Understanding Evolutionary Mutations of SARS-CoV-2

STAT1005 Group 4.2

OUTLINE

01. Introduction

02. General view of mutation

03. Object 1: Specific investigation of mutation

04. Object 2: Comparison of mutation between time and location

05. Object 3: Relationship between external factors and mutation

06. Object 4: Prediction of mutation rate in China

07. Conclusion and Prospect

05

Object 4:

Relationship between death rate and mutation

Find the effect of mutation rate and death rate

Data & Goals

Data Used

- Virus genome sequences with the information of location and time

- Death rate of top 5 countries

Goals

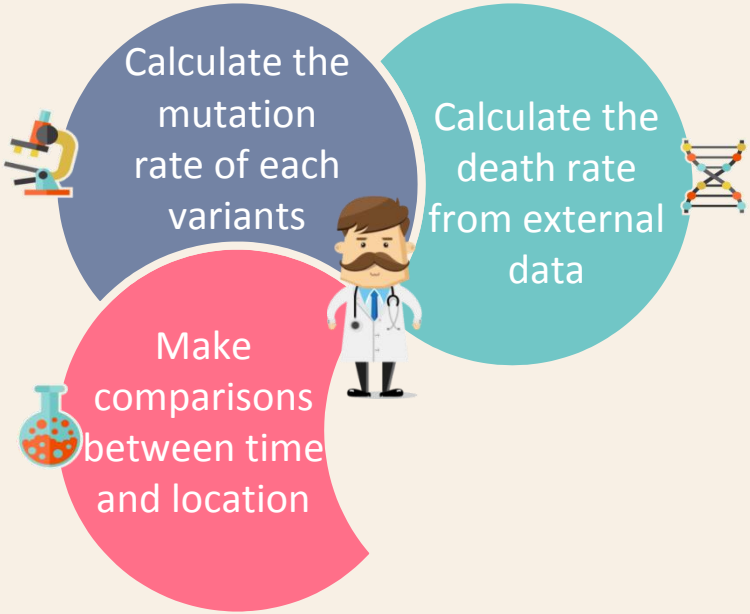
Find the correlation of mutation rate and death rate among countries

Find the correlation globally

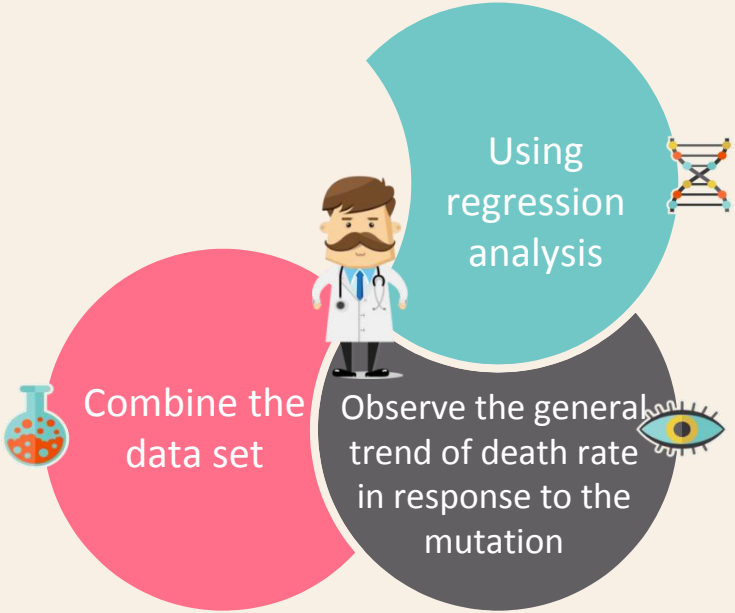
Find the effect of mutation rate and death rate

