COVID-19: An epidemiological perspective

P. Girardi

Dipartimento di Psicologia dello Sviluppo e della Socializzazione
Università degli Studi di Padova

April 15, 2020

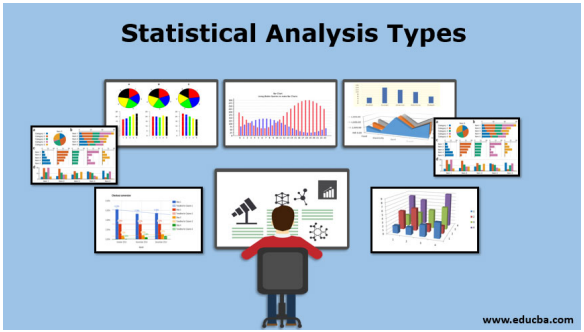
PSICOSTAT - COVID19 Talks

Sommario

E' solo un'altra analisi?

Un'analisi statistica è come una nuova sfida...
ma quando si procede con le analisi si segue uno standard più o meno consolidato

- Comprensione del fenomeno o dello studio... e del "meccanismo generatore dei dati". (Consulto con l'esperto, uno specialista, uno sciamano, un mago, etc...)
- Analisi descrittive (tanti grafici, tabelle, etc...)
- Scelta delle analisi statistiche più appropriate (test, regressioni, analisi correlazionali, etc...)
- Valutazione e discussione dei risultati ottenuti.

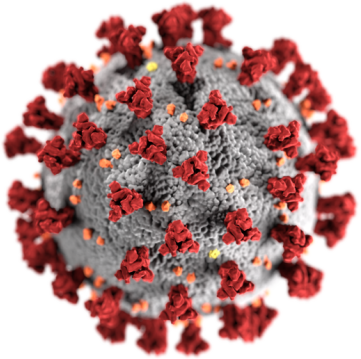


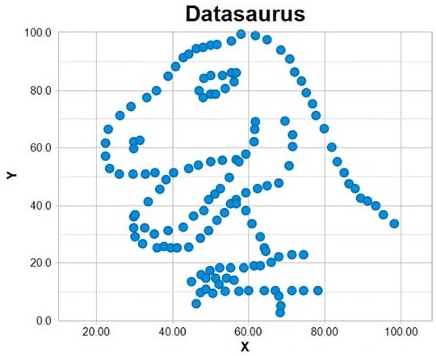
Raccomandazioni per un'analisi statistica

Alcune linee guida sulle analisi statistiche raccomandano che...

...se la storia è semplice mantienila semplice!

Il principio KISS: "*Keep It Simple, Stupid*", suggerisce di non introdurre più complessità di quanta ne serva.





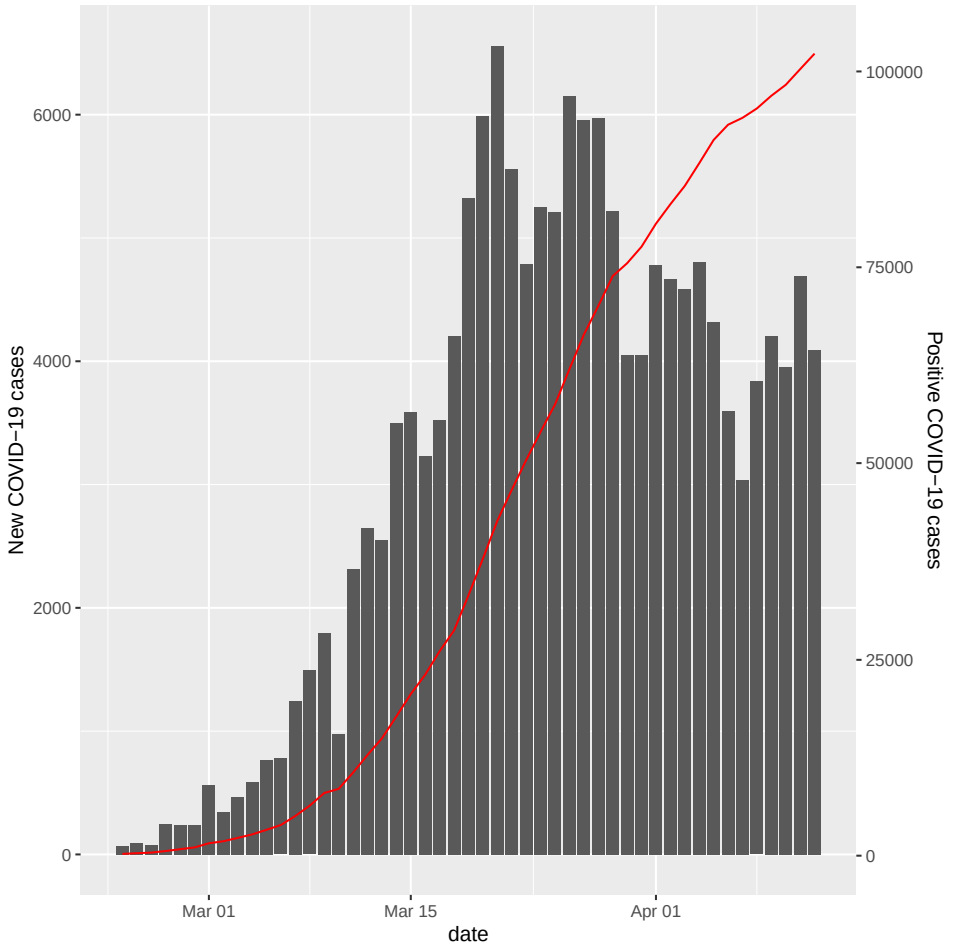
...se la storia è complicata... rendila semplice, ma non troppo!

