Lesson 10: Simple Regression Analysis

Victoria Blanes-Vidal

Manuella Lech Cantuaria

The Maersk Mc-Kinney Møller Institute

Applied AI and Data Science

Lesson	Week	Date	TOPICS	Teacher
1	35	1/Sep	Introduction to the course Descriptive statistics –Part I	MLC
2	36	8/sep	Descriptive statistics –Part II	MLC
3	37	15/Sep	Probability distributions	MLC
4	38	22/Sep	Hypothesis testing (one sample)	VBV
5	39	29/Sep	Hypothesis testing (two samples)	VBV
6	40	6/Oct	ANOVA one-way	VBV
7	41	13/Oct	R class (Introduction to R and descriptive statistics) Point giving activity	MLC
-	42	20/Oct	NO CLASS (Autum holidays)	
8	43	27/Oct	R class (hypothesis testing + ANOVA)	MLC
9	44	3/Nov	ANOVA two-way	VBV
-	45	10/Nov	NO CLASS	
10	46	17/Nov	Simple regression analysis	VBV
11	47	24/Nov	Experimental design Point giving activity	VBV
12	48	1/Dec	Multiple regression analysis and questions	VBV+MLC

VBV = Victoria Blanes-Vidal
MLC = Manuella Lech Cantuaria

Lesson 10

1. **Regression analysis**
2. Scatter plot
3. Correlation coefficient
4. Statistical significance of the correlation coefficient
5. Correlation vs. Causation
6. Determining the regression line
7. Plotting the regression line
8. Prediction
9. Coefficient of determination
10. In-class activities

Regression analysis

Regression Analysis is a statistical method used to describe how one (or more) quantitative variable is related to another quantitative variable.

In regression, there are **two types of variables**:

Independent variable/s, also called explanatory variable/s or predictor variable/s (input): x

Dependent variable, also called a response variable (output): y

$x \longleftrightarrow y$ **Simple regression analysis**

$x_1, x_2, x_3, x_4, \ldots \longleftrightarrow y$ **Multiple regression analysis**

Example: Basketball players

We want to study the relation between weight and height of NBA basketball players.

We randomly select 10 players and measure their height and weight.

Example: Basketball players

```
> Basketball_players <-
read.table("C:/Users/vbv/Desktop/My_documents/Teaching/Teaching/Sta
tistical_Data_Analysis/2022_2023/Exercises_in_R/Basketball_players.
txt", header=TRUE)
> Basketball_players
   X_Player_height Y_Player_weight
1           190.05           88.00
2           190.72           88.89
3           191.48           90.11
4           194.21           91.96
5           199.63           97.02
6           201.76          102.93
7           202.99          103.72
8           204.75          105.27
9           209.45          111.61
10          210.60          112.46
```

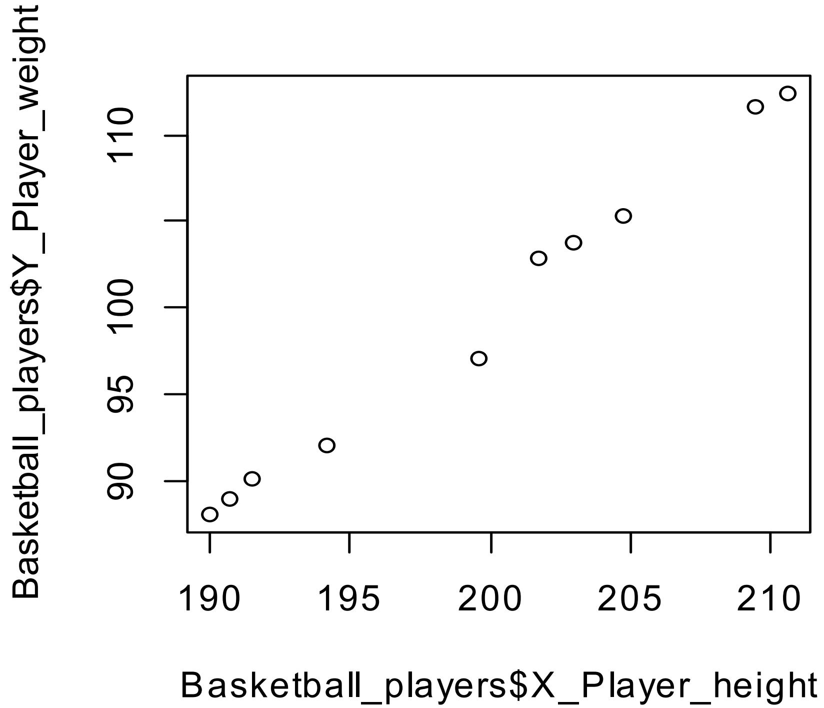
Lesson 10

1. **Regression analysis**
2. **Scatter plot**
3. Correlation coefficient
4. Statistical significance of the correlation coefficient
5. Correlation vs. Causation
6. Determining the regression line
7. Plotting the regression line
8. Prediction
9. Coefficient of determination
10. In-class activities

Scatter plot

- The relation between two quantitative variables can be visualized by plotting a ***scatter plot***.
- A scatter plot is a graph of the x and y pairs.

```
> plot(Basketball_players$X_Player_height,Basketball_players$Y_Player_weight)
```



- Usually, the variable *x* is plotted on the horizontal axis, and the variable *y* is plotted on the vertical axis.

Scatter plot

- The scatter plot is a visual way to describe the nature of the relationship between the two variables.

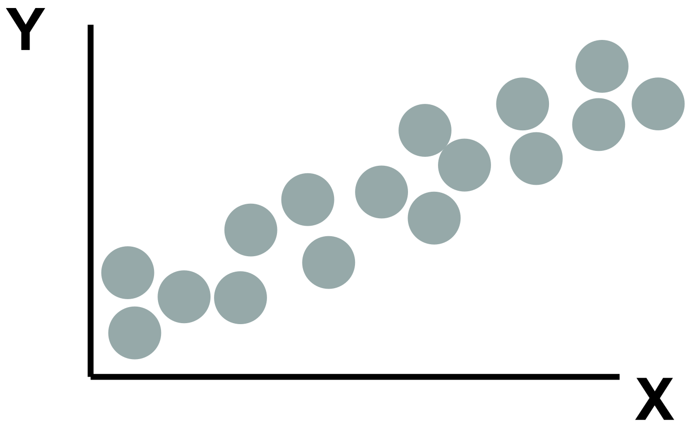
- A scatterplot displays the:

  - Form: Linear vs. non-linear
  - Direction: Positive vs. Negative
  - Strength: Weak vs. Strong

of the relationship between two variables (x and y).

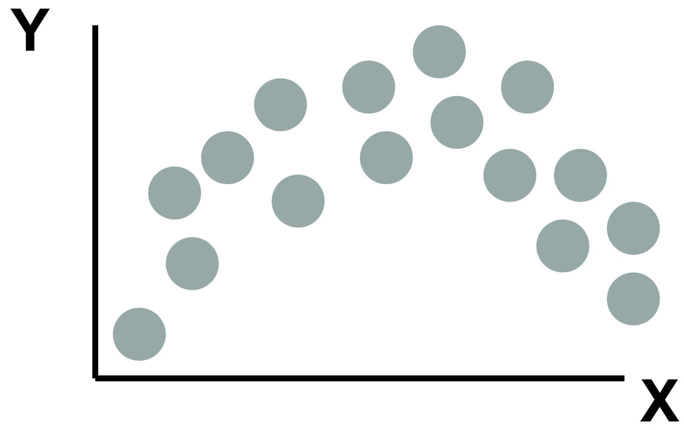
Form

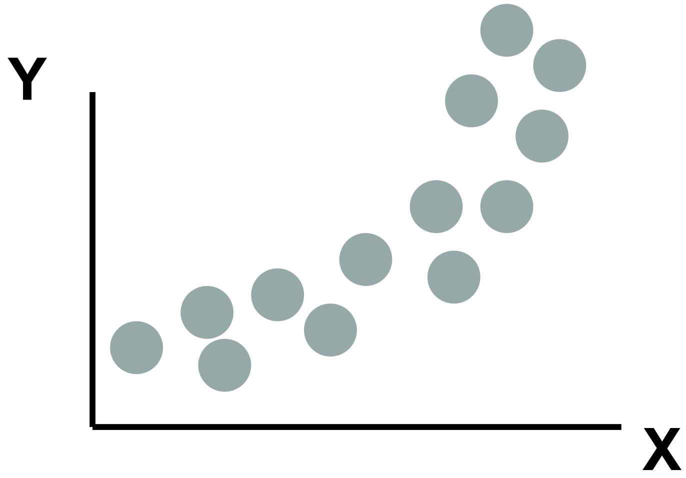
Linear



When the relationthip between X and Y is linear, the scatter plot gives a **straight line**.