Methodology

Find the correlation of mutation rate and death rate among countries



Find the correlation globally



Computation of mutation rate

```
lst2=[]
noise_ignore= [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,\
               21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38,\
               39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 29839, 29840, 29841, 29842,\
               29843, 29844, 29845, 29846, 29847, 29848, 29849, 29850, 29851, 29852,\
               29853, 29854, 29855, 29856, 29857, 29858, 29859, 29860, 29861, 29862,\
               29863, 29864, 29865, 29866, 29867, 29868, 29869, 29870, 29871, 29872,\
               29873, 29874, 29875, 29876, 29877, 29878, 29879, 29880, 29881, 29882,\
               29883, 29884, 29885, 29886, 29887, 29888, 29889, 29890, 29891, 29892,\
               29893, 29894, 29895, 29896, 29897, 29898, 29899, 29900, 29901, 29902,\
               29903, 25505, 29734, 29837]
for i in data["Nucleotides"]: #sample
            for j in range(29903):
                if j != noise_ignore:
                    if (i[j] == "-"):
                        noisenumber +=1
                        continue
```

Noise filtering

```
individual_mutationrate = diff/(29903)*100000
```

Transformation of data

```
for i in countries_dict:
    data_dr[i]=data_dr[i].values/countries_dict[i]*2*100000 #for a better process of data
```

Computation of mutation rate

	Virus Strain Name	Virus Strain ID	Sample Collection Date	Location	Nucleotides	Individual mutationrate (%x100000)
0	hCoV-19/Wuhan/IPBCAMS-WH-05/2020	EPI_ISL_403928	2020-01-01	China, Wuhan	attaaaggtttataccttcccaggtaacaaaccaaccaactttcga...	3.344146
1	hCoV-19/Wuhan/IVDC-HB-04/2020	EPI_ISL_402120	2020-01-01	China, Wuhan	attaaaggtttataccttcccaggtaacaaaccaaccaactttcga...	10.032438
2	hCoV-19/env/Wuhan/IVDC-HBF13-21/2020	EPI			tcccaggtaacaaaccaaccaactttcga...	20.064876
3	hCoV-19/env/Wuhan/IVDC-HBF13-20/2020	EPI			tcccaggtaacaaaccaaccaactttcga...	26.753169
4	hCoV-19/Wuhan/WH03/2020	EPI			----caggtaacaaaccaaccaactttcga...	26.753169
...	...				...	...
1495	hCoV-19/Bangladesh/CHRF_0029/2020	EPI			--ccaggtaacaaaccaaccaactttcga...	77968.765676
1496	hCoV-19/India/InStem_NCBS_0038/2020	EPI_ISL_477239	2020-06-17	India	---agacgtgtgctcttccgatcttaacaaaccaaccaactttcga...	78142.661271
1497	hCoV-19/USA/MO-WUSTL032/2020	EPI_ISL_476021	2020-06-19	USA	----------ataccttcccaggtaacaaaccaaccaactttcga...	78169.414440
1498	hCoV-19/USA/MO-WUSTL070/2020	EPI_ISL_476018	2020-06-19	USA	----------accttcccaggtaacaaaccaaccaactttcga...	78206.200047

```
USA                447
China              294
India              135
United Kingdom     133
Bangladesh          58
Name: Location, dtype: int64
```

