POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI

Predykcja liczby dziennych zakażeń na podstawie danych o COVID-19 portalu Our World In Data

Sprawozdanie z laboratorium

Autor:
Krzysztof Wróblewski
nr albumu: **260394**
kierunek: **Informatyka Stosowana**
12 czerwca 2022

Streszczenie

Praca przedstawia wykonanie predykcji liczby nowych zakażeń przy użyciu modelu funkcji utworzonego dzięki regresji liniowej dla wielu zmiennych.. Dane na temat COVID-19 w formacie .csv zostały pobrane z repozytorium Our World In Data na GitHubie (pod linkiem). Dataset następnie został oczyszczony względem przedziału czasowego oraz stopnia korelacji względem liczby nowych zakażeń. Po tym procesie został stworzony model liniowy metodą najmniejszych kwadratów. Model nie jest wysoce optymalny ale pozwala na sprawdzenie zależności liniowych pomiędzy liczbą nowych zakażeń, a pozostałymi danymi.

1. Wstęp

Autor potrzebuje przetestować skuteczność predykcji nowych zakażeń na względnie prostym modelu liniowym, oraz poznać wartości estymowanych danych .

2. Opis danych

a. Pierwotny zestaw danych:

Wielkość danych: 181224 wierszy, 67 kolumn.

Gdzie:

r – współczynnik korelacji Pearsona względem nowych przypadków zachorowań (*new_cases*)

Przedział wartości = X – dane nienumeryczne nie biorące udział w tworzeniu modelu

*_smoothed – wygładzonie danych za pomocą średniej ruchomej z 7 dni

Index	Nazwa	Typ	Opis	Przedział wartości	Liczba wystąpień
0	iso_code	string	ISO 3166-1 alpha-3 – trzyliterowy kod kraju	X	181224
1	continent	string	Kontynent lokacji geograficznej	X	170631
2	location	string	Lokacja geograficzna	X	181224
3	date	string	Data obserwacji	Od 01.01.2020 do 23.04.2022	181224
4	total_cases	float64	Suma potwierdzonych (również prawdopodobnych) przypadkow zachorowań na Covid-19	[1.00, 509195070.00]	174465
5	new_cases	float64	Nowe potwierdzone (również prawdopodobnych) przypadki zachorowań na Covid-19	[0.00, 4089061.00]	174261
6	new_cases_smoothed	float64	Nowe potwierdzone przypadki (wygładzone)	[0.00, 3437004.57]	173092
7	total_deaths	float64	Suma śmierci (prawdopodobnie) spowodowanych przez Covid-19	[1.00, 6217046.00]	156248
8	new_deaths	float64	Nowe śmierci (prawdopodobnie) spowodowane przez Covid-19	[0.00, 18144.00]	156259
9	new_deaths_smoothed	float64	Nowe śmierci (wygładzone)	[0.00, 14783.29]	155112
10	total_cases_per_million	float64	Liczba wszystkich zakażeń, przypadająca na milion osób	[0.00, 706541.90]	173658
11	new_cases_per_million	float64	Liczba nowych zakażeń, przypadająca na milion osób	[0.00, 51427.49]	173454
12	new_cases_smoothed_per_million	float64	Liczba nowych (wygładzonych) zakażeń,	[0.00, 16052.61]	172290

