PCTO in Coding & Data Science

CD: 50/50 - Coding Diversity

Liceo Scientifico S. Cannizzaro
18/03/2022

CD:50/50
CODING DIVERSITY

Nelle ultime puntate…

1. Descrivere i dati: media, mediana, varianza
2. Dai valori assoluti alle frequenze
3. Il concetto di distribuzione e la distribuzione normale
4. Grafici bivariato di tipo “scatter”
5. Grafici bivariato di tipo “a barre”
6. Grafici a barre raggruppate
7. Grafici nel tempo: le serie storiche
8. Grafici nello spazio: le mappe
9. Alcuni errori comuni nella data analysis e come evitarli

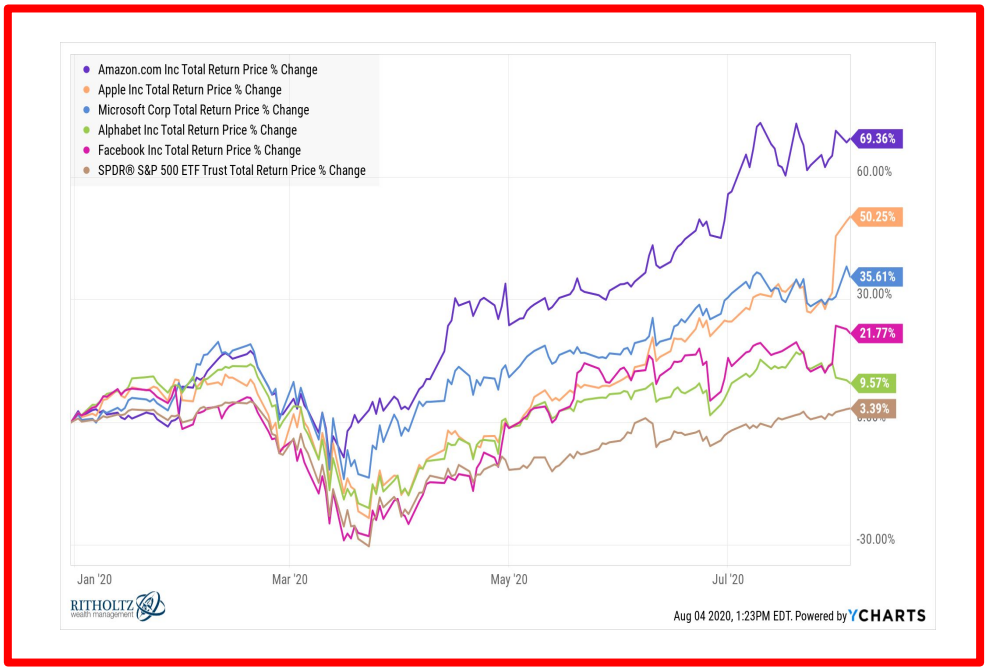
Modulo 1: Introduzione al coding

Modulo 2: Saper leggere e rappresentare i dati

Modulo 3: Basi di inferenza e analisi predittiva