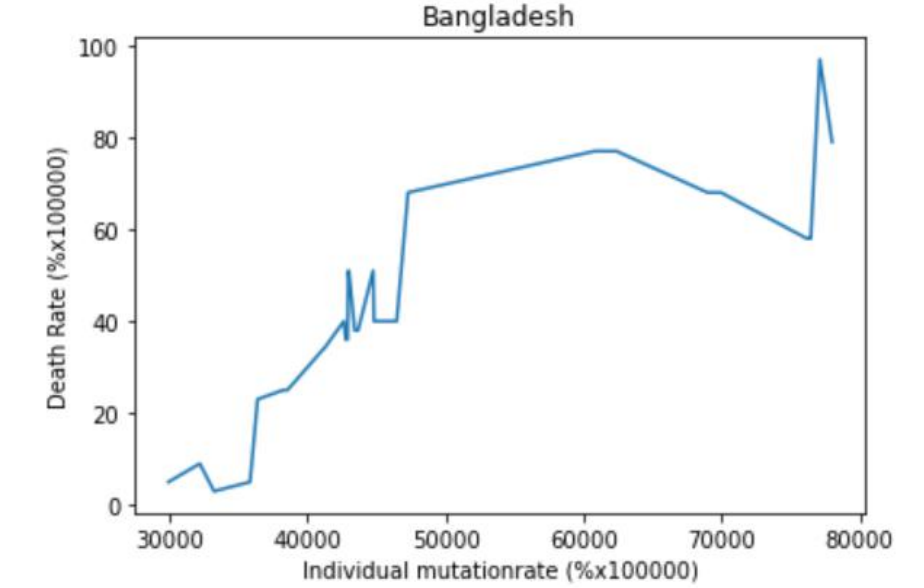
Combine data

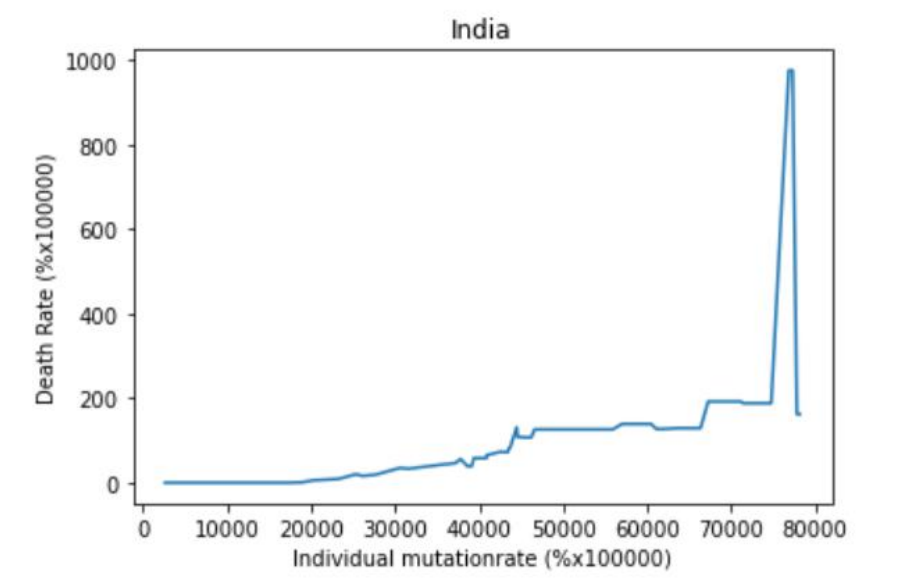
Combined data of Bangladesh

	Individual mutationrate (%x100000)	Sample Collection Date	Death rate (%x100000)
0	29900.010032	2020-04-28	5
1	32177.373508	2020-05-02	9
2	33207.370498	2020-05-03	3
3	35819.148580	2020-05-06	5
4	36364.244390	2020-05-07	23
5	38210.213022	2020-05-10	25
6	38297.160820	2020-05-10	25
7	38333.946427	2020-05-10	25
8	38487.777146	2020-05-10	25
9	38524.562753	2020-05-10	25

