УДК 519.233.5

П.С. КНОПОВ, А.С. КОРХИН

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДИНАМИКИ ИНФИЦИРОВАНИЯ КОРОНАВИРУСОМ С ПОМОЩЬЮ ПОШАГОВОЙ РЕГРЕССИИ С ПЕРЕКЛЮЧЕНИЯМИ[1]

Аннотация. Рассмотрено моделирование динамики инфицирования коронавирусом с использованием регрессии с переключениями, точки переключения которой неизвестны. Описан пошаговый процесс построения регрессии во времени. Исследована динамика инфицирования коронавирусом в Украине.

Ключевые слова: регрессия, точки переключения, параметры регрессии, пошаговое оценивание, инфицирование, коронавирус.

ВВЕДЕНИЕ

Регрессия с переключениями представляет собой совокупность регрессионных моделей, последовательно расположенных во времени, которые могут быть как не связанными, так и связанными между собой. Регрессии разделяются одна от другой точками переключений, которые часто неизвестны. Этот случай является предметом изучения исследователями. Особо следует выделить работы П. Перрона с соавторами (см., например, [1–3]). В них предлагается использовать алгоритм Р. Беллмана и Р. Рота [4] оценивания точек переключения методом динамического программирования. Разработки П. Перрона с соавторами использовались для решения экономических задач. Ряд результатов в построении регрессий с переключениями предложен также авторами настоящей статьи. В работах [5, 6] описаны методы оценивания точек переключения по заданной выборке, позволяющие учитывать ограничения на эти точки переключения и параметры регрессии, которые вытекают из априорной информации о моделируемом процессе. Такие ограничения невозможно учесть, используя схему динамического программирования.

В некоторых приложениях (например, экономике, здравоохранении) концепция фиксированного интервала наблюдения, которая используется в [1–3, 5, 6], не всегда приемлема ввиду непрерывного обновления данных. В качестве примера сошлемся на процесс инфицирования коронавирусом, являющийся актуальным в настоящее время.

В данной статье предлагается анализировать динамику инфицирования, опираясь на регрессионную модель с переключениями, а само построение модели выполнять по шагам во времени. Интервал наблюдения, длина которого фиксирована или увеличивается во времени, разбивается на последовательность перекрывающих один другого достаточно коротких интервалов I_j, $j = 1, 2, \ldots$, содержащих априори небольшое число точек переключения, например не более двух. Тем самым задача их оценивания значительно упрощается.

1. МЕТОДОЛОГИЯ ПОСТРОЕНИЯ МОДЕЛИ ВРЕМЕННОГО РЯДА ЧИСЛА ИНФИЦИРОВАННЫХ КОРОНАВИРУСОМ

На рис. 1, данные для которого взяты из [7], представлен временно́й ряд числа лиц, ежесуточно инфицированных коронавирусом (ЧИКВ) в Украине, начиная с 12.04.20, когда уровень ЧИКВ имел четко выраженную тенденцию роста. Из рисунка видно, что скорость изменения ЧИКВ не постоянная — она меняется со временем не только по величине, но и по знаку. Поэтому измене-

[1]Работа частично поддержана грантом № 2020.02/0121 Национального фонда исследований Украины.

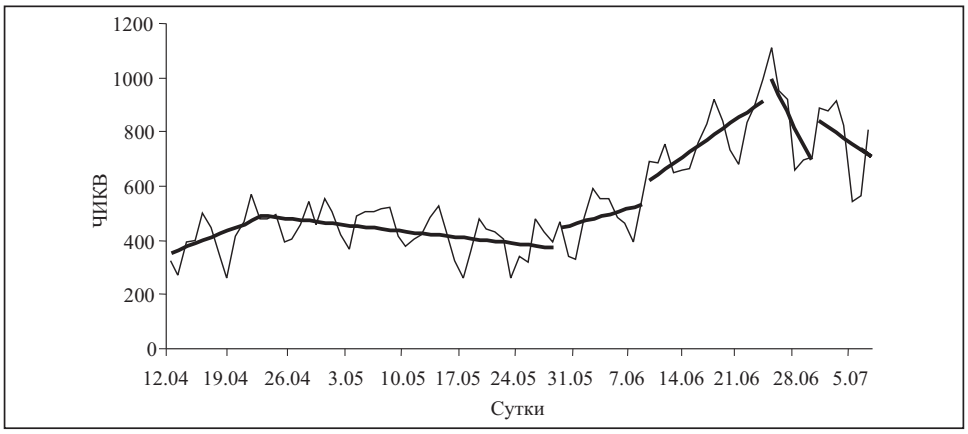

Рис. 1. Временной ряд ЧИКВ и аппроксимирующая его линия регрессии с переключениями (жирная линия)

ние ЧИКВ целесообразно описать линейной регрессией с переключениями. Анализ динамики ЧИКВ показал, что для нее характерны регулярные колебания, период которых составляет одну неделю. Этот факт следует учитывать при построении регрессии с переключениями.

Временной ряд ЧИКВ состоит из тренда TR, регулярных колебаний K и случайной компоненты E. Тренд представляет собой последовательность прямых, отделенных одна от другой точками переключения, которые не обязательно должны стыковаться между собой в этих точках. Регулярные колебания, по-видимому, связаны с трудовой активностью. Случайные колебания обусловлены действием множества второстепенных факторов, к которым можно отнести ошибки тестирования на коронавирус и появление случайных очагов инфекции.

Естественно предположить аддитивную структуру временного ряда ЧИКВ, определяемую моделью

$$y = TR + K + E, \tag{1}$$

где y — число инфицированных.

Определим величины в правой части (1). Вначале исключим тренд из y, используя скользящую среднюю с интервалом усреднения, равным семи – числу дней в неделе:

$$\bar{y}_t = \frac{\sum\limits_{i=t-3}^{t+3} y_i}{7}, \ t = 4, 5, \dots, \tag{2}$$

где y_i — ЧИКВ в t-е сутки ($t = 1$ соответствует дате 12.04.20).

