УДК 004.67

ПРОГНОЗИРОВАНИЕ РАСПРОСТРАНЕНИЯ COVID-19 НА ТЕРРИТОРИИ РЕСПУБЛИКИ БЕЛАРУСЬ С ПРИМЕНЕНИЕМ ЛИНЕЙНОЙ РЕГРЕССИИ

Ларькин А.Д.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Тонкович И.Н. – к.х.н, доцент, доцент кафедры ПИКС

Аннотация. Для прогнозирования заболеваемости и смертности от COVID-19 на территории Республики Беларусь использовалась модель линейной регрессии. Прогноз строился на основе статистических данных, предоставленных Всемирной Организации Здравоохранения.

Ключевые слова: COVID-19, прогнозирование распространения, линейная регрессия.

Введение. Целью данного исследования является разработка модели прогнозирования заболеваемости и смертности от COVID-19 в Республике Беларусь на основе линейной регрессии. Построение моделей прогнозирования – важный шаг в направлении развития методов анализа данных для более эффективного управления пандемией и защиты общественного здоровья.

Основная часть. Одним из самых простых методов прогнозирования является применение метода линейной регрессии. Линейная регрессия используется для прогнозирования непрерывных значений. Данный способ основан на построении линейной зависимости между входными признаками и целевой переменной [1]. Данная линейная зависимость представлена формулой 1:

$$y = a + b * x\,, \tag{1}$$

где y – объясняемая (или зависимая) переменная;

x – независимая переменная (регрессор);

a – коэффициент сдвига (интерсепт);

b – коэффициент наклона прямой линии регрессии.

Коэффициент a является значением зависимой переменной, когда независимая переменная равна нулю. Данный коэффициент вычисляется по формуле 2:

$$a = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \sum x_i \sum x_i}, \tag{2}$$

где n – количество периодов;

x_i – независимая переменная за конкретный период;

y_i – объясняемая переменная за конкретный период.

Коэффициент b показывает, насколько изменяется зависимая переменная при изменении независимой переменной на одну единицу. Данный коэффициент вычисляется по формуле 3:

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \sum x_i \sum x_i}, \tag{3}$$

где n – количество периодов;

x_i – независимая переменная за конкретный период;

y_i – объясняемая переменная за конкретный период.

Для прогноза использовались данные о новых случаях заболеваемости и смертности населения Республики Беларусь за период с 19 апреля 2020 года по 18 апреля 2021 года. Все данные брались из официальных источников, предоставленных Всемирной Организацией Здравоохранения (ВОЗ). В случае использования уравнения линейной регрессии, объясняемой переменной будут считаться данные о заболеваемости и смертности населения, в качестве независимой переменной будет браться номер периода (недели), в конце которой были получены сведения о заболеваемости и смертности [2]. Все данные за этот период и необходимые значения для расчета коэффициентов *a* и *b* представлены в таблице 1:

Таблица 1 – Статистические значения новых случаев заболеваемости и летального исхода и вспомогательные значения, необходимые для расчета коэффициентов a и b

Дата (последний день недели)	**Новые случаи летального исхода**	**Новые случаи заболеваемости**	**X^2(x – новые случаи заболеваемости)**	**X_iY_i (новые случаи заболеваемости)**	**X_iY_i (новые случаи летального исхода)**
19.04.2020	22	3063	1	3063	22
26.04.2020	22	4301	4	8602	44
03.05.2020	30	6238	9	18714	90
10.05.2020	34	7145	16	28580	136
17.05.2020	29	5708	25	28540	145
24.05.2020	34	6563	36	39378	204
31.05.2020	35	6414	49	44898	245
07.06.2020	30	5210	64	41680	240
14.06.2020	44	6373	81	57357	396
21.06.2020	40	4695	100	46950	400
28.06.2020	34	3159	121	34749	374
05.07.2020	41	2175	144	26100	492
12.07.2020	41	1494	169	19422	533
19.07.2020	36	1189	196	16646	504
26.07.2020	35	1049	225	15735	525
02.08.2020	33	944	256	15104	528
09.08.2020	22	792	289	13464	374
16.08.2020	22	686	324	12348	396
23.08.2020	30	861	361	16359	570
30.08.2020	34	1238	400	24760	680
06.09.2020	34	1140	441	23940	714
13.09.2020	39	1312	484	28864	858
20.09.2020	32	1486	529	34178	736
27.09.2020	37	1828	576	43872	888
04.10.2020	38	2563	625	64075	950
11.10.2020	40	3171	676	82446	1040
18.10.2020	34	4040	729	109080	918
25.10.2020	20	4104	784	114912	560
01.11.2020	40	8292	841	240468	1160
08.11.2020	19	5824	900	174720	570
15.11.2020	35	7587	961	235197	1085
22.11.2020	50	9565	1024	306080	1600
29.11.2020	54	10889	1089	359337	1782
06.12.2020	55	11955	1156	406470	1870
13.12.2020	56	13055	1225	456925	1960
20.12.2020	62	13245	1296	476820	2232

