СТЕПАНОВ В.С.,
к.ф.-м. н., с.н.с., ФГБУН Центральный экономико-математический институт РАН, г. Москва, Россия,
e-mail: vladstep0355@gmail.com

СВЯЗЬ ОНКОЛОГИЧЕСКОЙ ЗАБОЛЕВАЕМОСТИ С ВОЗРАСТОМ НАСЕЛЕНИЯ, ПРОЖИВАЮЩЕГО ПРИ НЕБЛАГОПРИЯТНЫХ ФАКТОРАХ ОКРУЖАЮЩЕЙ СРЕДЫ

DOI: 1025881/18110193_2021_3_38

Аннотация.

Исследуется вопрос: можно ли связать частоту онкозаболеваний в регионе с возрастным составом населения, а также — с тремя его эколого-гигиеническими переменными: по уровням загрязнений воздуха и радиационного загрязнения, сбросам загрязнённых сточных вод? Цель исследования — построение линейной регрессионной модели, которая устанавливает связь распространённости такой заболеваемости Y в год t с перечисленными факторами, взятыми с лагами. Один из факторов был выбран качественным, поэтому получилась модель с двумя фиктивными переменными.

Объекты и методы. Объектами исследования были данные Росстата по регионам за пять лет, а также экспертно оцененный уровень радиационного загрязнения их территории. Математическими методами были корреляционный и регрессионный анализ, методы проверки статистических гипотез. Параметры модели оценивались методом наименьших квадратов по пространственно-временной выборке, которая включала переменную Y за 2017, 2018 гг. Расчёты и статистический анализ выполнялись в Excel.

Результаты. Средняя ошибка аппроксимации модели получилась при обучении равной 2,2%; оценка коэффициента детерминации — 91%. На этапе экзамена модель применялась к новым данным с переменной Y за 2019 год. Ошибка оказалась равной 4,3%.

Заключение. Построена и оценена по точности линейная регрессионная модель с переменной структурой, связывающая онкологическую заболеваемость в регионе РФ с индикатором гигиены атмосферного воздуха, показателем загрязнённости его территории сточными водами, фиктивными переменными по радиационной безопасности и долей лиц старших возрастов. На её основе можно по общедоступной статистике оценивать распространенность случаев рака в ряде регионов (с горизонтом в 1 год).

Ключевые слова: *модель регрессии, загрязнение воздуха, загрязнение сточными водами, радиационное загрязнение, злокачественные новообразования, регионы России.*

Для цитирования*: Степанов В.С. Связь онкологической заболеваемости с возрастом населения, проживающего при неблагоприятных факторах окружающей среды. Врач и информационные технологии. 2021; 3: 38-49. doi: 1025881/18110193_2021_3_38.*

STEPANOV V.S.,
PhD, Central Economics and Mathematics Institute of the RAS, Moscow, Russia,
e-mail: vladstep0355@gmail.com

THE RELATIONSHIP OF CANCER PREVALENCE WITH AGE OF THE POPULATION LIVING UNDER ADVERSE ENVIRONMENTAL FACTORS

DOI: 1025881/18110193_2021_3_38

Abstract.

Is it possible to link the prevalence of cancer cases in a Russian region with the age of its population and several environmental variables, specifically by levels of atmospheric air pollution, emissions of radioactive elements, and discharges of contaminated wastewater? The aim of the study is to build a linear regression model that links the oncologic incidence of the population Y within the region in year t with the listed factors taken with lags. One of the variables was qualitative and therefore a variable structure model was obtained.

Objects and methods: The objects of the study were panel data for Russia's regions from the past five years, as well as the expertly estimated variable on the radiation pollution of its territories. The mathematical methods used were multivariate regression analysis for data with a few dummy variables. The model parameters were estimated by the ordinary least square method based on a spatiotemporal sample from the panel data, which included the variable Y for 2017 and 2018. Calculations and statistical data analysis were performed in MS-Excel.

Results: The mean relative error for the model was equal to 2.2%. Then, at the exam stage, the model was applied to new data, where the variable Y was for 2019. The error on the exam was equal to 4.3%.

Conclusion: A linear regression model with a variable structure was built and evaluated in terms of accuracy, linking cancer prevalence in Russia's regions with atmospheric air hygiene, an indicator of the pollution of its territory with wastewater, two fictitious variables for radiation safety and the proportion of older people. On this basis, it is possible to estimate the prevalence of cancer cases in a number of Russia's regions (with a horizon of one year).

***Keywords**: regression model, air pollution, contaminated waste water, radiation pollution, malignant neoplasms, Russia's region*

***How to cite**: Stepanov V.S. The relationship of canser prevalence with age of the population living under adverse environmental factors. Medical doctor and information technology. 2021; 3: 38-49. (In Russ.). doi: 1025881/18110193_2021_3_38.*

Развитие у граждан злокачественных новообразований (ЗНО) — одна из девяти социально-значимых болезней, смертность от которых стоит на втором месте в структуре смертности россиян, вслед за болезнями системы кровообращения: 16,4% против 46,8% в 2019 г. [1]. Также онкологические заболевания печально возглавляют эту девятку по ежегодному приросту пациентов. Заболеваемость (контингент страдающих ЗНО) увеличилась с 2005 г. за полтора десятка лет в 1,5 раза: с 1680 до 2677 пациентов на 10^5 россиян [1; 2]. Из сказанного ясно, что проблема борьбы с ЗНО весьма актуальна для страны.