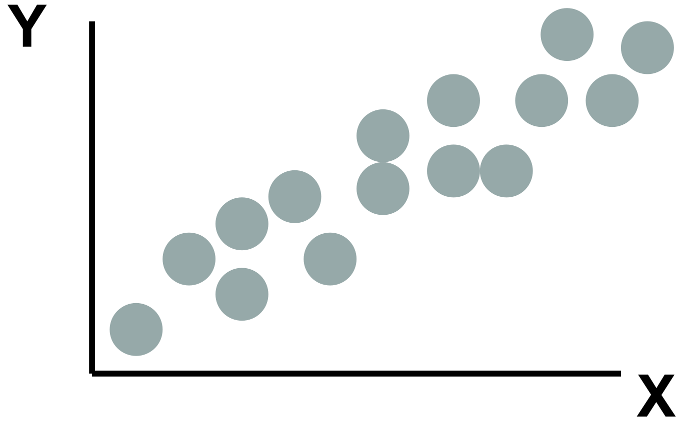
Non-linear





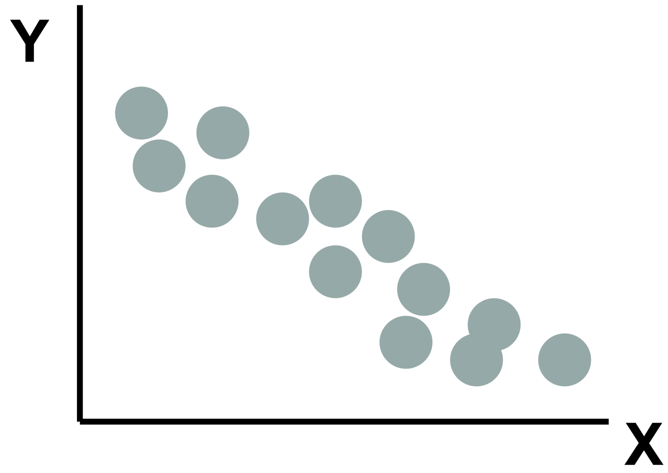
Direction

Positive



If there is a positive linear relationship, as **X increases, Y increases.**

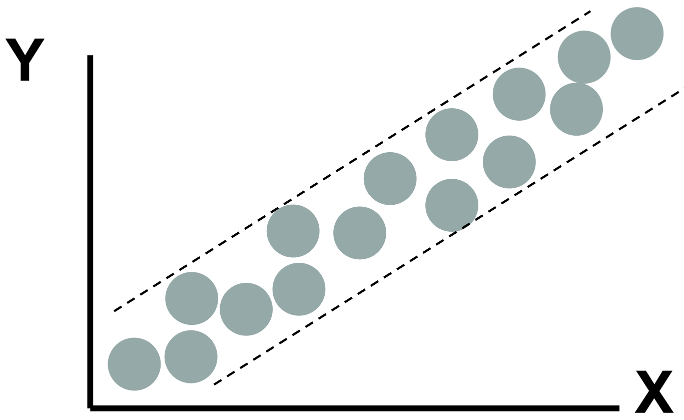
Negative



If there is a negative linear relationship, as **X increases, Y decreases.**

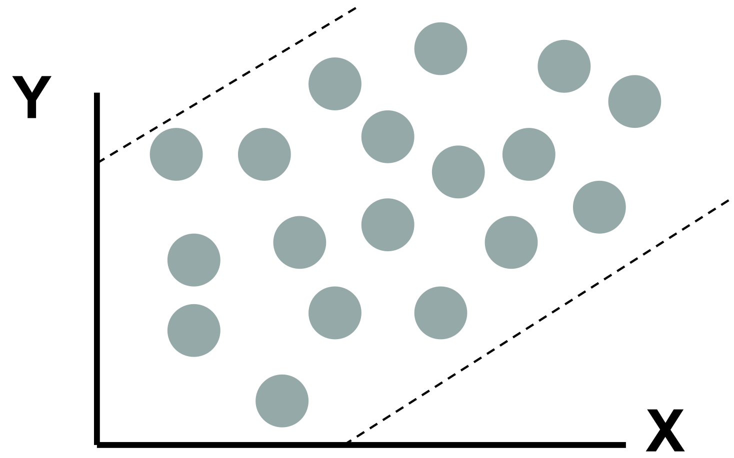
Strength

Strong

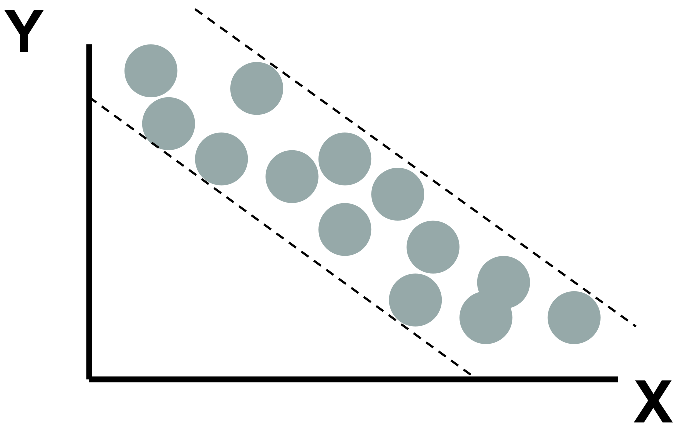


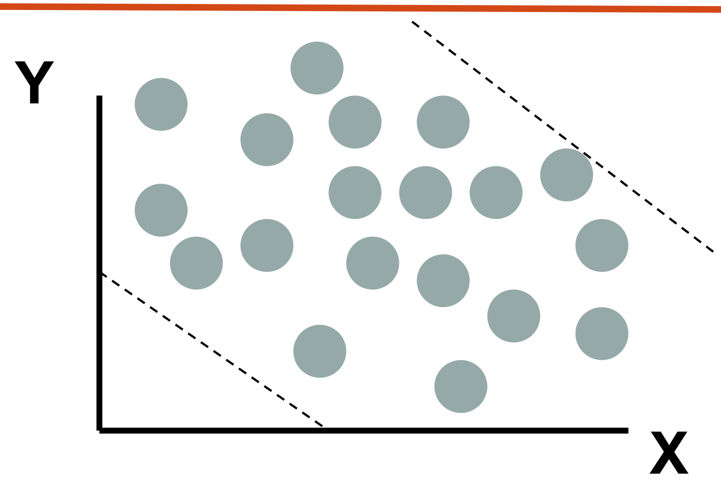
If there is a strong linear positive relationship, when the value of X increases, the value of Y increases in a reliable manner.

Weak

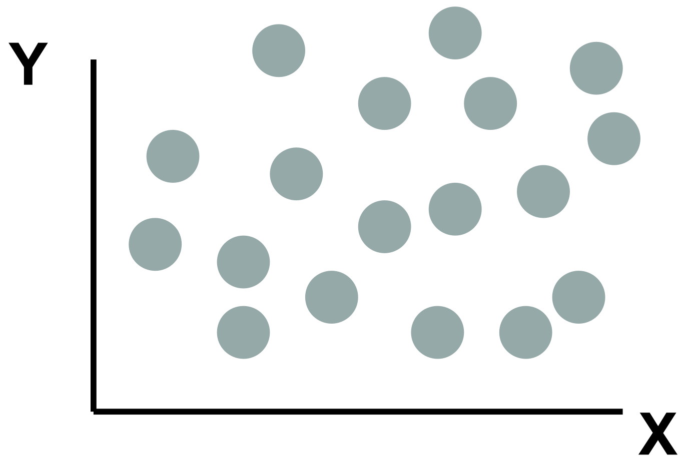


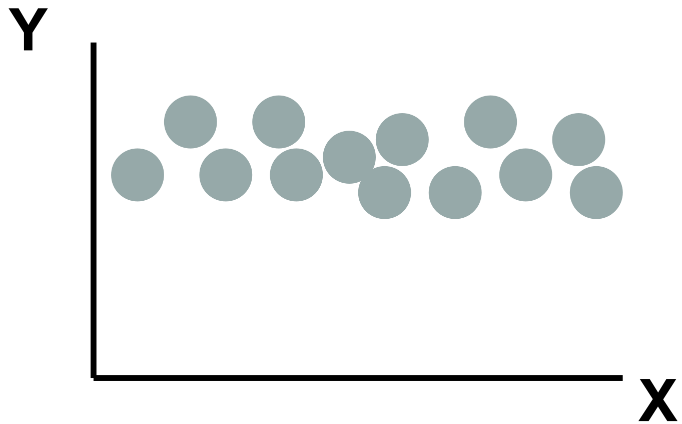
If there is a weak linear positive relationship, when the value of X increases, the other variable tends to increase as well, but in a weak or unreliable manner.





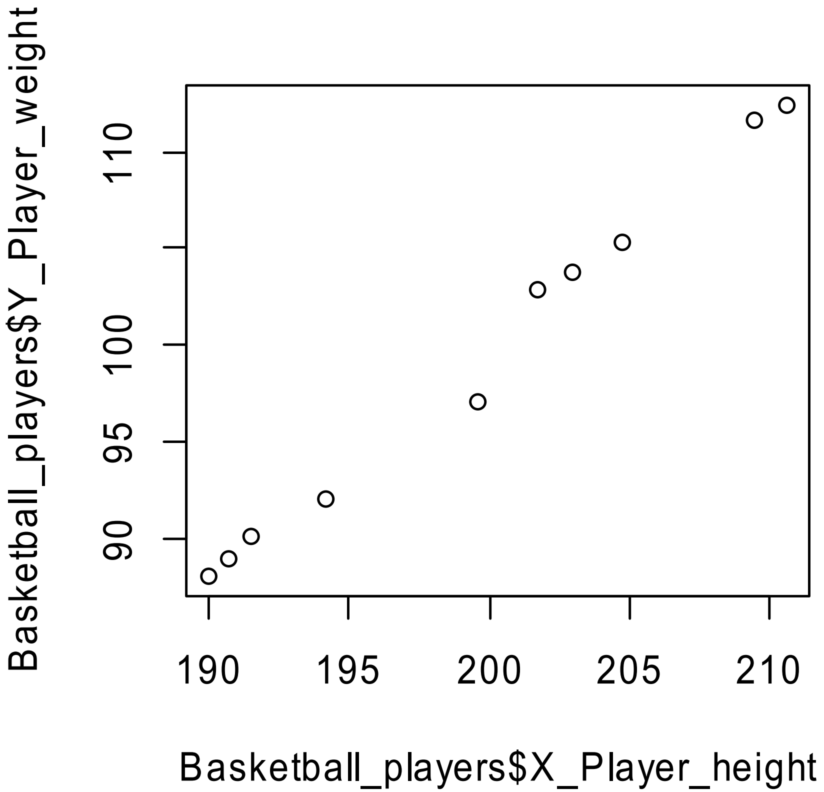
No relationship





Example: Basketball players

```
> plot(Basketball_players$X_Player_height,Basketball_players$Y_Player_weight)
```



What is the form, direction and strength of the relationship between basketball players height and weight?

