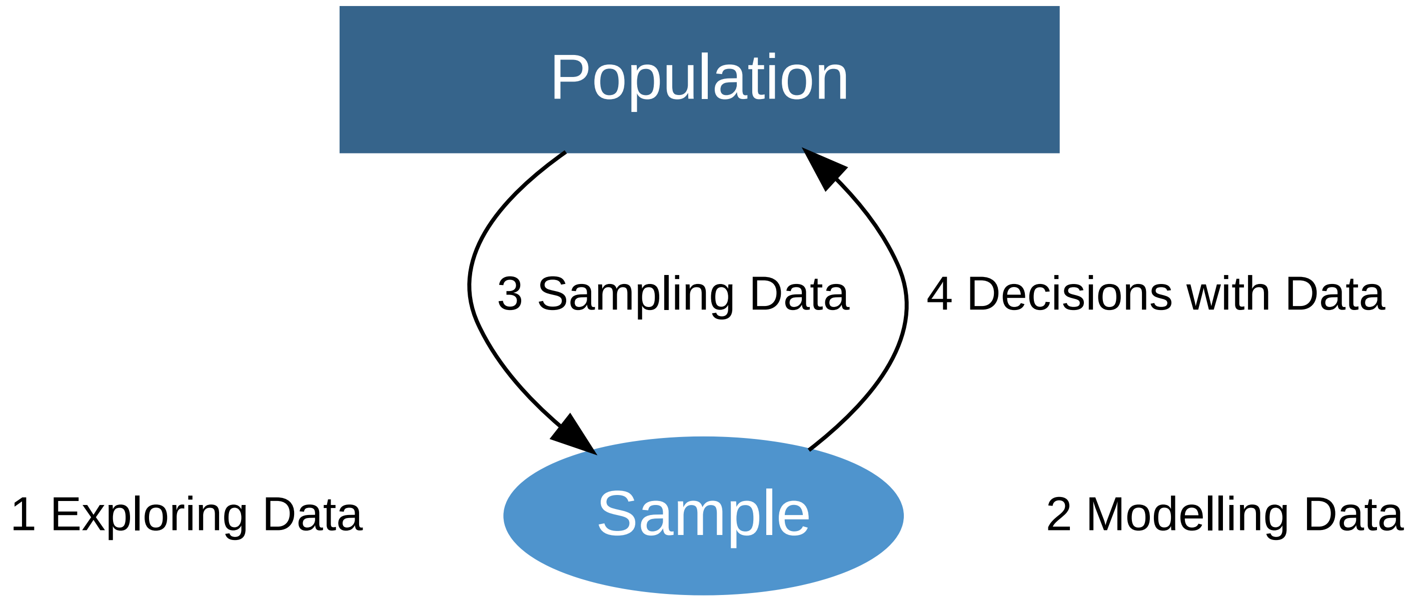
Scatter Plot & Correlation

Modelling Data | Linear Model

Unit Overview



Module2 Modelling Data

Normal Model

What is the Normal Curve? How can we use it to model data?

Linear Model

How can we describe the relationship between 2 variables? When is a linear model appropriate?

Scatter Plot & Correlation

Data Story | Can we predict a son's height from his father's height?

Bivariate Data & Scatter Plot

Correlation Coefficient

SD Line

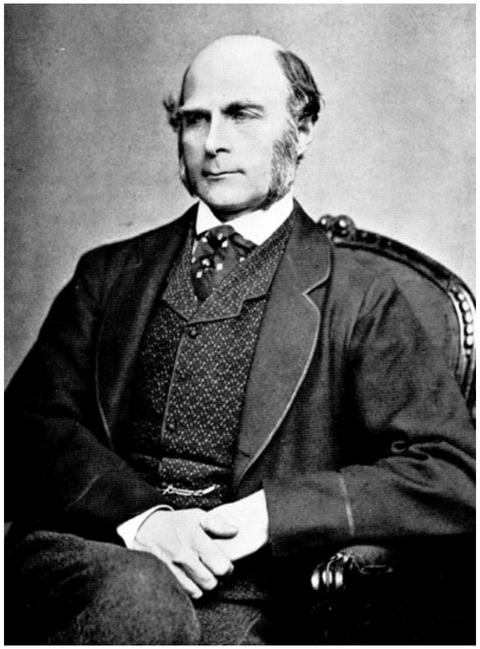
Summary

Data Story

Can we predict a son’s height from his father’s height?

