement of the covariance?

  If `x` was measured in u_x, and `y` in u_y, then the unit of measurement of the bivariate covariance of these variables would be $u_x * u_y$ - e.g. if we measure the joint variability of dollars and years, the covariance is measured in "dollar-years".

4 Example

- Let's go back to `xdata` and `ydata` considered before to illustrate measures of spread:

```
xdata <- c(2, 4.4, 3, 3, 2, 2.2, 2, 4)
ydata <- c(1, 4.4, 1, 3, 2, 2.2, 2, 7)
sd(xdata)  # small spread
sd(ydata)  # large spread
mean(xdata-ydata) # identical mean
```

```
[1] 0.9528
[1] 2.013
[1] 0
```

- Computing the sample covariance (`digits=4`):

$$\begin{aligned}
&\frac{(2-2.825)\times(1-2.285)+\ldots+(4-2.825)\times(7-2.825)}{7} \\
&= \frac{(-0.825)(-1.825)+\ldots+(1.175)(4.175)}{7} \\
&= \frac{10.355}{7} = 1.479
\end{aligned}$$

- `[ ]`

  Compute this using R "by hand":

```
m <- mean(xdata)
((2-m)*(1-m)+
 (4.4-m)*(4.4-m)+
 (3-m)*(1-m)+
 (3-m)*(3-m)+
 (2-m)*(2-m)+
 (2.2-m)*(2.2-m)+
 (2-m)*(2-m)+
 (4-m)*(7-m))/(length(xdata)-1)
```

```
[1] 1.479
```

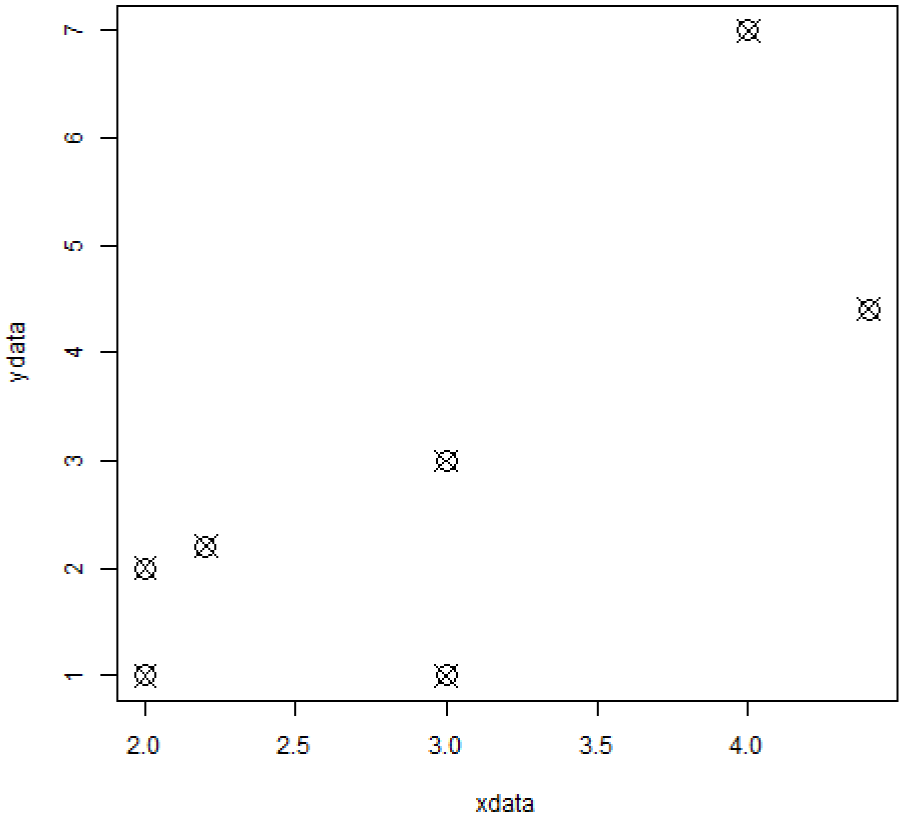
- Using the `cov` function:

```
options(digits=4)
cov(xdata,ydata)
```

```
[1] 1.479
```

- This suggests that there is a positive relationship based on the observations
- Plotting the vectors:

```
plot(ydata ~ xdata, pch=13, cex=2)
```



5 Correlation

- Correlation allows you to interpret the covariance further by identifying both *direction* and *strength* of any association
- Correlation measures association well under controlled conditions but it does not ever measure causation[2]
- The most common correlation coefficient is Pearson's product-moment correlation coefficient (the default in R) $\rho_{xy} \in (-1,1)$ computed with the respective standard deviations s_x and s_y:

$$\rho_{xy} = \frac{r_{xy}}{s_x s_y}$$

- When $\rho_{xy} = -1$ the relationship is perfectly negative
- The closer ρ_{xy} gets to 0, the weaker the relationship
- $\rho_{xy} = 0$ shows no relationship at all

- $\rho_{xy} = +1$ indicates a perfectly positive relationship
- Again, $\rho_{xy} \equiv \rho_{yx}$

- Computing $\rho_{xdata,ydata}$ by hand using $s_x = 0.953$ and $s_y = 2.013$:

```
cov(xdata,ydata)/(sd(xdata)*sd(ydata))
```

```
[1] 0.7714
```

- The result indicates a moderate to strong positive association between the observations in `xdata` and `ydata`