Согласно (1) разность $u_t = y_i - \bar{y}_t$, $t = 4, 5, \dots$, является суммой $K + E$ регулярных колебаний и случайной компоненты. Величина u_t зависит от дня недели, которому соответствует номер суток t. Она может быть представлена суммой $u_t = K^0_{i(t)} + E_t$, где $i(t) = i$ — номер дня недели, соответствующий номеру наблюдения (суток), для определенности считаем, что $i = 1$ соответствует понедельнику; $K^0_{i(t)}$ — истинное неизвестное значение регулярного отклонения ЧИКВ от тренда для i-го дня недели; E_t —- величина случайной компоненты исходного временного ряда.

Чтобы найти оценку K^0_i, усредним все значения u_t, соответствующие i-му дню недели:

$$\hat{K}_i = \sum_{t \in \Omega_i} u_t / |\Omega_i|, \ i = 1, \dots, 7. \tag{3}$$

где Ω_i — множество индексов суток во временном ряду ЧИКВ, которым соответствует день недели i; $|\Omega_i|$ — число элементов в Ω_i. Согласно (2) при вычислении $\hat{K}_i$ по (3) теряется шесть наблюдений: в начале и в конце интервала наблюдения по три наблюдения.

Из (3) имеем $\hat{K}_i = K_i^0 + \sum_{t \in \Omega_i} E_t / |\Omega_i|$. Если E_t, $t = 1, 2, \ldots$, — последовательность взаимно независимых одинаково распределенных случайных величин, причем $E\{E_t\} = 0$, то при $t \to \infty$ (поскольку $|\Omega_i| \to \infty$) второе слагаемое в формуле для $\hat{K}_i$ сходится к нулю в среднем квадратическом. Тогда $\hat{K}_i$ будет состоятельной оценкой K_i^0, $i = 1, \ldots, 7$.

В силу симметричности регулярных колебаний относительно тренда, которую можно предположить согласно рис. 1, имеем $\sum_{i=1}^{7} K_i^0 = 0$. Оценки регулярных колебаний для конечного числа наблюдений, определенные по (3), не будут удовлетворять этому условию. Поэтому их необходимо откорректировать так, чтобы получить

$$\sum_{i=1}^{7} \hat{K}_i = 0. \tag{4}$$

В табл. 1 приведены результаты расчета величин $\hat{K}_i$ для временного интервала $I_1 = [12.04.20, 12.05.20]$, имеющем 31 наблюдение.

В шестой строке табл. 1 величины $\hat{K}_i^-$ вычислены по формуле (3), а в последней строке они откорректированы в соответствии с условием (4). Отметим, что $\sum_{i=1}^{7} \hat{K}_i^- = 35{,}784$, что представляется большой величиной.

График $\hat{K}_i$, $i = 1, \ldots, 7$, не является гладким. Однако с ростом числа наблюдений он изменяется. Так, для ЧИКВ на интервале с 12.04.20 по 10.07.20, содержащем 90 наблюдений, из которых для вычислений по (3) используется 84 наблюдения, $\sum_{i=1}^{7} \hat{K}_i^- = -2{,}143$, а график этих величин становится близким к гладкому (рис. 2). Как видим, максимальная амплитуда колебаний наблюдается в середине недели; в конце и в начале недели имеем минимум — отрицательное число, которое определяет наименьшее число случаев инфицирования за неделю. Поэтому рис. 2 отражает ситуацию, когда на инфицирование в быту наслаивается инфицирование, связанное с выходом на работу и поездками в транспорте.

Таблица 1

Случаи	Результаты расчета $\hat{K}_i$, соответствующие дням недели						
	Понедельник $i = 1$	Вторник $i = 2$	Среда $i = 3$	Четверг $i = 4$	Пятница $i = 5$	Суббота $i = 6$	Воскресенье $i = 7$
1	−73,799	34,774	26,348	116,922	46,496	−67,93	−5,373
2	−22,783	15,791	105,365	12,365	13,365	27,365	−125,738
3	−63,635	−8,635	75,365	−9,635	85,365	37,365	−72,635
4	−98,635	22,365	42,365	39,365	50,365	57,365	−46,635
5	−89,635	−62,635	0	0	0	0	−46,635
$\hat{K}_i^-$	−69,698	0,332	62,361	39,754	48,897	13,541	−59,404
$\hat{K}_i$	−74,81	−4,78	57,249	34,642	43,786	8,429	−64,515

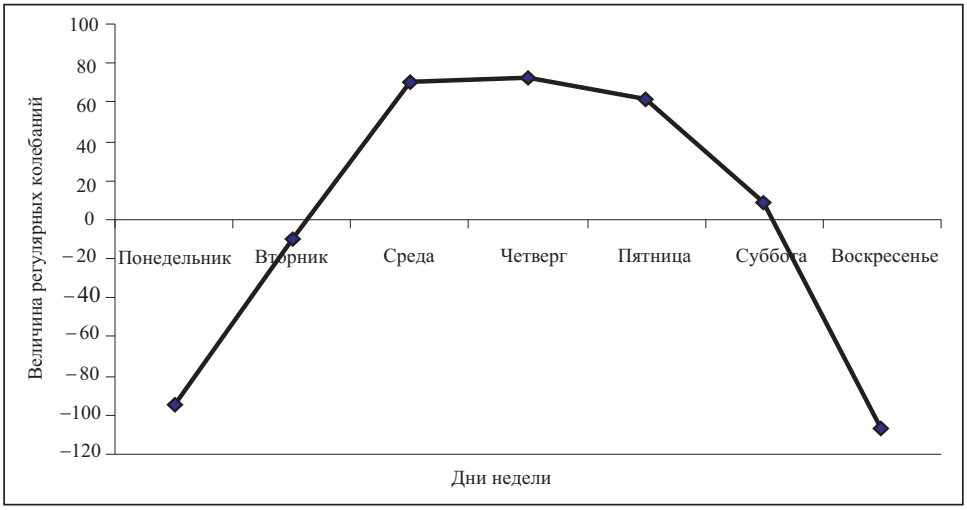

Рис. 2. График регулярных колебаний ЧИКВ по 84 наблюдениям

После выделения регулярных колебаний удалим их из исходного временного ряда ЧИКВ. В результате согласно (1) имеем $z = y - K = TR + E$. Уровни временного ряда z определяются формулой $z_t = y_t - \hat{K}_{i(t)}$.