History

- Sir Francis Galton (England, 1822–1911) studied the degree to which children resemble their parents (and wrote travel books on “wild countries”!)
- His work was continued by Galton’s student Karl Pearson (England, 1857–1936). Pearson measured the heights of 1,078 fathers and their sons at maturity.



Pearson’s data on heights

```
#install.packages("UsingR")  # Loads another collection of datasets
suppressMessages(library(UsingR))
library(UsingR)
data(father.son)  #This is Pearson's data.
data=father.son
dim(data)
```

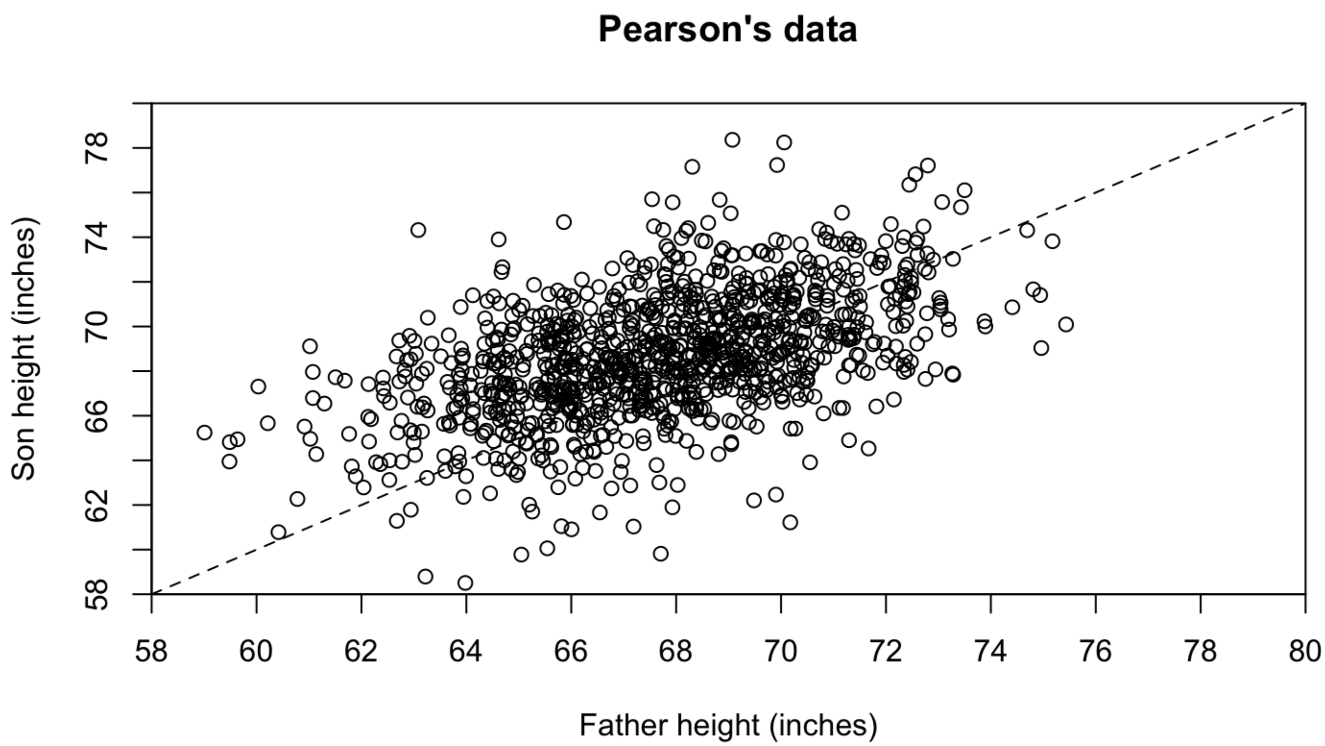
```
## [1] 1078    2
```

```
str(data)
```

```
## 'data.frame':    1078 obs. of  2 variables:
##  $ fheight: num  65 63.3 65 65.8 61.1 ...
##  $ sheight: num  59.8 63.2 63.3 62.8 64.3 ...
```

```
x = data$fheight
y = data$sheight
```

Pearson’s plot of heights



Statistical Thinking

What do you notice about the heights?

