УДК 004.942

Широкова, Светлана, Владимировна
Shirokova S.V.
канд. техн. наук, доцент
swchirokov@mail.ru
Цветкова, Анна, Денисовна
Tsvetkova A.D.
студент
adtsvetkova@yandex.ru

МОДЕЛИРОВАНИЕ ЭПИДЕМИЙ СРЕДСТВАМИ МАШИННОГО ОБУЧЕНИЯ
EPIDEMIC MODELING WITH MACHINE LEARNING METHODS

Санкт-Петербургский политехнический университет Петра Великого
Peter the Great St. Petersburg Polytechnic University

Аннотация. В данной статье рассматривается вопрос моделирования эпидемии коронавируса в России. Рассматривается возможность аппроксимации значений числа заболевших с помощью экспоненциальной и сигмоидной модели. Параметры находятся с помощью алгоритма машинного обучения – линейной регрессии. Оценка параметров производится с помощью байесовской регуляризации. Делаются выводы об описательной способности рассмотренных моделей.

Abstract. This article is about modeling coronavirus in Russia. Authors try to approximate number of ill people by using exponential and sigmoid models. Parameters are found by dint of linear regression machine learning algorithm. Parameters are evaluated using Bayesian regularization. Authors are drawing conclusions about descriptive power of reviewed models.

Ключевые слова: линейная регрессия, байесовская регуляризация, коронавирус, моделирование, экспонента, гауссиан.

Key words: linear regression, Bayesian regularization, coronavirus, modelling, exponent, gaussian.

Введение. В современном мире, где глобализация и свободное перемещение людей являются неотъемлемыми аспектами жизни и нормального функционирования, эпидемии и пандемии становятся все более распространенными и имеют значительное влияние на различные секторы общества. Одним из секторов, которые сильно подвержены последствиям эпидемий, является бизнес. Эпидемии могут приводить к сокращению спроса, проблемам с поставками, изменению потребительского поведения и непредсказуемым изменениям в экономической ситуации. В свете этих вызовов моделирование эпидемий становится неотъемлемым инструментом для успешного управления бизнесом. Прогнозирование эпидемий позволяет бизнесу адаптироваться к переменным условиям, предвидеть потенциальные риски и принимать рациональные решения. С другой стороны, актуальным и эффективным средством моделирования с развитием информационных технологий является машинное обучение. Оно позволяет автоматически обрабатывать и анализировать эпидемиологические данные больших объемов, выявлять скрытые закономерности и создавать точные прогнозы. Кроме того, модели машинного обучения могут быть обучены на исторических данных и использованы для автоматического и быстрого прогнозирования будущих эпидемий. Это позволяет бизнесу оперативно реагировать на изменяющиеся ситуации и принимать соответствующие меры для минимизации потенциальных негативных последствий. В данной статье рассматриваются подходы к моделированию эпидемий с помощью средств машинного обучения с целью дальнейшего прогнозирования числа заболевших на примере данных об эпидемии коронавируса SARS-CoV-2.

Цель и задачи. Целью данной работы является создание математической модели темпов роста заболеваемости коронавирусом в России с помощью регрессионного анализа.

Для достижения поставленных целей необходимо выполнить следующие задачи:

1. Предложить математическую модель описания данных на основе регрессионного анализа для данных по России;

2. Оценить параметры и точность аппроксимации данных с помощью построенной модели.

Входные данные. В данной работе используются данные по количеству заболевших в России на каждый день с начала пандемии [1].

В датасете содержатся данные за каждый день, начиная с 1 января 2020 года и заканчивая 8 марта 2023 года. В рассмотрение берутся данные

по России, начиная с 3 марта 2020, так как именно в этот момент впервые стало больше 2 заболевших.

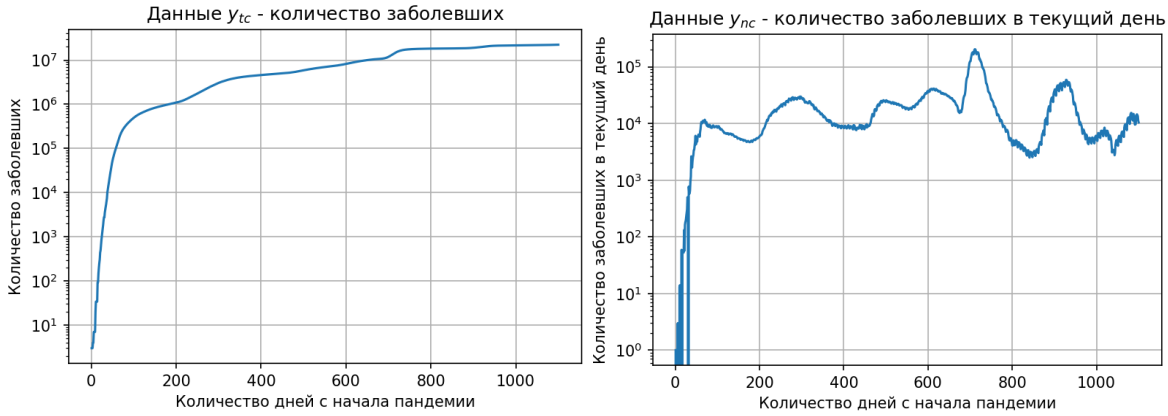
Для моделирования выбраны две переменные:

Y_{nc}– новые случаи заболеваний на текущий день;

Y_{tc} – всего заболевших с начала пандемии на текущий день;

T – номера дня.

Визуализация исходных данных показана на рис. 1 а) и б). Количество заболевших взято в логарифмических осях.



а) б)

Рис. 1. Распределение количества заболевших по дням: а) – для всех случаев, б) – для новых случаев

Математическая модель. 1. Исходя из визуального анализа исходных данных, можно сделать предположение, что рост числа заболевших происходит экспоненциально. Составим линейную регрессионную модель для временного ряда следующим образом:

$$Y_{tc} \sim e^{a_1 T + a_0}$$

$$\ln Y_{tc} \sim a_1 T + a_0$$

Коэффициенты модели оценим с помощью МНК-оценок [2]:

$$[a_1, a_0] = (X'X)^{-1} X' \ln Y_{tc},$$

где $X = [T\ 1]$.

2. Заранее понятно, что экспоненциальная модель может хорошо описывать тенденцию в начальные дни эпидемии, однако для прогнозирования она не подойдет – экспоненциальный рост в природе не может продолжаться вечно. Вводится широкий ряд мер по борьбе с вирусом, поэтому рост числа заболевших замедляется. Кривая общего числа заболевших во время эпидемии более вероятно имеет

сигмоидальный вид: после начальной фазы экспоненциального роста происходит насыщение. Согласно работе Мюррея [3] утверждается, что для пандемии коронавируса лучшим сигмоидом является гауссиан:

$$y_{tc} \sim \int_{-\infty}^{t} e^{a_2 t^2 + a_1 t + a_0}$$

Поскольку необходимо оценить коэффициенты в экспоненте с помощью линейной регрессии, выполним преобразование:

$$\ln \nabla y_{tc} \sim a_2 t^2 + a_1 t + a_0$$

Вспомним определение производной – это предел по приращению времени:

$$\lim_{\Delta t \to 0} \frac{y(t - \Delta t) - y(t)}{\Delta t}$$

Неожиданно заметим, что для $\Delta t = 1$:

$$\lim_{\Delta t \to 0} y_{tc}^{k+1} - y_{tc}^{k} = \lim_{\Delta t \to 0} y_{nc}^{k+1}$$

То есть производная для всех случаев заболевания есть приращение новых случаев:

$$\nabla y_{tc} \sim y_{nc}$$

Тогда МНК-оценки для параметров регрессии можно посчитать следующим образом:

$$[a_2, a_1, a_0] = (X'X)^{-1} X' \ln Y_{nc}$$

где $X = [T^2, T, 1]$.

Оценка качества модели. Оценку качества модели проведем с помощью коэффициента детерминации [4]:

$$R^2 = 1 - \frac{\sum_i \left(y_{tc}^i - \tilde{y}_{tc}^i\right)^2}{Y_{tc} - \mu} \to 1$$

где $\mu = \frac{\sum_i y_{tc}^i}{n}$, n – количество дней, в которые были сделаны наблюдения (дней с начала первого заражения).

Апостериорное распределение параметров модели. С помощью байесовского подхода оценим параметры модели. Для этого необходимо взять априорное распределение. Будем считать, что параметры a распределены нормально [4]:

$$p(a) = N(a|\mu_0, \Sigma_0)$$

Будем предполагать, что данные одинаково распределены и независимы:

$$p(Y_{tc}|X, a, \sigma^2) = \prod_{i=1}^{n} N(y_i|a^T\phi(x_i), \sigma^2)$$

Апостериорное распределение параметров будет выражаться как

$$p(a|Y_{tc}) \propto p(Y_{tc}|X, a, \sigma^2)p(a) = N(a|\mu_0, \Sigma_0)\prod_{i=1}^{n} N(y_i|a^T\phi(x_i), \sigma^2)$$

