

# Linear models

Master of Cognitive Science

Marco Maiolini & Steven Moran

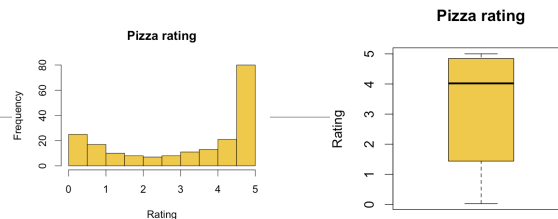
Nov 16, 2022

# Overview of visualization and summary techniques

- **one quantitative variable**

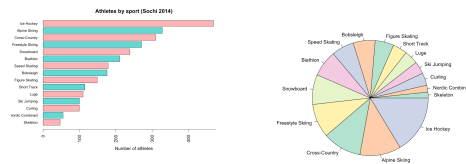
histogram, box plot

five-number-summary (min., max., median, mean, IQR), SD



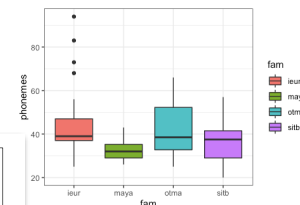
- **one categorical variable**

bar plot, pie chart + frequency distribution



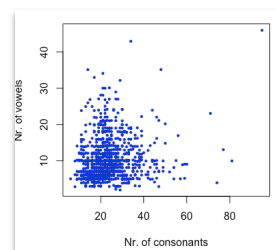
- **one categorical and one quantitative variable**

parallel box plots



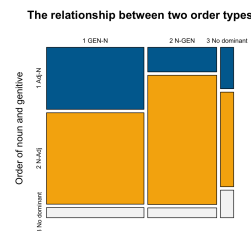
- **two quantitative variables**

scatter plot, correlation coefficients  $r$ ,  
regression line



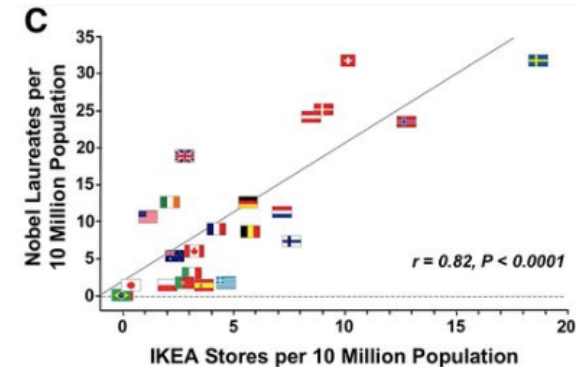
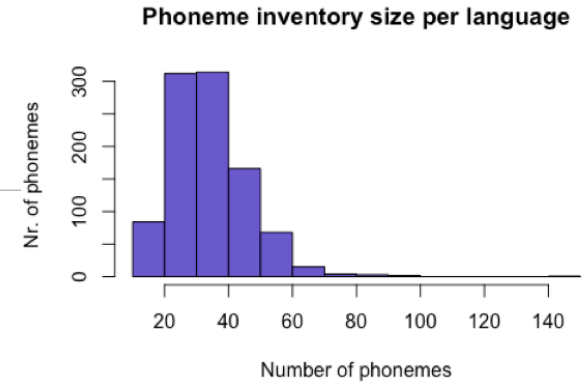
- **two categorical variables**

mosaic plot, grouped/stacked bar plots



## Bivariate distributions

- **Univariate** distributions involve one variable
- In addition to considering individual variables, we can look at the association (or relation) between two variables  
→ the question of **correlation**
- A correlation is exactly what its name suggests:  
a **co-relation** between two variables
- A bivariate distribution may show positive correlation, negative correlation, or zero correlation
- Statistically, the strength of the relation between two variables is indicated with the **correlation coefficient  $r$** .
- Visually the relation between two variables is represented as a **scatter plot**



## Bivariate distributions

- **Positive** correlation between two variables:
  - high measurements on one variable tend to be associated with high measurements on the other variable, and low measurements on one variable with low measurements on the other
    - e.g. tall fathers tend to have sons who grow up to be tall men; short fathers tend to have sons who grow up to be short men
- **Negative correlation:** increases in one variable are accompanied by decreases in the other variable (an inverse relationship)
- **Zero correlation:** no *linear* relationship between two variables  
High and low scores on the two variables are not associated in any predictable manner