По определению, контингент пациентов включает совокупность всех лиц, страдающих ЗНО с любой локализацией и обратившихся в лечебно-профилактическую организацию Минздрава в регионе в году t , как и в предыдущие годы [2]. Этот контингент в расчёте на 100 тыс. жителей был выбран нами в качестве зависимой переменной Y в модели регрессии. Очевидно, что на величину Y влияет не только заболеваемость ЗНО, но и качество диагностических процедур, а также уровень медицинской помощи в онкологии и частота онкологической смертности. На роль независимых факторных переменных X_1 ,..., X_4 были выбраны три эколого-гигиенические характеристики региона и один показатель по демографии. Если иметь в арсенале такую модель, имеющую приличную точность, то можно снижать заболеваемость ЗНО, воздействуя на управляемые переменные. Модель полезна при разработке противораковых программ; её также можно использовать в расчётах экологических платежей за вред, нанесенный здоровью [3; 4].

Далее обзорно опишем пять работ, где применялись примерно аналогичные методы многомерной статистики [5–9]. В [10] подробно рассмотрены основные этапы их применения и возможности использования в различных областях медицины. В первых двух анализировались региональные данные, а в оставшихся трёх — персонифицированные базы данных для пациентов с одной из форм ЗНО. Сначала использовалась пара программ *Statistica* и *Excel* [5]. Были взяты данные по всем регионам страны за 2015 г.; для каждого региона было взято более двух десятков показателей Росстата, включая заболеваемость формами ЗНО. Предварительно переменные были сгруппированы в ряд логических блоков: *а*) “обеспечение медицинской помощью”; *б*) “загрязнение окружающей среды”; *в*) “заболеваемость населения” и др. Затем набор показателей (переменных) каждого блока сворачивался линейным преобразованием в одну новую факторную переменную, называемую “первая главная компонента”. Затем на их основе была построена линейная регрессионная модель, в которой зависимой переменной была первичная заболеваемость ЗНО в регионе в 2015 году [5; таб. 6]. Соответствующий коэффициент детерминации модели получился 70%. В качестве пятёрки переменных блока “*б*)” выбирались удельные объёмы загрязнений: выбросы в окружающую среду региона от его стационарных источников загрязнения; сбросы загрязнённых сточных вод; объёмы образовавшихся отходов. Кроме того, выполнен кластерный анализ иерархического типа, в ходе которого строилась дендрограмма; для неё бралось пять экологических переменных и шестёрка показателей заболеваемости [5]. В итоге регионы разбились на четыре группы; причём заболеваемость была выше в тех группах, где хуже экологическая ситуация.

Во второй работе регрессионный анализ в среде *RStudio* применялся к оценке первичной заболеваемости ЗНО со всеми локализациями [6]. Объектом исследования были панельные данные за 11 лет по 78 регионам РФ. В итоге было построено несколько регрессионных моделей, в которых показатель онкопатологии играл роль зависимой переменной Y. В состав независимых факторов входила пятёрка показателей Росстата, которые весьма приближённо оценивают антропогенную нагрузку на регион (доля городского населения, средняя обеспеченность автомобилями, среднее потребление свежей воды в регионе и два удельных показателя для стационарных источников загрязнения атмосферы: по выбросам загрязняющих её веществ и по эффективности борьбы с ними). Здесь также было предложено изучить онкологическую обстановку по первичной заболеваемости в региональном разрезе по всем регионам, а также и по группам регионов, сформированным после кластерного анализа. В итоге была выбрана модель панельных данных с детерминированными эффектами (без учёта последнего

фактора), которая объяснила 63% дисперсии *Y*. После применения одного из алгоритмов кластерного анализа было получено 4 класса регионов по первичной заболеваемости ЗНО (но последняя группа содержала лишь два региона с "выпадающими" данными). Далее исследовалась зависимость показателя заболеваемости от экологических факторов для регионов каждого кластера. В итоге были построены три линейные регрессионные модели по панельным данным с детерминированными эффектами (с разными вариантами включения в модель перечисленных выше 4-х факторов). Эти модели объясняли: в 1-й группе регионов — 69% дисперсии *Y*, а также 74% — во 2-й и 44% — во 3-й [6].

В третьей работе для оценки вероятности развития колоректального рака (КРР) среди жителей Пермского края были построены *logit*-модели вероятности развития ЗНО в зависимости от влияния набора медико-социальных и средовых факторов [7]. В этой статье и паре следующих работ статистические пакеты использовались ещё и *как инструмент медико-биологического познания* [8]. Сначала были сформированы две группы лиц: группа "случай" включала 204 пациента с КРР, верифицированным по гистологии, а "контрольная" — 205 здоровых. Наличие возможных 33-х факторов риска оценивалось для каждого пациента в интервью. В качестве зависимой переменной использовался код группы: 1 — если человек был болен КРР, иначе — 0. Факторы риска, объясняющие переменные в логистической регрессии, чаще всего были бинарными, номинальными, качественными, а 7 переменных измерялись здесь по интервальной шкале. Сначала через процедуру включения-исключения факторов была снижена размерность, а затем специфицировано несколько вариантов *logit*-модели для пациентов обоего пола [7; табл. 2]. Затем все аналогично выполнялось для каждого пола. В итоге модель вероятности развития КРР у мужчин включала 14 средовых факторов риска, а модель для женщин — 10. Были установлены значительные различия: как в наборе, так и в степени значимости факторов риска КРР среди мужчин или женщин [7].

В четвёртой работе статистические методы из пакета *IBM SPSS-13* применялись к данным по заболеваемости раком желудка (РЖ) [8]. Было взято почти 2500 жителей Таганрога, для каждого из которых было рассмотрено p = 107 первичных потенциальных факторов риска РЖ (т.е. возможных причин, способствующих его развитию). Затем вычислялась корреляционная матрица размерности $p \cdot p$ и применялся факторный анализ (метод главных компонент) и выполнялась процедура снижения размерности. После чего, с использованием комбинации методов главных компонент и кластерного анализа были построены, проинтерпретированы и наглядно представлены три главных компоненты $Y_1,...Y_3$, что объяснили 84,5% суммарной дисперсии [8]. Компонента Y_1 связана с особенностями питания: всего 9 показателей по диете. Другая переменная Y_2 связана со злоупотреблением пациентом алкоголем и курением, наличием в его анамнезе хронических заболеваний желудка и других отделов пищеварительного тракта. Компонента Y_3 объединяет четыре первичных фактора риска РЖ; из них два — по наследственности и два показателя качества питьевой воды: по санитарно-химическим параметрам, а также использование воды из водопровода без её доочистки. В итоге приводятся весовые множители для каждого из важных факторов риска РЖ.