- Using the `cor` function:

```
cor(xdata,ydata)
```

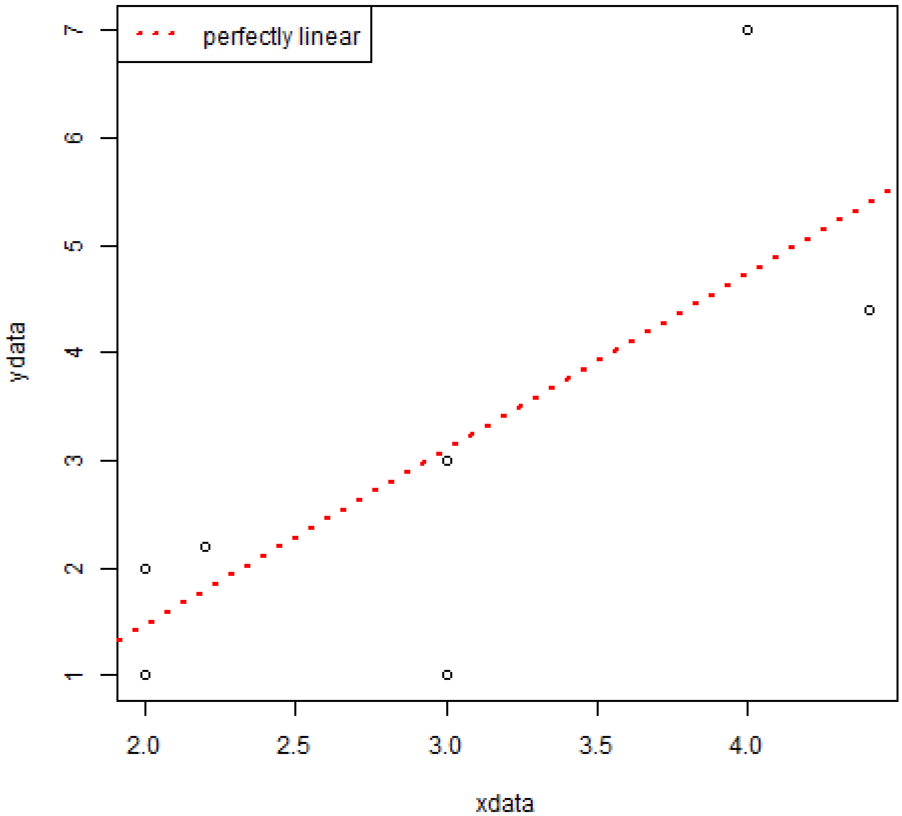
```
[1] 0.7714
```

- `[ ]` Check out the `help` for `cor` or `cov` (same vignette), and run the `example(cor)` programs

6 Checking the relationship with a linear model

- We can attempt to fit a line through the points using `lm`

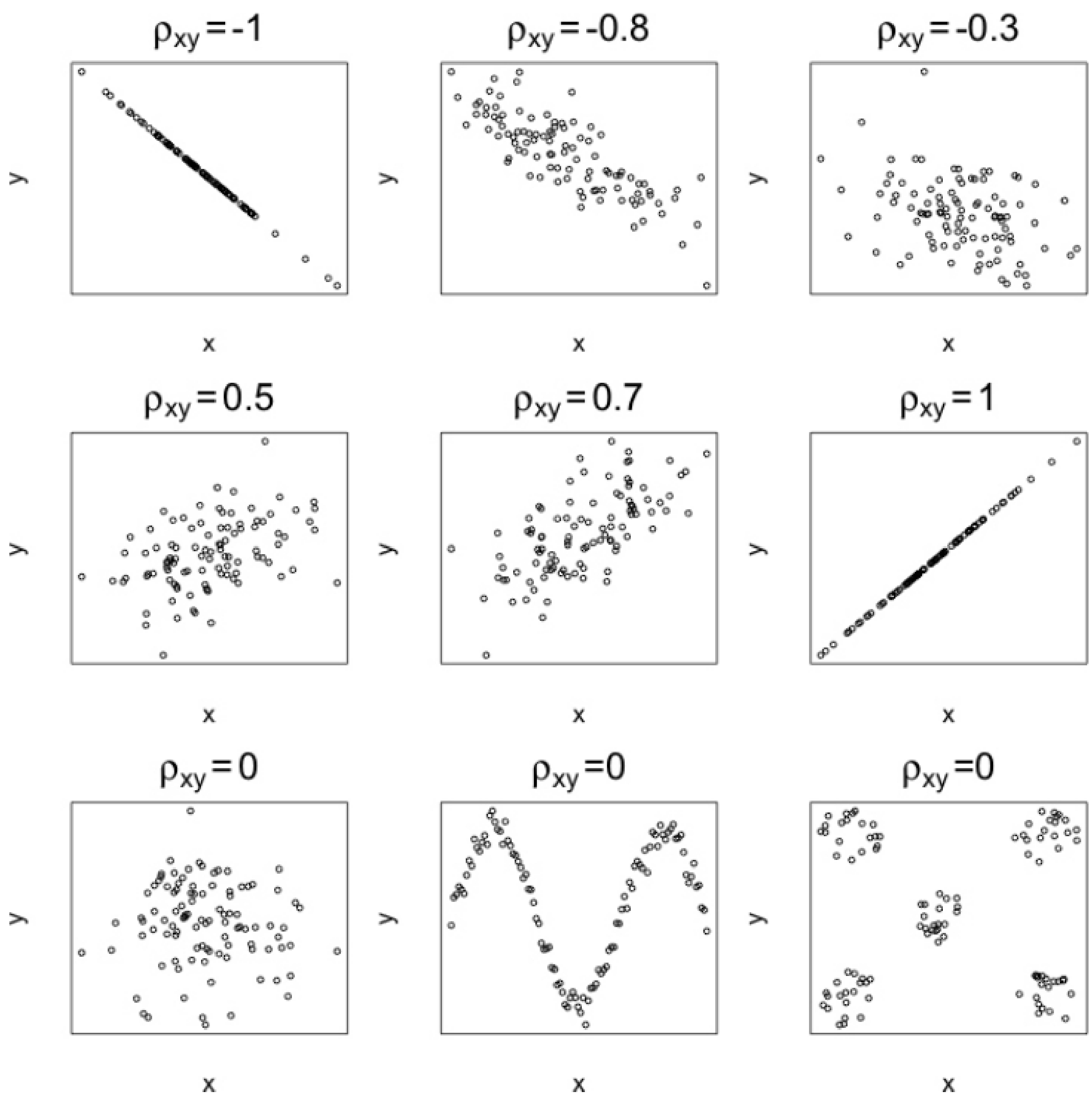
```
line <- lm(ydata ~ xdata)
plot(ydata ~ xdata)
abline(line, lty=3, lwd=2, col="red")
legend(x="topleft",
       legend="perfectly linear",
       lty=3,lwd=2,col="red")
```



- The correlation coefficient estimates the nature of the *linear* relationship between these variables: points closer to a perfect straight line have a value ρ_{xy} close to either -1 or 1.

