La Regressione, ovvero come passare da tanti brutti puntini ad una simpatica linea

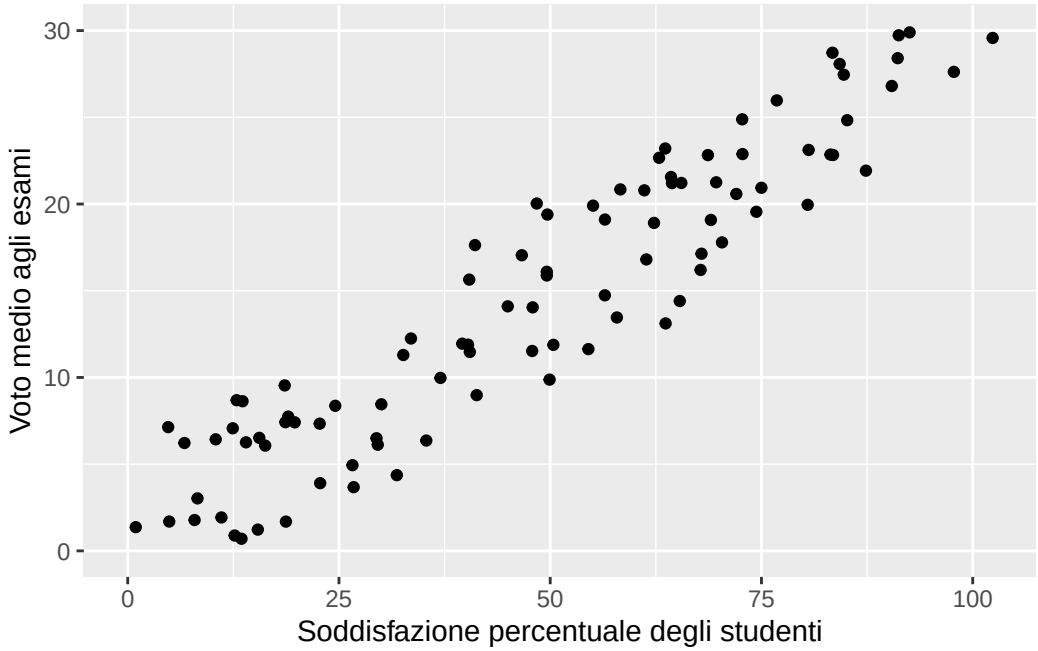
- Gaetano Scaduto, Analysis of Social and Economic Processes, University of Milan-Bicocca
- Per chiarimenti, reclami e schemi piramidali mi trovate all'indirizzo g.scaduto2@campus.unimib.it

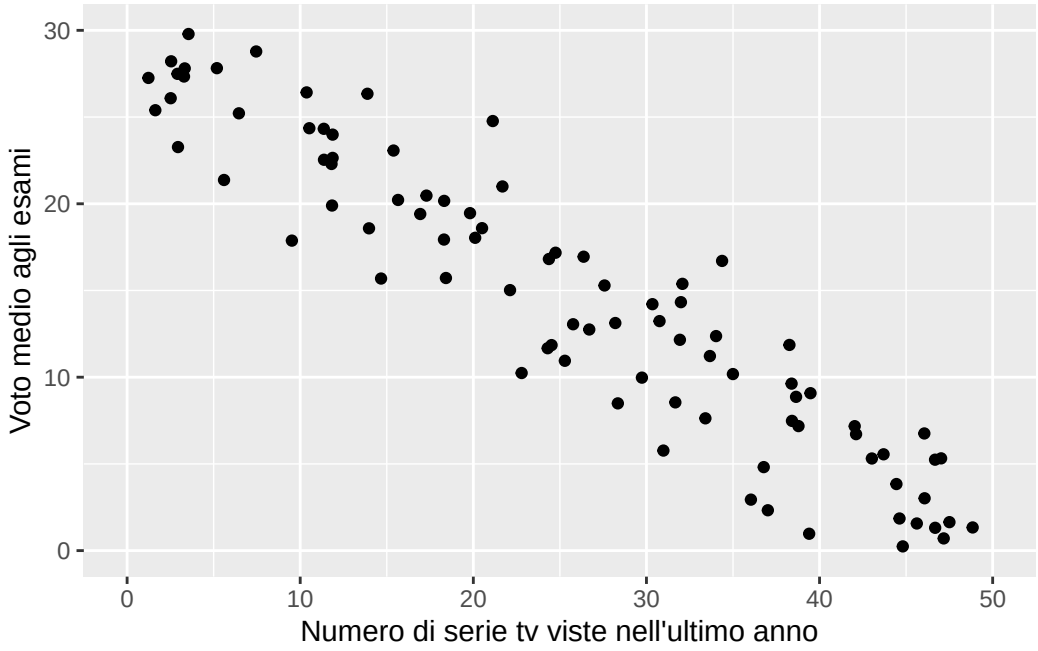
- **Nota**: tutti i dati utilizzati in questa lezione sono simulati.
- Ogni riferimento a cose o partiti realmente esistenti è volutamente fatta a scopi parodistici.

Oggi parliamo di **regressione**, ovvero una tecnica statistica che ci permette di trovare una possibile **relazione funzionale fra una variabile dipendente e una o più variabili indipendenti.**

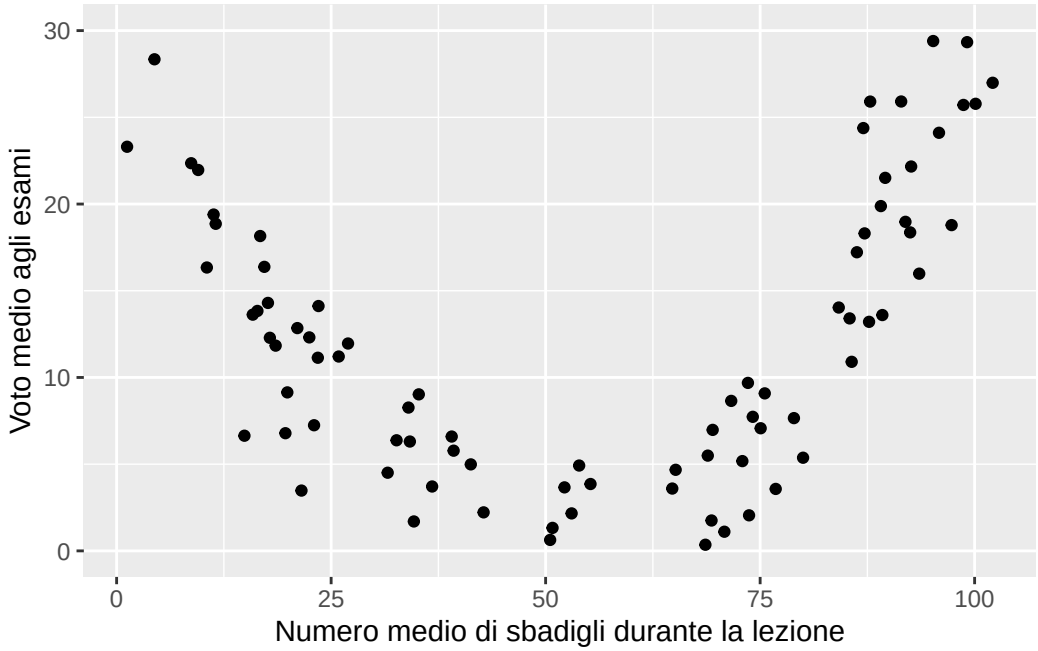
Prima di addentrarci in cosa sia la regressione, cosa significhi e perché la usiamo, cerchiamo di avere uno sguardo più dall'alto.