Form	Linear	Non-linear
Direction	Positive	Negative
Strength	Weak	Strong

Lesson 10

1. **Regression analysis**
2. **Scatter plot**
3. **Correlation coefficient**
4. Statistical significance of the correlation coefficient
5. Correlation vs. Causation
6. Determining the regression line
7. Plotting the regression line
8. Prediction
9. Coefficient of determination
10. In-class activities

Correlation coefficient

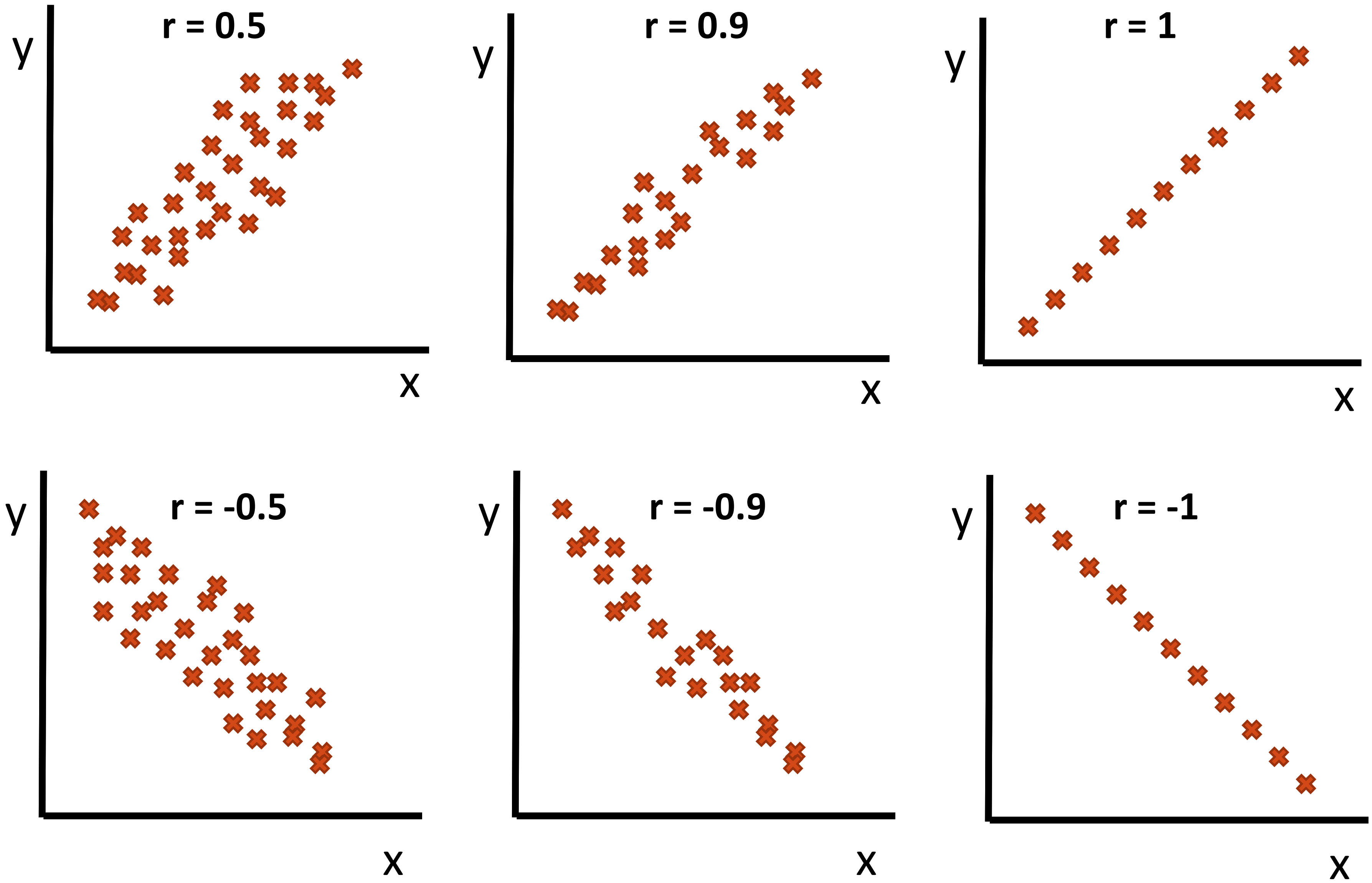
- The **correlation coefficient (r)** measures the strength and direction of a linear relationship between two variables.
- The range of the correlation coefficient (r) is **from -1 to 1**.

If there is a strong negative linear relationship between the variables, the value of r will be close to -1.

When there is no linear relationship between the variables or only a very weak relationship, the value of r will be close to 0.

If there is a strong positive linear relationship between the variables, the value of r will be close to 1.

Scatter plots and correlation coefficients

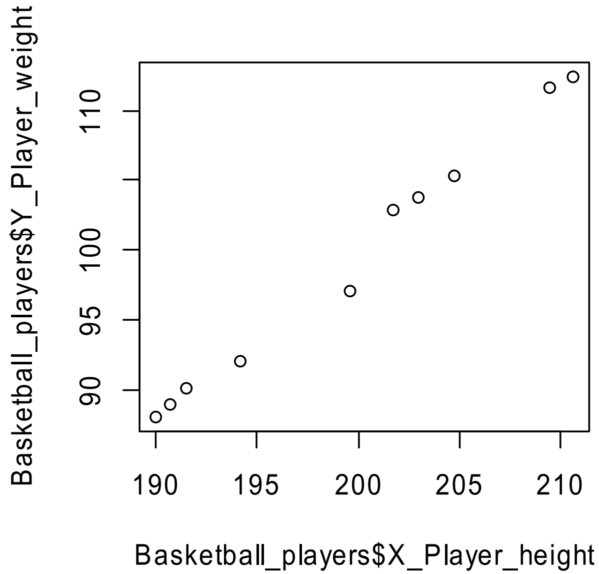


Example: Basketball players

```
> cor(Basketball_players$X_Player_height,Basketball_players$Y_Player_weight)
[1] 0.9948415
```

Based on the correlation coefficient (r = 0.994), the direction and strength of the relationship between basketball players height and weight is…

Direction	Positive	Negative
Strength	Weak	Strong



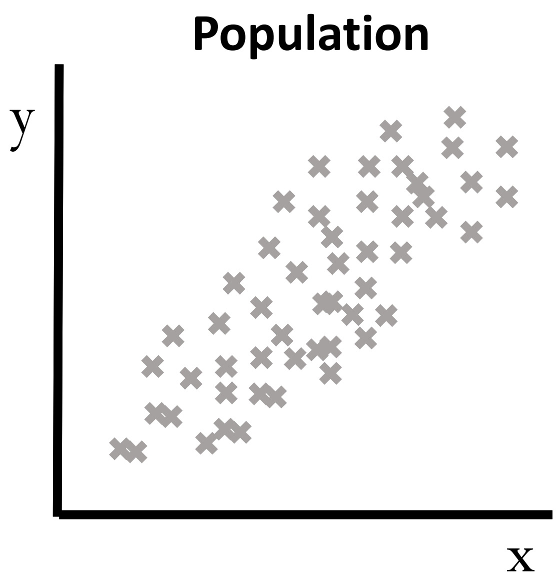
Lesson 10

1. **Regression analysis**
2. **Scatter plot**
3. **Correlation coefficient**
4. **Statistical significance of the correlation coefficient**
5. Correlation vs. Causation
6. Determining the regression line
7. Plotting the regression line
8. Prediction
9. Coefficient of determination
10. In-class activities

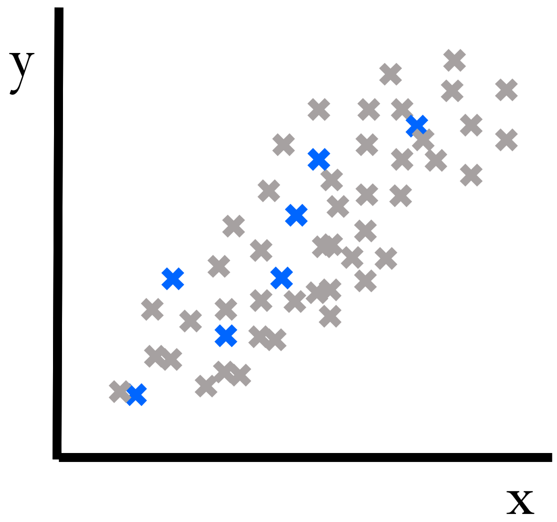
Statistical Significance of the Correlation Coefficient

- What we would like to know is whether there is a relation between X and Y <u>in the population</u> (not in the sample).

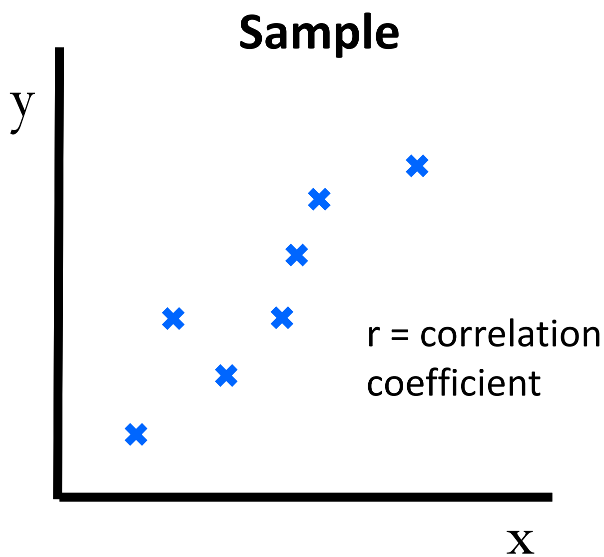
- However, the correlation coefficient (r) is calculated from data obtained from a sample (not from the population).



Out of all members of the population…



We have selected (sampled) only some of them, and measured X and Y only for those



We have calculated the correlation coefficient (r) using the X and Y measured in that sample (7 points)

Statistical Significance of the Correlation Coefficient

If the correlation coefficient in the sample (r) shows that there is a relationship between X and Y in the sample:

Does that mean that there is a relation between X and Y in the population?

We need to use a hypothesis test, where:

Null hypothesis (H0):
There is not a relation between X and Y in the population

Alternative hypothesis (H1):
There is a relation between X and Y in the population

Statistical Significance of the Correlation Coefficient


```
> Basketball_players
   X_Player_height Y_Player_weight
1           190.05           88.00
2           190.72           88.89
3           191.48           90.11
4           194.21           91.96
5           199.63           97.02
6           201.76          102.93
7           202.99          103.72
8           204.75          105.27
9           209.45          111.61
10          210.60          112.46
```

```
> 
cor.test(Basketball_players$X_Player_height,Basketball_players$Y_Player_weight)

        Pearson's product-moment correlation

data:  Basketball_players$X_Player_height and Basketball_players$Y_Player_weight
t = 27.738, df = 8, p-value = 3.079e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9775002 0.9988253
sample estimates:
      cor 
0.9948415 
```

This is the correlation coefficient between height and weight in the sample.