Изложенная здесь методика аналогична процедуре выделения сезонности из временного ряда, описываемого мультипликативной моделью (все компоненты ряда перемножаются) [8, гл. 15].

Полученную сумму тренда и случайной компоненты на каждом временном интервале I_j, $j = 1, 2, \ldots$, на котором проводится оценивание, опишем регрессией с переключениями, состоящей из $k+1$ линейной регрессии:

$$\left.\begin{aligned} &z_t = \alpha^0_{01} + \alpha^0_{11} t + \varepsilon_{t1},\ t = \tau_{0j}, \tau_{0j}+1, \ldots, t^0_1, \\ &z_t = \alpha^0_{02} + \alpha^0_{12} t + \varepsilon_{t2},\ t = t^0_1 + 1, \ldots, t^0_2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &z_t = \alpha^0_{0k} + \alpha^0_{1k} t + \varepsilon_{tk},\ t = t^0_{k-1} + 1, \ldots, t^0_k, \\ &z_t = \alpha^0_{0,k+1} + \alpha^0_{1,k+1} t + \varepsilon_{t,k+1},\ t = t^0_k + 1, \ldots, \tau_{1j}. \end{aligned}\right\} \qquad (5)$$

Здесь $k = k(j)$ — число точек переключений; α^0_{0i} и α^0_{1i}, $i = 1, \ldots, k+1$, — неизвестные параметры регрессий; t^0_i, $i = 1, \ldots, k$, — неизвестные точки переключений; τ_{0j} и τ_{1j} — соответственно начало и конец интервала I_j; ε_{ti} — случайные компоненты регрессий, характеризующие влияние второстепенных факторов на ЧИКВ. Верхний индекс 0 у перечисленных величин указывает на истинность величины параметра регрессии или точки переключения. Здесь и далее для упрощения обозначений индекс j интервала I_j опущен у k, параметров и точек переключения регрессии.

Поскольку случайные компоненты в (5) являются подпоследовательностями последовательности E_t, $t = 1, 2, \ldots$, то в соответствии с приведенными ее свойствами они отвечают стандартным предположениям регрессионного анализа: не коррелируют между собой, имеют нулевые математические ожидания и одинаковые дисперсии σ^2. Изложенное запишем в виде допущений.

Допущение 1. Случайные компоненты регрессий (5) имеют характеристики: математическое ожидание

$$E\{\varepsilon_{ti}\} = 0,\ t \in \theta^0_i = \{t^0_{i-1} + 1, \ldots, t^0_i\},\ i = 1, \ldots, k+1; \qquad (6)$$

дисперсия

$$E\{\varepsilon_{ti}^2\} = \sigma^2,\ t \in \theta_i^0 = \{t_{i-1}^0 + 1, \dots, t_i^0\},\ i = 1, \dots, k+1; \tag{7}$$

случайные компоненты одной регрессии некоррелированы:

$$E\{\varepsilon_{\tau_1 i}\varepsilon_{\tau_2 i}\} = 0,\ \tau_1, \tau_2 \in \theta_i^0 = \{t_{i-1}^0 + 1, \dots, t_i^0\},\ i = 1, \dots, k+1; \tag{8}$$

случайные компоненты разных регрессий также некоррелированы:

$$\begin{gathered} E\{\varepsilon_{\tau_1 i_1}\varepsilon_{\tau_2 i_2}\} = 0,\ \tau_1 \in \theta_{i_1}^0 = \{t_{i_1-1}^0 + 1, \dots, t_{i_1}^0\}, \\ \tau_2 \in \theta_{i_2}^0 = \{t_{i_2-1}^0 + 1, \dots, t_{i_2}^0\},\ i_1,\ i_2 = 1, \dots, k+1. \end{gathered} \tag{9}$$

Здесь θ_i^0 — интервал времени, на котором параметры регрессии α_{0i}^0, α_{1i}^0 постоянны; для j-го интервала $t_0^0 = \tau_{0j}$, $t_{k+1}^0 = \tau_{1j}$.

Допущение 2. Случайные компоненты регрессий (5) нормально распределены.

Из рис. 1 видно, что возможна существенная разница между ЧИКВ в двух соседних сутках. Поэтому отрезки прямых $\alpha_{0i}^0 + \alpha_{1i}^0 t$, $i = 1, \dots, k+1$, не должны обязательно образовывать непрерывную кусочно-линейную функцию t, т.е. в точках переключений возможен разрыв этой функции.

Свойства случайных компонент регрессии (6)–(9) обусловливают задачу оценивания точек переключений и параметров регрессии с переключениями на каждом интервале оценивания I_j, $j = 1, 2, \dots$:

$$S = \sum_{i=1}^{k+1} \sum_{t=t_{i-1}+1}^{t_i} (z_t - \alpha_{0i} - \alpha_{1i} t)^2 \to \min, \tag{10}$$

$$\tau_{0j} \le t_i \le \tau_{1j},\ t_i - t_{i-1} \ge 2,\ i = 1, \dots, k+1,\ t_0 = \tau_{0j},\ t_{k+1} = \tau_{1j}. \tag{11}$$

Минимизация в задаче (10), (11) осуществляется по параметрам α_{0i}, α_{1i}, $i = 1, \dots, k+1$, — непрерывным величинам и целочисленным точкам переключений t_i, $i = 1, \dots, k$. Поскольку эти величины варьируемые, верхний индекс 0 у них опущен. Переменным точкам переключений в (10), (11) соответствуют интервалы θ_i постоянства параметров, концы которых изменяются:

$$\theta_i = \{t_{i-1} + 1, \dots, t_i\},\ i = 1, \dots, k+1. \tag{12}$$

Ограничение (11) устанавливает минимальное число наблюдений — два для оценивания двух параметров каждой из $k+1$-й прямой на интервалах θ_i, $i = 1, \dots, k+1$, постоянства параметров регрессии с переключениями.

