УДК 51-76

МЕТОДЫ ДИСКРЕТНОГО МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ УРОВНЯ ЗАБОЛЕВАЕМОСТИ НАСЕЛЕНИЯ

Анисов А.С.

студент,

ФГБОУ ВО «Казанский национальный исследовательский технический университет им. А.Н. Туполева–КАИ»,

Казань, Россия

Кремлева Э.Ш.

к.т.н., доцент,

ФГБОУ ВО «Казанский национальный исследовательский технический университет им. А.Н. Туполева–КАИ»,

Казань, Россия

Аннотация

В работе предложена модель множественной линейной регрессии для прогнозирования уровня заболеваемости гриппом, учитывающая краткосрочные (недельные) и долгосрочные (сезонные годовые) тенденции. С опорой на данные за 2016-2025 гг. показан пример выбора оптимальных параметров модели и определения коэффициентов регрессии, на основе которых строится модель, демонстрирующая высокую точность.

Ключевые слова: математическое моделирование, множественная линейная регрессия, прогнозирование, заболеваемость, эпидемия.

DISCRETE MATHEMATICAL MODELING METHODS FOR PREDICTING THE MORDIBITY RATE OF THE POPULATION

Anisov A.S.

student,

Federal State Budgetary Educational Institution of Higher Education «Kazan National Research Technical University named after A.N. Tupolev–KAI», Kazan, Russia

Kremleva E.Sh.

Ph.D., Associate Professor,

Federal State Budgetary Educational Institution of Higher Education «Kazan National Research Technical University named after A.N. Tupolev–KAI», Kazan, Russia

Abstract

The paper proposes a multiple linear regression model for predicting the incidence of influenza, taking into account short-term (weekly) and long-term (seasonal annual) trends. Based on the data for 2016-2025, an example of choosing optimal model parameters and determining regression coefficients is shown, on the basis of which a model is built that demonstrates high accuracy.

Keywords: mathematical modeling, multiple linear regression, forecasting, morbidity, epidemic.

В современном мире прогнозирование уровня заболеваемости является актуальной задачей в области медицины. Сезонные эпидемии – например, гриппа – оказывают большое влияние на здоровье населения, создавая высокую нагрузку на систему здравоохранения и социальную сферу. Для своевременного и качественного решения проблемы требуются высокоточные методы для выявления закономерностей и прогноза уровня заболеваемости.

В данной работе рассматривается решение данной задачи при помощи методов дискретного математического моделирования. Такие методы позволяют выявлять скрытые закономерности в исходных данных: как краткосрочные (за последние недели или месяцы), так и долгосрочные (связанные в первую очередь с сезонностью заболеваемости) [4].

Цель работы – спроектировать математическую модель для прогноза уровня заболеваемости населения. Для создания модели и оценки её эффективности будут использоваться данные о заболеваемости населения в 2016–2025 годах, предоставленные НИИ гриппа им. А.А. Смородинцева [1]. В этом источнике отсутствуют данные за летние периоды, однако это не препятствует построению модели.

Модель должна давать прогноз об уровне заболеваемости гриппом на выбранную неделю с точностью не менее 75% – данная величина обусловлена тяжёлой предсказуемостью вспышек, изменчивостью вируса и сильного влияния внешних факторов.

Для прогноза уровня заболеваемости гриппом необходимо определить наиболее влияющие на данную величину факторы. Построение модели осуществлялось с опорой на следующие предположения:

1) Уровень заболеваемости y на текущей неделе напрямую зависит от уровней заболеваемости x в течение n последних недель;
2) Уровень заболеваемости y на текущей неделе будет в той или иной степени зависеть от уровней заболеваемости z на точно такой же неделе в течение m последних лет.

Первое предположение позволяет строить модель на основе краткосрочной динамики уровня заболеваемости. Второе же предположение указывает на периодичный характер изменения уровня заболеваемости: в одно и то же время года он приблизительно одинаков, и имеет свою «годовую» динамику.

В качестве вида модели была выбрана модель множественной линейной регрессии [2, 3]. С учётом вышеобозначенных предположений модель должна иметь вид

$$y = f(x, z) = \sum_{i=1}^{n} a_i x_i + \sum_{j=1}^{m} b_j z_j + c,$$

где

x_i – уровень заболеваемости i недель назад;

y_j – уровень заболеваемости j лет назад на такой же неделе;

a_i – степень влияния уровня заболеваемости i недеьной давности на текущее прогнозируемое значение;

b_j – степень влияния уровня заболеваемости j годовой давности на текущее прогнозируемое значение;

c – постоянная составляющая уровня заболеваемости.