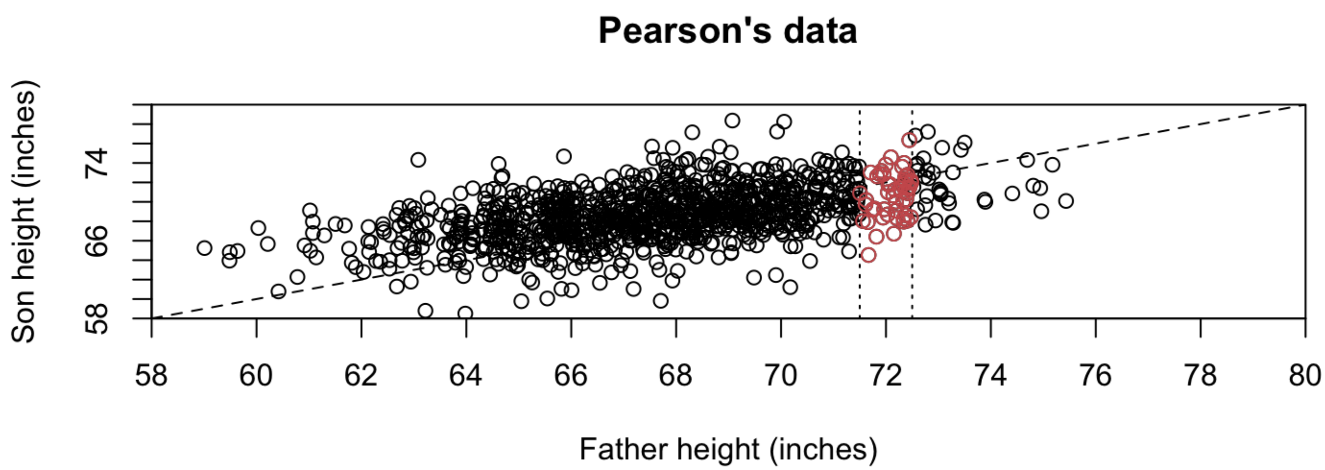
- Plotting the pairs of heights creates a cloud of points.
- Generally, taller fathers tend to have taller sons.

What is the dotted line?

- It joins together the points where the son has exactly the **same height** as the father.
- If a son’s height is close to his father’s height, the father-son point is close to the line.
- As there is a lot of spread around this line, there is a weak relationship between father’s height and son’s height.

Guessing a son’s height

- Suppose you want to guess the height of a son, when the father was 72 inches tall.
- Draw a vertical “chimney” containing the father-son pairs where the father is 72 inches tall to the nearest inch.



Notice: There is a lot of variability in the “chimney” with heights of the sons, because the relationship is weak.

Bivariate Data & Scatter Plots

Bivariate Data

Bivariate data

Bivariate data involves a **pair** of variables. We are interested in the relationship between the 2 variables. Can one variable be used to predict the other?

- Formally, we have (x_i, y_i) for $i = 1, 2, \ldots, n$.
- X is called the **independent** variable (or explanatory variable, predictor or regressor).
- Y is called the **dependent** variable (or response variable).

What are examples?

Scatter Plot

Scatter Plot

A **scatter plot** is a graphical summary of 2 variables on the same 2D plane, resulting in a cloud of points.

Linear association

Linear association

- The **linear assocation** (or **association**) between 2 variables describes how tightly the points cluster around a line.

- If there is a **strong** association, the cloud of points are **tightly clustered** around a line, and this allows for good predictions from 1 variable to the other.

- If one variable tends to **increase** with the other, then we have **positive** association.

How do we measure this association?