			przypadająca na milion osób		
13	total_deaths_per_million	float64	Liczba wszystkich śmierci przez Covid-19 na milion osób	[0.00, 6376.73]	155454
14	new_deaths_per_million	float64	Liczba nowych śmierci przez Covid-19 na milion osób	[0.00, 550.40]	155465
15	new_deaths_smoothed_per_million	float64	Liczba nowych (wygładzonych) śmierci przez Covid-19 na milion osób	[0.00, 144.17]	154323
16	reproduction_rate	float64	Zakładany stopień rozprzestrzeniania się wirusa Covid-19	[-0.03, 6.15]	135364
17	icu_patients	float64	Liczba pacjentów na intensywnej terapii z Covid-19	[0.00, 28891.00]	24850
18	icu_patients_per_million	float64	Liczba pacjentów na intensywnej terapii z Covid-19 na milion osób	[0.00, 177.28]	24850
19	hosp_patients	float64	Liczba pacjentów hospitalizowanych z Covid-19	[0.00, 154540.00]	25571
20	hosp_patients_per_million	float64	Liczba pacjentów hospitalizowanych z Covid-19 na milion osób	[0.00, 1544.08]	25571
21	weekly_icu_admissions	float64	Tygodniowa nowa liczba pacjentów na intensywnej terapii z Covid-19	[0.00, 4838.00]	5859
22	weekly_icu_admissions_per_million	float64	Tygodniowa nowa liczba pacjentów na intensywnej terapii z Covid-19 na milion osób	[0.00, 221.21]	5859
23	weekly_hosp_admissions	float64	Tygodniowa nowa liczba pacjentów hospitalizowanych z Covid-19	[0.00, 153995.00]	11733
24	weekly_hosp_admissions_per_million	float64	Tygodniowa nowa liczba pacjentów hospitalizowanych z Covid-19 na milion osób	[0.00, 645.81]	11733
25	total_tests	float64	Liczba wszystkich testów wykonanych na Covid-19	[0.00, 862700218.00]	75084
26	new_tests	float64	Liczba nowych testów wykonanych na Covid-19	[1.00, 35855632.00]	72421
27	total_tests_per_thousand	float64	Liczba wszystkich testów wykonanych na Covid-19 na 1000 osób	[0.00, 32925.90]	75084
28	new_tests_per_thousand	float64	Liczba nowych testów wykonanych na Covid-19 na 1000 osób	[0.00, 534.01]	72421
29	new_tests_smoothed	float64	Liczba nowych testów wykonanych na Covid-19 (wygładzone)	[0.00, 5471529.00]	94292
30	new_tests_smoothed_per_thousand	float64	Liczba nowych testów wykonanych na Covid-19 na 1000 osób (wygładzone)	[0.00, 147.60]	94292
31	positive_rate	float64	Odsetek testów pozytywnych z 7 dni	[0.00, 1.00]	91195
32	tests_per_case	float64	Ilość testów przypadających na ilość zakażeń z 7 dni	[1.00, 199914.90]	89945
33	tests_units	string	Jednostki testowania	**people tested, samples tested, tests performer, units unclear**	96981
34	total_vaccinations	float64	Całkowita liczba zarejestrowanych szczepień przeciwko Covid-19	[0.00, 11517071804.00]	49115
35	people_vaccinated	float64	Całkowita liczba osób, którzy przyjęli co najmniej 1 dawkę szczepionki	[0.00, 5125207873.00]	46718
36	people_fully_vaccinated	float64	Całkowita liczba osób zaszczepionych wszystkimi dawkami szczepionki	[1.00, 4632178666.00]	44180
37	total_boosters	float64	Całkowita ilość dawek boosterów	[1.00, 1821801574.00]	21548
38	new_vaccinations	float64	Liczba nowych szczepień	[0.00, 54503105.00]	40359
39	new_vaccinations_smoothed	float64	Liczba nowych szczepień (wygładzona)	[0.00, 43545621.00]	94709
40	total_vaccinations_per_hundred	float64	Liczba wszystkich szczepień na 100 osób	[0.00, 354.93]	49115