Modulo 4: Basi di machine learning

Un’importante funzione della **data analysis**: la **previsione**



CD:50/50
CODING DIVERSITY

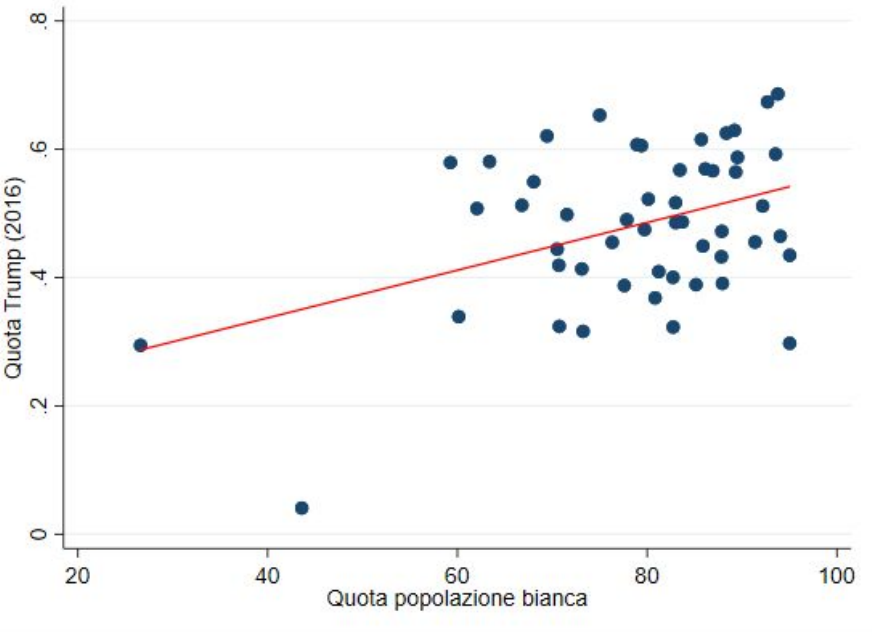
Alcuni esempi di previsione

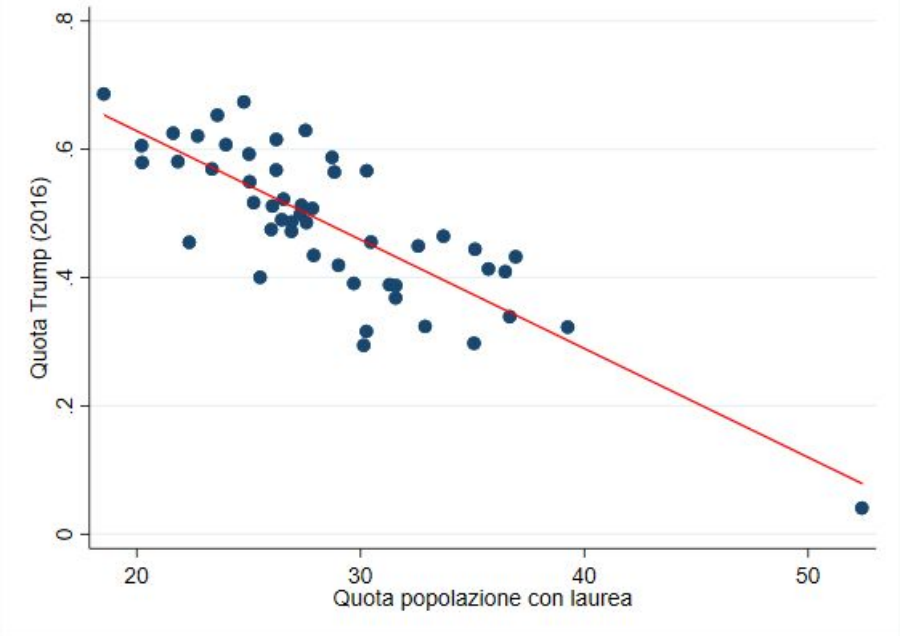
- andamento mercato azionario
- evoluzione surriscaldamento globale
- tendenza nei pazienti a sviluppare effetti collaterali
- tendenza alla recidiva (quando un ex-detenuto ricommette un crimine)
- tendenza ad evadere le tasse
- click-through rate

ALTRI ESEMPI?

In tempi recenti i dati sono diventati uno strumento fondamentale in occasione delle elezioni

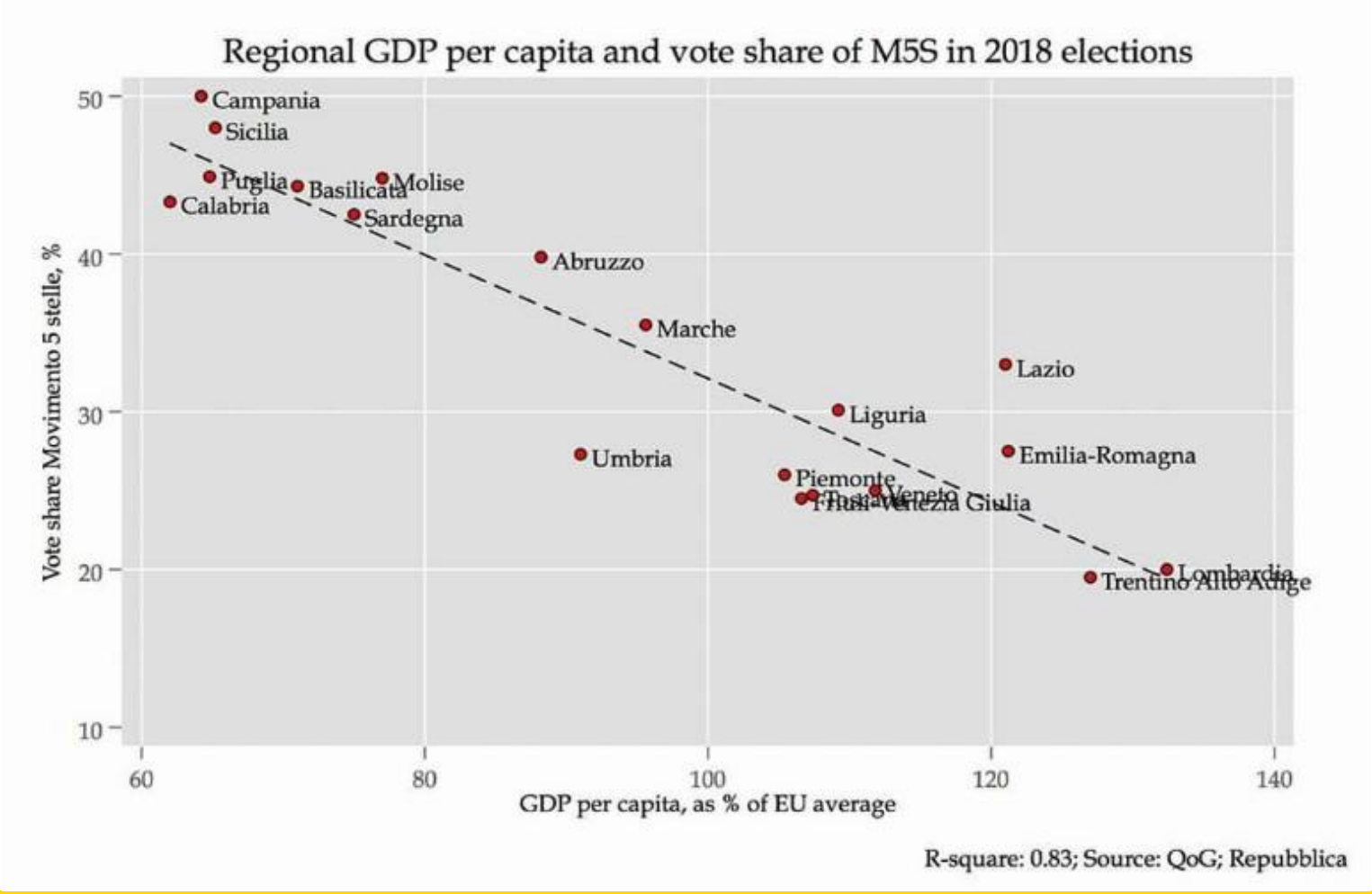
Uno sguardo alle elezioni americane del 2016