La realtà è piena di associazioni fra variabili per i più svariati motivi. Talvolta queste associazioni posssono essere positive, talvolta negative e talvolta possono non esserci proprio. Talvolta possono seguire una legge semplice e lineare, talvolta complessa, logaritmica, quadratica, sinusoidale (non nelle scienze sociali) e così via... A volte queste associazioni sono esattamente quello che ci aspettiamo.

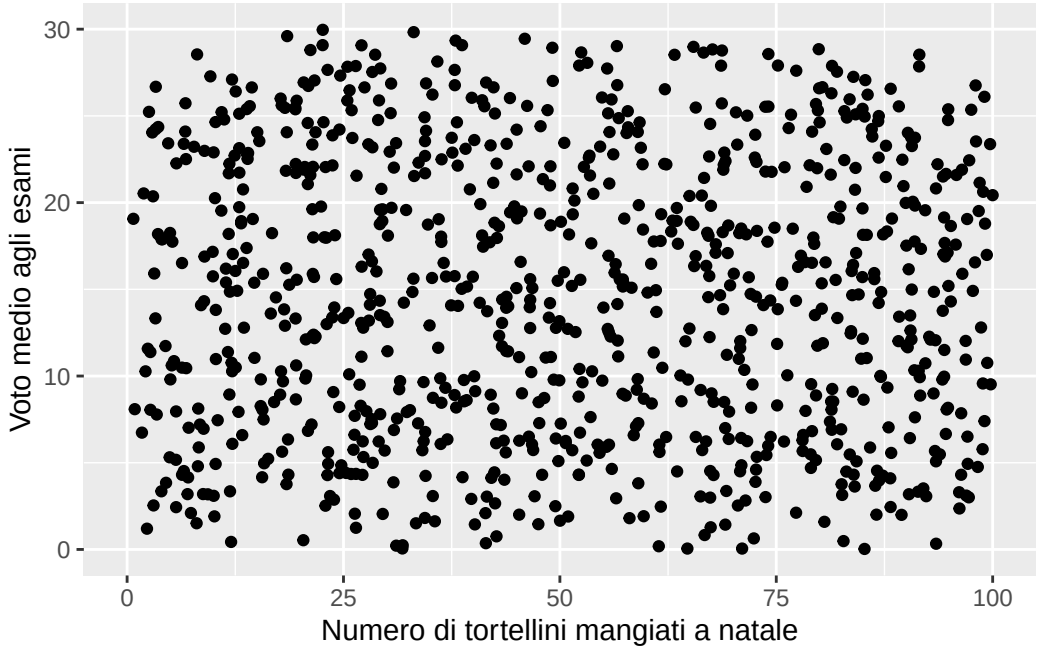




A volte ci sorprendono e dobbiamo sforzarci di spiegarle.

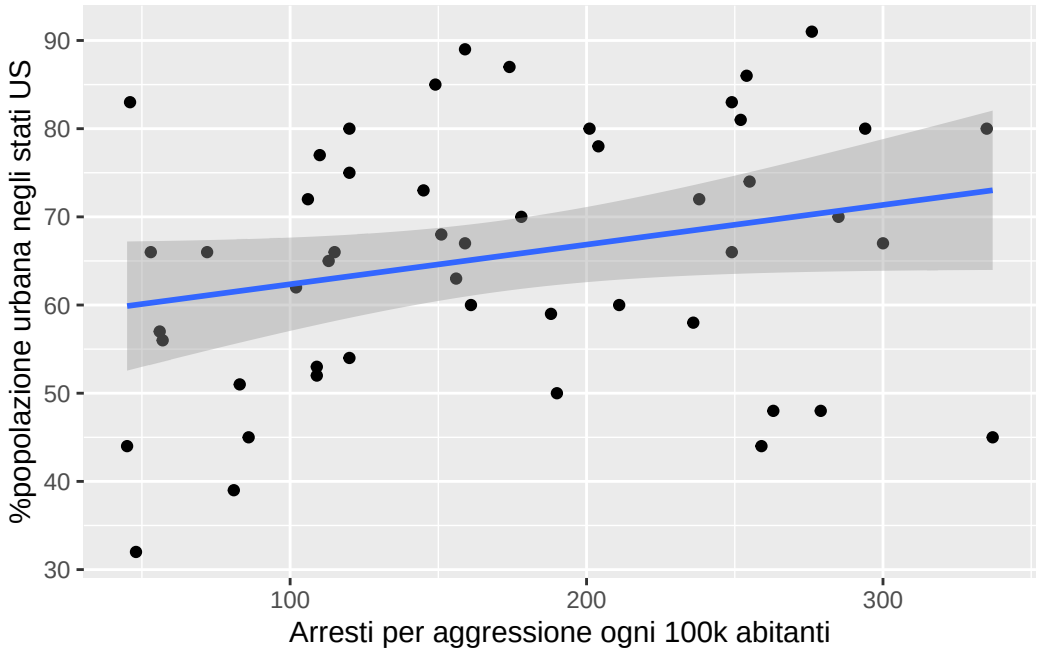


A volte dobbiamo accettare l'assenza di relazioni e darci pace.



Tutti i grafici precedenti che vi ho mostrato sono fatti su dati **creati ad hoc**. Spiegare dati simulati non ha alcun senso, ma questi ci permettono di aprire un discorso importante riguardo

come possiamo passare dall'osservare dei semplici puntini allineati a vederci delle leggi, senza lavorare troppo di fantasia, ma con una regola chiara e rigorosa. **Noi non cerchiamo carri fra le stelle, ma leggi in mezzo ai dati**.



Okay, questi ultimi sono dati reali. Gli unici dati reali che vedrete in questa lezione.

La regressione lineare

Nel nostro caso, noi parleremo solamente di **regressione lineare**.

Ma cosa significa **lineare**?

- In matematichese, lineare significa "che sta in una linea dritta", che di per sé non sembra una cosa speciale, ma invece lo è. Le rette hanno un'importante proprietà: variano sempre allo stesso modo in ogni punto! (La "crescita" o "decrescita" è costante = La derivata è nulla).
- Bisogna tenere a mente che la linearità riguarda i coefficienti e non le variabili. Quelli che tra poco presenteremo.
- Bisogna anche ricordare che nella regressione lineare **la y è sempre una variabile numerica,** mentre le variabili indipendenti possono essere di qualunque tipo: numeriche, categoriali, dicotomiche, ordinali. Ma ci ritorneremo.

Le rette, in generlale, seguono sempre la stessa semplice equazione:

$$y = \beta_0 + \beta_1 * x$$

Ogni retta possibile sul piano è identificata dai due valori.

- β_0 si chiama "termine noto", "**intercetta**" o anche "ordinata all'origine", perché corrisponde al valore di y quando x=0
- β_1 si chiama **coefficiente angolare** ed è quello che ci interessa davvero , perché ci dice *quanto varia y al variare di x*

Fissati questi due parametri, fra tutte le infinite e immaginabili rette che passano sullo stesso piano in cui ci sono i nostri puntini, ne identifichiamo una e una sola. La nostra retta di regressione.

Per farci un'idea di come cambia una retta al variare di β_0 e β_1 possiamo utilizzare questo simpatico tool. Potete farlo anche voi, basta inserire y=a+b*x nella calcolatrice e potere giocare su come cambia la retta al variare di intercetta e coefficiente angolare.