7 Different values of ρ_{xy}

- The figure displays different scatterplots, each showing 100 points
- Observations have been generated randomly and artificially to follow the preset values of ρ_{xy} labeled above each plot



- The last row shows that Pearson's correlation coefficient can only detect linear relationships: the two last plots show distinct patterns but no linear correlation

8 Practice: `quakes`

OPEN YOUR ORG-MODE PRACTICE FILE IN EMACS AND START R

We are interested in the correlation between the number of detecting earthquake stations and the magnitude of earthquakes detected by them.

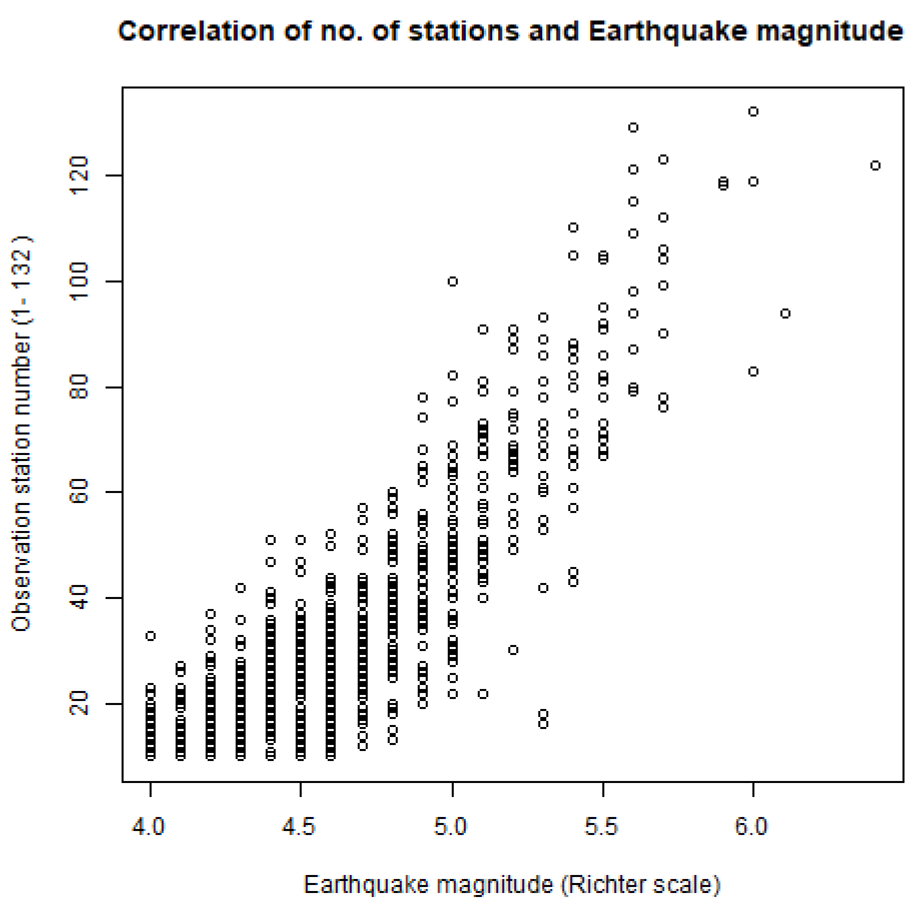
1. **Look** at the built-in `quakes` data set

```
str(quakes)
head(quakes)
stations <- quakes$stations
magnitudes <- quakes$mag
```

```
'data.frame':	1000 obs. of  5 variables:
 $ lat     : num  -20.4 -20.6 -26 -18 -20.4 ...
 $ long    : num  182 181 184 182 182 ...
 $ depth   : int  562 650 42 626 649 195 82 194 211 622 ...
 $ mag     : num  4.8 4.2 5.4 4.1 4 4 4.8 4.4 4.7 4.3 ...
 $ stations: int  41 15 43 19 11 12 43 15 35 19 ...
     lat  long depth mag stations
1 -20.42 181.6   562 4.8       41
2 -20.62 181.0   650 4.2       15
3 -26.00 184.1    42 5.4       43
4 -17.97 181.7   626 4.1       19
5 -20.42 182.0   649 4.0       11
6 -19.68 184.3   195 4.0       12
```

2. **Plot** the observation `stations` against the earthquake magnitude `mag`, label the axes using the `xlab` and `ylab` parameters, and `title` it.

```
plot(stations ~ magnitudes,
     xlab = "Earthquake magnitude (Richter scale)",
     ylab = paste("Observation station number (1-",
                  max(quakes$stations),")"))
title("Correlation of no. of stations and Earthquake magnitude")
```



3. What **insights** do you get from this plot?

   - What does a single point tell you?
   - What do vertical point groups mean?
   - What correlations can you see?
   - A single point corresponds to a pair of values: how many stations have detected an earthquake of a particular magnitude?
   - There are lot of points on top of one another: a single magnitude value seems to have been detected to different levels of precision (it's difficult to measure this exactly)
   - The plot shows a positive relationship: more stations tend to detect events of higher magnitude.

