MPSMF_lez08

Cosa facciamo oggi?

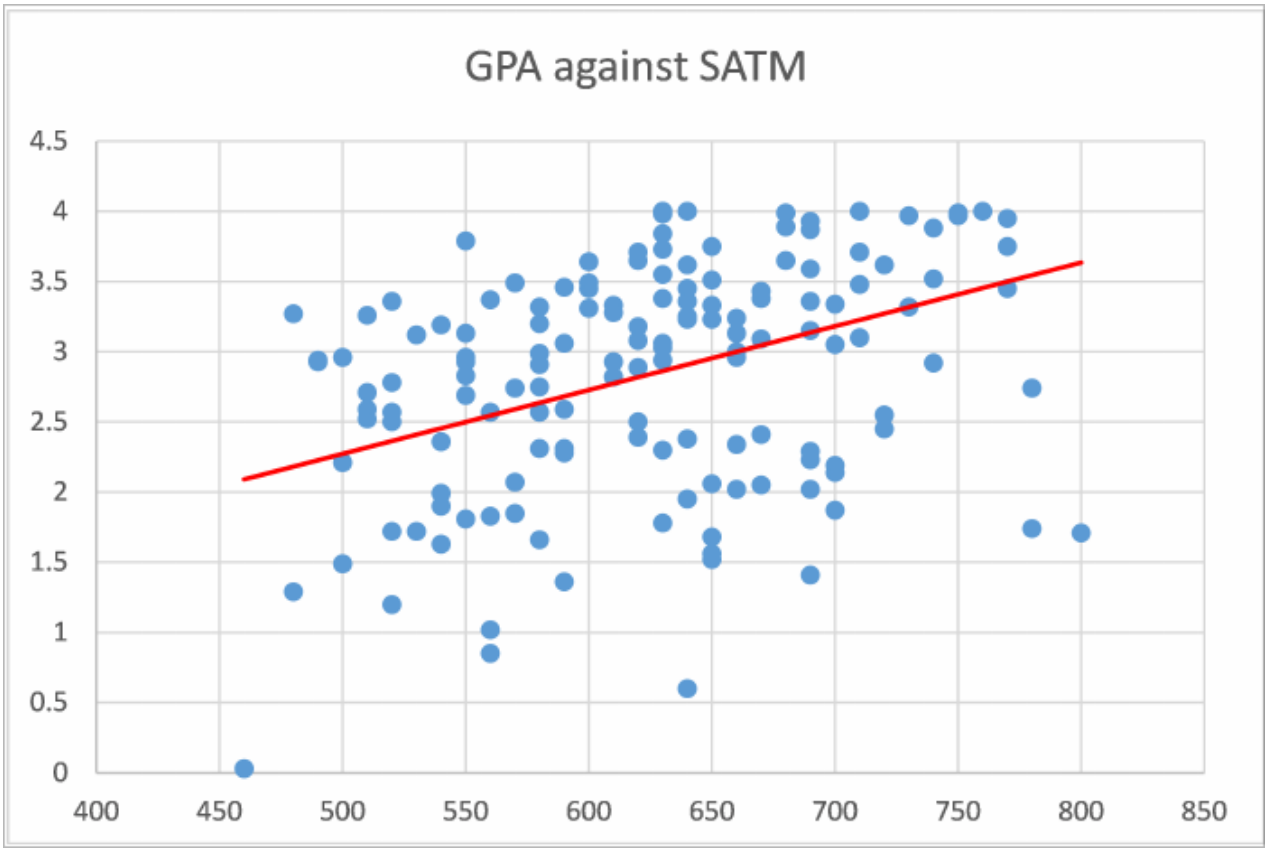
Discussione note sulla regressione lineare:

Sostanzialmente con la regressione vogliamo banalmente cercare di approssimare una v.a
La **speranza condizionata** è la funzione di regressione per eccellenza.
Nella realtà però trovare la speranza condizionata è un qualcosa di complesso, quindi noi in pratica procederemo con un *procedimento a priori*.
Ovvero noi diremo *a priori* che l'approssimatore è una retta, ad esempio, e quindi andiamo a cercare tra tutte le rette la migliore.
Nella maggior parte dei casi ci dovremo "accontentare", ma nel caso in cui le v.a. solo delle v.a. gaussiane e disgiunte allora l'approssimazione mediante una retta è anche la migliore.
Chiaramente bisogna sempre fare attenzione all'**overfitting**.

Per certe realtà che si verificano nei Mercati Finanziari ad esempio ci conviene fissare come famiglia, quella delle funzioni esponenziali.