Продолжение таблицы 1

1	2	3	4	5	6
27.12.2020	60	13343	1369	493691	2220
03.01.2021	66	13203	1444	501714	2508
10.01.2021	65	12243	1521	477477	2535
17.01.2021	66	13169	1600	526760	2640
24.01.2021	66	12322	1681	505202	2706
31.01.2021	69	10711	1764	449862	2898
07.02.2021	65	10389	1849	446727	2795
14.02.2021	67	10070	1936	443080	2948
21.02.2021	63	9961	2025	448245	2835
28.02.2021	63	8969	2116	412574	2898
07.03.2021	63	8473	2209	398231	2961
14.03.2021	58	6896	2304	331008	2784
21.03.2021	61	7965	2401	390285	2989
28.03.2021	63	8338	2500	416900	3150
04.04.2021	65	8434	2601	430134	3315
11.04.2021	68	8798	2704	457496	3536
18.04.2021	69	8060	2809	427180	3657

Проведем прогнозирование новых случаев заболеваемости населения Республики Беларусь. Период (n) составляет 53 недели, значение коэффициента a = 2014,06, значение коэффициента b = 163,48. График линейной зависимости представлен на рисунке 1:

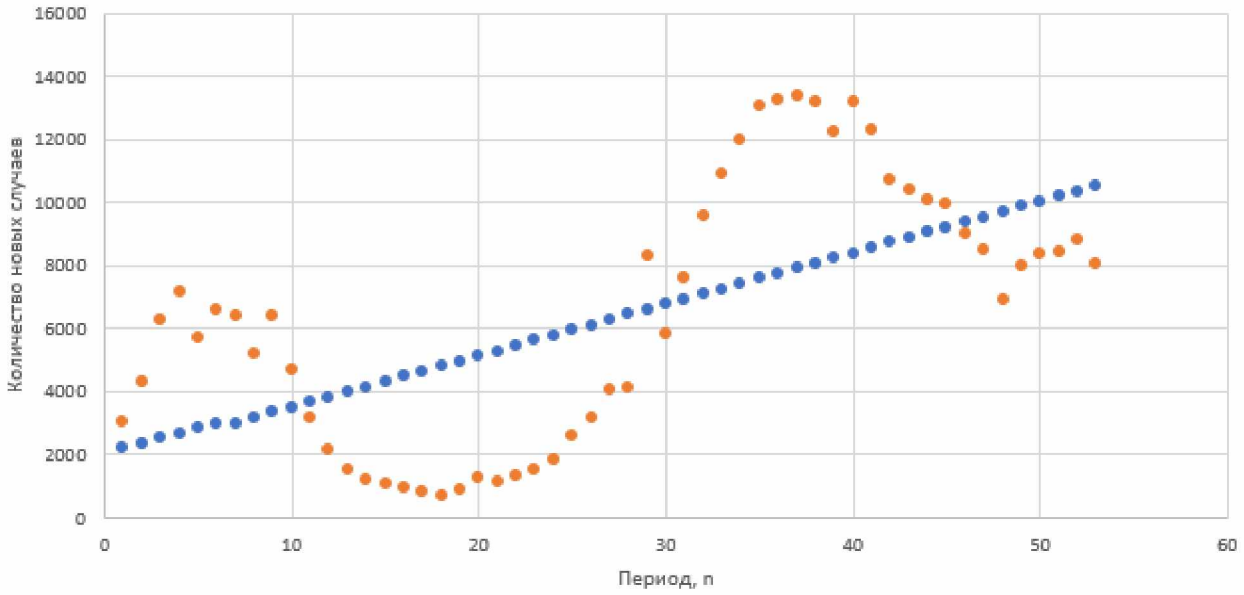

Рисунок 1 – Линейная регрессия новых случаев заболеваемости COVID-19 на территории Республики Беларусь (в период с 19.04.2020 по 18.04.2021 г.)