41	people_vaccinated_per_hundred	float64	Całkowita liczba osób, którzy przyjęli co najmniej 1 dawkę szczepionki na 100 osób	[0.00, 124.87]	46718
42	people_fully_vaccinated_per_hundred	float64	Całkowita liczba osób zaszczepionych wszystkimi dawkami szczepionki	[0.00, 122.88]	44180
43	total_boosters_per_hundred	float64	Całkowita ilość dawek boosterów na 100 osób	[0.00, 107.17]	21548
44	new_vaccinations_smoothed_per_million	float64	Liczba nowych szczepień (wygładzona) na milion osób	[0.00, 117497.00]	94709
45	new_people_vaccinated_smoothed	float64	Liczba nowych osób przyjmujących 1 dawkę szczepienia (wygładzona)	[0.00, 21367752.00]	93713
46	new_people_vaccinated_smoothed_per_hundred	float64	Liczba nowych osób przyjmujących 1 dawkę szczepienia (wygładzona) na 100 osób	[0.00, 11.75]	93713
47	stringency_index	float64	Stopień reakcji rządu na stan pandemii (100 – najbardziej restrykcyjna)	[0.00, 100.00]	141297
48	population	float64	Liczba ludności	[47.00, 7874965730.00]	180100
49	population_density	float64	Gęstość zaludnienia	[0.14, 20546.77]	161469
50	median_age	float64	Mediana wieku społeczeństwa (predykcja ONZ na 2020)	[15.10, 48.20]	149626
51	aged_65_older	float64	Odsetek populacji powyżej 65 roku życia	[1.14, 27.05]	148034
52	aged_70_older	float64	Odsetek populacji powyżej 70 roku życia	[0.53, 18.49]	148838
53	gdp_per_capita	float64	Wskaźnik PKB	[661.24, 116935.60]	149276
54	extreme_poverty	float64	Odsetek ludności żyjącej w skrajnym ubóstwie	[0.10, 77.60]	97500
55	cardiovasc_death_rate	float64	Stopień śmierci na atak serca	[79.37, 724.42]	149308
56	diabetes_prevalence	float64	Odsetek cukrzyków w społeczeństwie w wieku 20-79 lat	[0.99, 30.53]	156607
57	female_smokers	float64	Odsetek palących kobiet	[0.10, 44.00]	113282
58	male_smokers	float64	Odsetek palących mężczyzn	[7.70, 78.10]	111729
59	handwashing_facilities	float64	Odsetek ludności z podstawowym dostępem do pomieszczeń z możliwością mycia rąk	[1.19, 100.00]	73582
60	hospital_beds_per_thousand	float64	Ilość łóżek szpitalnych przypadających na 1000 osób	[0.10, 13.80]	132435
61	life_expectancy	float64	Oczekiwana długość życia	[53.28, 86.75]	169457
62	human_development_index	float64	Indeks rozwoju człowieka	[0.39, 0.96]	145872
63	excess_mortality_cumulative_absolute	float64	Całkowita różnica pomiędzy liczbą rzeczywistych a przewidywanych śmierci, na podstawie danych od początku pandemii	[-37726.10, 1163660.50]	6229
64	excess_mortality_cumulative	float64	Procentowa różnica pomiędzy liczbą rzeczywistych a przewidywanych śmierci, na podstawie danych od początku pandemii	[-28.45, 111.01]	6229
65	excess_mortality	float64	Procentowa różnica pomiędzy tygodniową lub miesięczną ilością rzeczywistą oraz przwidywaną śmierci, na podstawie danych z zeszłego roku	[-95.92, 375.00]	6229
66	excess_mortality_cumulative_per_million	float64	Całkowita różnica pomiędzy liczbą rzeczywistych a przewidywanych śmierci, na podstawie danych od początku pandemii, przypadająca na milion osób	[-1826.60, 9620.68]	6229

3. Opis rozwiązania

Dane zostały pobrane z repozytorium na GitHubie, oraz została ona zapisana jako DataFrame biblioteki Pandas. Zawiera ona informację na temat 181224 dziennych przypadków Covid-19 na podstawie 67 zmiennych. Aby utworzyć liniowy model na podstawie metody najmniejszych kwadratów, potrzebujemy dane numeryczne. Proces przekształcania datasetu prezentuje się następująco:

- Ograniczenie danych do lat 2021 - 2022
- Wydzielenie danych nienumerycznych, niebiorących udziału w modelowaniu
- Zmiana reprezentacji danej *tests_units* na reprezentacje one-hot
- Wypełnienie pustych wartości danych zerami
- Analizę liniowego wpływu poszczególnych danych na wartość danej *new_cases* poprzez wartość współczynnika korelacji Pearsona
- Używając Variance Inflation Factor (VIF) wydzielenie zbędnych zmiennych niezależnych w zbiorze danych. VIF określa ilościowo stopień nasilenia wieloliniowości w normalnej analizie regresji najmniejszych kwadratów. Dla optymalizajci modelu, przy wyborze danych najlepiej jak dla każdej zmiennej zachodzi warunek:

$$VIF_{factor} < 10$$

- Odrzucenie danych, które mają mały wpływ na dzienną wartość nowych zakażeń po warunku:

$$|r| < 0{,}5$$

- Ustalenie części wspólnej dwóch poprzednich procesów

Ograniczyliśmy w ten sposób wielkość danych do modelu do 6 zmiennych i 113114 przypadków (prób)

Dane numeryczne ze współczynnikiem korelacji Pearsona (na zielono zaznaczono dane biorące udział w tworzeniu modelu):

Indeks	Nazwa	r
0	total_cases	0.81
1	new_cases	1.00
2	new_cases_smoothed	0.98
3	total_deaths	0.74
4	new_deaths	0.66
5	new_deaths_smoothed	0.65
6	total_cases_per_million	0.07
7	new_cases_per_million	0.08
8	new_cases_smoothed_per_million	0.09
9	total_deaths_per_million	0.08
10	new_deaths_per_million	0.03
11	new_deaths_smoothed_per_million	0.04
12	reproduction_rate	-0.07
13	icu_patients	0.09
14	icu_patients_per_million	0.01
15	hosp_patients	0.10
16	hosp_patients_per_million	0.00
17	weekly_icu_admissions	0.04

18	weekly_icu_admissions_per_million	0.01
19	weekly_hosp_admissions	0.11
20	weekly_hosp_admissions_per_million	0.05
21	total_tests	0.08
22	new_tests	0.08
23	total_tests_per_thousand	-0.01
24	new_tests_per_thousand	-0.01
25	new_tests_smoothed	0.09
26	new_tests_smoothed_per_thousand	-0.01
27	positive_rate	-0.02
28	tests_per_case	-0.02
29	total_vaccinations	0.65
30	people_vaccinated	0.64
31	people_fully_vaccinated	0.65
32	total_boosters	0.66
33	new_vaccinations	0.53
34	new_vaccinations_smoothed	0.52
35	total_vaccinations_per_hundred	0.19
36	people_vaccinated_per_hundred	0.18
37	people_fully_vaccinated_per_hundred	0.18
38	total_boosters_per_hundred	0.18
39	new_vaccinations_smoothed_per_million	0.03
40	new_people_vaccinated_smoothed	0.41
41	new_people_vaccinated_smoothed_per_hundred	-0.01
42	stringency_index	-0.14
43	population	0.59
44	population_density	-0.03
45	median_age	-0.09
46	aged_65_older	-0.04
47	aged_70_older	-0.04
48	gdp_per_capita	-0.04
49	extreme_poverty	-0.04
50	cardiovasc_death_rate	-0.12
51	diabetes_prevalence	-0.10
52	female_smokers	-0.02
53	male_smokers	-0.04
54	handwashing_facilities	-0.01
55	hospital_beds_per_thousand	-0.04
56	life_expectancy	-0.26
57	human_development_index	-0.10
58	excess_mortality_cumulative_absolute	0.01
59	excess_mortality_cumulative	-0.01
60	excess_mortality	-0.01
61	excess_mortality_cumulative_per_million	-0.01
62	tests_units_people tested	-0.02
63	tests_units_samples tested	-0.02
64	tests_units_tests performed	-0.08
65	tests_units_units unclear	-0.01

Na podstawie tych danych tworzymy model liniowy.