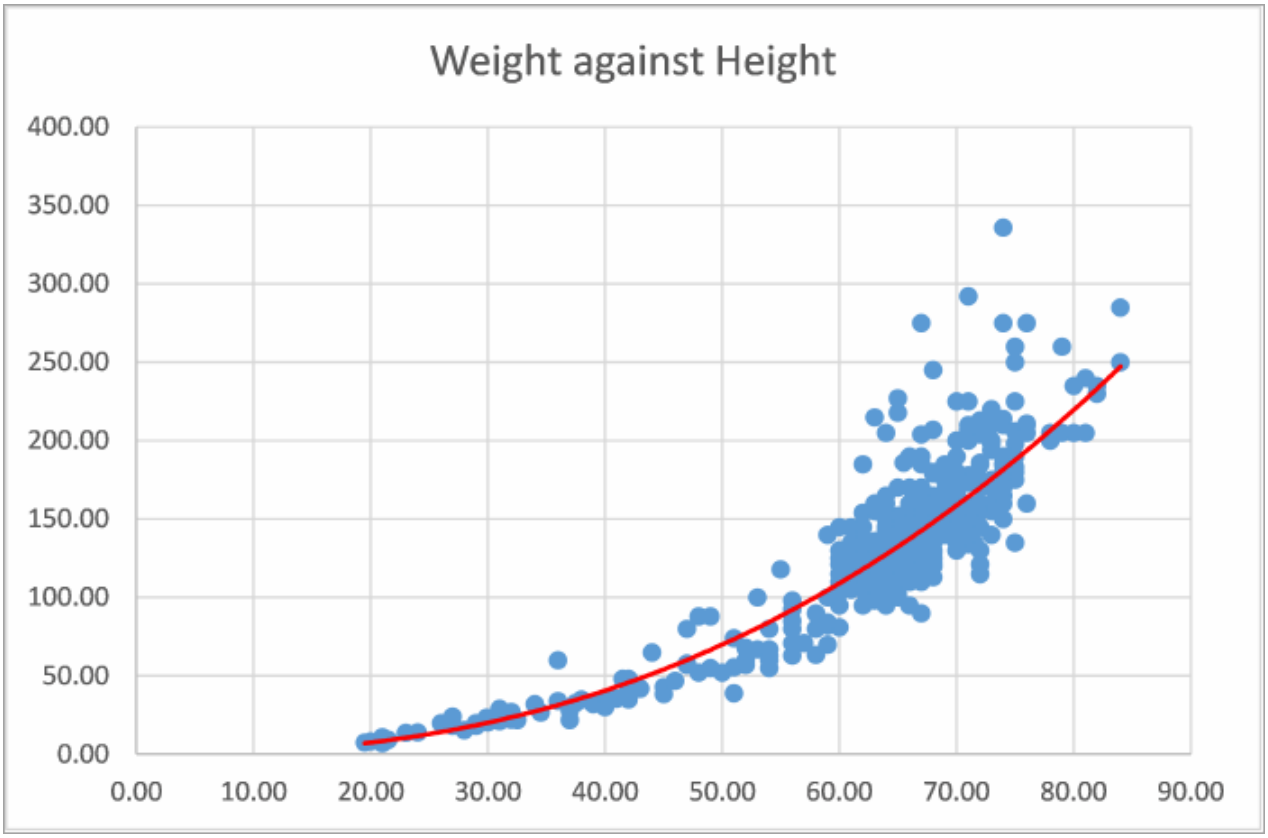
> **Esempio**: la legge di Hooke ci dice che:
>
> $$L = L_0 + KF$$
>
> Quindi ci dice qual'è il comportamento di una molla ideale.
> Possiamo quindi prendere una legge che prevede anche l'aggiunta di un rumore:
>
> $$L = L_0 + KF + U$$
>
> dove U modella il rumore.
>
> **Esempio**: due statistici si sono posti il seguente problema: "le performance di studenti universitari dato la loro valutazione in matematica alle superiori".
> Quello che si sono accorti è che la retta di regressioni di questo dataset era una retta con coefficiente angolare positivo. Quindi hanno potuto affermare che c'era una sorta di correlazione. Chiaramente non conosciamo la legge teorica, stiamo

semplicemente usando una legge empirica.