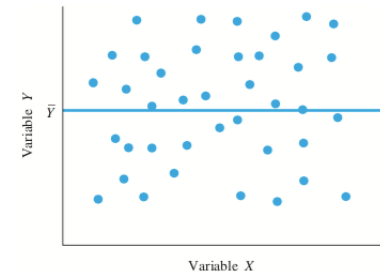
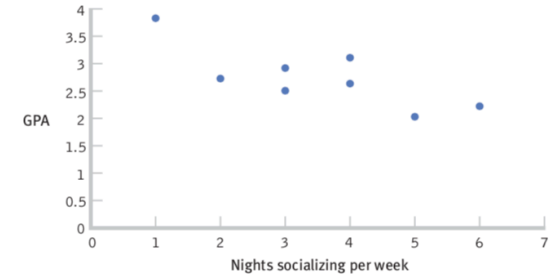
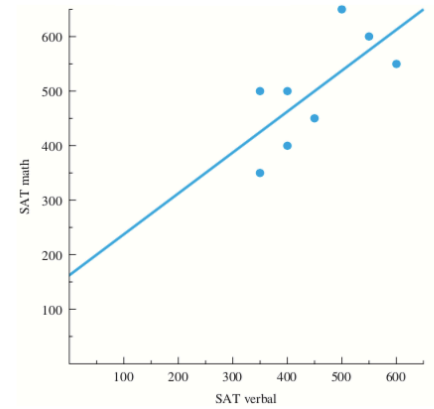
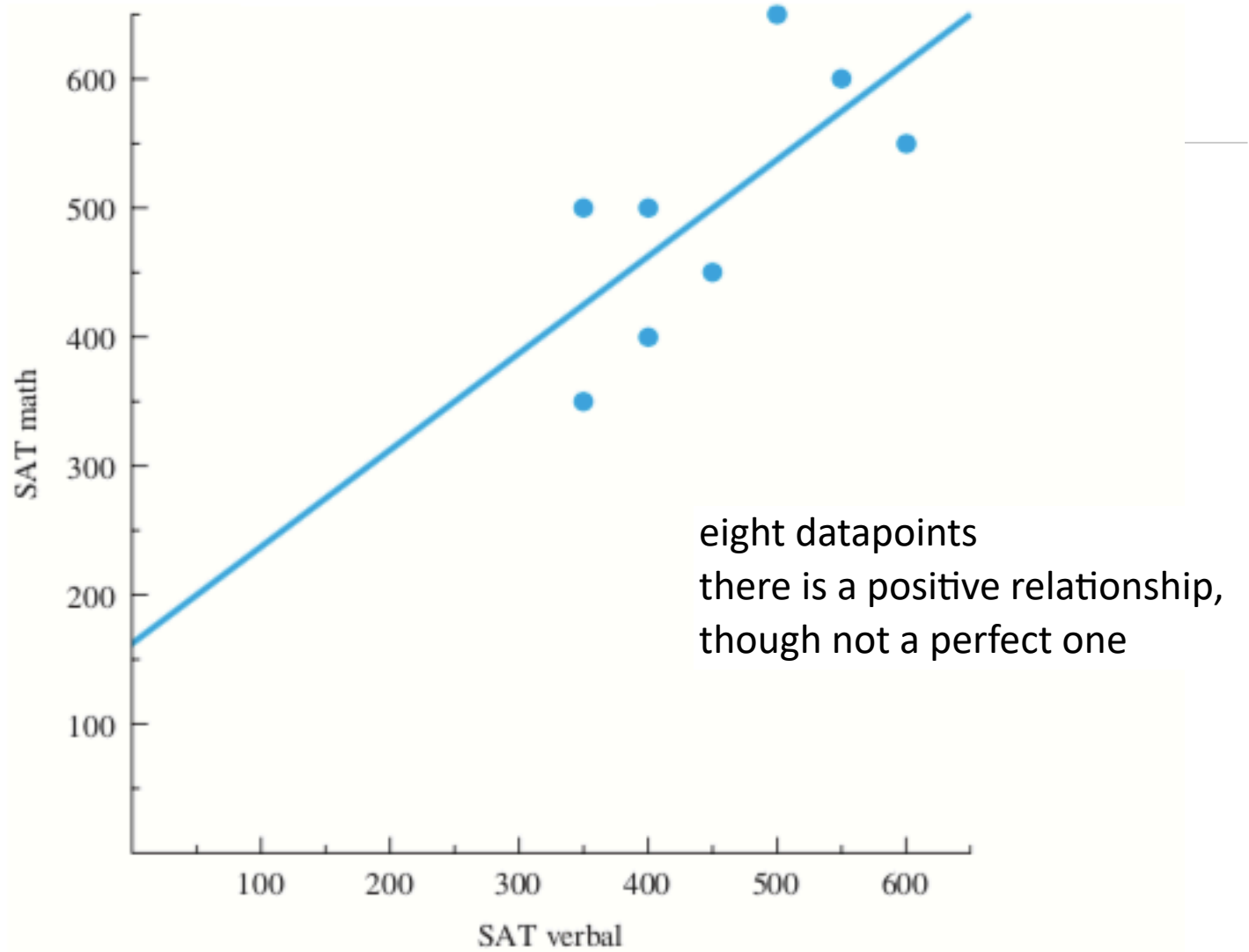
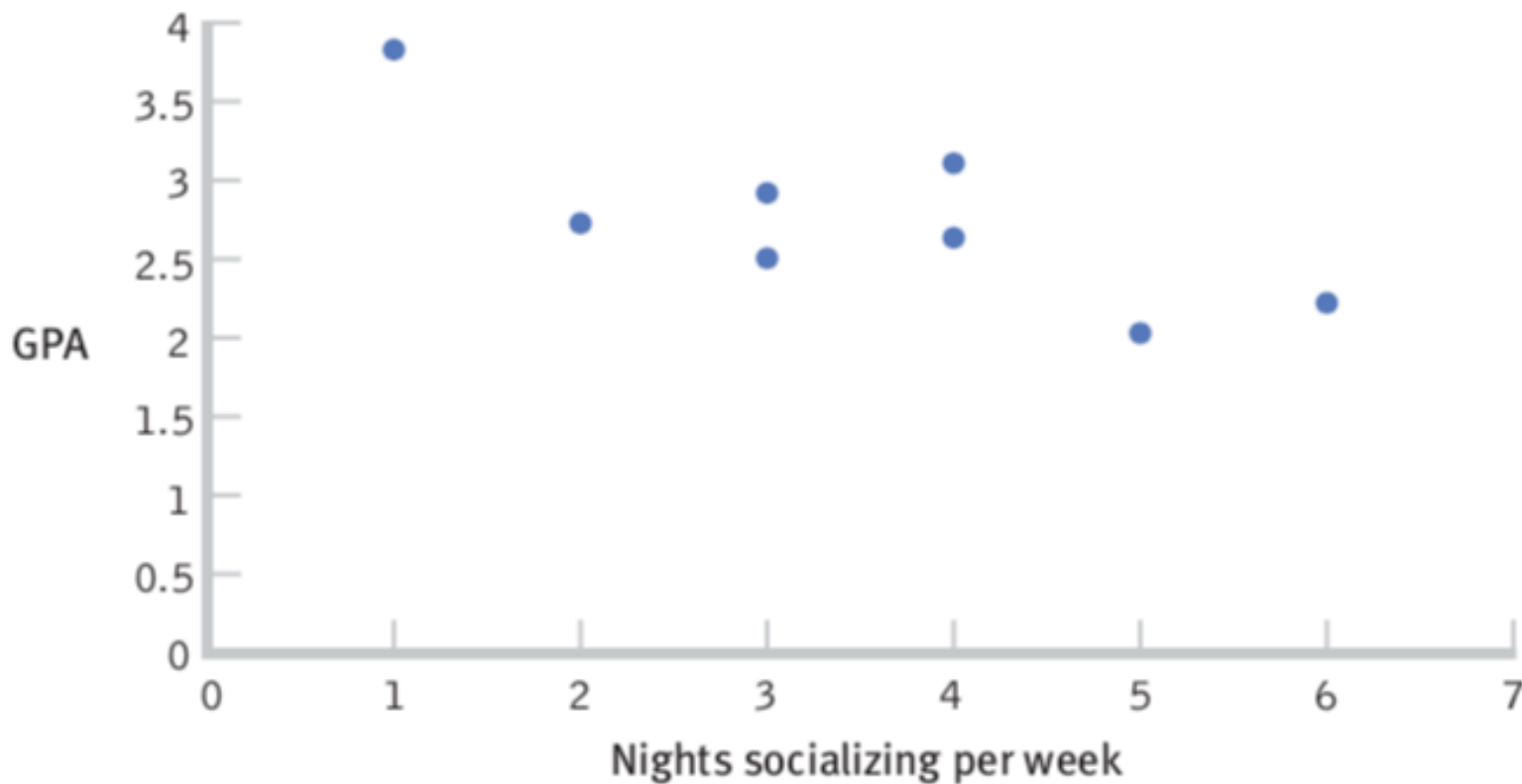
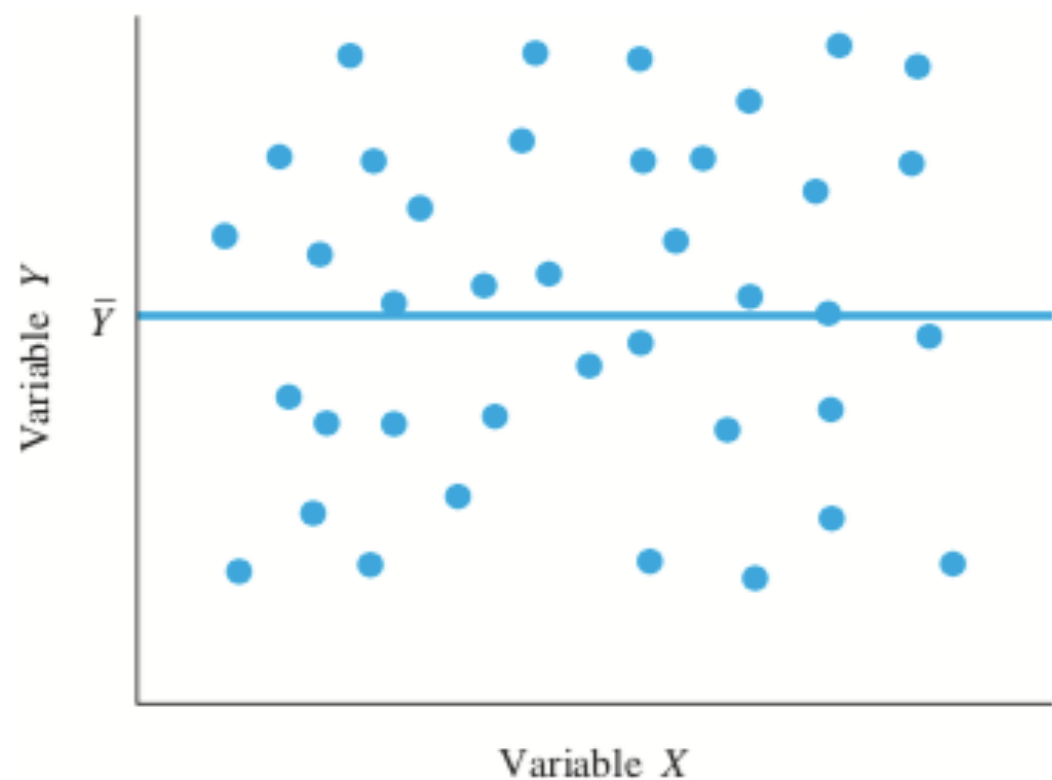


FIGURE 5.5 Scatterplot and regression line for a zero correlation



These data points depict a negative correlation between nights socializing per week and GPA. Those who go out more tend to have lower GPAs, whereas those who go out less tend to have higher GPAs.