Наконец, в последней работе многомерные статистические методы применялись к данным по заболеваемости РЖ в Алтайском крае [9]. Были сформированы две группы пациентов: "случай" (или основная) включала 667 пациентов с РЖ, а "контрольная" — 50 пациентов без любых форм ЗНО. Для каждого пациента измерялось p = 131 потенциальных факторов риска развития РЖ. Группа "случай" была сформирована в рамках дискриминантного анализа (метода, близкого к логистической регрессии); параллельно рассчитывалась корреляционная матрица размерности $(p+1) \cdot (p+1)$. В ней находились коэффициенты парных корреляций между каждым фактором и индикатором РЖ, затем выявлены факторы риска, которые выше 0.3 по модулю [9] Таким образом, размерность p была снижена до 25 информативных для риска РЖ факторов. Для выявления различий между этими двумя группами строились линейные комбинации этих 25-ти переменных (факторов риска). Они задаются через линейную дискриминантную функцию, коэффициенты которой приводятся в [9; таб.1]. Целью этой работы было построить такую функцию, на

основе которой оценивается вероятность наличия РЖ у пациента по всем 25 информативным факторам. Точность построения этой функции проверялась как на материале обучения, так и методом "скользящий экзамен"; уровень ошибок оказался равным 5%. Аналогичные исследования авторы провели с другими основными формами ЗНО, создали региональный регистр с почти 10 тыс. пациентов. В результате заметно возросла эффективность диспансеризации населения и улучшилась ранняя диагностика онкологических заболеваний ЗНО. Кроме того, в познавательном плане интересны выявленные веса факторов.

Также следует отметить, что в большинстве работ исследователи чаще анализируют не распространённость ЗНО, а первичную или общую онкологическую заболеваемость (или в "грубом", или в "стандартизованном" виде). Намного реже рассматривается распространённость ЗНО, называемая также "болезненностью" [11], [12; Рис. 2]. Для неё оказывается, что после подгонки тренда к региональным точкам за последние 15 лет, зависимость от времени (по годам с номером отсчёта t) имеет с высокой точностью вид трёхчлена $Y = at^2+bt+c$ или, явно реже, — прямой линии: $Y = bt+c$ [2; 12].

В [13] предложена долгосрочная стратегия борьбы с раком; первым её звеном является *первичная профилактика*. Она является важнейшей компонентой стратегии, требуя лишь модификации образа жизни и минимума денежных затрат. Ещё одним звеном профилактики является создание персонифицированных cancer-регистров, с их последующей обработкой алгоритмами многомерного статистического анализа [9]. Это позволяет значительно повысить эффективность диспансеризации населения в плане ЗНО и улучшить раннюю их диагностику. Из изложенного далее следует, что в ряде регионов также необходимо *улучшать качество экологии*: снижать уровень загрязнений воздуха, усиливать степень очистки сточных вод и уменьшать объёмы их сброса (это повышает безопасность питьевой воды, снижает загрязненность продуктов), улучшать радиационную безопасность жизни.

Рост заболеваемости ЗНО связан, кроме состояния окружающей среды, со старением граждан и перечнем других показателей [11; 14; 15]. Заболеваемость ЗНО часто бывает выше в крупных промышленных центрах, на территориях с высокой долей пожилых [11]. Так, если взять из каждого региона лишь "долю лиц старшего возраста" к концу предыдущего года X_1^{t-1} , а затем подогнать к нашим обучающим данным модель простой регрессии, то получится уравнение $Y^t = 567{,}1+83{,}085 \cdot X_1^{t-1}$. Оно объясняет почти 58% разброса переменной Y^t вокруг её среднего значения. Как показано ниже, дополнительное использование тройки эколого-гигиенических переменных X_2 ,..., X_4, включённых в соответствующее уравнение множественной регрессии, позволяет повысить эту долю разброса Y^t уже до 91%. Переменная X_4 была ординального типа, поэтому получается модель с переменной структурой.

Далее чуть подробнее рассмотрим эти три фактора окружающей среды и их связь с развитием ЗНО у граждан РФ. Во многих регионах РФ довольно острой проблемой является доступность доброкачественной, на 100% безопасной *питьевой воды*; особенно для сельских поселений. Также свыше ½ сельского населения РФ использует колодезную воду, качество которой, особенно по загрязнению нитратами, нередко хуже, чем у воды из централизованных сетей. Чем сильнее загрязнена вода в поверхностных водных объектах, в том числе из-за сброса нормативных стоков, тем более высокие концентрации соединений хлора будут, скорее всего, применяться при водоподготовке (на станциях водоочистки). А для таких органов человека как ободочная и прямая кишка, мочевой и желчный пузырь, операция хлорирования воды является канцерогенной [16]. Методом статистического анализа В.М. Боев с соавторами недавно установили, что риск развития ЗНО в ободочной кишке тесно связан с концентрациями хрома и кадмия в питьевой воде. Эта локализация рака весьма важна и для мужчин, и для женщин, особенно с учётом риска смертности [1]. С другой стороны, китайскими учёными доказано, что ржавчина на трубах, изготовленных из железа, легированного хромом, может вступать в химическую реакцию с остаточными соединениями хлора, что имеется в водопроводных сетях. В результате в питьевой воде возникает канцероген Cr^{VI}. Отсюда призывается сократить использование таких труб либо защищать их от коррозии,

использовать в процессе водоподготовки менее активные средства [17]. Этот вывод согласуется с рядом российских региональных данных по указанной форме рака [1]. Кроме того, сброс загрязненных сточных вод ухудшает гигиену почв, что негативно влияет на сельскохозяйственную продукцию, а поэтому — здоровье граждан.