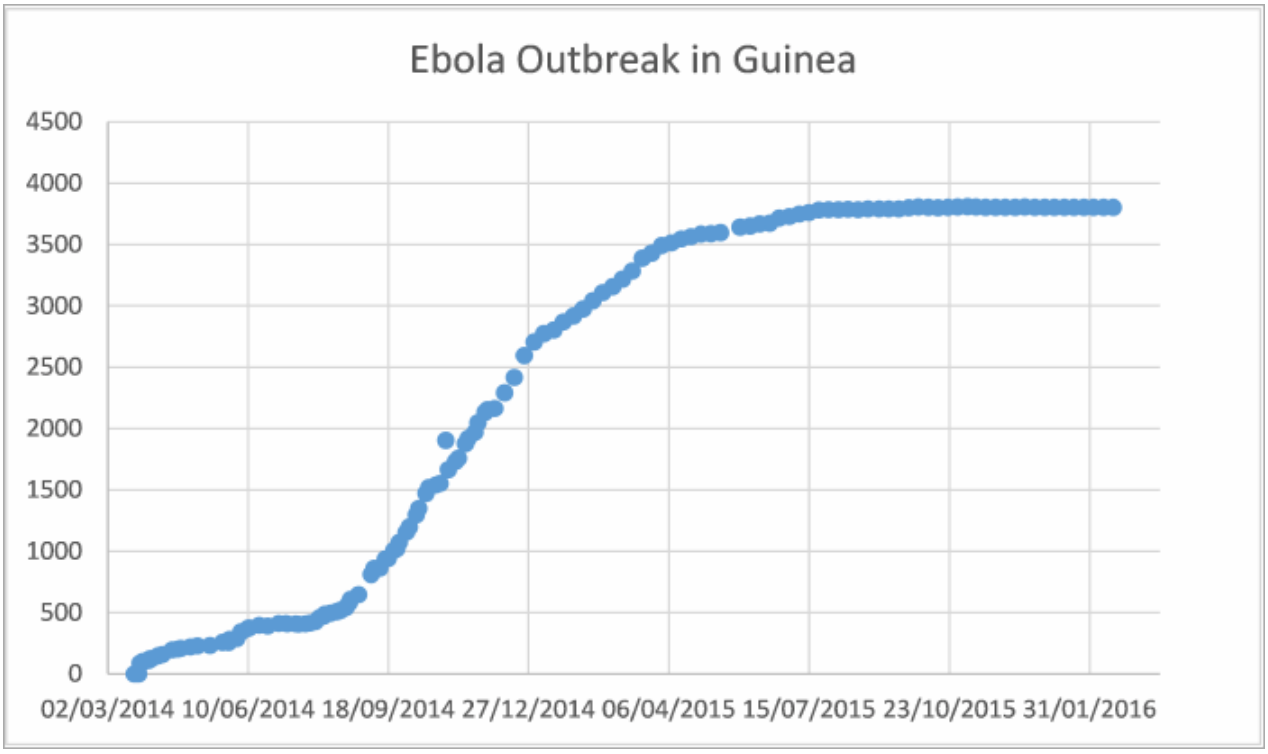


Esempio: un professore dell'Università dello Utah ha fatto il seguente esperimento: "chiedere a tutti gli studenti di portare una serie di coppie peso e altezza". Quello che ha dimostrato è che la "legge del BMI" è in prima approssimazione corretta. Quello che si nota osservando il grafico sotto (asse x altezza, asse y peso) però è che abbiamo una variabilità minima per valori di altezza bassi (a sinistra) e una variabilità via via crescente al crescere delle altezze. Questo fenomeno si chiama