Frase attribuita ad A. Einstein: "*Make everything as simple as possible, but not simpler*". Il tentativo di rendere semplice non deve far perdere di significato l'analisi o tralasciare caratteristiche importanti.

Covid19 - Breve intro... ma sapete già

- L'epidemia di COVID19 (a livello internazionale SARS CoV-2) è dovuta ad un nuovo coronavirus che sviluppa una malattia respiratoria.
- I primi casi di COVID19 risalirebbero al mese di dicembre 2019 dovuti a ricoveri da polmonite da causa sconosciuta tra la popolazione di Wuhan, nella provincia di Hubei in Cina.
- In Italia il virus arriva prima tramite casi indiretti provenienti dalla Cina (turisti cinesi a Roma) poi con casi diretti in alcune località della Lombardia e del Veneto (20-25 Febbraio).
- Al 13 Aprile (dopo 50 giorni di epidemia) in Italia sono stati accertati un totale di 159,516 casi e 20,465 deceduti attribuibili al COVID19.

Covid19 - Breve intro... ma sapete già

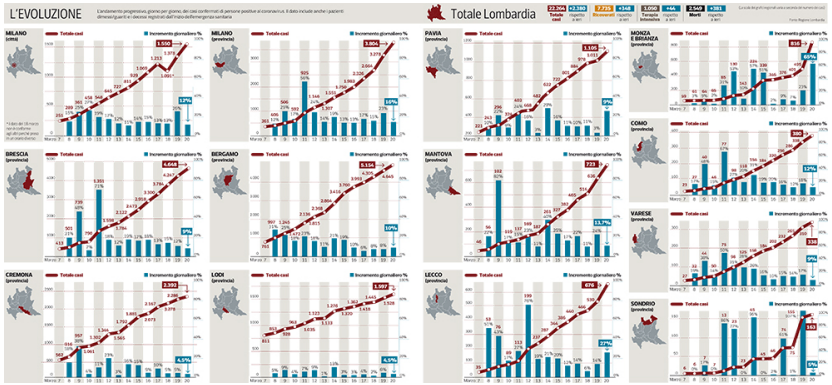


Tante analisi... tanta metodologia...

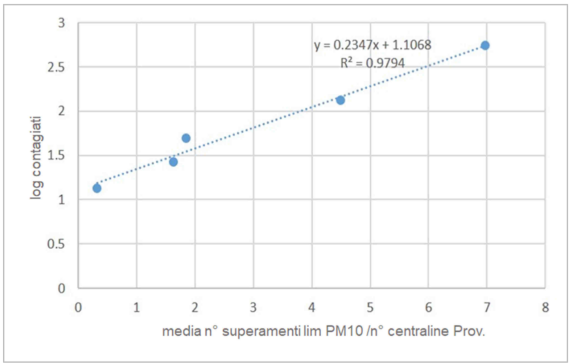
Dal punto di vista applicativo l'epidemia di COVID19 offre l'opportunità di analizzare in tempo reale dei dati che possono essere di pubblica utilità e fornire alle amministrazioni supporto in condizione di incertezza.