Кроме водного фактора, риск развития ЗНО связан с загрязнениями атмосферного *воздуха* [14; 15; 18]. Из-за таких загрязнений могут порождаться до 30% случаев появления ЗНО среди жителей промышленных регионов [14]. Дополнительно отметим, что загрязнение атмосферы сильно влияет и на риск развития болезней системы кровообращения и других классов болезней [16]. Поэтому для снижения рисков развития ЗНО и других заболеваний, а также для повышения качества жизни в 12 городах и близлежащих поселениях, выделены большие средства на проект "Чистый воздух" [18]. Первые три фактора X_1,..., X_3 брались не синхронно с Y, а с небольшим отставанием, причём с разными временны*м*и лагами. На этот момент, применительно к заболеваемости ЗНО, ранее уже обращалось внимание [19].

Канцерогенность ионизирующей *радиации* была показана в эпидемиологических исследованиях среди различных групп населения (после аварий на "Маяке", на ряде АЭС, при облучении на производстве, после бомбардировок в Японии) [14; 20]. Известно, что рак может возникать при воздействии сравнительно невысоких доз облучения через 10–20 лет после воздействия или позднее. Более того, учёными из Великобритании в опытах на мышах недавно установлено, что даже малые дозы радиации, эквивалентные дозе трёх компьютерных томографий, могут дать раковым клеткам конкурентное преимущество перед здоровыми клетками [21]. Такие уровни излучения увеличивают количество клеток, имеющих мутации в гене *p*-53. В организме человека имеются мутантные клетки, способные породить ЗНО (в том числе и с мутациями в *p*-53), причём по мере старения человека число их увеличивается. С возрастом его иммунная система ослабевает и поэтому хуже борется с мутантными клетками, так что риск развития рака возрастает.

Итак, ниже будет строиться линейная регрессионная модель, связывающая контингент лиц с ЗНО в регионе с долей населения старших возрастов и комплексом наблюдаемых в нём эколого-гигиенических факторов. Почти все эти переменные также оказались значимыми в модели распространения онкологических заболеваний [15]. Авторы этой работы нашли корреляции между заболеваемостью ЗНО и каждым из многих показателей среды обитания. Затем на основе методов теории информации было установлено, что сильнее всего на частоту появления ЗНО повлияли: демографическая структура (фактор X_1); качество питьевой воды; сброс в природу загрязненной воды (X_3), загрязнение атмосферного воздуха (X_2) [15].

Целью исследования является построение (с оценкой по точности) модели линейной регрессии, которая связывает распространённость заболеваемости ЗНО в регионе России в год t с рядом факторов, взятых с небольшими лагами.

ОБЪЕКТЫ И МЕТОДЫ

Объектами были ежегодные данные Росстата по регионам за период 2014–2019 гг., а также переменная по степени радиационного загрязнения территории регионов, экспертно оцененная после изучения открытых литературных источников. Основными математическими методами были корреляционный и регрессионный анализ, методы теории проверки гипотез. Все расчеты и статистический анализ выполнялись средствами Excel.

В материал обучения или выборку из информационной базы, на основе которого оценивались неизвестные параметры, зависимая переменная Y входит за 2017–18 годы [2; таб. 2.5]. Все необходимые данные имеются в сборниках Росстата, в открытом доступе на сайте (факторы X_2 и X_3 легко рассчитываются по публикуемым данным):

- доля лиц старше трудоспособного возраста, или X_1 (в [%]), [22];
- характеристика состояния атмосферного воздуха в населенных пунктах — доля исследованных проб воздуха, где превышена предельно допустимая концентрация (ПДК) вредных веществ, или X_2 , (в [%]), [23; таб. 3.6] ;
- $X_3=(V_{зсв}/S)^{0.5}$ — квадратный корень из удельного сброса загрязненных сточных вод в поверхностные водные объекты региона [23], где $V_{зсв}$ — объём сброса этих вод в году t ,

Таблица 1 — Значения двух фиктивных переменных

Регион / Region	d_1	d_2	Регион / Region	d_1	d_2
Алтайский край / Altay Territory	0	1	Омская / Omsk Region	1	0
Архангельская / Arkhangelsk Region	1	0	Орловская область / Orel Region	0	1
Брянская / Bryansk Region	0	1	Мурманская / Murmansk Region	0	1
Воронежская / Voronezh Region	1	0	Новосибирская / Novosibirsk Region	0	1
Забайкальский край / Trans-Baikal Territory	1	0	Нижегородская / Nizhny Novgorod Region	0	1
Ивановская / Ivanovo Region	1	0	Пензенская / Penza Region	1	0
Калужская / Kaluga Region	0	1	Рязанская / Ryazan Region	1	0
Курганская / Kurgan Region	1	0	Тверская / Tver Region	1	0
Курская / Kursk Region	0	1	Томская / Tomsk Region	1	0
Липецкая / Lipetsk Region	1	0			

[млн. м^3], S — площадь территории региона в [тыс. км^2], за вычетом площади его лесов, дорог и населенных пунктов [22]). При этом в числителе в X_3 берётся сумма объемов вод, сброшенных "вообще без очистки", вместе с загрязнёнными водами, очищенными слабее норматива, причём объёмы последних делились нами на число 4,5 ;

– X_4 — качественная переменная с тремя градациями: "0", "1", "2", характеризующими уровень радиационного загрязнения территории после различных причин (по данным из [20]) (здесь "0" — ситуация в регионе, сравнительно благополучна, "1" — явно хуже, "2" — самая острая; далее вместо X_4 используется пара фиктивных переменных d_1 , d_2; см. таб. 1).