Далее проведем прогнозирование новых случаев летального исхода у населения Республики Беларусь. Период (n) также составляет 53 недели, значение коэффициента a = 21,87, значение коэффициента b = 0,86. График линейной зависимости представлен на рисунке 2:

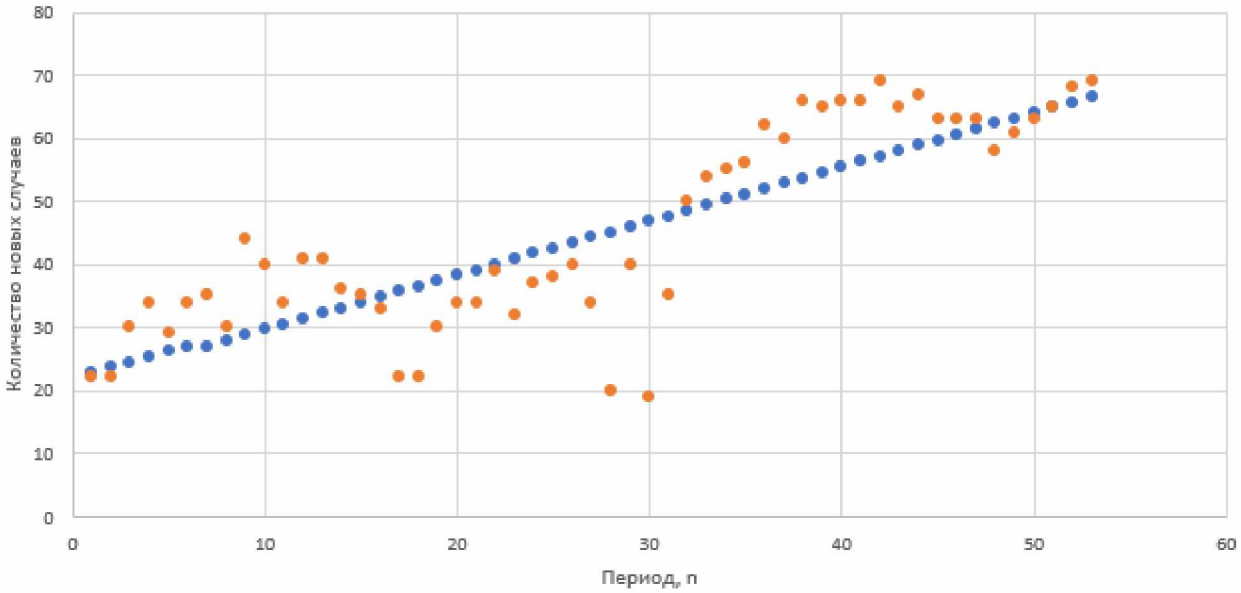

Рисунок 2 – Линейная регрессия новых случаев летального исхода от COVID-19 на территории Республики Беларусь (в период с 19.04.2020 по 18.04.2021 г.)

Как видно из графиков, линейная регрессия построена успешно. На следующем шаге выполнялось прогнозирование на основе модели линейной регрессии с использованием инструментария Microsoft Excel. С помощью встроенной функции linest мы сможем спрогнозировать значения на следующие 37 периодов (недель) и получим количество новых случаев заболеваемости и летального исхода на 02.01.2022 года. Данная дата была выбрана вследствие того, что самый пик заболеваемости и смертности населения приходится на зимний период, ввиду климатических условий на территории Республики Беларусь.

Функция linest принимает массив прогнозируемой величины (новых случаев заболеваемости и смертности), массив периодов и принимает коэффициенты a и b, рассчитанные выше. Полученные значения необходимых для прогноза величин представлены в таблице 2.