Задачу (10), (11) можно рассматривать как обобщение задачи оценивания параметров нелинейной регрессии с ограничениями [9, ch. 1, 10–13]. Ввиду наличия в ней целочисленных искомых величин она решалась специальным методом [5]. Процесс решения и результаты приводятся ниже для перекрывающих один другого интервалов оценивания I_1, I_2, I_3: $I_1 \cap I_2 \ne \varnothing$, $I_1 \cap I_3 = \varnothing$, $I_2 \cap I_3 \ne \varnothing$. При этом величины $\hat{K}_i$, $i = 1, \dots, 7$, находились по ЧИКВ на конец соответствующего интервала. Так, при построении регрессии с переключениями на интервале I_1 эти величины вычислялись по наблюдениям за этот интервал: с 12.04.20 по 12.05.20. При построении регрессии на интервале I_2 использовались наблюдения, начиная с 12.04.20 по конец этого интервала 21.06.20. Такой подход позволил, в частности, имитировать постепенное поступление данных

о ЧИКВ. что и имело место фактически. Следует отметить, что для решения рассмотренной задачи могут быть предложены и другие подходы, основанные на методах дискретной оптимизации [14, 15].

2. ПОСТРОЕНИЕ РЕГРЕССИИ С ПЕРЕКЛЮЧЕНИЯМИ НА ИНТЕРВАЛЕ I_1

На интервале $I_1 = [12.04.20, 12.05.20]$, содержащем 31 наблюдение, согласно рис. 3 имеется не более одной точки переключения, в которой может измениться ЧИКВ — убывать или возрастать. Добавив одну точку переключения, где, возможно, происходит изменение скорости, получим $k = 2$ в задаче оценивания (10), (11). Ее решение — оценки точек переключения: $\hat{t}_1 = 4$, $\hat{t}_2 = 19$. Оценка длины первого интервала постоянства параметров регрессии θ_1^0 равна четырем. Она слишком мала, что позволяет выдвинуть нулевую гипотезу H_0: $\alpha_{01}^0 = \alpha_{02}^0$; $\alpha_{11}^0 = \alpha_{12}^0$ (первая и вторая регрессии в (5) совпадают). Считая, что вторая точка переключения фиксирована, для проверки сформулированной гипотезы на основе допущений 1 и 2 используем критерий из работы [16]. Согласно ему определим S — сумму квадратов отклонений линейной регрессии с n параметрами от наблюдений на интервале времени длиной T; S_1 и S_2 — суммы квадратов отклонений двух других линейных регрессий — первой и второй с n параметрами каждая от тех же наблюдений на интервалах времени длиной соответственно T_1 и T_2, причем $T_1 + T_2 = T$. Вычислим статистику

$$F^* = \frac{S - (S_1 + S_2)}{(S_1 + S_2)} \frac{T - 2n}{n}. \tag{13}$$

Гипотеза H_0 отклоняется, если $F^* > F_p(n, T-2n)$, где $F_p(n, T-2n)$ есть $100p$ %-я точка F-распределения со степенями свободы n и $T-2n$. В данном случае $T = 19$, $T_1 = 4$, $T_2 = 15$, $n = 2$, $S = 34936{,}83$, $S_1 + S_2 = 31107{,}86$. Имеем $F^* = 0{,}925$, $F_{0,05}(2, 15) = 3{,}682$. Таким образом, на 5%-м уровне гипотеза о совпадении первой и второй регрессий не отклоняется. Значит, не отклоняется гипотеза о том, что на интервале [1, 19] имеется одна точка переключения. Определим ее, решив задачу оценивания (10), (11) для $k = 1$, $j = 1$, $\tau_{01} = 1$, $\tau_{11} = 31$. Полученной оценке единственной точки переключения $\hat{t}_1 = 11 < 19$ соответствует дата 22.04.20. При этом сумма квадратов невязок в (10) составляет $S = 42251{,}76$, а в

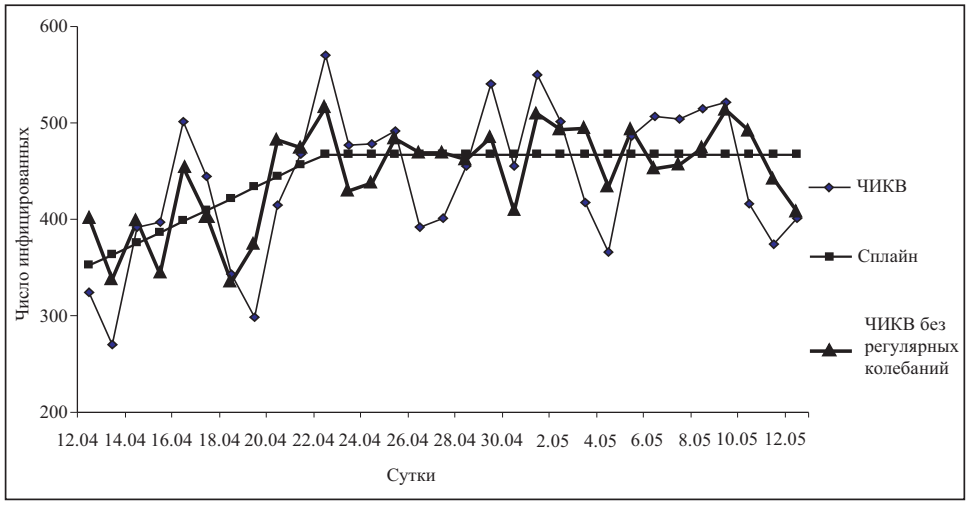

Рис. 3. Построение регрессии с переключениями на интервале I_1

точке $t = \hat{t}_1$ прямые $\hat{\alpha}_{01} + \hat{\alpha}_{11} t$ и $\hat{\alpha}_{02} + \hat{\alpha}_{12} t$, где $\hat{\alpha}_{kl}$ — оценка α_{kl}^0, $k = 0, 1$; $l = 1, 2$, попарно близки. Поэтому к задаче (10), (11) было добавлено ограничение