4. **Compute** the *covariance* of these two features.

```
cov(stations,magnitudes)
```

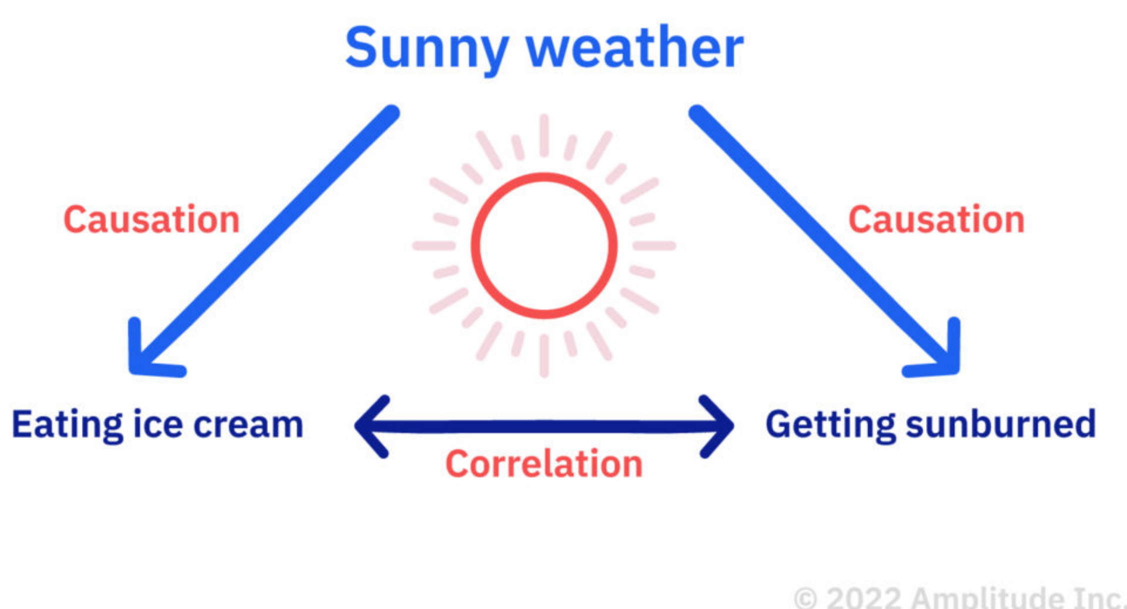
```
[1] 7.508
```

5. **Compute** Pearson's linear correlation coefficient.

```
cor(stations,magnitudes)
```

```
[1] 0.8512
```

9 Correlation and causation



- The correlation measures *association* not *causation*
- Causation is difficult to prove even in controlled situations
- If two variables are highly correlated, you only need one
- This "dimension reduction" is important in machine learning

10 The Strangeness of Outliers

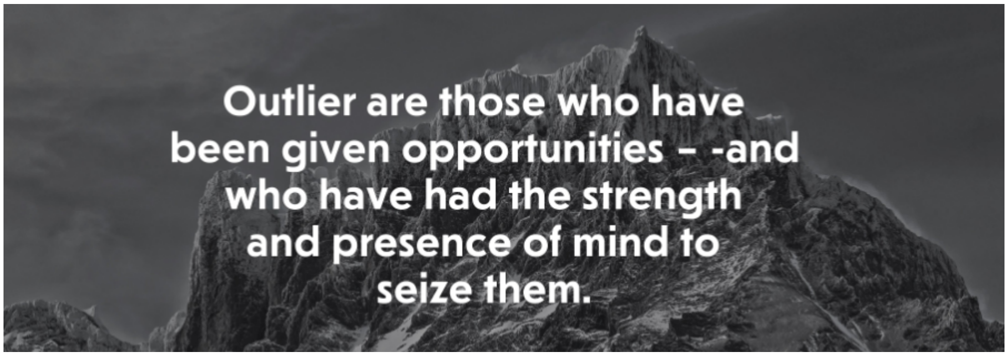
»People are strange When you're a stranger Faces look ugly When you're alone

Women seem wicked When you're unwanted Streets are uneven When you're down

When you're strange Faces come out of the rain When you're strange No one remembers your name When you're strange When you're strange When you're strange People are strange All right, yeah«

THE DOORS

11 Definition and example



- An *outlier* is an *anomalous* observation that does not appear to "fit" with the bulk of the data (and that may be hard to explain)
- Outliers correspond to extreme values but there is no numeric rule as to what constitutes an 'extreme' event
- Observing human "outliers" (or eccentrics) also leads to a lot of extreme qualitative values (what do you think they are?)
    1. extreme kindness
    2. extreme work ethics (+/-)
    3. extremely talented
    4. extreme intelligence
    5. extreme hoarder
    6. extreme height
    7. extreme wealth
    8. extreme occupation
    9. extremely funny
    10. extremely consistent
- **Whom do you know who would qualify as a "human outlier"? Why?**

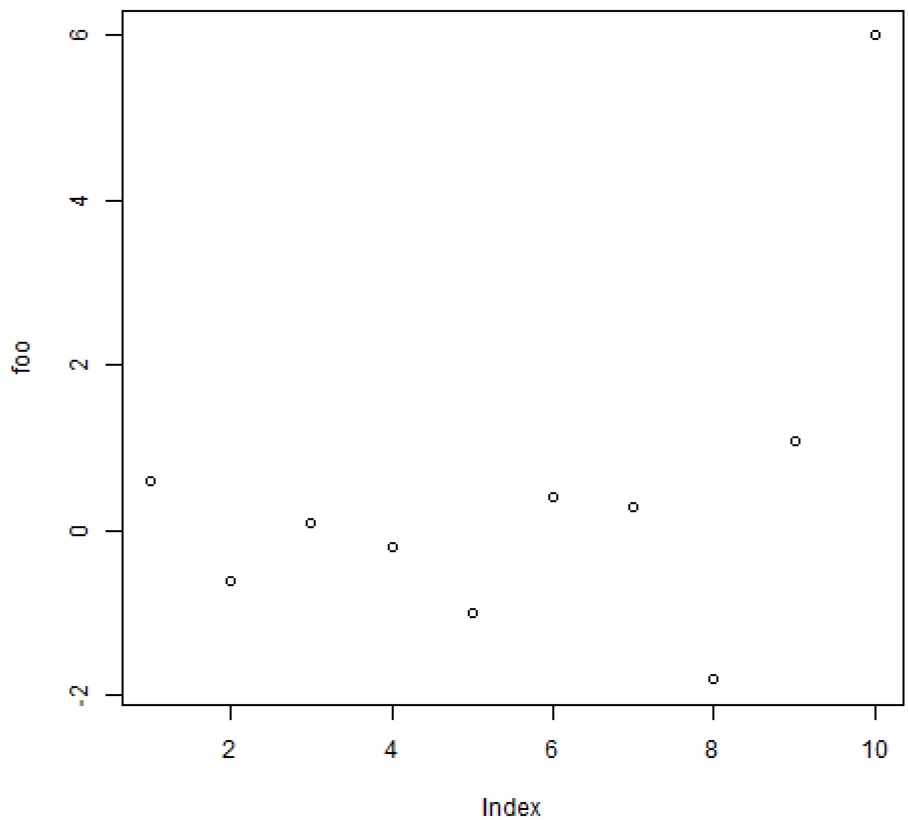
12 Univariate example

- Ten hypothetical data points

```
foo <- c(0.6, -0.6, 0.1, -0.2, -1.0, 0.4, 0.3, -1.8, 1.1, 6.0)
```