This is the p-value of the statistical significance of the correlation coefficient. In this case, the hypothesis test shows that there is a significant relation between X and Y <u>in the population</u> (p-value<0.05).

Note that the strength and the significance of the relationship between X and Y, are two different things:

A relationship can be **strong** for the sample (r close to 1 or -1), and yet **not significant** for the population, because the result of the statistical test **depends also on the sample size**.

Lesson 10

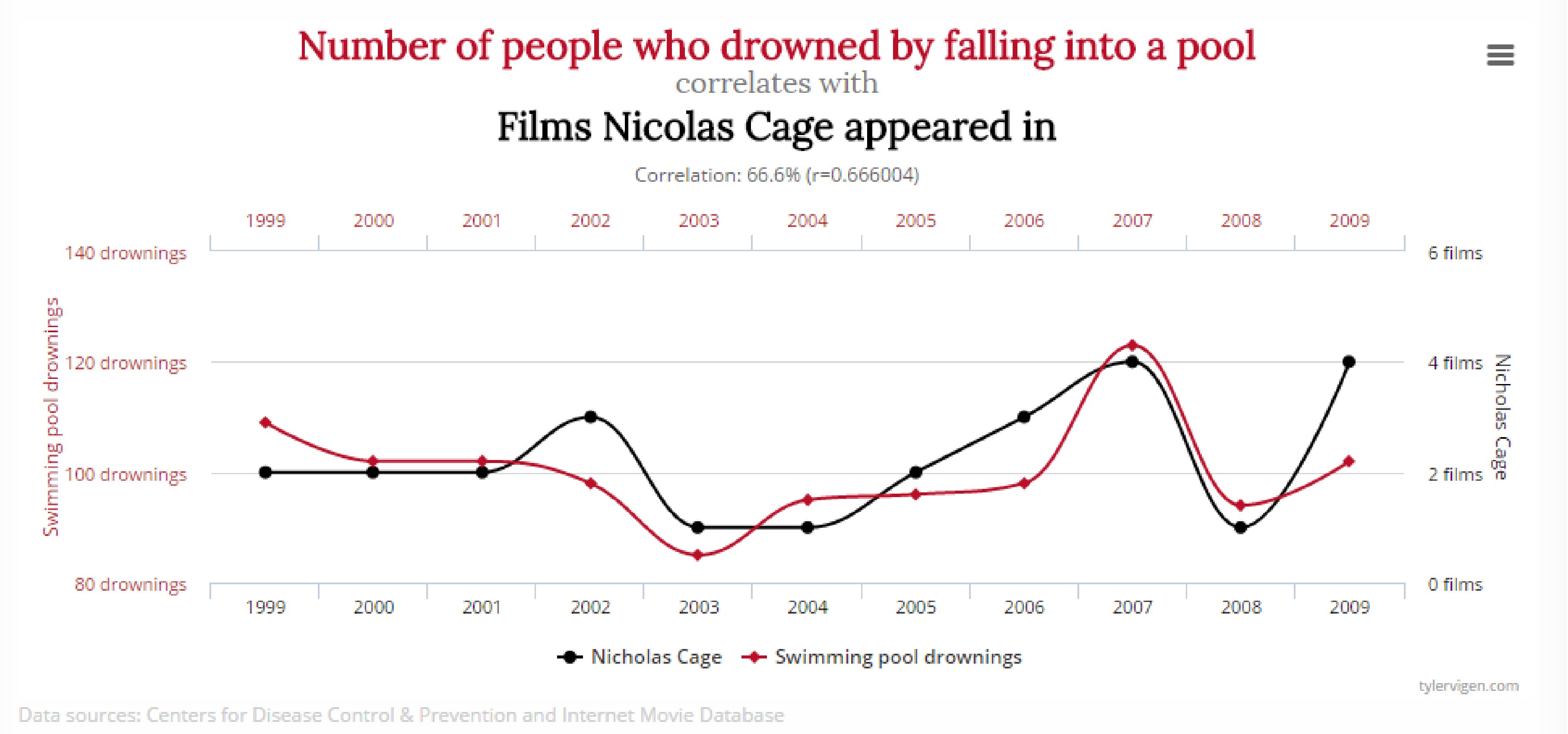
1. **Regression analysis**
2. **Scatter plot**
3. **Correlation coefficient**
4. **Statistical significance of the correlation coefficient**
5. **Correlation vs. Causation**
6. Determining the regression line
7. Plotting the regression line
8. Prediction
9. Coefficient of determination
10. In-class activities

Correlation vs. causation

When analyzing data, many people confuse the concepts of correlation and causation.

- **Correlation** - When researchers find a correlation, which can also be called an association, what they are saying is that they found **a significant relationship between two variables**.

- **Causation** - When an article says that causation was found, this means that the researchers found that **changes in one variable they measured, *directly caused* changes in the other**.



Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
Correlation: 66.6% (r=0.666004)
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com
Data sources: Centers for Disease Control & Prevention and Internet Movie Database

```
> Nicolas_cage_drowning <-
read.table("C:/Users/vbv/Desktop/My_documents/Teaching/Teaching/Statistical_D
ata_Analysis/2022_2023/Exercises_in_R/Nicolas_cage.txt", header=TRUE)
> Nicolas_cage_drowning
   Nicolas_cage_films People_drowned
1                   2            110
2                   2            103
3                   2            101
4                   3             98
5                   1             85
6                   1             93
7                   2             95
8                   3             98
9                   4            122
10                  1             95
11                  4            102

cor.test(Nicolas_cage_drowning$Nicolas_cage_films,Nicolas_cage_drowning$People_
drowned)

        Pearson's product-moment correlation

data:  Nicolas_cage_drowning$Nicolas_cage_films and
Nicolas_cage_drowning$People_drowned
t = 2.6502, df = 9, p-value = 0.02647
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1031497 0.9032183
sample estimates:
      cor
0.6620586
```

There is a significant positive relation between Nicolas Cage films and People drowned in swimming pools (p-value<0.05).