- Stati con un'elevata % di popolazione "bianca" hanno votato **più** per Trump
- Stati con un'elevata % di popolazione laureata hanno votato **meno** per Trump

(Anche in Italia, territori con caratteristiche simili hanno tendenze di voto simili)



CD:50/50
CODING DIVERSITY

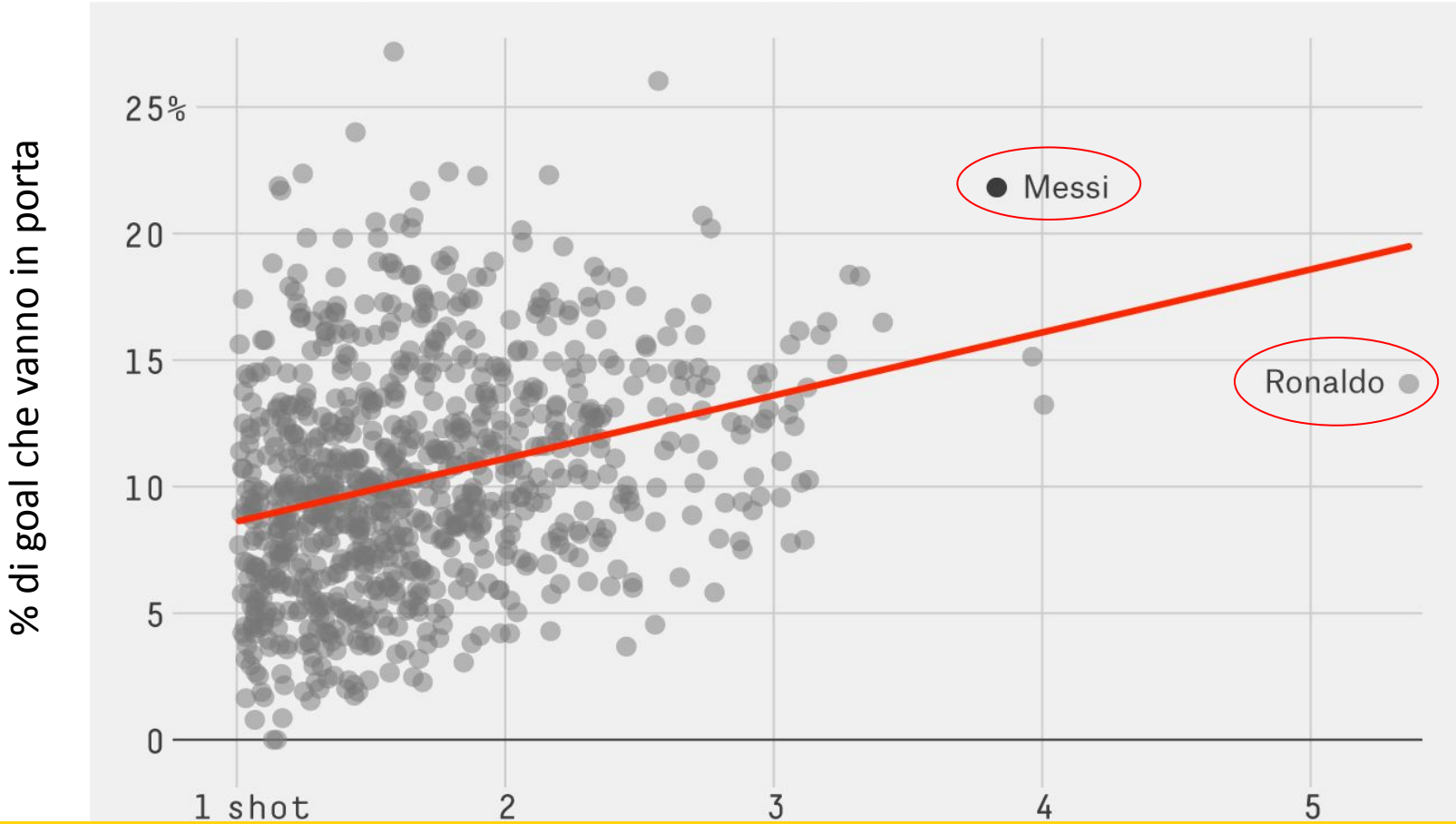
La **covarianza** misura che tipo di **relazione** esiste tra due variabili, X e Y:

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

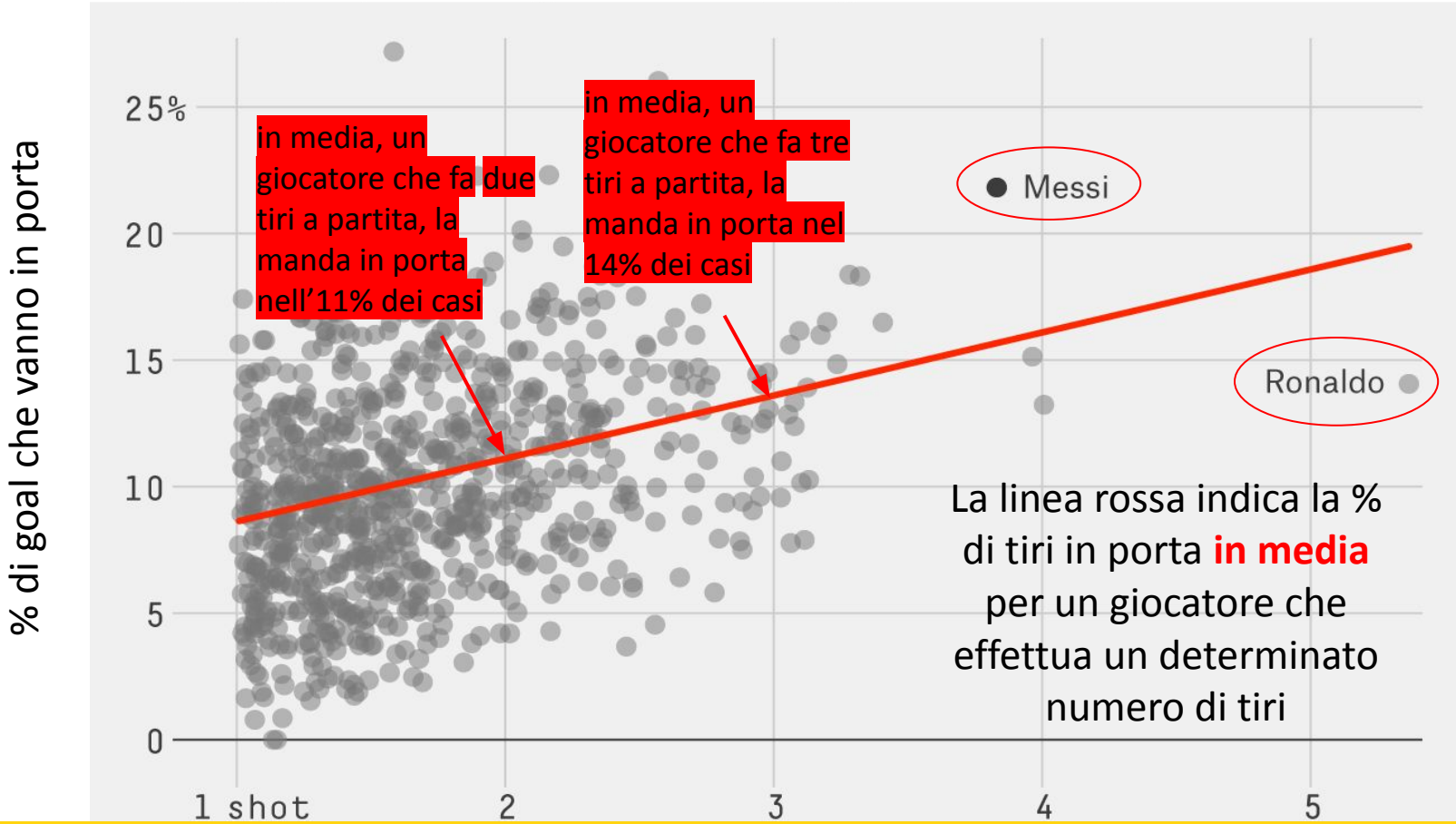
Diciamo che le **due variabili Y e X** sono:

- Correlate **positivamente** (covarianza>0) se all'aumentare di Y aumenta anche X
- Correlate **negativamente** (covarianza>0) se all'aumentare di Y diminuisce anche X
- altrimenti diciamo che **non sono correlate** (covarianza=0)

Le correlazioni esprimono sempre delle relazioni tra due serie di dati che valgono in media



Le correlazioni esprimono sempre delle relazioni tra due serie di dati che valgono in media



Scelta del modello e margine d'errore

Cos'è un modello?