$$\alpha_{01} + \alpha_{11} t_1 = \alpha_{02} + \alpha_{12} t_1. \tag{14}$$

В результате получена регрессия с переключениями, у которой линия регрессии представляет линейный сплайн (см. рис. 3). Сумма квадратов отклонений (10) для этого случая $S = 42559{,}85$ незначительно увеличилась при добавлении ограничения в задачу оценивания. Это свидетельствует о том, что отмеченная выше нестыковка двух прямых в $\hat{t}_1$ была обусловлена случайными факторами.

Приближенный анализ точности оценок параметров полученного сплайна исходил из того, что найденная точка переключения совпадает с истинной. Такое допущение не является сильноограничивающим для рассматриваемой задачи, так как независимой переменной является время, а не ее некоторая сложная функция. Тогда в соответствии с [8, гл. 15] полученные две регрессии будем рассматривать как одну, варьируемые параметры которой объединены равенством (14), где $t_1 = \hat{t}_1$ — фиксированная величина. Тогда ковариационная матрица оценок параметров регрессии определяется выражением

$$\mathbf{V} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} [\mathbf{J}_{2(k+1)} - \mathbf{G}'\mathbf{S}\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}], \tag{15}$$

где штрих означает транспонирование, $\mathbf{J}_{2(k+1)}$ — единичная матрица порядка $2(k+1)$, $\mathbf{S} = [\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}']$.

В рассматриваемом случае $\mathbf{X} = \operatorname{diag}(\mathbf{X}_1, \mathbf{X}_2)$, $\mathbf{G} = [1 \;\; \hat{t}_1 \;\; -1 \;\; -\hat{t}_1]$, причем

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & \hat{t}_1 \end{bmatrix}, \; \mathbf{X}_2 = \begin{bmatrix} 1 & \hat{t}_1 + 1 \\ 1 & \hat{t}_1 + 2 \\ \vdots & \vdots \\ 1 & \tau_{11} \end{bmatrix}, \; \tau_{11} = 31. \tag{16}$$

В общем случае величины в (15) для j-го интервала оценивания имеют вид

$$\mathbf{X} = \operatorname{diag}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{k+1}), \; \mathbf{X}_1 = \begin{bmatrix} 1 & \tau_{0j} \\ 1 & \tau_{0j} + 1 \\ \vdots & \vdots \\ 1 & \hat{t}_1 \end{bmatrix}, \; \mathbf{X}_{k+1} = \begin{bmatrix} 1 & \hat{t}_k + 1 \\ 1 & \hat{t}_k + 2 \\ \vdots & \vdots \\ 1 & \tau_{1j} \end{bmatrix},$$

$$\mathbf{X}_l = \begin{bmatrix} 1 & \hat{t}_{l-1} + 1 \\ 1 & \hat{t}_{l-1} + 2 \\ \vdots & \vdots \\ 1 & \hat{t}_l \end{bmatrix}, \; l = 2, \ldots, k; \; \mathbf{G} = \begin{bmatrix} 1 & \hat{t}_1 & -1 & -\hat{t}_1 & 0 & 0 & \mathbf{O}_{1m} \\ 0 & 0 & 1 & \hat{t}_2 & -1 & -\hat{t}_2 & \mathbf{O}_{1m} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & \hat{t}_k & -1 & -\hat{t}_k \end{bmatrix}, \tag{17}$$

где $\mathbf{O}_{1m}$ — строка с m столбцами, $m = 2(k+1) - kn$. При $m = 0$ $\mathbf{O}_{10}$ означает отсутствие строки.

Если в некоторых точках отсутствует стыковка прямых, то соответствующие этим точкам строки матрицы $\mathbf{G}$ удаляются. В случае нулевой матрицы $\mathbf{G}$, что соответствует отсутствию ограничений на параметры регрессии, из (15) имеем $\mathbf{V} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$.

Значимость оценок параметров регрессии для фиксированных точек переключения определялась на основе (15) и допущений 1, 2 известными методами. Проверка некоррелированности случайных компонент регрессии проводилась по кри-

терию Дарбина–Уотсона, а их нормальности — по критерию, основанному на коэффициентах асимметрии и эксцесса [17, п. 3.2], его применение см. в [18, п. 9.3].

Для случая $k=1$ согласно формуле (15) была определена оценка ковариационной матрицы $\hat{\mathbf{V}}$ заменой σ^2 ее оценкой $s^2=1464$.

Оценки параметров сплайна оказались значимыми на 5%-м уровне, кроме оценки $\hat{\alpha}_{12}=-1{,}374$. Поэтому была решена задача оценивания (10), (11) с добавленными ограничениями (14) и $\alpha_{12}=0$. В результате получен сплайн, описывающий плавный переход от роста ЧИКВ к плато — горизонтальному участку (см. рис. 3) с высокозначимыми ненулевыми оценками его параметров: $\hat{\alpha}_{01}=340{,}69$; $\hat{\alpha}_{11}=11{,}56$; $\hat{\alpha}_{02}=467{,}81$. Расчет оценки ковариационной матрицы $\mathbf{V}$ проводился согласно исходным данным (16), где матрица $\mathbf{X}_2$ представляет столбец из единиц.