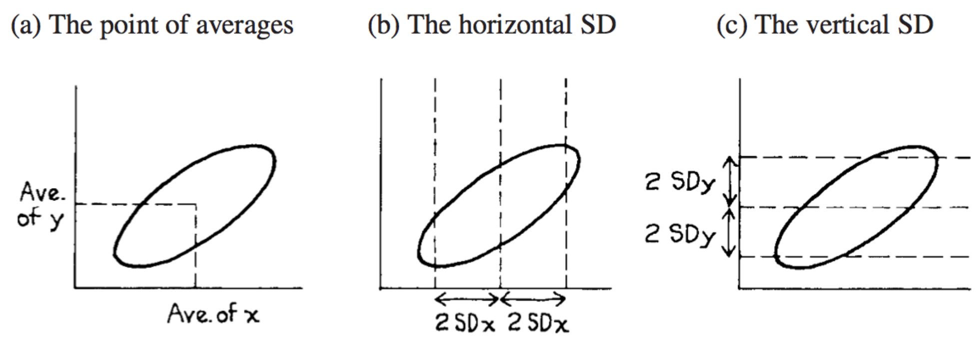
Correlation Coefficient

How can we summarise a scatter plot?

The scatter plot can be summarised by the following 5 numerical summaries:

- mean and SD of X ($\bar{x}$, SD_x)
- mean and SD of Y ($\bar{y}$, SD_y)
- correlation coefficient (r).

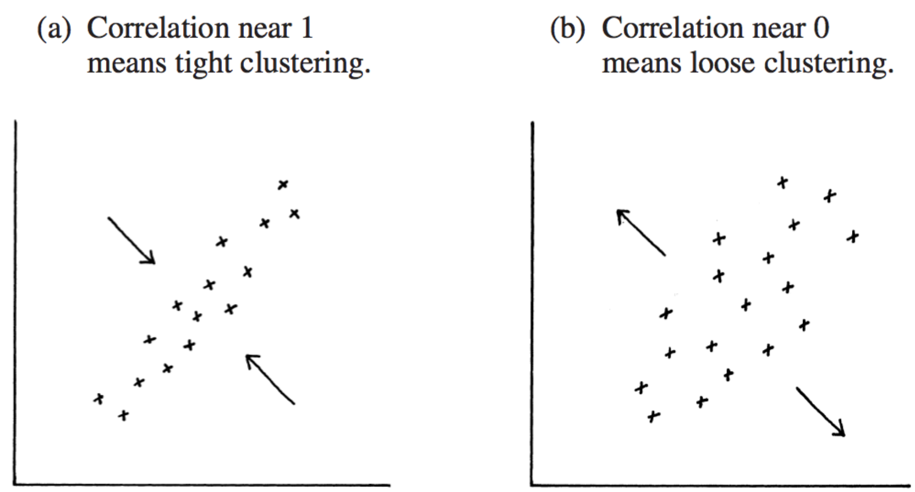
Centre and spread of the cloud



Source: Freedman et al, Statistics p125

- The **centre** of the cloud is represented by the point of averages ($\bar{x}$, $\bar{y}$).
- The **horizontal spread** of the cloud is measured by SD_x. We expect most of the points to fall with 2 SDs from $\bar{x}$.
- The **vertical spread** of the cloud is measured by SD_y. We expect most of the points to fall with 2 SDs from $\bar{y}$.