**FIGURE 5.5** Scatterplot and regression line for a zero correlation

# Bivariate relationship

When we have two numerical variables, we can distinguish:



# Bivariate relationship

When we have two numerical variables, we can distinguish:

- *Response variable*: dependent variable, as known as Y

# Bivariate relationship

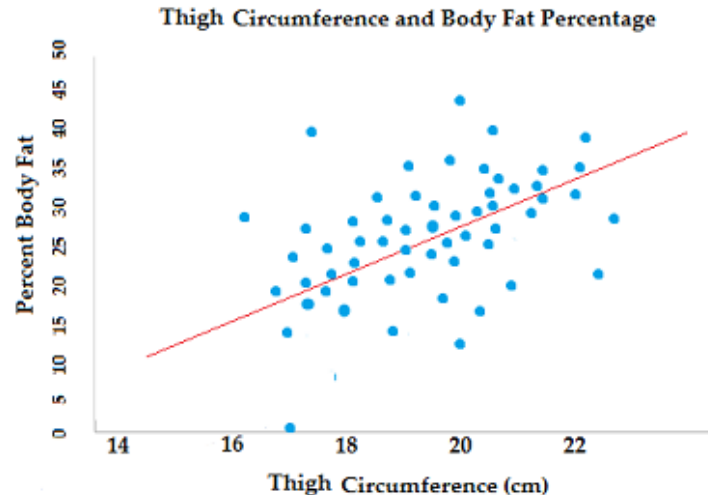
When we have two numerical variables, we can distinguish:

- *Response variable*: dependent variable, as known as Y
- *Explanatory variable*: independent variable, as known as X

# Bivariate relationship

When we have two numerical variables, we can distinguish:

- *Response variable*: dependent variable, as known as Y
- *Explanatory variable*: independent variable, as known as X



## Response (dependent) and explanatory (independent) variable

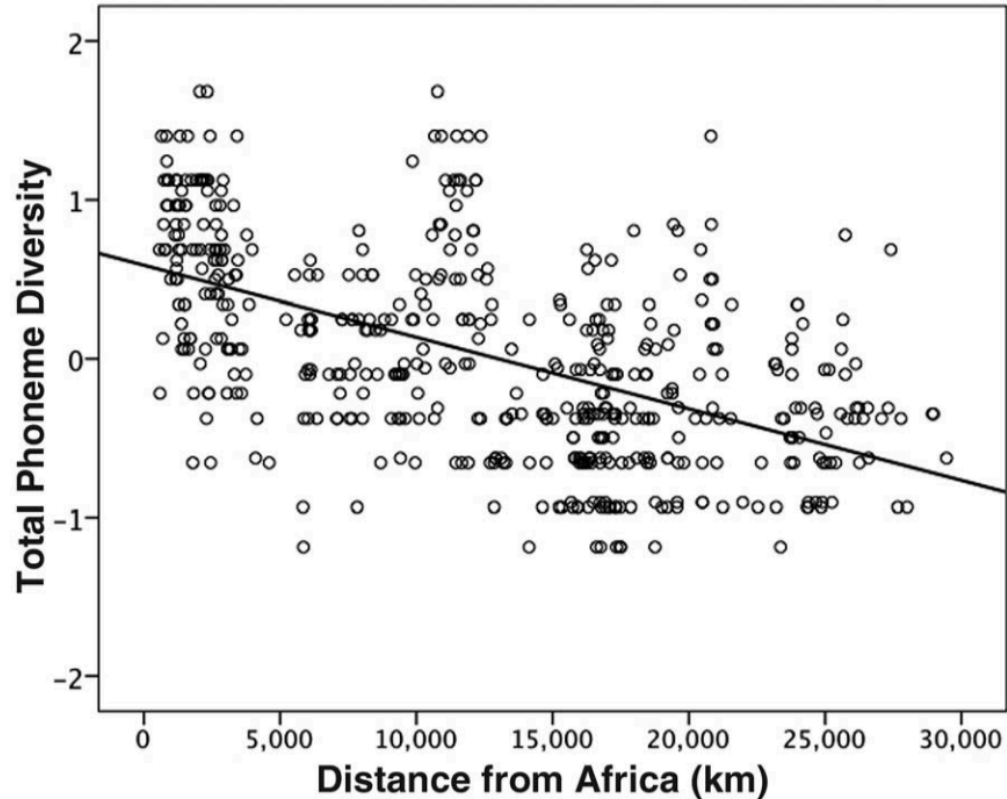
---



- Warming Reducing Rice Yields (2004, *PNAS*) reported that
  - “an average daily temperature increase of  $1^{\circ}\text{C}$  resulted in a 10% reduction in the rice crop.”
- In the US, corn and soybean yields were found to reduce in a manner similar to that of rice in the Philippines.
  - a. A positive or a negative correlation?
  - b. Which of the variables, rice yield or temperature, is the **explanatory** variable?

## Scatter plot

- Visually the relation between two variables (X and Y) can be represented as a **scatter plot**



# Relationship between X and Y

Techniques based on fitting a straight line to the data:

# Relationship between X and Y

Techniques based on fitting a straight line to the data:



Linear regression



Correlation analysis

# Linear regression

## Example

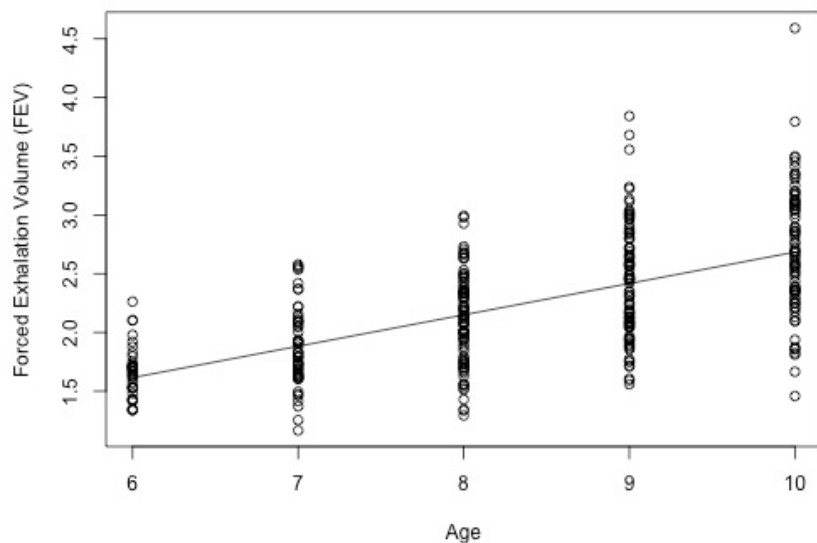
You want to test Lung Function in children. You measure the forced exhalation volume (FEV), the measure of how much air somebody can forcibly exhale from their lungs, from 6 to 10 year old children. You survey 345 children.



# Linear regression

## Example

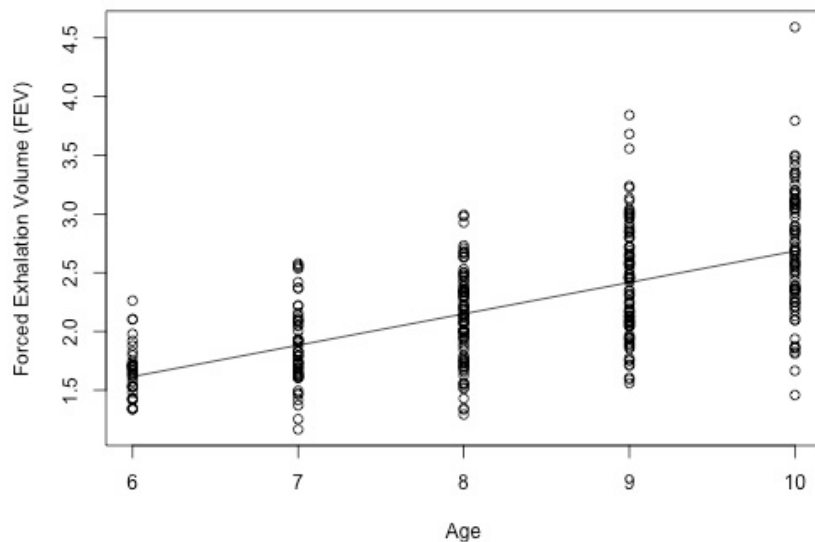
You want to test Lung Function in children. You measure the forced exhalation volume (FEV), the measure of how much air somebody can forcibly exhale from their lungs, from 6 to 10 year old children. You survey 345 children.



# Linear regression

## Example

You want to test Lung Function in children. You measure the forced exhalation volume (FEV), the measure of how much air somebody can forcibly exhale from their lungs, from 6 to 10 year old children. You survey 345 children.