Тогда, окончательно,

$$p(a|Y_{tc}) = N\left(a\middle|\tilde{\mu}_n, \tilde{\Sigma}_n\right),$$

где

$$\tilde{\Sigma}_n = \left(\Sigma_0^{-1} + \frac{X^T X}{\sigma^2}\right)^{-1}$$

$$\tilde{\mu}_n = \tilde{\Sigma}_n\left(\Sigma_0^{-1}\tilde{\mu}_0 + \frac{X^T Y_{tc}}{\sigma^2}\right)$$

Значение шума σ посчитаем как стандартное отклонение реальных значений и тех, которые получаются при МНК-оценке коэффициентов линейной регрессии:

$$\sigma = \sqrt{\frac{\sum_i^l \left(z_i - \overline{Z}\right)^2}{l}},$$

где l – количество элементов в Z,

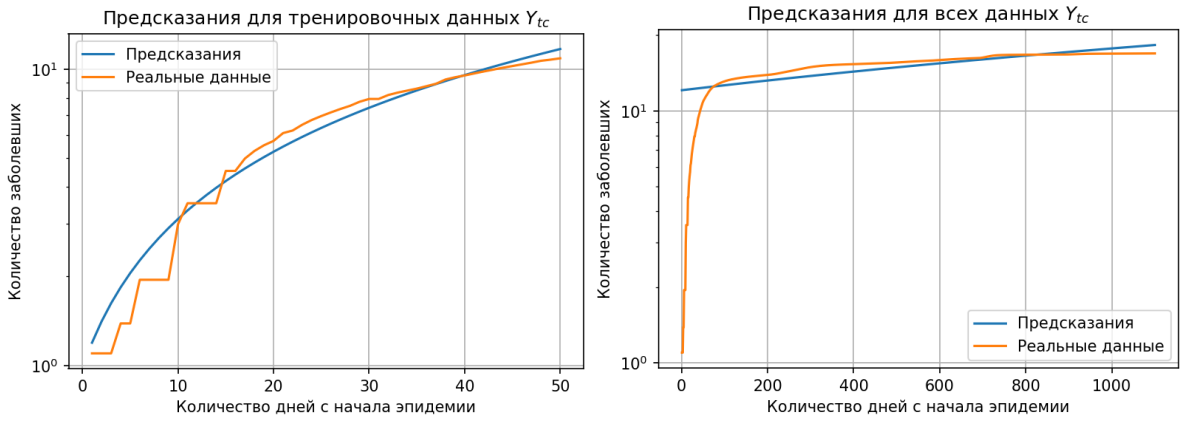
$$Z = Y_{tc} - X((X'X)^{-1}X'Y_{tc}$$

Результаты. В ходе работы первые 50 дней эпидемии считались основными тренировочными данными.

Более детальный анализ данных показал, что общее число заболеваний изменяется по следующему сценарию: во время вспышки вируса тенденция приобретает экспоненциальный рост, а затем выходит на плато.

1. Экспоненциальная модель, ожидаемо, более-менее неплохо описывает только начало заболевания во время его вспышки. Этот результат можно объяснить тем, что в данный период нет возможности оперативно принять меры по предотвращению заболевания. Согласно рис.

2 (б), понятно, что при долгосрочном прогнозировании результаты получаются чересчур пессимистичными. Кроме того, точность экспоненциальной модели даже на начальных этапах эпидемии оставляет желать лучшего, о чем говорят отрицательные коэффициенты детерминации.