Association between the 2 variables



Source: Freedman et al, Statistics p125

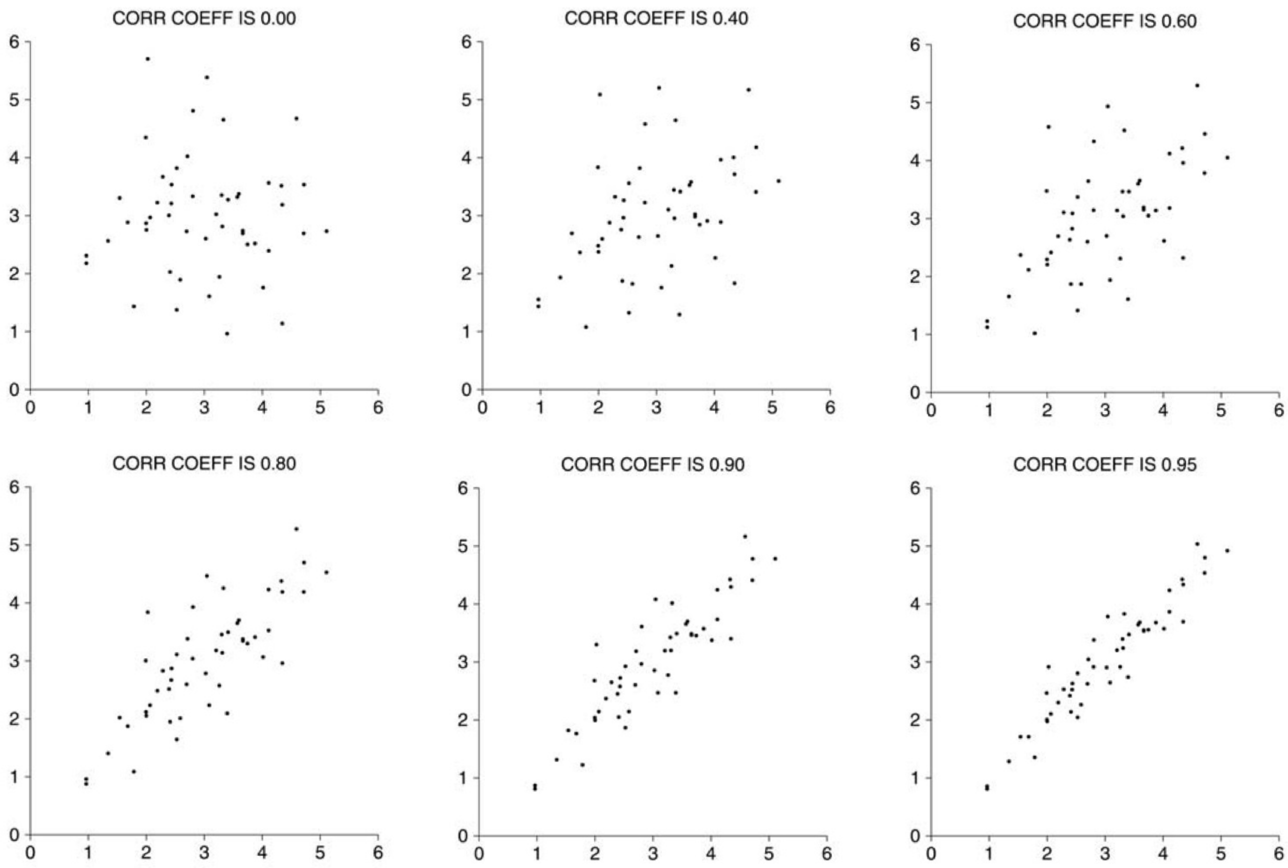
- Note that both clouds have the same centre and horizontal and vertical spread.
- However they have different clustering around a line (linear association). How do we measure this?

The correlation coefficient

Correlation coefficient

- The **correlation coefficient** r is a numerical summary which measures the clustering around the line.
- It indicates both the sign and strength of the linear association.
- The correlation coefficient is between -1 and 1.
  - If r is positive: the cloud slopes up.
  - If r is negative: the cloud slopes down.
  - As r gets closer to ± 1: the points cluster more tightly around the line.

Examples



Source: Freedman et al, Statistics p127

Calculating the correlation coefficient

Population correlation coefficient

The population correlation coefficient (r_{pop}) is the mean of the product of the variables in standard units.

Calculation by hand

```
head(data,3)
```

```
##    fheight  sheight
## 1 65.04851 59.77827
## 2 63.25094 63.21404
## 3 64.95532 63.34242
```

```
c(mean(data$fheight),sd(data$fheight))
```

```
## [1] 67.687097  2.744868
```

```
c(mean(data$sheight), sd(data$sheight))
```

```
## [1] 68.684070  2.814702
```

Here, for illustration, we round data to 1 decimal place to make calculations simpler.

x (father's heights)	y (son's heights)	standard units	standard units	product	quadrant
		$\frac{x-67.7}{2.7}$	$\frac{y-68.7}{2.8}$	$\left(\frac{x-67.7}{2.7}\right)\left(\frac{y-68.7}{2.8}\right)$	
65.0	59.8	-1.0	-3.2	3.2	lower left
63.3	63.2	-1.6	-2.0	3.2	lower left
65.0	63.3	-1.0	-1.3	1.3	lower left
70.3	67.0	1.0	-0.6	-0.6	lower right
$\vdots$					
				mean=+0.5	

Quick calculation in R

```
SU_x=(data$fheight-mean(data$fheight))/sd(data$fheight)
SU_y=(data$sheight-mean(data$sheight))/sd(data$sheight)
mean(SU_x*SU_y)
```

```
## [1] 0.5008732
```

Even quicker calculation in R

```
cor(data$fheight,data$sheight)
```

```
## [1] 0.5013383
```

Why is this slightly different?