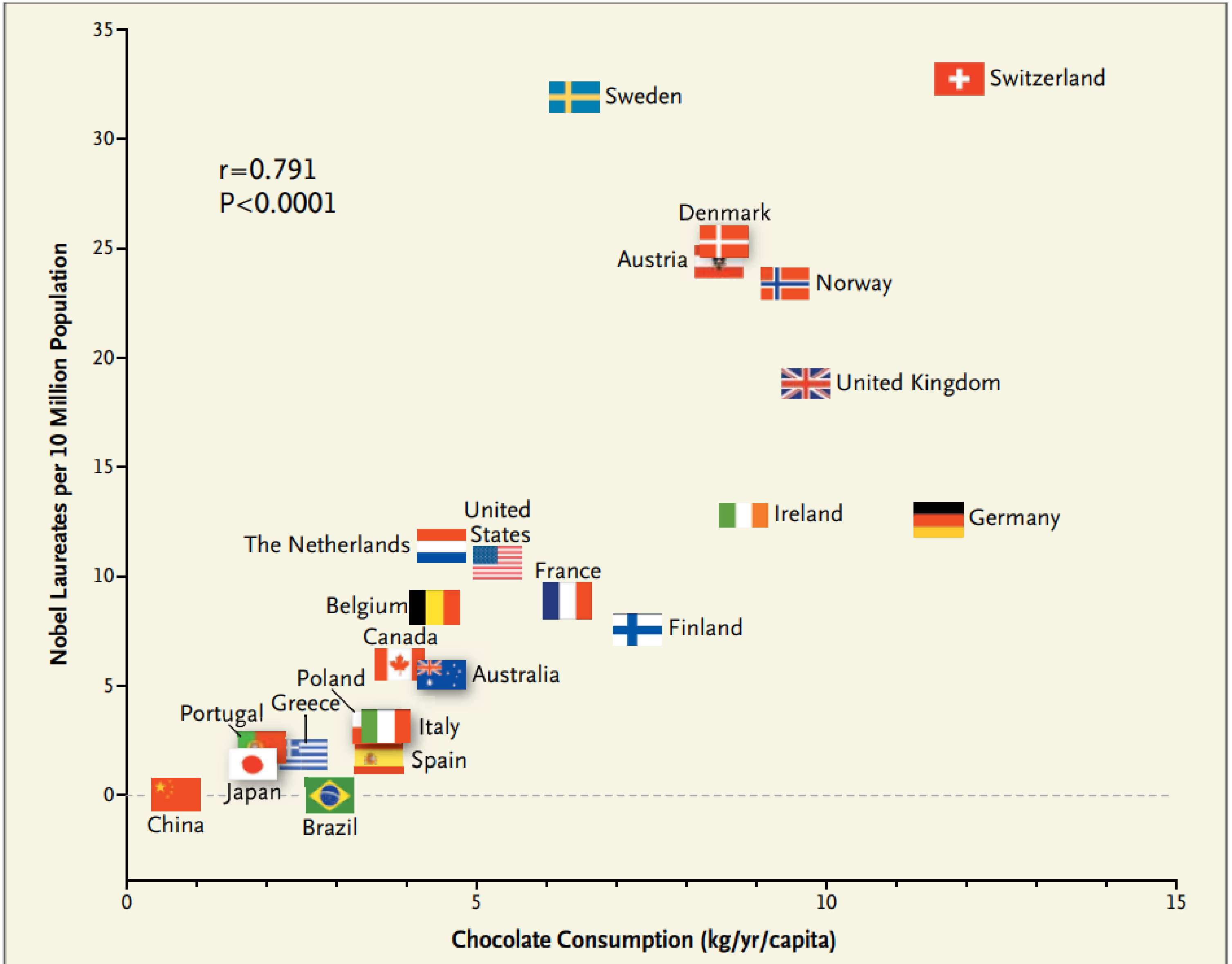


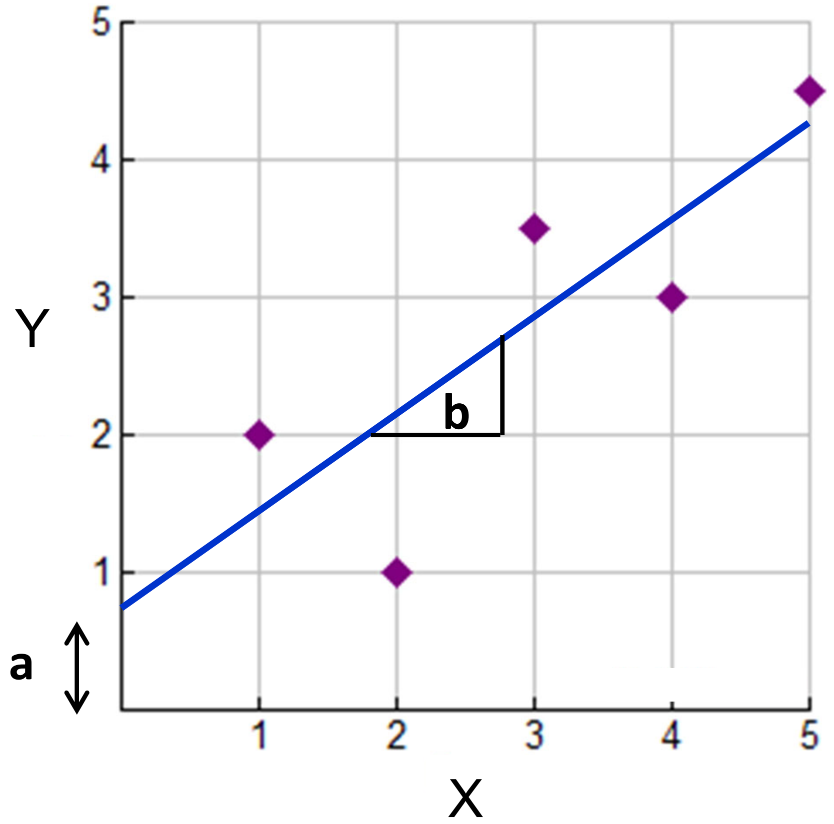
Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Lesson 10

1. **Regression analysis**
2. **Scatter plot**
3. **Correlation coefficient**
4. **Statistical significance of the correlation coefficient**
5. **Correlation vs. Causation**
6. **Determining the regression line**
7. Plotting the regression line
8. Prediction
9. Coefficient of determination
10. In-class activities

Determining the regression line

- If the value of the correlation coefficient is significant, the next step is to determine the equation of the **regression line**.



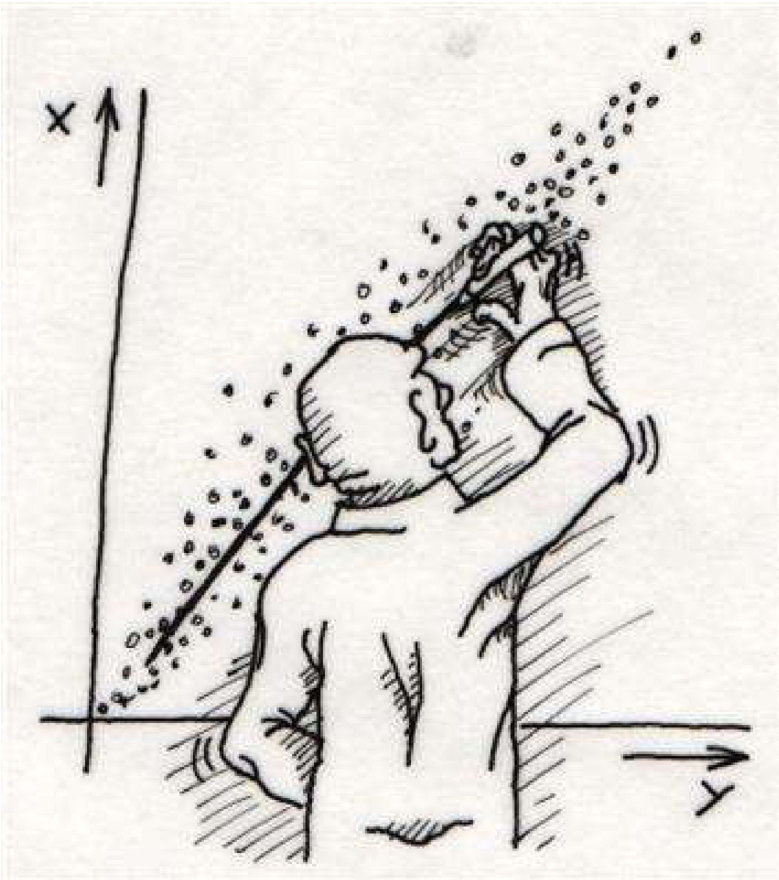
a = intercept

b = slope

The equation of the regression line is:

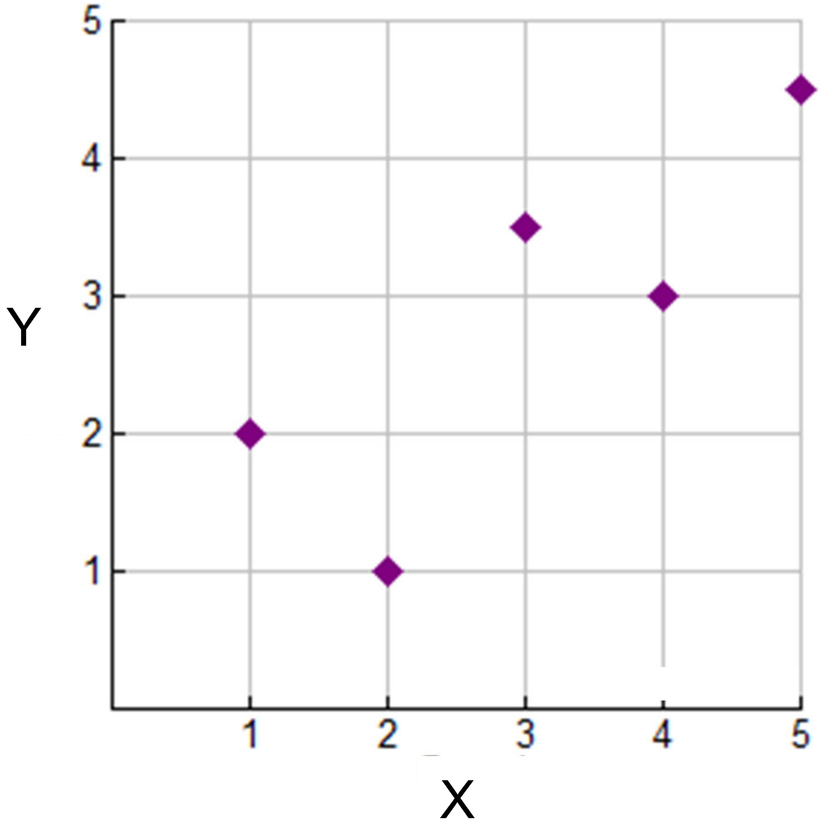
$$Y = a + b * X$$

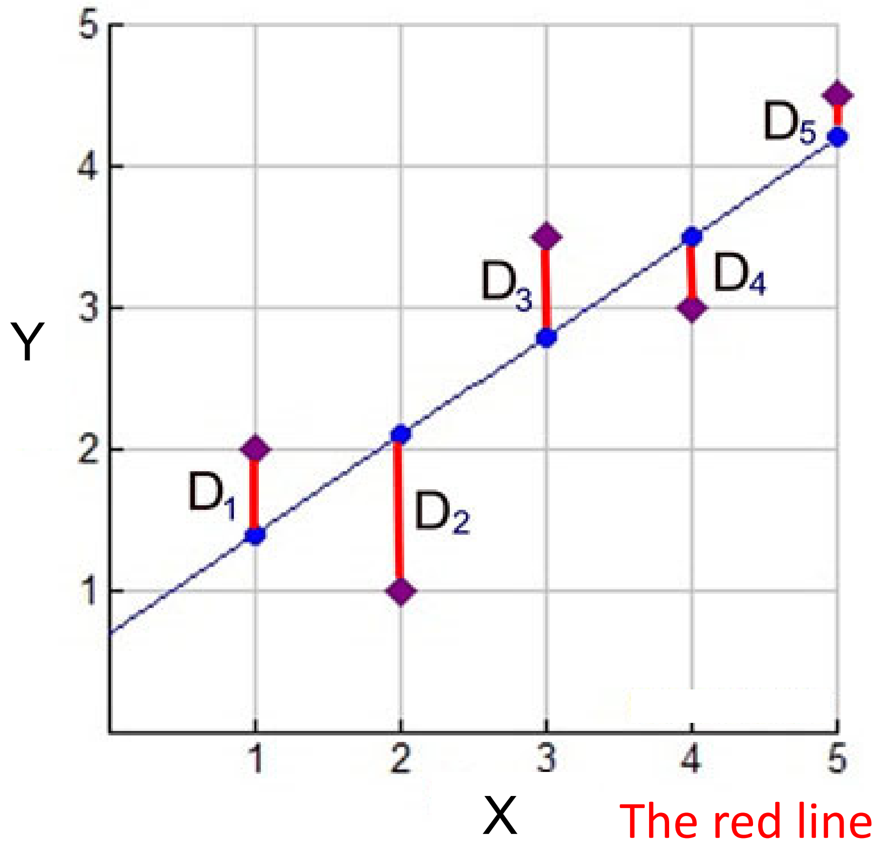
Determining the regression line



Out of all the possible straight lines we could draw…

which one is the regression line?