Types of correlation

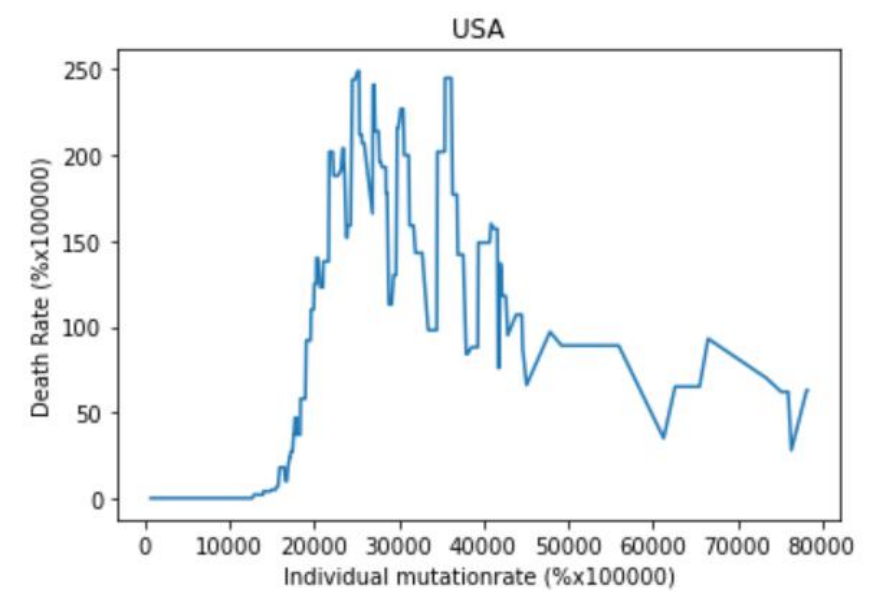
Increasing type

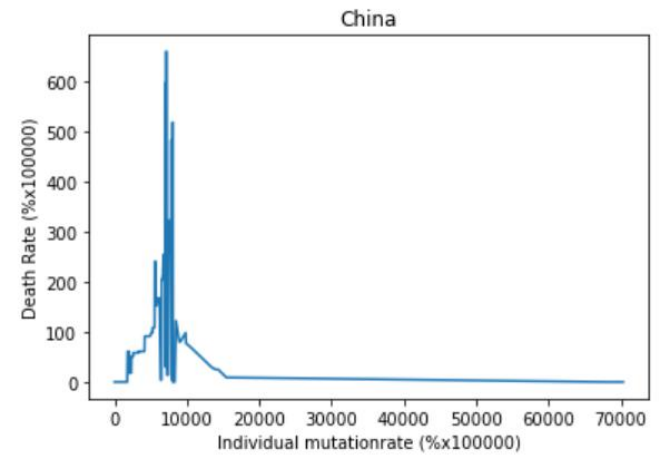


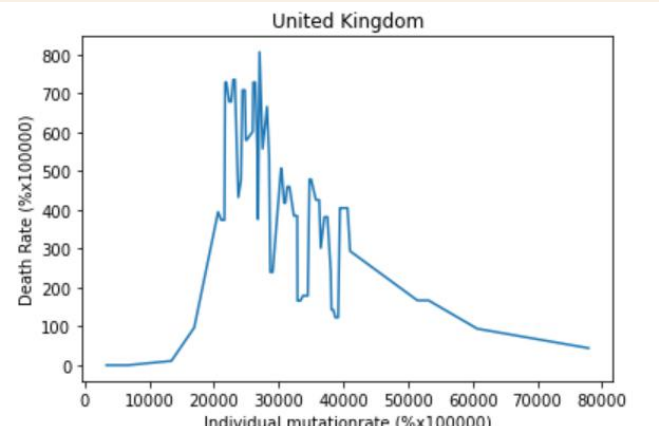


Types of correlation

Recovery type







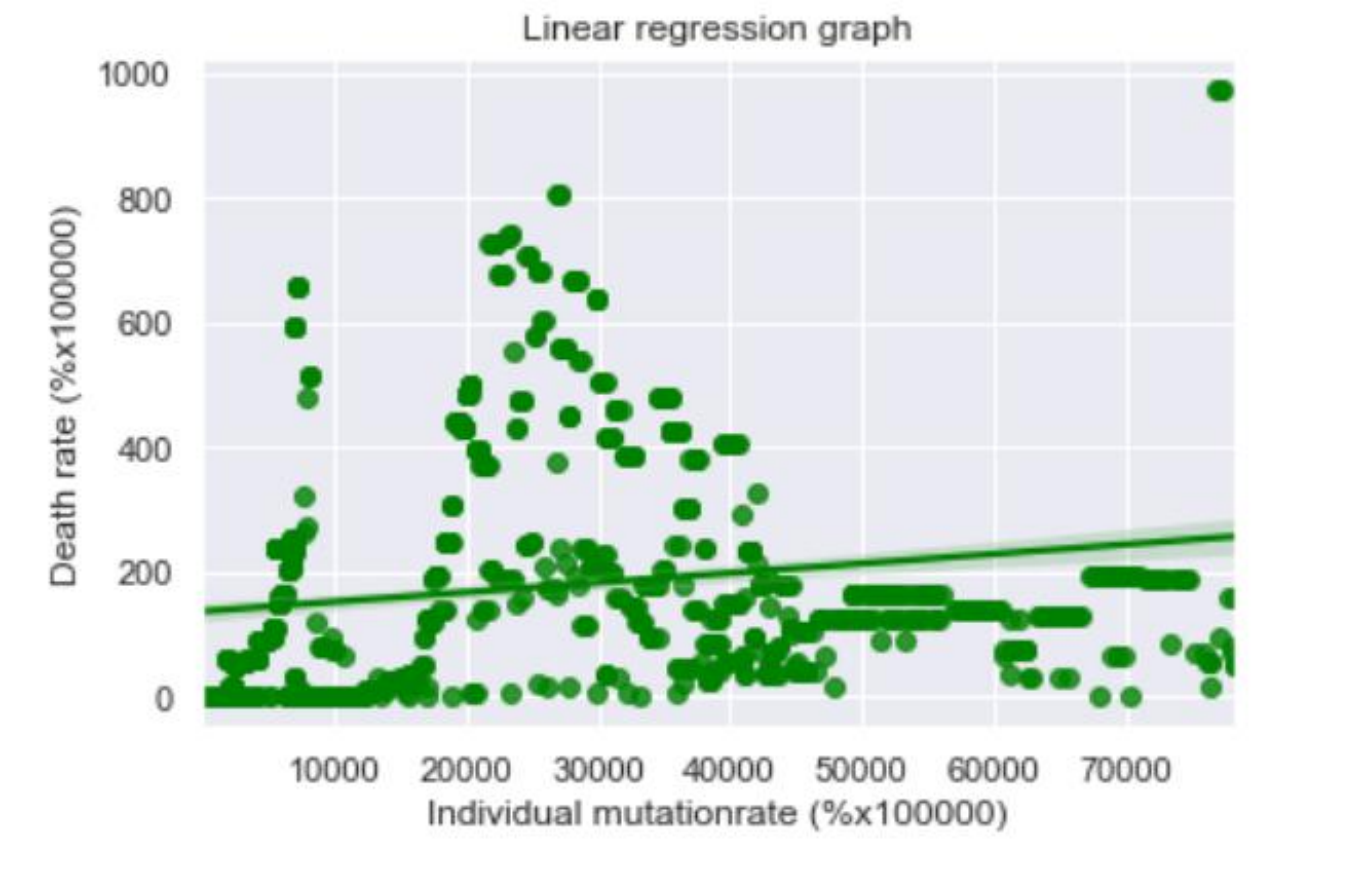
Globally analysis

All_combine

	Individual mutationrate (%x100000)	Sample Collection Date	Death rate (%x100000)
0	29900.010032	2020-04-28	5
1	32177.373508	2020-05-02	9
2	33207.370498	2020-05-03	3
3	35819.148580	2020-05-06	5
4	36364.244390	2020-05-07	23
...	...	...	...
443	75982.342909	2020-06-13	70
444	76360.231415	2020-06-14	17
445	78169.414440	2020-06-19	50
446	78206.200047	2020-06-19	50
447	78246.329800	2020-06-19	50