Adesso basta matematica! Entriamo nella statistica!

$$y = \beta_0 + \beta_1 * x + \epsilon$$

Abbiamo un intruso! Il temibile **epsilon!** Ne parleremo nel dettaglio dopo, ma possiamo anticipare che epsilon rappresenta **tutto ciò che non sappiamo**, tutto ciò che non ci permette di estrarre delle **leggi deterministiche.** Questo significa che quando osserveremo qualunque coppia di dati "umani" non li avremo mai davvero distribuiti lungo una linea. Ma li avremo sparsi un po' in giro. E **toccherà a noi cercare di trovare la legge nascosta sotto il rumore.**

Come se non bastasse, abbiamo un altro problema: noi non potremo mai conoscere i parametri **reali** della nostra legge, ma il massimo che potremo fare è avere **delle stime** di quei parametri, attraverso degli **stimatori**. Nel nostro caso, gli stimatori saranno sempre gli stimatori ottenuti dal **metodo dei minimi quadrati (OLS - Ordinary Least Squares).** Vedremo dopo cosa sono per bene. Per ora è importante avere chiaro la differenza fra questa:

$$y = \beta_0 + \beta_1 * x + \epsilon$$

Che rappresenta il nostro "ideale", la legge che noi immaginiamo esista nella realtà ma che non possiamo davvero osservare. Una sorta di iperuranio che possiamo solo stimare con una certa **confidenza**, ma mai ottenenere davvero.

$$y = \hat{\beta}_0 + \hat{\beta}_1 * x + \epsilon$$

E questa, che invece rappresenta ciò che noi ci andremo sempre a calcolare. Quel triangolino sui coefficienti ci serve a ricordarci che quelli sono **gli stimatori che ci siamo calcolati a partire dai nostri dati**, non i parametri "reali". Noi calcoliamo sempre e solo gli stimatori, mai i parametri reali!

Andiamo a vedere un primo esempio con dei dati simulati per capire meglio di cosa stiamo parlando.

Un esempietto

Supponiamo di voler investigare la relazione fra **X=il numero di giorni dedicati alla preparazione dell'esame di politiche pubbliche** e **Y= il voto preso all'esame**.

Supponiamo che in questa classe siate **200 persone**, che **in media** la gente studi **10 giorni** per preparare questo esame, e che il numero di giorni passati a studiare sia distribuito **normalmente** con una **varianza** pari a **9** (deviazione standard = **3**). Notate che vale la seguente proprietà.

Per la proprietà ricordata sopra, il 99,7% di voi dedicherà a questa materia fra i 10+3*3=**19** e i 13-3*3=**1** giorni allo studio di questa materia. Che direi che ci sta.

Mi aspetto che il punteggio atteso all'esame sia una funzione dei giorni passati a studiare nella forma:

$$Punteggio = punteggiolavorodigruppo + 2 * giornistudiati$$

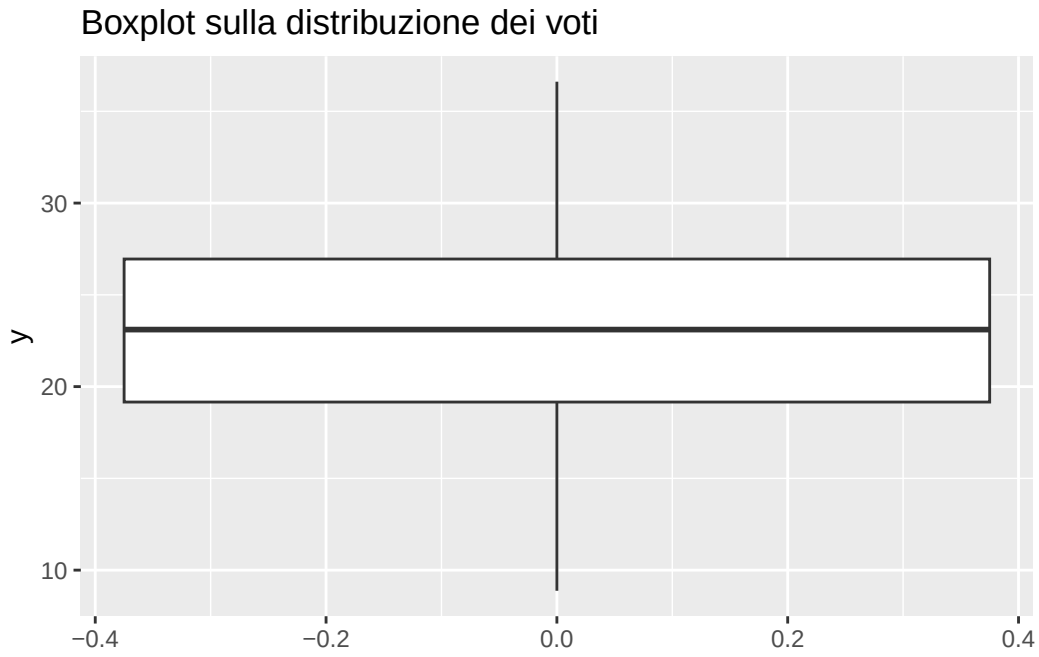
Notate che in questa legge non c'è alcun termine di errore. Quindi mi sto aspettando che il voto all'esame sia solo e solamente una conseguenza diretta del vostro punteggio al lavoro di gruppo e dei giorni che avete passato a studiare. Nient'altro. Non conta l'abilità, non conta la fortuna, non conta l'intelligenza. Solo giorni di studio, come se fosse una legge della fisica.

Cosa ci dice questa formula? Proviamo a guardarla bene.

- Questa formula significa che **se studio 0 giorni, il mio punteggio sarà finale sarà pari solo al mio punteggio al lavoro di gruppo**.
  - Questa cosa in effetti può aver senso: se non studio nemmeno un giorno non riuscirà a rispondere a nessuna domanda all'esame, e dunque lì farò zero punti e l'unica mia "fonte di punti" sarà il punteggio del lavoro di gruppo.
- **Se studio 15 giorni, posso prendere 33=30L.**
  - Ogni giorno in più che studio il mio punteggio finale migliora di due punti.
- Notate che stiamo supponendo che siate state tutti bravissimi e bravissime nei vostri lavori di gruppo e che dunque **tutti i gruppi abbiano preso 3 punti**.

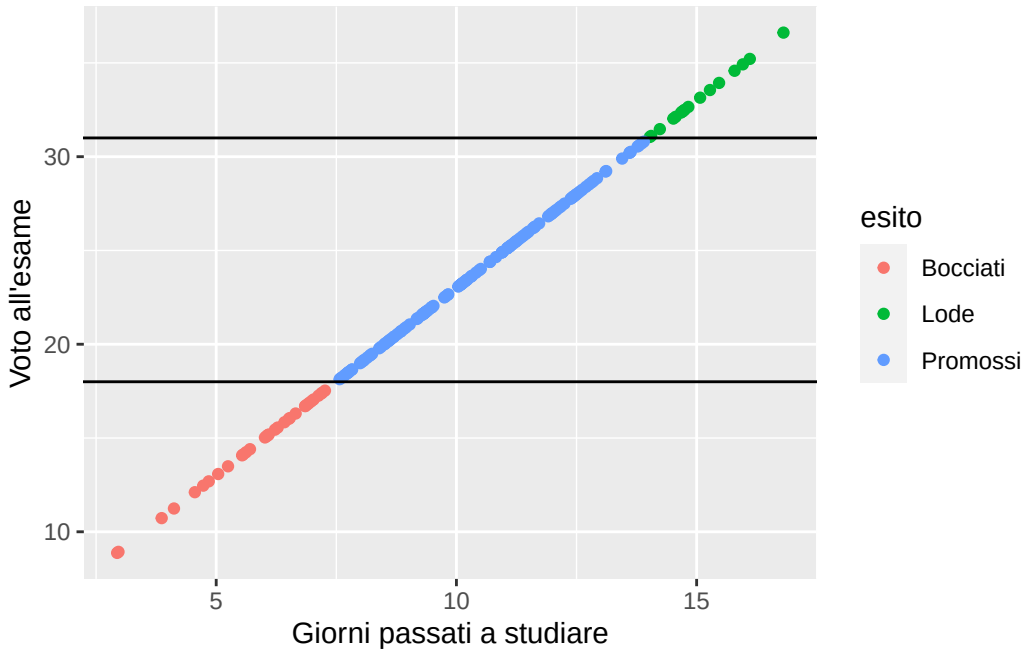
– Potremmo pure supporre che in realtà non sia così, e che in realtà voi siate distribuiti in diversi sottogruppi di studenti che prendono 0, 1, 2 o 3 punti al lavoro di gruppo. Non approfondiremo questo caso perché è più complicato, ma può valer la pena rifletterci un po' su cosa ci potrebbe aspettare in quel caso…