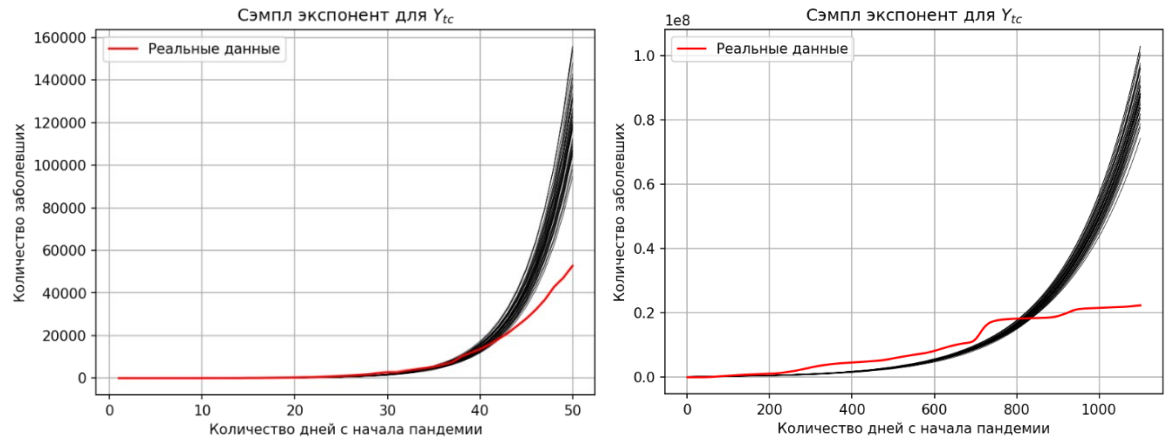
а) $R^2 = -0.361$ б) $R^2 = -1.486$

Рис. 2. Предсказания для экспоненциальной модели: а) – для тренировочных данных, б) – для всех данных

Для применения байесовской регуляризации использовались следующие параметры нормального априорного распределения:

$$\mu_0 = [0, 0]$$

$$\Sigma_0 = \begin{bmatrix} 250 & 0 \\ 0 & 250 \end{bmatrix}$$



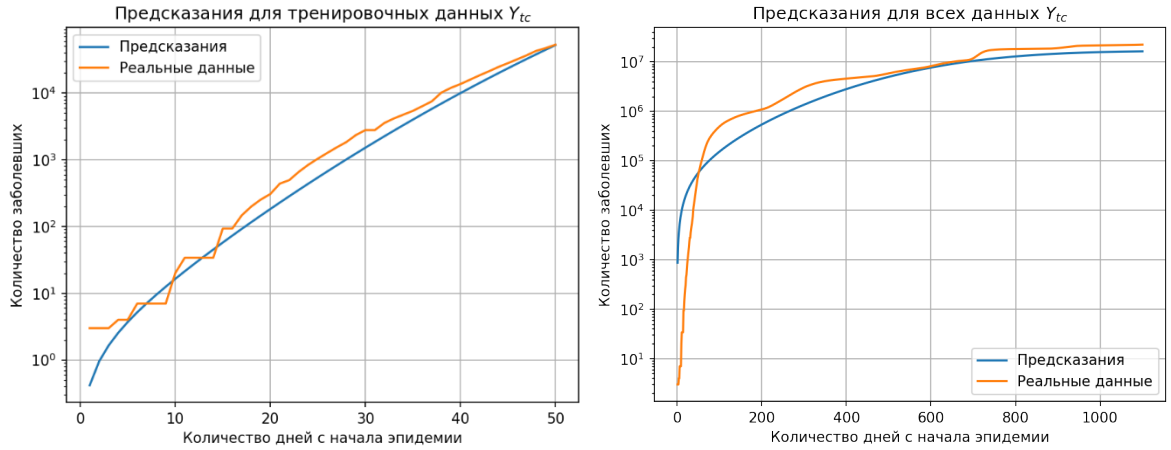
а) $\mu_n = [0.214, 0.979]$ б) $\mu_n = [5.628e^{-3}, 12.083]$

$$\Sigma_n = \begin{bmatrix} 1.944e^{-5} & -4.956e^{-4} \\ -4.956e^{-4} & 1.669e^{-2} \end{bmatrix} \quad \Sigma_n = \begin{bmatrix} 2.214e^{-8} & -1.219e^{-5} \\ -1.219e^{-5} & 8.942e^{-3} \end{bmatrix}$$

Рис. 3. Сэмпл экспонент по оценкам байесовской регуляризации: а) – для тренировочных данных, б) – для всех данных

На рис. 3 показано сэмплирование согласно байесовским оценкам параметров линейной регрессии. Видно, что уже через 40 дней с начала эпидемии экспоненциальное распределение чересчур завышает предсказания по числу заболеваний.

2. Сигмоида (гауссиан) дает высокий предсказательный результат на начальных этапах заболевания (рис. 4 (а)) с высоким коэффициентом детерминации, однако не может описать данные полностью верно в долгосрочной перспективе, так как не может учесть внешние изменяющиеся факторы. Таким образом, при описании всего набора данных происходят «провалы» (рис. 4 (б)), при которых предсказания занижаются в отличие от реальных данных.