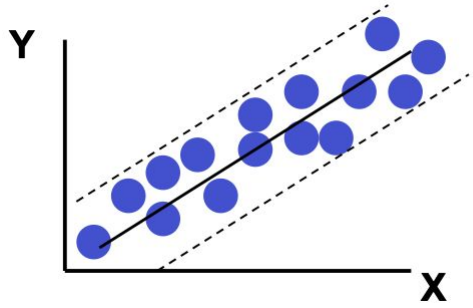
"...una **semplificazione della realtà** **necessaria** per aiutarci a comprenderla".

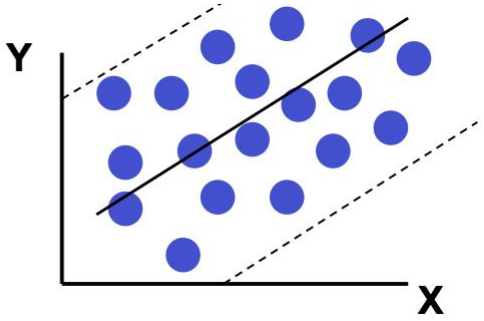
Un **modello** è essenzialmente una **funzione** che, data una serie di dati (**"input"**), li analizza e ci restituisce un **"output"** o **stima** a cui siamo interessati che presenta però **sempre un margine d'errore**

L'obiettivo di chi fa previsioni è quello di trovare un modello che sia preciso "abbastanza"

Scelta del modello e margine d'errore

I dati che osserviamo non sono generati da un computer secondo una rigida regola matematica → per esempio, quando **stimiamo** una **correlazione** che ci aiuta a **prevedere** una variabile con il supporto di un'altra variabile che osserviamo, la **previsione non è mai perfetta.**

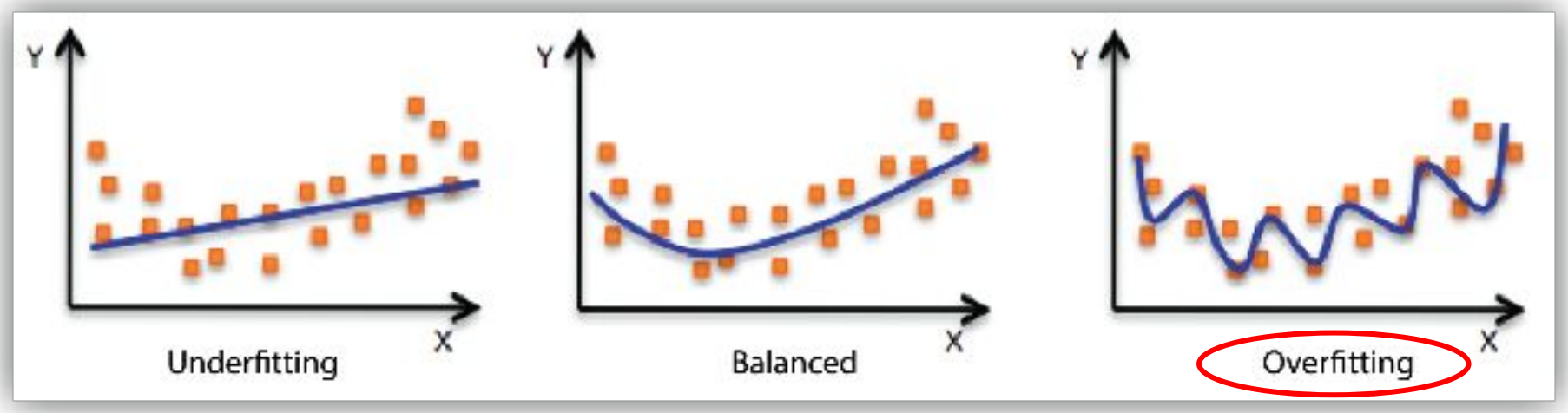




Quale di queste due correlazioni è più precisa?

Come ridurre il margine d'errore legato alla scelta del modello?

Secondo voi qual è il modello migliore dei tre in basso?



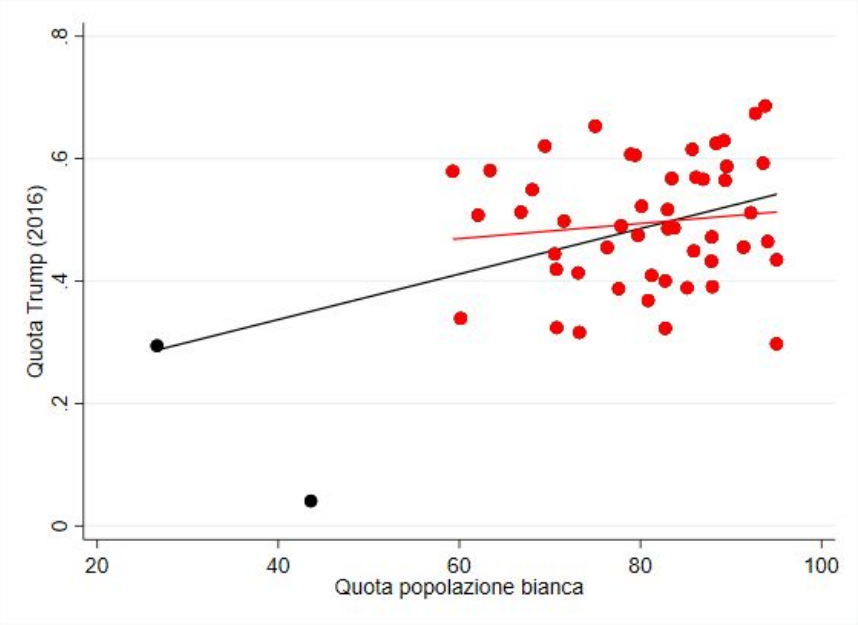
Overfitting: quando si costruisce un modello che replica "troppo bene" un insieme di dati.