Cosa succede se la popolazione segue davvero esattamente questa legge?




```
<ScaleContinuousPosition>
 Range:
 Limits:    0 --    1
```

Notate che in questo momento stiamo parlando solo di punteggio all'esame! Non di voto eh! Il voto ha un limite superiore (30L) e inferiore (18) ma questo ci scombinerebbe un po' le cose. Quindi supponiamo per ora di star parlando solamente di punteggio. Adesso abbiamo dei dati, li mappiamo.



La retta di regressione

E la sua equazione...

Adesso vedremo per la prima volta in questa lezione l'esito di una regressione da parte di un programma statistico. Ci soffermeremo un po' per imparare a leggere questi numeri, e man mano che andiamo avanti, sperabilmente, acquisteranno significato.

```
========================
              Model 1
------------------------
(Intercept)    3.00 ***
               (0.00)
x              2.00 ***
               (0.00)
------------------------
R^2            1.00
Adj. R^2       1.00
Num. obs.    200
========================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

Premio grazie al cavolo 2023.

La realtà è ben diversa...

Adesso andiamo a simulare dei dati un po' più **realistici**.

- **La realtà in generale non è deterministica**, ma è piena di rumore (errore causale) e bias (errore sistematico). Per chi volesse approfondire leggete questo libro.
- Ovviamente la relazione fra giorni di studio e voto sarà influenzata da tante cose che non consideriamo (metodo di studio, quoziente intellettivo, se avete una vita piena di distrazioni, capacità di concentrarvi, stato di salute etc etc).
- Dobbiamo allora ripresentare il convitato di pietra della regressione. Il temibile EPSILON (che non a caso sembra il nome di alieno cattivo).

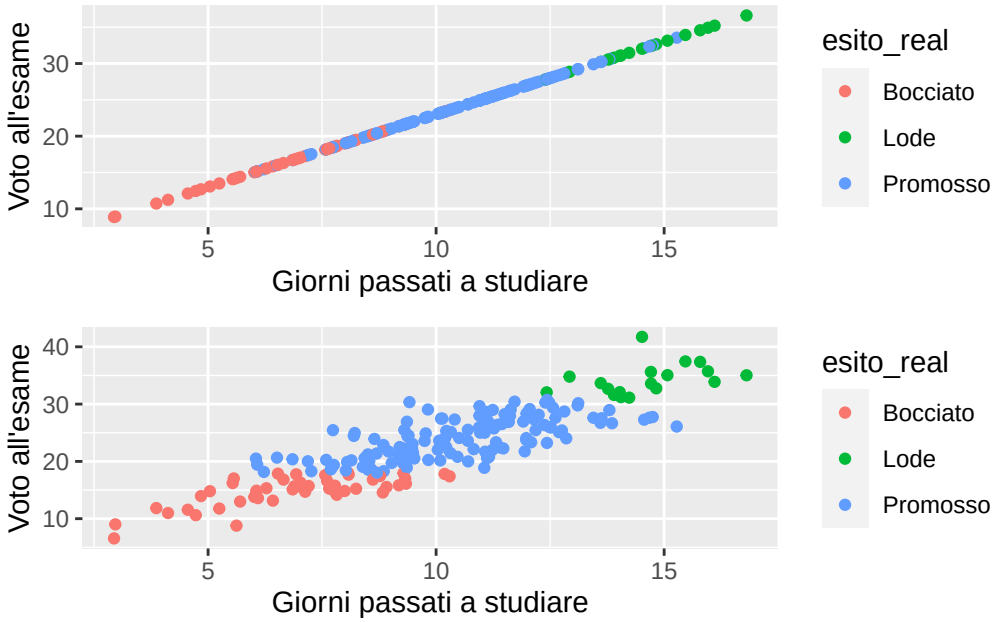
Il temibile epsilon Ovvero *L'errore in un mondo di giusti*

$$y = \beta_0 + \beta_1 * x_1 + \epsilon$$

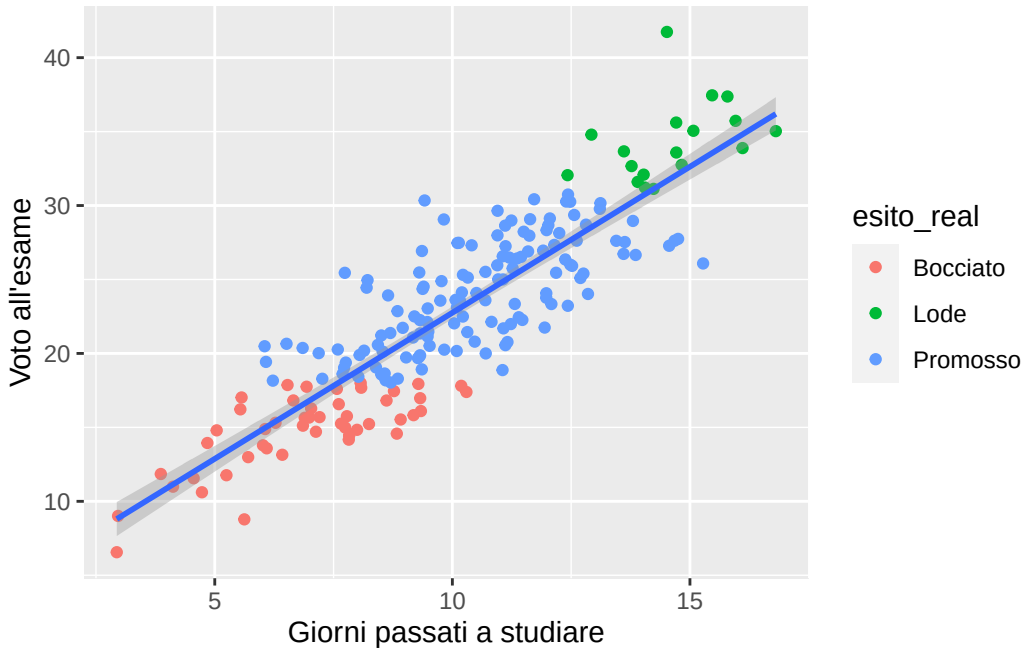
Per ora assumeremo l'errore come **normalmente distribuito**. Assumiamo che non ci sia bias (errore sistematico, ad esempio nel nostro strumento di misurazione) e che la deviazione standard dell'errore sia pari a 3. La nostra equazione adesso sarà:

$$Punteggio = punteggiolavorodigruppo + 2 * giornistudiati$$

Mondo deterministico vs Mondo reale

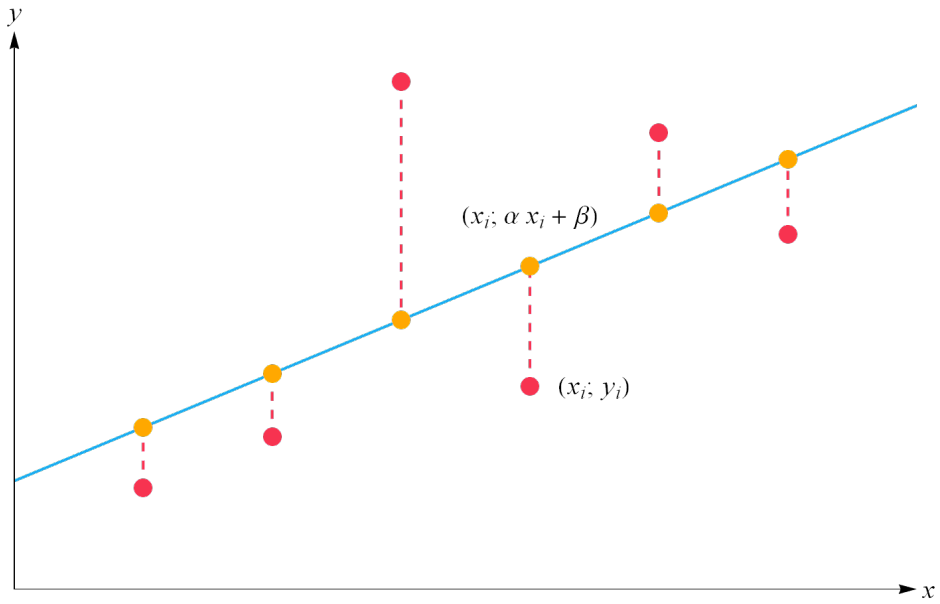


La regressione nel mondo reale



Ma come abbiamo trovato questa retta? Come fa il computer a calcolarsela?

Detour: *I minimi quadrati* Ovvero *Distanze al quadrato delle mie brame chi è la retta più minimizzante del reame*



Arrivato a questa parte della lezione progetto di passare dieci minuti a commentare questa immagine e approfondire solo se c'è tempo e voglia da parte vostra. In ogni caso, **se proprio aveste voglia di approfondire** potreste cliccare su questo link o se foste proprio tanto ma tanto interessati su questo qui, che spiegano queste cose in maniera molto più semplice e diretta di come sono in grado di farlo io. Resto comunque a disposizione se volete approfondire il discorso di persona, potete sempre scrivermi per mail. Alla fine, comunque, si arriva a questi due bad boys.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

I coefficenti nel caso "realistico" sono:

```
============================
                  Model 1
----------------------------
(Intercept)        3.00 ***
                  (0.80)
giorni_studiati    1.98 ***
                  (0.08)
----------------------------
R^2                0.77
Adj. R^2           0.77
Num. obs.        200
============================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

Vale la pena qui soffermarci un attimo meglio su cosa sono lo **standard error**, il **p-value** e **l'R quadro**.

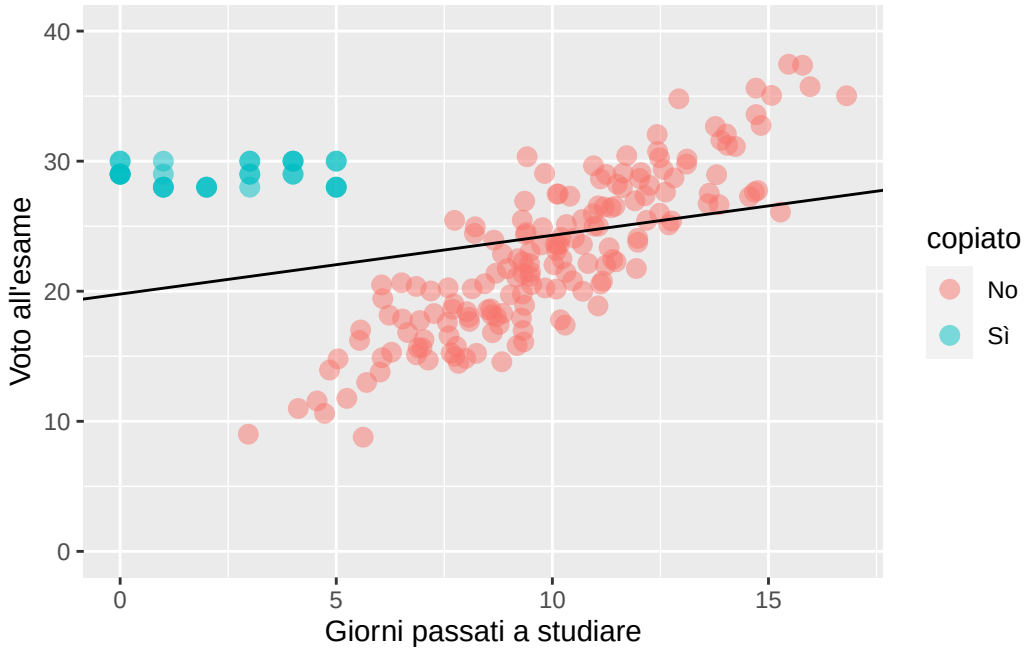
In maniera molto intuitiva:

- Lo **standard error** (il numero tra parentesi) ci aiuta a capire più o meno *quanto è precisa* la nostra stima dei coefficienti rispetto agli ignoti parametri reali
- Il **p-value** (gli asterischi!) ci dice più o meno *quanto siamo sicuri* che la nostra stima sia precisa.
  - Per i pignoli: il p-value è "*la probabilità che avremmo di osservare un valore così diverso o anche di più dall'ipotesi nulla (0) se l'ipotesi nulla fosse vera*". Quindi la probabilità che avremmo di ottenere queste stime dei parametri, o delle stime ancora più lontane da zero se i coefficienti reali fossero uguali zero (fra le variabili non esistesse alcuna relazione di qualsivoglia sorta).
- L'**R-quadro** ci dice *quanto bene il nostro modello di regressione spiega* [la covarianza tra] *i dati*

E se qualcuno copia? Ovvero *Il mio scopo nella vita*

Notate che ora i coefficienti non sono esattamente gli stessi di prima perché c'è l'errore, però la regressione cerca di fare del suo meglio.