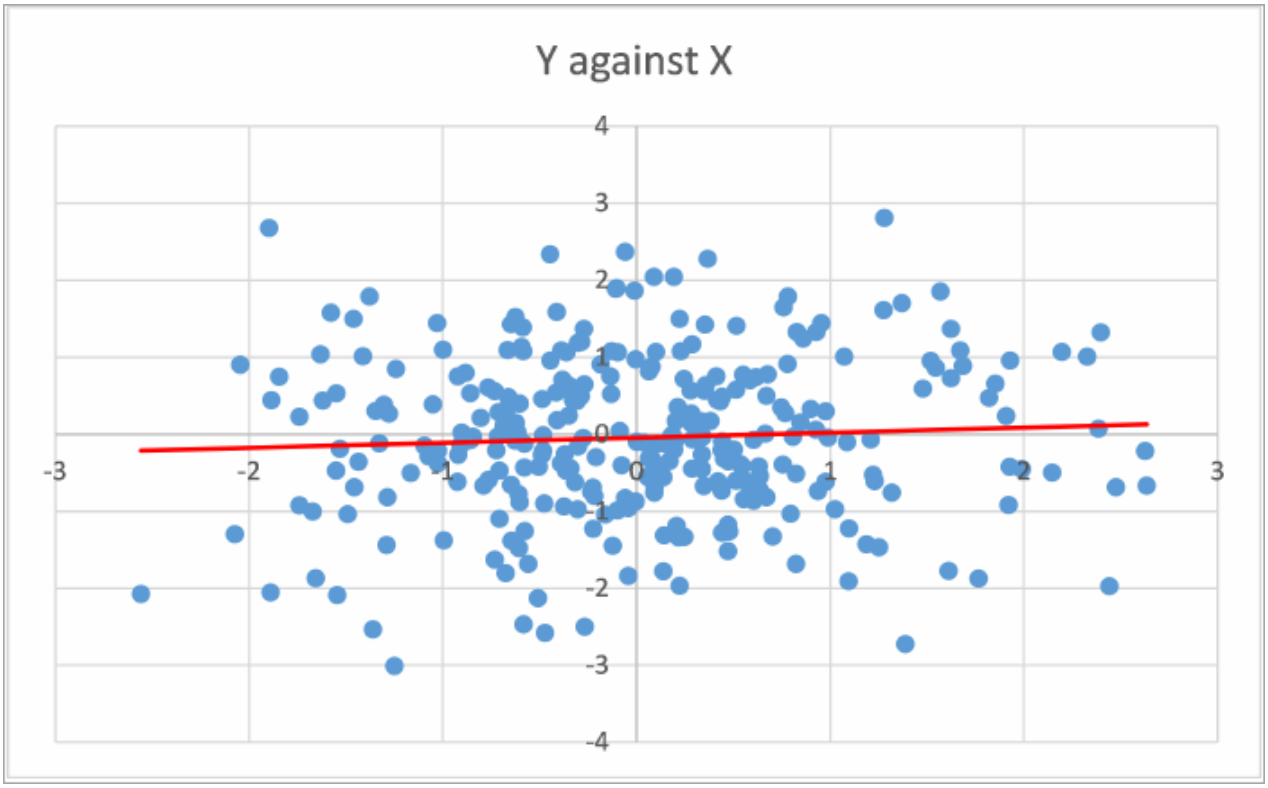
eteroschelasticità.



Esempio: una cosa interessante che ha trovato Monte riguardo la regressione logistica è un suo collegamento con l'Ebola.
Tracciando i grafici dell'epidemia produce delle sigmoidi molto evidenti:



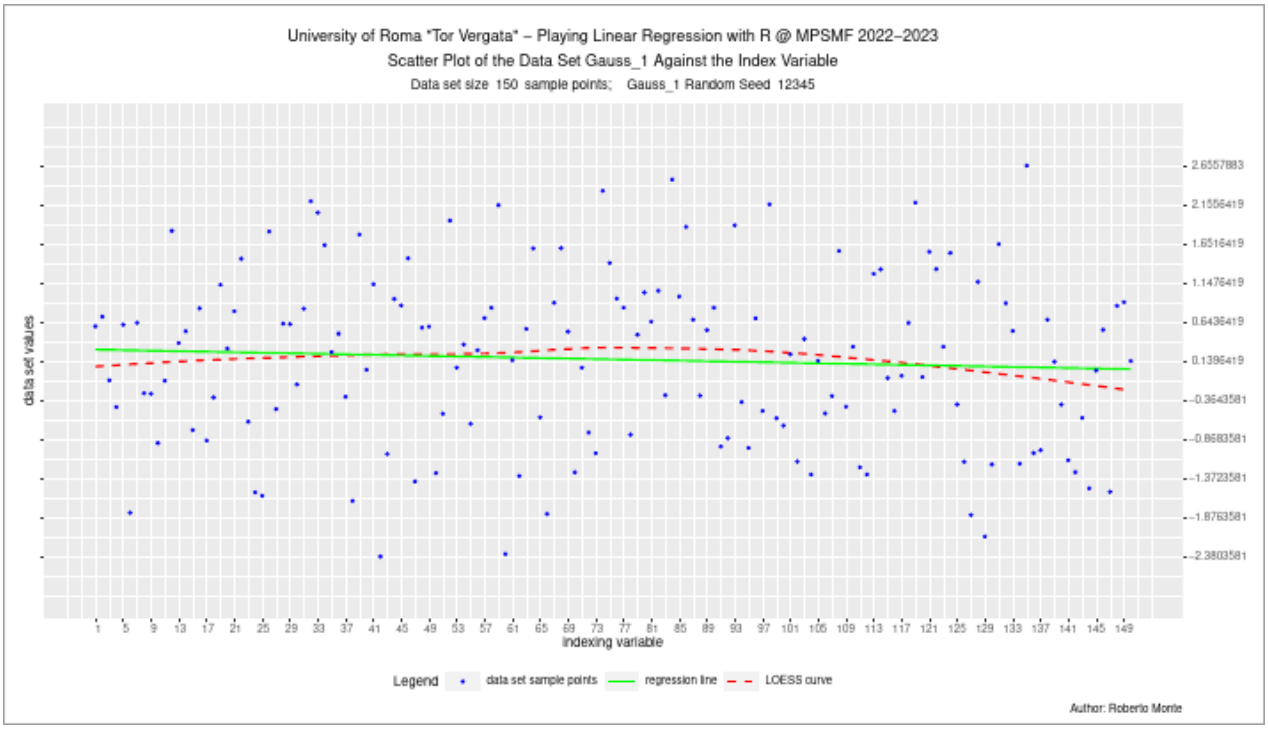
La derivata di questa curva è un qualcosa di molto simile ad una curva a campana di Gauss.
La curva dell'espansione del Covid-19, nelle sue prime fasi, è praticamente molto simile.
Sempre allegato all'esempio dell'Ebola troviamo quest'altro grafico:



Cosa ci dice questo grafico? Ci dice chiaramente che le due variabili aleatorie Y e X non hanno alcuna correlazione.

Esempio: al minuto 54:00 circa siamo andati a osservare l'esempio 3.1.
i>Questo è un esempio sulla regressione lineare.
L'esempio viene generato un dataset Gaussiano e su questo dataset andremo a generare un'approssimatore lineare.

Questo è uno dei grafici generati:



Esempio: Biscotti Gentilini. I proprietari della ditta hanno forniti questi dati a Monte. (non sembra niente di interessante).

Quando si fa un'analisi statistica in cui si cerca di evidenziare una certa correlazione statistica, formalmente quello che si sta facendo è cercare di scrivere la relazione:

$$Y = f(X;\theta) + U,\ \theta \in R^M$$

Perché la regressione non sia cattiva, si dovranno creare certe condizioni. Queste condizioni sono:

- **Stazionarietà**: per verificarla si effettuano dei **test di stazionarietà** (esempi di questi test possono essere: test di Dickey Fuller (ADF), test KPSS)).
  Il test di ADF prende come ipotesi nulla H_0 la seguente: "il processo che ha generato i dati contiene una componente di Random-Walk (detto anche: trend stocastico o Unit Root)".
  Le ipotesi alternative invece sono in genere:
  - $H_1^{(I)}$: "processo stazionario con intercetta senza trend";
  - $H_1^{(II)}$: "...";
  - $H_1^{(III)}$: "...".
    Con ADF vogliamo il rigetto dell'ipotesi nulla.

    KPSS invece ha come H_0: "il processo che ha generato i dati è stazionario". Quindi quando io ho il rigetto dell'ipotesi nulla significa che il processo non è stazionario.

    Se applicando i due test ci danno un esito contraddittorio (ovvero entrambi Rigettano o entrambi non rigettano). In questo caso non possiamo dire nulla. Nel caso in cui ADF Rigetta e KPSS non rigetta allora possiamo dire che il processo è stazionario. Il viceversa invece ci permette di stabilire che il processo non è stazionario.
- **Autocorrelazione (correlazione seriale)**: i residui non sono generati in maniera indipendente, ma la generazione dell'uno influenza la generazione di un altro. Per fare emergere questa autocorrelazione si possono usare due tecniche:
  - **Metodo grafico (autocorrelogramma)**: si va a calcolare la correlazione ...;
  - **Test** $L_{jung} - Box$: a differenza di quello di prima che è un metodo grafico, questo è un metodo computazionale.
- **Homoschedasticità**: si accerta che non si verifichino situazioni con dati fortemente variabili. I test per verificare questa proprietà è:
  - **Breish-Pagah**:
  - **White**.
- **Gaussianità**: alla fine andremo a verificare questa proprietà. Quindi andremo a vedere se i residui sono Gaussianamente distribuiti. Nelle applicazioni, nella

maggior parte dei casi reali, non abbiamo che questa relazione sia valida (quindi magari i residui sono distribuiti come una Student, come una Logistica, ...). Uno cosa può fare in questi casi? L'unico problema di questa casistica è che gli intervalli di confidenza cambiano forma; tutto qui. Anche qui ci sono varie tecniche.