Si sono susseguite un importante numero di analisi statistiche.

Alcune utili

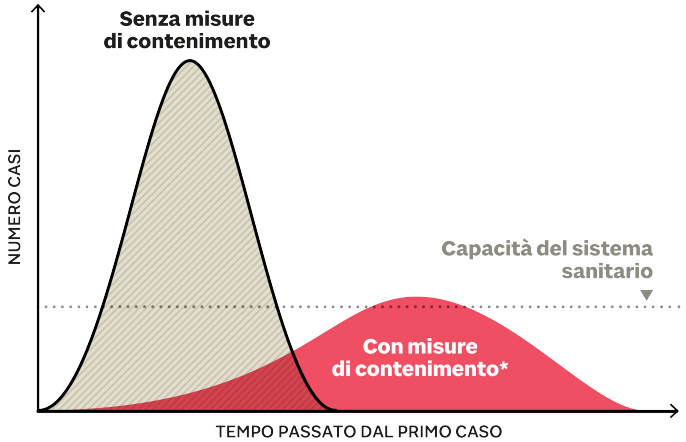


Altre un po' meno



Obiettivo dell'analisi

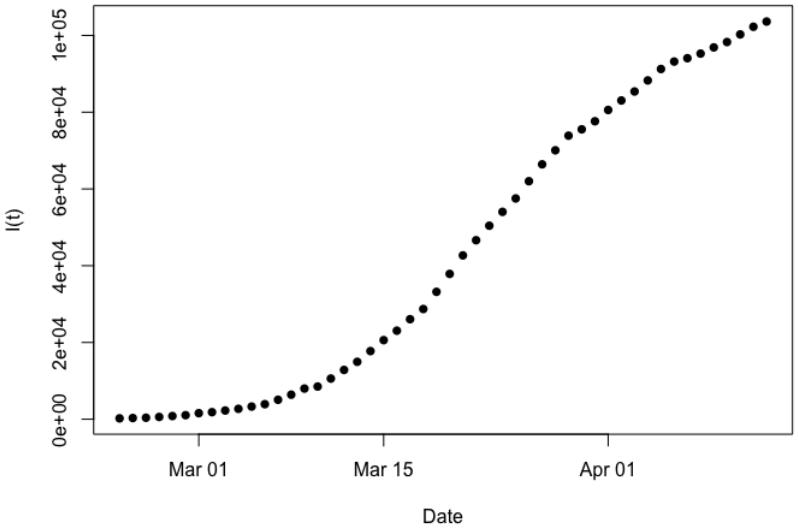
- Vogliamo studiare l'evoluzione della pandemia di coronavirus in Italia da una prospettiva statistica usando dati aggregati.
- L'obiettivo di rilevare importanti cambiamenti nel processo sottostante (casuale?) il più presto possibile dopo che si è verificato. Le misure di contenimento sono servite?



- Utilizziamo i dati forniti dal Dipartimento della Protezione Civile:
  `https://github.com/pcm-dpc/COVID-19`

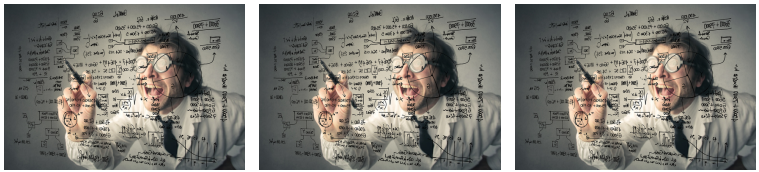
Analisi descrittiva: Covid19 in Italia

Il dataset della Protezione Civile contiene molte quantità su cui svolgere delle analisi statistiche: **Decessi, Ricoveri, Nuovi Positivi, Totale Positivi attuali, Casi totali, Guariti, Tamponi, etc....**



Ci concentriamo sul numero di casi **attualmente positivi** che è una delle misure più importanti per verificare il superamento dell'epidemia.

Covid19 - Con che metodi?



L'analisi del numero cumulato di casi positivi è un conteggio temporale. Come si analizza?