Модель оценивалась не по всему набору регионов информационной базы, сначала был исключён ряд из них. Почти во всех исключенных регионах общие коэффициенты смертности в 2018 г. были относительно низкими: слабее 26-го места [2; таб. 2.9]; ещё в таких регионах нередко проживает повышенная доля молодежи [22]. Также оказалось, что ¾ регионов, включённых нами в материал обучения, имеют долю этноса "русские" среди населения выше, чем 90% [24]. И ещё почти все эти регионы попали в кластеры 1, 2 из работы [6]. В обучающую таблицу вошло 75 строк с наборами переменных, относящихся к республикам: Карелия, Крым, Татарстан, Удмуртия; краям: Алтайский, Забайкальский, Краснодарский, Пермский, Приморский, Хабаровский и многим областям. Они часто были в составе ЦФО, реже — СЗФО, ПФО, ЮФО и совсем редко в УФО, СФО, ДФО (Амурская, Архангельская, Белгородская, Брянская, Владимирская, Волгоградская, Вологодская, Воронежская, Ивановская, Иркутская, Калужская, Кировская, Костромская, Курганская, Курская, Липецкая, Московская, Мурманская, Нижегородская, Новгородская, Омская, Орловская, Пензенская, Псковская, Ростовская, Рязанская, Самарская, Саратовская, Сахалинская, Тамбовская, Тверская, Томская, Ярославская). Итак, оценивается линейная модель регрессии

$$Y = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_4 \cdot X_4 + \varepsilon, \quad (1)$$

где: Y — распространённость страдающих ЗНО в регионе в году t ; X_1,..., X_4 — вышеописанные факторы; ε отражает влияние на Y неучтённых переменных; β_0,..., β_4 — неизвестные параметры, оцениваемые по выборке. Величина ε определяется как случайная с нулевым средним и дисперсией d^2, которая неизвестна. Оценки b_0,..., b_5 для β-параметров и d^2 находятся по методу наименьших квадратов (МНК) после обработки таблицы с обучающими данными средствами программы *Excel* [25]. Эта таблица содержит в каждой строке набор переменных Y, X_1,...., X_4 для региона России, причём значения факторов X_1,...., X_3 входят в неё с небольшим отставанием относительно t — года наблюдения заболеваемости Y (оно равно от 1-го до 3-х лет). При этом категории переменной X_4 кодируются парой индикаторных переменных d_1, d_2 : "0"– (0,0), "1" — (1,0), "2" — (0,1). Переменные Y, X_1,...., X_3 имеют индекс, привязанный к году t ; его нет в выражении (1) для упрощения записи. Выбирая лаги по t, мы предполагали, что факторы X_2 , X_3 способствует росту заболеваемости ЗНО не сразу, а через ряд лет относительно

момента своего воздействия. Это связано с тем, что продолжительность скрытого периода между воздействием канцерогена на организм и развитием ЗНО зависит от ряда причин (возраст, пол и др.) [11; 19]. В таблицу обучения вошло $n = 75$ строк; около половины соответствовало Y^t за 2017 г., а все прочие были — 2018 г.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Исходя из выше описанного, формула (1), с подстановкой оценок b_j вместо неизвестных параметров β_j, явным обозначением всех выбранных лагов и введением пары dummy- переменных, будет иметь вид

$$Y^t = b_0 + b_1 \bullet X_1^{t-1} + b_2 \bullet X_2^{t-2} + b_3 \bullet X_3^{t-3} + b_4 \bullet d_1 + b_5 \bullet d_2 , \quad (2)$$

где: Y^t (или Y) — контингент больных со ЗНО к концу года t в некотором регионе, который имеет значения X_1, ..., X_3, взятые с соответствующими лагами относительно текущего года t, а b_j — оценка параметра β_j модели (1).

В таб. 2, в 1-й строке приводятся значения для b_0,..., b_5 из (2) и несмещённая оценка d; во 2-й и 3-й строке — значения t-статистик при проверке гипотез Ho: "$\beta_j = 0$" (для каждого j), а также достигнутые уровни значимости (p-values, где 0,001* означает ниже 0,001); в 4-й и 5-й строке — результаты дисперсионного анализа (статистика Фишера, число степеней свободы и достигнутый уровень значимости при проверке Ho: "$\beta_j = 0$", для всех $j = 1, ..., 5$) [25].

Оценка коэффициента детерминации R^2 была 91,8%, а её подправленное, несмещённое значение R_{adj}^2 около 91% [25]. Применение модели (2) к 75ти наблюдениям обучающих данных даёт среднее значение модуля относительной ошибки, равное 2,2%. Эта ошибка равна, по определению: $Err = 100 \bullet (Y_{est} - Y_f) / Y_f$, то есть — относительной разности между оценкой Y_{est}, полученной через (2) и фактическим значением Y_f контингента страдающих ЗНО к концу года t, (%) [2]. Для варианта её расчета через (2) по 38 регионам России, имеющим значения Y_f за 2019 год, она равна 4,3% [26].

Далее сначала исследуется вопрос мультиколлинеарности. Этим термин обозначают сильную корреляцию между факторами; её наличие ухудшает качество оценок параметров [25]. Этот вопрос, как и другие проверки ("есть ли гомоскедастичность?", "есть ли автокорреляция?" и др.), важен для обоснования корректности использования МНК- похода к оцениванию параметров. То есть желательно выяснить, выполняются ли все "предпосылки МНК" (или условия теоремы Гаусса-Маркова, которая доказывает, в каких случаях МНК- оценки обладают хорошими свойствами [25]). Поэтому ниже кратко описана проверка остатков из (2) на гомоскедастичность, а также приводятся результаты для критериев согласия распределения остатков с нормальным законом, излагается результат теста на автокорреляцию.