Again, like the SD, there are 2 slightly different formulas for the population and sample.

```
n = length(data$fheight)
cor(data$fheight,data$sheight)*(n-1)/n  # to match up with hand calculation (Ext)
```

```
## [1] 0.5008732
```

Overall Summary: population vs sample

Summary	Formula	In R
Population correlation coefficient r_{pop}	Mean of the product of the variables in standard units.	
Sample correlation coefficient r_{sample}	Adjusted mean of the product of the variables in standard units.	`cor(x,y)`

Note:

- In what follows, we'll assume a sample and simply use `cor()`.
- Formally, $\mathrm{r}_{pop} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{SD_x}\frac{y_i - \bar{y}}{SD_y}$ and $\mathrm{r}_{sample} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{SD_x}\frac{y_i - \bar{y}}{SD_y}$

Classic mistakes

Mistake 1:

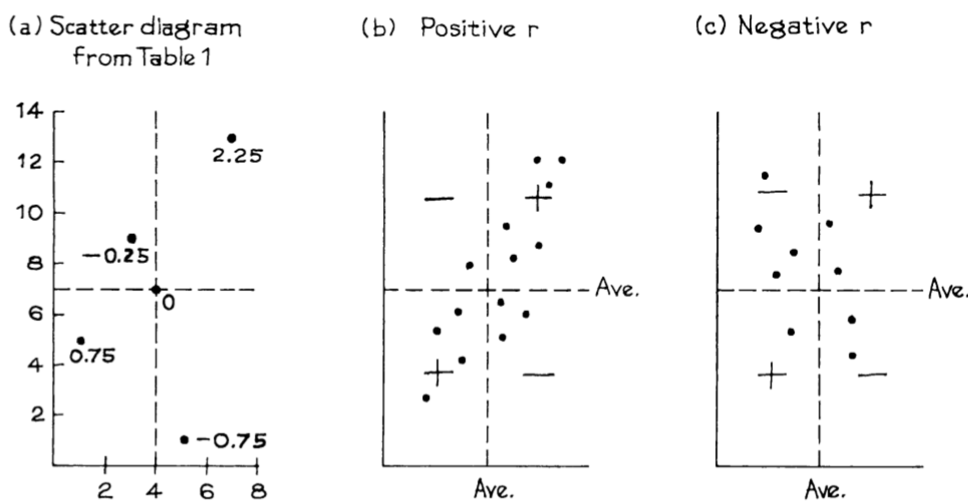
$r = 0.8$ means that 80% of the points are tightly clustered around the line.

Mistake 2:

$r = 0.8$ means that the points are twice as tightly clustered as $r = 0.4$.