- **Esperto 1**: Sono serie temporali... realizzo grafici e ne estrapolo i trend con una regressione!!
- **Esperto 2**: Utilizzo metodi per le serie storiche: Trend, Stagionalità... tutto è spiegato da queste componenti!
- **Esperto 3**: E' un dato "tempo dipendente": c'è un autocorrelazione? Posso utilizzare modelli autoregressivi!
- **Esperto 4**: E' un conteggio quindi è una realizzazione di una variabile casuale di Poisson, modelli GLM... fatto!
- **Esperto 5**: Sono arrivato tardi.... hanno già fatto tutto loro!

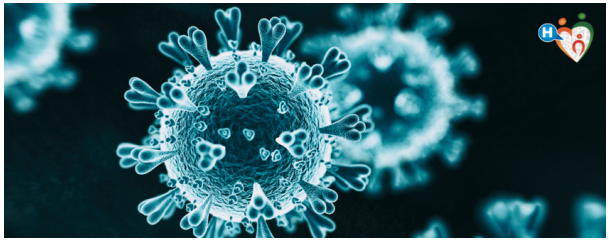
Ed è quello che sta succedendo...

Ecco un esempio (mooolto limitato) di articoli nell'ultimo mese:

- **Esperto 1**: Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020 (Roosa et al., 2020);
- **Esperto 2**: Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models (Dehesh et al., 2020);
- **Esperto 3**: A Poisson autoregressive model to understand COVID-19 contagion dynamics. (Agosto and Giudici, 2020);
- **Esperto 4**: Inferring the number of COVID-19 cases from recently reported deaths (Jombart et al., 2020);
- **Esperto 5**: Coronavirus Disease (COVID-19)–Statistics and Research (Roser et al., 2020).

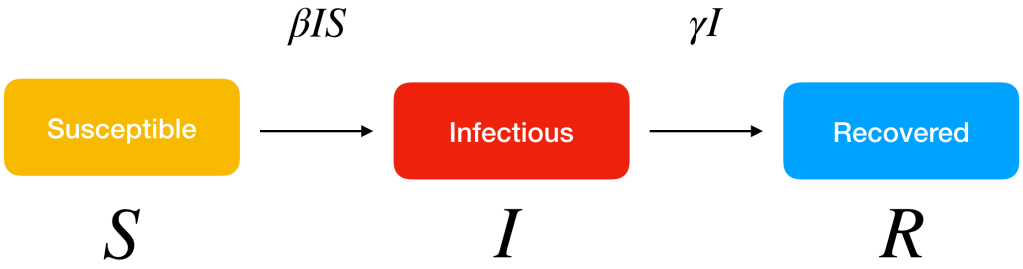
Facciamo un passo indietro...

Il dato che osserviamo è "conseguenza" di un fattore latente che è, in prima istanza, il coronavirus.



Ripartiamo non dal dato, ma dal modello che descrive il fenomeno!

Una possibile soluzione è usare un modello **SIR**, modello che divide la popolazione rispetto al virus e regola il passaggio tra gli stati.



Modello SIR

Il modello SIR può essere risolto con un approccio:

- stocastico (probabilistico) a tempi discreti;
- modello deterministico a tempi continui.

Nel modello SIR deterministico il passaggio tra stati è determinato da questa serie di equazioni differenziali:

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = \beta SI - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases}$$

Una misura di infezione - l'indice R_0

Al modello SIR è collegato un indice, detto R_0, che l'indice base di riproducibilità e controlla la trasmissione della malattia e viene calcolato dal rapporto:

$$R_0 = \frac{\beta}{\gamma}$$

Mi dice quanti nuovi infetti genera un infetto. Con R_0 <1 l'epidemia è in controllo. Dai casi osservati e ipotizzando una crescita esponenziale può essere stimato partendo dall'equazione precedente così:

$$I(t) \approx I(t-1)\, e^{(R_0-1)\,(\gamma)\,t}$$

Passando ai logaritmi ottengo

$$\log I(t) \approx \log I(t-1) + (R_0-1)\,(\gamma)\,t,$$

che può essere visto come un modello di regressione lineare

$$y(t) \approx \alpha + \beta_1 t$$

dove $\alpha = \log I(t-1)$ e $\beta_1 = (R_0-1)\,(\gamma)$

Una misura di infezione - l'indice R_0

Ripartiamo da

$$\log I(t) \approx \log I(t-1) + (R_0 - 1)(\gamma)\, t,$$

e ricordandoci di

$$\beta_1 = (R_0 - 1)(\gamma)$$

In sostanza una stima di R_0 può essere ottenuta da

$$\hat{R}_0 = 1 + \frac{\hat{\beta}_1}{\gamma}$$

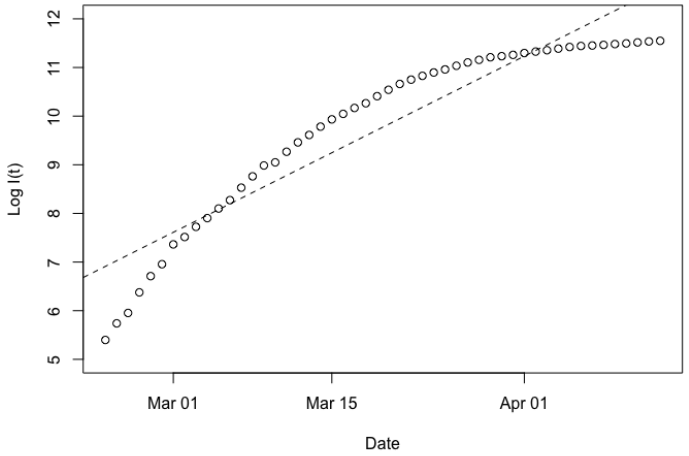
dove γ può essere fissato da precedenti pubblicazioni in 1/18 (Wang et al., 2020), mentre β_1 viene stimato con il metodo dei minimi quadrati (o altri metodi).

Una misura di infezione - l'indice R_0

	Stima	Errore Standard
α	6.79^{***}	(0.18)
β_1	0.12^{***}	(0.01)
R^2 0.88	RMSE (Num. obs.)	0.64 (50)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

$\hat{R}_0 = 1 + \frac{0.12}{1/18}$=3.16 [CI 95%: 2.98-3.34]. Ma non è soddisfacente!