Из анализа матрицы корреляций $\boldsymbol{R}$ переменных следует, что главным фактором, определяющим распространённость ЗНО в регионе, является "доля лиц старшего возраста". Мультиколлинеарности не выявляется, ибо наблюдаются немалые корреляции r_{Yj} для пар (Y, X_j) и одновременно низкие r_{ij} — между факторами. Изучение диаграммы рассеяния остатков (1) в случае расположения Y^t по оси абсцисс, показывает, что разброс остатков, откладываемых здесь по ординате, был примерно одинаковым. Отсюда можно предполагать, что сильного нарушения

Таблица 2 — Результаты оценивания параметров и статистики критериев

	Оценки параметров / Estimators of all parameters	b_0	b_1	b_2	b_3	b_4	b_5	
	Обозначение / Short name		Age	Air	Water	d_1	d_2	d 78,2373
1	Значения / Its values	318,13	80,019	73,056	66,026	124,99	331,92	
2	t-статистики / t-statistics	(3,03)	(21)	(5,5)	(12,8)	(5,8)	(13,2)	
3	p-значения / p-values	0,0034	0,001*	0,001*	0,001*	0,001*	0,001*	
4	F-статистика / F-statistic = 154,3 ; Степени свободы / Degrees of freedom v = 69							
5	p-значение / p-value 0,001* <0,001							

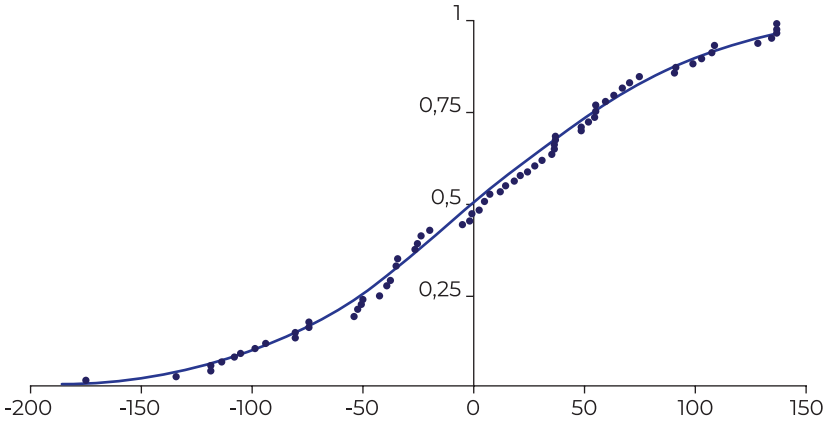

Рисунок. 1 — Эмпирическая функция распределения остатков (2).

условия гомоскедастичности нет [25]. Нормальность их распределения изучалась по критериям согласия χ^2 , w^2 и Лиллиефорса [27]. Верхняя 5%-ная точка Q для распределения χ^2 с параметром v = 6 равна 12,6 и поэтому наблюдаемое значение 2,44 для χ^2-статистики не попадает в критическую область (при разбиении диапазона 75-ти остатков на 8 интервалов, в каждый из которых попало от 9 до 10 наблюдений). Поэтому гипотеза о том, что остатки модели распределены нормально (со средним 0 и неизвестной дисперсией d^2) по критерию χ^2 не отвергается. Тоже относится и к двум прочим критериям, статистика w^2 была равна 0,027, а максимальное расстояние между функциями распределения равно 0,053 (Рис. 1; ось ординат – вероятность, абсцисс – число больных со ЗНО на 10^5 жителей).

Отсутствие автокорреляции 1-го порядка проверялось *DW* критерием Дарбина –Уотсона [25]. Значение статистики *DW* = 2,05 не попадает в критические интервалы для этого теста при уровне значимости α = 5%. Отсюда гипотеза H_0: "коэффициент автокорреляции остатков ρ = 0", не отвергается. Исходя из выше изложенного, все предпосылки теоремы Гаусса-Маркова выполнены, поэтому оценивание параметров посредством МНК корректно.

Для иллюстрации, наконец, приведём пример для одного из регионов с данными за 15 лет из [2; 26] (Таб. 3). Если по первым 14-ти отсчётам подогнать по МНК квадратичную параболу $Y = at^2 + bt + c$, то получатся значения a = 1,4205, b = 63,053 *и* c = 1820,3; это объясняет 97,2% разброса Y вокруг своего среднего. Если подставить значение следующего 15-го отсчёта, то оценка Y для 2019 г. равна 3071.7, что превышает значение из [26] примерно на 1,4%.

Значение X_1 для этой области за 2018 г. равно 29,3% [22; таб. 2.5]; а X_2 для неё в 2017 г. равно 0,4417; X_3 для неё было в 2016 г. равно $(5,226)^{0,5}$ или 2,286. Наконец из-за наличия АЭС на её территории и того факта, что территория региона несколько пострадала после давней аварии на ЧАЭС, радиационная обстановка была оценена нами "1", что кодируется так: d_1 = 1, d_2 = 0. Перемножение этих пяти значений согласно (2) и Таб. 2 даёт оценку Y на 2019 г., равную 2971,4, что ниже значение значения из [26] на 1,9%. Если подставить другие значения d_1 = 0, d_2 = 1, то оценка Y получается выше, чем 3029.7, на 4,9%. Отсюда можно предполагать, что уровень радиационного загрязнения в Воронежской обл. был слегка выше, чем тот уровень, что кодируется нами в данных меткой "1".

Таблица 3 — Данные по распространённости ЗНО в Воронежской обл.: 1-2005, ..., 15-2019

t	1	2	3 (2017)	4	5	6	7	8
Y	1914,5	2000,5	–	1946,4	2220,6	2175,9	2361,6	2462,3
t	9	10	11	12	13	14	15	
Y	2480,7	2632,5	2708,2	2816,6	2865,2	2938,8	3029,7	

ЗАКЛЮЧЕНИЕ

Построена и оценена по точности линейная регрессионная модель, связывающая онкологическую заболеваемость жителей региона с парой показателей по загрязнениям воздуха и сбросам загрязненной воды, долей граждан старшего возраста, фиктивными переменными по радиационной безопасности. На её основе можно по открытой статистике Росстата оценивать распространенность ЗНО в ряде регионов к концу года *t*, с горизонтом в 1 год. Показана важность борьбы за снижение загрязнений воздуха в населенных пунктах, улучшение очистки загрязнённых сточных вод, снижение объёма их сброса, а также — повышение радиационной безопасности жизни.