- Plot the points with `plot` - you can already see the outlier, but it's hard to see any clustering effects for univariate data unless they're printed on a line.

```
plot(foo)
```



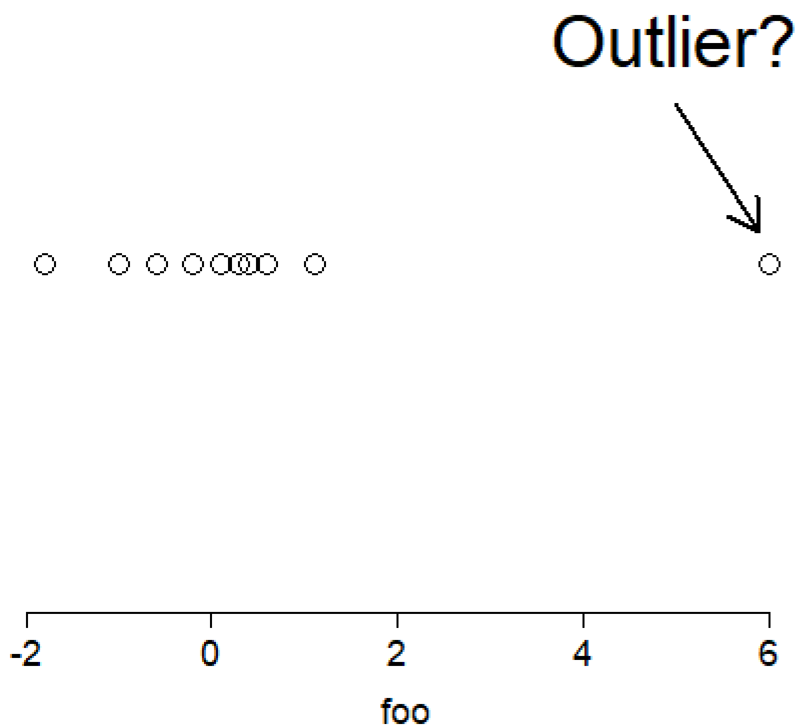
- Plot the points on a line to see the distribution more clearly: most of the observations are centered around 0 but one value is out at 6.

  Save the plot in the file `outlier1.png`.

```
plot(
  x = foo,        # univariate data
  y = rep(0,10),  # substitute 2nd dimension
  yaxt = "n",     # no y-axis
  ylab = "",      # no y-label
  bty = "n",      # no frame
  cex = 2,        # double point size
  cex.axis = 1.5, # increase axis and label size
  cex.lab = 1.5)

arrows(x0=5,y0=0.5,  # arrow starting point
       x1=5.9,y1=0.1, # arrow end point
       lwd=2)  # double line width

text(x=5,y=0.7, # location of textbox
     labels="Outlier?",
     cex=3)
```



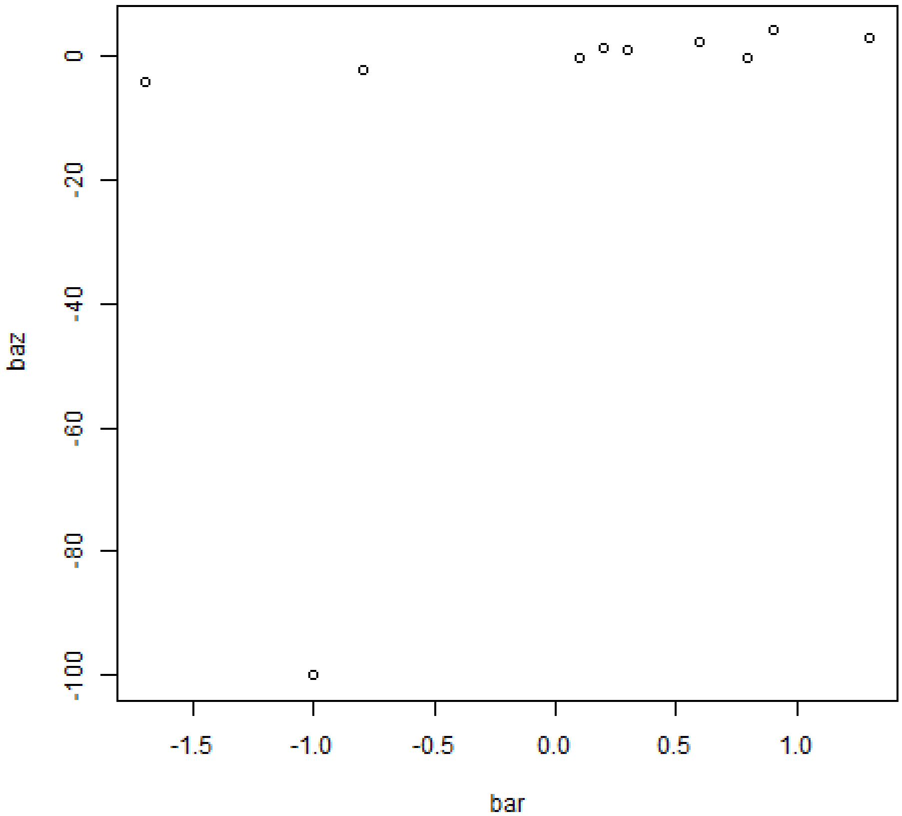
13 Bivariate example

- We define two more ten-element example vectors, `bar` and `baz`:

```
bar <- c(0.1, 0.3, 1.3, 0.6, 0.2, -1.7, 0.8, 0.9, -0.8, -1.0)
baz <- c(-0.3, 0.9, 2.8, 2.3, 1.2, -4.1, -0.4, 4.1, -2.3, -100.0)
```