Таким образом, построение модели подразумевает подбор коэффициентов a_i $(i = \overline{1,n})$, b_j $(j = \overline{1,m})$ и c, которые будут удовлетворять условию оптимальности. Под таким условием будем подразумевать минимум суммы квадратов отклонений реальных значений от прогнозируемых:

$$S(a,b,c) = \sum_{k=1}^{N} \left(y^k - f(x^k, z^k)\right)^2 \to \min \ ,$$

т.е.

$$\sum_{k=1}^{N} \left(y^k - \left(\sum_{i=1}^{n} a_i x_i^k + \sum_{j=1}^{m} b_j z_j^k + c \right) \right)^2 \to \min \ .$$

Из данного условия можно получить систему $n + m + 1$ нормальных уравнений с $n + m + 1$ неизвестными, которые будут однозначно определять искомые коэффициенты [3]:

$$\sum_{i=1}^{n} \left(\sum_{k=1}^{N} x_i^k x_s^k \right) a_i + \sum_{j=1}^{m} \left(\sum_{k=1}^{N} z_j^k x_s^k \right) b_j + \left(\sum_{k=1}^{N} x_s^k \right) c = \sum_{k=1}^{N} y^k x_s^k \,, \qquad s = \overline{1,n},$$

$$\sum_{i=1}^{n} \left(\sum_{k=1}^{N} x_i^k z_s^k \right) a_i + \sum_{j=1}^{m} \left(\sum_{k=1}^{N} z_j^k z_s^k \right) b_j + \left(\sum_{k=1}^{N} z_s^k \right) c = \sum_{k=1}^{N} y^k z_s^k \,, \qquad s = \overline{1,m},$$

$$\sum_{i=1}^{n} \left(\sum_{k=1}^{N} x_i^k \right) a_i + \sum_{j=1}^{m} \left(\sum_{k=1}^{N} z_j^k \right) b_j + Nc = \sum_{k=1}^{N} y^k.$$

Таким образом, поиск коэффициентов регрессии был сведён к решению системы $n+m+1$ нормальных уравнений с $n+m+1$ неизвестными.

Построенная математическая модель множественной линейной регрессии имеет следующие параметры:

1) Число n учитываемых последних недель;
2) Число m учитываемых последних лет.

Очевидно, что чем больше n, тем менее зависит y от фактора x_n, а потому нет смысла брать слишком большое значение параметра n. Аналогичные рассуждения справедливы и для оценки параметра m, причём, учитывая недавнюю эпидемию COVID-19, неразумно брать $m > 3$ для оценки уровня заболеваемости текущей недели, поскольку в таком случае всякая линейная зависимость практически отсутствует (рис. 1).