ЛИТЕРАТУРА/REFERENCES

1. Злокачественные новообразования в России в 2019 г. / Под ред. Каприна А.Д., Старинского В.В., Шахзадовой А.О. М.: Изд-во МНИОИ им П.А.Герцена, 2020. — 252 с. [Malignant neoplasms in Russia at 2019 year. M.: MNIOI. 2020. 252 p. (In Russ).]
2. Здравоохранение в РФ. Приложение по регионам. М.: Росстат, 2007–2019. [Healthcare in the Russian Federation: information by constituent entities of the Russian Federation. Statistical collection. M.: Rosstat, 2007–2019 (In Russ).]
3. Заботина А. Организация медицинской помощи онкологическим больным в России // Эксперт. 2020. [Zabotina *A.* Organization of medical care for cancer patients in the Russian Federation. Ekspert. 2020. (In Russ).] Доступно по: https://expertnw.com/naglyadno/otchyet-po-nozologii-onkologiya. Ссылка активна на 12.04.2021.
4. Дядик В.В., Дядик Н.В., Ключникова Е.М. Экономическая оценка ущерба здоровью населения от негативных экологических воздействий: обзор основных методологических подходов // Экология человека. — 2021. — №2. — С.57-64. [Dyadik VV, Dyadik NV, Klyuchnikova EM. Economic assessment of environmental effects on public health: a review of methods. Human Ecology. 2021; 2: 57-64. (In Russ).] doi: 10.33396/1728-0869-2021-2-57-64.
5. Мирасова В.М., Малыгина Н.В. Определение зависимости уровня заболеваемости граждан регионов РФ от состояния окружающей среды с помощью многомерных статистических методов // XXI век: итоги прошлого и проблемы настоящего плюс. — 2017. — №1(35). — С.58-66. [Mirasova VM, Malygina NV. Determination of dependence of the incidence rate of citizens in the regions of the Russian Federation on the environment state by means of multivariate statistical methods. XXI vek: itogi proshlogo i problemy nastoyashchego plyus = XXI century: the results of the past and problems of the present plus. 2017; 1(35): 58-66. (In Russ).]
6. Емцева Е.Д., Кику П.Ф., Мазелис А.Л. Использование методов многомерного статистического анализа для оценки динамики заболеваемости онкологическими новообразованиями // Экология человека. — 2019. — №2. — С.45-51. [Emtseva ED, Kiku PF, Mazelis AL. Assessment of temporal trends of malignant neoplasm using multivariate statistical analysis. Ekologiya cheloveka. 2019; 2: 45-51. (In Russ).] doi: 10.33396/1728-0869-2019-2-45-51.
7. Фельдблюм И.В., Алыева М.Х., Радионова М.В. Комплексное влияние медико-социальных и средовых факторов риска на вероятность развития колоректального рака // Тихоокеанский медицинский журнал. — 2018. — №3(73). — С.24-28. [Feldblyum IV, Alyeva MH, Radionova MV. Complex impact of medico-social and environmental risk factors on probability of colorectal canser development. Tikhookeanskiy meditsinskiy zhurnal. 2018; 3(73): 24-28. (In Russ).] doi: 10.17238/PmJ1609-1175.2018.3.24-28.
8. Айдинов Г.Т., Марченко Б.И., Софяникова Л.В., Синельникова Ю.А. Применение многомерных статистических методов при выполнении задач совершенствования информационно-аналитического обеспечения системы социально-гигиенического мониторинга // Здоровье человека и среда обитания. — 2015. — №7(268). — С.4-8. [Aydinov GT, Marchenko BI, Sofyanikova LV, Sinelnikova YuA. Application of multivariate statistical methods in the tasks of improving of information and analytical providing of the socio-hygienic monitoring system. Zdorov'ye naseleniya i sreda obitaniya. 2015; 7(268): 4-8. (In Russ).]