Ma se ci fosse qualcuno che copia? **Supponiamo che fra di voi, trenta persone abbiano copiato.** Queste persone prendono un voto alto (28,29,30) studiando meno della media (da 0 a 5 giorni).




```
============================
                  Model 1
----------------------------
(Intercept)       19.78 ***
                  (1.06)
giorni_studiati    0.45 ***
                  (0.11)
----------------------------
R^2                0.08
Adj. R^2           0.07
Num. obs.        200
============================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

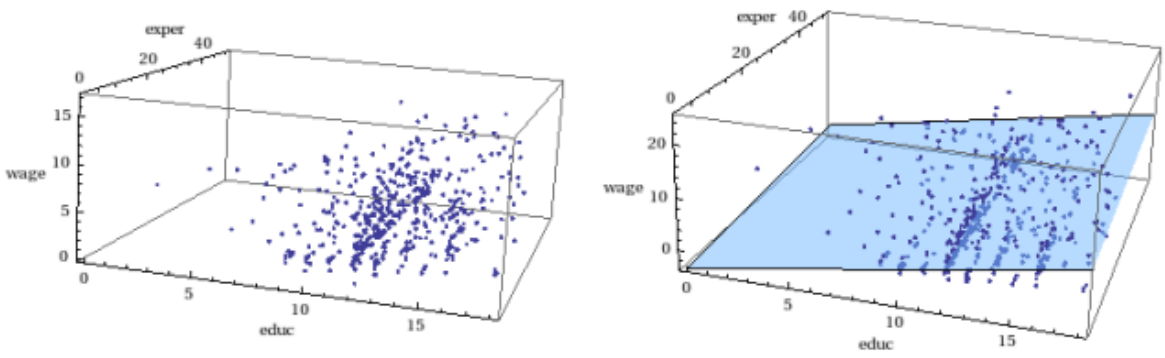
Qui abbiamo ciò che viene chiamato **omitted variable bias**. Il nostro modello non tiene conto del fatto che esiste un fattore, ovvero il **copiare,** che è fortemente correlato al fatto di prendere un voto alto e al fatto di studiare poco. Questo fattore ci fa sottostimare l'effetto che

i giorni di studio hanno sul voto e ci fa entrare nel mondo del **multivariato**. Introduciamo un nuovo modello

$$Voto = \beta_0 + \beta_1 * giornistudiati + \beta_2 * copiato + \epsilon$$

Notiamo un po' di cose:

- Per prima cosa teniamo presente che **copiato è una variabile nominale,** in particolare **dicotomica**. Come accennato prima, nella regressione lineare non abbiamo problemi a inserire **fra i predittori** variabili che non siano numeriche. Per mettere invece una variabile non numerica al posto della variabile dipendente abbiamo bisogno di tecniche più avanzate che oggi non trattiamo.
  - Le persone più volenterose fra voi che volessero approfondire potrebbero partire dal capitolo sulla **regressione logistica** di questo libro, a mio tempo l'ho trovato molto chiaro.
- Adesso abbiamo due variabili indipendenti. ciò significa che non siamo più nel mondo del **piano cartesiano**, ma in questo caso siamo in uno spazio. E non si parla più di una retta che minimizza, ma di un **piano.**
  - In generale, quando avremo più di 2 variabili indipendenti, ci troveremo in degli iperspazi e non potremo più avere una rappresentazione grafica. Per ora, con due variabili, possiamo ancora (con fatica) visualizzare.



- Il significato dei coefficienti cambia un po':
  - Se prima, quando avevamo solo una variabile indipendente:
    - L'intercetta rappresentava il valore della variabile dipendente quando la variabile indipendente stava a zero.

    - Il coefficiente angolare rappresentava il cambiamento nella variabile dipendente che si otteneva "aumentando di un'unità" la variabile indipendente.

- Ora le cose cambiano un po':

    - β_0 rappresenta il valore di ***Voto* quando *sia giornistudiati che copiato sono uguali a zero***, ovvero rappresenta il voto che prenderebbe uno che ha studiato zero giorni e non ha copiato.

    - β_1 rappresenta **quanto aumenta *Voto al variare di un'unità di giornistudiati e tenendo fermo copiato***, ovvero rappresenta quanto aumenterebbe il voto di una persona che ha copiato se copiasse comunque ma studiasse un giorno in più, oppure quanto aumenterebbe il voto di una persona che non ha copiato se non copiasse comunque ma studiasse un giorno in più.

    - β_2 rappresenta **quanto aumenterebbe *Voto* al variare di un'unità di copiato e tenendo fermo giornistudiati**, ovvero quanto aumenta il voto di una persona che studia un certo numero di giorni e non copia se studiasse lo stesso numero di giorni ma copiasse.

Calcoliamo ora i coefficienti per questo modello.

```
============================
                  Model 1
----------------------------
(Intercept)         4.66 ***
                   (0.89)
giorni_studiati     1.82 ***
                   (0.09)
copiato_r          20.03 ***
                   (0.90)
----------------------------
R^2                 0.74
Adj. R^2            0.73
Num. obs.         200
============================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

Possiamo notare che l'**omitted variable bias non c'è più** e che una volta incluso **copiato** nel modello siamo di nuovo in grado di vedere l'effetto "reale" di **giornistudiati**.

Un esempio, con dati sempre simulati, di regressione multivariata

Adesso vedremo un esempio di regressione lineare multivariata. Quindi una regressione che, genericamente, ha la forma…

$$y = \beta_0 + \beta_1 * x_1 + ... + \beta_k * x_k + \epsilon$$

Siccome non ci vogliamo male, noi faremo la regressione solo con due variabili, come abbiamo fatto prima quando abbiamo usato "copiato" e "giornistudiati".

Dobbiamo stare attenti però ad una cosa:

- Quando abbiamo una variabile indipendente **numerica "x"**, il coefficiente ci dice la differenza in y che si avrebbe aumentando x e tenendo ferme **tutte le altre variabili indipendenti**
- Quando abbiamo una variabile indipendente **dicotomica,** il coefficiente ci dice la differenza in y che si avrebbe passando da uno dei due possibili stati di x all'altro (da "copia" a "non copia" e viceversa - ogni variabile dicotomica in fondo nasconde una *dummy…*) tenendo ferme tutte le altre variabili indipendenti.
- Ma cosa succede quando abbiamo una variabile che ha più livelli, ma non è numerica, tipo una variabile **categoriale** (nazionalità) o **ordinale** (titolo di studio)?
  - In questo caso, le cose si complicano un po'. Vediamolo con un esempio e poi ne parliamo.

Altro simpatico esempio

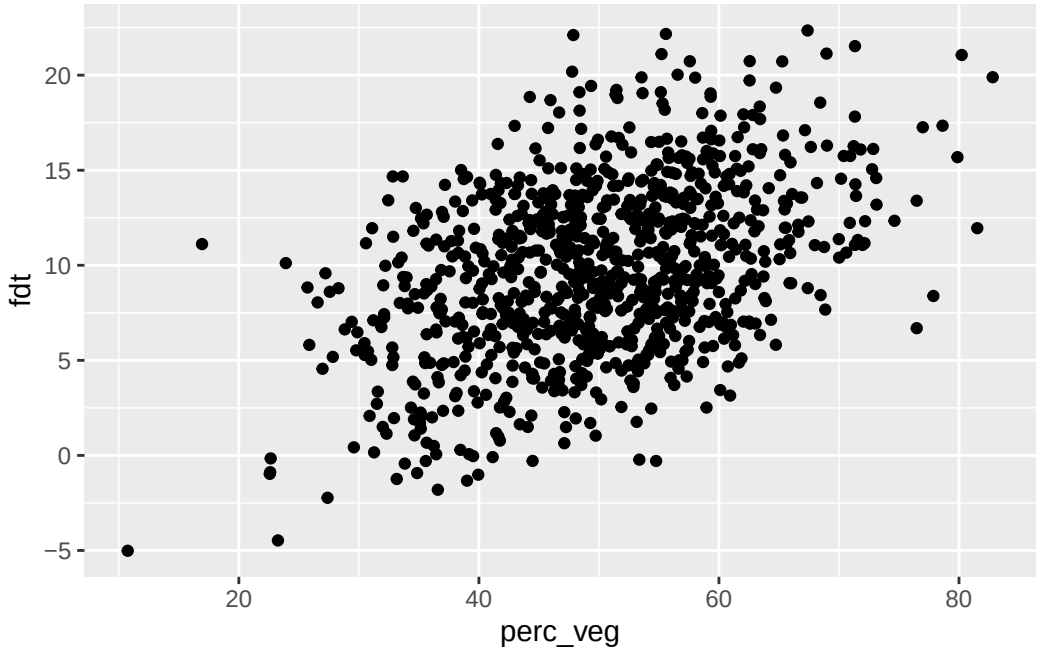
Supponiamo di avere un database di risultati elettorali. Abbiamo 900 comuni italiani a caso, di cui conosciamo solo queste informazioni:

- La **zona geografica** (nord, centro, sud).
- La **dimensione** (piccolo, medio, grande, metropoli).
- la **percentuale di vegani** in quel comune.
- la **percentuale di voti presi** in quel comune da tre partiti:
  - Il partito di sinistra "**Fratelli di Tofu**" (FdT)
  - Il partito di centro "**ColAzione**" (C).
  - Il partito di destra "**Italia dei Sapori**" (IdS).