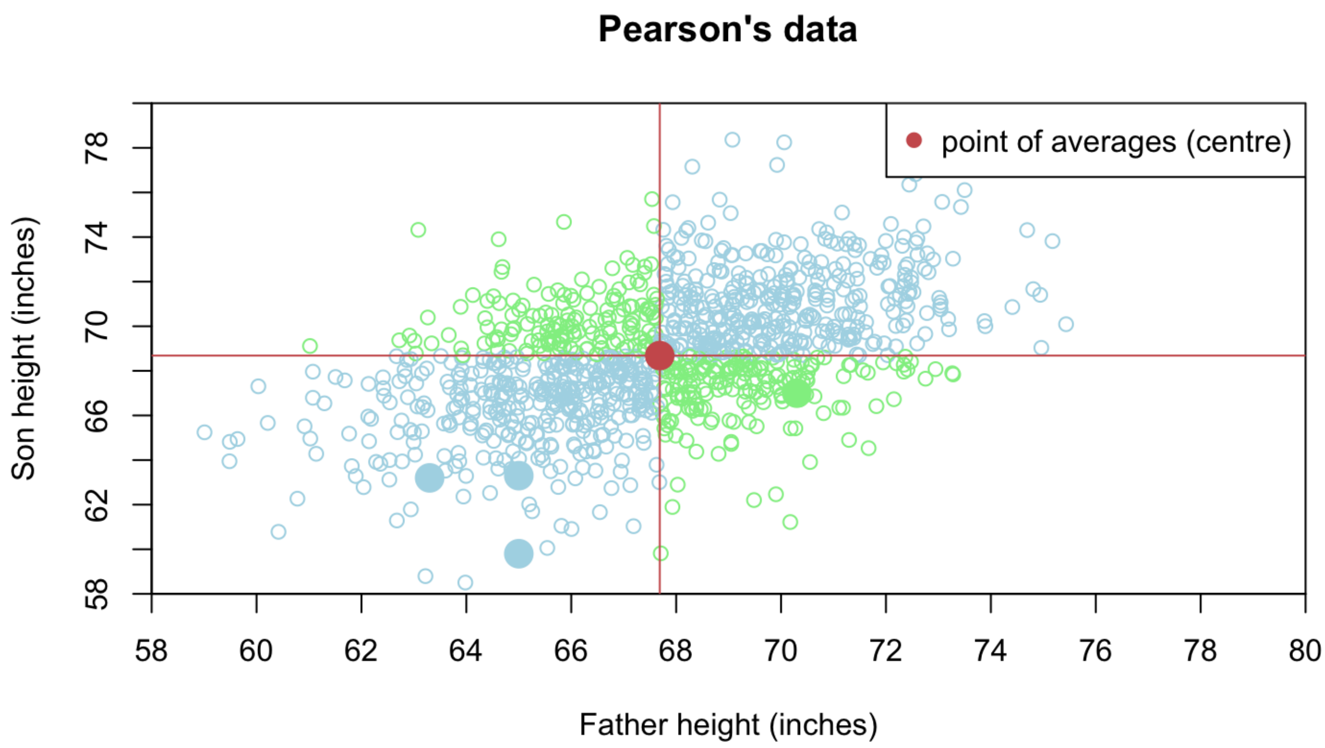
Why does r measure association?

- It divides the scatter plot into 4 quadrants, at the point of averages (centre).



Source: Freedman et al, Statistics p127

- Hence a majority of points in the upper right (+) and lower left quadrants (+) will be indicated by an overall + value of r.

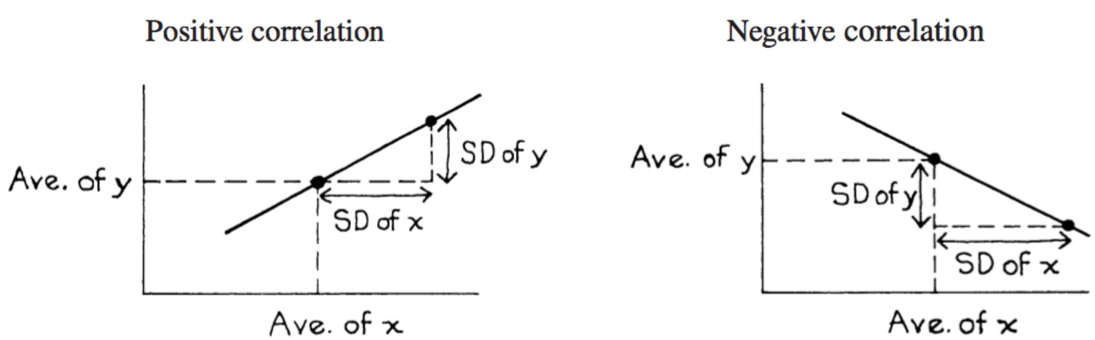


Pearson's data
point of averages (centre)
Son height (inches)
58
62
66
70
74
78
58
60
62
64
66
68
70
72
74
76
78
80
Father height (inches)

SD Line

What is the ‘line’?