The red line segments are called **residuals.**

The regression line is:
<u>the line with the minimum sum of the squares of the residuals</u>.

$$\min\left(\sum D_n^2\right)$$

$$\textit{In the example}: \ \min(D_1^2 + D_2^2 + D_3^2 + D_4^2 + D_5^2)$$

Determining the regression line

```
> Regression_Basketball <- lm(Y_Player_weight~X_Player_height,
data=Basketball_players)
> summary(Regression_Basketball)

Call:
lm(formula = Y_Player_weight ~ X_Player_height, data =
Basketball_players)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2566 -0.1474  0.3176  0.4659  1.0847

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -141.47114    8.68205  -16.30 2.03e-07 ***
X_Player_height   1.20597    0.04348   27.74 3.08e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9969 on 8 degrees of freedom
Multiple R-squared:  0.9897,	Adjusted R-squared:  0.9884
F-statistic: 769.4 on 1 and 8 DF,  p-value: 3.079e-09
```

$Y = a + b * X$

$a = -141.5$

$b = 1.206$

$Y = -141.5 + 1.206 * X$

$Weight = -141.5 + 1.206 * Height$

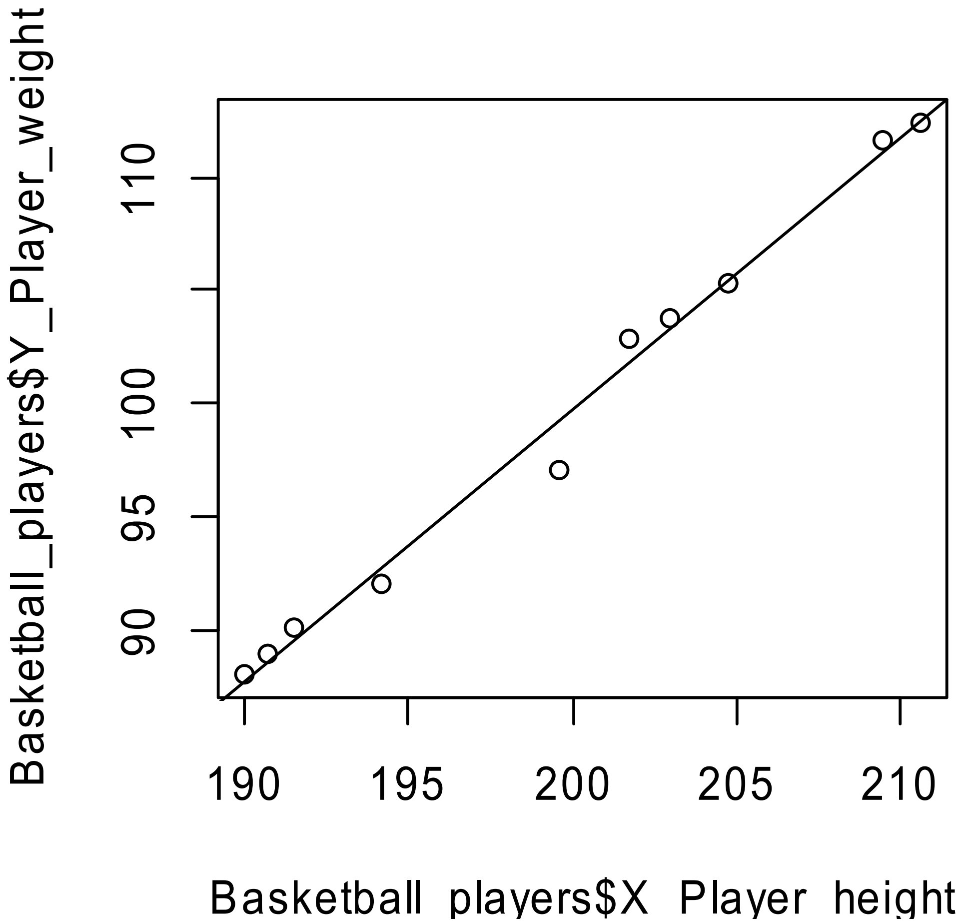
Lesson 10

1. **Regression analysis**
2. **Scatter plot**
3. **Correlation coefficient**
4. **Statistical significance of the correlation coefficient**
5. **Correlation vs. Causation**
6. **Determining the regression line**
7. **Plotting the regression line**
8. Prediction
9. Coefficient of determination
10. In-class activities

Plotting the regression line

Example: Basketball players

```
> plot(Basketball_players$X_Player_height,Basketball_players$Y_Player_weight)
> abline(Regression_Basketball)
```



Lesson 10

1. **Regression analysis**
2. **Scatter plot**
3. **Correlation coefficient**
4. **Statistical significance of the correlation coefficient**
5. **Correlation vs. Causation**
6. **Determining the regression line**
7. **Plotting the regression line**
8. **Prediction**
9. **Coefficient of determination**
10. **In-class activities**

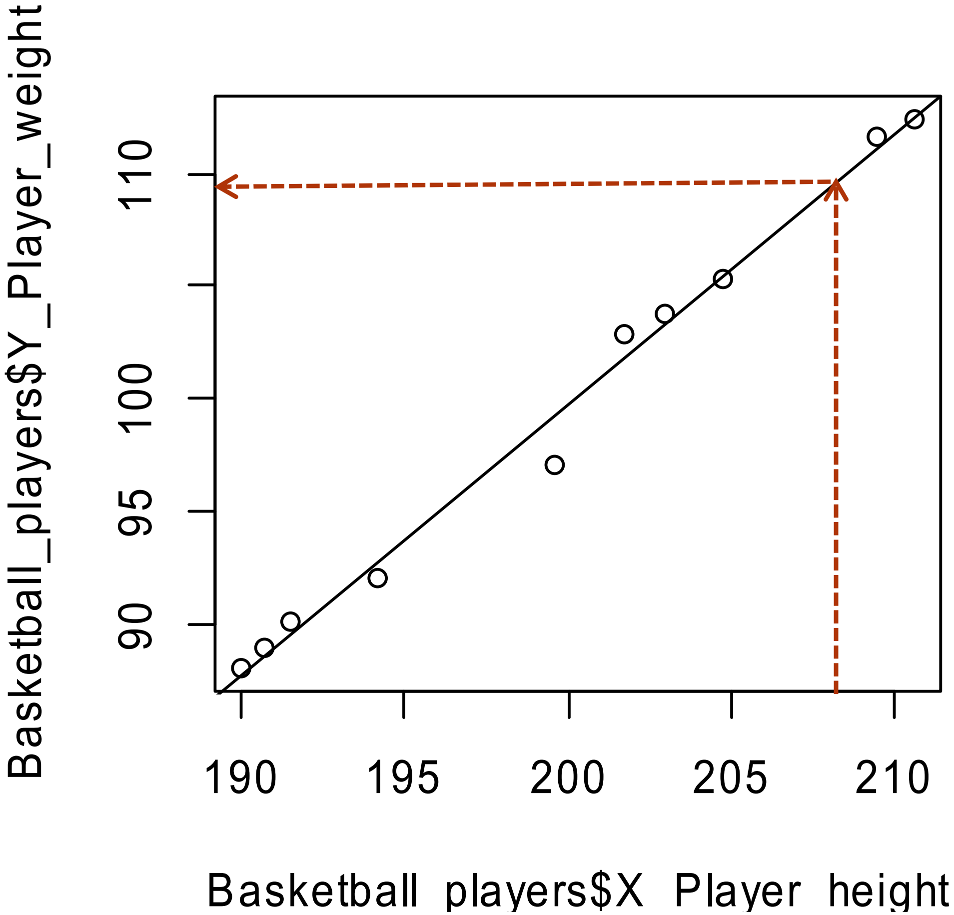
Prediction

The equation of the **regression line** can be used for **prediction**.

Example: Basketball players

We measure the height of a new basketball player: Height = 208 cm

What is the predicted weight?



The equation of the **regression line** can be used for **prediction**.

Example: Basketball players

We measure the height of a new basketball player: Height = 208 cm

What is the predicted weight?

Weight = -141.5 + 1.206 * Height

Weight = -141.5 + 1.206 * 208

Weight = 109.35 kg

```
> predict(Regression_Basketball,data.frame(X_Player_height=208))
       1
109.3706
```

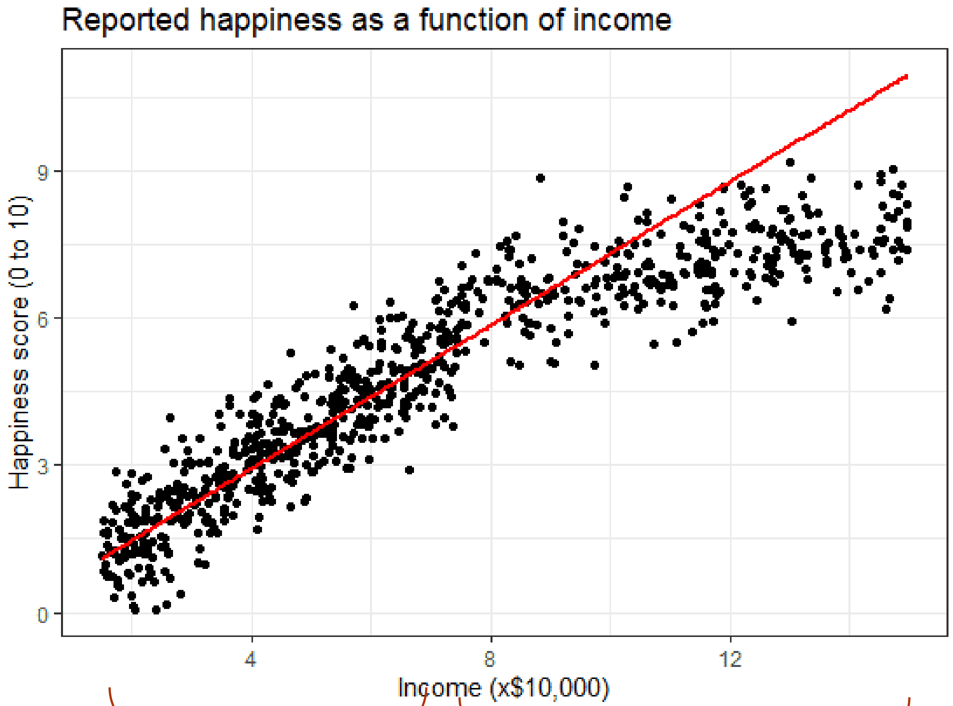
Prediction

Can you predict values outside the range of your data?

No.

We said that the regression equation can be used to **predict** the value of the dependent variable at certain values of the independent variable.

However, this is only true for the range of values where we have actually measured.