4. Rezultaty obliczeń

a. Plan badań

Na podstawie kryterium $|r| < 0{,}5$, oraz $VIF_{factor} < 10$ otrzymujemy ostateczny zbiór danych biorących udział w tworzeniu modelu:

Indeks	Nazwa	r	VIF_{factor}
0	new_cases_smoothed	0.98	3,396689
1	new_deaths	0.66	4,238341
2	people_fully_vaccinated	0.65	7,21206
3	total_boosters	0.66	4,943077
4	new_vaccinations	0.53	9,751024
5	population	0.59	7,059157

Po podzieleniu danych na testowe i treningowe w stosunku 1:5, za pomocą metody train_test_split biblioteki Sklearn. Ustawiamy stan losowy na 42. Tworzymy model za pomocą obiektu LinearRegression(fit_intercept=True) biblioteki Sklearn i używamy metody fit() do stworzenia modelu.

Wielkość próbki treningowej: 90491
Wielkość próbki testowej: 22623

b. Wyniki obliczeń

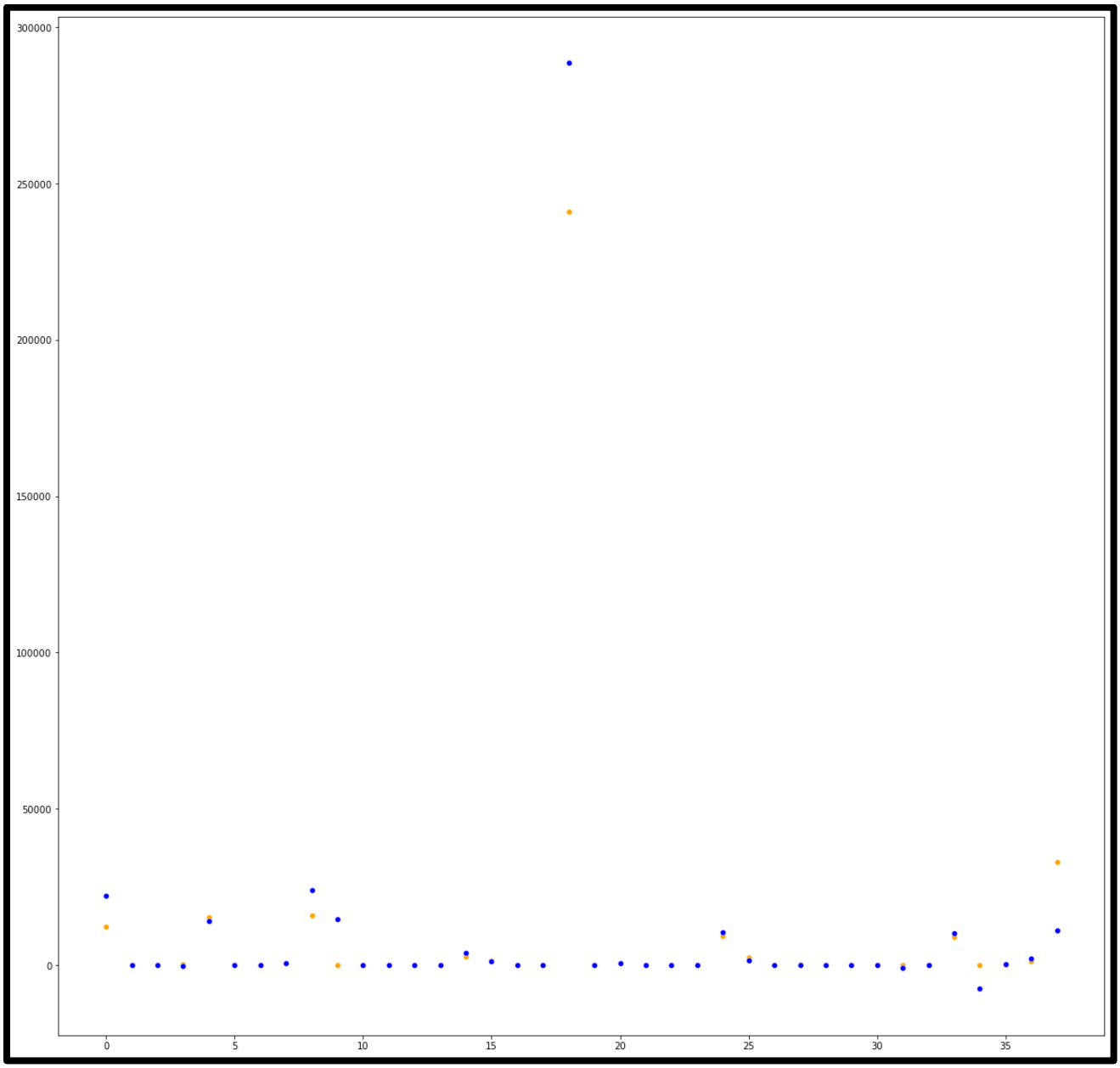
Otrzymany model liniowy prezentuje się następująco:

$$y = \sum_{i=0}^{n} a_i x_i + C$$

Gdzie (parametry podawane są z przybliżeniem do 6 miejsc po przecinku):
$n = 5$ – od 0 do n ilość zmiennych w modelu
$C = -32.426642$ – wyraz wolny w modelu

i	x_i	a_i
0	new_cases_smoothed	0.965225
1	new_deaths	9.375904
2	people_fully_vaccinated	0.000009
3	total_boosters	-0.000006
4	new_vaccinations	0.001723
5	population	-0.000014

c. Wykres i stopień dopasowania modelu do danych rzeczywistych



Opis:

Wykres prezentuje jedynie fragment porównań wartości estymowanej z rzeczywistą.

Oś x – ilość prób testowych
Oś y – ilość nowych zakażeń

- - rzeczywista liczba nowych zakażeń
- - estymowana liczba nowych zakażeń

Dane modelu:

MSE (błąd średniokwadratowy) = 466204305.605721

MEA (Średnie odchylenie bezwzględne) = 162.13890454042294

R^2 (Współczynnik determinacji) = 0.9598131627180571

5. Wnioski

Jednoznacznie można stwierdzić, że do tak skomplikowanych i złożonych danych model liniowy nie jest najbardziej optymalnym wyborem, widać to chociażby po ujemnych wartościach estymowanej dziennej liczby zakażeń przy niskich wartościach rzeczywistych, oraz przy wysokiej wartości średniego błędu kwadratowego. Wygładzone dane o nowych zakażeniach średnią ruchomą mają współczynnik praktycznie równy 1, co oznacza że dane do predykcji są praktycznie niezmienione i wartość estymacji jest silnie związana z danymi zakażeń z poprzednich dni. Gdy ilość dziennych śmierci rośnie, również rośnie dzienna ilość zakażeń. Jest to korelacja związana z okresem pandemicznym, gdzie w szczycie pandemii, dużej ilości zakażeń towarzyszyła duża liczba zgonów. Wysokim ilościom zakażeń towarzyszyła zwiększona ilość szczepień, a także widzimy, że duża wielkość populacji nieznacznie wpływała na zmniejszenie się liczby nowych zakażeń, tak samo jak całkowita ilość przyjętych boosterów.

6. Dodatek

Kody i pliki źródłowe dokonanych operacji, wraz z krótkimi opisami załączam wraz ze sprawozdaniem:

- covid_cases_regression_model.ipynb
- vif.xlsx
- owid-covid-data.csv