а) $R^2 = 0.854$ б) $R^2 = 0.368$

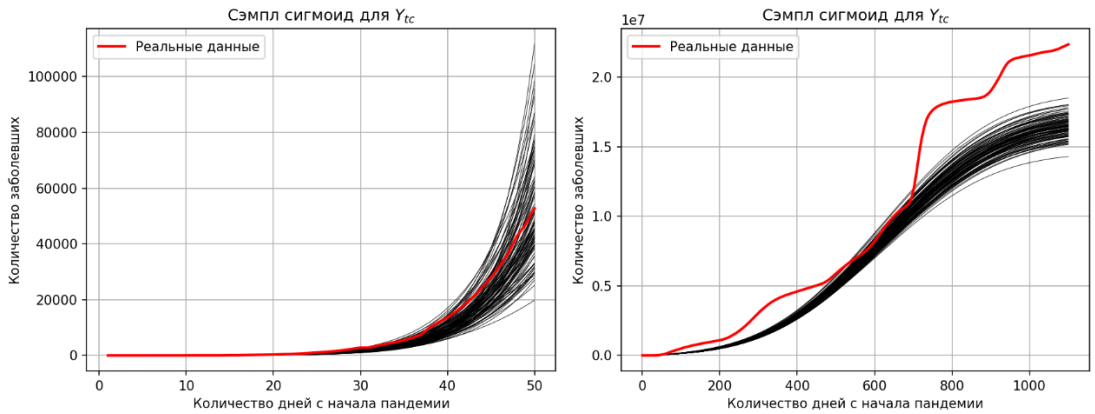
Рис. 4. Предсказания для гауссиана: а) – для тренировочных данных, б) – для всех данных

Оценки с помощью байесовского подхода выполнялись из предположения, что априорное распределение является нормальным со следующими параметрами:

$$\mu_0 = [0, 0, 0]$$

$$\Sigma_0 = \begin{bmatrix} 1000 & 0 & 0 \\ 0 & 1000 & 0 \\ 0 & 0 & 1000 \end{bmatrix}$$

На рис. 5 изображено сэмплирование предсказательных кривых в соответствии с оценками параметров линейной регрессии. Можем заметить, что, действительно, гауссиан имеет хорошую описательную способность в первые дни эпидемии. Однако при попытке описать полный набор данных за три года становится понятно, что полная тенденция скорее угадывается (на 450-700 дни эпидемии – см. рис. 5 (б)). При этом предсказания в общем случае склонны быть заниженными.



а) $\mu_n = [0.260, -0.001, -1.132]$

$$\Sigma_n = \begin{bmatrix} 2.285e^{-3} & -4.211e^{-5} & -2.211e^{-2} \\ -4.211e^{-5} & 8.257e^{-7} & 3.649e^{-4} \\ -2.211e^{-2} & 3.649e^{-4} & 2.792e^{-1} \end{bmatrix}$$

б) $\mu_n = [1.123e^{-2}, -9.016e^{-6}, 6.755]$

$$\Sigma_n = \begin{bmatrix} 2.067e^{-7} & -1.760e^{-10} & -4.270e^{-5} \\ -1.760e^{-10} & 1.599e^{-13} & 3.233e^{-8} \\ -4.270e^{-5} & 3.233e^{-8} & 1.175e^{-2} \end{bmatrix}$$

Рис. 5. Сэмпл сигмоид по оценкам байесовской регуляризации: а) – для тренировочных данных, б) – для всех данных

Выводы. Исходя из данных, полученных в ходе экспериментов, можно сделать вывод, что экспоненциальное распределение не подходит для описания эпидемиологических данных. Кроме того, предположение о хорошей описательной способности гауссиана также не подтвердилось: на начальных этапах эпидемии данные приближаются достаточно хорошо, однако в долгосрочной перспективе тенденция меняется непредсказуемо. Стоит обратить внимание на моделирование с помощью вероятностных моделей SIR, SEIR. При моделировании также необходимо учитывать дополнительные внешние факторы: сезонность, введение массовой вакцинации, появление новых штаммов вируса и другие.

Список литературы.

[1] Coronavirus Pandemic (COVID-19) [Электронный ресурс] // Our World in Data URL: https://ourworldindata.org/coronavirus (дата обращения: 08.03.2023)

[2] Econometric Theory. / Goldberger, Arthur Stanley // New York: John Wiley & Sons. 1964. P. 158

[3] Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months // Christopher JL Murray // medRxiv. 2020.

[4] Бахрушин В. Е. Методы оценивания характеристик нелинейных статистических связей // Системные технологии. — 2011. — №2(73). — С. 9—14.

[5] С. Николенко. Байесовский вывод в линейной регрессии [Электронный ресурс] // Сергей Николенко, личный сайт. URL: https://logic.pdmi.ras.ru/~sergey/teaching/mlspsu22/04-linbayes.pdf (дата обращения: 08.03.2023)