3. ПОСТРОЕНИЕ РЕГРЕССИИ С ПЕРЕКЛЮЧЕНИЯМИ НА ИНТЕРВАЛЕ I_2

Начало интервала $I_2=[23.04.20,\ 21.06.20]$ совпадает с началом плато — точкой переключения, определенной на I_1, плюс 1. Далее возможен сход с плато в сторону увеличения или уменьшения ЧИКВ. Не исключено повторное изменение направления динамики ЧИКВ. Поэтому (как и в разд. 2) положим в задаче оценивания $k=2$. Установим $\tau_{02}=1$, $\tau_{12}=60$. Таким образом, отсчет времени в I_2 начинается с единицы. Решив задачу оценивания для I_2, получим оценки двух точек переключения: $\hat{t}_1=37$ (29.05.20) и $\hat{t}_2=48$ (9.06.20), а также оценки параметров трех регрессий, образующих регрессию с переключениями:

$$\begin{gathered}\hat{\alpha}_{01}=493{,}65;\ \hat{\alpha}_{11}=-3{,}335;\ \hat{\alpha}_{02}=121{,}71;\ \hat{\alpha}_{12}=8{,}498;\\ (0)\qquad (0)\qquad (0{,}471)\qquad (0{,}033)\\ \hat{\alpha}_{03}=-102{,}3;\ \hat{\alpha}_{13}=15{,}275\\ (0{,}585)\qquad (0)\end{gathered}\tag{18}$$

(цифры в скобках означают значимость параметра, которая определялась так же, как в разд. 2).

Согласно (18) в точке $t=37$ закончилось плато. Оно началось в $t=1$ и имеет небольшой наклон со значимым угловым коэффициентом, оценка которого $-3{,}335$. Согласно разд. 2 его величина незначима, а оценка равна –1,374; такое расхождение можно объяснить большим числом данных по плато на интервале I_2.

Прямые линии второй и третьей регрессий согласно (18) имеют угловые коэффициенты одинаковых знаков, причем угловой коэффициент второй прямой незначим на 1%-м уровне. Поэтому была выдвинута нулевая гипотеза о равенстве параметров указанных регрессий, которая проверялась с помощью статистики (13), где $T=23$, $n=2$. Получено $F^*=4{,}97$. Так как $F_{0{,}05}(2,19)=3{,}52$, то нулевая гипотеза была отклонена.

Определим теперь прогноз ЧИКВ по полученной регрессии с переключениями на рассматриваемом интервале. Он заключается в экстраполяции прямой линии третьей регрессии. Отсюда имеем точечный прогноз ЧИКВ

$$\hat{y}_t=\hat{\alpha}_{03}+\hat{\alpha}_{13}t+\hat{K}_{i(t)}=\mathbf{w}_t'\hat{\boldsymbol{\alpha}}_3+\hat{K}_{i(t)},\ \hat{\boldsymbol{\alpha}}_3=[\hat{\alpha}_{03}\ \hat{\alpha}_{13}\ ]',\ \mathbf{w}_t=[1\ t]',\ t>T,\tag{19}$$

где штрих означает траспонирование.

В силу допущений 1 и 2 имеем интервальный прогноз

$$\hat{y}_t-u_p(q)\hat{\sigma}_f(t)\le y_t\le\hat{y}_t+u_p(q)\hat{\sigma}_f(t),\ t>T,\tag{20}$$

где $u_p(q)$ — $100p\,\%$-я точка распределения Стьюдента с числом степеней свободы $q=T-(k+1)n$; $\hat{\sigma}_f(t)$ — оценка с.к.о. прогноза. Найдем ее, исходя из (1);

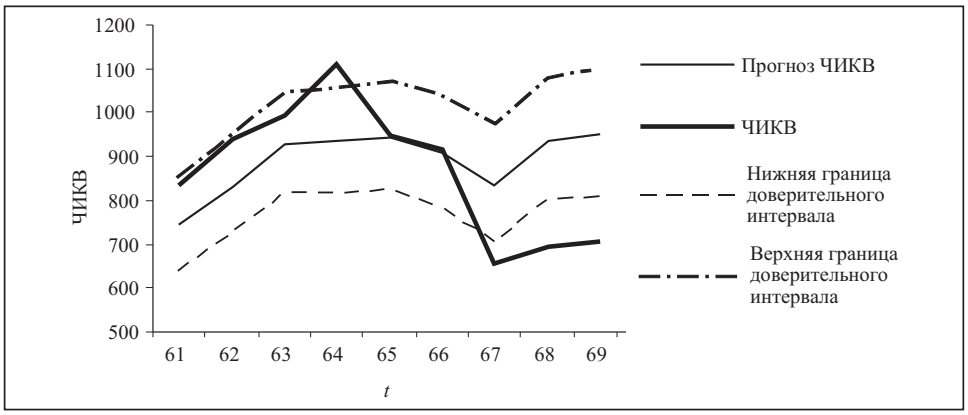

Рис. 4. Прогноз ЧИКВ по результатам оценивания на I_2

в результате имеем $y_t = \mathbf{w}_t'\boldsymbol{\alpha}_3^0 + K_{i(t)}^0 + E_t$, $\boldsymbol{\alpha}_3^0 = [\alpha_{03}^0 \;\; \alpha_{13}^0]'$, $t > T$. Из данного равенства и (19) вытекает ошибка прогноза $\hat{y}_t - y_t = \mathbf{w}_t'(\hat{\boldsymbol{\alpha}}_3 - \boldsymbol{\alpha}_3^0) + (\hat{K}_{i(t)} - K_{i(t)}^0) - E_t$. Отсюда, поскольку $\hat{\boldsymbol{\alpha}}_3$ и $\hat{K}_{i(t)}$ не зависят от E_t, пренебрегая зависимостью $\hat{K}_{i(t)}$ от $\hat{\boldsymbol{\alpha}}_3$, получаем дисперсию прогноза ЧИКВ $\sigma_f^2(t) = \mathbf{w}_t'\mathbf{K}_3\mathbf{w}_t + \sigma^2(\hat{K}_{i(t)}) + \sigma^2$, где $\mathbf{K}_3$ — ковариационная матрица $\hat{\boldsymbol{\alpha}}_3$; $\sigma^2(\hat{K}_{i(t)})$ — дисперсия $\hat{K}_{i(t)}$. Заменив в ней все величины их оценками, получим