Таблица 2 – Значения, необходимые для анализа и прогноза заболеваемости и смертности населения от COVID-19 на территории Республики Беларусь

Новые случаи заболеваемости			**Новые случаи смертности**		
Угловой коэффициент	Свободный член	Коэффициент детерминации	Угловой коэффициент	Свободный член	Коэффициент детерминации
163,49	2014,06	0,379	0,86	21,87	0,69

Коэффициенты детерминации (особенно в случае с прогнозом новых случаев заболеваемости) крайне малы, отчего прогноз будет недостоверным. Это подтверждает тот факт, что модель линейной регрессии крайне неэффективна при ограниченном количестве параметров. Также линейная регрессия не предусматривает спада новых случаев заболевания и смертности среди населения. График прогноза новых случаев заболеваемости к моменту 02.01.2022 показан на рисунке 3:

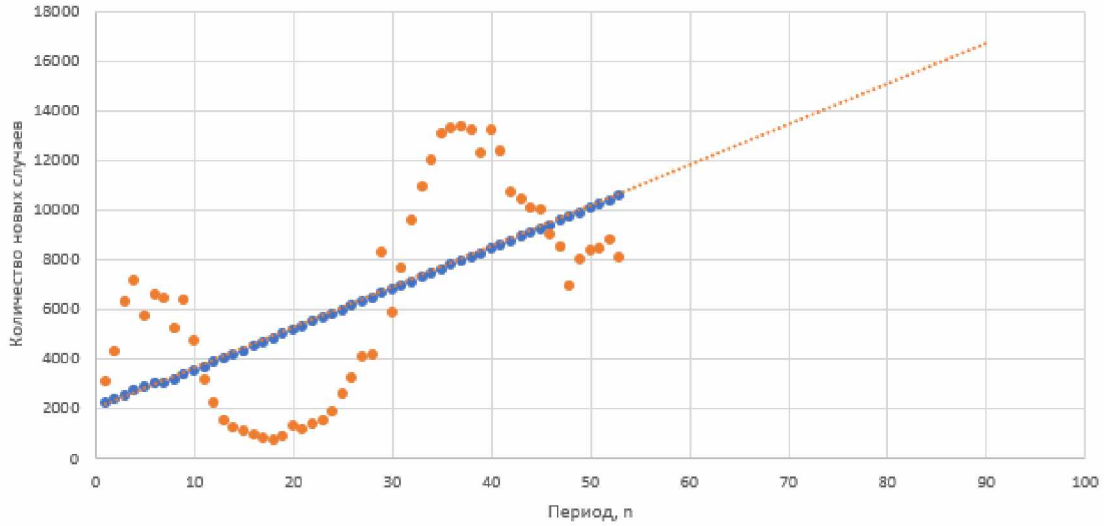

Рисунок 3 – Прогноз возникновения новых случаев заболеваемости COVID-19 на территории Республики Беларусь

Данный график показывает, что на начало 2022 года количество заболевших будет линейно расти и достигнет 16700 случаев, что в корне неверно. На начало 2022 года, согласно данным ВОЗ, количество новых заболевших составило 7820. Это произошло из-за обязательной вакцинации на территории Республики Беларусь после августа 2021 года, соблюдения мер предостережения распространения болезни, выполнения правил личной гигиены и других факторов, которые не допустили линейного распространения заболеваемости на территории страны.

Заключение. Проведенное исследование позволило сделать вывод, что линейная регрессия крайне неэффективна при прогнозировании распространения заболеваемости и смертности от COVID-19. Также на неэффективность модели повлияло малое количество параметров для оценки и прогноза. Для более точного и достоверного прогноза необходимо использовать модели машинного обучения, которые принимают во внимание большее количество параметров и имеют в своей основе более сложную логику работы.

Список литературы

1. *Ali Sendur, Zafer Cakir A comparative study for COVID-19 forecasting models // International Conference on Scientific and Innovative Studies. - 2023. – Vol.1, №1. – С.195-199.*

2. WHO COVID-19 dashboard data [Электронный ресурс]. – Режим доступа: https://data.who.int/dashboards/covid19/data?n=c.

UDC 004.67

FORECASTING THE SPREAD OF COVID-19 ON THE TERRITORY OF THE REPUBLIC OF BELARUS WITH THE APPLICATION LINEAR REGRESSION

Larkin A.D

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Tonkavich I.N. – Cand. of Che., associate professor, associate professor of the department of ICSD

Annotation. A linear regression model was used to predict morbidity and mortality from COVID-19 in the Republic of Belarus. The forecast was based on statistical data provided by the World Health Organization.

Keywords: COVID-19, spread forecasting, linear regression.