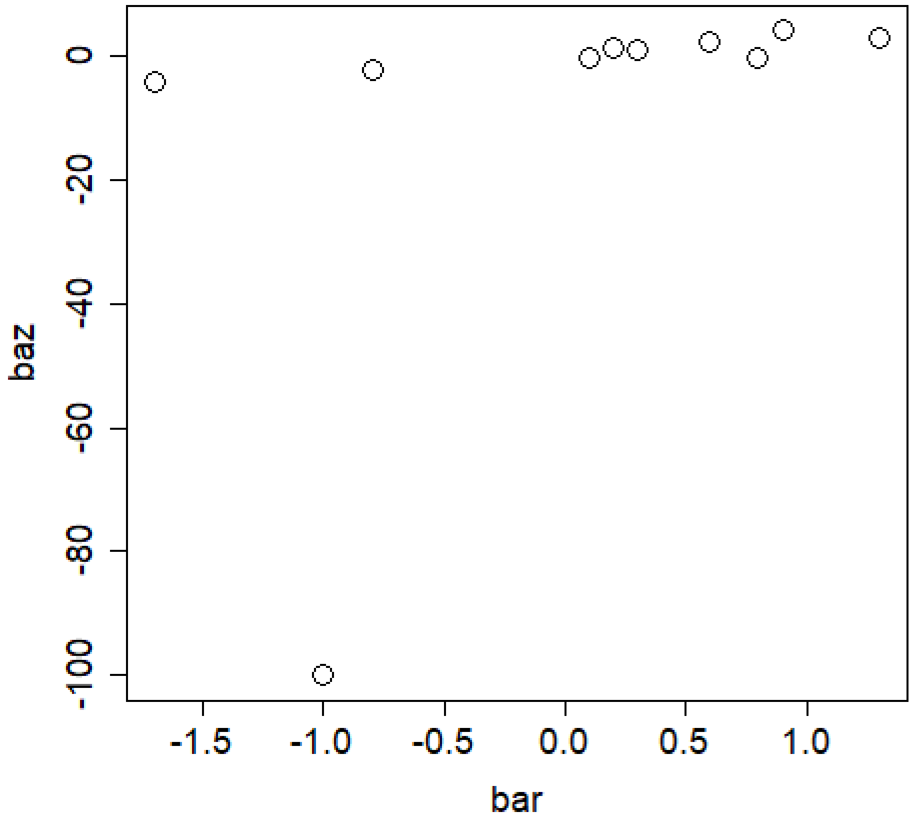
- Plot `x = bar` and `y = baz` without any customizations at first and save the plot in `outlier2.png`

```
plot(
  x = bar,
  y = baz)
```



- Add the commands to double the point size `cex` and increase axis and label size, `cex.axis` and `cex.lab` by `1.5`.

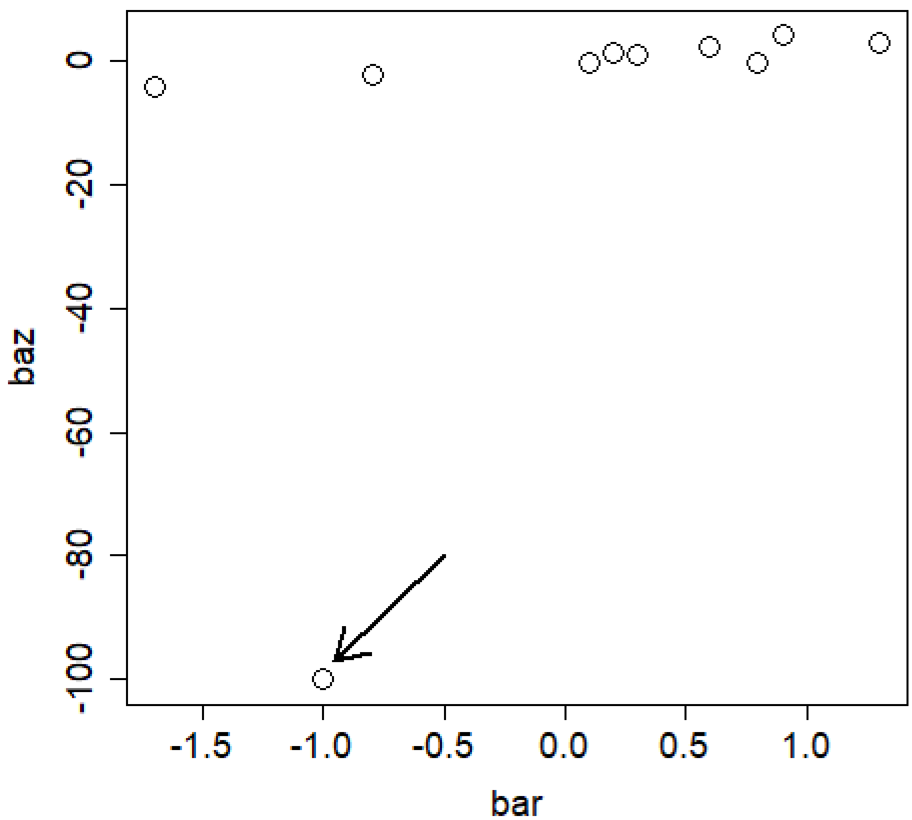
```
plot(
  x = bar,
  y = baz,
  cex = 2,
  cex.axis = 1.5,
  cex.lab = 1.5)
```



- Add an arrow that points at the outlier in the lower half of the plot using the function `arrows`. Double the line width `lwd`.