Supponiamo di avere una teoria molto semplice: che possiamo spiegare la percentuale di voti a Fratelli di Tofu con il seguente modello:

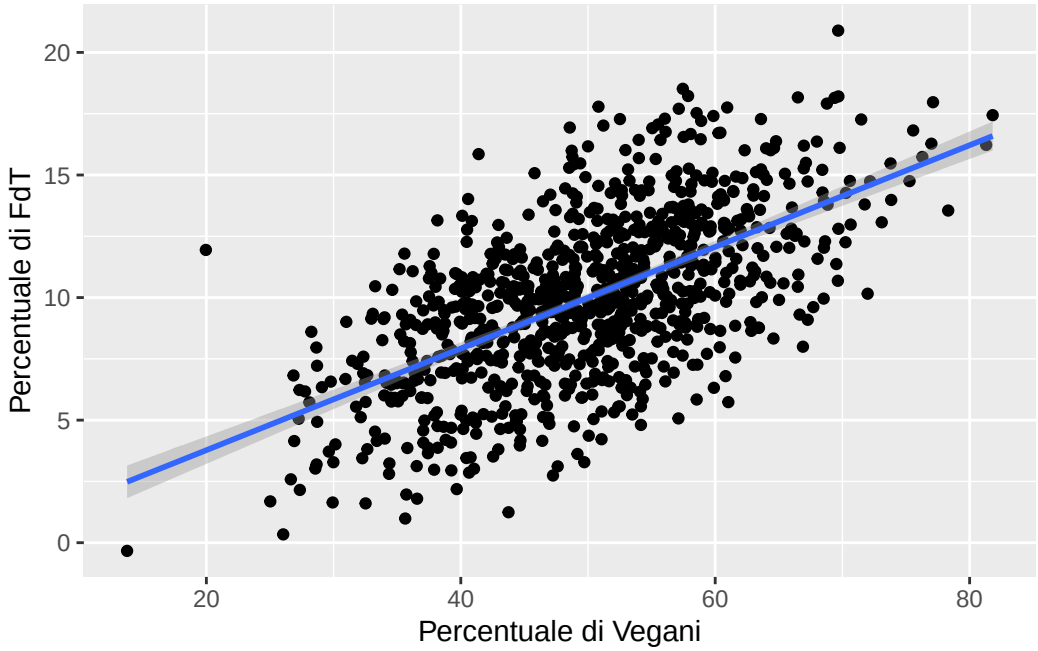
$$FdT = \beta_0 + \beta_1 * percvegani + \beta_2 zona + \epsilon$$

Andiamo un po' a esplorare la relazione fra percentuale di vegani e voto per FDT.



Questo grafico così com'è non ci dice molto.

Proviamo a fittarci ad occhio **una linea**…



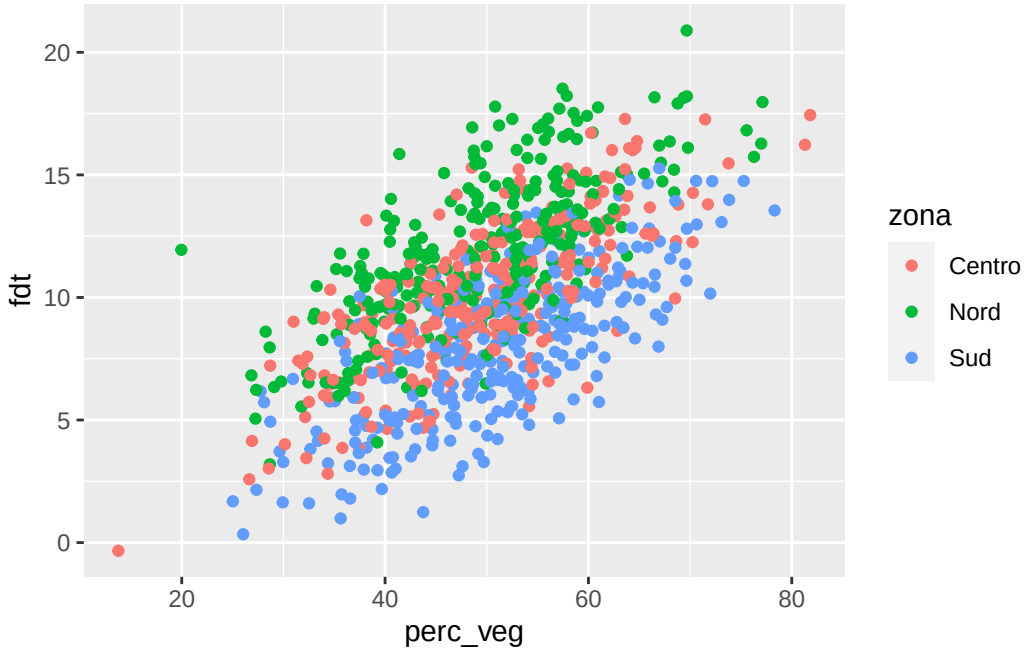
L'effetto è complessivamente positivo, ma non abbiamo ancora minimamente tenuto conto della variabile zona**.** Intanto però vediamo come sarebbe il coefficiente di regressione se tenessimo come unico predittore la percentuale di vegani.

```
========================
              Model 1
------------------------
(Intercept)    -0.37
               (0.46)
perc_veg        0.21 ***
               (0.01)
------------------------
R^2             0.37
Adj. R^2        0.37
Num. obs.     900
========================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

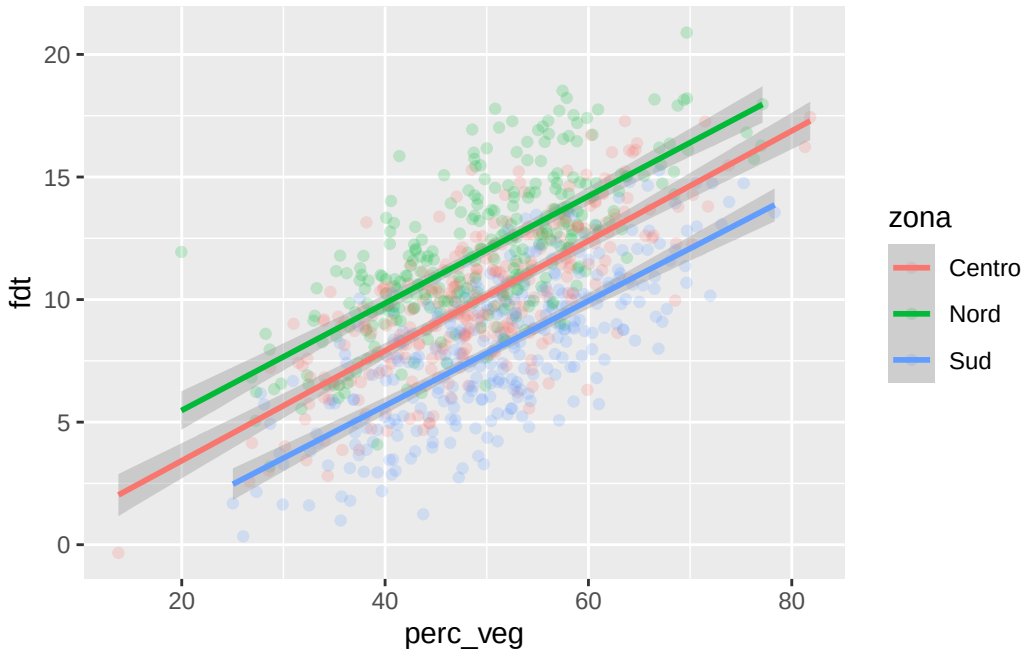
In questo caso siamo in un caso piuttosto fortunato di regressione multivariata. Siccome la nostra "terza variabile", ovvero "zona" (nord, centro o sud) è una **variabile categoriale** con pochi livelli, con un bel trucchetto possiamo ancora visualizzare la regressione. Questo trucchetto sono **i colori.** Mi chiedo dunque:

- E se **a seconda della zona** i diversi gruppi seguissero delle leggi **così diverse** da confondermi i dati?

Proviamo a visualizzare questa cosa...



Adesso mi sembra di intravedere un pattern... Per esplorarlo meglio proviamo a fare una cosa. Proviamo a trattare ogni gruppo geografico come se fosse un dataset a sé e fittiamo una regressione su questi dati. Vediamo se esce fuori qualcosa di interessante...



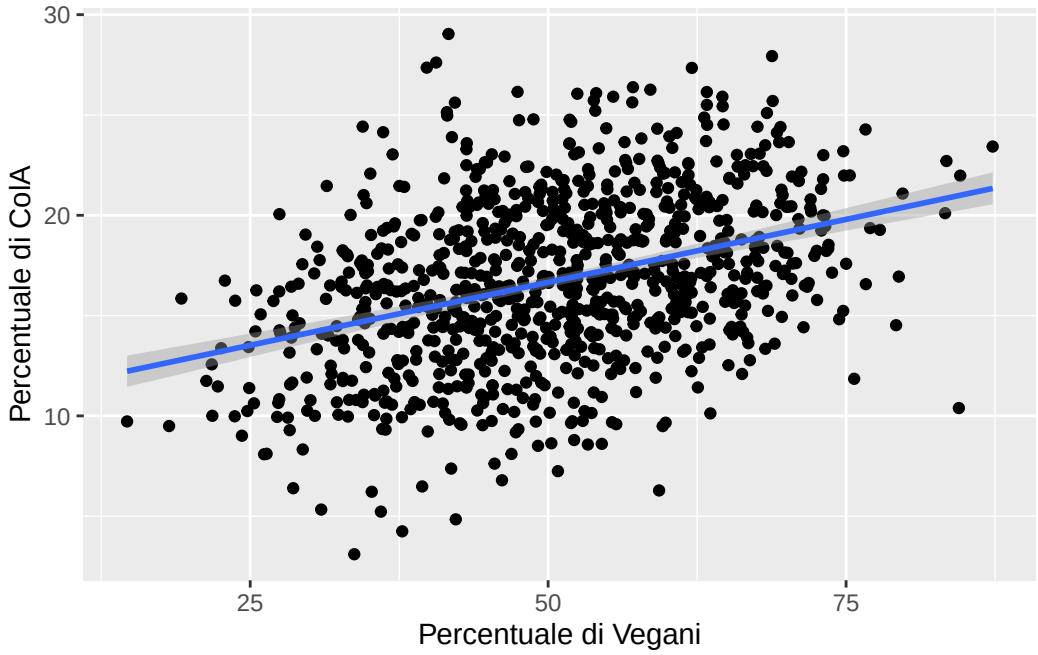
Wow! Quello che esce fuori è clamoroso! A seconda della zona in cui sono cambia moltissimo la mia probabilità di votare per FdT a parità di percentuale di vegani. Andiamo a vedere quanto facendo una bella regressione.