The scatter plot suggests a definite age-relationship, with larger X tending to be associated with bigger values of Y

# Correlation analysis

## Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

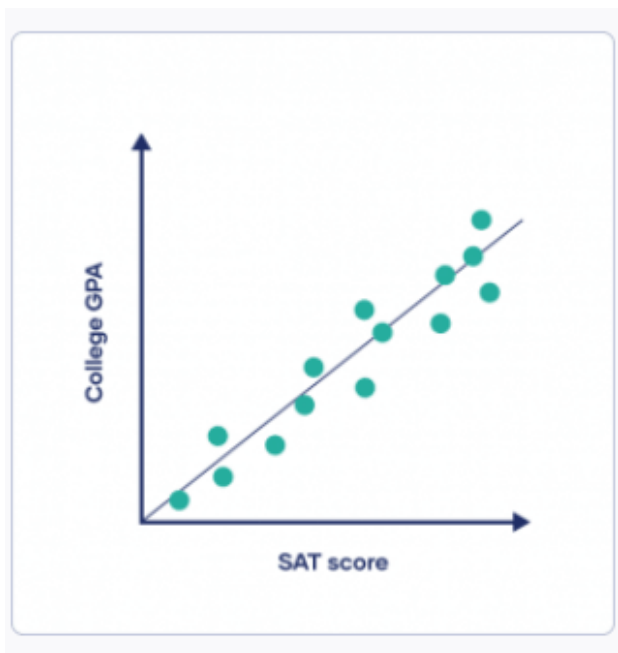
You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.

# Correlation analysis

## Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.



# Correlation analysis

## Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.



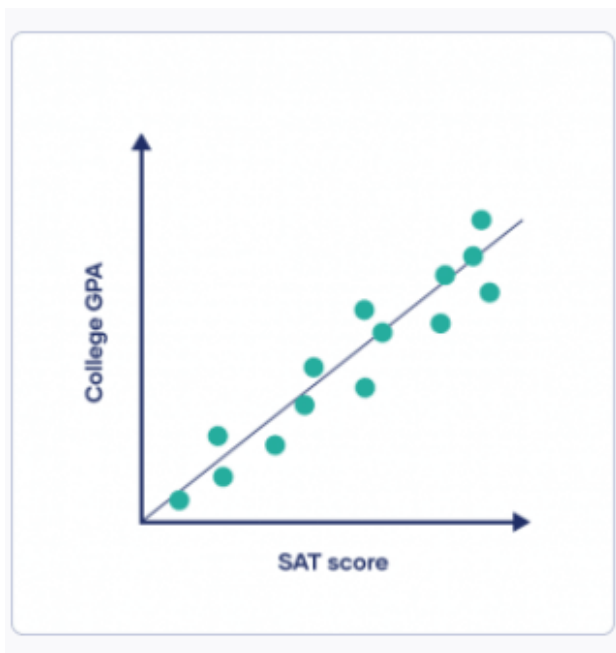
Correlation coefficient = 0.58

# Correlation analysis

## Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.



Correlation coefficient = 0.58

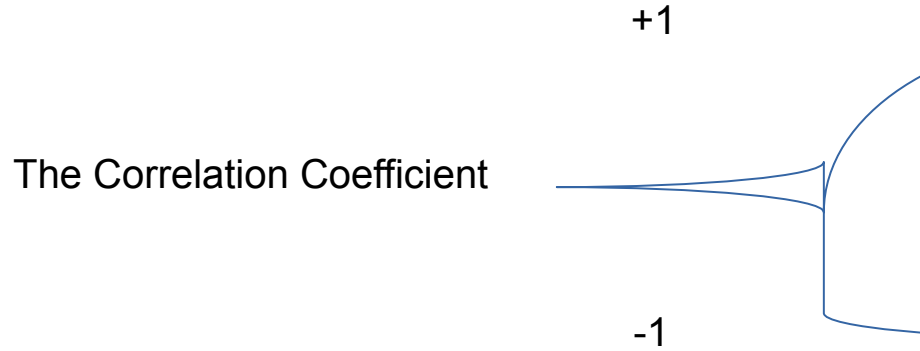
The scatter plot seems to confirm our prediction, with higher SAT scores associated with higher GPA values.

# The Correlation Coefficient

The Correlation Coefficient measures the strength of linear association between the two variables.

# The Correlation Coefficient

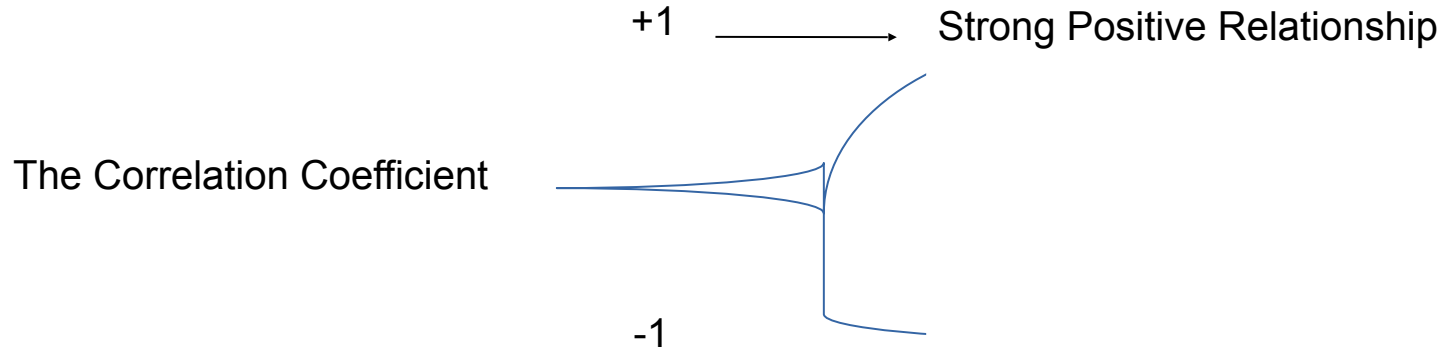
The Correlation Coefficient measure the strength of linear association between the two variables.





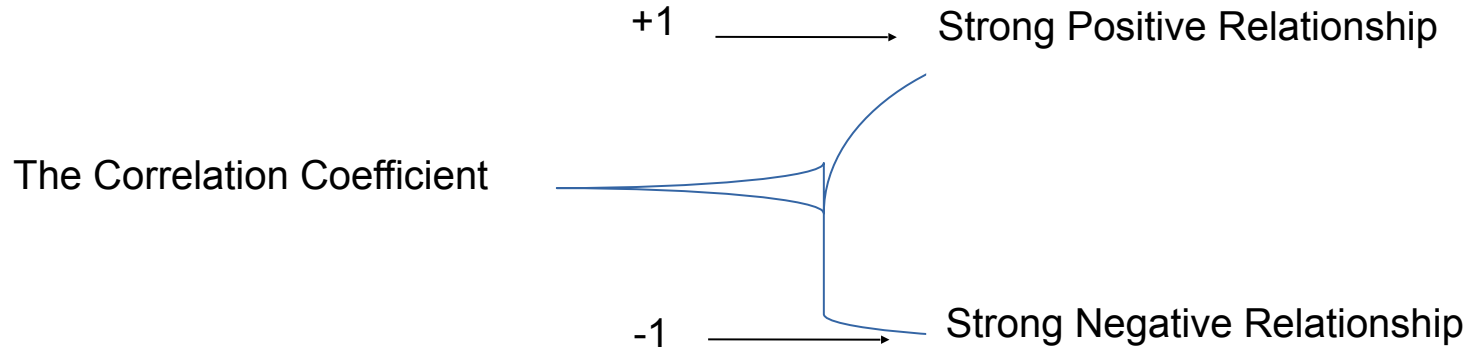
# The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.



# The Correlation Coefficient

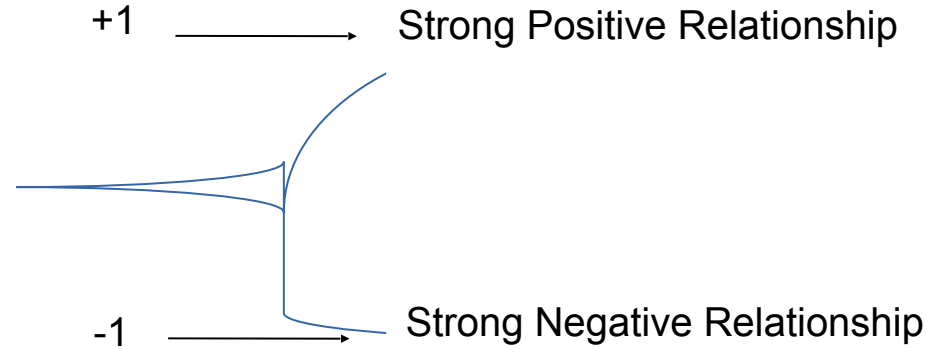
The Correlation Coefficient measure the strength of linear association between the two variables.