1022 rows × 3 columns

Linear Regression



Linear Regression

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.372
Model:	OLS	Adj. R-squared (uncentered):	0.372
Method:	Least Squares	F-statistic:	605.2
Date:	Fri, 26 Nov 2021	Prob (F-statistic):	2.63e-105
Time:	14:52:43	Log-Likelihood:	-11840.
No. Observations:	1022	AIC:	2.368e+04
Df Residuals:	1021	BIC:	2.369e+04
Df Model:	1		
Covariance Type:	nonrobust		

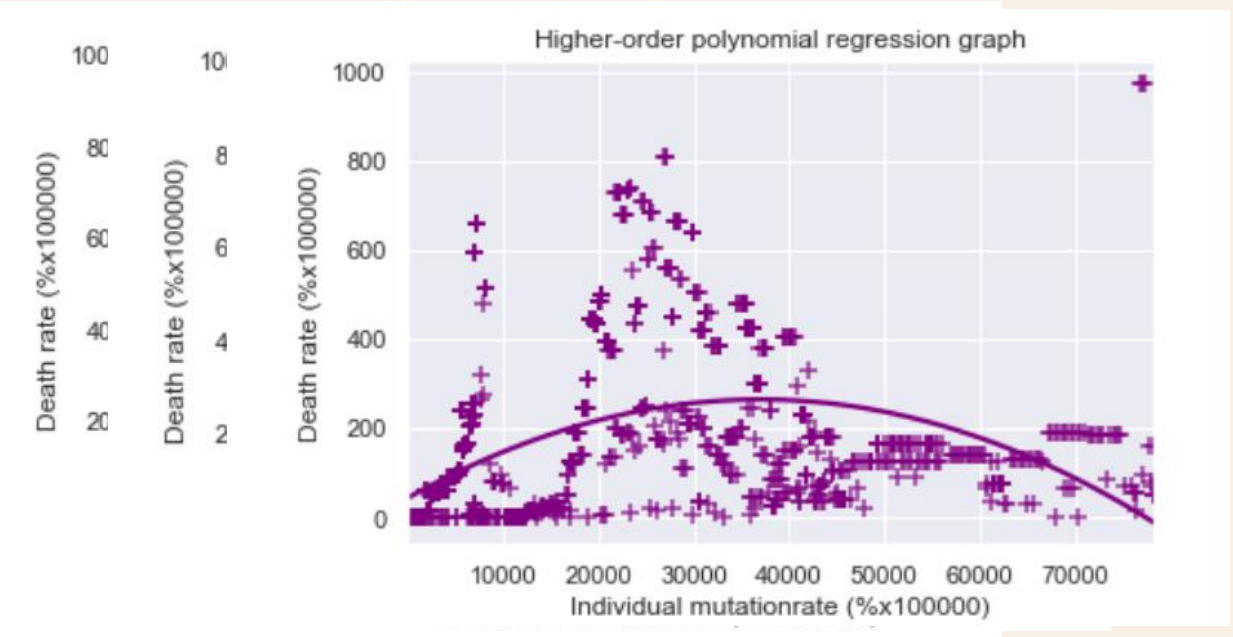
	coef	std err	t	P>|t|	[0.025	0.975]
Death rate (%x100000)	77.7003	3.159	24.600	0.000	71.502	83.898

Omnibus:	32.739	Durbin-Watson:	0.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35.265
Skew:	0.449	Prob(JB):	2.20e-08
Kurtosis:	2.851	Cond. No.	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

More Models Application

Higher-order polynomial regression



Summary

Find the effect of mutation rate and death rate

Individual categorization: China USA UK

01

AS we can observe, these 3 countries has similar pattern of correlation between death rate and mutation rate. As the virus mutate, the death rate will reach to a peak then return to a lower level, which indicate that while the virus mutate, their fitness or harmfulness drop

Individual categorization: India Bangladesh

02

There are similar pattern between these 2 countries. As they death rate keep rising while the virus in their countries mutate. It indicate that the viruses keep on mutating in a trend to obtain a higher fitness which is more harmful to human beings