Perché secondo voi può essere un problema?

Il problema dell'overfitting quando si analizza un campione

In generale, è raro **osservare** l'intera "**popolazione**", ovvero l'intero insieme di osservazioni (per es. l'intera popolazione italiana).

Molto più spesso si osserva un **campione**, ossia un **sottoinsieme**.



Le stime basate su campioni diversi possono essere diverse tra loro, spesso per la presenza di **osservazioni anomale**.

Due buone pratiche del data analyst:

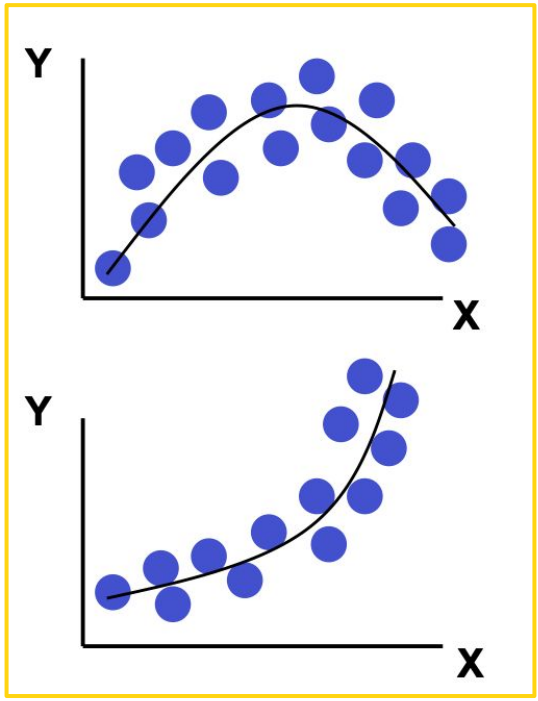
1. rimuovere osservazioni anomale
2. evitare modelli che fanno overfitting

Nell’esempio delle presidenziali americane di prima…

Supponiamo di voler prevedere l’esito delle elezioni americane sulla base di un semplice modello **lineare**

Quale variabile sceglieremmo?

Correlazione	Variabile
0.15	Tasso di over-65
0.26	Tasso povertà
-0.32	Tasso di popolazione femminile
0.41	Tasso di popolazione "white"
0.46	Tasso di U18
-0.59	Densità di popolazione
-0.73	Reddito pro capite
-0.78	Tasso di imprese femminili
-0.82	Tasso di laureati



Più alta è la correlazione (in valori assoluti), maggiore è la precisione del modello!

Un modello molto comune: il modello di regressione lineare

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

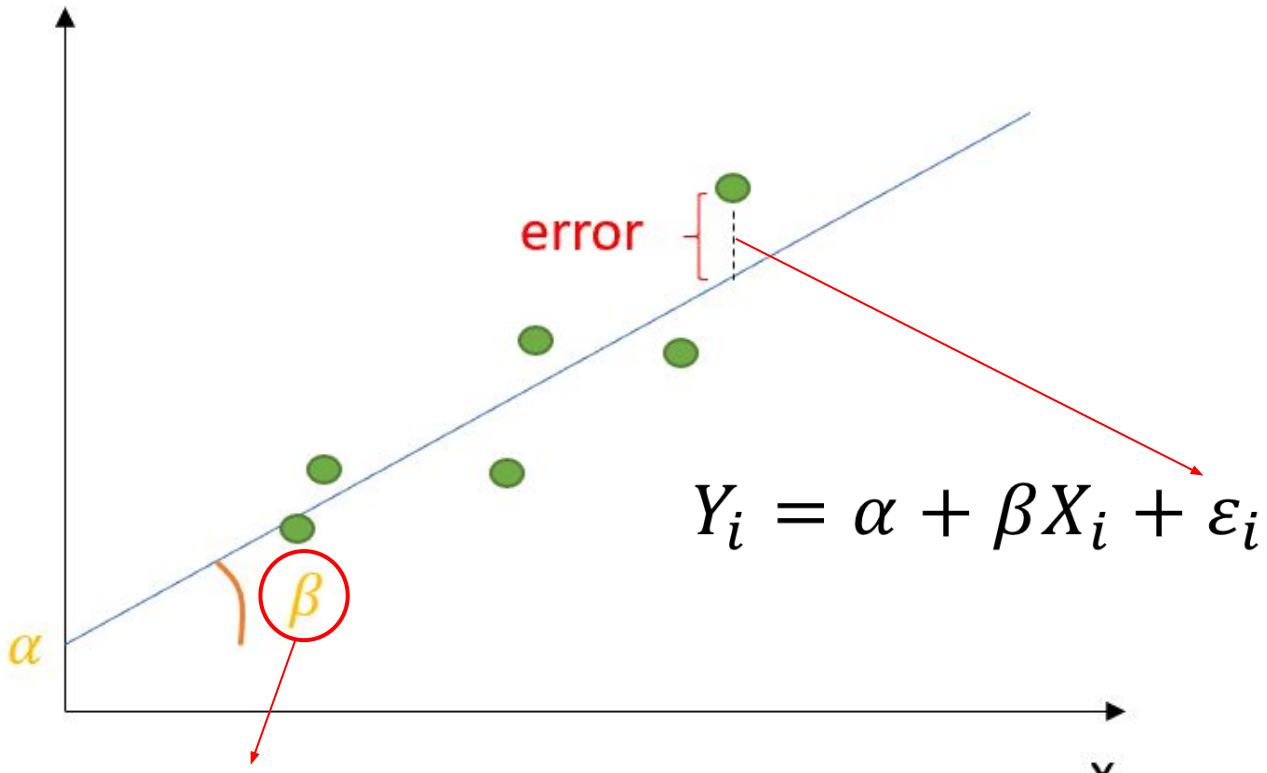
Dove:

- Y_i indica il valore della variabile «y» o variabile dipendente per l'osservazione i
- X_i indica il valore della variabile «x» o variabile indipendente per l'osservazione i
- α è l'intercetta, ovvero il valore medio di «y» se il valore medio di «x» fosse zero
- β indica il coefficiente di correlazione tra le due variabili:
  - → β>0 : le due variabili hanno una correlazione positiva
  - → β<0 : le due variabili hanno una correlazione negativa
- ε_i è un termine d'errore, dovuto al fatto che nessuna correlazione statistica vale per tutte le osservazioni, quanto piuttosto vale per una media

Esempio:

$$salario_i = \alpha + \beta ore\ lavorate_i + \varepsilon_i$$

Quando c'è solo una "x" il coefficiente "beta" è sostanzialmente analogo alla correlazione tra due variabili vista finora



di quanto aumenta la Y se X aumenta di 1?

Anticipando i risultati del comando in Colab…

Il nostro primo modello è molto semplice:

$$logY = \alpha + \beta_1 X_1 + \varepsilon$$

Dove:

$logY$: il logaritmo del reddito di una famiglia

X_1 : se la mamma lavora (1 se lavora, 0 se non lavora)

I numeri importanti sono pochi:

1. il **coefficiente di correlazione** (beta)
2. il livello di **precisione** con cui è stimato
3. la **bontà del modello a "predire" il reddito**

```
ols = sm.OLS(logy, X)
ols_result = ols.fit()
ols_result.summary()
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.141
Model:	OLS	**Adj. R-squared:**	0.141
Method:	Least Squares	**F-statistic:**	341.3
Date:	Fri, 11 Mar 2022	**Prob (F-statistic):**	1.08e-70
Time:	12:35:37	**Log-Likelihood:**	-1666.0
No. Observations:	2083	**AIC:**	3336.
Df Residuals:	2081	**BIC:**	3347.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.4397	0.024	18.475	0.000	0.393	0.486
const	10.2062	0.018	570.935	0.000	10.171	10.241

Omnibus:	278.915	**Durbin-Watson:**	1.838
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1562.844
Skew:	-0.494	**Prob(JB):**	0.00
Kurtosis:	7.127	**Cond. No.**	2.80

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

(opzionale) La statistica non è una scienza certa nel dover prendere delle decisioni sulla base dell'analisi dei dati a disposizione si corre sempre il rischio di commettere errori.

Come valutiamo se un coefficiente di correlazione è stimato abbastanza precisamente da poter concludere che esiste una correlazione tra due variabili?

(opzionale) L’approccio statistico: il test delle ipotesi

Ipotesi: **Fare esercizio fisico fa dimagrire**

Estraiamo 10 persone a caso e riscontriamo che chi ha fatto più attività fisica negli ultimi due mesi ha perso più peso.

Possiamo concludere che fare attività fisica fa dimagrire?

Per dare “validità scientifica” alla nostra **ipotesi**, cominciamo con l’**assumere che sia falsa**, e assumere che invece vera quella contraria.

Ipotesi contraria: Fare attività fisica non fa dimagrire.

In questo caso, ci aspettiamo di trovare una **correlazione** tra i due fenomeni con una **bassa probabilità**.

(opzionale) L'approccio statistico: il test delle ipotesi

Se però estraiamo il nostro campione di persone e scopriamo che la correlazione tra i due fenomeni è elevata, ci sono 2 possibilità:

- la correlazione è elevata ma non troppo:
  vuol dire che c'è un'elevata **probabilità che sia frutto del caso**
  → **non rifiutiamo l'ipotesi contraria**

- la correlazione è "troppo" elevata:
  vuol dire che c'è una bassa **probabilità che sia frutto del caso → l'ipotesi contraria è falsa, conviene rigettarla** in favore di quella che fare attività fisica faccia perdere peso =)

Questa probabilità si chiama p-value:
se è basso (meno di 0.1) esiste una correlazione tra Y e X)

Un test delle ipotesi molto comune di questi tempi

Il livello di **precisione** nella stima del coefficiente **beta**

$$logY = \alpha + \beta_1 X_1 + \varepsilon$$

Dove:

$logY$: il logaritmo del reddito di una famiglia

X_1 : se la mamma lavora (1 se lavora, 0 se non lavora)

Il **p-value è inversamente legato alla precisione delle stime**: tanto più è basso, tanto più il coefficiente è stimato con precisione.

Se è minore di **0.1** concludiamo che esiste una correlazione tra le due variabili **"statisticamente significativa"**.