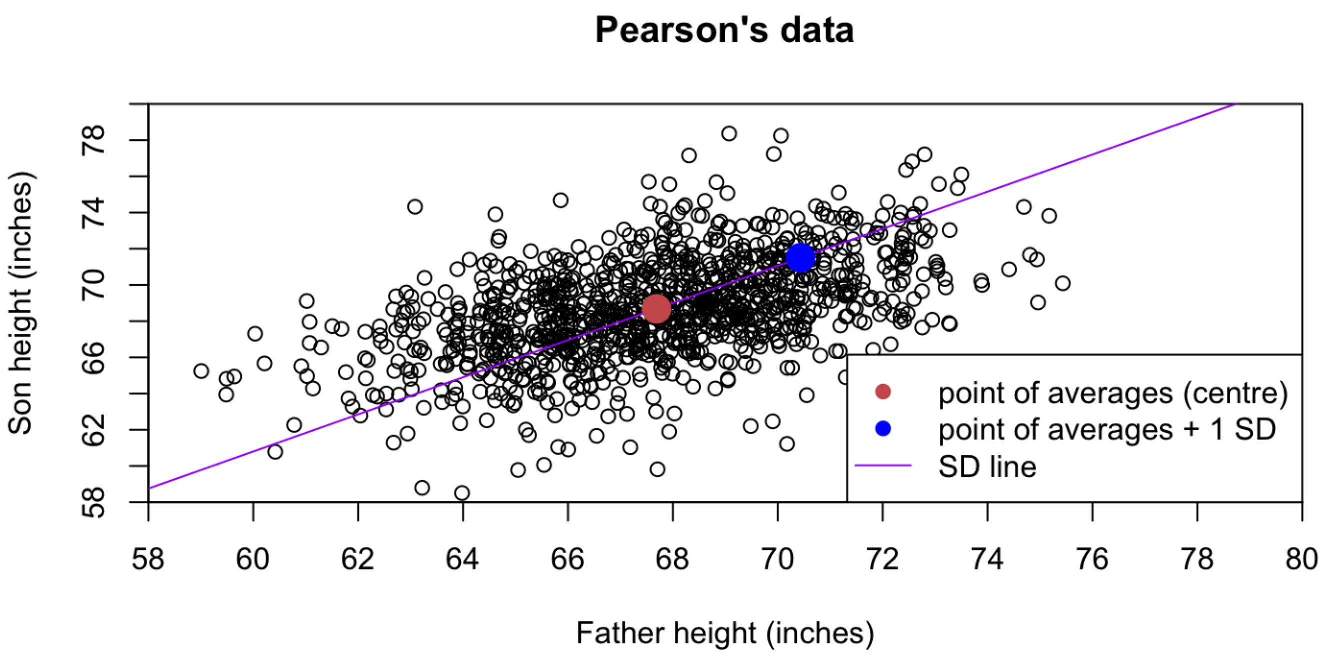
- What line are the points clustered around?
- We begin by considering the **SD line**, because the data points generally seem to cluster around it.
- The SD line connects the point of averages ($\bar{x}$, $\bar{y}$) to ($\bar{x} + \mathrm{SD}_x$, $\bar{y} + \mathrm{SD}_y$) (for $r > 0$) or ($\bar{x}$, $\bar{y}$) to ($\bar{x} + \mathrm{SD}_x$, $\bar{y} - \mathrm{SD}_y$) (for $r < 0$) .



Source: Freedman et al, Statistics p131

Features of the SD Line

- The SD line goes through the point of averages.
- A father-son pair where both are 0.5 SDs above the mean would lie on the SD line.



Experiment here.

Limitations of the SD Line

As the SD line does not use r:

- It is insensitive to clustering. The SD Line does not take into account how tightly the points are clustered in the cloud.
- At the extremes with positive [or negative] correlation, the SD Line will over-estimate in RHS [LHS] and under-estimate in LHS [RHS].
- We need something better for predictions!

Summary

The scatter plot is a cloud of points which represents bivariate data (a pair of variables). The scatter plot is summarised by the point of averages, the SD of the 2 variables and the correlation coefficient. The population correlation coefficient is the mean of the product of the variables in standard units. The sample correlation coefficient can be found using `cor()`.

Key Words

cloud, bivariate data, independent, dependent, scatter plot, linear association, correlation coefficient, horizontal spread, vertical spread, quadrants, SD line