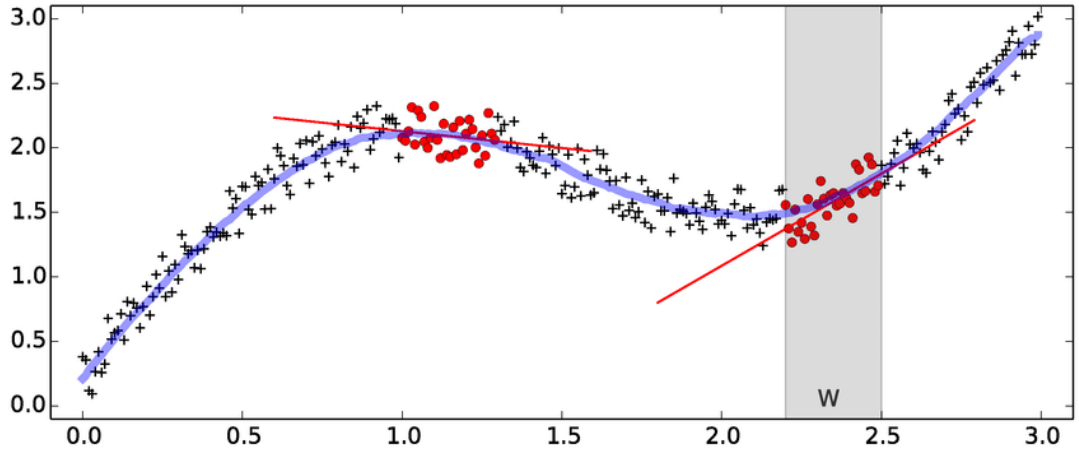


Una misura di infezione - l'indice R_0

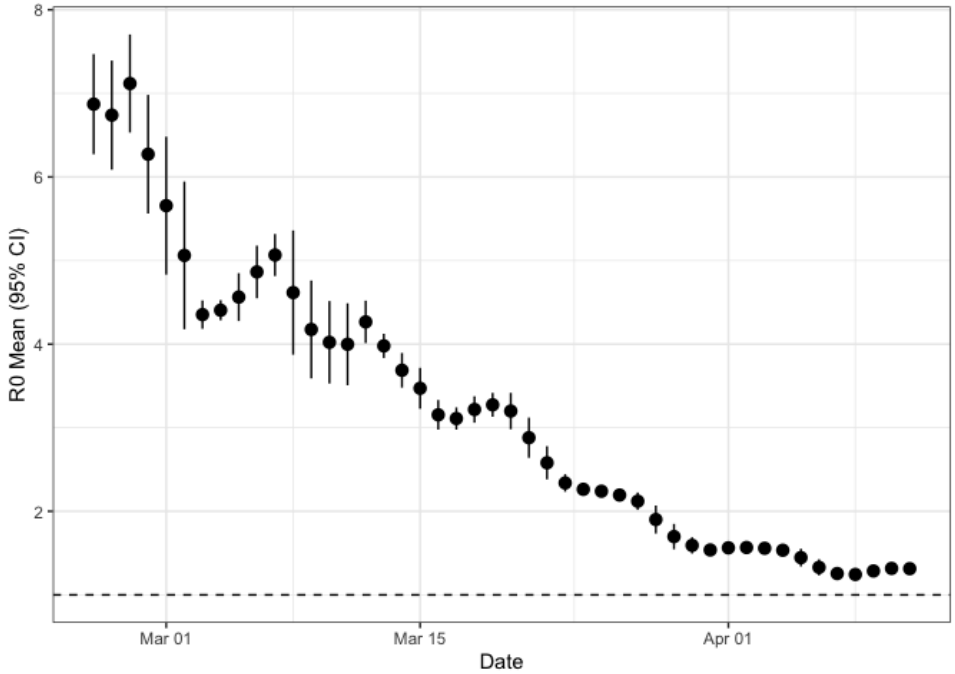
Questo risultato è dovuto al fatto che in una epidemia R_0 però non è costante, varia nel tempo! Ipotizziamo la seguente equazione

$$\hat{R}_0(t) = 1 + \frac{\hat{\beta}_1(t)}{\gamma}$$

dove $\beta_1(t)$ è una pendenza che può essere calcolata con diversi metodi. Con un approccio "molto flessibile" stimiamo $\beta_1(t)$ con una regressione locale sulla base di una finestra mobile di 5 giorni (da t-2 a t+2) ed otteniamo la seguente stima di $R_0(t)$.



Una misura di infezione - l'indice R_0



I valori partono da valori molto elevati, per assestarsi a valori di poco sopra l'unità. L'epidemia è quasi "sotto-controllo".

Previsione con l'indice R_0

La maggior parte di voi si sta chiedendo... quale sarà l'evoluzione? quanti casi avremo domani?

Per fare previsione ci basiamo sull'indice $R_0(t)$ o meglio su $\beta(t)$ che è il tasso di trasmissione (ricordo $R_0 = \frac{\beta}{\gamma}$).

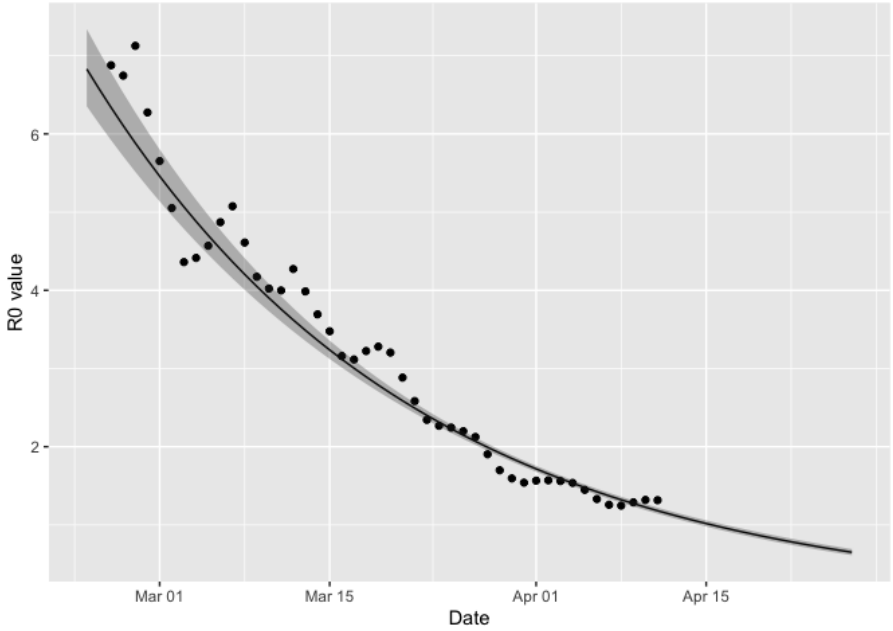
Supponiamo per $\beta(t)$ una distribuzione log-normale così ottenuta

$$\beta_{LogN}(t) \sim LogN(\beta_1(t) + \frac{1}{\gamma}, \sigma^2_{\beta_1(t)})$$

dato che può assumere solo valori positivi.