# The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.

The Correlation Coefficient



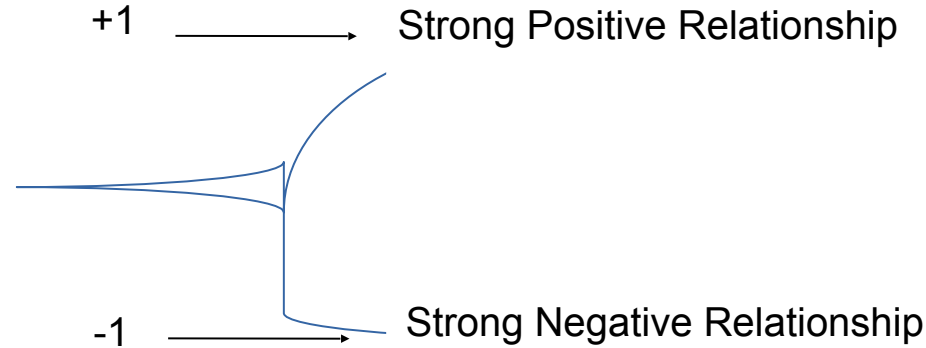
The Pearson's  $r$  correlation test:

- Variables are quantitative
- Variables normally distributed

# The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.

The Correlation Coefficient

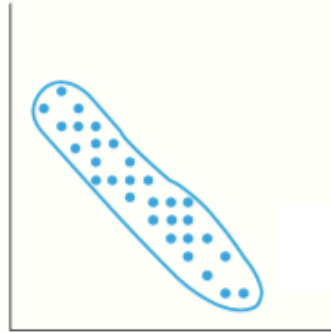
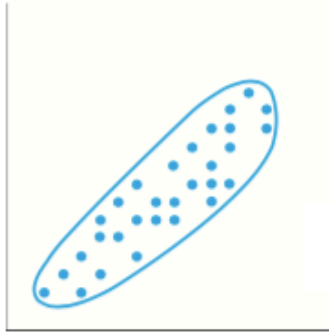
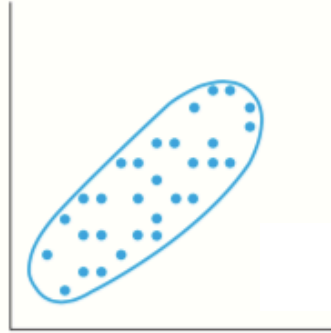
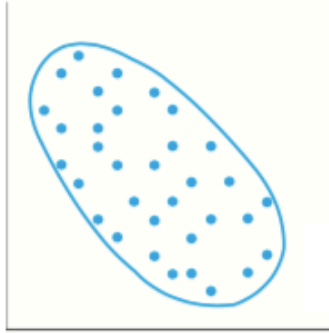
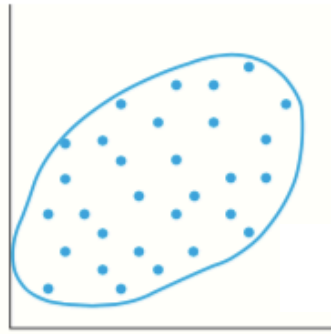
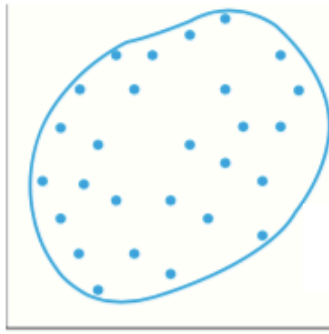


The Pearson's  $r$  correlation test:

- Variables are quantitative
- Variables normally distributed

In R you see if two variables are correlated by:

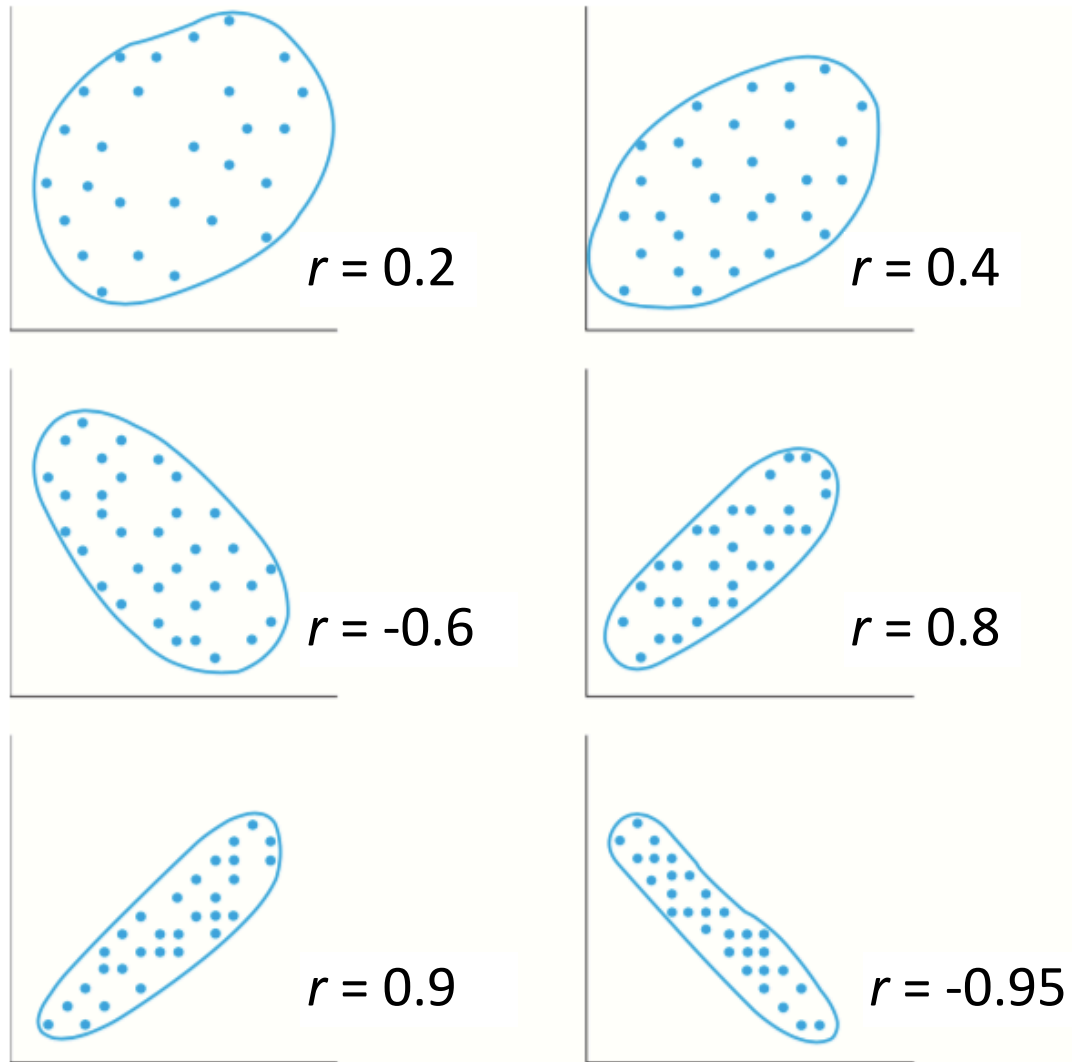
`cor(x, y)`



---

$r = 0.4$   
 $r = -0.6$   
 $r = 0.8$   
 $r = -0.95$   
 $r = 0.2$   
 $r = 0.9$

**FIGURE 5.6** Scatterplots of data in which  $r = .20, .40, -.60, .80, .90$ , and  $-.95$



**FIGURE 5.6** Scatterplots of data in which  $r = .20, .40, -.60, .80, .90$ , and  $-.95$

# The Correlation Coefficient

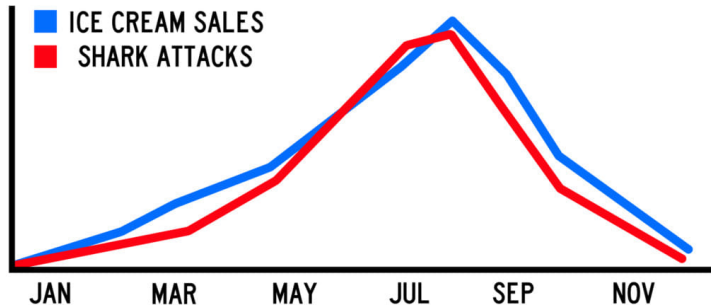


A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.