$$\hat{\sigma}_f^2(t) = \mathbf{w}_t'\hat{\mathbf{K}}_3\mathbf{w}_t + \hat{\sigma}^2(\hat{K}_{i(t)}) + s^2. \tag{21}$$

На рис. 4 представлены результаты прогноза для $t = 61, \ldots, 69$ (22.06.20–30.06.20), рассчитанного по (19). Границы доверительного интервала для прогноза определены с помощью (20), (21) для $p = 0{,}05$, $q = 60 - 6 = 54$. Согласно рисунку истинная величина ЧИКВ не попадает в доверительный интервал в точках t, равных 64, 67, 68, 69, что может свидетельствовать о появлении в окрестности этих периодов времени одной или двух новых точек переключения и, следовательно, необходимости перехода к новому интервалу оценивания I_3.

Средняя ошибка прогнозирования на интервале времени [61, 66], где была только одна величина ЧИКВ, не попавшая в доверительный интервал, определялась как отношение оценки с.к.о. прогноза к средней истинной величине ЧИКВ на интервале $[61, 66]$. Она составила 9,9 %. Как показали расчеты, ошибка прогноза может быть уменьшена до 7,4 %, если сделать оценку $\hat{\alpha}_{03}$ значимой, т.е. увеличить ее точность. Это достигается стыкованием второй и третьей регрессий в точке $t = 48$.

Для сравнения средняя относительная ошибка прогноза для восьми суток по модели, полученной на интервале I_1, составила 19 %. Такую достаточно большую погрешность можно объяснить невысокой точностью оценивания регулярных колебаний ЧИКВ по малому числу данных, равному 31.

4. ПОСТРОЕНИЕ РЕГРЕССИИ С ПЕРЕКЛЮЧЕНИЯМИ НА ИНТЕРВАЛЕ I_3

На интервале $I_3 = [10.06.20,\ 7.07.20]$ нумерация суток выполняется в продолжение нумерации на предыдущем интервале. Поэтому $\tau_{03} = 49$: начало I_3 находится в последней точке переключения на интервале I_2 плюс 1. Конец интервала $\tau_{13} = 76$. На случай возможного спада или подъема ЧИКВ, обнаруженного в разд. 3, полагаем $k = 2$.

Решение задачи оценивания (10), (11) определяет оценки точек переключения $\hat{t}_1 = 63$ (24.06.20), $\hat{t}_2 = 69$ (30.06.20) и оценки параметров регрессии:

$$\underset{(0,034)}{\hat{\alpha}_{01}} = -411{,}77;\ \underset{(0)}{\hat{\alpha}_{11}} = 21{,}11;\ \underset{(0)}{\hat{\alpha}_{02}} = 4810{,}11;\ \underset{(0)}{\hat{\alpha}_{12}} = -59{,}55;$$
$$\underset{(0)}{\hat{\alpha}_{03}} = 2277{,}1;\ \underset{(0{,}059)}{\hat{\alpha}_{13}} = -20{,}55, \qquad (22)$$

где обозначения в скобках имеют тот же смысл, что и в (18).

Согласно (22) все угловые коэффициенты, за исключением $\hat{\alpha}_{13}$, высокозначимы, в точке $\hat{t}_1$ начался резкий спад, который в $\hat{t}_2$ сменился медленным убыванием. Ввиду близости этих точек была проверена гипотеза о равенстве параметров второй и третьей регрессий, считая точку $\hat{t}_1 = 63$ приближенно известной, так как в ее окрестности началось резкое снижение ЧИКВ. Согласно критерию (13) это предположение было отвергнуто на 5 %-м уровне: $F^* = 5{,}36$, $F_{0,05}(2{,}9) = 4{,}26$. Резкое снижение ЧИКВ можно объяснить ликвидацией одного или нескольких очагов инфекции. Незначимость $\hat{\alpha}_{13}$ может быть объяснена тем, что динамика инфицирования, возможно, ненадолго вышла на очередное плато.

ЗАКЛЮЧЕНИЕ

В настоящей статье рассмотрена модель динамики инфицирования коронавирусом в виде регрессии с переключениями, точки переключения которой неизвестны.

Предложено пошаговое решение задачи ее построения. На каждом шаге вначале решалась задача оценивания для двух точек переключения по наблюдениям на некотором небольшом интервале времени I_j, $j = 1, 2, 3$. Затем проводился статистический анализ построенной части регрессии, что значительно упростило задачу оценивания. В результате искомая регрессия определилась на трех последовательных интервалах времени: $I_1 \setminus I_1 \bigcap I_2$, $I_2 \setminus I_2 \bigcap I_3$, I_3. Целостно линия этой регрессии получена в виде кусочно-линейной функции времени, поскольку концы данных интервалов совпадают с последними точками переключения предыдущего интервала плюс один или концами интервала наблюдения (см. рис. 1).

Использование прогноза позволяет получать удовлетворительное приближение к ЧИКВ, если процесс находится между двумя точками переключения, и определять попадание его в область, где может находиться новая точка переключения. Последнее важно при принятии решений об усилении (ослаблении) карантина и связанных с ним мероприятий.