Previsione con l'indice R_0

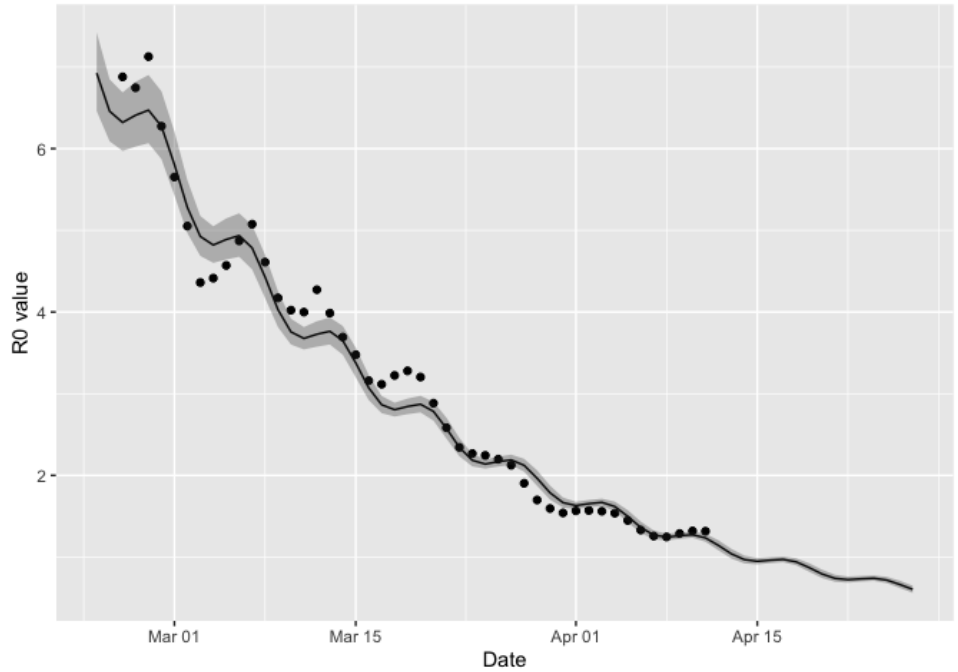
Tralasciando parte (noiosa, ma importante) di notazione per passare da una distribuzione log-normale ad una distribuzione normale, ottengo la seguente previsione con intervallo di confidenza al 95%.



Tuttavia mi sembra presente una stagionalità.
Proviamo a considerarla nell'analisi.

Previsione con l'indice R_0

Il seguente fitting viene ottenuto aggiungendo una stagionalità periodica a ritardo 7 con un modulation model (Eilers et al., 2008).



Sembra che il fitting sia migliorato.
Ma si può ancora migliorare con altri modelli piu' complessi!