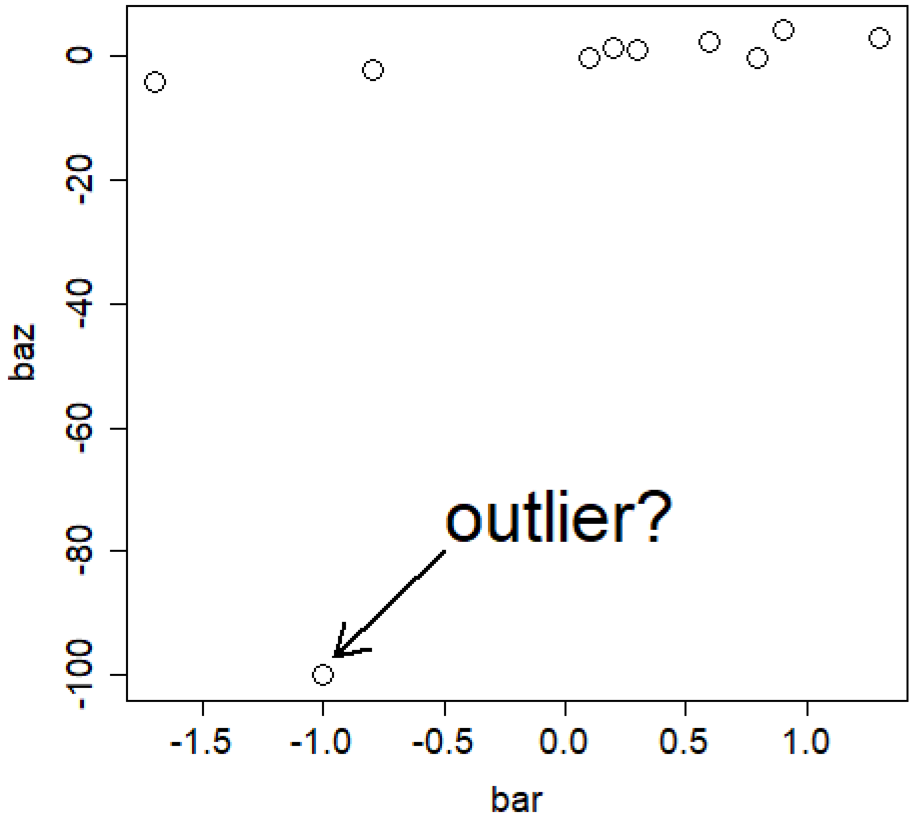
```
plot(
  x = bar,
  y = baz,
  cex = 2,
  cex.axis = 1.5,
  cex.lab = 1.5)

arrows(x0=-0.5,y0=-80,
       x1=-0.95,y1=-97,
       lwd=2)
```



- Add the text `"outlier?"` at the start of the arrow using `text`. Triple the text size `cex`. To left-justify the textbook, add `adj=0`.

```
plot(
  x = bar,
  y = baz,
  cex = 2,
  cex.axis = 1.5,
  cex.lab = 1.5)
arrows(x0=-0.5,y0=-80,
       x1=-0.95,y1=-97,
       lwd=2)
text(x=-0.5,y=-74,
     labels="outlier?",
     cex=3,
     adj=0)
```



14 Removing outliers

- Data scientist will try to remove outliers before computing results
- Outliers can occur naturally (outlier is an accurate observation), or unnaturally (as the result of a contamination or false input)

- To check this, look at both plots in one image:

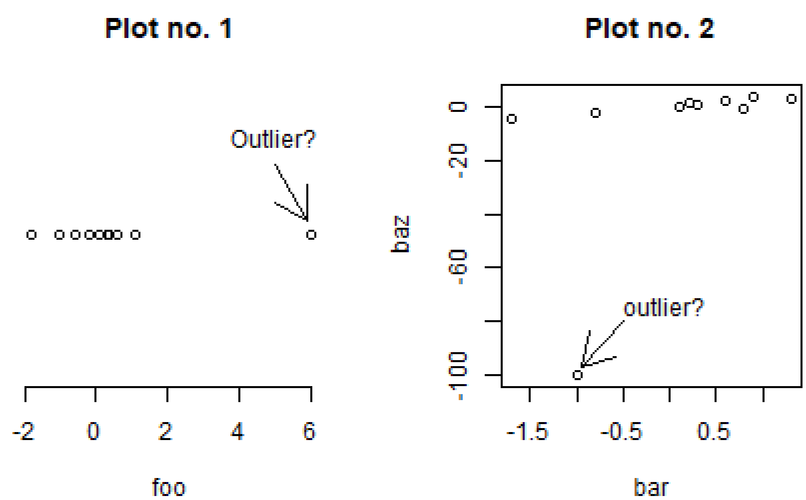
```
ls()  # check which variables are in the work environment
```

```
 [1] "bar"         "baz"         "foo"         "line"        "m"
 [6] "magnitudes"  "outlier_bar" "outlier_baz" "outlier_foo" "stations"
[11] "xdata"       "ydata"
```

```
par(mfrow=c(1,2), pty='s')  # set up two square plots side by side

plot(foo,rep(0,10),yaxt="n",ylab="",bty="n")
arrows(5,0.5,5.9,0.1)
text(5,0.7,labels="Outlier?")
title("Plot no. 1")
```

```
plot(bar,baz)
arrows(-0.5,-80,-0.95,-97)
text(-0.5,-74,labels="outlier?",adj=0)
title("Plot no. 2")
```