```
ols = sm.OLS(logy, X)
ols_result = ols.fit()
ols_result.summary()
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.141
Model:	OLS	**Adj. R-squared:**	0.141
Method:	Least Squares	**F-statistic:**	341.3
Date:	Fri, 11 Mar 2022	**Prob (F-statistic):**	1.08e-70
Time:	12:35:37	**Log-Likelihood:**	-1666.0
No. Observations:	2083	**AIC:**	3336.
Df Residuals:	2081	**BIC:**	3347.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.4397	0.024	18.475	0.000	0.393	0.486
const	10.2062	0.018	570.935	0.000	10.171	10.241

p-value

Omnibus:	278.915	**Durbin-Watson:**	1.838
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1562.844
Skew:	-0.494	**Prob(JB):**	0.00
Kurtosis:	7.127	**Cond. No.**	2.80

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Veniamo all'ultima numero: come scegliere tra due modelli?

$$logY = \alpha + \beta_1 X_1 + \varepsilon$$

Dove:

$logY$: il logaritmo del reddito di una famiglia

X_1 : se la mamma lavora (1 se lavora, 0 se non lavora)

L'ultimo numero che prendiamo in considerazione si chiama **R-squared.**

Questo rappresenta una **misura globale della capacità predittiva** (=precisione) del modello.

```
ols = sm.OLS(logy, X)
ols_result = ols.fit()
ols_result.summary()
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.141
Model:	OLS	**Adj. R-squared:**	0.141
Method:	Least Squares	**F-statistic:**	341.3
Date:	Fri, 11 Mar 2022	**Prob (F-statistic):**	1.08e-70
Time:	12:35:37	**Log-Likelihood:**	-1666.0
No. Observations:	2083	**AIC:**	3336.
Df Residuals:	2081	**BIC:**	3347.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.4397	0.024	18.475	0.000	0.393	0.486
const	10.2062	0.018	570.935	0.000	10.171	10.241

Omnibus:	278.915	**Durbin-Watson:**	1.838
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1562.844
Skew:	-0.494	**Prob(JB):**	0.00
Kurtosis:	7.127	**Cond. No.**	2.80

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Cosa succede se aggiungiamo un'altra variabile X?

Abbiamo inserito un'altra variabile tra le nostre "X", ovvero se la madre ha una laurea (1 se ce l'ha, 0 se non ce l'ha).

Il modello di regressione lineare è infatti utile soprattutto perché ci consente di analizzare come **più di una variabile influenzano simultaneamente la nostra Y.**

Cosa notate dalla tabella a fianco?

```
famiglie_redux = famiglie[['madre_lavora', 'madre_laurea']]
X = famiglie_redux.to_numpy()
X = np.append(X,np.ones((len(famiglie_redux),1)),axis=1)
ols = sm.OLS(logy, X)
ols_result = ols.fit()
ols_result.summary()
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.182
Model:	OLS	**Adj. R-squared:**	0.181
Method:	Least Squares	**F-statistic:**	231.1
Date:	Fri, 11 Mar 2022	**Prob (F-statistic):**	2.27e-91
Time:	12:36:19	**Log-Likelihood:**	-1615.2
No. Observations:	2083	**AIC:**	3236.
Df Residuals:	2080	**BIC:**	3253.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.3728	0.024	15.442	0.000	0.325	0.420
x2	0.3285	0.032	10.199	0.000	0.265	0.392
const	10.1895	0.018	581.408	0.000	10.155	10.224

Omnibus:	296.224	**Durbin-Watson:**	1.829
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1716.786
Skew:	-0.527	**Prob(JB):**	0.00
Kurtosis:	7.321	**Cond. No.**	3.48

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Un recap…

1. Uno degli usi più comuni della data science: la previsione
2. Il concetto di correlazione
3. Scelta del modello e il margine d'errore/precisione nelle stime
4. Il concetto di popolazione e quello di campione statistico
5. Due problemi quando si ha a che fare con il campionamento statistico: overfitting e valori anomali
6. Il modello di regressione lineare univariato (una sola X)
7. Il test delle ipotesi per capire
8. L'R-squared, misura complessiva della precisione di un modello
9. Il modello di regressione lineare bivariato (più di una X)