The regression line was calculated using these values

It cannot be used to predict Y for this range of values of X

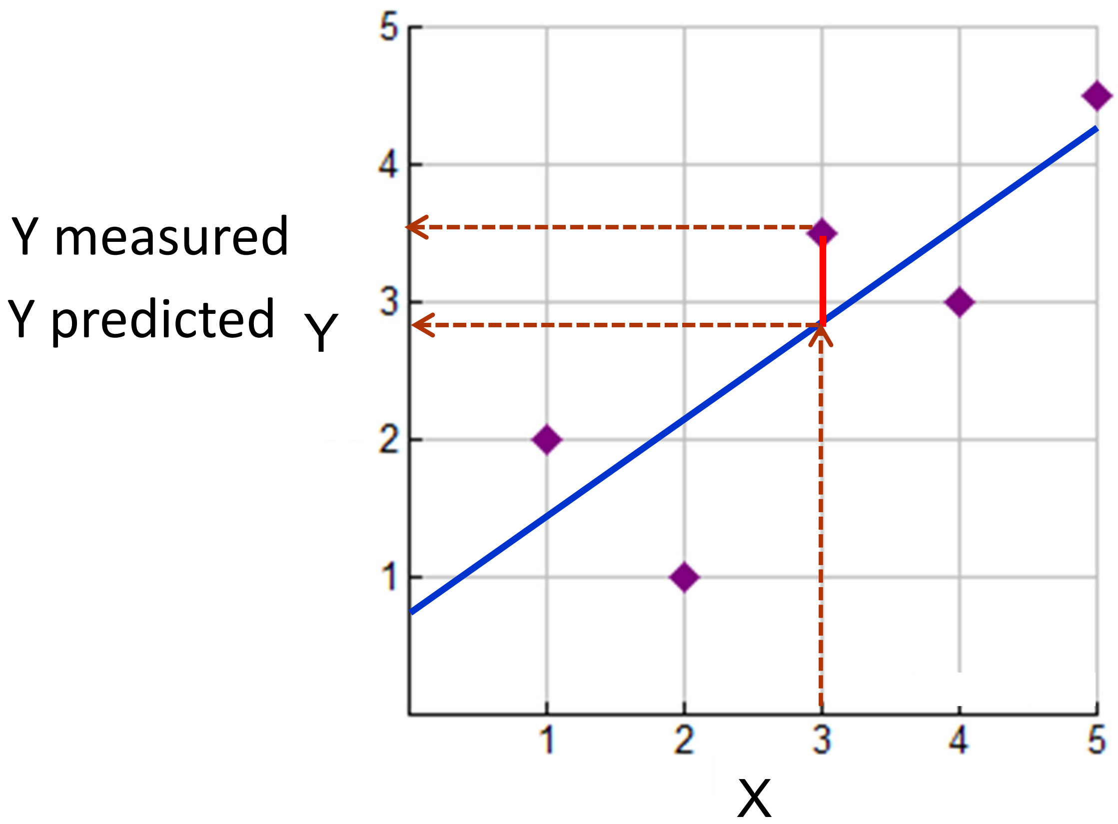
Lesson 10

1. Regression analysis
2. Scatter plot
3. Correlation coefficient
4. Statistical significance of the correlation coefficient
5. Correlation vs. Causation
6. Determining the regression line
7. Plotting the regression line
8. Prediction
9. Coefficient of determination
10. In-class activities

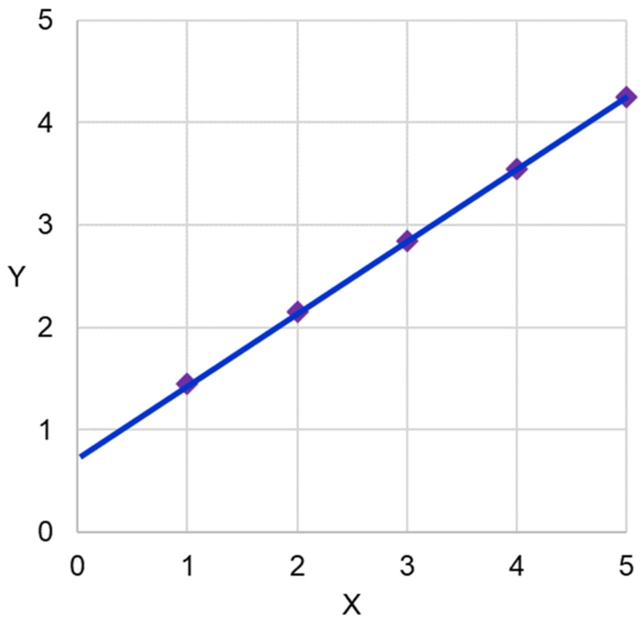
Coefficient of determination (R^2)

There is a difference between the location of a specific point in the scatter plot (Y measured), and what the regression equation predicts (Y predicted).

This is called residual.



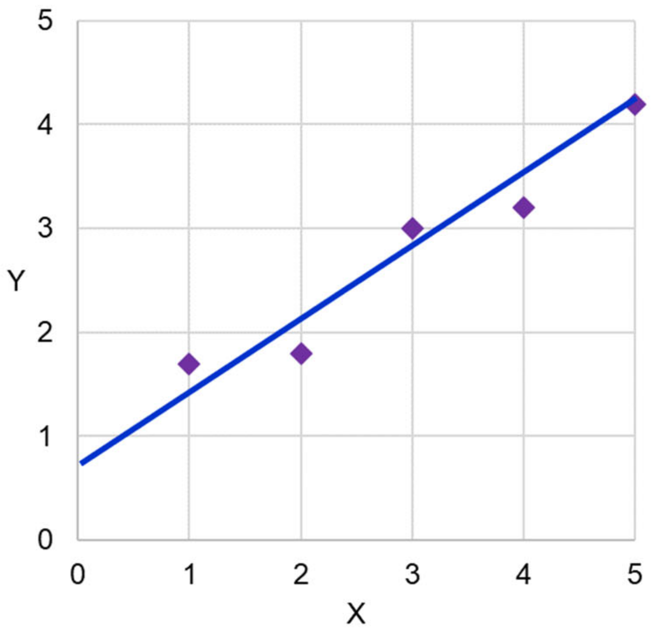
The size of these residuals provide information about the existence of other variables (variables other than X), that also affect the Y (and that have not been included in the regression analysis).



If the residuals are zero, that means that:

”X explains ALL the variation we see in Y”

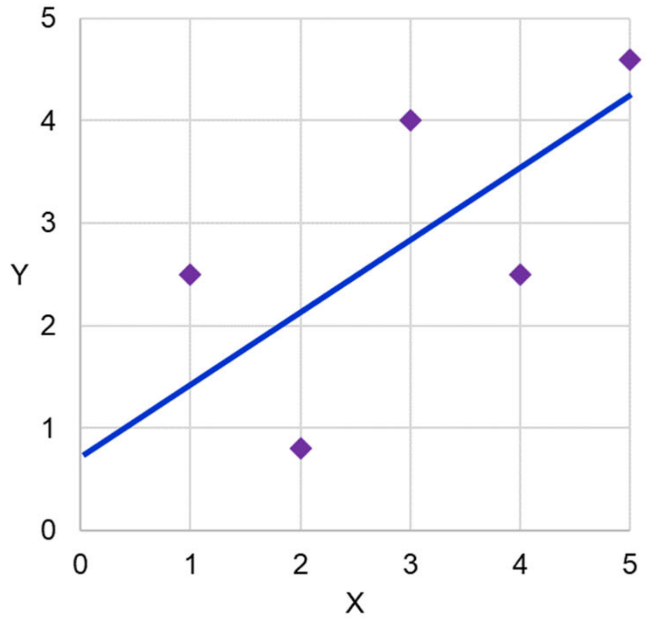
X perfectly predicts Y



If the residuals are small, that means that:

”X explains almost all the variation we see in Y”

X is good at predicting Y, but it is not perfect
(because there are other variables, other than X, that also affect Y and these other variables are not included in this analysis)

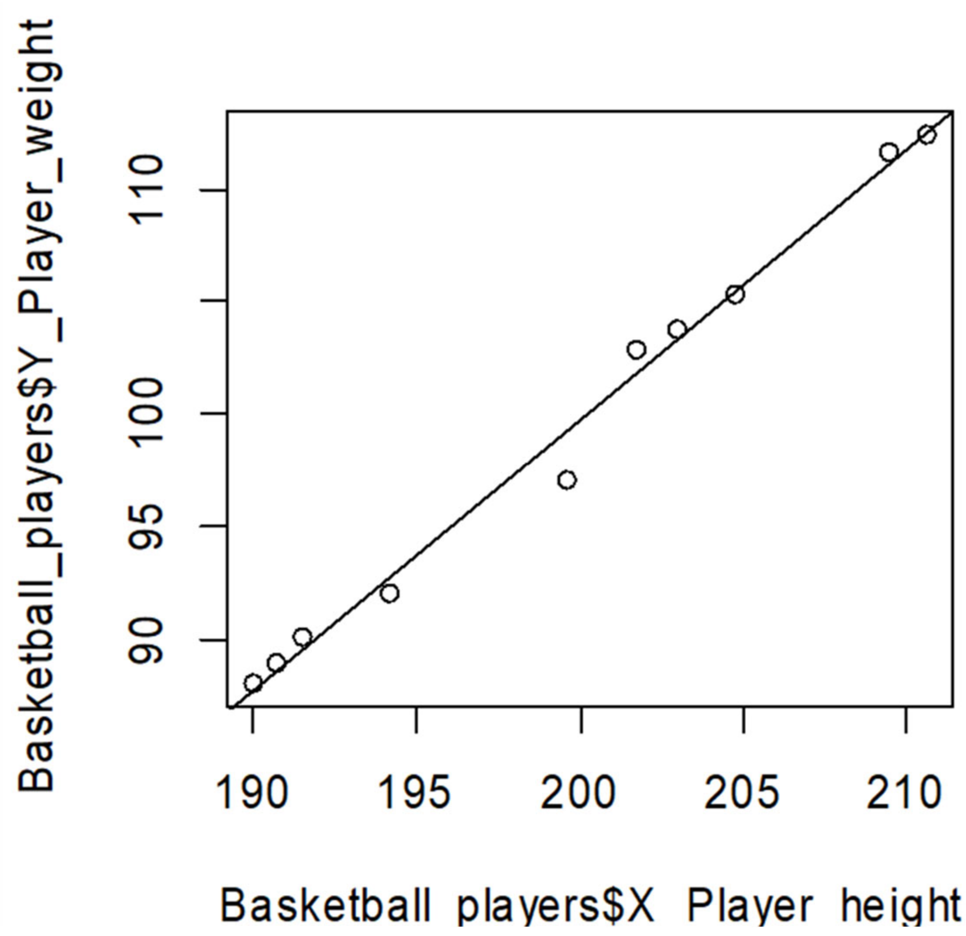


If the residuals are large, that means that:

”X explains very little of the variation we see in Y”.