```
============================
                  Model 1
----------------------------
(Intercept)        -0.79 *
                   (0.37)
perc_veg            0.22 ***
                   (0.01)
factor(zona)Nord    1.88 ***
                   (0.17)
factor(zona)Sud    -2.35 ***
                   (0.17)
----------------------------
R^2                 0.63
Adj. R^2            0.62
Num. obs.         900
============================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

Notiamo che in questo caso la **categoria di riferimento è il centro**! Quindi i coefficienti di Nord e Sud ci dicono, rispettivamente, quanto varia la percentuale di voti presi da FdT, a parità di percentuale di vegani, spostandoci da un comune del centro a uno del nord o da un comune del centro a uno del sud.

La correlazione non significa causazione

Andiamo adesso ad esaminare il voto per ColAzione in relazione alla percentuale di vegani.



La linea di regressione mi dice che effettivamente la correlazione fra le due variabili è positiva. Infatti, posso anche calcolarmi esplicitamente la correlazione in questi dati e otterrei:

```
[1] 0.3722813
```

Quindi, calcolandomi la regressione, concludo che…

```
========================
             Model 1
------------------------
(Intercept)   10.39 ***
              (0.54)
```

```
perc_veg       0.13 ***
              (0.01)
----------------------
R^2            0.14
Adj. R^2       0.14
Num. obs.    900
======================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

Quindi ogni punto perentuale in più di vegani che ci sono in un comune mi aspetto che ColAzione prenda lo 0.13% (circa, la stima precisa non ve la so dire prima di far eseguire la regressione al programma) in più. In un comune col 100% di vegani, ad esempio, ColAzione prenderà circa il…

$$ColA = 10.13 + 0.13 * 100 = 23,13$$

Quindi in quei comuni ColAzione prenderà il 23,13%!

C'è però un fatto da considerare: immaginiamo che alle ultime elezioni, ColAzione abbia preso un punteggio altissimo a Milano (sopra il 20%) dove il numero di vegani è tutto sommato nella media, ma ha preso un punteggio più basso a Reggio Emilia (10%) dove invece il numero di vegani è altissimo. Questo ci suggerisce che forse dovremmo scavare un po' di più nei dati. Ci chiediamo dunque cosa potrebbe succedere se includessimo come predittore, ad esempio, la dimensione della città (piccola, media, grande, metropoli).

```
========================
              Model 1
------------------------
(Intercept)    13.08 ***
               (0.50)
perc_veg        0.00
               (0.01)
dimMedia        2.38 ***
               (0.31)
dimGrande       4.91 ***
               (0.34)
dimMetropoli    7.09 ***
               (0.37)
------------------------
R^2             0.40
Adj. R^2        0.40
```

```
Num. obs.     900
=======================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

Inserendo la variabile "dimensione" ecco che improvvisamente l'effetto della percentuale di vegani presenti nel comune sul voto a ColAzione diventa sostanzialmente irrisoriso! Solo guardando i dati in questo modo, attraverso una regressione, potevamo renderci conto del fatto che le variabili si relazionano così.

- Notate bene che la "**categoria di riferimento**" della variabile dimensione è "piccolo". Quindi quei coefficienti indicano "quanto è la differenza nell'effetto passando da piccolo a quell'altro livello della variabile dimensione".

Occhio però! Anche questa regressione non è assolutamente abbastanza per dire che la dimensione della città ***causa*** il voto a ColAzione, per tutte le motivazioni che avete esplorato ed esplorerete questo semestre!

Tanto per dirne una, potremmo anche ipotizzare il rapporto causale sia **inverso**: chi abita in una grande città vota ColAzione perché lì ha la possibilità di provare tanti posti dove fare colazione o è chi vota ColAzione che preferisce trasferirsi in una grande città, dove ha un vasto assortimento di posti dove fare colazione?

Questo genere di domande sono il pane di chi fa analisi dei dati e distinguono lo scienziato sociale dal data scientist puro.

Grazie!

- Gaetano Scaduto, Analysis of Social and Economic Processes, University of Milan-Bicocca g.scaduto2@campus.unimib.it