9. Лазарев А.Ф., Петрова В.Д., Терехова С.А., Синкина Т.В. Многофакторный анализ при формировании групп высокого онкологического риска // Бюллетень медицинской науки. — 2017. — №1(5). — С. 37-43. [Lazarev AF, Petrova VD, Terekhova SA, Sinkina TV. Multivariate statistical analysis when forming groups of high oncological risk // Byulleten meditsinskoy nauki. 2017; 1(5): 37-43. (In Russ).] doi: 10.31684/2541-8475.2017.1(5).37-43.
10. Карякина О.Е., Добродеева Л.К., Мартынова Н.А., Красильников С.В., Карякина Т.И. Применение математических моделей в клинической практике // Экология человека. — 2012. — №7. — С.55-64. [Karyakina OE, Dobrodeeva LK, Martynova NA, Krasilnikov SV, Karyakina TI. Use of mathematical models in clinical practice. Human Ecology. 2012; 7: 55-64. (In Russ).]
11. Сулейманов Р.А., Бакиров А.Б., Валеев Т.К., Давлетнуров Н.Х., Степанов Е.Г., Туктарова И.О. Анализ заболеваемости и смертности населения Республики Башкортостан злокачественными новообразованиями // Медицина труда и экология человека. — 2019. — №2. — С.14-23. [Suleimanov RA, Bakirov AB, Valeev TK, Davletnurov NKh, Stepanov EG, Tukhtarova IO. Analysis of morbidity and mortality of the population of the Republic of Bashkortostan malignant neoplasmsю. Meditsina truda i ekologiya cheloveka. 2019; 2: 14-23. (In Russ).] doi: 10.24411/2411-3794-2019-10016.
12. Басова О.М., Басов М.О., Исаев Н.И. Оценка гигиенических факторов риска онкологической заболеваемости в условиях малых промышленных городов // Анализ риска здоровью. — 2013. — №3. — С.34-40. [Basova OM, Basov MO, Isaev NI. Assessment of hygienic risk factors of oncologic diseases in conditions of small industrial towns. Health Risk Analysis. 2013; 3: 34-40. (In Russ).]
13. Суконко О.Г., Красный С.А. Роль научных исследований в улучшении онкологической службы и направления дальнейшего совершенствования медицинской науки // Онкоурология. — 2015. — Т.11. — №2. — С.14-22. [Sukonko OG, Krasny SA. Role of researches in improving a cancer care service and a direction for further improvement of medical science. Cancer Urology. 2015; 11(2): 14-22. (In Russ).] doi: 10.17650/1726-9776-2015-11-2-14-22.
14. Мешков Н.А. Приоритетные факторы риска окружающей среды в развитии онкопатологии // Научный альманах. — 2016. — Т.5. — №3. — С.309-318. [Meshkov NA. Major environmental risk factors for canser development. Nauchnyy al'manakh. 2016; 5(3): 309-318. (In Russ).] doi: 10.17117/na.2016.05.03.309.
15. Веремчук Л.В., Кику П.Ф., Жерновой М.В. Системное моделирование экологической зависимости распространения онкологических заболеваний в Приморском крае // Бюллетень физиологии и патологии дыхания. — 2011. — №41. — С. 48-53. [Veremchuk LV, Kiku PF, Zhernovoi MV. System modeling of ecological dependence in distribution of oncologic diseases within the Primorye Territories. Byulleten' fiziologii i patologii dykhaniya. 2011; 41: 48-53. (In Russ).] Доступно по: https://cfpd.elpub.ru/jour/article/view/411/389. Ссылка активна на 07.05.2021.
16. Кику ПФ, Бениова СН, Морева ВГ, Горборукова ТВ, Измайлова ОА, Сухова АВ, Сабирова КМ, Богданова ВД. Эколого-гигиенические факторы и распространенность болезней системы кровообращения // Здравоохранение Российской Федерации. — 2019. — Т.63. — №2. — С. 92-97. [Kiku PF, Beniova SN, Moreva VG, Gorborukova TV, Izmaylova OA, Sukhova AV, Sabirova KM, Bogdanova VD. Ecological and hygienic factors and prevalence of the diseases of the circulatory system. Health Care of the Russian Federation. 2019; 63(2): 92-97. (In Russ).] doi: 10.18821/0044-197X-2019-63-2-92-97.
17. Tan Ch, Avasarala S, Liu H. Hexavalent chromium release in drinking water distribution systems: new insights into zerovalent chromium in iron corrosion scales. Environmental Science and Technology. 2020; 54(20): 13036-13045. doi: 10.1021/acs.est.0c03922.
18. Ревич Б.А., Харькова Т.Л., Кваша Е.А. Некоторые показатели здоровья жителей городов федерального проекта «Чистый воздух» // Анализ риска здоровью. — 2020. — №2. — С.16-27. [Revich BA, Khar'kova TL, Kvasha EA. Selected health parameters of people living in cities included into the «Clean air» federal project. Health Risk Analysis. 2020; 2: 16-27. (In Russ).] doi: 10.21668/health.risk/2020.2.02.

19. Шкуратова Т.А. Анализ и моделирование онкологической заболеваемости на основе устранения мультиколлинеарности и определения лагов: Автореф. дис. ... к.м.н. — Воронеж; 2006. [Shkuratova T.A. Analysis and modeling of cancer incidence based on elimination of multicollinearity and the determination of lags. [Autoreferat dissertation] Voronezh; 2006. (In Russ).]

20. Александров Ю.А. Основы радиационной экологии. Йошкар-Ола: Изд-во Марийского госуниверситета, 2007. — 268 с. [Aleksandrov YuA. Fundamentals of radiation ecology. Yoshkar-Ola: Izd. Mariyskogo gosuniversiteta, 2007. 268 p. (In Russ).]

21. Outcompeting *p*53-Mutant Cells in the Normal Esophagus by Redox Manipulation / Fernandez-Antoran D., Piedrafita G., Murai K., Ong S.H., Herms A., Jones P.H., Frezza C. Cell Stem Cell. 2019; 25(3): 329-341.e6. doi: 10.1016/j.stem.2019.06.011.

22. Регионы России. Социально-экономические показатели. М.: Росстат, 2016–2020. [Regions of Russia. Socio-economic indicators. M.: Rosstat, 2016–2020 (In Russ).]

23. Охрана окружающей среды в России / Приложение по регионам. М.: Росстат,2014–2020. [Environmental protection in Russia: Appendix (Information on the regions of Russia). M.: Rosstat, 2014–2018. (In Russ).]

24. Регионы России Основные характеристики субъектов РФ. М.: Росстат, 2020. [Regions of Russia. Major characteristics of subjects of the Russian Federation. M.: Rosstat, 2020. (In Russ).]

25. Карлберг К. Регрессионный анализ в Microsoft Excel / Пер. с англ. — М., 2017. — 400 с. [Carlberg C. Regression Analysis Microsoft Excel. M., 2017. 400 p. (In Russ).]

26. Социально-значимые заболевания населения России в 2019 году (статистические материалы). — М.: ЦНИИОИЗ МЗ РФ, 2020. — 76 с. [Socially significant diseases of the population of Russia in 2019. M.: TsNIIOIZ MZ RF, 2020. 76 p. (In Russ).]

27. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983. — 471 с. [Aivazyan SA, Yenyukov IS, Meshalkin LD. Applied statistics: Fundamentals of modeling and initial data processing. M.: Finansy i statistika, 1983. 471 p. (In Russ).]