# The Correlation Coefficient



A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.



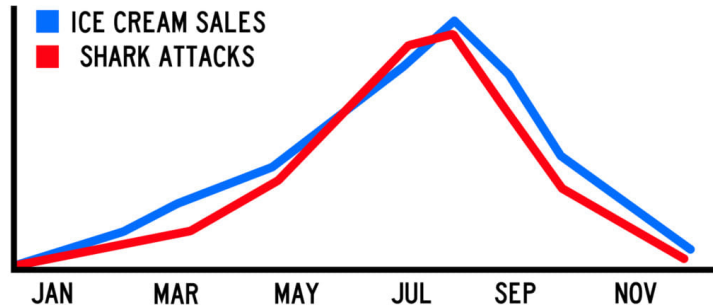
Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)



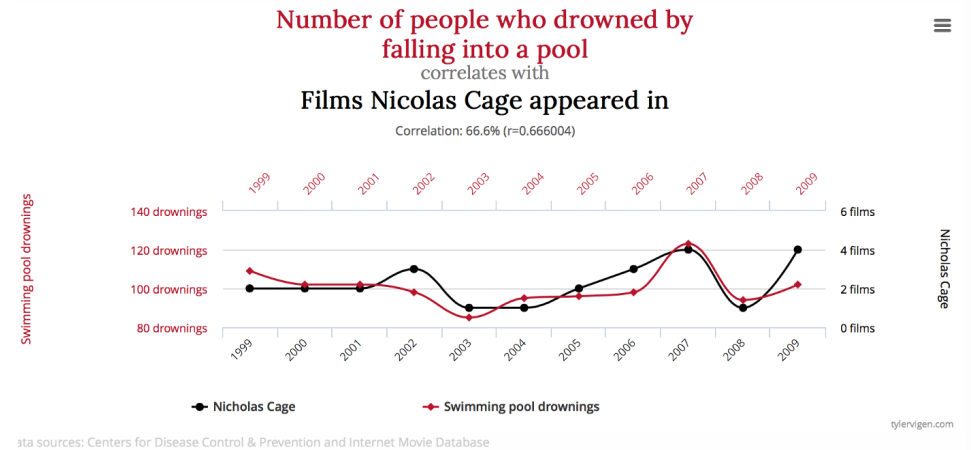
# The Correlation Coefficient



A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)



# The Correlation Coefficient



A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.



# The Correlation Coefficient



A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.



The cat didn't crush the awning

# The Regression Line

In perfect linear relationships the line that fits exactly the data have slope  $S_y/S_x$  and passes through the point  $(\bar{x}, \bar{y})$  or SD line.

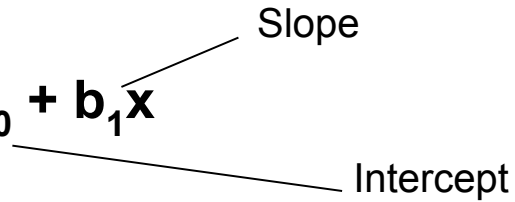
# The Regression Line

In perfect linear relationships the line that fits exactly the data have slope  $S_y/S_x$  and passes through the point  $(\bar{x}, \bar{y})$  or SD line. When there is not linear relationship:

$$Y = b_0 + b_1x$$

# The Regression Line

In perfect linear relationships the line that fits exactly the data have slope  $S_y/S_x$  and passes through the point  $(\bar{x}, \bar{y})$  or SD line. When there is not linear relationship:

$$Y = b_0 + b_1x$$


Slope

Intercept

# The Regression Line

In perfect linear relationships the line that fits exactly the data have slope  $S_y/S_x$  and passes through the point  $(\bar{x}, \bar{y})$  or SD line. When there is not linear relationship:

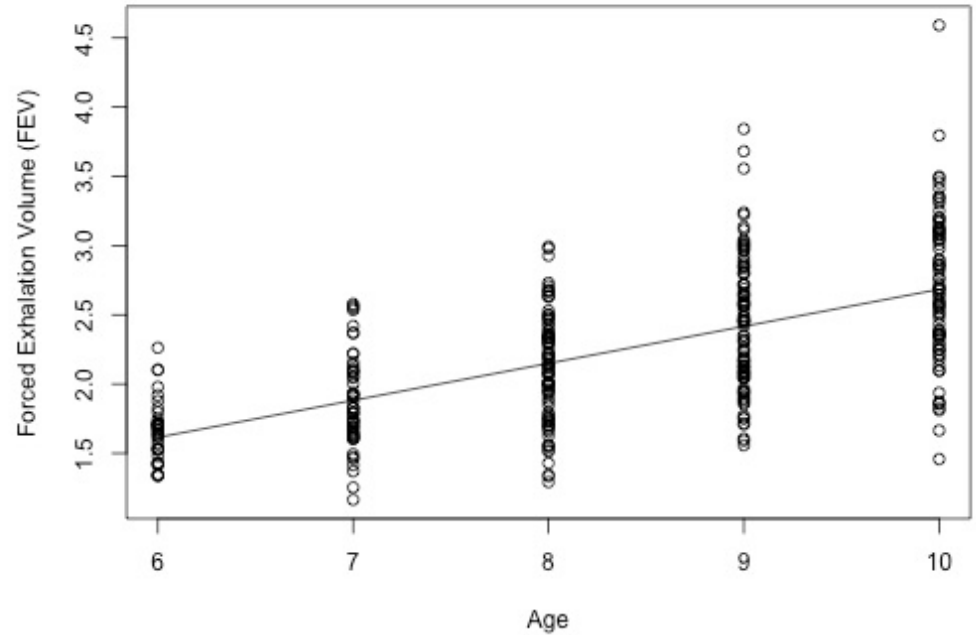
$$Y = b_0 + b_1x$$

Slope

Intercept

Intercept = 0.01165

Slope = 0.26721



# The Regression Line

In perfect linear relationships the line that fits exactly the data have slope  $S_y/S_x$  and passes through the point  $(\bar{x}, \bar{y})$  or SD line. When there is not linear relationship:

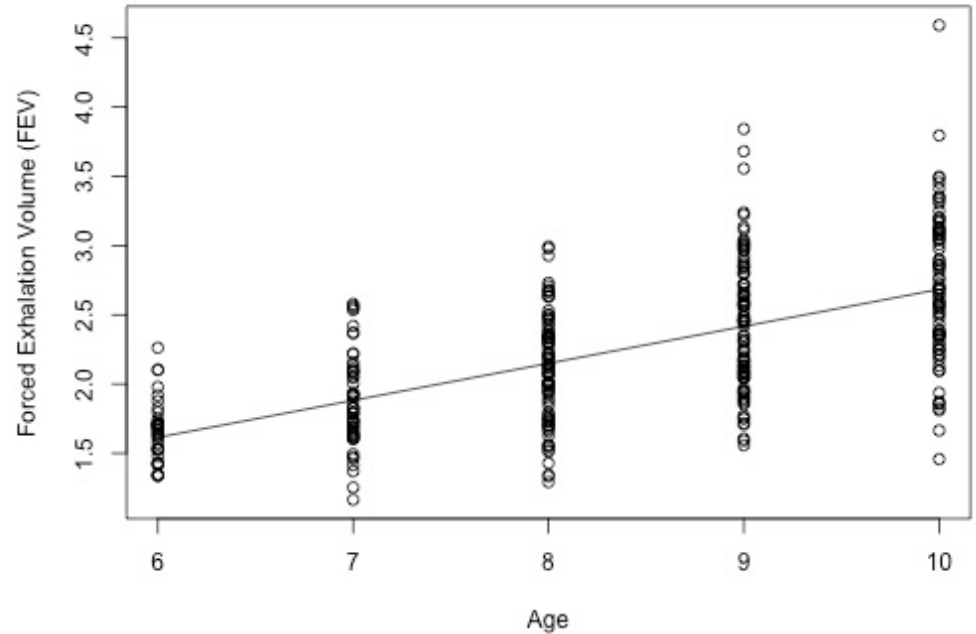
$$Y = b_0 + b_1x$$

Slope

Intercept

Intercept = 0.01165      Slope = 0.26721

$$FEV = 0.01165 + 0.26721 * \text{Age}$$





# The Regression Line

In perfect linear relationships the line that fits exactly the data have slope  $S_y/S_x$  and passes through the point  $(x,y)$  or SD line. When there is not linear relationship:

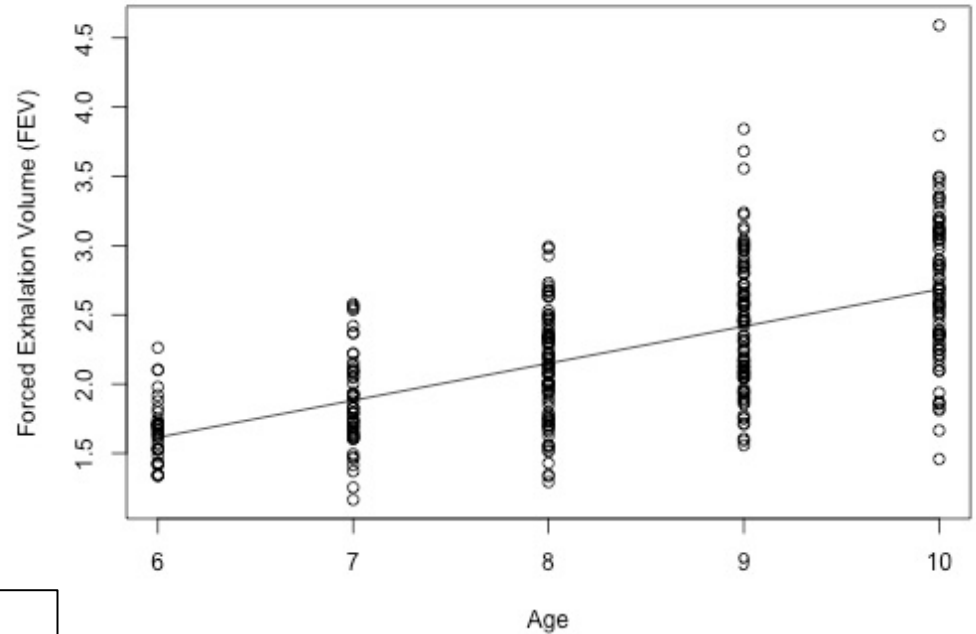
$$Y = b_0 + b_1x$$

Slope

Intercept

Intercept = 0.01165      Slope = 0.26721

$$FEV = 0.01165 + 0.26721 * Age$$



In R you can estimate slope & intercept:

```
lm(formula = Response ~ Explanatory, data = dataset)
```

# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

- Simple Regression  $\longrightarrow Y = b_0 + b_1x$

# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

- Simple Regression       $\longrightarrow$        $Y = b_0 + b_1x$
- Multiple Regression       $\longrightarrow$        $Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3\ldots$

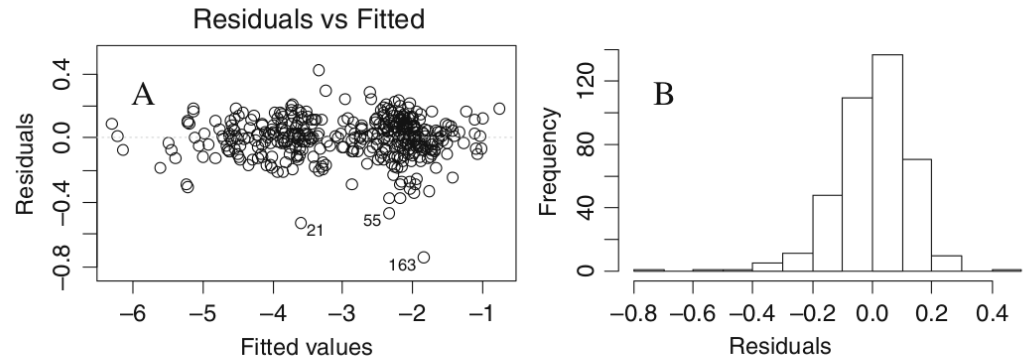
# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

- Simple Regression  $\longrightarrow Y = b_0 + b_1x$
- Multiple Regression  $\longrightarrow Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3\ldots$

## Assumption

- Linearity
- Normality of residuals
- Homoscedasticity  
(Homogeneity of variance)



# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

## **Simple regression**

# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

**Simple regression**       $\longrightarrow$       One explanatory variables

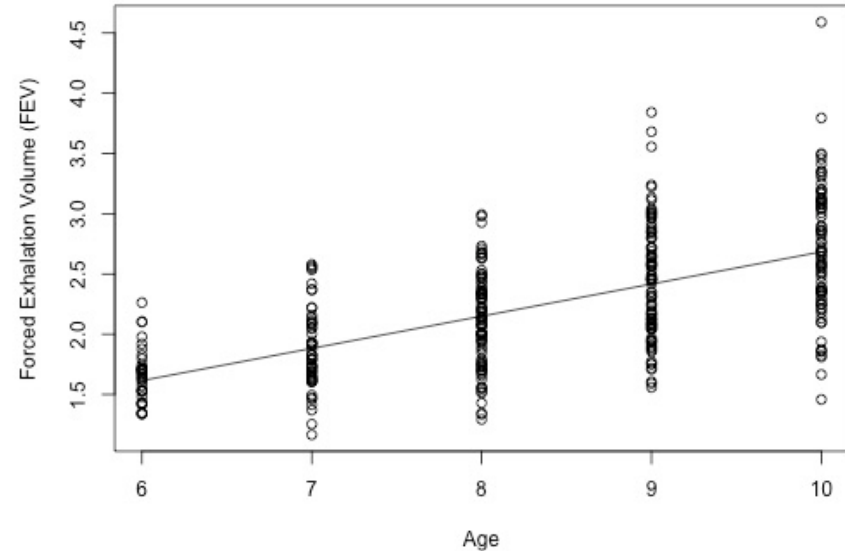
$$Y = b_1X + b_0$$

# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

**Simple regression**       $\longrightarrow$       One explanatory variables

$$Y = b_1X + b_0$$





# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

**Simple regression**  $\longrightarrow$  One explanatory variables

$$Y = b_1X + b_0$$

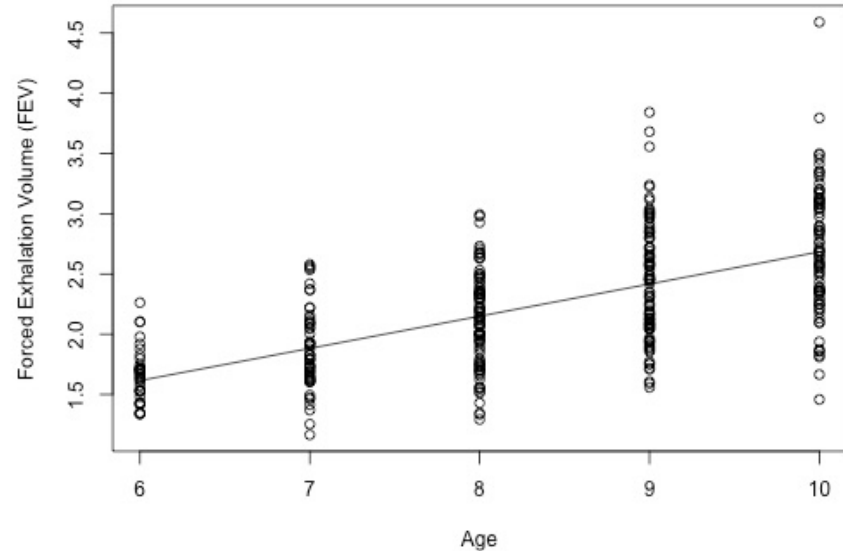
## Linear Regression

**X**



X explain Y

$X \sim Y$



# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

**Multiple regression**

# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

**Multiple regression**       $\longrightarrow$       Have multiple explanatory variables

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_0$$

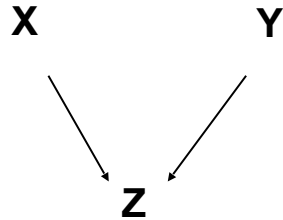
# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

**Multiple regression** → Have multiple explanatory variables

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_0$$

**Additive independent effects**



X and Y explain the variation in Z independently

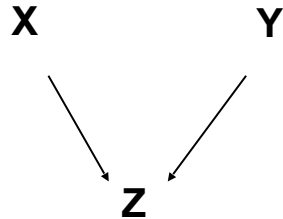
$$Z \sim X + Y$$

# (General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

**Multiple regression** → Have multiple explanatory variables  $Y = b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_0$