X is not good at predicting Y

Example: Basketball players



The residuals are small, that means that:

- The weight of basketball players (Y) can be almost fully explained by their height (X).

- The variable height can predict very well the weight of the player, but the prediction is not perfect (there are some residuals), because:

There are other variables (variables other than height), that also affect weight, for example: diet, exercise, genetics,...

(and these other variables are not included in the simple regression analysis).

Coefficient of determination (R^2)

- The coefficient of determination (R^2) is the ratio of the explained variation to the total variation.

$$R^2 = \frac{Explained\ variation}{Total\ variation}$$

Coefficient of determination (R^2)

```
> Regression_Basketball <- lm(Y_Player_weight~X_Player_height,
data=Basketball_players)
> summary(Regression_Basketball)

Call:
lm(formula = Y_Player_weight ~ X_Player_height, data =
Basketball_players)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2566 -0.1474  0.3176  0.4659  1.0847

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -141.47114    8.68205  -16.30 2.03e-07 ***
X_Player_height   1.20597    0.04348   27.74 3.08e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9969 on 8 degrees of freedom
Multiple R-squared:  0.9897,      Adjusted R-squared:  0.9884
F-statistic: 769.4 on 1 and 8 DF,  p-value: 3.079e-09
```

Coefficient of determination (R^2)

The coefficient of determination is: $R^2 = 0.9897$

The coefficient of determination (R^2) is the ratio of the explained variation to the total variation.

Interpreting R^2:

If $R^2 = 0.9897$, that means that:

98.97% of the variation of weight of basketball players can be explained by their height.

The remaining 1.03% is unexplained; that is, it is not explained by X (height).

It could be explained by other variables such as diet, exercise, genetics,…

Coefficient of determination (R^2)

Another way to obtain the value of the coefficient of determination R^2 is to square the correlation coefficient (r).

$$(\text{Correlation coefficient})^2 = \text{Coefficient of determination}$$

```
>
cor.test(Basketball_players$X_Player_height,Basketball_players$Y_Player_weight)

	Pearson's product-moment correlation

data:  Basketball_players$X_Player_height and Basketball_players$Y_Player_weight
t = 27.738, df = 8, p-value = 3.079e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9775002 0.9988253
sample estimates:
      cor
0.9948415
```

$$0.99484^2 = 0.9897$$

The coefficient of determination is: $R^2 = 0.9897$

Lesson 10

1. Regression analysis
2. Scatter plot
3. Correlation coefficient
4. Statistical significance of the correlation coefficient
5. Correlation vs. Causation
6. Determining the regression line
7. Plotting the regression line
8. Prediction
9. Coefficient of determination
10. In-class activities

We want to study the relationship between the number of times a student is absent in the class and the final grade.

We register data on number days absent and final grade of 20 students.

The correlation coefficient is r= -0.975 (p-value<0.05)

That means:

- There is a strong positive relationship, and this relationship is statistically significant (at $\alpha=0.05$). **TRUE** **FALSE**
- There is a strong negative relationship, and this relationship is statistically significant (at $\alpha=0.05$). **TRUE** **FALSE**
- There is a strong negative relationship in the sample, but this relation is not statistically significant (at $\alpha=0.05$). **TRUE** **FALSE**

We want to study the relationship between the number of times a student is absent in the class and the final grade.

We register data on number days absent and final grade of 20 students.

The correlation coefficient is $r = -0.975$ (p-value < 0.05)

That means:

- About 97.5% of the variation of final grades can be explained by the number of times a students is absent in class. The other 2.5% is unexplained; that is, it is not explained by x, but it could be explained by other variables such as amount of time studied, intelligence, etc.

  TRUE **FALSE**

- About 95% of the variation of final grades can be explained by the number of times a students is absent in class. The other 5% is unexplained; that is, it is not explained by x, but it could be explained by other variables such as amount of time studied, intelligence, etc.

  TRUE **FALSE**

Instagram App Users, Smartphone Users and Total Population

Below, we see the number of Instagram users, smartphone users and total population of 9 countries.

```
> Instagram <-
read.table("C:/Users/vbv/Desktop/My_documents/Teaching/Teaching/Statis
tical_Data_Analysis/2022_2023/Exercises_in_R/Instagram.txt",
header=TRUE)
> Instagram
         Country Total_population Smartphone_owners Instagram_users
1            USA            329.0             260.0             120
2         Brazil            212.0              96.9              77
3      Indonesia            270.0              83.9              63
4         Russia            144.0              95.4              44
5         Turkey             83.0              44.8              38
6          Japan            127.0              72.6              29
7 United_Kingdom             67.0              55.5              24
8         Mexico            132.0              65.6              24
9        Germany             82.4              65.9              21
```

```
> cor.test(Instagram$Total_population,Instagram$Instagram_users)

	Pearson's product-moment correlation

data:  Instagram$Total_population and Instagram$Instagram_users
t = 5.9425, df = 7, p-value = 0.0005744
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6341979 0.9819273
sample estimates:
      cor
0.9135475
```

```
> Regression_Instagram <- lm(Instagram_users~Total_population,
data=Instagram)
> summary(Regression_Instagram)

Call:
lm(formula = Instagram_users ~ Total_population, data = Instagram)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9064  -8.7790   0.6185  11.2082  15.6494

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.07543   10.10060  -0.403 0.698633
Total_population  0.32956    0.05546   5.943 0.000574 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.26 on 7 degrees of freedom
Multiple R-squared:  0.8346,      Adjusted R-squared:  0.8109
F-statistic: 35.31 on 1 and 7 DF,  p-value: 0.0005744


> predict(Regression_Instagram,data.frame(Total_population=175))
       1
53.59797
```

There is a weak positive relationship between total population and number of Instagram users.

TRUE FALSE

About 83.5% of the variation in the number of Instagram users in a country can be explained by the number of habitants.

TRUE FALSE

The predicted number of Instagram users in a country with 175 million habitants is 76.3 millions.

TRUE FALSE

Videogames critics score and number of copies sold.

We want to study the relationship between critics score of videogames and number of copies sold.

We collect data from 30 videogames on:

- Rank - Ranking of overall sales
- Name - The games name
- Genre - Genre of the game
- Publisher - Publisher of the game
- Critic score - Score of the game on metacritic.com
- Copies sold – in millions

```
> Video_games <-
read.table("C:/Users/vbv/Desktop/My_documents/Teaching/Teaching/Statistical_Data_Analysis/2022_2023/Exerc
ises_in_R/Video_games.txt", header=TRUE)
> Video_games
   Rank                          Name        Genre              Publisher Critic_score Copies_sold_mill
1     1                    Wii_Sports       Sports               Nintendo          7.7            82.86
2     2              Super_Mario_Bros.    Platform               Nintendo          9.4            40.24
3     3                 Mario_Kart_Wii      Racing               Nintendo          8.2            37.14
4     4 PlayerUnknown's_Battlegrounds      Shooter        PUBG_Corporation          8.5            36.60
5     5              Wii_Sports_Resort      Sports               Nintendo          8.0            33.09
6     6             Pokemon_RGB_Version Role-Playing             Nintendo          9.4            31.38
7     7            New_Super_Mario_Bros.   Platform              Nintendo          9.1            30.80
8     8                          Tetris     Puzzle               Nintendo          7.3            30.26
9     9          New_Super_Mario_Bros_Wii   Platform             Nintendo          8.6            30.22
10   10                       Minecraft       Misc                 Mojang          9.3            30.01
11   11                       Duck_Hunt    Shooter               Nintendo          7.1            28.31
12   12                        Wii_Play       Misc               Nintendo          5.9            28.02
13   13               Kinect_Adventures      Party Microsoft_Game_Studios          6.7            24.00
14   14                      Nintendogs  Simulation              Nintendo          8.4            23.96
15   15                   Mario_Kart_DS     Racing               Nintendo          9.1            23.60
16   16              Pokemon_Gold_Silver Role-Playing            Nintendo          9.2            23.10
17   17                         Wii_Fit     Sports               Nintendo          7.9            22.67
18   18                    Wii_Fit_Plus     Sports               Nintendo          8.0            21.13
19   19               Super_Mario_World   Platform               Nintendo          8.5            20.61
20   20              Grand_Theft_Auto_V     Action         Rockstar_Games          9.4            20.00
21   21                       Brain_Age       Misc               Nintendo          8.1            19.01
22   22                      Garrys_Mod       Misc                Unknown          6.5            18.58
23   23                Super_Mario_Land   Platform               Nintendo          7.0            18.14
24   24                    Mario_Kart_7     Racing               Nintendo          8.2            18.11
25   25            Pokemon_Diamond_Pearl Role-Playing            Nintendo          8.6            17.67
26   26      Grand_Theft_Auto_San_Andreas     Action         Rockstar_Games          9.5            17.30
27   27               Super_Mario_Bros_3   Platform              Nintendo          7.5            17.28
28   28                    Pokemon_X/Y Role-Playing              Nintendo          8.9            16.37
29   29            Pokemon_Ruby_Sapphire Role-Playing            Nintendo          8.8            16.22
30   30                Pokemon_Sun_Moon Role-Playing             Nintendo          9.0            16.14
```

```
> cor.test(Video_games$Critic_score,Video_games$Copies_sold_mill)

	Pearson's product-moment correlation

data:  Video_games$Critic_score and Video_games$Copies_sold_mill
t = -0.30182, df = 28, p-value = 0.765
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4088272  0.3096771
sample estimates:
        cor 
-0.05694534 
```

```
> Regression_videogames <- lm(Copies_sold_mill~Critic_score,
data=Video_games)
> summary(Regression_videogames)

Call:
lm(formula = Copies_sold_mill ~ Critic_score, data = Video_games)

Residuals:
   Min     1Q Median     3Q    Max 
-9.799 -8.319 -3.112  4.289 56.009 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.6746    20.8329   1.568    0.128
Critic_score  -0.7563     2.5059  -0.302    0.765

Residual standard error: 12.92 on 28 degrees of freedom
Multiple R-squared:  0.003243,	Adjusted R-squared:  -0.03236 
F-statistic: 0.09109 on 1 and 28 DF,  p-value: 0.765
```

There is a significant positive relationship between critics score of videogames and number of copies sold.

TRUE

FALSE

We can use critics score to predict the number of copies sold.

TRUE

FALSE