Previsione - Modello SIR

Il modello SIR con parametro $\beta(t)$ dinamico è così formulato

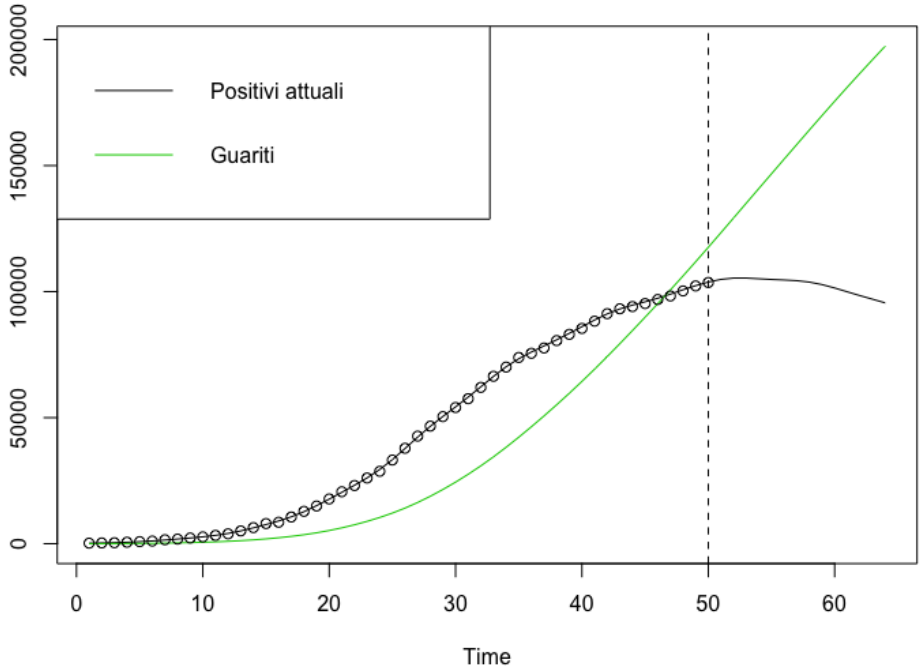
$$\frac{d}{dt}S(t) = -\beta(t)\frac{S(t)I(t)}{N}$$
$$\frac{d}{dt}I(t) = \beta(t)\frac{S(t)I(t)}{N} - \gamma I(t)$$
$$\frac{d}{dt}R(t) = \gamma\, I(t)$$

Devo definire gli stati iniziali e i parametri.

Stati iniziali	**Parametri**
I(0)=155 Infetti al giorno 0 R(0)=0 guariti S(0)=N-I(0)-R(0) N=60,317,000 (Istat 2019)	$\beta(t) = \hat{R}_0(t) * \gamma$ γ=1/18 vedi Wang et al. (2020)

Con questi parametri il modello è completamente identificabile.

Modello SIR stimato e previsione a 14 giorni



Il picco di casi positivi è previsto nella prossima settimana. Tuttavia i **guariti** attuali (circa 90.000) stimati dal modello sono molti di più del dato fornito dalla protezione civile (35.000).

Piccola riflessione

- L'analisi considerata parte da una dinamica di popolazione, non dai dati.
- L'analisi SIR è l'analisi più semplice, ma modelli più complessi possono essere considerati (SEIR, SEIRD, etc...).
- Stimare un R_0 tempo dipendente permette di verificare quando le misure adottate hanno cominciato a dare il loro effetto.
- Non va bene per previsioni a lungo termine (l'incubazione media del coronavirus è di 5.2 giorni (Wang et al., 2020), quindi variazioni sulle restrizioni avrebbero un effetto veloce).
- Il modello SIR adottato è deterministico... ma per fare inferenza è preferibile usare un modello stocastico.

Grazie per l'attenzione

Prossimo incontro: ultima settimana di aprile. Ulteriori informazioni saranno inviate tramite la mailing list
(per iscriversi scrivere a psicostat.dpss@unipd.it)

References

Agosto, A. and Giudici, P. (2020). A poisson autoregressive model to understand covid-19 contagion dynamics. *Available at SSRN 3551626*.

Dehesh, T., Mardani-Fard, H., and Dehesh, P. (2020). Forecasting of covid-19 confirmed cases in different countries with arima models. *medRxiv*.

Eilers, P. H., Gampe, J., Marx, B. D., and Rau, R. (2008). Modulation models for seasonal time series and incidence tables. *Statistics in Medicine*, 27(17):3430–3441.

Jombart, T., van Zandvoort, K., Russell, T., Jarvis, C., Gimma, A., Abbott, S., Clifford, S., Funk, S., Gibbs, H., Liu, Y., et al. (2020). Inferring the number of covid-19 cases from recently reported deaths. *medRxiv*.

Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J., Yan, P., and Chowell, G. (2020). Real-time forecasts of the covid-19 epidemic in china from february 5th to february 24th, 2020. *Infectious Disease Modelling*, 5:256–263.

Roser, M., Ritchie, H., and Ortiz-Ospina, E. (2020). Coronavirus disease (covid-19)–statistics and research. *Our World in Data*.

Wang, H., Wang, Z., Dong, Y., Chang, R., Xu, C., Yu, X., Zhang, S., Tsamlag, L., Shang, M., Huang, J.,