Как показано в статье, без априорной информации о числе точек переключений найдено пять таких точек и определен характер изменения ЧИКВ (рост, спад, стабилизация) справа от этих точек по небольшому числу наблюдений, имеющихся в этой области. Таким образом, регрессия с переключениями позволяет не только получать краткосрочный прогноз динамики эпидемии, но и оперативно определять направление ее развития.

Отметим, что если даже в самом начале исследования имелось 87 наблюдений, то одновременное нахождение всех точек переключения, когда их число неизвестно, представляло бы сложную задачу. Ее решение упростилось с использованием идеи пошагового оценивания.

СПИСОК ЛИТЕРАТУРЫ

1. Bai J., Perron P. Estimating and testing linear models with multiple structural changes. *Econometrica*. 1998. Vol. 66, N 1. P. 47–78.
2. Bai J., Perron P. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*. 2003. Vol. 18, Iss. 1. P. 1–22.
3. Casini A., Perron P. Structural breaks in time series. Preprint. Boston University, 2018. URL: https://arxiv.org/abs/1805.03807.
4. Bellman R., Roth R. Curve fitting by segmented straight lines. *Journal of the American Statistical Association*. 1969. Vol. 64. P. 1079–1084.

5. Korkhin A.S. Constructing a switching regression with unknown switching points. *Cybernetics and Systems Analysis.* 2018. Vol. 54, N 3. P. 443–455. https://doi.org/10.1007/s10559-018-0045-9.
6. Knopov P.S., Korkhin A.S. Continuous-time switching regression method with unknown switching points. *Cybernetics and Systems Analysis.* 2020. Vol. 56, N 1. P. 68–80. https://doi.org/10.1007/ s10559-020-00222-z.
7. Коронавирусная инфекция (COVID-19). Статистика в Украине. URL: https://www.google.com/search?q=%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0+%D0%BA%D0%BE%D1%80%D0%BE% D0%BD%D0% B0%D0%B2%D0%B8%D1%80%D1%83%D1%81%D0%B0&rlz=1C1AOHY_enUA793UA793&oq=%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0+%D0%BA%D0%BE&aqs=chrome.2.69i59j69i57j35i39j0l3.24667j0j7&sourceid=chrome&ie=UTF-8.
8. Корхин А.С., Минакова Е.П. Компьютерная статистика. Ч. 2. Дніпро: Національний гірничий університет, 2009. 239 с.
9. Knopov P.S., Korkhin A.S. Ch. 1. Estimation of regression model parameters with specific constraints. In: *Regression Analysis under a Priori Parameter Restrictions*. New York: Springer, 2012. P. 1–28.
10. Golodnikov A.N., Knopov P.S., Pepelyaev V.A. Estimation of reliability parameters under incomplete primary information. *Theory and Decision*. 2004. Vol 57, N 4. P. 331–344.
11. Mikhalevich V.S., Knopov P.S., Golodnikov A.N. Mathematical models and methods of riks assessment in ecologically hazardous industries. *Cybernetics and Systems Analysis*. 1994. Vol. 30, N 2. P. 259–273.
12. Ermoliev Yu.M., Knopov P.S. Method of empirical means in stochastic programming problems. *Cybernetics and Systems Analysis*. 2006. Vol. 42, N 6. P. 773–785.
13. Knopov P.S., Pepelyaev V.A. Nonparametric estimate of almost periodic signals. *Cybernetics and Systems Analysis*. 2007. Vol. 43, N 3. P. 362–367.
14. Sergienko I.V., Shylo V.P. Problems of discrete optimization: Challenges and main approaches to solve them. *Cybernetics and Systems Analysis*. 2006. Vol. 42, N 4. P. 465–482.
15. Butenko S., Pardalos P., Sergienko I., Shylo V., Stetsyuk P. Estimating the size of correcting codes using extremal graph problems. In: *Optimization. Springer Optimization and Its Applications*. Pearce C., Hunt E. (Eds.). New York: Springer, 2009. Vol. 32. P. 227–243.
16. Chow G.C. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*. 1960. Vol. 28, N 3. P. 591–605.
17. Кокс Д., Хинкли Д. Теоретическая статистика. Москва: Мир, 1978. 560 с.
18. Корхин А.С., Минакова Е.П. Компьютерная статистика. Ч. 1. Днепропетровск: Национальный горный университет, 2008. 239 с.

Надійшла до редакції 17.07.2020

П.С. Кнопов, А.С. Корхін

СТАТИСТИЧНИЙ АНАЛІЗ ДИНАМІКИ ІНФІКУВАННЯ КОРОНАВІРУСОМ ЗА ДОПОМОГОЮ ПОКРОКОВОЇ РЕГРЕСІЇ З ПЕРЕМИКАННЯМИ

Анотація. Запропоновано моделювати динаміку інфікування коронавірусом з використанням регресії з перемиканнями, точки перемикання якої невідомі. Описано покроковий процес побудови регресії у часі. Досліджено динаміку інфікування коронавірусом в Україні.

Ключові слова: регресія, точки перемикання, параметри регресії, покрокове оцінювання, інфікування, коронавірус.

P.S. Knopov, A.S. Korkhin

STATISTICAL ANALYSIS OF THE CORONAVIRUS INFECTION DYNAMICS USING STEPWISE SWITCHING REGRESSION

Abstract. It is proposed to model the coronavirus infection dynamics using switching regression whose switching points are unknown. The step-by-step process of constructing the regression in time is described. The dynamics of the coronavirus infection in Ukraine is analyzed.

Keywords: regression, switch points, regression parameters, stepwise estimation, infection, coronavirus.

Кнопов Павел Соломонович,
чл.-кор. НАН Украины, заведующий отделом Института кибернетики им. В.М. Глушкова НАН Украины, Киев, e-mail: knopov1@yahoo.com.

Корхин Арнольд Самуилович,
доктор физ.- мат. наук, профессор Приднепровской академии строительства и архитектуры, Днепр, e-mail: a.s.korkhin@gmail.com.