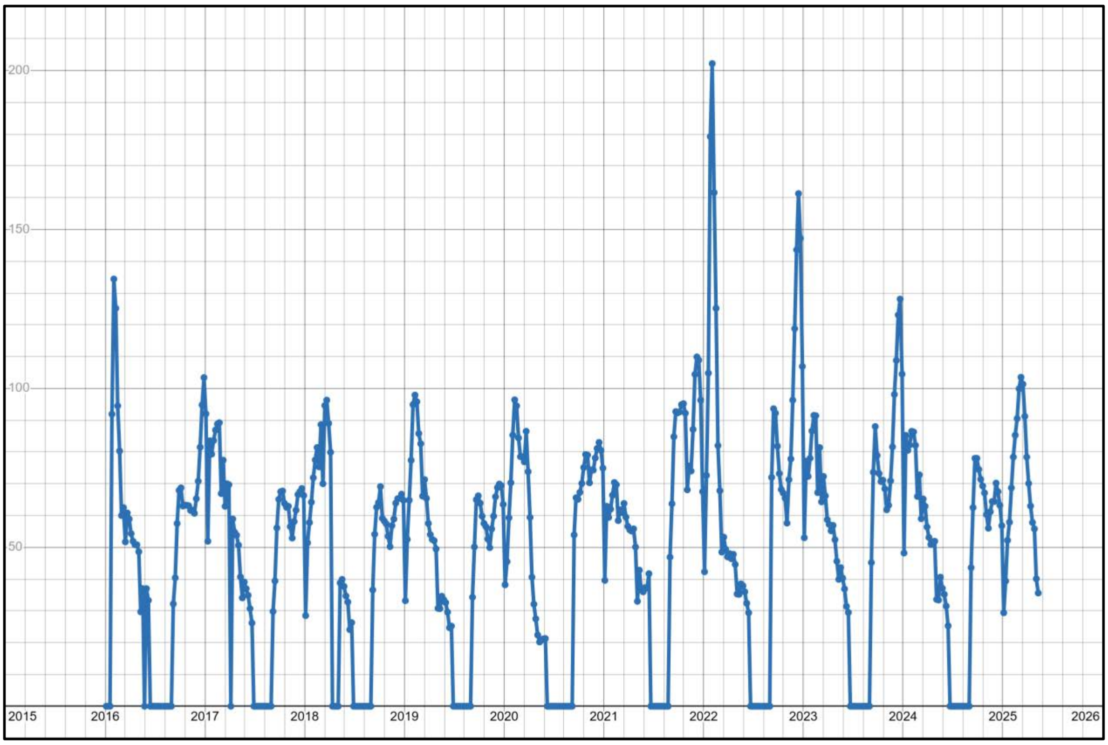

Рис. 1 – Изменение уровня заболеваемости гриппом

Для выбора наиболее оптимальных значений параметров использовался коэффициент детерминации, который рассчитывался для моделей с разными значениями параметров (таблица 1).

Таблица 1 – Коэффициенты детерминации моделей, %

n \ m	1	2	3
1	82.563	80.681	65.100
2	85.724	84.110	72.010
3	85.329	83.434	67.215
4	84.824	83.301	72.284
5	85.020	83.783	73.330
6	84.794	83.591	73.338
7	84.488	83.387	73.043

Таким образом, для прогнозирования на основе выбранных данных наиболее оптимальными будут параметры с $n = 2$ или $n = 3$ и $m = 1$. В дальнейшем будем использовать модель с параметрами $n = 3$, $m = 1$, т.е. будем учитывать в ней данные за 3 последние недели и 1 прошлый год.

Решая систему нормальных уравнений относительно искомых коэффициентов, получим следующую модель:

$$y = 0.957x_1 - 0.111x_2 - 0.133x_3 + 0.195z_1 + 5.705,$$

имеющую коэффициент детерминации $R^2 = 85.329\%$.

Для прогноза выберем 20-ю неделю 2025-го года. В таком случае входной вектор для модели будет равен $(x_1, x_2, x_3, z_1) = (35.6, 40.1, 55.8, 40.6)$. Для таких аргументов получаем прогноз $y = 35.84040$ при реальном уровне заболеваемости $y^* = 40.1$, т.е. имеем точность 89.377%.

Построенная модель достаточно хорошо описывает зависимость уровня заболеваемости от номера недели и года, о чём свидетельствует высокий коэффициент детерминации 85.329%, а также точность 89.377% для прогноза на 20-ю неделю 2025-го года.

Представленная в работе модель имеет следующие достоинства:

1) Модель учитывает как краткосрочные, так и долговременные тенденции, а также сезонность.
2) Параметры модели можно адаптировать под конкретные условия, в которых выполняется прогноз. Например, для разной местности, социальных групп или при анализе других заболеваний необходимые значения m и n можно выбрать иначе. Также можно легко изменить данные для определения коэффициентов a_i, b_j и c.
3) Прогнозы с использованием модели обладают высокой точностью.
4) Модель способна адаптироваться к хаотичному влиянию внешних факторов.
5) Простота вычислений и обучения модели.

Вместе с тем у модели есть и свои недостатки:

1) Для прогноза модели необходимы полные данные за прошедшие недели, а также за несколько последних лет.
2) Модель плохо работает в ситуациях, когда рост заболеваемости имеет нелинейный характер.
3) Модель не способна учитывать влияние иных факторов.

Для устранения первых двух ограничений модели возможны следующие её модификации:

1) Внедрение аппроксимации недостающих данных различными математическими методами: например, интерполяцией, методом скользящего среднего или экспоненциального сглаживания.
2) Усложнение структуры модели за счёт включения нелинейных членов.

Библиографический список:

1. Еженедельный бюллетень по гриппу // НИИ гриппа имени А.А. Смородинцева. [Электронный ресурс]. – Режим доступа – URL:

https://www.influenza.spb.ru/surveillance/flu-bulletin (дата обращения: 24.05.2025).
2. Кондратьев, М. А. Методы прогнозирования и модели распространения заболеваний / М. А. Кондратьев // Компьютерные исследования и моделирование. – 2013. – Т. 5, № 5. – С. 863-882.
3. Курс теории вероятностей и математической статистики : учеб. пособие / Н.Е. Роднищев ; Мин-во образования РФ, КГТУ им. А.Н. Туполева. – Казань : Изд-во КГТУ им. А.Н. Туполева, 2001. – 156 с.
4. Тарасова, С. А. Прогнозирование временного ряда инфекционной заболеваемости / С. А. Тарасова // Программные продукты и системы. – 2019. – № 2. – С. 337-342.

Оригинальность 81%