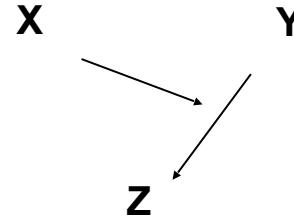
## Additive independent effects



X and Y explain the variation in Z independently

$$Z \sim X + Y$$

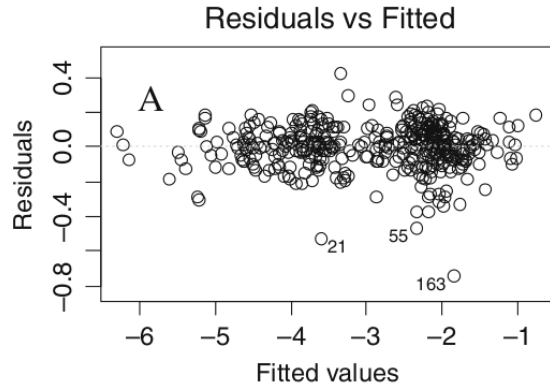
## Interaction among variable



X modifies how Y affects Z

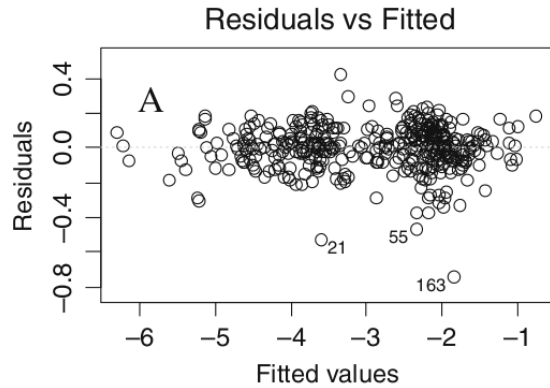
$$Z \sim X + Y + X*Y$$

# Residuals

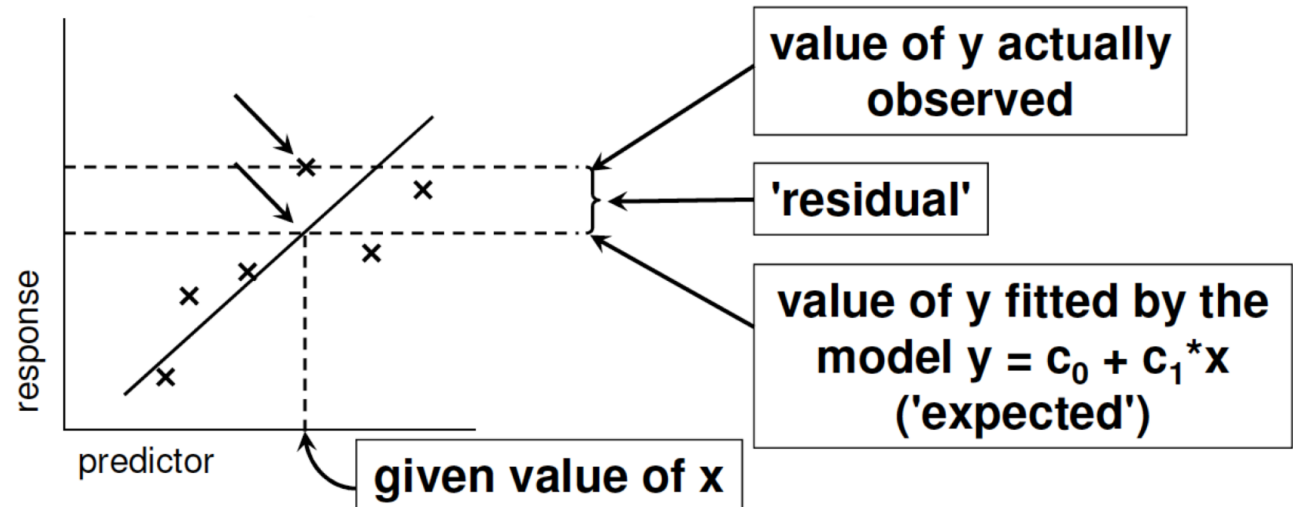


- *Residuals*: Difference between observation and fitted values
- *Fitted values*: *Estimation of an observation using all previous ones*

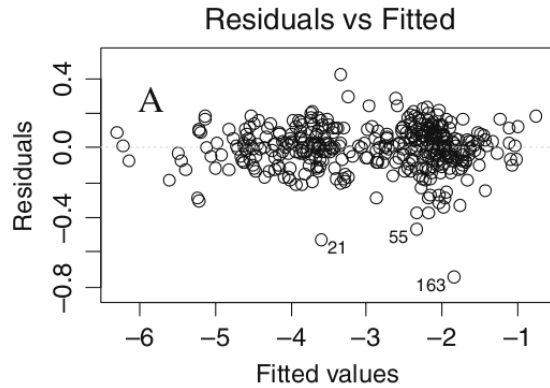
# Residuals



- *Residuals*: Difference between observation and fitted values
- *Fitted values*: Estimation of an observation using all previous ones



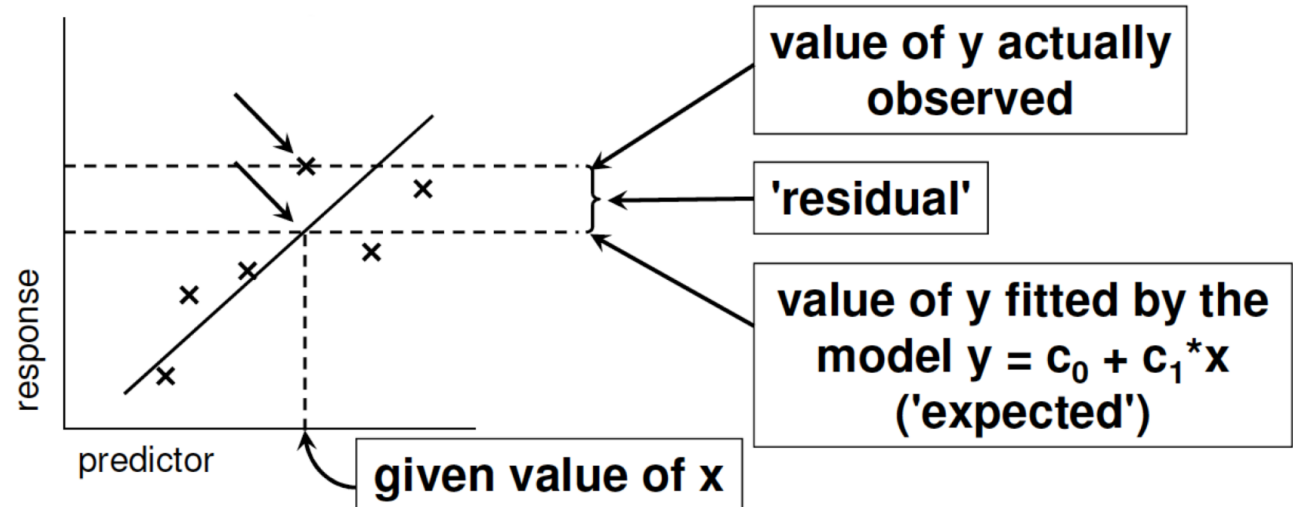
# Residuals



- *Residuals*: Difference between observation and fitted values
- *Fitted values*: Estimation of an observation using all previous ones

Error (so called Residual)

$$Y = b_1X + b_0 + e$$





# Generalized Linear Models

If we don't have the normality of residuals, we can use the Generalized Linear Models (GLM).

# Generalized Linear Models

If we don't have the normality of residuals, we can use the Generalized Linear Models (GLM).

- Can be used with residuals with distribution normal, binomial, poisson...
- Have the same features of General Linear Models

# Generalized Linear Models

If we don't have the normality of residuals, we can use the Generalized Linear Models (GLM).

- Can be used with residuals with distribution normal, binomial, poisson...
- Have the same features of General Linear Models

In R you can fit your data in a General Linear Model:

*lm(formula = Response ~ Explanatory + Z + Z\*Y, data = dataset)*

In R you can fit your data in a GLM:

*glm(formula = Response ~ Explanatory + Z + Z\*Y, family = binomial, data = dataset)*

# Summary

Model	Variables	Distribution	R code
Linear Regression	$Y = b_0 + b_1x$	Normal	<i>lm(formula, data)</i>
General Linear Models	$Y = b_0 + b_1x_1 + b_2x_2 + \dots$	Normal	<i>lm(formula, data)</i>
Generalized Linear Models (GLM)	$Y = b_0 + b_1x_1 + b_2x_2 + \dots$	Any	<i>glm(formula, family, data)</i>