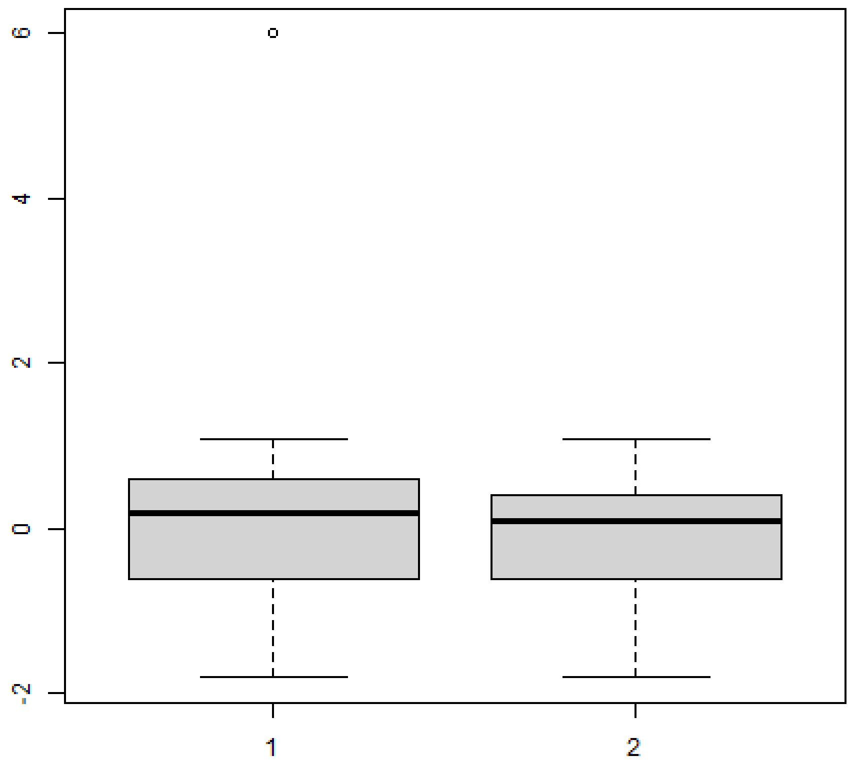
- Compute the sample mean and the mean once the outlier has been removed for `foo` in the left plot. Tip: use `which` to get the outlier's index.

```
mean(foo)
outlier_foo <- which(foo==max(foo))
mean(foo[-outlier_foo])
```

```
[1] 0.49
[1] -0.1222
```

- The sample `mean` is greatly affected. It is not "robust" against outliers.
- The boxplot of both vectors shows that the 5-point-summary is little affected - these measures are considered "robust" against outliers

```
boxplot(foo, foo[-outlier_foo])
```



- In the `boxplot`, the outliers are identified as IQR * 1.5 (can be changed, see `help(boxplot)`
- Without additional information, it is impossible to say if removing the outlier is sensible or not.

- Compute the correlation coeffient of `bar` with `baz` shown in the right plot:

```
cor(bar,baz)
outlier_baz <- which(baz==min(baz))
outlier_bar <- which(bar==min(bar))
cor(bar[-outlier_bar], baz[-outlier_baz])
```

```
[1] 0.4566
[1] -0.01537
```

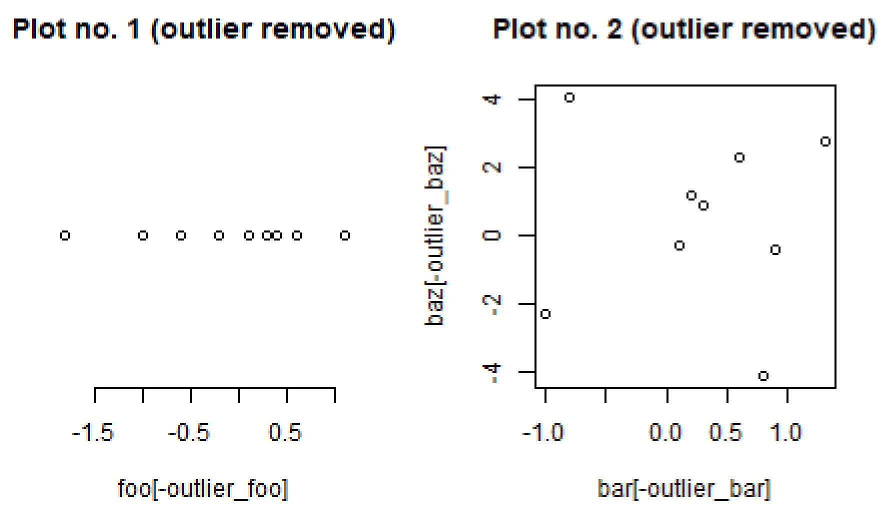
- The correlation is much stronger without that outlier.

- Check via plot that the outliers were actually removed:

```
par(mfrow=c(1,2), pty='s')  # set up two square plots side by side

plot(foo[-outlier_foo],rep(0,length(foo)-1),yaxt="n",ylab="",bty="n")
arrows(5,0.5,5.9,0.1)
title("Plot no. 1 (outlier removed)")

plot(bar[-outlier_bar],baz[-outlier_baz])
arrows(-0.5,-80,-0.95,-97)
title("Plot no. 2 (outlier removed)")
```



15 Practice: covariance, correlation, outliers

- Download the practice file from tinyurl.com/45fjmyxy
- Get the dataset from tinyurl.com/494vdr56
- Complete the practice file in Emacs
- Upload the completed practice file to Canvas

16 Glossary: concepts

TERM	MEANING

TERM	MEANING
Linearity	Functions of the form y = ax + b
Model	Abstraction after removing detail
Fit/trend	Aligning data with a curve
Intercept	Intercept with the y-axis
Slope	Gradient of a curve
Covariance	Measure of point variability
Univariate	Single variable
Bivariate	Two variables
Multivariate	Multiple variables
Outlier	Anomaly or extreme value
Causation	Causal mechanism
Correlation	$cov_{xy} / sd_x * sd_y$

17 Glossary: code

CODE	MEANING
`lm`	linear model
`formula`	e.g. `y ~ x`
`cov(x,y)`	covariance of x with y
`cor(x,y)`	correlation of x with y
`cex`	point scale
`cex.axis`	axis label scale
`cex.lab`	label scale
`axes`	Draw axes or not (T/F)
`bty`	box type (no box: `"n"`)
`yaxt`	y-axis type
`arrows`	place arrow
`text`	place textbook
`labels`	text in `text`
`adj`	textbox justification
`las`	axis label orientation
`boxplot`	box-and-whiskers plot

18 References

- Davies TD (2016). Book of R. NoStarch Press. URL: nostarch.com

Footnotes:

[1] The covariance formula carries the same correction in the denominator n-1 for samples vs. n for populations as the variance.

[2] This begs the question: how can you measure causation?

Author: MARCUS BIRKENKRAHE
Created: 2022-11-03 Thu 10:02